

National Aeronautics and
Space Administration

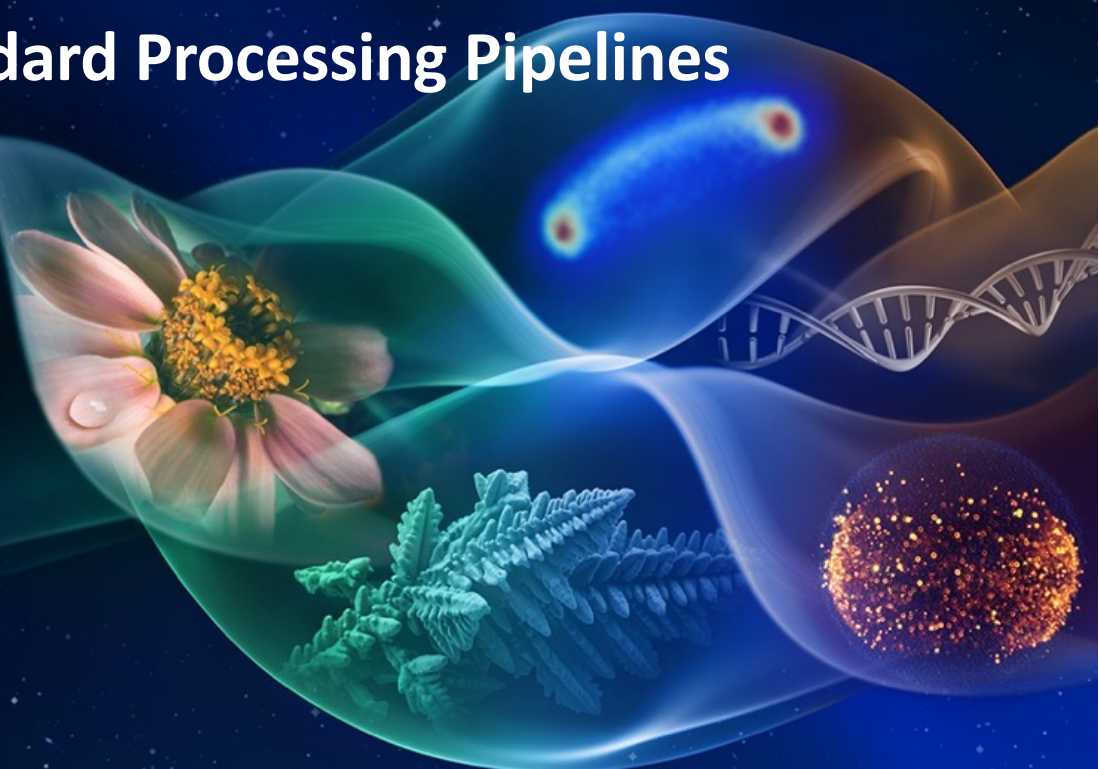


GeneLab Standard Processing Pipelines

Biological and Physical
Sciences

Mike Lee, Ph.D
Amanda M. Saravia-Butler, Ph.D.
GeneLab Data Processing Team
NASA Ames Research Center

GeneLab Analysis Working Group Workshop (10/2021)



