Vall d'Hebron
Institut de Recerca

DATA FORMATS IN NGS. INTRODUCTION TO GALAXY

Bioinformàtica per a la Recerca Biomèdica

Ricardo Gonzalo Sanz
ricardo.gonzalo@vhir.org

Vall d'Hebron Hospital
Generalitat de Catalunya Departament de Salut
Institut Català de la Salut
UAB Universitat Autònoma de Barcelona
Generalitat de Catalunya Departament d'Economia i Coneixement
Institució CERCA Centres de Recerca de Catalunya
Instituto de Salud Carlos III

Vall d'Hebron
Institut de Recerca

1. Data formats used in NGS
2. Introduction to Galaxy

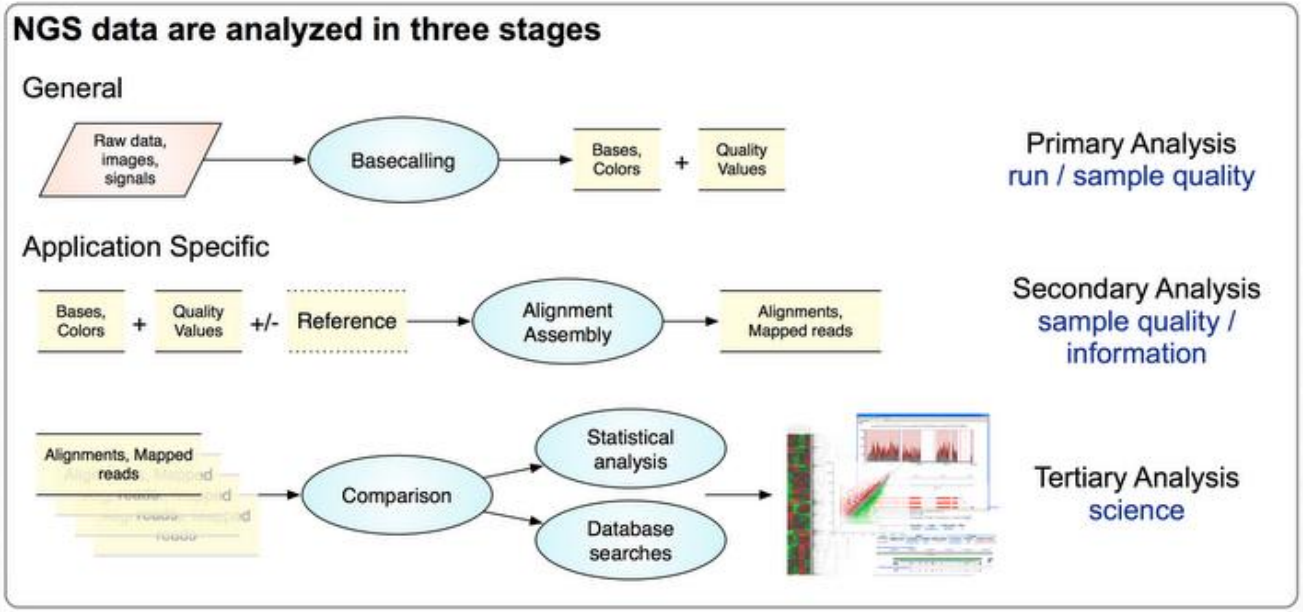
Vall d'Hebron
Institut de Recerca

1. **Data formats used in NGS**
2. **Introduction to Galaxy**

1. Data formats used in NGS

There are many different types of file formats depending on:

- Type of information they contain
  - Raw Sequence files
  - Co-ordinate files
  - Parameter files
  - Annotation files
  - Metadata files
- Sequencing platform
- Analysis stage
- Data source



1. Data formats used in NGS

- Formats are designed to hold sequence data and other information about sequence
- All Sequence formats are ASCII text containing sequence ID, Quality Scores, Annotation details, comments, and other descriptions about sequence

1. Data formats used in NGS

FASTA format

- FASTA format is a text-based format for representing either nucleotide sequences or peptide sequences, in which base pairs or amino acids are represented using single-letter codes
- Header line starts with “>” followed by a sequence ID, and followed by lines of sequence data

```
>NG_016798.2 Homo sapiens DNA polymerase alpha 1, catalytic subunit (POLA1), RefSeqGene on
chromosome X
CCCTCAGTTGGTGCCAGTAACTGTTGTTCCCTTCTTTGTGTCCATTTGTACTCAATGTTTACCTCCCACT
TATAAGTGAGAACATGTGGTATTTGGTTTTCTGTTCCTATGTTAGTTCGCTTAGGATAATGGCTTCCAAC
TCCATCCATGTTGCTGCAGACGTGATCTCATTCTTTTTTTTTTTTTTTTTTTGAGACAGAGTCGTGCTC
TGTCGCCCAGGCTGGAGTGCAGTGGTGCGATCTCGGCTCACTGCAACCTCTGCCTCCTGGGTTGAAGTGA
CTCTCCTGGTTCAGCCTCCTGAGTAGCTAGGATTATAGGTGCCCGCCACCATGCCTGGCTAATTTTTGTA
TTTTTAGTAGAGATGGGGTTTTGCCATGTTGGCCAGGCTGATCTCGAACTCCTGACCTCAGGTGATCTGC
CCACCCAGAGTGGCCTCCCAAAGTGTTTGGAATACAGGCGTGAGCCACTGCACCTGGTTTCTTTTTTATG
GCTGTAAATTAGTTCCACCATTGTGGAAGACAGTGTGGTGATTCACATAAAAGTAGAAGTCTAAGAATCA
AACCCTAAGTCTGACTCTACCTGAGTCTTTAATCCTTCCAATATAATATTAAAGAGGACAAATTATAAAC
AAAAAGAGTCTATAATTCTATCATCCTGGCAAAATATACTCCATTTTGCATATTGCTTTCAGGTAAATAA
```

```
>NP_001365232.1 DNA polymerase alpha catalytic subunit isoform 3 [Homo sapiens]
MAPVHGDDCEIGASALSDSGSFVSSRARREKKSKKGRQEALERLKKAKAGEKYKYEVEDFTGVYEEVDEE
QYSKLVQARQDDDWIVDDDGIGYVEDGREIFDDDLEDDALDADEKGKDGKARNKDKRNVKKLAVTKPNNI
KSMFIACAGKKTADKAVDLSKDGLLGDILQDLNTETPQITPPPVMILKKKRSIGASPNPFSVHTATAVPS
GKIASPVSRKEPPLTPVPLKRAEFAGDDVQVESTEEEQESGAMEFEDGDFDEPMEVEEVDLEPMAAKAWD
KESEPAEEVKQEADSGKGTVSYLGSFLPDVSCWDIDQEGDSSFSVQEVQVDSSHLPLVKGADEEQVFHFY
WLDAYEDQYNQPGVVFLFGKVWIESAETHVSCCVMVKNIERTLYFLPREMKIDLNTGKETGTPISMKDVY
EEFDEKIATKYKIMKFKSKAEMPQLPQDLKGETFSHVFGTNTSSLELFLMNRKIKGPCWLEVKSPQLLNQ
PVSWCKVEAMALKPDLVNVIKDVSPPPLVVMAFSMKTMQNAKNHQNEIIAMAALVHHSFALDKAAPKPPF
QSHFCVVSKPKDCIFPYAFKEVIEKKNVKVEVAATERTLLGFFLAKVHKIDPDIIVGHNIYGFELEVLLQ
RINVCKAPHWSKIGRLKRSNMPKLGGRSGFGERNATCGRMICDVEISAKELIRCKSYHLSELVQQILKTE
```

