ACEMFS FUT Minna Bioinformatics Workshop

Sequence Retrieval & Quality Control using Galaxy

Itunuoluwa Isewon PhD
Covenant University

Workshop Outline

By the end of this session, participants will be able to:

- Retrieve sequences of mycotoxins & fungal enzymes from databases.
- Upload & organize datasets in **Galaxy**.
- Perform QC using Galaxy tools (**FastQC**)

Introduction to Galaxy

- Web-based platform for bioinformaticians.
- No coding required.
- Supports reproducible research.
- Widely used for genomics, proteomics and transcriptomics analysis.

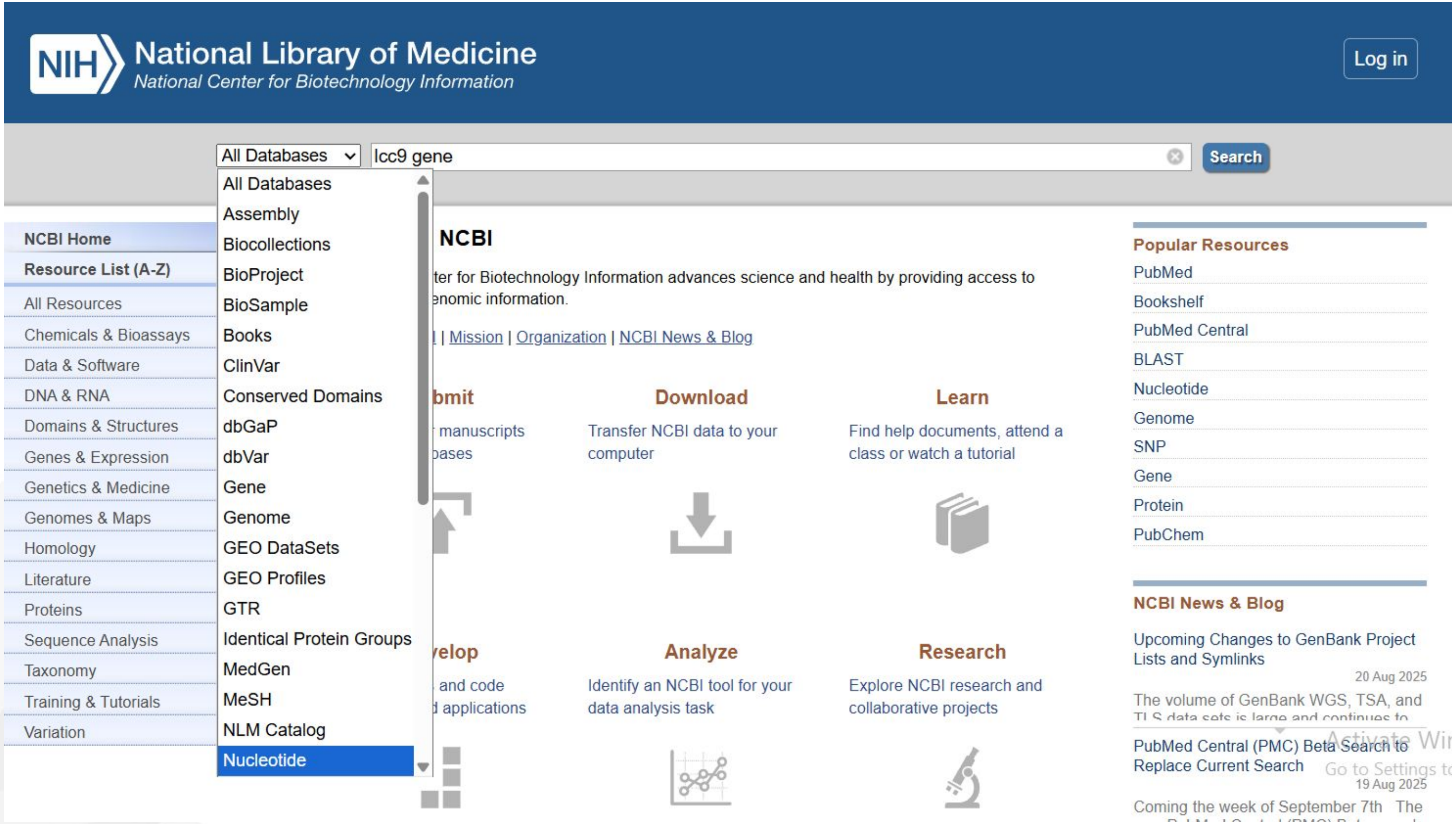
Sequence Retrieval

Sequence Retrieval

NCBI Exploration:

1. Go to **NCBI**
2. In the search bar, paste the **lcc9 gene**.
3. Click on the dropdown arrow next to the search bar and select **Nucleotide**.
4. Click the search button.

Sequence Retrieval



All Databases: Icc9 gene

NCBI

ter for Biotechnology Information advances science and health by providing access to enomic information.

| Mission | Organization | NCBI News & Blog

bmit

manuscripts
bases

Download

Transfer NCBI data to your computer

Learn

Find help documents, attend a class or watch a tutorial

velop

and code
applications

Analyze

Identify an NCBI tool for your data analysis task

Research

Explore NCBI research and collaborative projects

NCBI News & Blog

Upcoming Changes to GenBank Project Lists and Symlinks

20 Aug 2025

The volume of GenBank WGS, TSA, and TLS data sets is large and continues to

PubMed Central (PMC) Beta Search to Replace Current Search

19 Aug 2025

Coming the week of September 7th The

Sequence Retrieval

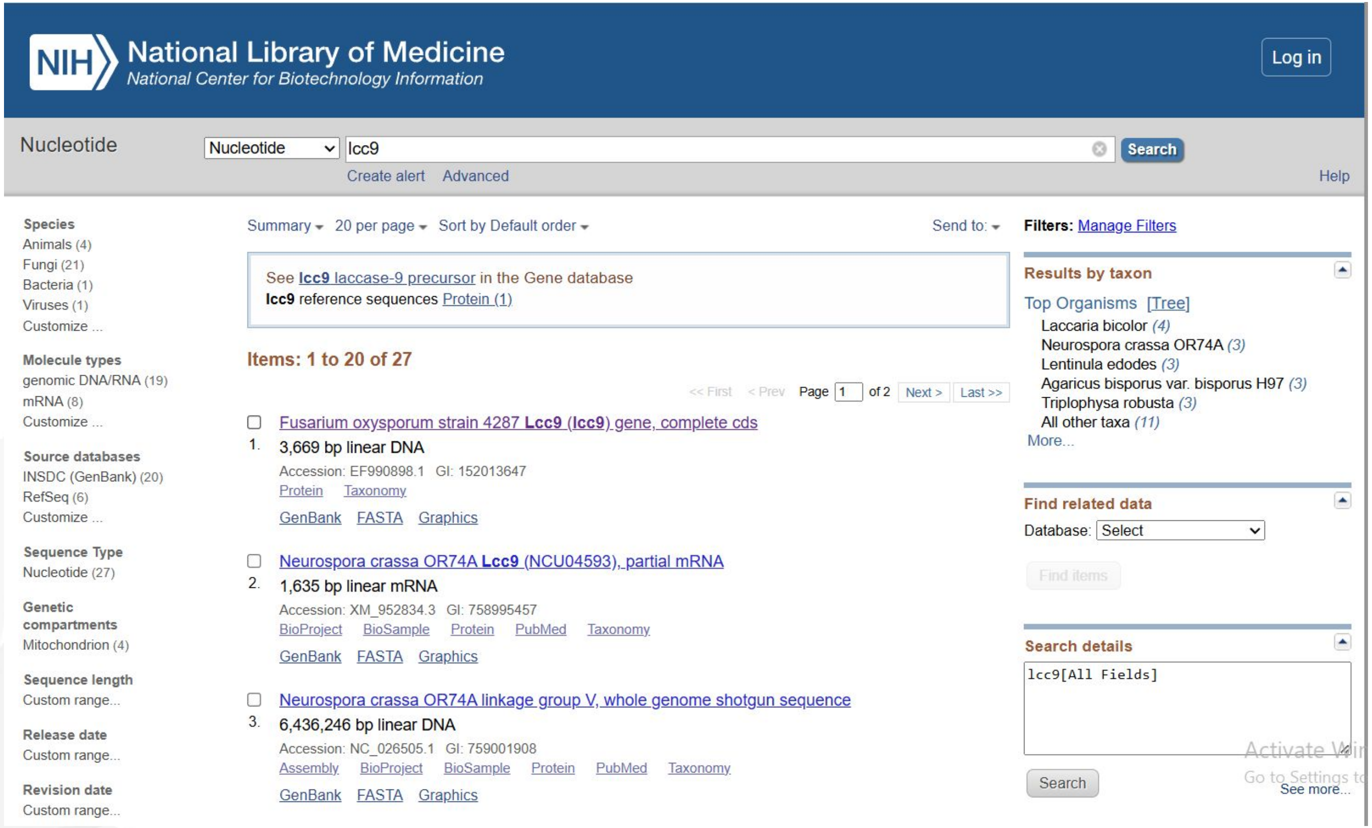
5. On the results page click the **first hit**.

