

# Introduction to Galaxy and preprocessing of sequences

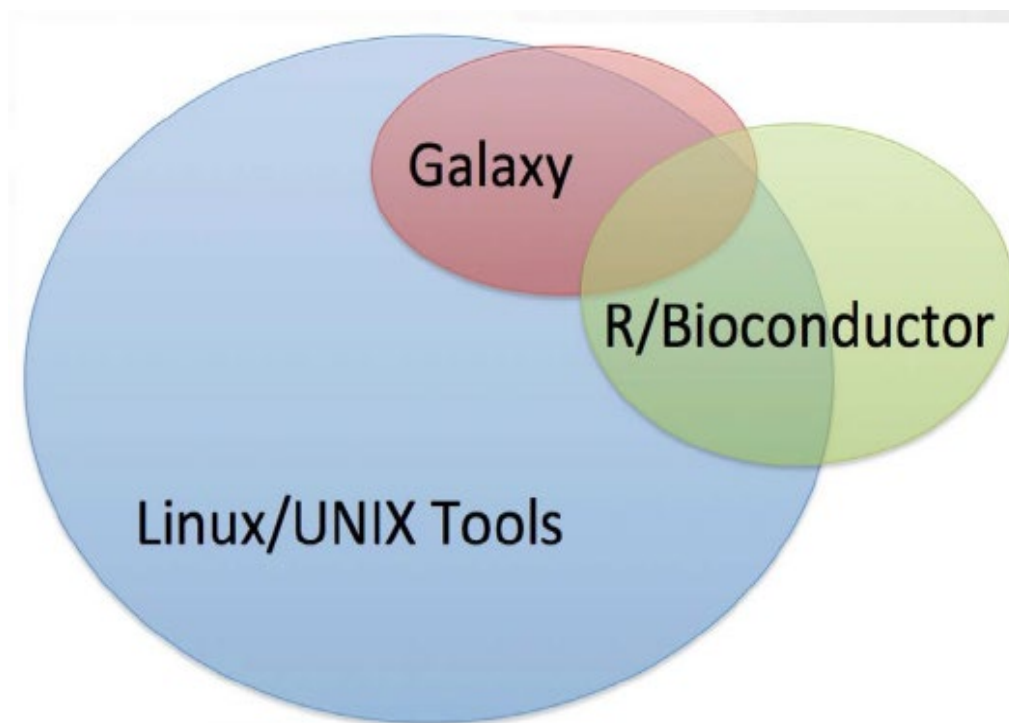
Bioinformatics Course UEB-VHIR  
November 2020

Ricardo Gonzalo<sup>1</sup>, Mireia Ferrer<sup>1</sup>, Àlex Sánchez<sup>1,2</sup>  
Berta Miró<sup>1</sup>, Angel Blanco<sup>1,2</sup>

1 Unitat d'Estadística i Bioinformàtica (UEB) VHIR

2 Departament de Genètica Microbiologia i Estadística, UB

Highly efficient and fast processing tools are required to handle large volume of datasets



# Introduction to Galaxy

# Introduction to Galaxy

## Galaxy Project

<https://galaxyproject.eu/>

- An open, web-based platform integrating many popular tools and resources for intensive biomedical research.
- **What can be done?**
  - Obtain data from many data sources like UCSC Table Browser, Biomart, WormBase, or your own data
  - Prepare data for further analysis by rearranging or cutting data columns, filtering data and many other options
  - Analyze data by finding overlapping regions, determining statistics, preprocessing NGS data and much more
  - Share data and workflows

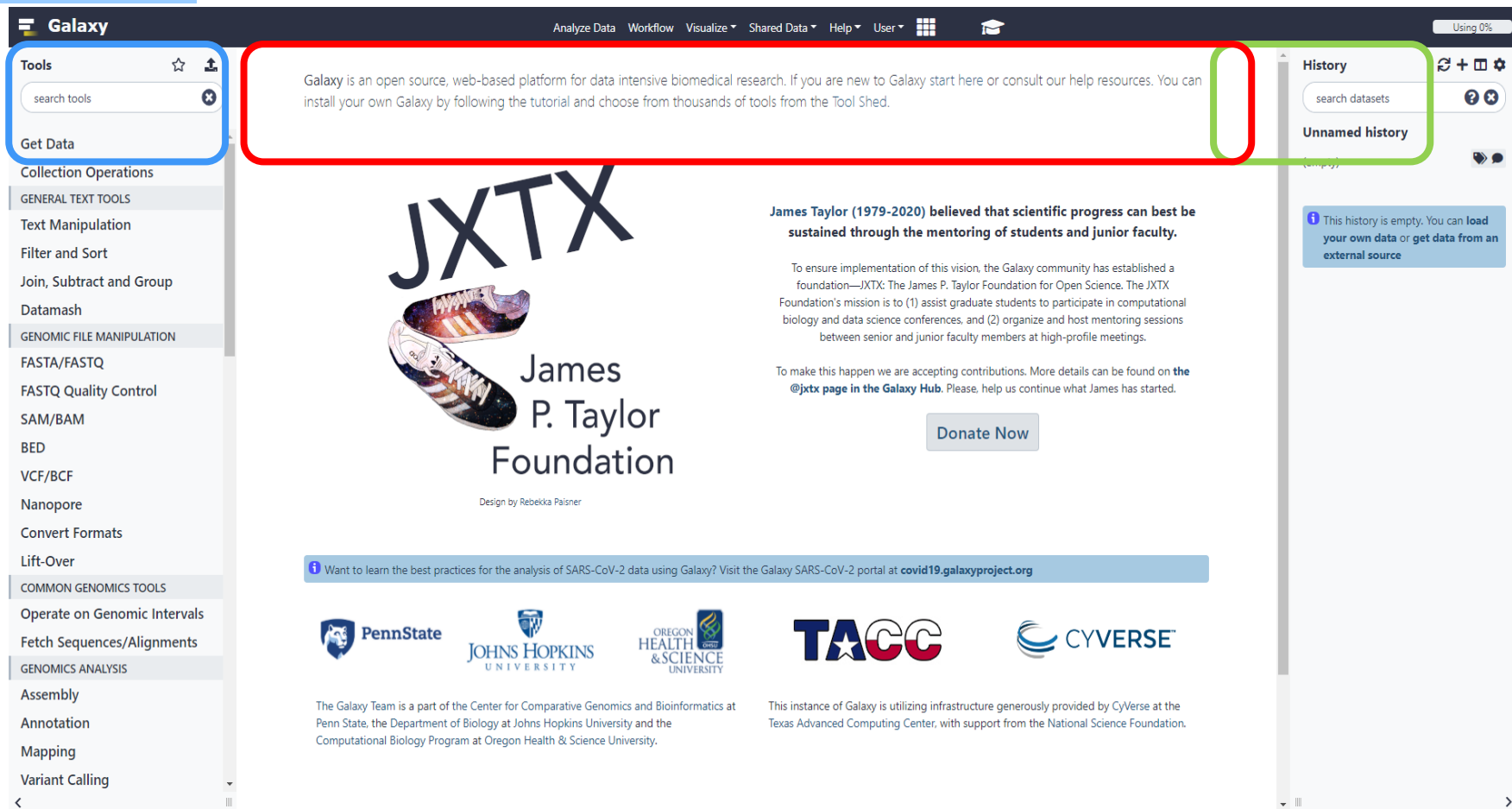
# Introduction to Galaxy

- The Galaxy page is divided into three panels:

**Tools** for uploading, processing and analysis

**Viewing panel**  
(menus, data, results)

**History** of analysis steps and datasets



The screenshot displays the Galaxy web interface with three main panels highlighted by colored boxes:

- Tools Panel (Left):** Contains a search bar and a list of tool categories including "Get Data", "Collection Operations", "GENERAL TEXT TOOLS", "Text Manipulation", "Filter and Sort", "Join, Subtract and Group", "Datamash", "GENOMIC FILE MANIPULATION", "FASTA/FASTQ", "FASTQ Quality Control", "SAM/BAM", "BED", "VCF/BCF", "Nanopore", "Convert Formats", "Lift-Over", "COMMON GENOMICS TOOLS", "Operate on Genomic Intervals", "Fetch Sequences/Alignments", "GENOMICS ANALYSIS", "Assembly", "Annotation", "Mapping", and "Variant Calling".
- Viewing Panel (Center):** Displays a message about Galaxy being an open source platform, followed by a large graphic for the "JXTX James P. Taylor Foundation". The graphic includes the text "JXTX" and "James P. Taylor Foundation" along with an image of sneakers. Below the graphic is a "Donate Now" button. At the bottom, there are logos for PennState, Johns Hopkins University, Oregon Health & Science University, TACC, and CyVerse.
- History Panel (Right):** Shows a search bar for datasets and a message stating "This history is empty. You can load your own data or get data from an external source".