1. Data formats used in NGS

Vall d'Hebron
Institut de Recerca

FASTQ format

- Output of most actual sequencing platforms for raw data
- A text-based format for storing both a **nucleotide sequence** and its corresponding **quality scores**
- Standard file extension for a FASTQ file are .fq and .fastq
- FASTQ files are uncompressed and quite large because they contain the following information for every single sequencing read.
- Compressed files are also possible: fastq.gz

1. Data formats used in NGS

FASTQ format

- File structure. 4 lines:
  - @ followed by the read ID and possibly information about the sequencing run
  - sequenced bases
  - + (perhaps followed by the read ID again, or some other description)
  - quality scores for each base of the sequence (ASCII-translated Phred scores)

```
@Seq description
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*(((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

Vall d'Hebron Institut de Recerca

FASTQ format

Phred Scores

- Sequencing systems assign quality scores to each peak, that represents the error probability that an individual base call is incorrect.
- Phred scores provide log(10)-transformed error probability values:

If p is probability that the base call is wrong the Phred score is $Q = .10 \cdot log_{10} p$

PHRED Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90 %
20	1 in 100	99 %
30	1 in 1000	99.9 %
40	1 in 10000	99.99 %
50	1 in 100000	99.999 %

1. Data formats used in NGS

- The base calling (A, T, G or C) is performed based on Phred scores.
- Ambiguous positions with Phred scores <= 20 are labeled with N.
- To assign each base a unique score identifier (instead of numbers of varying character length), Phred scores are typically represented as ASCII characters.
- Different sequencing platforms may use different ASCII ranges for Phred encoding

Description	ASCII characters		Quality score	
	Range	Offset	Type	Range
Solexa/early Illumina (1.0)	59 to 126 (; to ~)	64	Solexa	-5 to 62
Illumina 1.3+	64 to 126 (@ to ~)	64	Phred	0 to 62
Sanger standard/Illumina 1.8+	33 to 126 (! to ~)	33	Phred	0 to 93

Base call quality scores are represented with the Phred range. Different Illumina (formerly Solexa) versions used different scores and ASCII offsets. Starting with Illumina format 1.8, the score now represents the standard Sanger/Phred format that is also used by other sequencing platforms and the sequencing archives.

1. Data formats used in NGS

SAM / BAM formats

- The **Sequence Alignment/Map (SAM)** format is a generic nucleotide alignment format that describes the alignment of sequencing reads to a reference.
- SAM files typically contain:
  - a short header section with information about the genomic loci of each read
  - a very long alignment section where each row represents a single read alignment.
  - Each alignment line has 11 mandatory fields for essential alignment information such as mapping position, and variable number of optional fields for flexible or aligner specific information

https://samtools.github.io/hts-specs/SAMv1.pdf

1. Data formats used in NGS

Vall d'Hebron
Institut de Recerca

Mandatory Alignment Section Fields

Position	Field	Description
1	QNAME	Query template (or read) name
2	FLAG	Information about read mapping (see next section)
3	RNAME	Reference sequence name. This should match a @SQ line in the header.
4	POS	1-based leftmost mapping position of the first matching base. Set as 0 for an unmapped read without coordinate.
5	MAPQ	Mapping quality of the alignment. Based on base qualities of the mapped read.
6	CIGAR	Detailed information about the alignment (see relevant section).
7	RNEXT	Used for paired end reads. Reference sequence name of the next read. Set to "=" if the next segment has the same name.
8	PNEXT	Used for paired end reads. Position of the next read.
9	TLEN	Observed template length. Used for paired end reads and is defined by the length of the reference aligned to.
10	SEQ	The sequence of the aligned read.
11	QUAL	ASCII of base quality plus 33 (same as the quality string in the Sanger FASTQ format).
12	OPT	Optional fields (see relevant section).

1. Data formats used in NGS

Vall d'Hebron
Institut de Recerca

```
@HD VN:1.6 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

1. Data formats used in NGS

Vall d'Hebron
Institut de Recerca

SAM / BAM formats

- A **BAM file** is a binary version of a SAM file.
- Both contain identical information about reads and their mapping.
- A BAM file requires a header but a SAM file may not have one.
- Many operations (such as sorting and indexing) work only on BAM files.
- For almost any application that requires SAM input, this can be created on the fly from a BAM,
- BAM files take up much less space than SAM files.
- For archiving purposes, keep only the BAM file. The SAM file can easily be regenerated (if ever needed).

Vall d'Hebron
Institut de Recerca

BED / GFF / GTF formats

- Formats for genome annotations
- One line per genomic feature
- The **BED format** is the simplest way to store annotation tracks. It has three required fields (chromosome, start, end) and up to 9 optional fields (name, score, strand, thickStart, thickEnd, itemRgb, blockCount, blockSizes, blockStarts).

```
# 6-column BED file defining transcript loci
chr1   66999824   67210768   NM_032291 0 +
chr1   33546713   33586132   NM_052998 0 +
chr1   25071759   25170815   NM_013943 0 +
chr1   48998526   50489626   NM_032785 0 -
```

1. Data formats used in NGS

BED / GFF / GTF formats

- The **General Feature Format (GFF) and General Transfer Format (GTF)** has nine required fields; the first three fields form the basic name, start, end tuple that allows for the identification of the location in respect to the reference genome.