6. What is the name of the **organism**?

7. How many **base pairs** does it have?

8. What is its **accession** and **version numbers**?

Sequence Retrieval



NIH
National Library of Medicine
National Center for Biotechnology Information
Log in

Nucleotide | Nucleotide | lcc9 | Search

Create alert Advanced Help

Summary ▾ 20 per page ▾ Sort by Default order ▾ Send to: ▾

Species
Animals (4)
Fungi (21)
Bacteria (1)
Viruses (1)
Customize ...

Molecule types
genomic DNA/RNA (19)
mRNA (8)
Customize ...

Source databases
INSDC (GenBank) (20)
RefSeq (6)
Customize ...

Sequence Type
Nucleotide (27)

Genetic compartments
Mitochondrion (4)

Sequence length
Custom range...

Release date
Custom range...

Revision date
Custom range...

See **lcc9** laccase-9 precursor in the Gene database
lcc9 reference sequences Protein (1)

Items: 1 to 20 of 27

<< First < Prev Page 1 of 2 Next > Last >>

□ 1. Fusarium oxysporum strain 4287 **Lcc9** (**lcc9**) gene, complete cds
3,669 bp linear DNA
Accession: EF990898.1 GI: 152013647
Protein Taxonomy
GenBank FASTA Graphics

□ 2. Neurospora crassa OR74A **Lcc9** (NCU04593), partial mRNA
1,635 bp linear mRNA
Accession: XM_952834.3 GI: 758995457
BioProject BioSample Protein PubMed Taxonomy
GenBank FASTA Graphics

□ 3. Neurospora crassa OR74A linkage group V, whole genome shotgun sequence
6,436,246 bp linear DNA
Accession: NC_026505.1 GI: 759001908
Assembly BioProject BioSample Protein PubMed Taxonomy
GenBank FASTA Graphics

Filters: Manage Filters

Results by taxon

Top Organisms [Tree]
Laccaria bicolor *(4)*
Neurospora crassa OR74A *(3)*
Lentinula edodes *(3)*
Agaricus bisporus var. bisporus H97 *(3)*
Triplophysa robusta *(3)*
All other taxa *(11)*
More...

Find related data

Database: Select

Find items

Search details

```
lcc9[All Fields]
```

Search See more...

Sequence Retrieval

9. On the right, click on the dropdown arrow next to Send to:

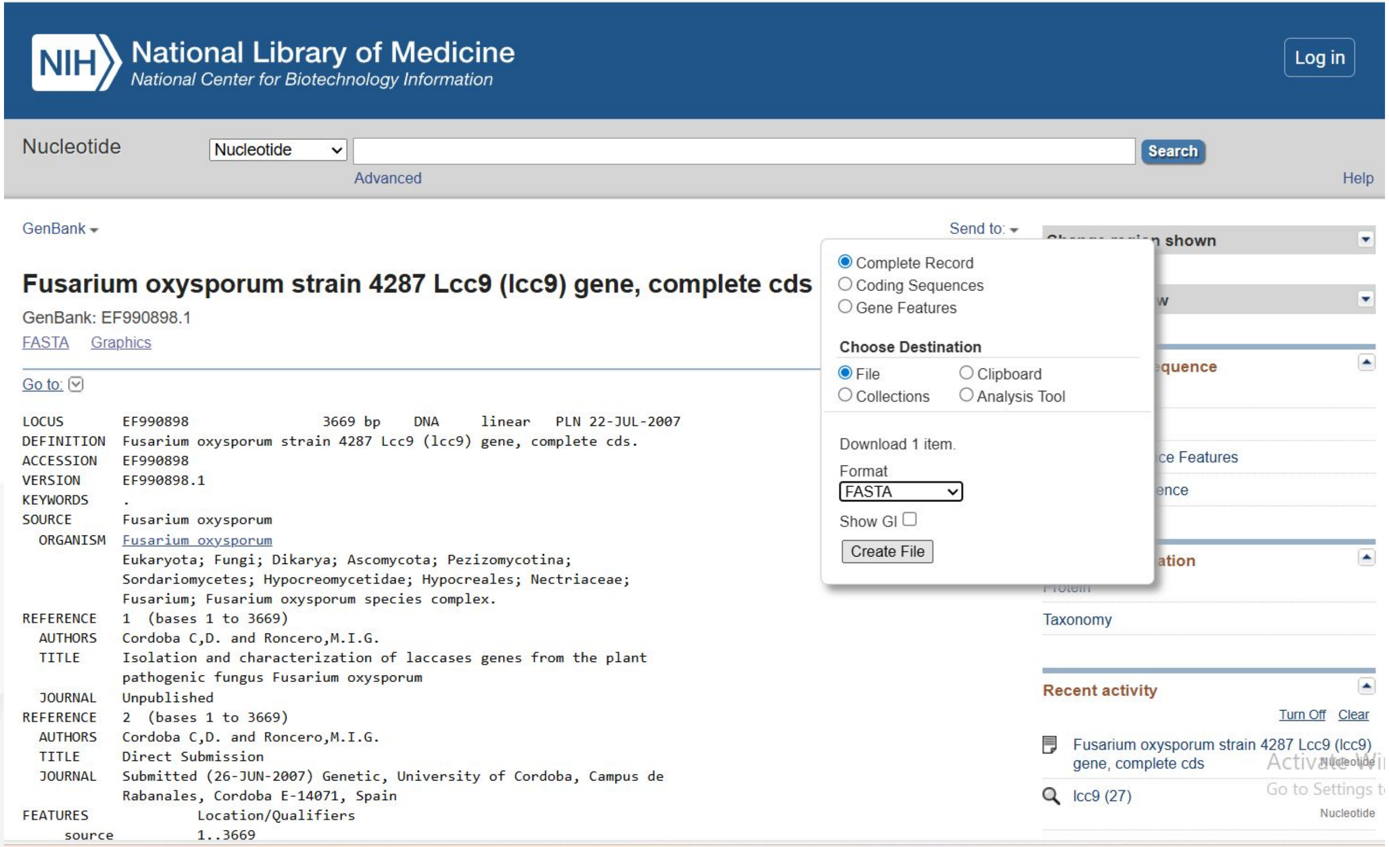
10. Select file under, Choose **Destination**.

