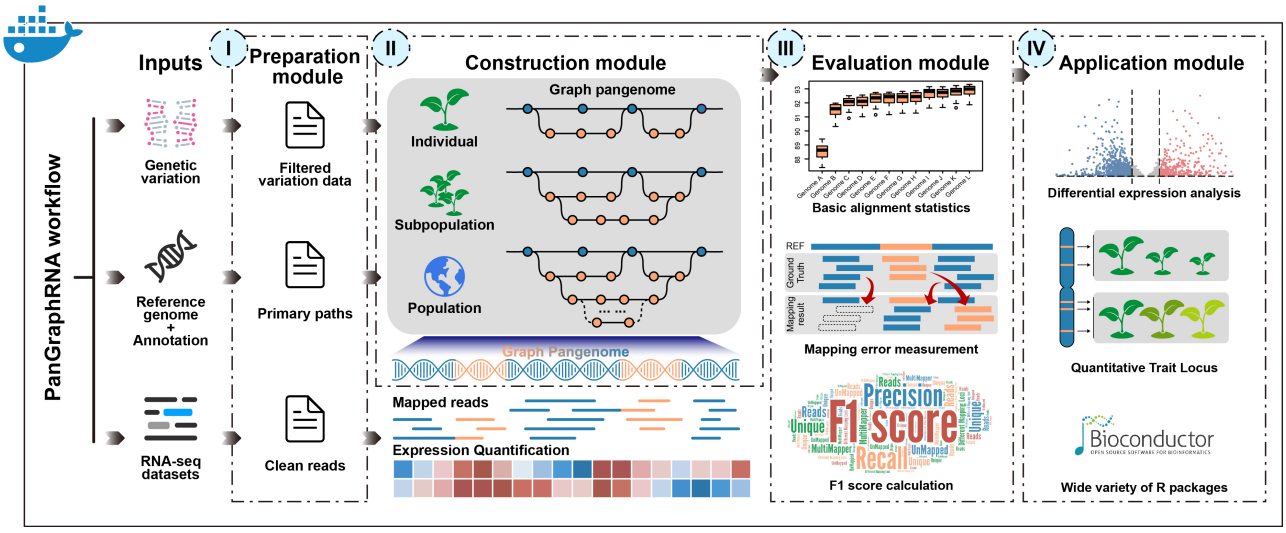
PanGraphRNA: an efficient and flexible bioinformatics platform for graph pangenome-based RNA-seq data analysis

Introduction

- PanGraphRNA is an efficient, flexible and web-based Galaxy platform that can be easily used to construct graph pangenomes from genetic variations at individual, subpopulation, and population levels. It can assist researchers to select appropriate graph pangenomes using various performance metrics for both real and simulation experiments. Currently, PanGraphRNA is composed of four graph pangenome-related functional modules: **Preparation Module, Construction Module, Evaluation Module, and Application Module**.
- The PanGraphRNA project is hosted on https://github.com/cma2015/PanGraphRNA.
- The PanGraphRNA Docker image can be obtained from https://hub.docker.com/r/malab/pangraphrna.



How to use PanGraphRNA

- Test data and tutorial of PanGraphRNA are presented at https://github.com/cma2015/PanGraphRNA/tree/main/Tutorials
- The installation of PanGraphRNA
- The Graph Pangenome Preparation Module
- The Graph Pangenome Construction Module
- The Graph Pangenome Evaluation Module
- The Graph Pangenome Application Module

How to cite PanGraphRNA

- Bu, Y., Qiu, Z., Sun, W., Han, Y., Liu, Y., Yang, J., Song, M., Li, Z., Zhang, Y., & Ma, C. (2024). PanGraphRNA: an efficient and flexible bioinformatics platform for graph pangenome-based RNA-seq data analysis. (Submitted)

How to access help

- Comments/suggestions/bugs/issues are welcome reported here or contact: Yifan Bu b761307648@163.com, Minggui Song smg@nwafu.edu.cn or Chuang Ma chuangma2006@gmail.com

Change log

- 2024.10 Release PanGraphRNA v1.0
- 2022.09 we launched PanGraphRNA project

Quick start

- **Step 1**: Docker installation

  i) Docker installation and start (Official installation tutorial)

  For **Windows (Only available for Windows 10 Prefessional and Enterprise version):**

  - Download [Docker](https://download.docker.com/win/stable/Docker for Windows Installer.exe) for windows;
  - Double click the EXE file to open it;
  - Follow the wizard instruction and complete installation;
  - Search docker, select **Docker for Windows** in the search results and click it.