Sample GTF output from Ensembl data dump:

```
1 transcribed_unprocessed_pseudogene  gene        11869 14409 . + . gene_id "ENSG00000223972"; gene_name "DDX11L1"; gene_source "havana"; gene_biotype "transcribed_unprocessed_pseudogene";
1 processed_transcript                transcript  11869 14409 . + . gene_id "ENSG00000223972"; transcript_id "ENST00000456328"; gene_name "DDX11L1"; gene_sourc e "havana"; gene_biotype "transcrib
```

Sample GFF output from Ensembl export:

```
X       Ensembl Repeat  2419108 2419128 42      .       .       hid=trf; hstart=1; hend=21
X       Ensembl Repeat  2419108 2419410 2502    -       .       hid=AluSx; hstart=1; hend=303
X       Ensembl Repeat  2419108 2419128 0       .       .       hid=dust; hstart=2419108; hend=2419128
X       Ensembl Pred.trans.     2416676 2418760 450.19  -       2       genscan=GENSCAN00000019335
X       Ensembl Variation       2413425 2413425 .       +       .
X       Ensembl Variation       2413805 2413805 .       +       .
```

http://m.ensembl.org/info/website/upload/gff.html

1. Data formats used in NGS

1. **reference sequence**: coordinate system of the annotation (e.g., "Chr1")
2. **source**: describes how the annotation was derived (e.g., the name of the annotation software)
3. **method**: annotation type (e.g., gene)
4. **start position**: 1-based integer, always less than or equal to the stop position
5. **stop position**: for zero-length features, such as insertion sites, start equals end and the implied site is to the right of the indicated base
6. **score**: e.g., sequence identity
7. **strand**: "+" for the forward strand, "-" for the reverse strand, or "." for annotations that are not stranded
8. **phase**: codon phase for annotations linked to proteins; 0, 1, or 2, indicating the frame, or the number of bases that should be removed from the beginning of this feature to reach the first base of the next codon
9. **group**: contains the class and ID of an annotation which is the logical parent of the current one ("feature is composed of")

1. Data formats used in NGS

VCF format

- Variant Call Format (VCF) is a text file format.
- It contains meta-information lines, a header line, and then data lines each containing information about a position in the genome.
- Also has the ability to contain genotype information on samples for each position.
- The header line names the 8 fixed, mandatory columns.

1. #CHROM
2. POS
3. ID
4. REF
5. ALT
6. QUAL
7. FILTER
8. INFO

1. Data formats used in NGS

Vall d'Hebron
Institut de Recerca

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

https://samtools.github.io/hts-specs/VCFv4.2.pdf

Vall d'Hebron
Institut de Recerca

To sum up

- **Raw data**: .fastq (.fastq.gz)
- **Aligned data**: .sam / .bam
- **Annotation data**: .gtf / .gff / .bed
- **Results data**: .vcf

Vall d'Hebron
Institut de Recerca

1. Data formats used in NGS
2. Introduction to Galaxy

2. Introduction to Galaxy

- 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

2. Introduction to Galaxy

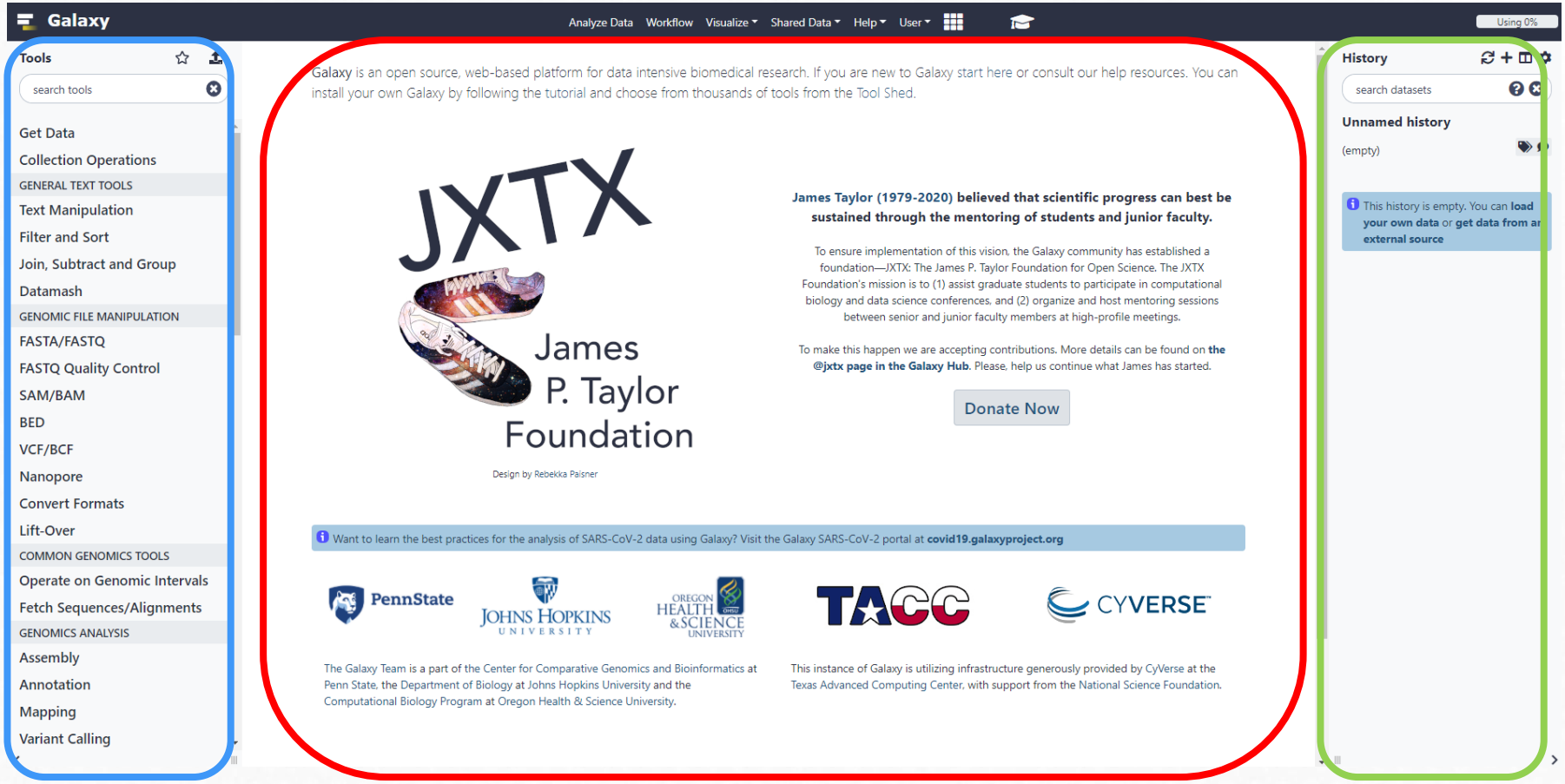
Vall d'Hebron Institut de Recerca

The Galaxy page is divided into three panels:

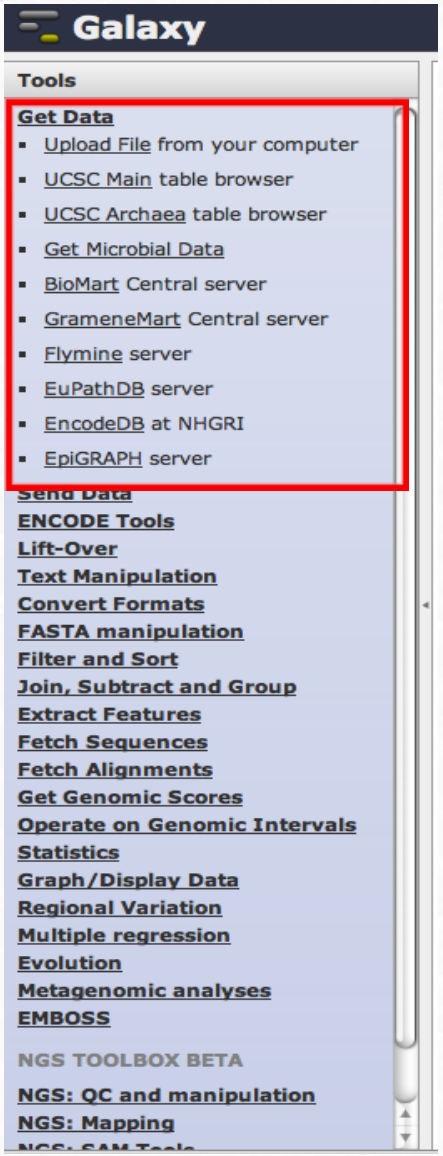
Tools for uploading, processing and analysis

Viewing panel (menus, data, results)