11. Change the format to **FASTA**

12. Click on **Create File** to download the FASTA File.

13. **Repeat** this process for a gene interesting to you.

Sequence Retrieval



NIH National Library of Medicine
National Center for Biotechnology Information

Log in

Nucleotide | Nucleotide ▾ | Search
Advanced | Help

GenBank ▾ | Send to: ▾

Fusarium oxysporum strain 4287 Lcc9 (lcc9) gene, complete cds

GenBank: EF990898.1

FASTA Graphics

Go to: ▾

```
LOCUS       EF990898                3669 bp    DNA     linear   PLN 22-JUL-2007
DEFINITION  Fusarium oxysporum strain 4287 Lcc9 (lcc9) gene, complete cds.
ACCESSION   EF990898
VERSION     EF990898.1
KEYWORDS    .
SOURCE      Fusarium oxysporum
  ORGANISM  Fusarium oxysporum
            Eukaryota; Fungi; Dikarya; Ascomycota; Pezizomycotina;
            Sordariomycetes; Hypocreomycetidae; Hypocreales; Nectriaceae;
            Fusarium; Fusarium oxysporum species complex.
REFERENCE   1  (bases 1 to 3669)
  AUTHORS   Cordoba C,D. and Roncero,M.I.G.
  TITLE     Isolation and characterization of laccases genes from the plant
            pathogenic fungus Fusarium oxysporum
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 3669)
  AUTHORS   Cordoba C,D. and Roncero,M.I.G.
  TITLE     Direct Submission
  JOURNAL   Submitted (26-JUN-2007) Genetic, University of Cordoba, Campus de
            Rabanales, Cordoba E-14071, Spain
FEATURES             Location/Qualifiers
     source          1..3669
```

● Complete Record
○ Coding Sequences
○ Gene Features

Choose Destination

● File ○ Clipboard
○ Collections ○ Analysis Tool

Download 1 item.

Format: FASTA ▾

Show GI □

Create File

...n shown ▾

...w ▾

...equence

...ce Features

...ence

...ation

Protein

Taxonomy

Recent activity

Turn Off Clear

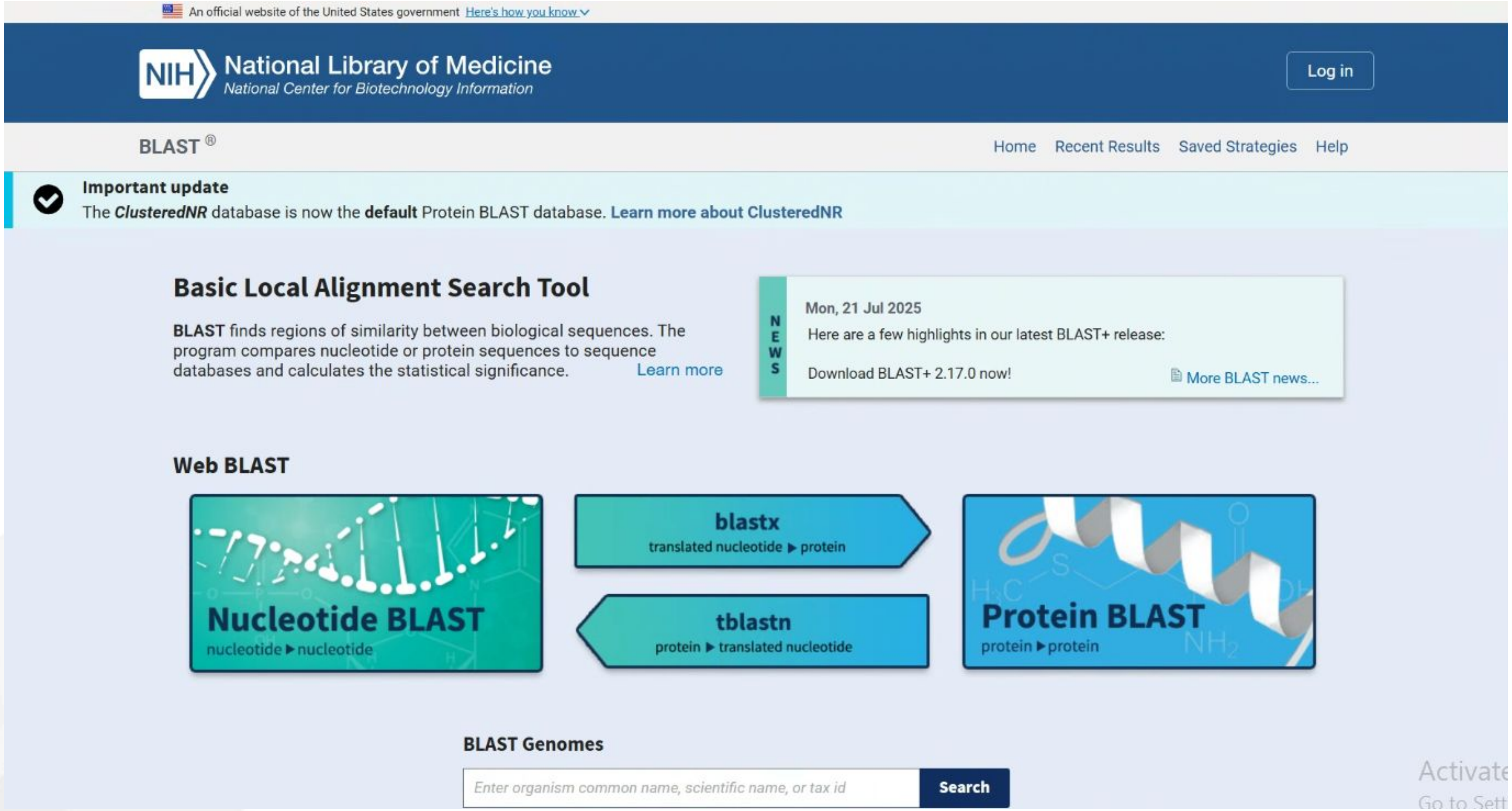
Fusarium oxysporum strain 4287 Lcc9 (lcc9) gene, complete cds — Nucleotide

lcc9 (27) — Nucleotide

BLAST

BLAST

- **BLAST** stands for Basic Local Alignment Search Tool.
- It identifies similarities between biological sequences by comparing nucleotide or protein sequences to a database of sequences.



An official website of the United States government Here's how you know
NIH
National Library of Medicine
National Center for Biotechnology Information
Log in
BLAST ®
Home
Recent Results
Saved Strategies
Help

Important update

The ***ClusteredNR*** database is now the **default** Protein BLAST database. **Learn more about ClusteredNR**

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. Learn more

NEWS

Mon, 21 Jul 2025

Here are a few highlights in our latest BLAST+ release:

Download BLAST+ 2.17.0 now!

More BLAST news...