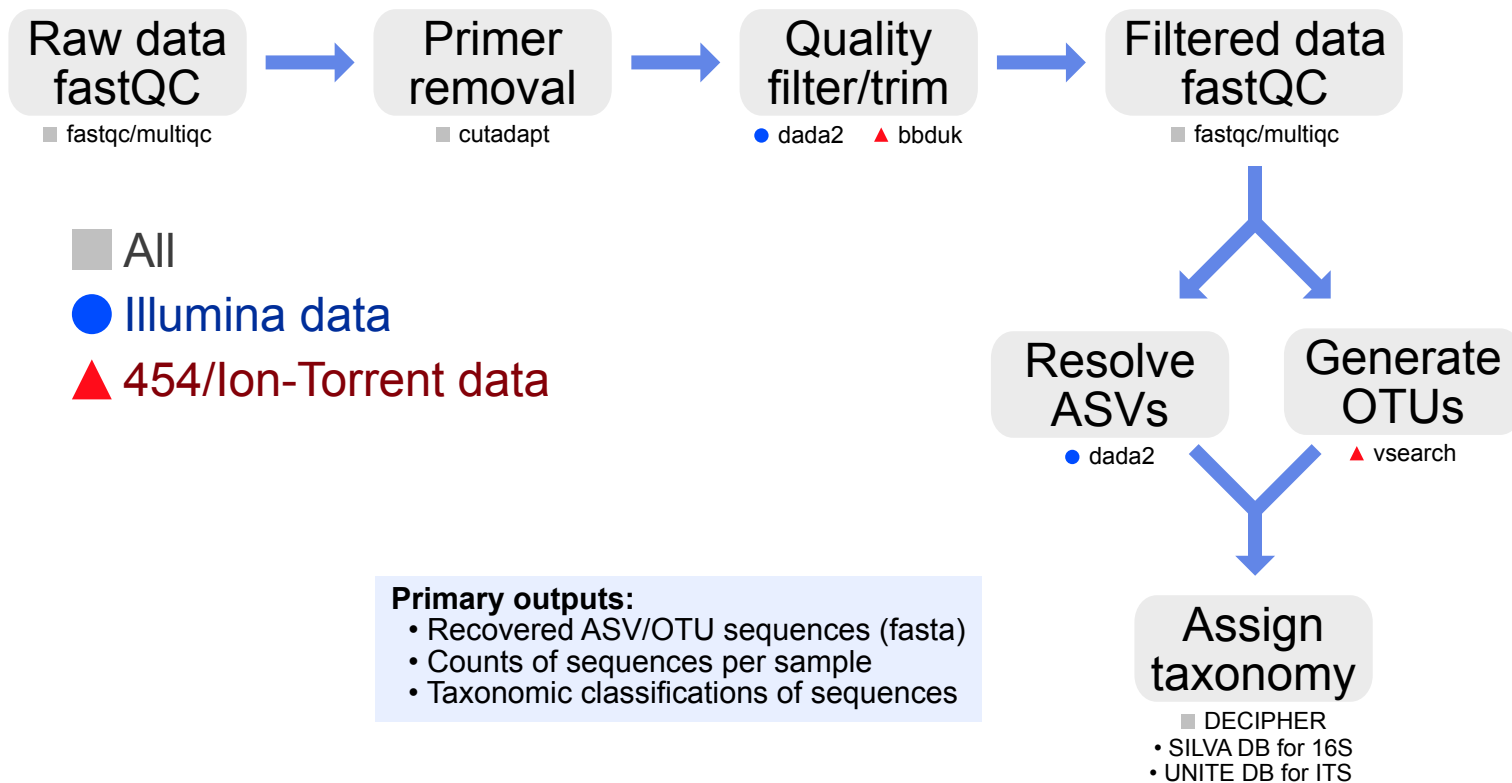
Amplicon Seq



Amplicon Seq Data Processing Pipeline



Amplicon Processing Overview



Amplicon Seq Data Processing Pipeline on GitHub



 [nasa / GeneLab_Data_Processing](#) Public

[Code](#) [Issues](#) [Pull requests](#) 1 [Actions](#) [Projects](#) [Wiki](#) [Security](#) [Insights](#)

 master [GeneLab_Data_Processing / Amplicon / Illumina /](#)

...

 asaravia-butler Adding NASA and 3rd party software licensing info for Amplicon and Me... ..	on Sep 11 History
..	
 GLDS_Processing_Scripts	last month
 workflow-template	last month
 GL-DPPD-7104-A.md	last month
 README.md	last month
 workflow-usage.md	last month

Repository links

- [Current processing protocol \(GL-DPPD-7104-A.md\)](#)
- [Processed dataset files](#)
- [Template workflow files](#)
- [Workflow information and usage instructions](#)

https://github.com/nasa/GeneLab_Data_Processing

Amplicon Seq Data Processing Workflow on GitHub



1. Install conda and `genelab-utils` package

We recommend installing a Miniconda, Python3 version appropriate for your system, as exemplified in [the above link](#).

Once conda is installed on your system, you can install the genelab-utils conda package in a new environment with the following:

```
conda create -n genelab-utils -c conda-forge -c bioconda -c defaults -c astrobiomike genelab-utils
```

The environment then needs to be activated:

```
conda activate genelab-utils
```

2. Download the workflow template files

All files required for utilizing the GeneLab workflow for removing human reads from metagenomics data are in the [workflow-template](#) directory. To get a copy of that directory on to your system, run the following command:

```
GL-get-Illumina-amplicon-wf
```

3. Modify the variables in the `config.yaml` file

Once you've downloaded the workflow template, you can modify the variables in the [config.yaml](#) file as needed. For example, you will have to provide a text file containing a single-column list of unique sample identifiers (see an example of how to set this up below). You will also need to indicate the paths to your input data (raw reads) and, if necessary, modify each variable to be consistent with the study you want to process.

Note: If you are unfamiliar with how to specify paths, one place you can learn more is [here](#).

```
14 ## single-column file with unique sample identifiers:
15 sample_info_file:
16     "unique-sample-IDs.txt"
17
18 ## input reads directory (can be relative to workflow directory, or needs to be full path)
19 raw_reads_dir:
20     "../Raw_Sequence_Data/"
```

https://github.com/nasa/GeneLab_Data_Processing

Workflow information and usage instructions

General workflow info

The current processing protocol is implemented as a [Snakemake](#) workflow and utilizes [conda](#) environments. The workflow can be used even if you are unfamiliar with Snakemake and conda, but if you want to learn more about those, [this Snakemake tutorial](#) within [Snakemake's documentation](#) is a good place to start for that, and an introduction to conda with installation help and links to other resources can be found [here at Happy Belly Bioinformatics](#).