# Introduction to Galaxy



## Tools for data analysis

### Get Data

- From databases (UCSC Table Browser, ...)
- From uploaded files
- From urls

### Text manipulation

### Filter and Sort

### Operate on Genomic Intervals

### FASTA manipulation

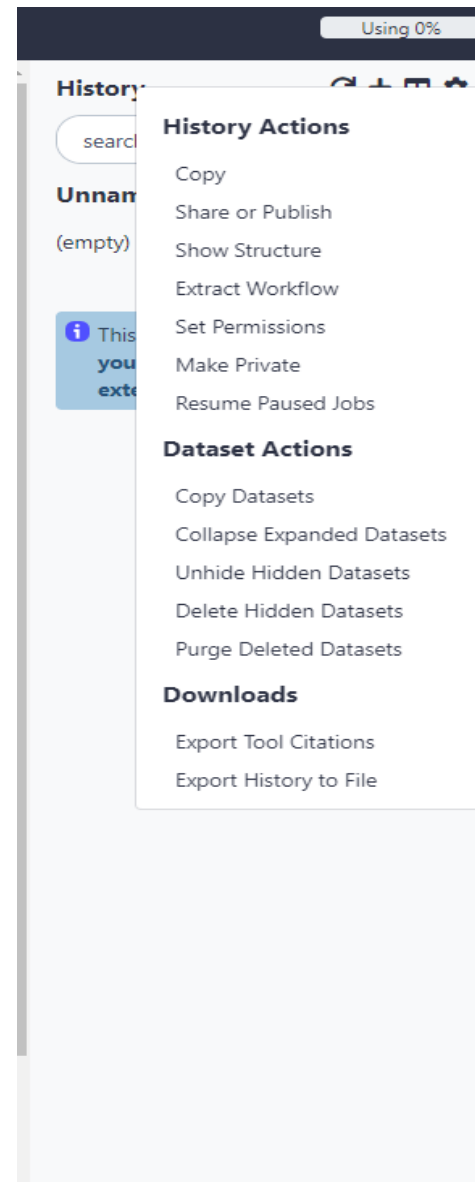
### NGS analysis

- QC
- Fastq file pre-processing
- Read Alignment / Mapping
- SAM tools

# Introduction to Galaxy


## Histories


List saved histories and shared histories.  
Work on Current History, create new, clone, share, create workflow, set permissions, show deleted datasets or delete history.




# Introduction to Galaxy

## Workflows


**Galaxy**

Analyze Data
 **Workflow**
 Visualize ▾
 Shared Data ▾
 Help ▾
 User ▾
 

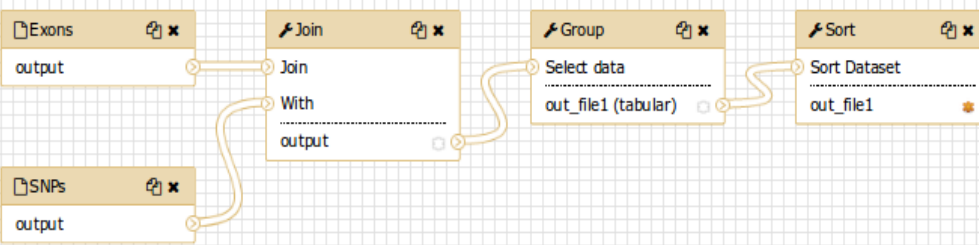
Using 2%

 Galaxy will be down for six hours beginning at 2:30 PM UTC, Tuesday, November 20 for filesystem maintenance.

**Tools**

**Inputs**  
[Get Data](#)  
[Send Data](#)  
[Lift-Over](#)  
[Collection Operations](#)  
[Text Manipulation](#)  
[Datamash](#)  
[Convert Formats](#)  
[Filter and Sort](#)  
[Join, Subtract and Group](#)  
[Fetch Alignments/Sequences](#)  
[NGS: QC and manipulation](#)  
[NGS: DeepTools](#)  
[NGS: Mapping](#)  
[NGS: RNA Analysis](#)  
[NGS: SAMtools](#)  
[NGS: BamTools](#)  
[NGS: Picard](#)  
[NGS: VCF Manipulation](#)

**Workflow Canvas | Coding Exon SNPs**



```

graph LR
    Exons[Exons output] --> Join[Join]
    SNPs[SNPs output] --> Join
    Join --> Group[Group]
    Group --> Sort[Sort]
    Sort --> out_file1[out_file1]
  
```

**Details**

**Edit Workflow Attributes**

**Name:**  
 Coding Exon SNPs

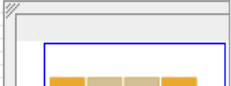
**Version:**  
 Version 1, 5 steps (active) ▾

**Tags:**

Apply tags to make it easy to search for and find items with the same tag.

**Annotation / Notes:**  
 Describe or add notes to workflow  
 Add an annotation or notes to a workflow; annotations are available when a workflow is viewed.

Workflows with all the analysis steps, allows user to repeat analysis using different datasets





# Introduction to Galaxy

## Register for a Galaxy account

This will only take a moment, and will allow all the work that you do to persist between sessions and allow you to name, save, share, and publish Galaxy histories, workflows, datasets and pages.

It allows you to store up to 250GB of data on this public server.

<https://usegalaxy.eu/>



# Introduction to Galaxy

## Training Infrastructure as a Service

We want to help you conduct your training seminars. You provide the training, we provide you training infrastructure *at no cost*.

Why use UseGalaxy.eu training infrastructure?

- Free
- Private queue, no wait times
- No Galaxy Maintenance
- No Galaxy Administration
- Official Galaxy Training Materials guaranteed to work



Simply fill out the infrastructure request form and we'll get back to you shortly.

[Find out more](#)

After registration in [European Galaxy server](#)



[https://usegalaxy.eu/join-training/ueb\\_bi2020](https://usegalaxy.eu/join-training/ueb_bi2020)

# Introduction to Galaxy

## Importing data into Galaxy

1. From **database** queries: eg. obtain a BED-formatted dataset of all RefSeq genes from platypus using the UCSC Table Browser.

### Get Data > UCSC Main – Table Browser tool

- Set genome, RefSeq Genes, and BED output format (send to Galaxy)

**Table Browser**

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve [Browser](#) for a description of the controls in this form, and the [User's Guide](#) for general information and sample queries. For more information on the biological function of your set through annotation enrichments, send the data to [GREAT](#). Send data to [GenomeSpace](#) for use with restrictions associated with these data. All tables can be downloaded in their entirety from the [Sequence and Annotation Download](#)

**clade:** Mammal **genome:** Platypus **assembly:** Feb. 2007 (ASM227v2/ornAna2)

**group:** Genes and Gene Predictions **track:** RefSeq Genes
 add custom tracks
track hubs

**table:** refGene describe table schema

**region:** ☒ genome ☐ position chrX5:870777-1056769
 lookup
define regions

**identifiers (names/accessions):**

**filter:**

**intersection:**

**correlation:**

**output format:** BED - browser extensible data
 ☒ Send output to ☒ Galaxy
 ☐ GREAT
 ☐ GenomeSpace

**output file:**  (leave blank to keep output in browser)

**file type returned:**
☒ plain text
 ☐ gzip compressed

### Output refGene as BED

- ☐ Include [custom track](#) header:

name= tb\_refGene

description= table browser query on refGene

visibility= pack

url=

### Create one BED record per:

- ☒ Whole Gene

☐ Upstream by 200 bases

☐ Exons plus 0 bases at each end

☐ Introns plus 0 bases at each end

☐ 5' UTR Exons

☐ Coding Exons

☐ 3' UTR Exons

☐ Downstream by 200 bases

Note: if a feature is close to the beginning or end of a chromosome and upstream/downstream

## Importing data into Galaxy







2. From a **File** on your computer / FTP file:



**Get Data > Upload File**








Download from web or upload from disk

Regular Composite Collection Rule-based

You added 1 file(s) to the queue. Add more files or click 'Start' to proceed.