Web BLAST

BLAST Genomes

BLAST

- **BLASTn (Nucleotide BLAST):** compares one or more nucleotide query sequences to a subject nucleotide sequence or a database of nucleotide sequences.
- **BLASTx (translated nucleotide sequence searched against protein sequences):** compares a nucleotide query sequence that is translated in six reading frames against a database of protein sequences.
- And **many others.**

BLAST

1.Choose **blastn.**

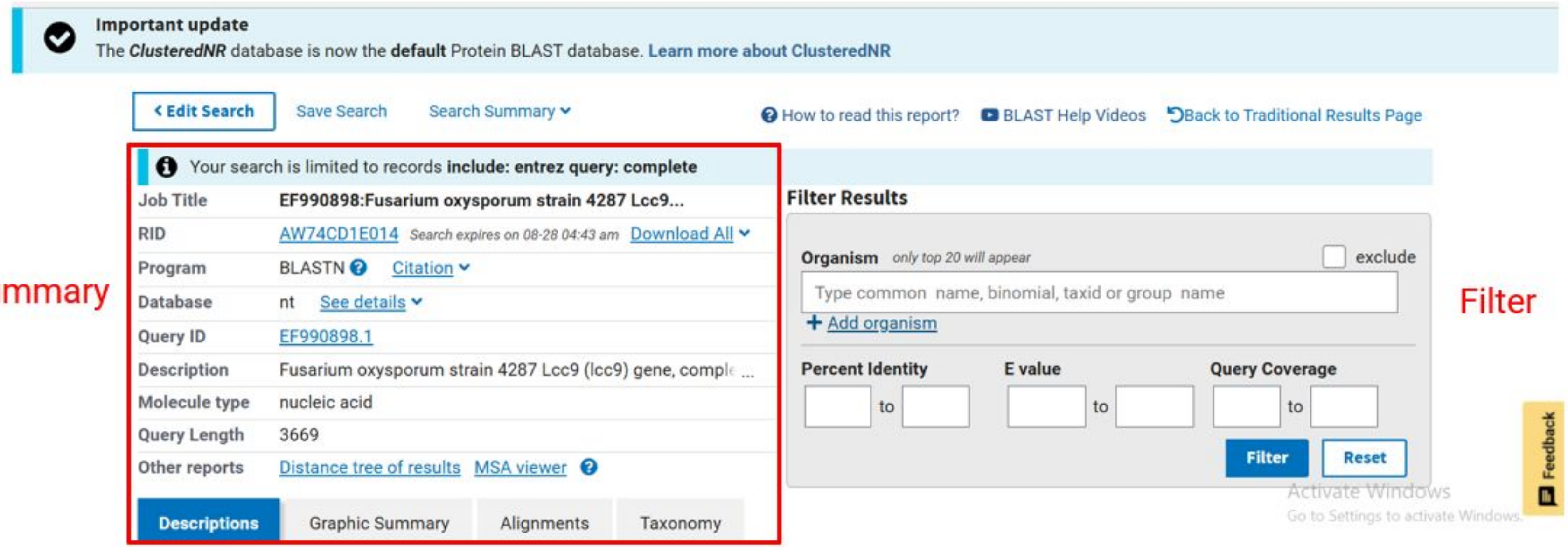
2.Paste the accession number **EF990898.1** in the Enter Query Sequence box.

3.Select **Nucleotide collection nr/nt** under Database in Choose Search Set section.

4.Under Program Selection choose **Highly similar sequences(megablast).**

5.Set the max target sequences under General Parameters to **50.**

6.Click the **BLAST** button.



Important update

The *ClusteredNR* database is now the **default** Protein BLAST database. **Learn more about ClusteredNR**

Your search is limited to records **include: entrez query: complete**

Job Title	**EF990898:Fusarium oxysporum strain 4287 Lcc9...**
RID	AW74CD1E014 *Search expires on 08-28 04:43 am* Download All
Program	BLASTN Citation
Database	nt See details
Query ID	EF990898.1
Description	Fusarium oxysporum strain 4287 Lcc9 (lcc9) gene, comple ...
Molecule type	nucleic acid
Query Length	3669
Other reports	Distance tree of results MSA viewer

Filter Results

Organism *only top 20 will appear* □ exclude

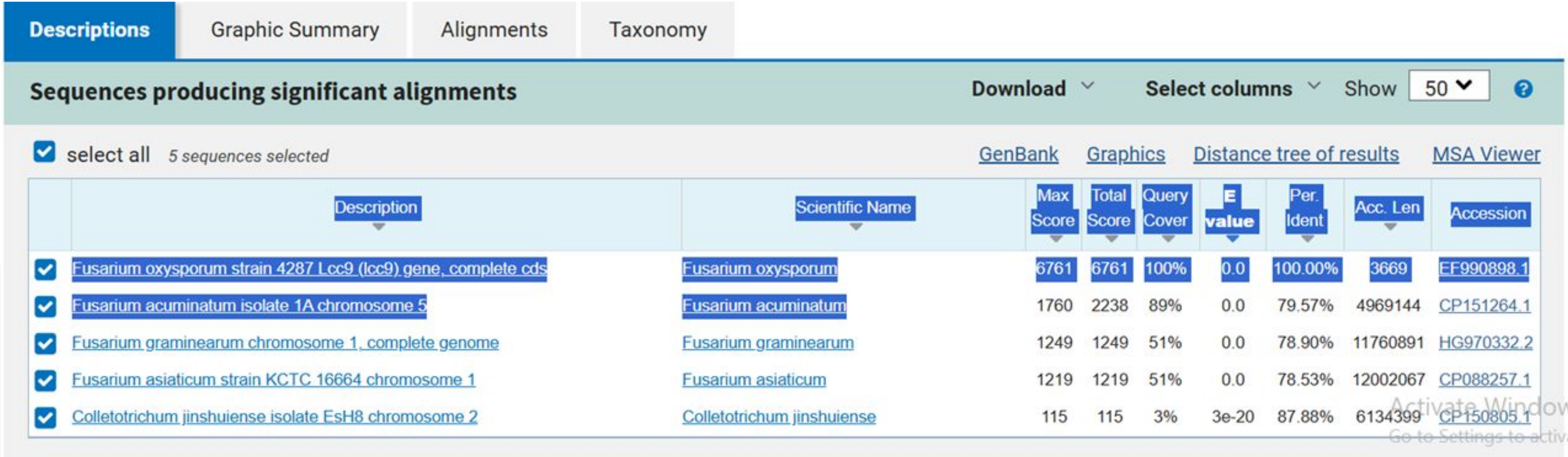
Type common name, binomial, taxid or group name