Utilizing the workflow

1. [Install conda and `genelab-utils` package](#)
2. [Download the workflow template files](#)
3. [Modify the variables in the `config.yaml` file](#)
4. [Run the workflow](#)

4. Run the workflow

While in the directory holding the Snakefile, config.yaml, and other workflow files that you downloaded in step 2, here is one example command of how to run the workflow:

```
snakemake --use-conda --conda-prefix ${CONDA_PREFIX}/envs -j 2 -p
```

Parameter Definitions:

- `--use-conda` – specifies to use the conda environments included in the workflow (these are specified in the [envs](#) directory)
- `--conda-prefix` – indicates where the needed conda environments will be stored. Adding this option will also allow the same conda environments to be re-used when processing additional datasets, rather than making new environments each time you run the workflow. The value listed for this option, `${CONDA_PREFIX}/envs`, points to the default location for conda environments (note: the variable `${CONDA_PREFIX}` will be expanded to the appropriate location on whichever system it is run on).
- `-j` – assigns the number of jobs Snakemake should run concurrently
- `-p` – specifies to print out each command being run to the screen

See `snakemake -h` and [Snakemake's documentation](#) for more options and details.

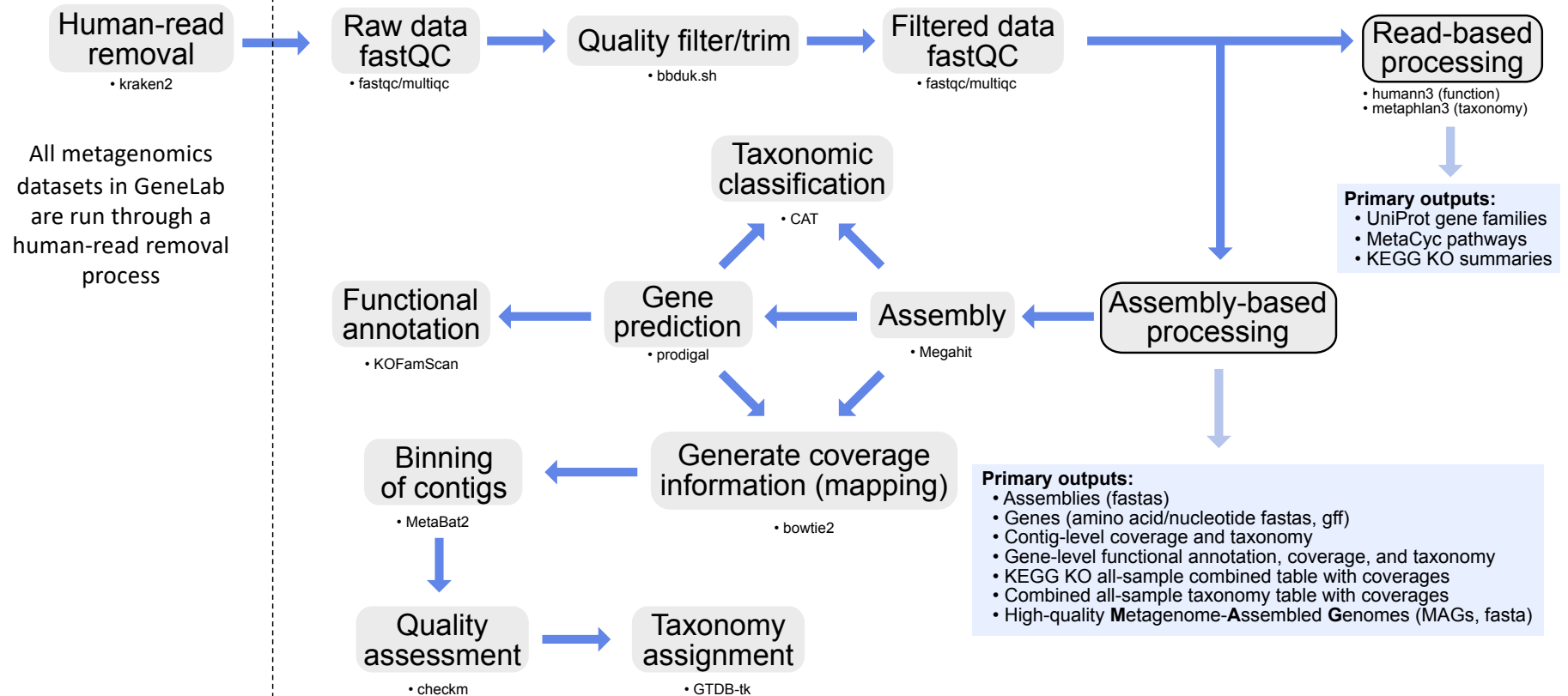
Metagenomics



Metagenomics Data Processing Pipeline



Illumina Metagenomics Overview



Metagenomics Data Processing Pipeline on GitHub



nasa / [GeneLab_Data_Processing](#) Public

[Code](#) [Issues](#) [Pull requests](#) 1 [Actions](#) [Projects](#) [Wiki](#) [Security](#)

[master](#) [GeneLab_Data_Processing / Metagenomics / Illumina /](#)

 asaravia-butler Update GL-DPPD-7107.md ...	on Sep 15 History
..	
 GLDS_Processing_Scripts	last month
 workflow-template	last month
 GL-DPPD-7107.md	last month
 README.md	last month
 workflow-usage.md	last month

Repository links

- [Current processing protocol \(GL-DPPD-7107.md\)](#)
- [Processed dataset files](#)
- [Template workflow files](#)
- [Workflow information and usage instructions](#)

https://github.com/nasa/GeneLab_Data_Processing

Metagenomics Data Processing Workflow on GitHub



1. Install conda and `genelab-utils` package

We recommend installing a Miniconda, Python3 version appropriate for your system, as exemplified in [the above link](#).