Name	Size	Type	Genome	Settings	Status
 New File	72 b	fastqsang...  	----- Additional Sp... 		0% 
<p>You can tell Galaxy to download data from web by entering URL in this box (one per line). You can also directly paste the contents of a file.</p> <p><a href="http://chagall.med.cornell.edu/galaxy/rnaseq/GM12878_rnaseq1.fastqsanger">http://chagall.med.cornell.edu/galaxy/rnaseq/GM12878_rnaseq1.fastqsanger</a></p>					

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

 Choose local file  Choose FTP file  Paste/Fetch data  Pause  Reset  Start  Close

# Introduction to Galaxy

## Importing data into Galaxy

### 3. From a website:

#### Get Data > Upload File

- Copy this URL into the text-entry box:  
[https://zenodo.org/record/582600/files/mutant\\_R1.fastq](https://zenodo.org/record/582600/files/mutant_R1.fastq)

Regular Composite Collection

You added 1 file(s) to the queue. Add more files or click 'Start' to proceed.

Name	Size	Type	Genome	Settings	Status
New File	-	Auto-det...	unspecified (?)		

You can tell Galaxy to download data from web by entering URL in this box (one per line). You can also directly paste the contents of a file.

← 2. Paste file address in this box

1. click Paste/Fetch data

3. Start 4. Close

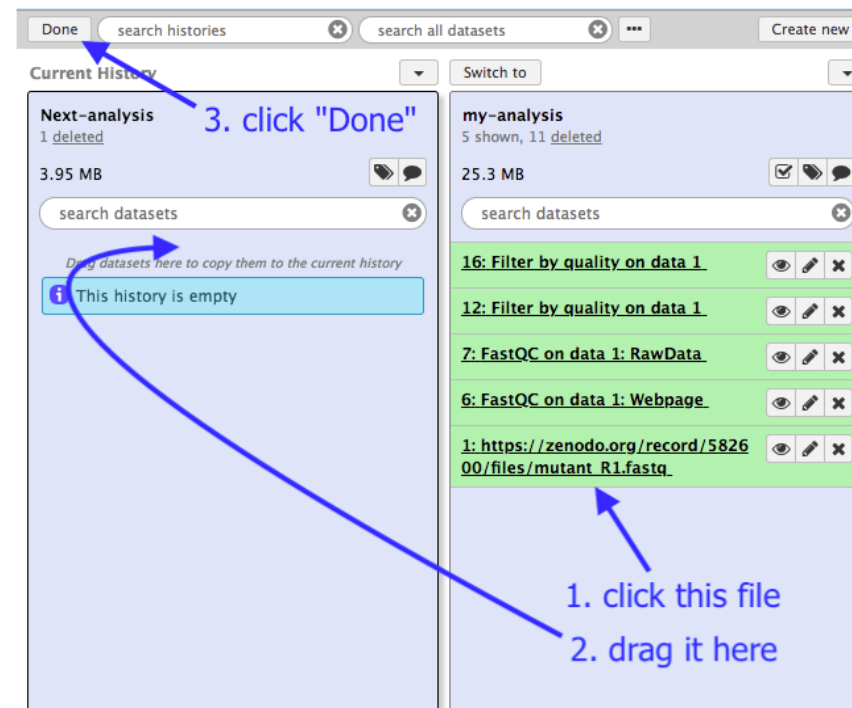
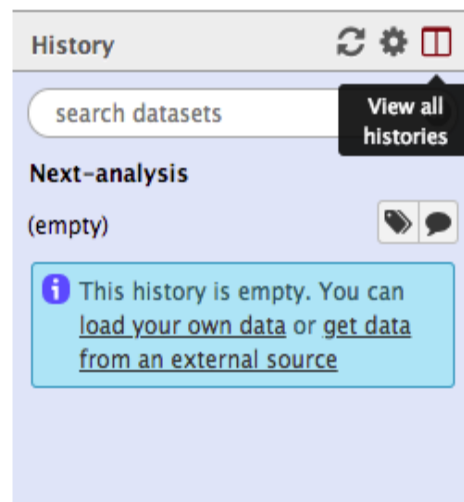
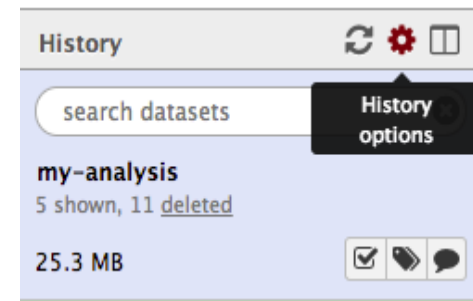
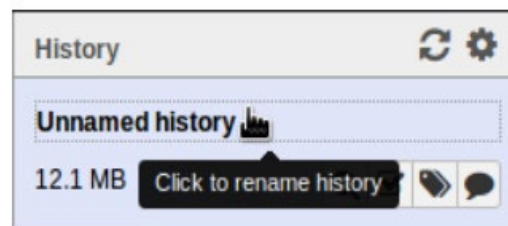
Type (set all): Auto-detect Genome (set all): unspecified (?)

Choose local file Paste/Fetch data Pause Reset Start Close

# Introduction to Galaxy

## Managing histories

- Name your current history
- Create new history and rename it
- Manage datasets and histories:
- View all histories
- Drag files between histories (**new history must be set to current**)



# Introduction to Galaxy

## Visualizing the dataset

- You can view file content clicking the eye icon in history.

The mutant\_R1.fastq file contains DNA sequencing reads from a bacteria, in FASTQ format:

```
@mutant-no_snps.gff-24960/1
AATGTTGTCACTTGGATTCAAATGACATTTTAAATCTAATTATTCATGAATCGAACTAGTACGAAATGCAATGAG
+
5??A9?BBBDDDBEDDBFF+FGHHIIHHHEIHHIIHIIAHDHIIHIG#IIHIFHHHFGIII*IHHHIIHFIIHGICI
@mutant-no_snps.gff-24958/1
CAAAGTCGTTGGTCATATAAAAAACCGCGTACAGTCAACTATAGATAACAATCAAGATAAACTCATGCACAGATTG
+
?A????@?DDDABDE9FGGGFGICFHIIIBGHIIIGICHHIFH=IHAFIHHHHHIFCIIIEIHAIFGIHIDDIHE
@mutant-no_snps.gff-24956/1
TATAAATTCAACTTTGCAACAGAACCATCTAATCTTCAACAACTGGCCCGTTTGTGAACTACTCTTTAATAAA
+
?????BBADD5DDDDDGFGCFEECFBBCIIII,IIHIICHIHIIIFHHHHHHIHHIIIIIIAHHIHHH5FHDHHHE
```

History

search datasets

my-analysis  
1 shown

3.95 MB

1: [https://zenodo.org/record/582600/files/mutant\\_R1.fastq](https://zenodo.org/record/582600/files/mutant_R1.fastq)

View data

## Create workflow from history

- From history options: Export 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

Workflow constructed from history 'prova'

Create Workflow Check all Uncheck all

### Tool

Upload File

This tool cannot be used in workflows

FastQC

☒ Include "FastQC" in workflow

Filter by quality

☒ Include "Filter by quality" in workflow

### History Items created

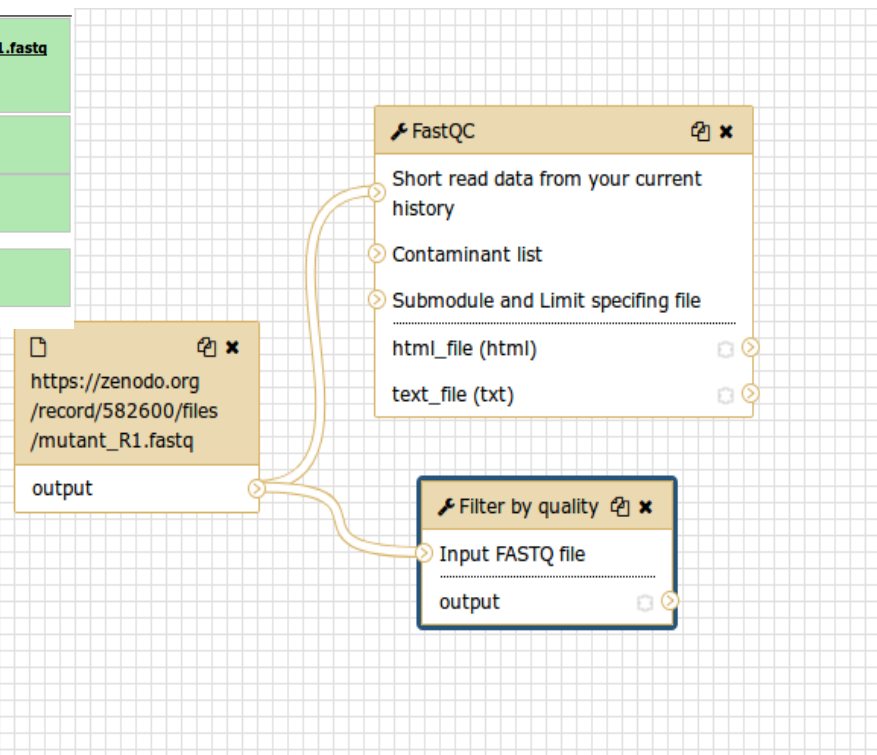
1 [https://zenodo.org/record/582600/files/mutant\\_R1.fastq](https://zenodo.org/record/582600/files/mutant_R1.fastq)