+ Add organism

Percent Identity to **E value** to **Query Coverage** to

Summary

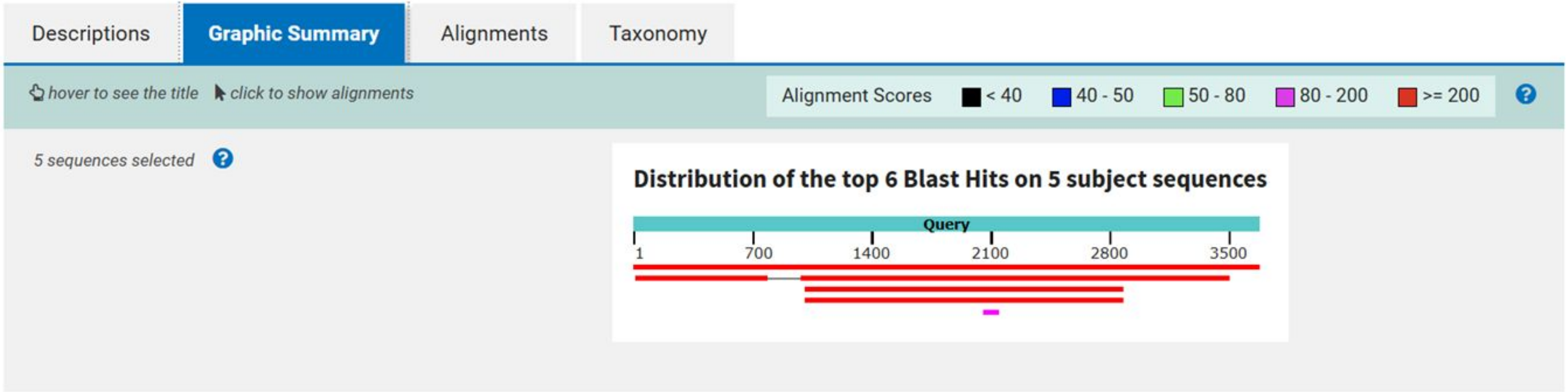
Filter



Sequences producing significant alignments

☑ select all *5 sequences selected*

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	Fusarium oxysporum strain 4287 Lcc9 (lcc9) gene, complete cds	Fusarium oxysporum	6761	6761	100%	0.0	100.00%	3669	EF990898.1
☑	Fusarium acuminatum isolate 1A chromosome 5	Fusarium acuminatum	1760	2238	89%	0.0	79.57%	4969144	CP151264.1
☑	Fusarium graminearum chromosome 1, complete genome	Fusarium graminearum	1249	1249	51%	0.0	78.90%	11760891	HG970332.2
☑	Fusarium asiaticum strain KCTC 16664 chromosome 1	Fusarium asiaticum	1219	1219	51%	0.0	78.53%	12002067	CP088257.1
☑	Colletotrichum jinshuiense isolate EsH8 chromosome 2	Colletotrichum jinshuiense	115	115	3%	3e-20	87.88%	6134399	CP150805.1

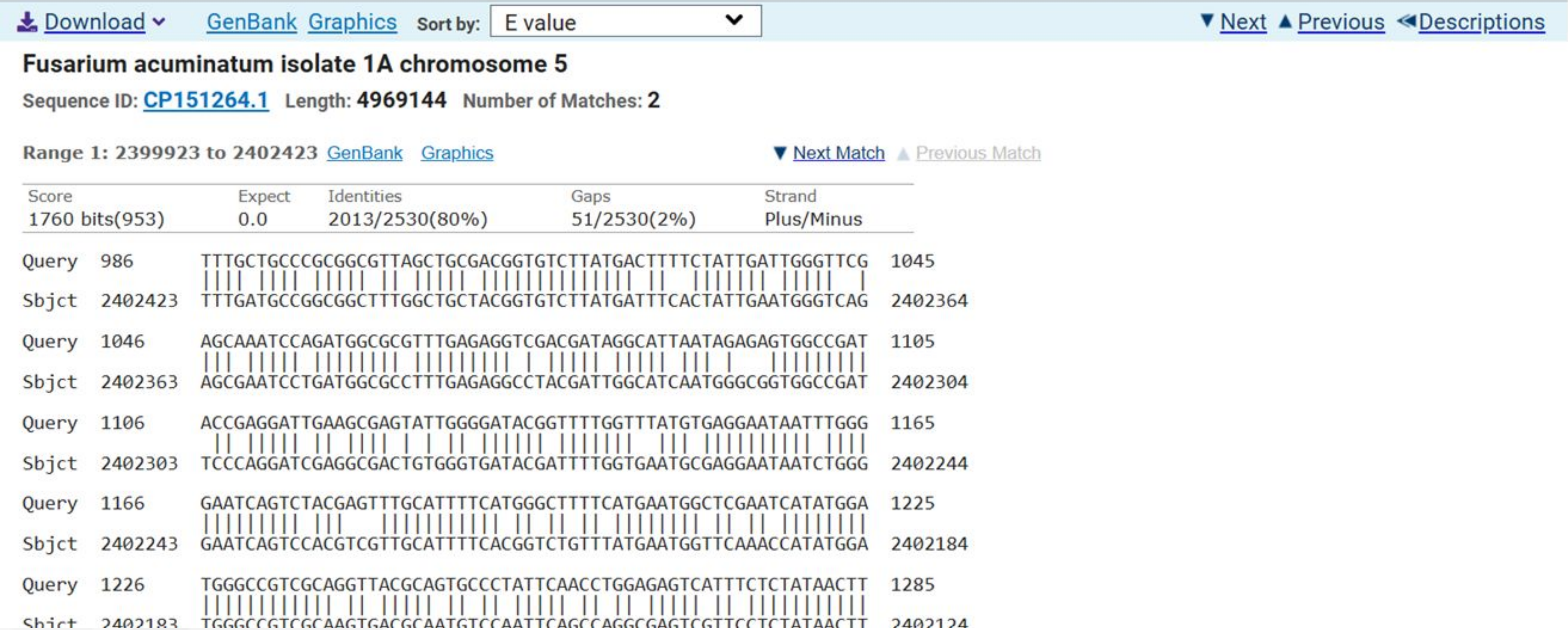


hover to see the title click to show alignments

Alignment Scores ■ < 40 🟦 40 - 50 🟩 50 - 80 🟪 80 - 200 🟥 >= 200

5 sequences selected

Distribution of the top 6 Blast Hits on 5 subject sequences



Sort by: E value

Fusarium acuminatum isolate 1A chromosome 5

Sequence ID: CP151264.1 Length: 4969144 Number of Matches: 2

Range 1: 2399923 to 2402423 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
1760 bits(953)	0.0	2013/2530(80%)	51/2530(2%)	Plus/Minus

```
Query  986      TTTGCTGCCCGCGGCGTTAGCTGCGACGGTGTCTTATGACTTTTCTATTGATTGGGTTCG  1045
                |||| |||| ||||| || ||||| |||||||||||||| ||  ||||||| |||||  |
Sbjct  2402423  TTTGATGCCGGCGGCTTTGGCTGCTACGGTGTCTTATGATTTCACTATTGAATGGGTCAG  2402364

Query  1046     AGCAAATCCAGATGGCGCGTTTGAGAGGTCGACGATAGGCATTAATAGAGAGTGGCCGAT  1105
                ||| ||||| |||||||| ||||||||| | ||||| ||||| ||| |   |||||||||
Sbjct  2402363  AGCGAATCCTGATGGCGCCTTTGAGAGGCCTACGATTGGCATCAATGGGCGGTGGCCGAT  2402304

Query  1106     ACCGAGGATTGAAGCGAGTATTGGGGATACGGTTTTGGTTTATGTGAGGAATAATTTGGG  1165
                 || ||||| || |||| | | || |||||| |||||||  ||| |||||||||| ||||
Sbjct  2402303  TCCCAGGATCGAGGCGACTGTGGGTGATACGATTTTGGTGAATGCGAGGAATAATCTGGG  2402244

Query  1166     GAATCAGTCTACGAGTTTGCATTTTCATGGGCTTTTCATGAATGGCTCGAATCATATGGA  1225
                ||||||||| |||   ||||||||||| || || || |||||||| || || ||||||||
Sbjct  2402243  GAATCAGTCCACGTCGTTGCATTTTCACGGTCTGTTTATGAATGGTTCAAACCATATGGA  2402184

Query  1226     TGGGCCGTCGCAGGTTACGCAGTGCCCTATTCAACCTGGAGAGTCATTTCTCTATAACTT  1285
                |||||||||||| || ||||| || || ||||| || || ||||| || |||||||||||
Sbjct  2402183  TGGGCCGTCGCAAGTGACGCAATGTCCAATTCAGCCAGGCGAGTCGTTCCTCTATAACTT  2402124
```