Once conda is installed on your system, you can install the genelab-utils conda package in a new environment with the following:

```
conda create -n genelab-utils -c conda-forge -c bioconda -c defaults -c astrobiomike genelab-utils
```

The environment then needs to be activated:

```
conda activate genelab-utils
```

2. Download the workflow template files

All files required for utilizing the GeneLab workflow for removing human reads from metagenomics data are in the [workflow-template](#) directory. To get a copy of that directory on to your system, run the following command:

```
GL-get-Illumina-metagenomics-wf
```

3. Modify the variables in the `config.yaml` file

Once you've downloaded the workflow template, you can modify the variables in the [config.yaml](#) file as needed. For example, you will have to provide a text file containing a single-column list of unique sample identifiers (see an example of how to set this up below). You will also need to indicate the paths to your input data (raw reads) and, if necessary, modify each variable to be consistent with the study you want to process.

Note: If you are unfamiliar with how to specify paths, one place you can learn more is [here](#).

```
14 ## single-column file with unique sample identifiers:
15 sample_info_file:
16     "unique-sample-IDs.txt"
17
18 ## input reads directory (can be relative to workflow directory, or needs to be full path)
19 raw_reads_dir:
20     "../Raw_Sequence_Data/"
```

https://github.com/nasa/GeneLab_Data_Processing

Workflow information and usage instructions

General workflow info

The current processing protocol is implemented as a [Snakemake](#) workflow and utilizes [conda](#) environments. The workflow can be used even if you are unfamiliar with Snakemake and conda, but if you want to learn more about those, [this Snakemake tutorial](#) within [Snakemake's documentation](#) is a good place to start for that, and an introduction to conda with installation help and links to other resources can be found [here at Happy Belly Bioinformatics](#).

Utilizing the workflow

1. [Install conda and `genelab-utils` package](#)
2. [Download the workflow template files](#)
3. [Modify the variables in the `config.yaml` file](#)
4. [Run the workflow](#)

4. Run the workflow

While in the directory holding the Snakefile, config.yaml, and other workflow files that you downloaded in step 2, here is one example command of how to run the workflow:

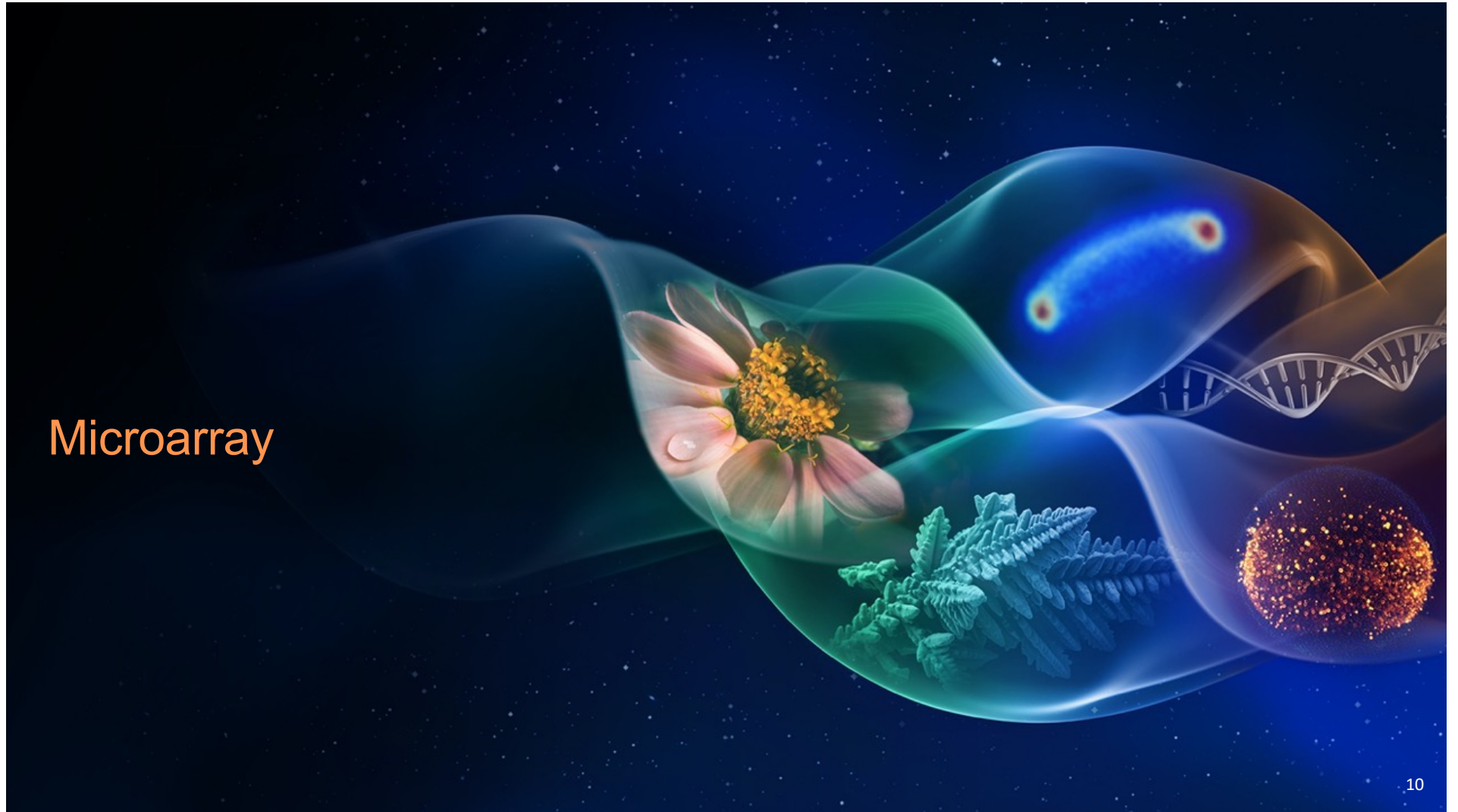
```
snakemake --use-conda --conda-prefix ${CONDA_PREFIX}/envs -j 2 -p
```