☒ Treat as input dataset <https://zenodo.org/record/58>

2 FastQC on data 1: Webpage

3 FastQC on data 1: RawData

4 Filter by quality on data 1





# Introduction to Galaxy

- <https://galaxyproject.org/learn/>

## Learn Galaxy

There are many approaches to learning how to use Galaxy. The most popular is probably to just dive in and use it. Galaxy is simple enough to use that you can do many analyses just by exploring the interface. However, you may miss much of the power this way.

Have you created or know of a resource that is useful for teaching with Galaxy? Then please share it! This will help others and also help get the word out about your resource. Use [this Google form](#) to describe your resource. **Also:** consider joining Galaxy Training Network and contributing your tutorial as described [here](#)!

## Tutorials by Galaxy Training Network

Thanks to a large group of wonderful contributors there is a constantly growing set of tutorials maintained by the Galaxy Training Network. These include:

### Introductory Tutorials

- [Introduction to Galaxy Analyses](#)
- [Data Manipulation](#)
- [User Interface and Features](#)

### Scientific Analyses

- [Assembly](#)
- [Computational chemistry](#)
- [Ecology](#)
- [Epigenetics](#)
- [Genome Annotation](#)
- [Imaging](#)
- [Metabolomics](#)
- [Metagenomics](#)
- [Proteomics](#)
- [Sequence analysis](#)
- [Statistics and machine learning](#)
- [Transcriptomics](#)
- [Variant Analysis](#)

## Material

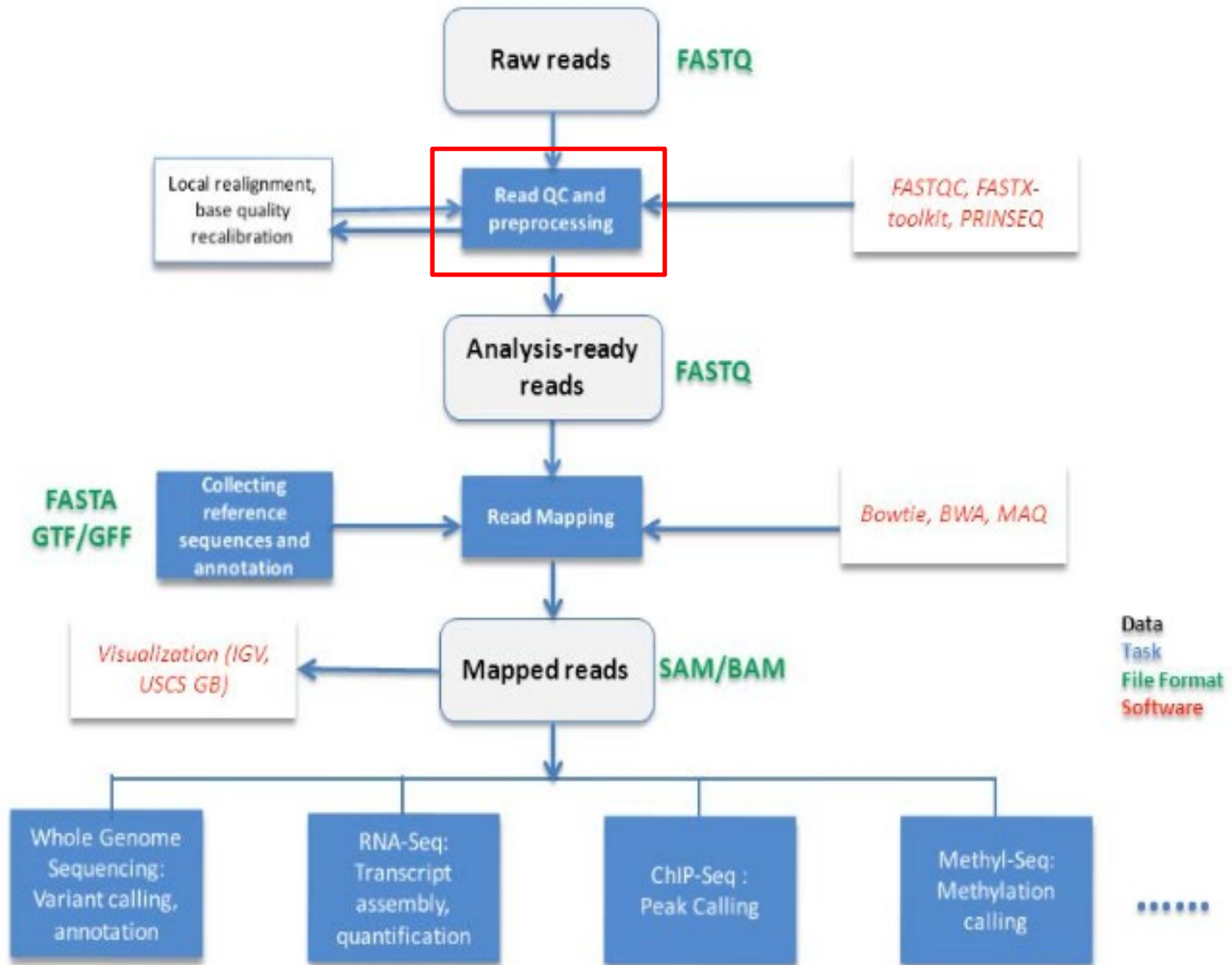
Q Search x

Lesson	Slides	Hands-on	Input dataset	Workflows	Galaxy tour	Galaxy instances
Introduction to metagenomics						
16S Microbial Analysis with mothur (extended)						
16S Microbial Analysis with mothur (short)						
Analyses of metagenomics data - The global picture						
Antibiotic resistance detection 						
Metatranscriptomics analysis using microbiome RNA-seq data 						
Metatranscriptomics analysis using microbiome RNA-seq data (short) 						

# **First steps in NGS analysis with Galaxy:**

## **Quality Control and preprocessing of reads**

# Steps in NGS analysis



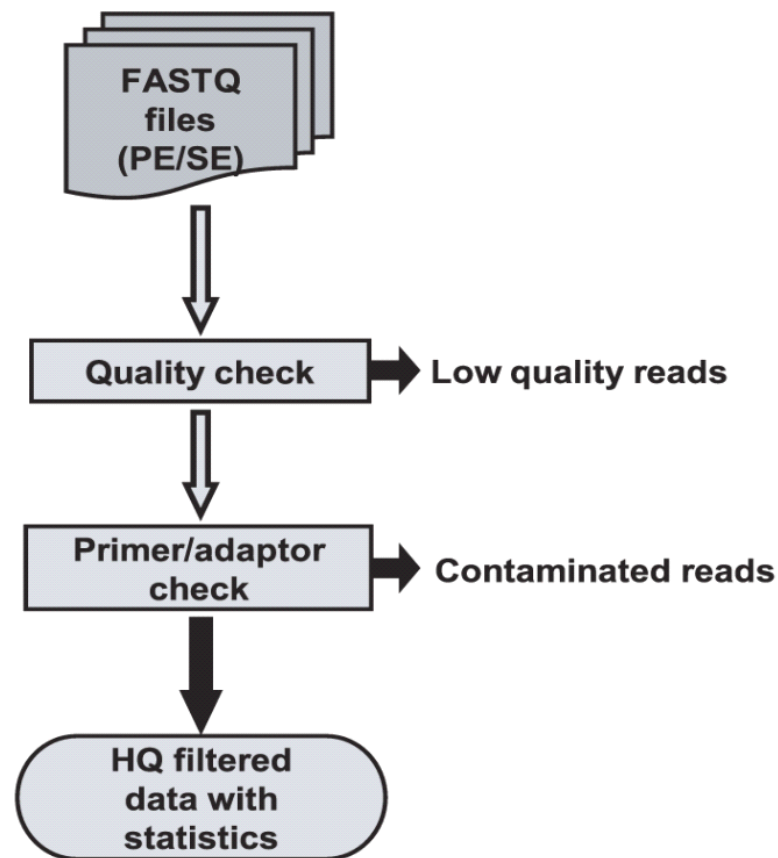
# Steps in NGS analysis

## Quality Control

- Quality Control analysis of sequence data is extremely important for meaningful downstream analysis