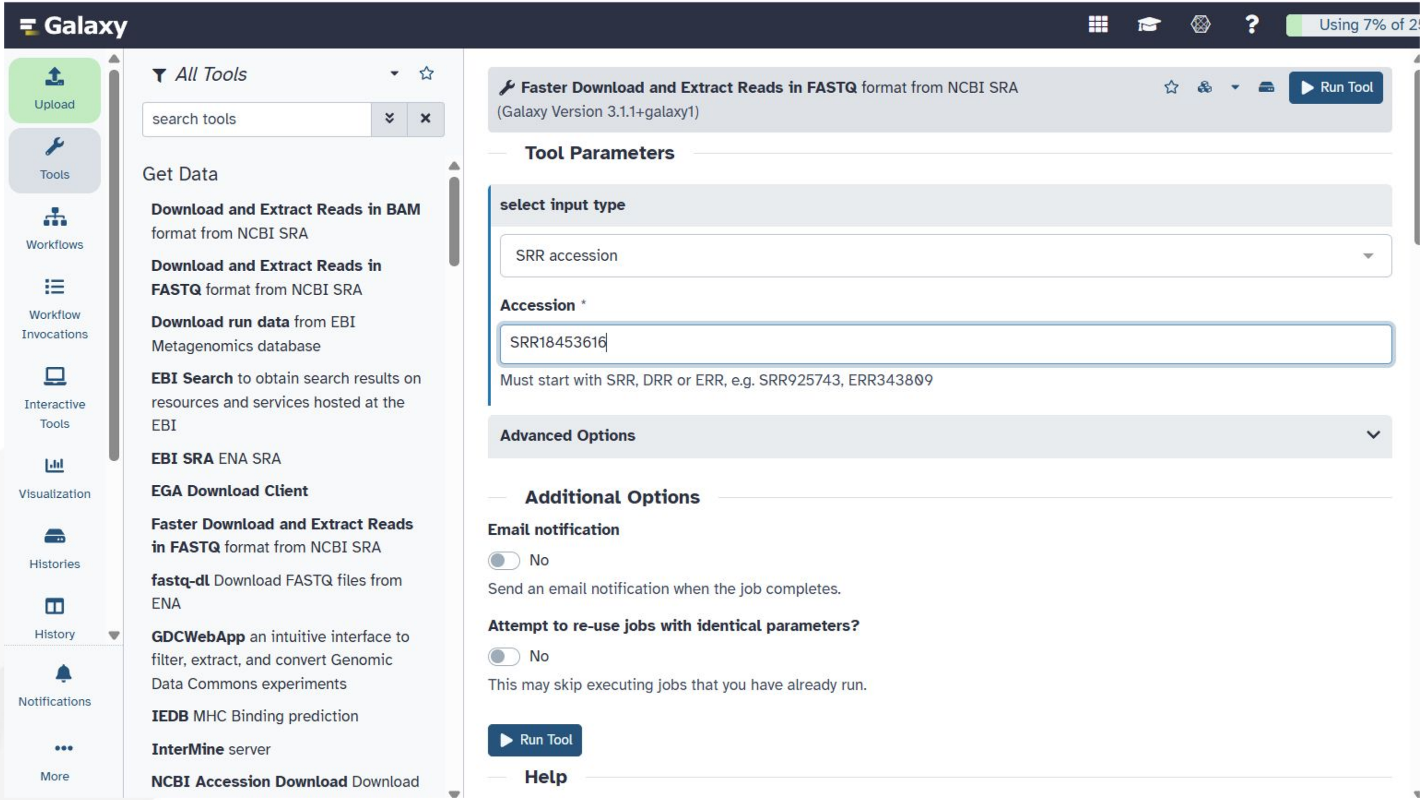
Quality Control

Quality Control in Galaxy

Obtain **FASTQ** files:

1. Go to **Galaxy**
2. In the tool Bar, click on **Get Data**
3. Choose "**Faster Download and Extract Reads in FASTA/Q format from NCBI SRA**".
4. In the Accession tab, write the accession number of the fastq file: **SRR18453616**.

Quality Control in Galaxy



Get Data

Download and Extract Reads in BAM format from NCBI SRA

Download and Extract Reads in FASTQ format from NCBI SRA

Download run data from EBI Metagenomics database

EBI Search to obtain search results on resources and services hosted at the EBI

EBI SRA ENA SRA

EGA Download Client

Faster Download and Extract Reads in FASTQ format from NCBI SRA

fastq-dl Download FASTQ files from ENA

GDCWebApp an intuitive interface to filter, extract, and convert Genomic Data Commons experiments

IEDB MHC Binding prediction

InterMine server

NCBI Accession Download Download

Faster Download and Extract Reads in FASTQ format from NCBI SRA (Galaxy Version 3.1.1+galaxy1)

Run Tool

Tool Parameters

select input type

SRR accession

Accession *

SRR18453616

Must start with SRR, DRR or ERR, e.g. SRR925743, ERR343809

Advanced Options

Additional Options

Email notification

No

Send an email notification when the job completes.

Attempt to re-use jobs with identical parameters?

No

This may skip executing jobs that you have already run.

Run Tool

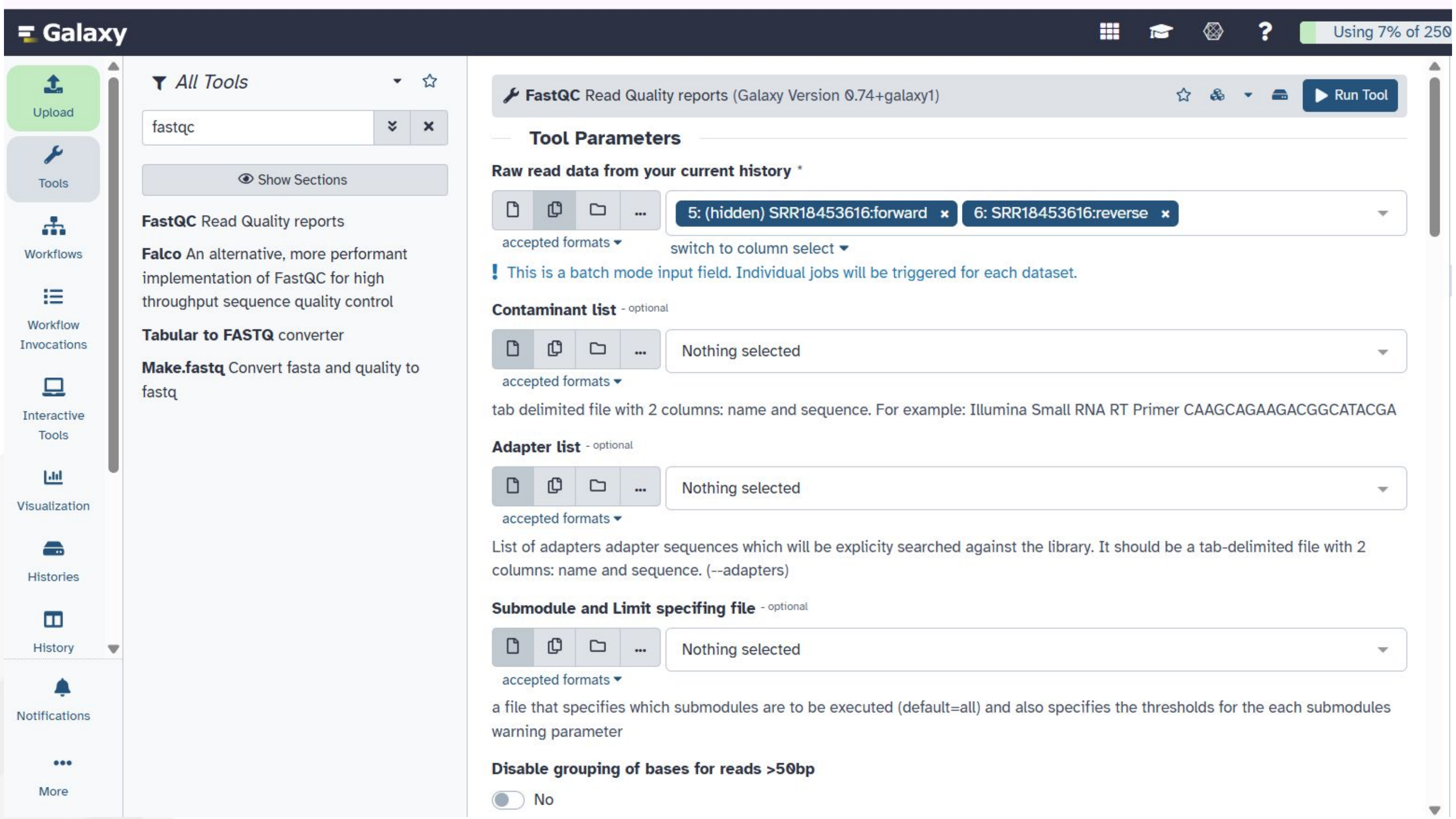
Help

Quality Control in Galaxy

Perform **QC**:

1. On the search bar, type **fastqc**.
2. **Choose** the desired fastq file (paired end) in the raw read tab.
3. **Leave** all other tabs unchanged.