Parameter Definitions:

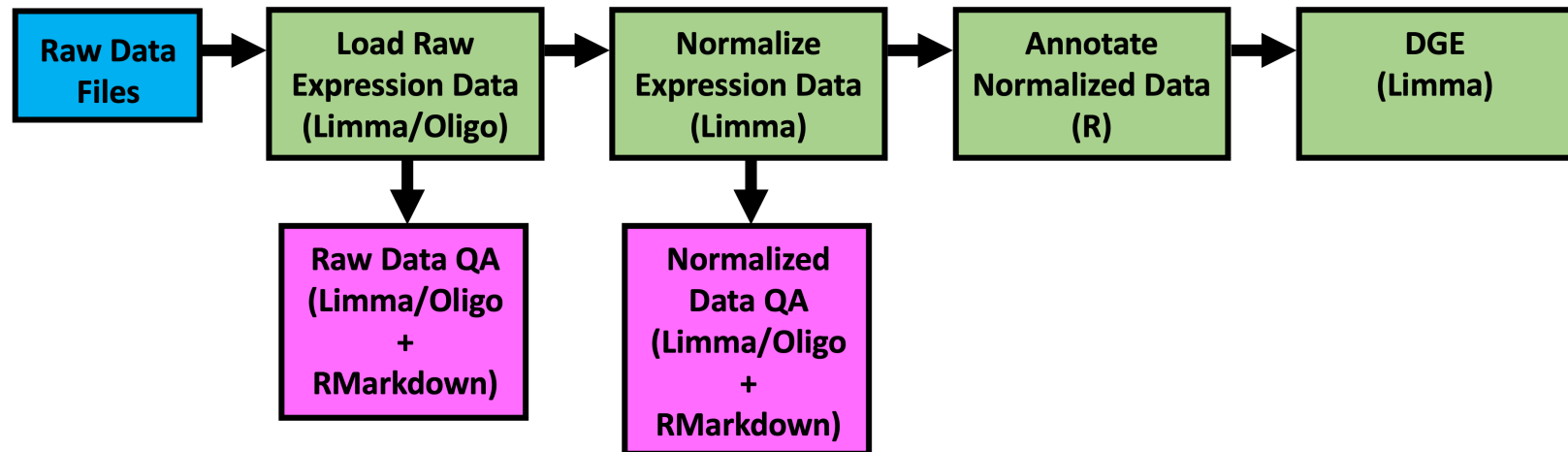
- `--use-conda` – specifies to use the conda environments included in the workflow (these are specified in the [envs](#) directory)
- `--conda-prefix` – indicates where the needed conda environments will be stored. Adding this option will also allow the same conda environments to be re-used when processing additional datasets, rather than making new environments each time you run the workflow. The value listed for this option, `${CONDA_PREFIX}/envs`, points to the default location for conda environments (note: the variable `${CONDA_PREFIX}` will be expanded to the appropriate location on whichever system it is run on).
- `-j` – assigns the number of jobs Snakemake should run concurrently
- `-p` – specifies to print out each command being run to the screen

See `snakemake -h` and [Snakemake's documentation](#) for more options and details.