- To analyze problems in quality scores/ statistics of sequencing data
- To check whether further analysis with sequence is possible
- To remove redundancy (filtering)
- To remove low quality reads from analysis
- To remove adapter contamination

Highly efficient and fast processing tools are required to handle large volume of datasets



## FastQC tool

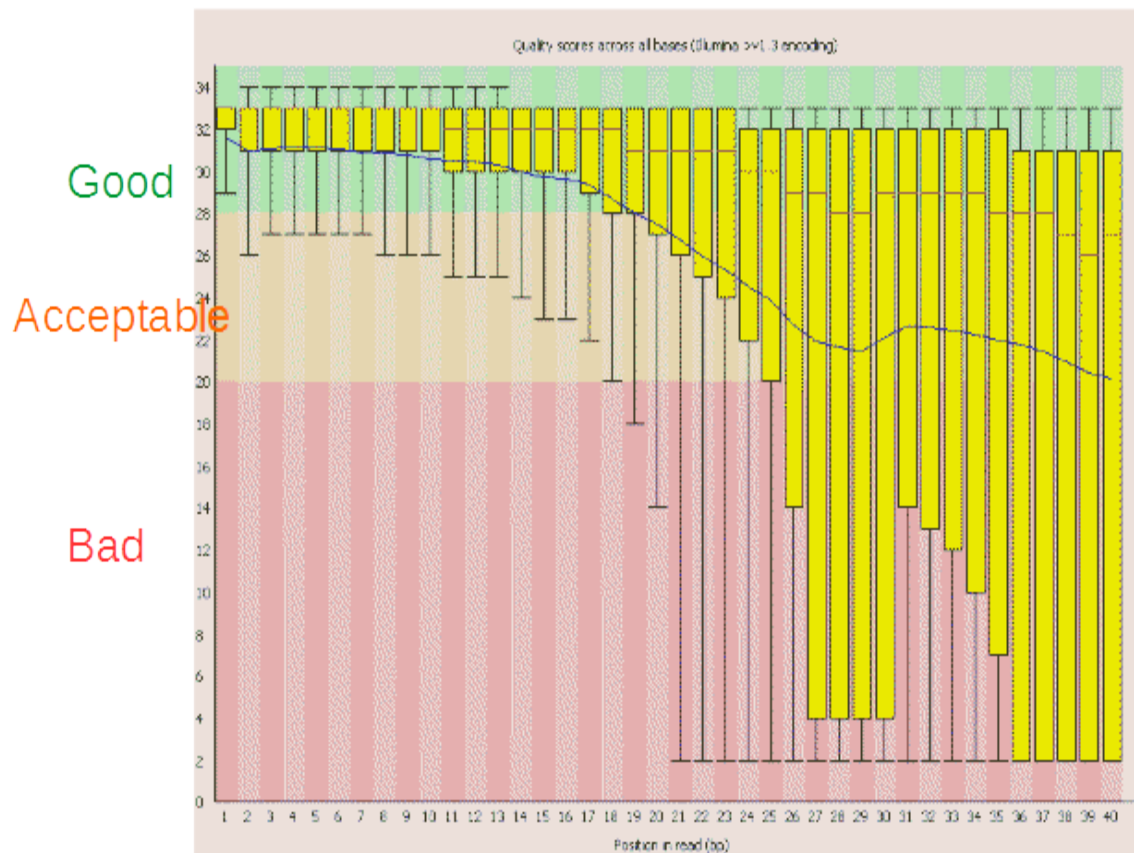
<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

- Basic statistics
- Quality- Per base position
- Per Sequence Quality Distribution
- Nucleotide content per position
- Per sequence GC distribution
- Per base GC distribution
- Per base N content
- Length Distribution
- Overrepresented/ duplicated sequences
- K-mer content

# Quality Control

## FastQC

### Per base sequence quality (Boxplot)



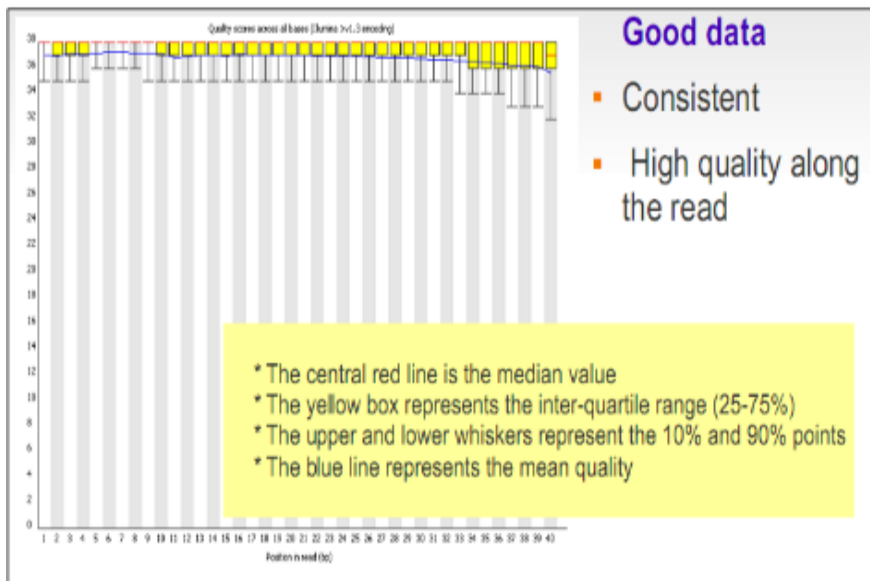
shows an overview of the range of quality values across all bases at each position in the FastQ file

Y axis- Quality Score  
X axis- Base position

# Quality Control

## FastQC

### Per base sequence quality (Boxplot)

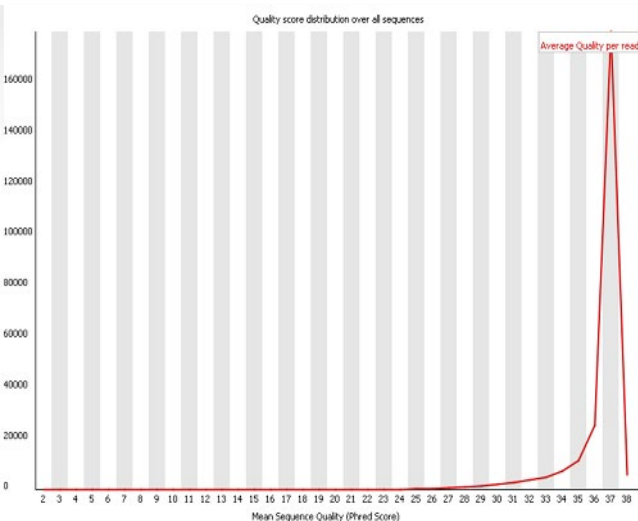


# Quality Control

## FastQC

### Per sequence quality scores

allows you to see if a subset of your sequences have universally low quality values.