History of analysis steps and datasets



2. 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

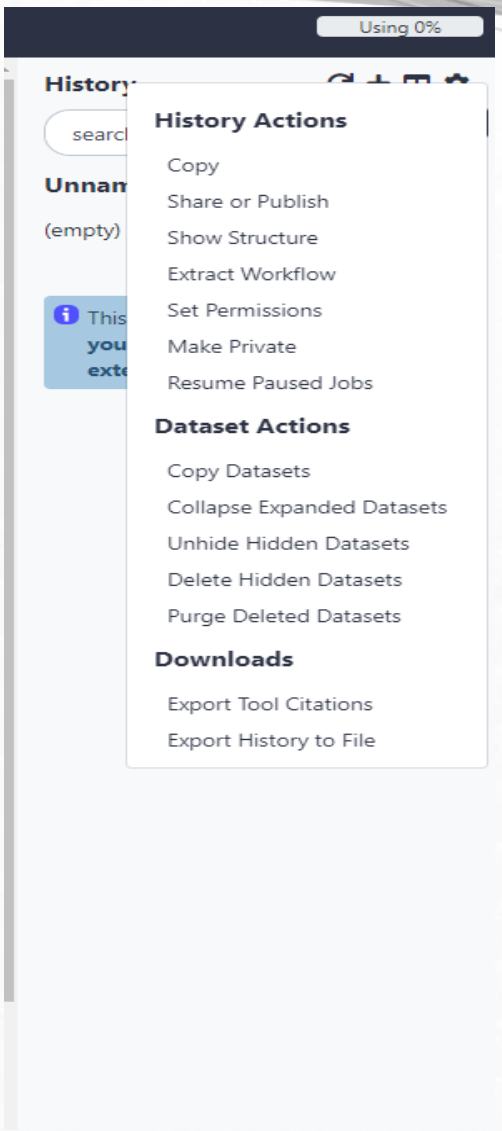
FASTA manipulation

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

2. Introduction to Galaxy

Histories

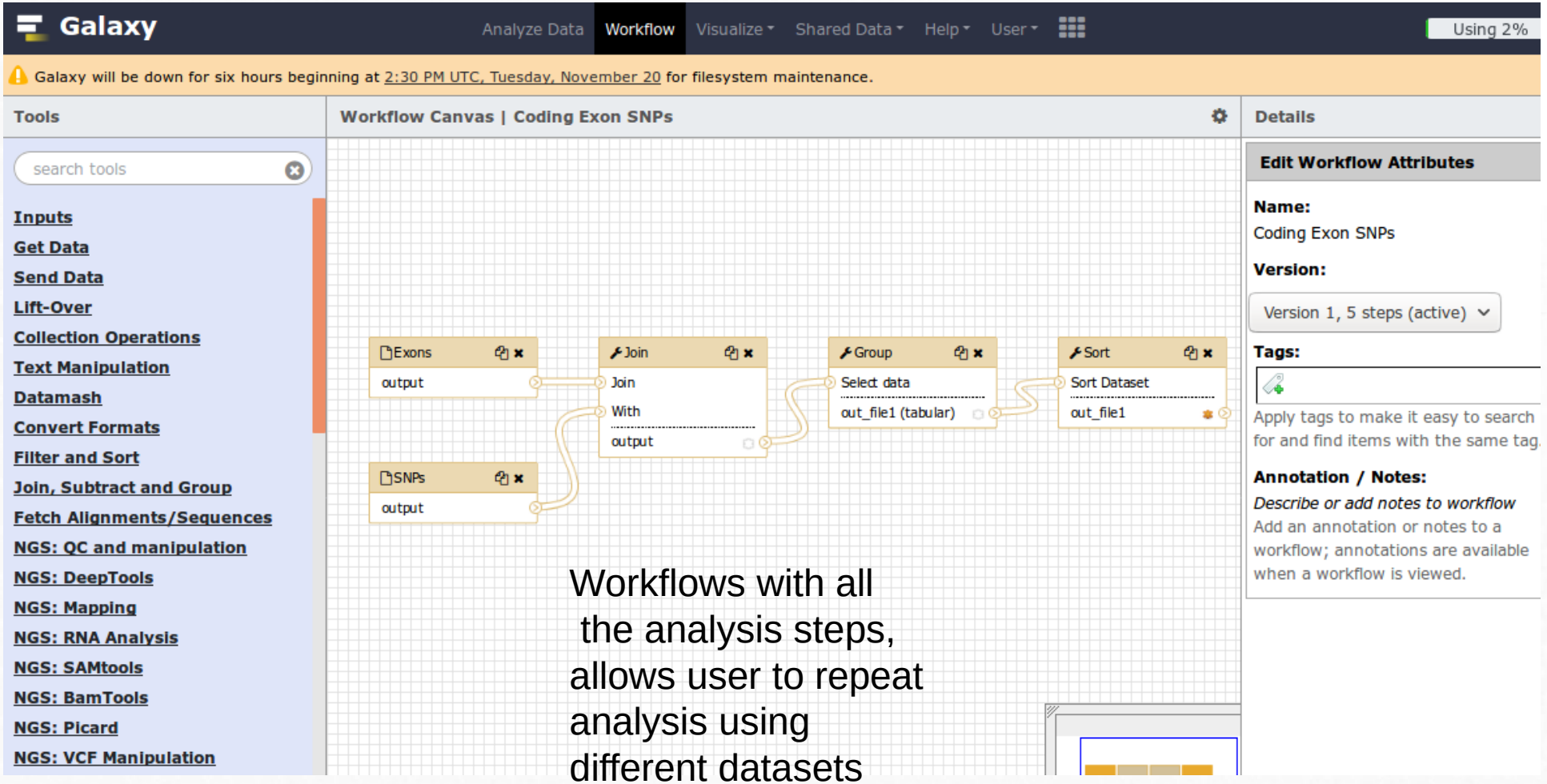
List saved histories and shared histories.
Work on Current History, create new, clone, share, create workflow, set permissions, show deleted datasets or delete history.