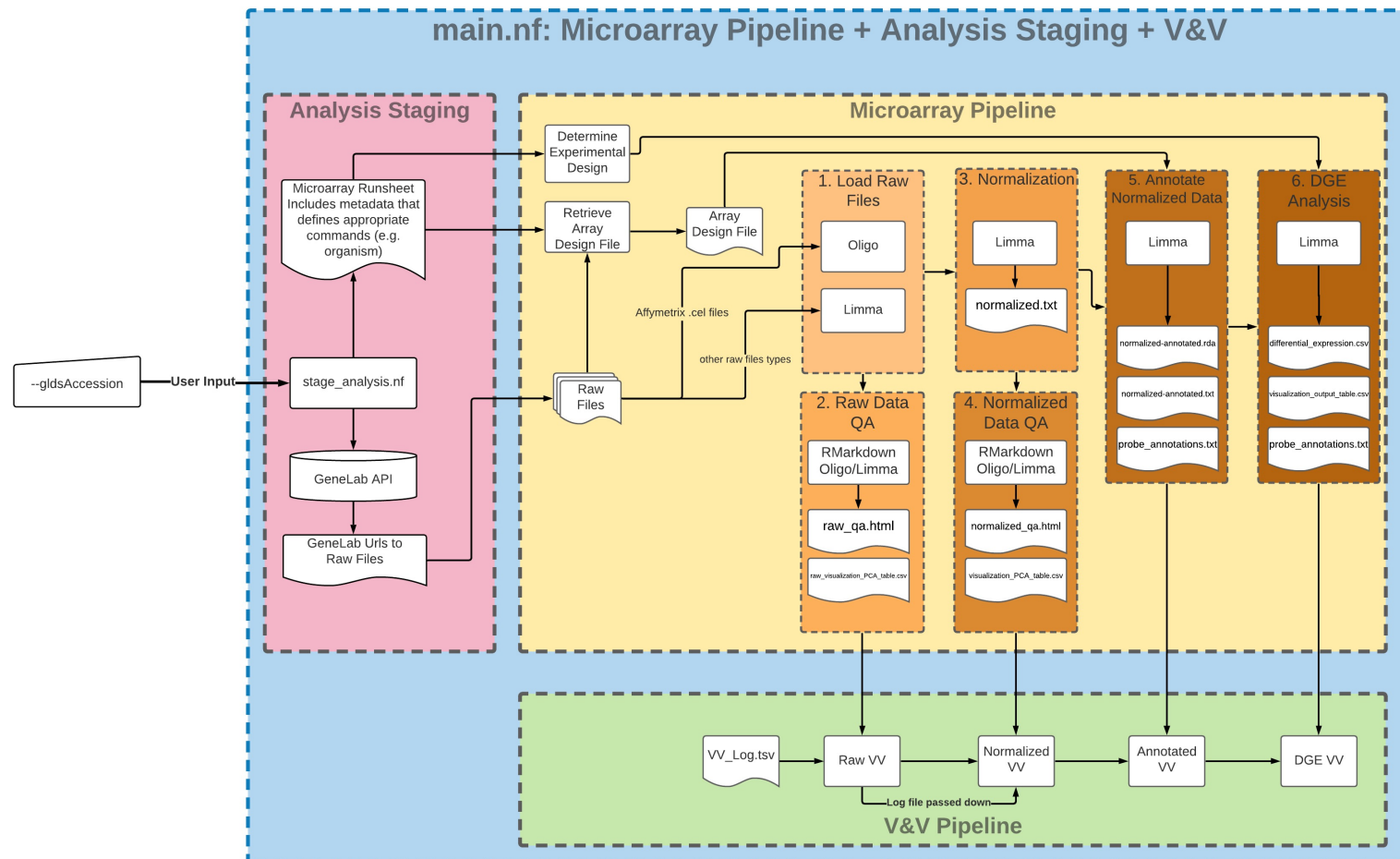
Microarray



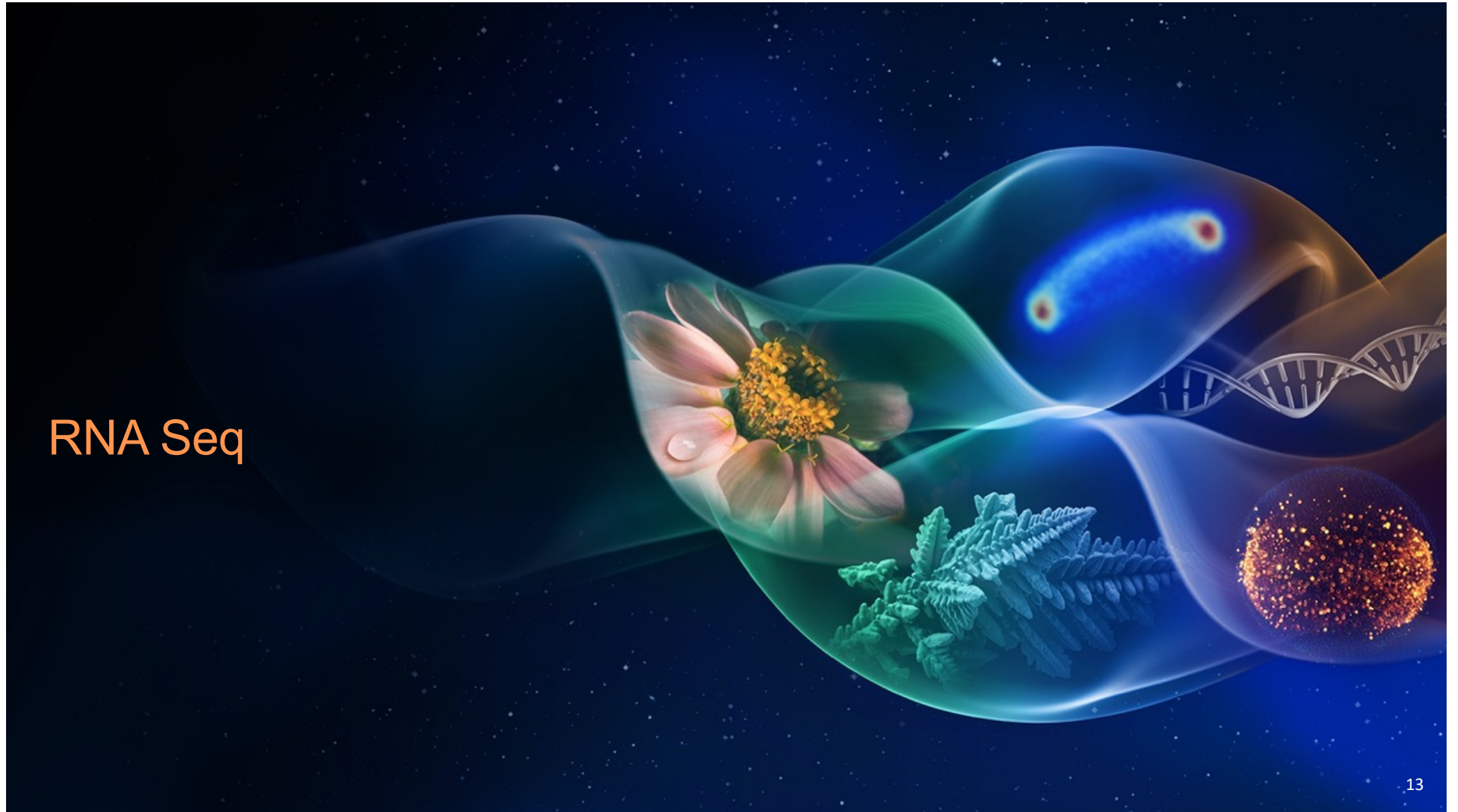
Microarray Data Processing Pipeline



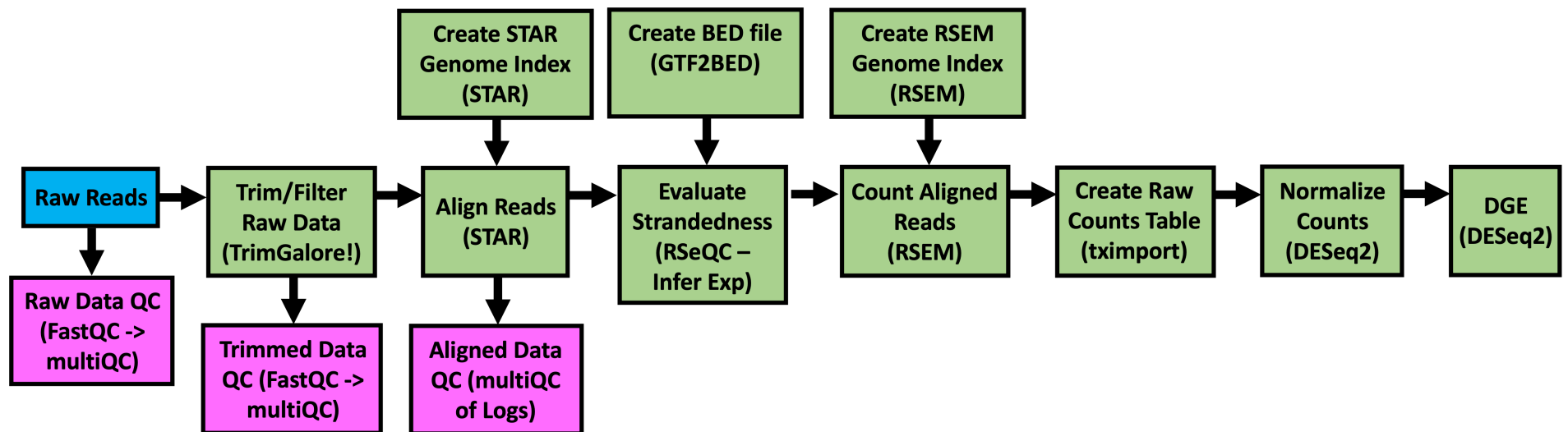
Microarray Data Processing Workflow with V&V – Coming Soon



RNA Seq



RNA Seq Data Processing Pipeline



RNA Seq Data Processing Pipeline on GitHub



nasa / [GeneLab_Data_Processing](#) Public

[Code](#) [Issues](#) [Pull requests](#) 1 [Actions](#) [Projects](#) [Wiki](#) [Security](#)

master [GeneLab_Data_Processing / RNAseq /](#)

...



asaravia-butler Update README.md ...