#### Good data

- Most are high-quality sequences



#### Bad data

- Not uniform distribution

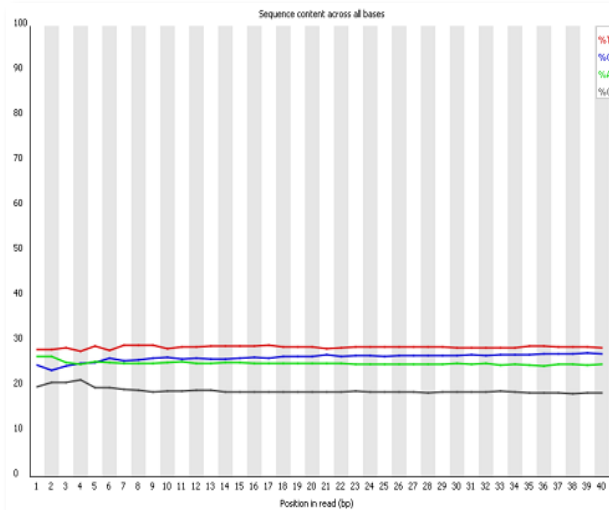


# Quality Control

## FastQC

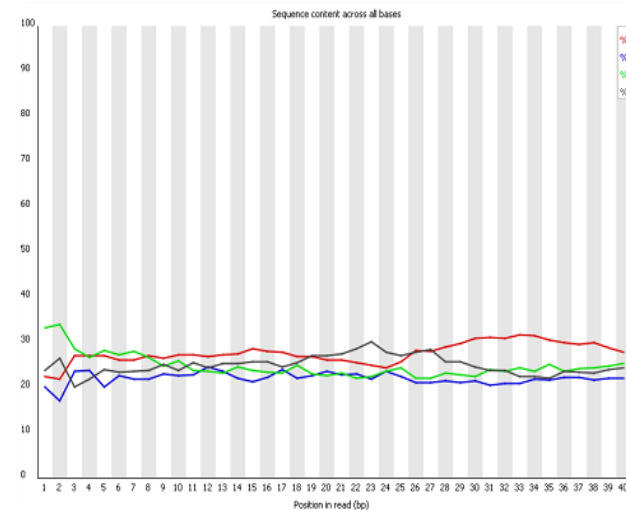
### Per base sequence content

proportion of each base  
position in a file for  
which each of the four  
normal DNA bases has  
been called



#### Good data

- Smooth over length
- Organism dependent (GC)



#### Bad data

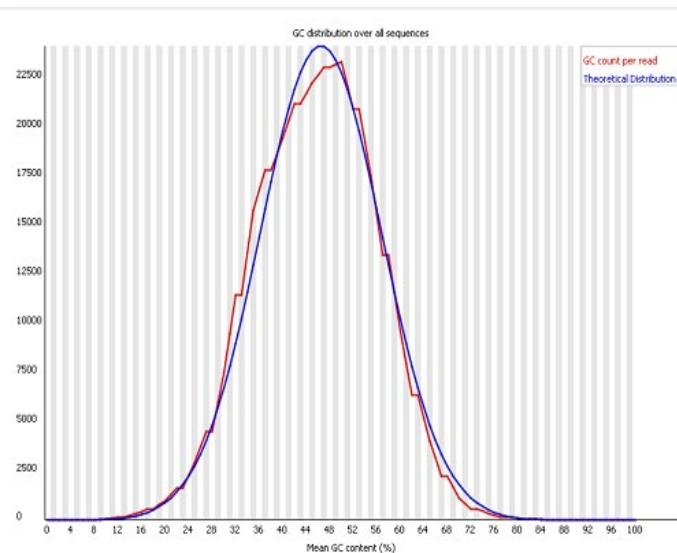
- Sequence position bias

# Quality Control

## FastQC

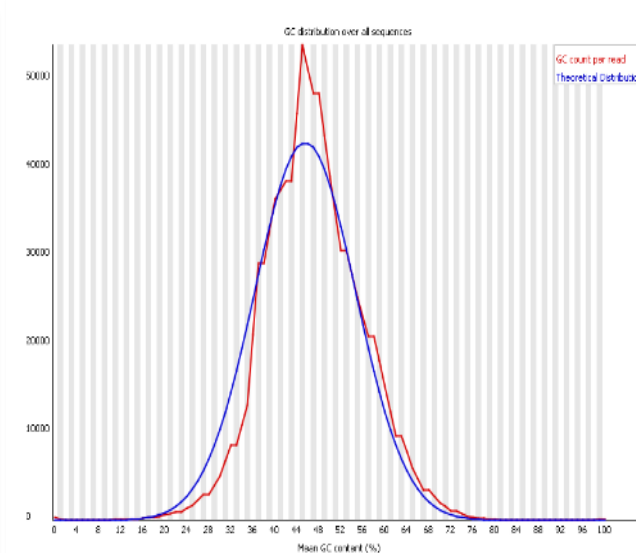
### Per sequence GC content

measures the GC content across the whole length of each sequence in a file and compares it to a modelled normal distribution of GC content



#### Good data

- Fits with the expected
- Organism dependent



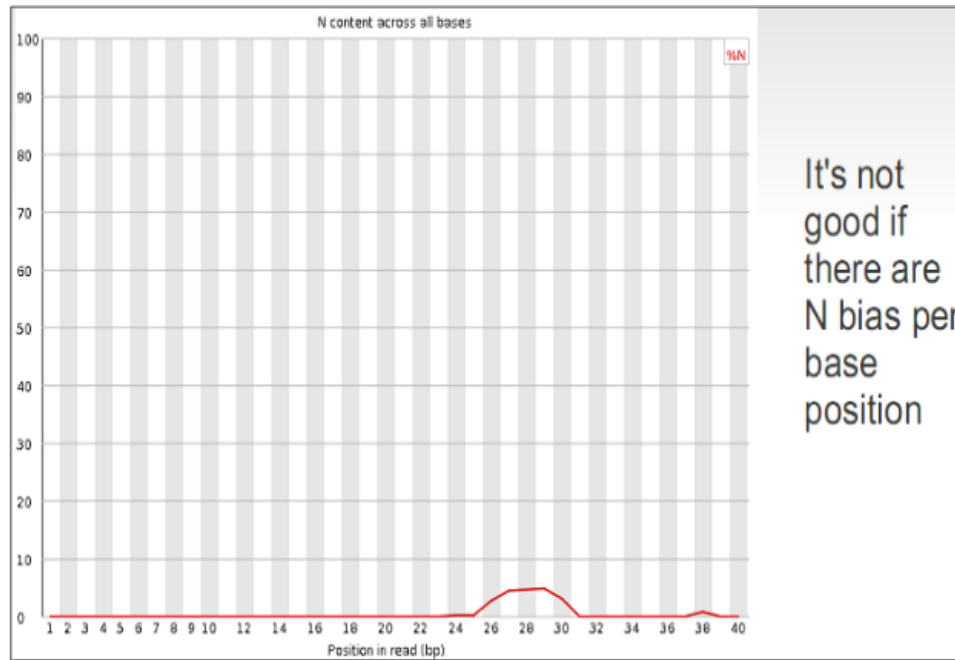
#### Bad data

- It does not fit with expected
  - Organism dependent
- Library contamination?

# Quality Control

## FastQC

### Per base N content

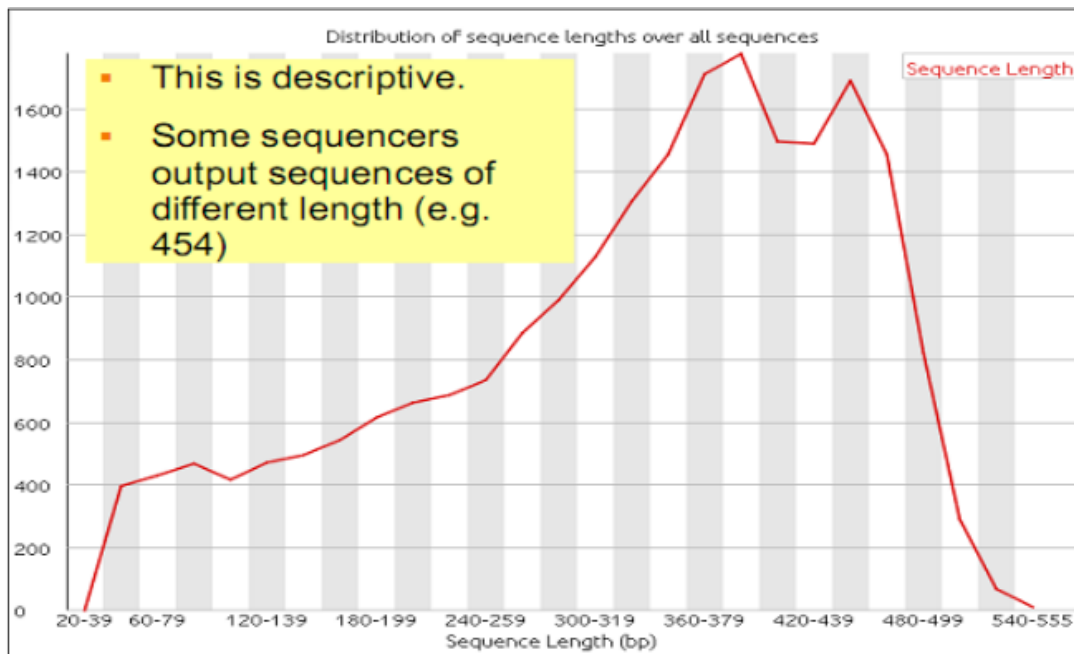


If a sequencer is unable to make a base call with sufficient confidence then it will normally substitute an N rather than a conventional base call. It plots out the percentage of base calls at each position for which an N was called.