2. Introduction to Galaxy

Vall d'Hebron Institut de Recerca

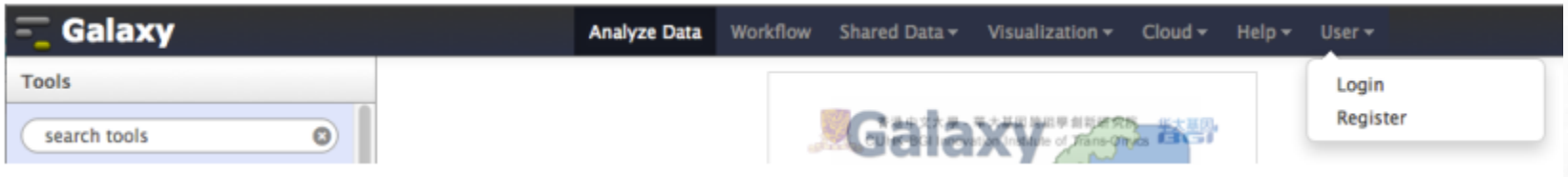
Workflows



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

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/

2. Introduction to Galaxy

Vall d'Hebron
Institut de Recerca

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

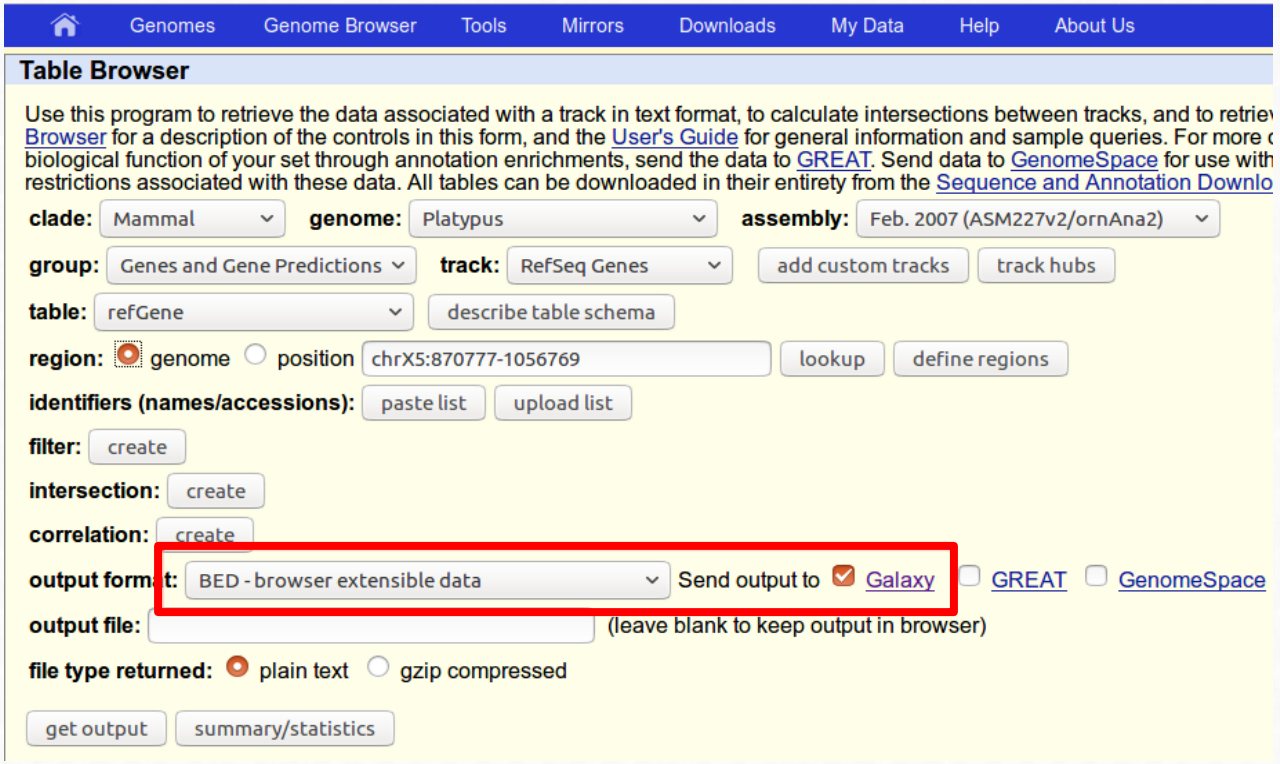
Vall d'Hebron Institut de Recerca

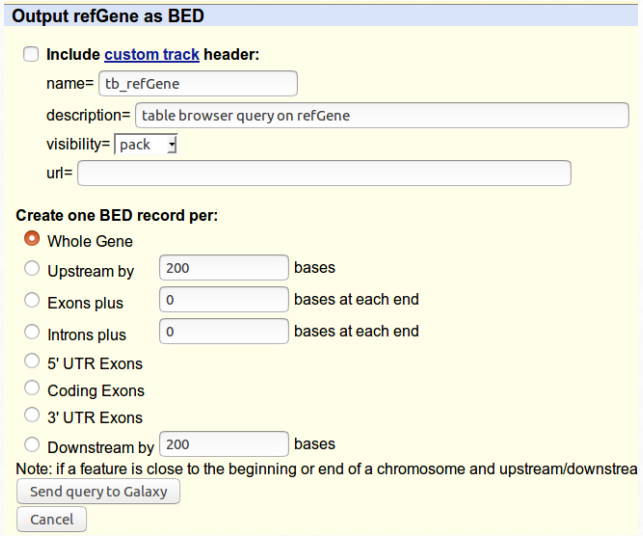
Importing data into Galaxy

1. From database queries (eg. UCSC): obtain a BED-formatted dataset of all RefSeq genes from platypus.

   Get Data > UCSC Main – Table Browser tool

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



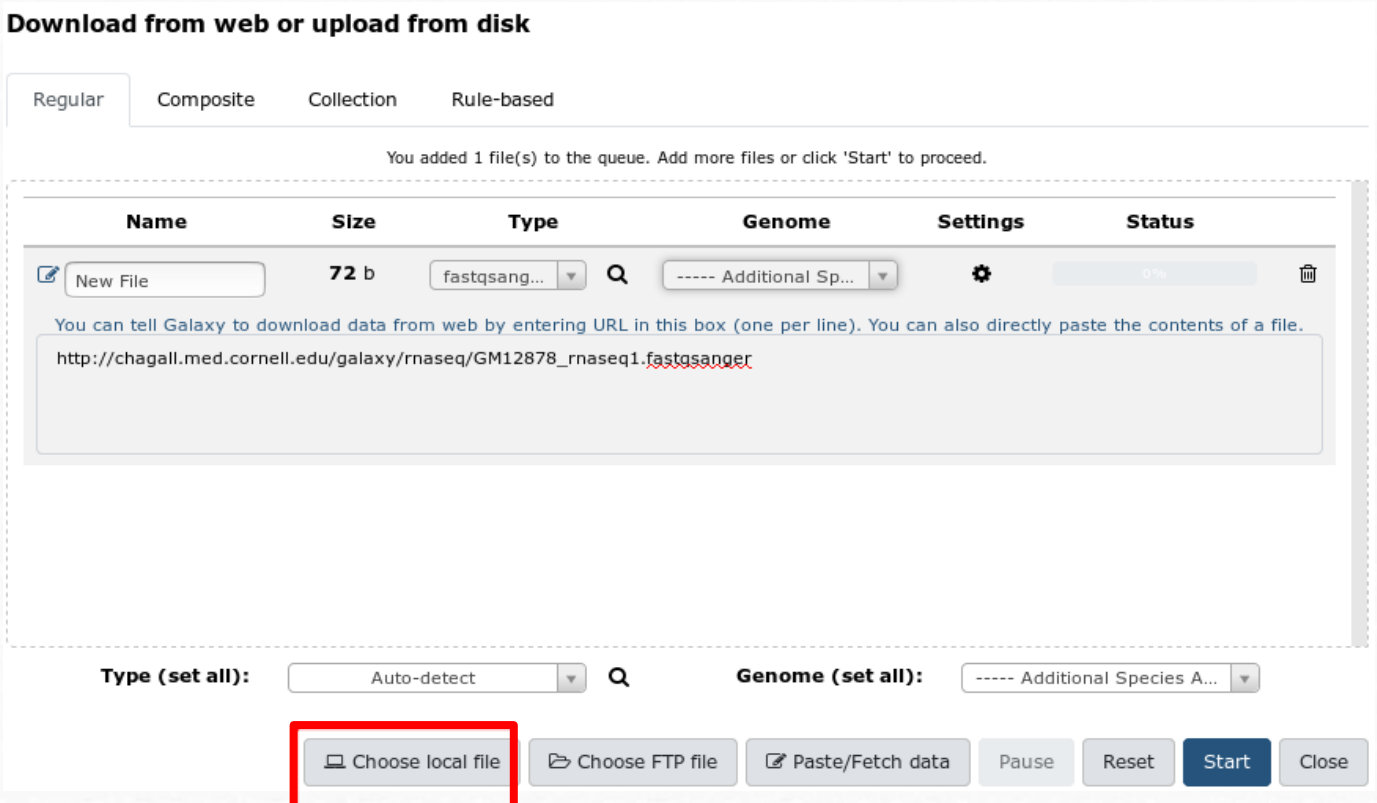


Vall d'Hebron
Institut de Recerca

Importing data into Galaxy

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