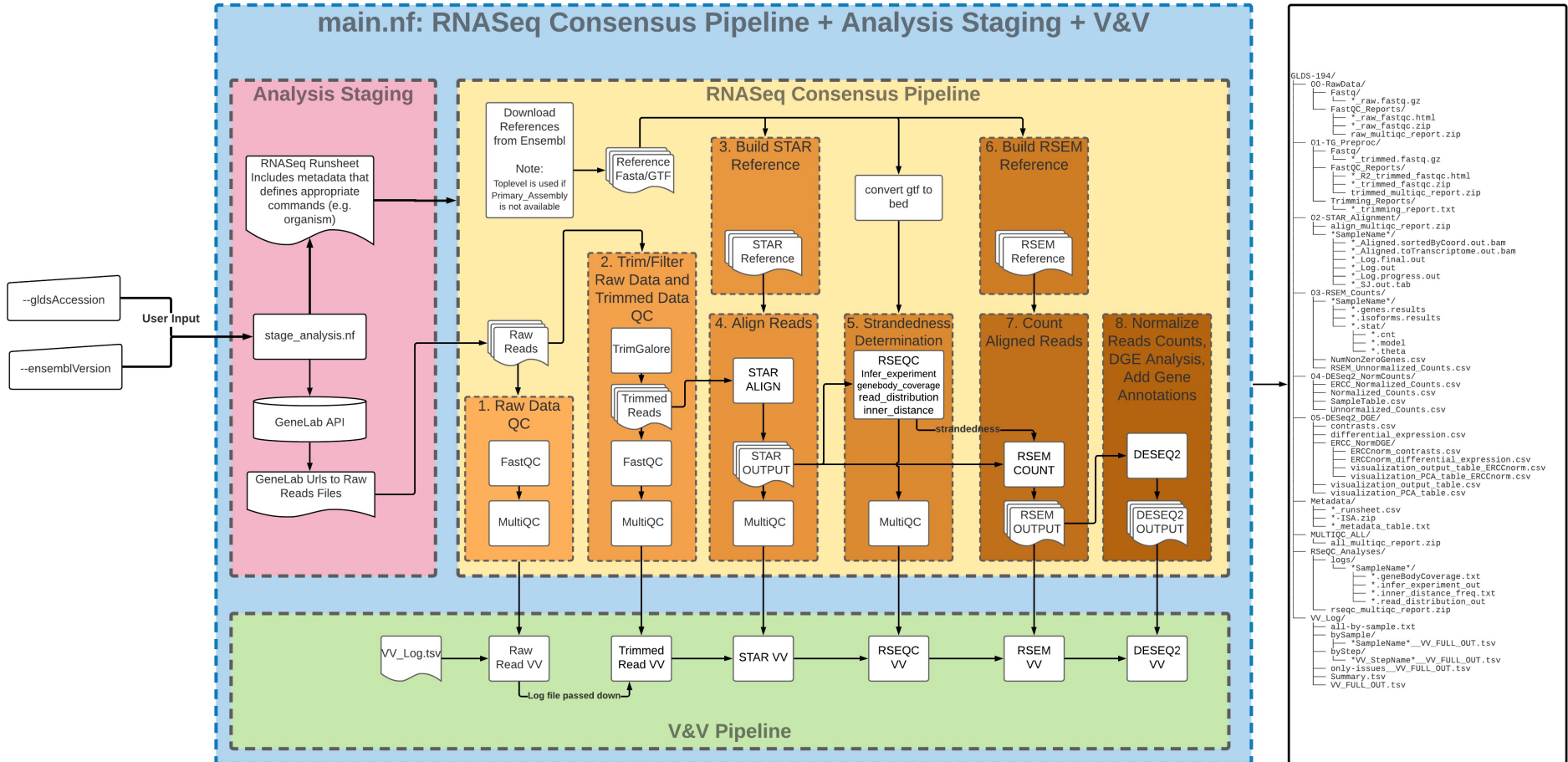
1 hour ago [History](#)

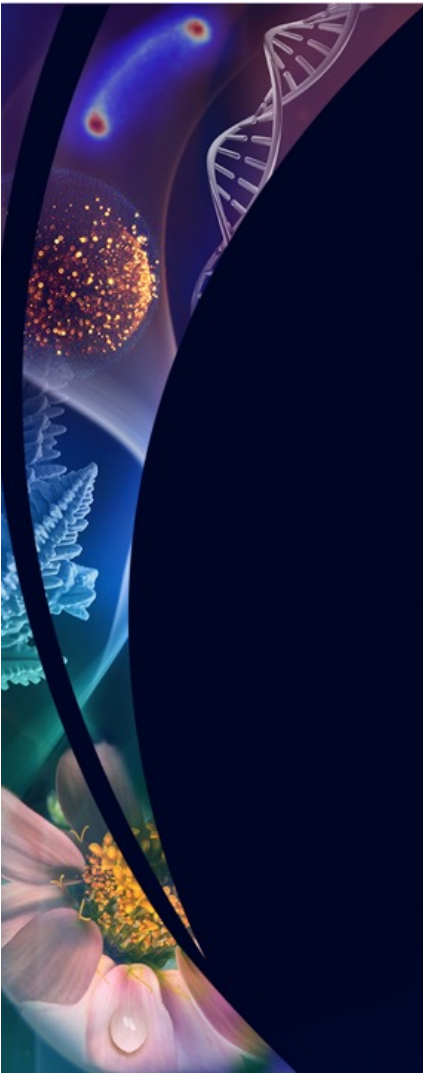
..

GLDS_Processing_Scripts	4 months ago
GeneLab_Reference_and_Annotation_Files	11 months ago
Previous_GL-DPPD-7101_Revisions	13 months ago
RNAseq_Tool_Install	9 months ago
images	1 hour ago
GL-DPPD-7101-C.md	27 days ago
README.md	1 hour ago
organisms.csv	13 months ago

https://github.com/nasa/GeneLab_Data_Processing

RNA Seq Data Processing Workflow with V&V – Coming Soon





Questions?