# Quality Control

## FastQC

### Sequence length distribution

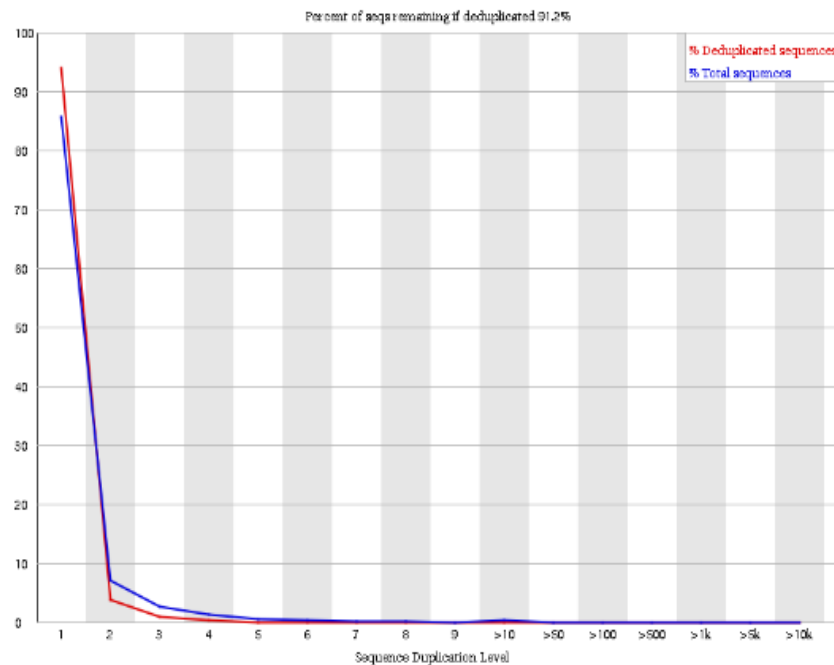


In many cases it will produce a simple graph showing a peak only at one size, but for variable length FASTQ files, it will show the relative amounts of each different size of sequence fragment.

# Quality Control

## FastQC

### Sequence duplication level



Counts the degree of duplication for every sequence. Too many duplicate regions in the sequence may indicate contamination or technical problems

## FastQC

### Overrepresented sequences

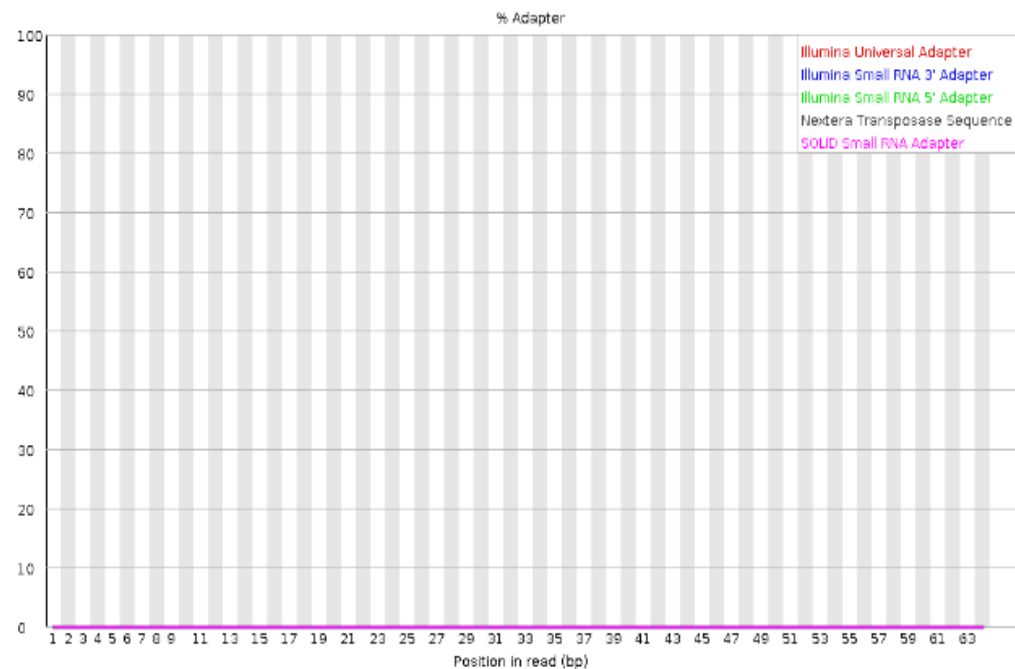
Sequence	Count	Percentage	Possible Source
AAGATCCGAGTCGTCCGGAAATCCATTGCCCGTGTCTCACAGTTATTAA	432	0.43585733743631133	No Hit
AGATCCGAGTCGTCCGGAAATCCATTGCCCGTGTCTCACAGTTATTAA	335	0.33799122231750994	No Hit
TGGCAGAAGTAGAGCAGAAGAAGAAGCGGACCTTCCGCAAGTTCACCTAC	250	0.25223225546082834	No Hit
CAGAAGTAGAGCAGAAGAAGAAGCGGACCTTCCGCAAGTTCACCTACCGC	237	0.23911617817686526	No Hit
GTAGAGCAGAAGAAGAAGCGGACCTTCCGCAAGTTCACCTACCGCGCGT	223	0.22499117187105888	No Hit
AAGAAATCTGACCCGGTCGTCTCGTACCGGAGACGGTCAGTGAAGAGTC	204	0.2058215204560359	No Hit
AAGTAGAGCAGAAGAAGAAGCGGACCTTCCGCAAGTTCACCTACCGCGGC	151	0.1523482822983403	No Hit
CACCTGGAGATCTGCCTGAAGGACCTGGAGGAGGACCACGCCTGCATCCC	147	0.14831256621096706	No Hit
TCTGCCTGAAGGACCTGGAGGAGGACCACGCCTGCATCCCATCAAGAAA	146	0.14730363718912376	No Hit

Lists all of the sequence which make up more than 0.1% of the total. Finding that a single sequence is very overrepresented in the set either means that is highly biologically significant, or that the library is contaminated. For each overrepresented sequence it will look for matches in a database of common contaminants.

# Quality Control

## FastQC

### Adapter content



Does a generic analysis of all the Kmers in the library to find those that don't have even coverage through the length of the reads.

# Quality Control

## FastQC

- Good (Illumina) quality data:

[http://www.bioinformatics.babraham.ac.uk/projects/fastqc/good\\_sequence\\_short\\_fastqc.html](http://www.bioinformatics.babraham.ac.uk/projects/fastqc/good_sequence_short_fastqc.html)

- Bad (Illumina) quality data:

[http://www.bioinformatics.babraham.ac.uk/projects/fastqc/bad\\_sequence\\_fastqc.html](http://www.bioinformatics.babraham.ac.uk/projects/fastqc/bad_sequence_fastqc.html)



# Quality Control

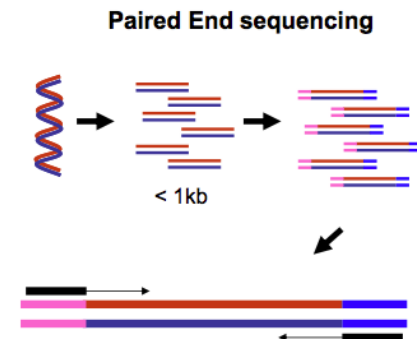
## Your turn!