Get Data > Upload File



Vall d'Hebron
Institut de Recerca

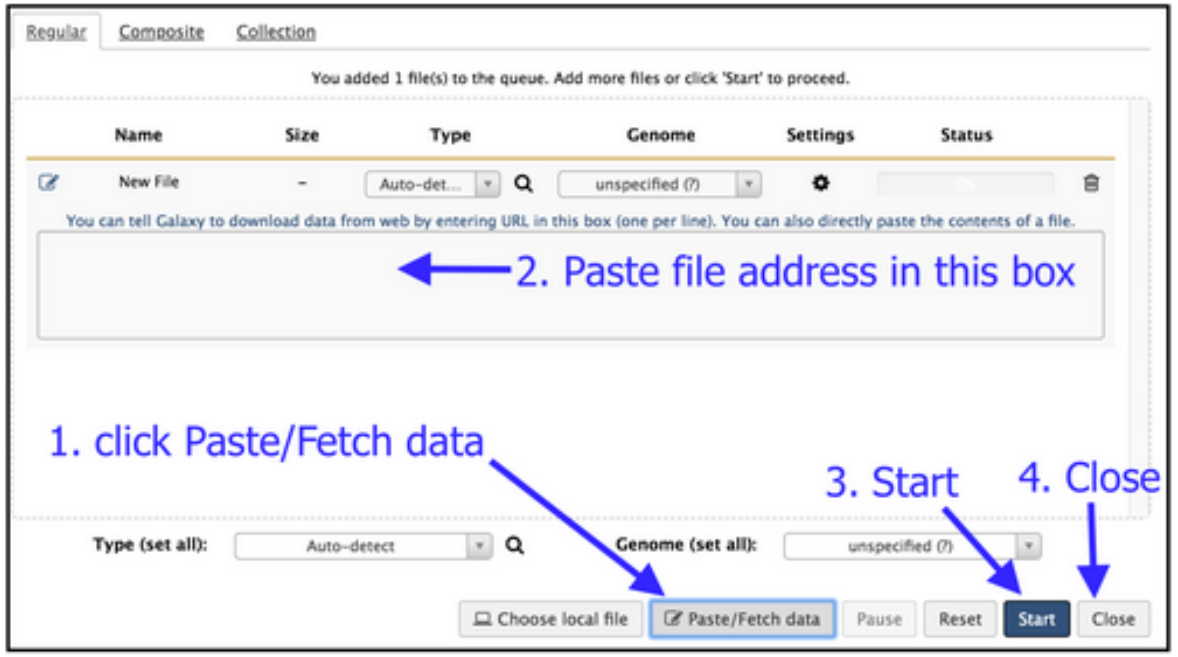
Importing data into Galaxy

3. From a website:

Get Data > Upload File

Copy this URL into the text-entry box:

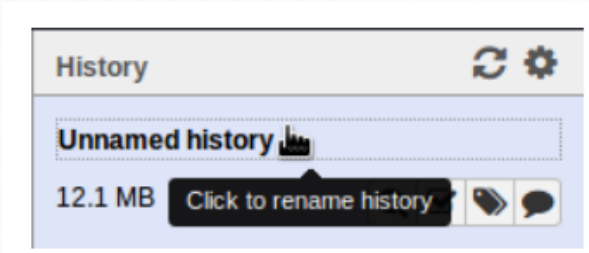
url: https://zenodo.org/record/582600/files/mutant_R1.fastq

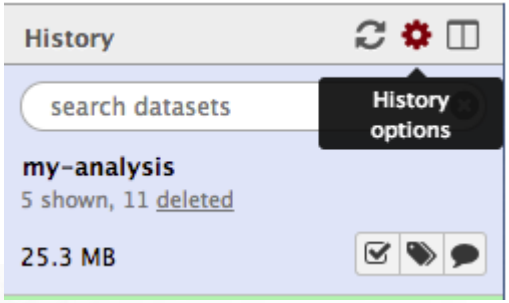


Vall d'Hebron Institut de Recerca

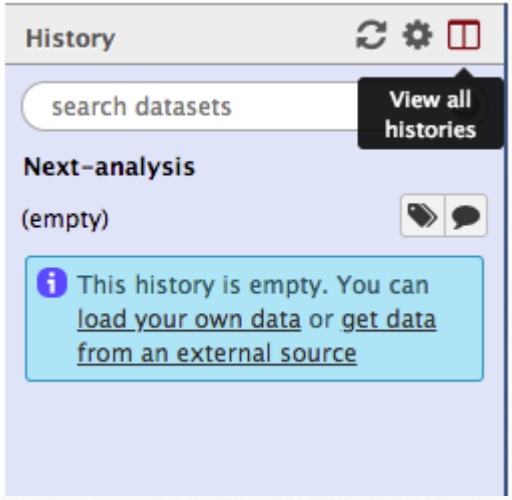
Managing histories

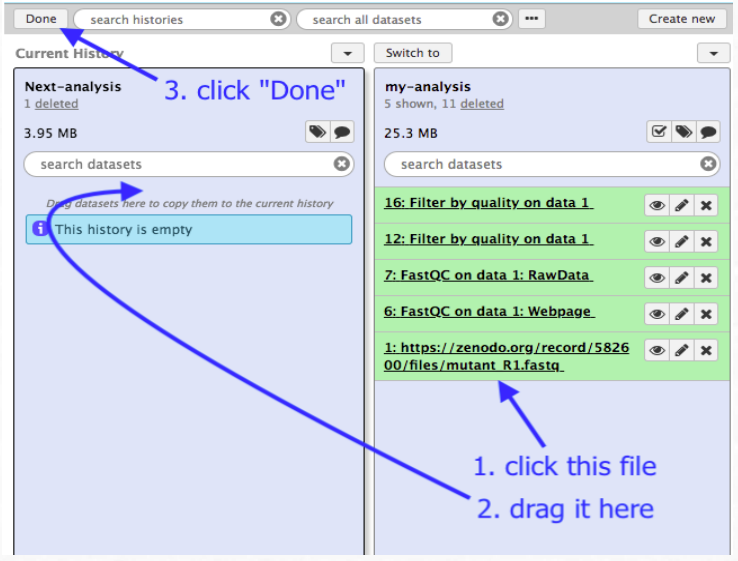
- Name your current history
- Create new history and rename it





- Manage datasets and histories:
- View all histories
- Drag files between histories (new history must be set to current)





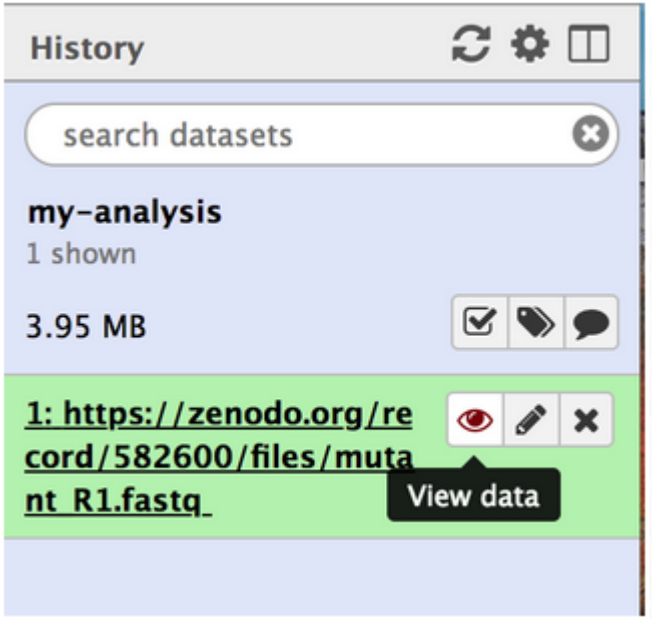
Vall d'Hebron
Institut de Recerca

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:

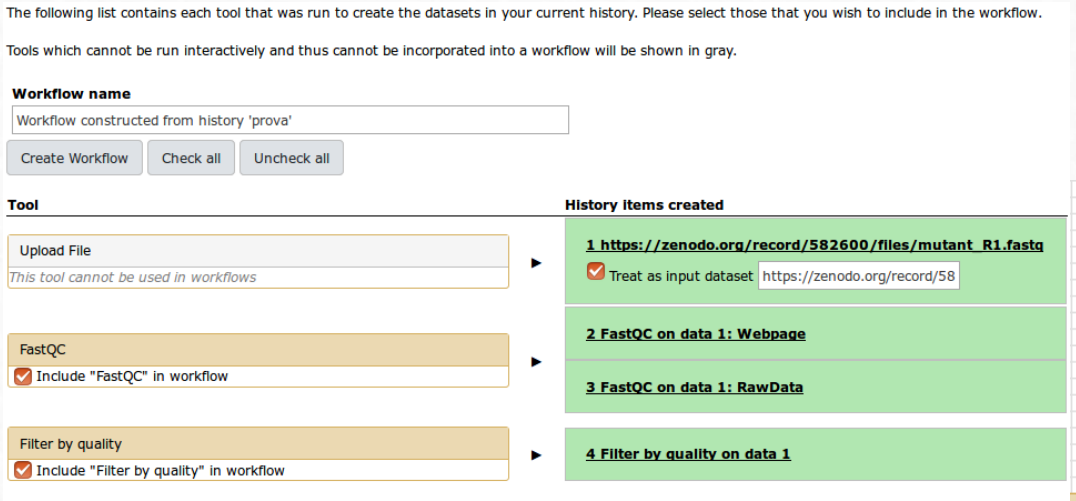
```
@mutant-no_snps.gff-24960/1                                    read 1 sequence
AATGTTGTCACTTGGATTCAAATGACATTTTAAATCTAATTATTCATGAATCGAACTAGTACGAAATGCAATGAG
+                                                              read 1 quality scores
5??A9?BBBDDDBEDDBFF+FGHHIIHHHEIHIIHIIAHDHIIHIG#IIHIFHHHFGIII*IHHHIHFIIHGICI
@mutant-no_snps.gff-24958/1
CAAAGTCGTTGGTCATATAAAAAACCGCGTACAGTCAACTATAGATACAATCAAGATAAACTCATGCACAGATTG
+
?A????@?DDDABDE9FGGGFGICFHIIIBGHIIIGICHHIFH=IHAFIHIHHHHIFCIIEIHAIFGIHIDDIHE
@mutant-no_snps.gff-24956/1
TATAAATTCAACTTTGCAACAGAACCATCTAATCTTCAACAAACTGGCCCGTTTGTTGAACTACTCTTTAATAAA
+
?????BBADD5DDDDDGFGCFFEECFBBICIII,IIHIICHIHIIFHHHHHIHIIIIIIIAHHIHHH5FHDHHHH
```

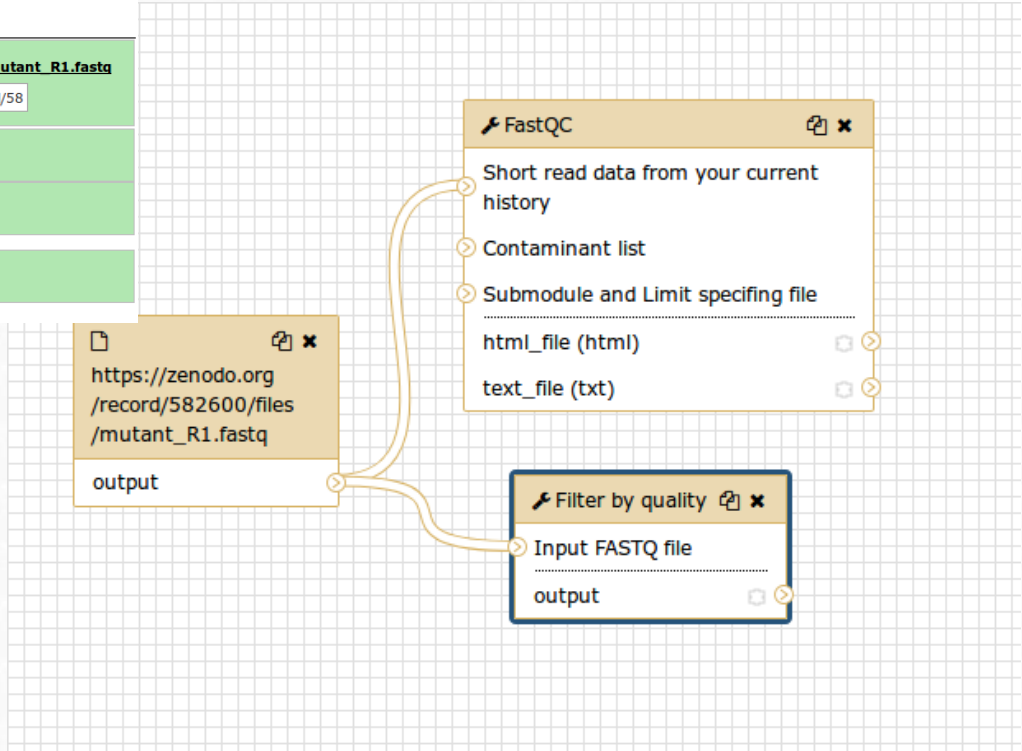


Vall d'Hebron Institut de Recerca

Create workflow from history

- From history options: Export workflow





2. 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

Search

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 nanopore plasmids						
Metatranscriptomics analysis using microbiome RNA-seq data metatranscriptomics						
Metatranscriptomics analysis using microbiome RNA-seq data (short) metatranscriptomics						