Quality Control in Galaxy



fastqc

FastQC Read Quality reports

Falco An alternative, more performant implementation of FastQC for high throughput sequence quality control

Tabular to FASTQ converter

Make.fastq Convert fasta and quality to fastq

FastQC Read Quality reports (Galaxy Version 0.74+galaxy1)

Tool Parameters

Raw read data from your current history *

5: (hidden) SRR18453616:forward × 6: SRR18453616:reverse ×

accepted formats ▾ switch to column select ▾

! This is a batch mode input field. Individual jobs will be triggered for each dataset.

Contaminant list - optional

Nothing selected

accepted formats ▾

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

Adapter list - optional

Nothing selected

accepted formats ▾

List of adapters adapter sequences which will be explicity searched against the library. It should be a tab-delimited file with 2 columns: name and sequence. (--adapters)

Submodule and Limit specifing file - optional

Nothing selected

accepted formats ▾

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

Disable grouping of bases for reads >50bp

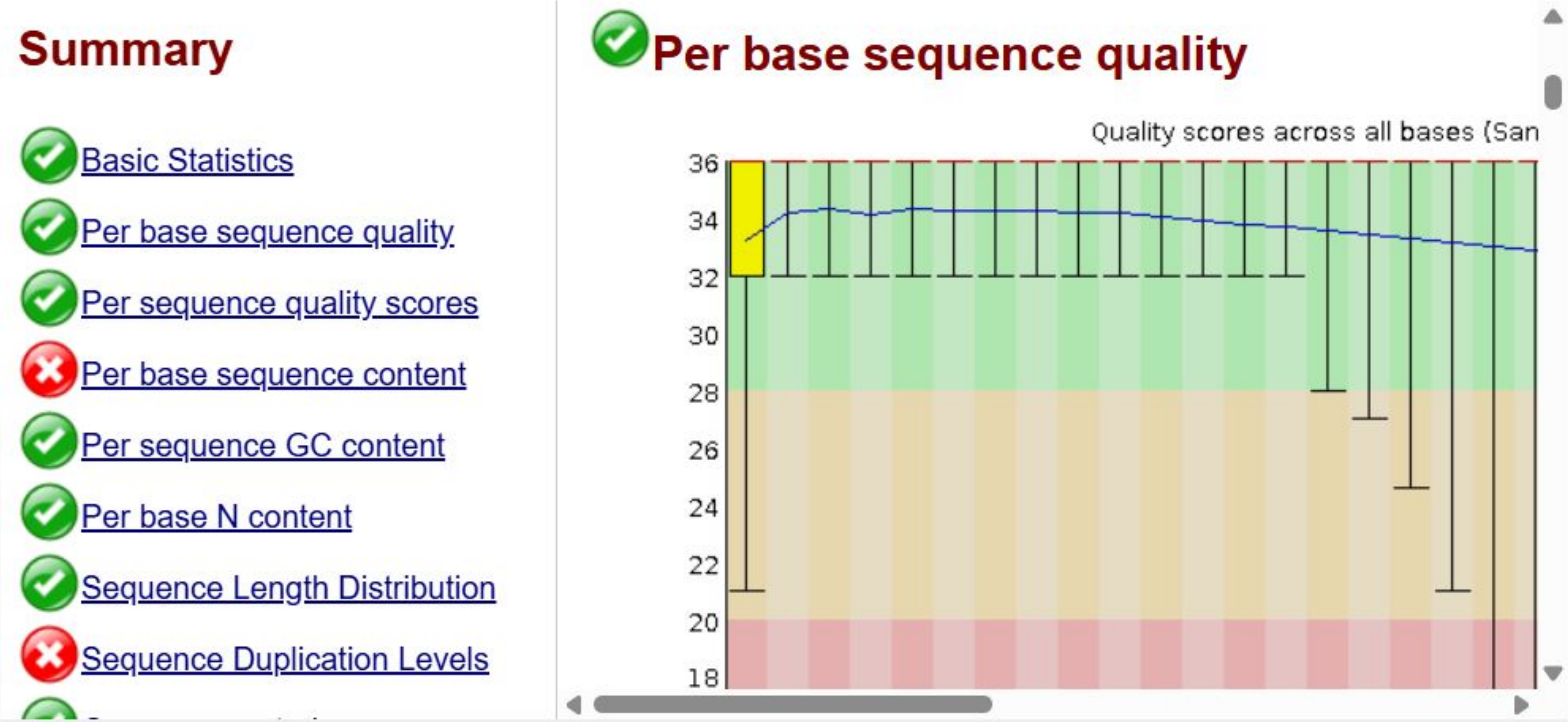
○ No

Quality Control in Galaxy

4. Once it runs, two files are generated, a raw data file and a Web Page file.

5. **View** the result by clicking on the web page file produced.

Quality Control Result



Summary

- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✓ Per sequence quality scores
- ✗ Per base sequence content
- ✓ Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ✗ Sequence Duplication Levels

✓ Per base sequence quality

Quality Control (MultiQC)

Why: It helps us to obtain a more intuitive comparison

1. On the search bar, type **multiqc**
2. On the “Which tool was used generate logs?” tab, choose **Fastqc**
3. Then click on “**Insert FastQC output**”

Quality Control (MultiQC)