- We will analyze exome sequencing data from a study that aimed to identify genetic variants associated to a disease. The data comes from paired-end sequencing, each file corresponding to the forward or reverse, respectively:

[https://zenodo.org/record/3243160/files/proband\\_R1.fq.gz](https://zenodo.org/record/3243160/files/proband_R1.fq.gz)

[https://zenodo.org/record/3243160/files/proband\\_R2.fq.gz](https://zenodo.org/record/3243160/files/proband_R2.fq.gz)

**Paired-end data:** a single physical piece of DNA/RNA is sequenced from two ends and so generates two reads. These can be represented as separate files (two fastq files with first and second reads) or a single file where reads for each end are interleaved.



# Quality Control

## Your turn!

- We will analyze exome sequencing data from a study that aimed to identify genetic variants associated to a disease. The data comes from paired-end sequencing, each file corresponding to the forward or reverse, respectively:

[https://zenodo.org/record/3243160/files/proband\\_R1.fq.gz](https://zenodo.org/record/3243160/files/proband_R1.fq.gz)

[https://zenodo.org/record/3243160/files/proband\\_R2.fq.gz](https://zenodo.org/record/3243160/files/proband_R2.fq.gz)

1. Create a new history and name it as you want (eg. Practica1)
2. Upload the fastq files into Galaxy from the urls copied above
3. Update the attributes of the two datasets (pencil icon):
  - a) Rename the datasets to “sample-f.fq.gz” and “sample-r.fq.gz”, respectively.
  - b) Check data type is set to “fastqsanger”
  - c) Associate the dataset with the human hg38 genome in the Database/Build field.
4. Run a quality control on each dataset using the FastQC tool.
  - a) What is the length of reads?
  - b) Are sequences of good quality? Any adapter that should be removed?
5. What would be the next step in the analysis workflow?

# Quality Control

## Preprocessing of raw data

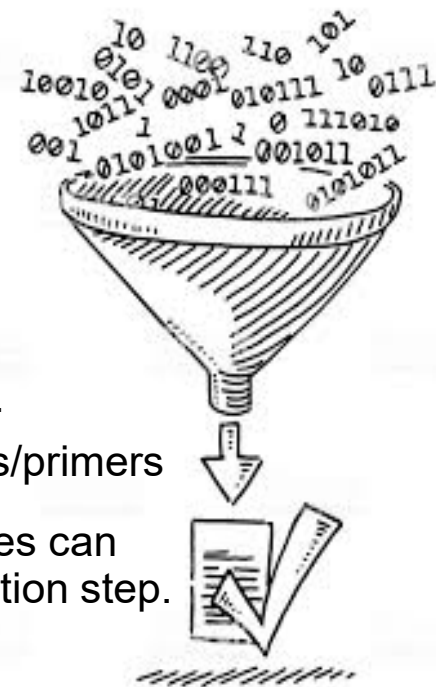
Based on the information provided by the QC graphs, the sequences may be treated to reduce bias in downstream analysis:

### •Filtering sequences

- with low mean quality score
- too short
- with too many ambiguous (N) bases
- based on their GC content
- Biological contamination: polyA-tails, rRNA or mtDNA sequences,...
- Technical contamination: PhiX internal control sequences, adapters/primers
- Removing duplicate reads is not advised since high expressed genes can have genuine duplicate reads that are not due to the PCR amplification step.

### •Cutting/Trimming sequences

- from low quality score regions
- beginning/end of sequence
- removing adapters, primers



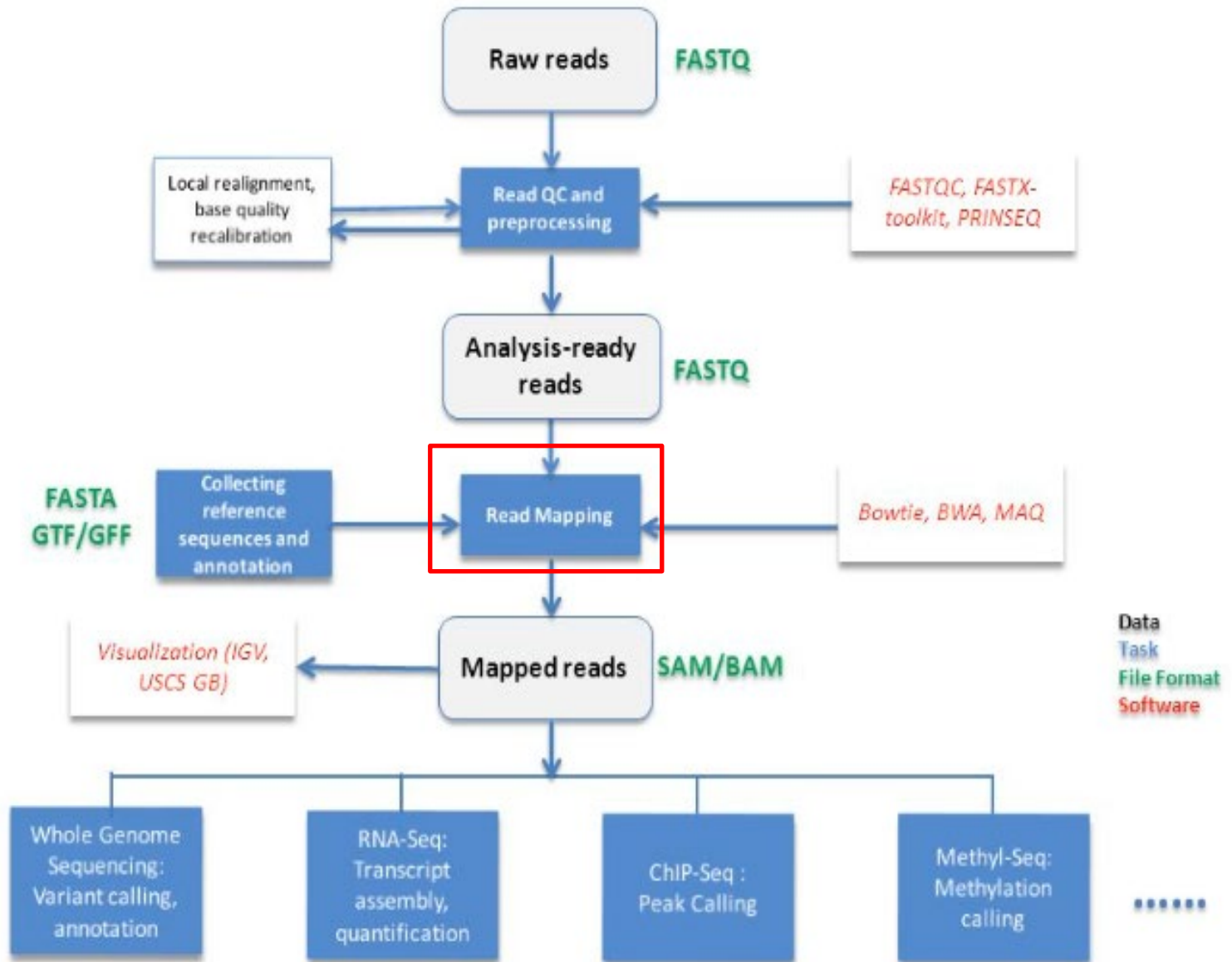
When preprocessing paired-end data coming from separate files, this must be taken into account so that reads are preprocessed “in pairs”

# Quality Control

## Your turn!

6. Trim the reads in each dataset using **Cutadapt** tool. Set the parameters:
  - a) Paired-end data  
File 1: sample-f (forward)  
File 2: sample-r (reverse)
  - b) Determine from the FastQC boxplot where the quality of the reads begins to drop off sharply. Calculate how many bases have to be trimmed from the end and use that number as the Offset from 3' end.
  - c) Output options: Report=yes
7. Inspect the results:
  - a) How many datasets do we get? Rename them to sample-f-trim / sample-r-trim, respectively. What is their format?
  - b) Do they have the same number of reads?
8. Re-run FastQC on the trimmed data, and inspect the new FastQC report. Has the sequence quality been improved?
9. Convert your analysis history into a workflow
10. What would be the next step in the analysis workflow?

# Steps in NGS analysis



## References and resources

### Bibliography

Goecks J, Nekrutenko A, Taylor J; Galaxy Team. *Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences*. *Genome Biol.* 2010

### Links and resources

#### Galaxy tutorials

<https://galaxyproject.github.io/training-material/topics/variant-analysis/>

<https://galaxyproject.github.io/training-material/topics/variant-analysis/tutorials/exome-seq/tutorial.html>