4. Type of output is raw data

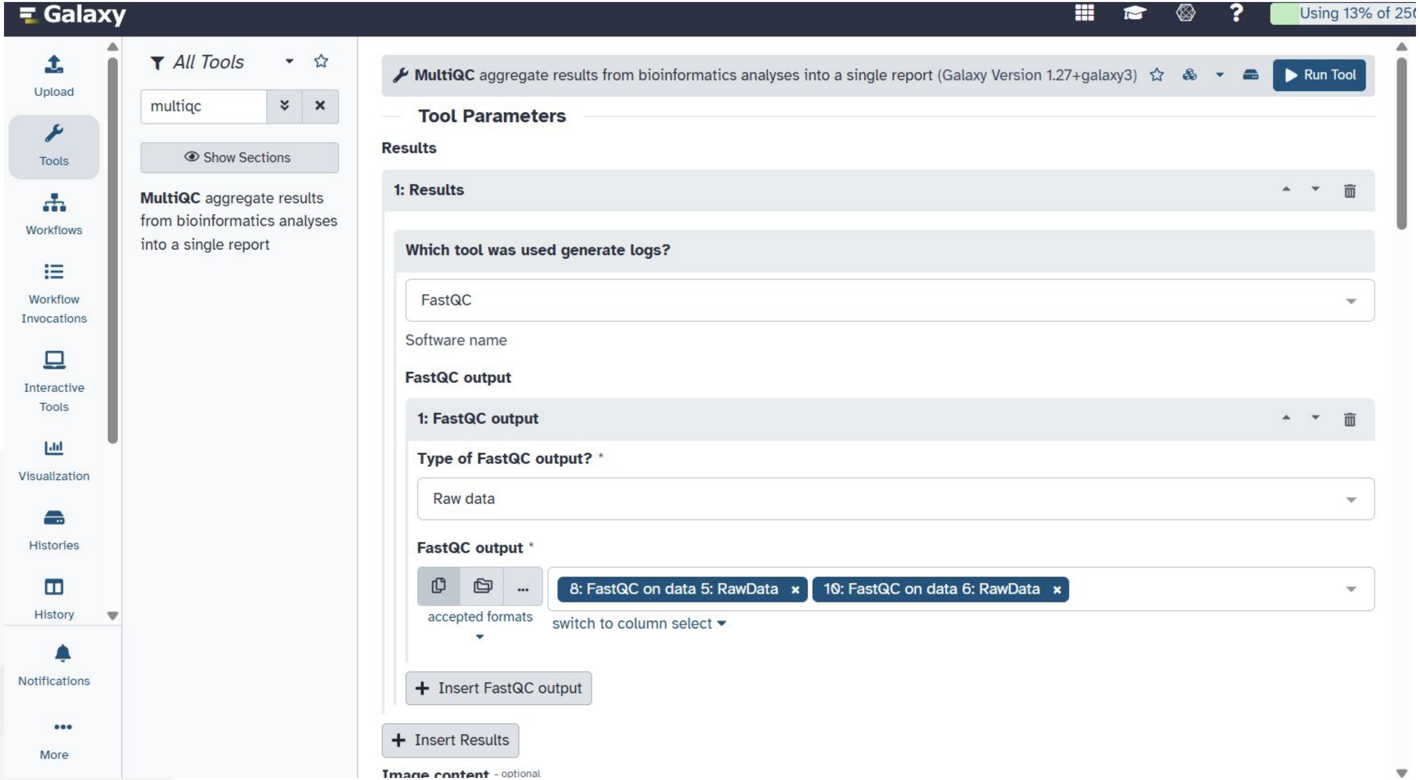
5. Add the raw data files generated earlier

6. **Leave** all other parameters at default

7. **Run** tool

8. **View** the result by clicking on the web page file produced

Quality Control (MultiQC)



multiqc

MultiQC aggregate results from bioinformatics analyses into a single report

MultiQC aggregate results from bioinformatics analyses into a single report (Galaxy Version 1.27+galaxy3)

Tool Parameters

Results

1: Results

Which tool was used generate logs?

FastQC

Software name

FastQC output

1: FastQC output

Type of FastQC output? *

Raw data

FastQC output *

8: FastQC on data 5: RawData × 10: FastQC on data 6: RawData ×

accepted formats

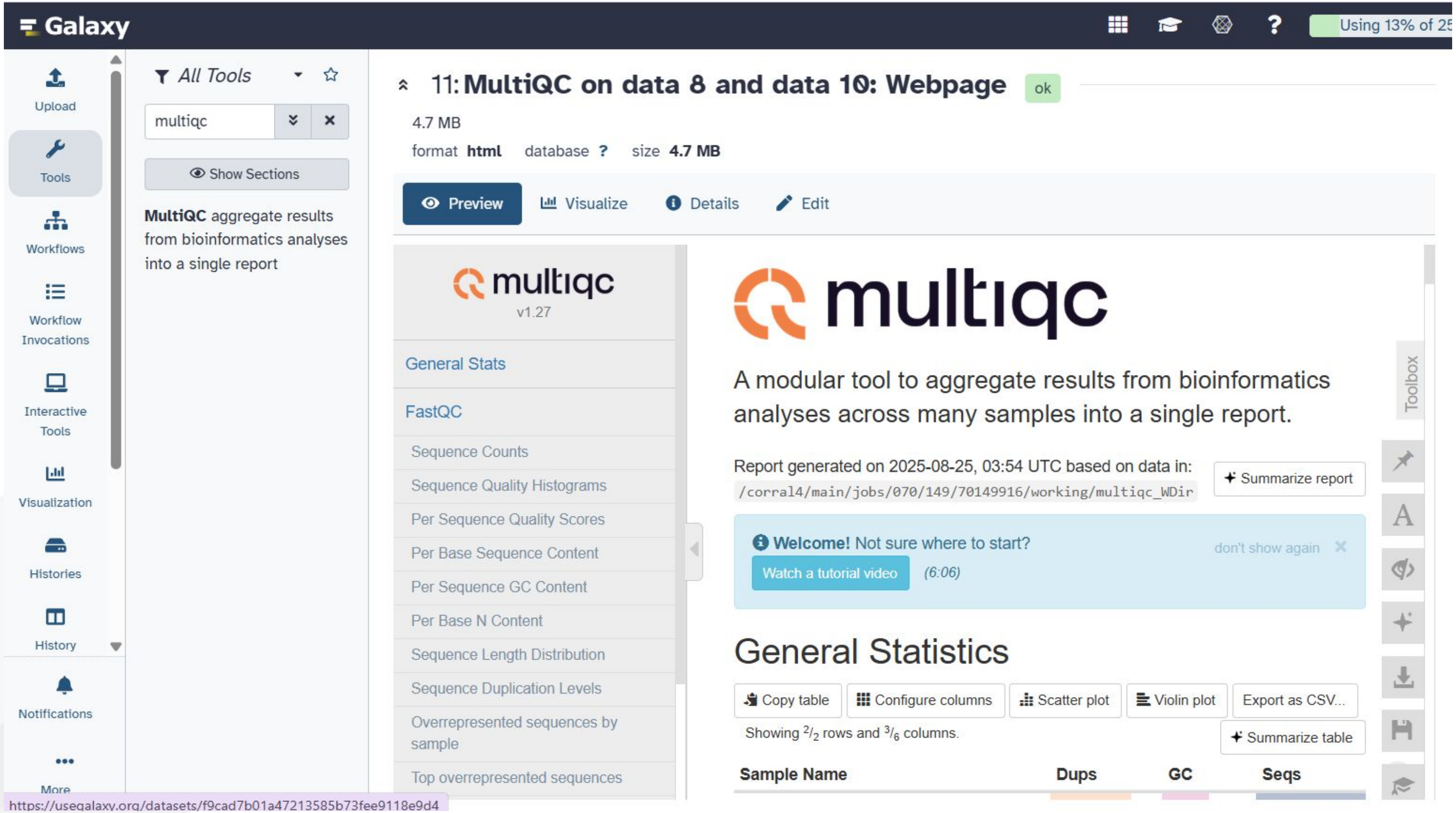
switch to column select

+ Insert FastQC output

+ Insert Results

Image content - optional

Quality Control Result



MultiQC aggregate results from bioinformatics analyses into a single report

11: MultiQC on data 8 and data 10: Webpage ok

4.7 MB

format **html** database **?** size **4.7 MB**

multiqc v1.27

multiqc

A modular tool to aggregate results from bioinformatics analyses across many samples into a single report.

Report generated on 2025-08-25, 03:54 UTC based on data in:
`/corral4/main/jobs/070/149/70149916/working/multiqc_WDir`

Welcome! Not sure where to start? don't show again ×

(6:06)

General Statistics

Showing $^2/_2$ rows and $^3/_6$ columns.

Sample Name	Dups	GC	Seqs

https://usegalaxy.org/datasets/f9cad7b01a47213585b73fee9118e9d4

Why QC is Important

- Ensures data integrity before downstream analysis.
- Detects contamination, errors, or poor-quality sequences.
- Prevents misleading results.

Why QC is Important

- Ensures data integrity before downstream analysis.
- Detects contamination, errors, or poor-quality sequences.
- Prevents misleading results.