  For **Mac OS X (Test on macOS Sierra version 10.12.6 and macOS High Sierra version 10.13.3):**

  - Download Docker for Mac OS;
  - Double click the DMG file to open it;
  - Drag the docker into Applications and complete installation;
  - Start docker from Launchpad by click it.

  For **Ubuntu (Test on Ubuntu 18.04 LTS):**

- Go to Docker, choose your Ubuntu version, browse to **pool/stable** and choose **amd64, armhf, ppc64el or s390x**. Download the **DEB** file for the Docker version you want to install;
- Install Docker, supposing that the DEB file is download into following path:**"/home/docker-ce~ubuntu_amd64.deb"**

```
$ sudo dpkg -i /home/docker-ce<version-XXX>~ubuntu_amd64.deb
$ sudo apt-get install -f
```

ii) Verify if Docker is installed correctly

Once Docker installation is completed, we can run `hello-world` image to verify if Docker is installed correctly. Open terminal in Mac OS X and Linux operating system and open CMD for Windows operating system, then type the following command:

```
$ docker run hello-world
```

Note: root permission is required for Linux operating system.

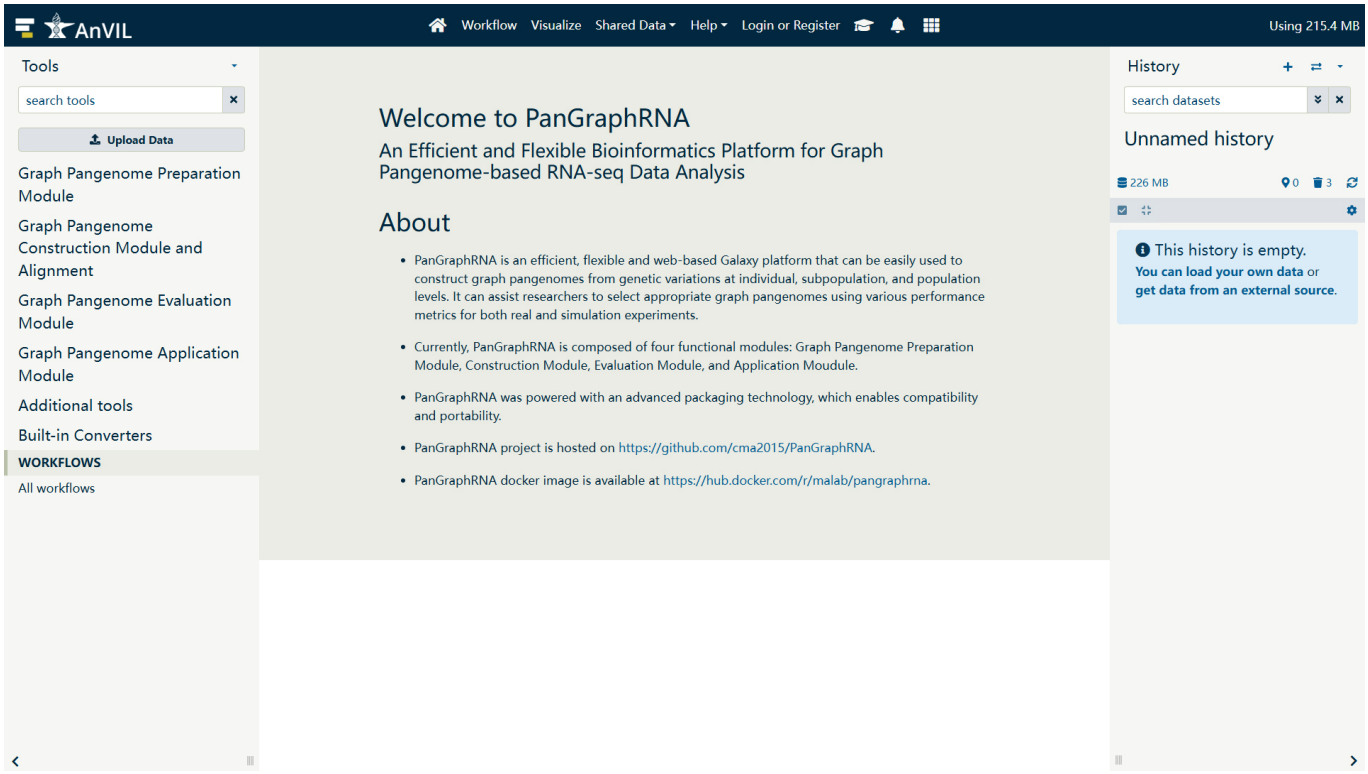
- **Step 2**: PanGraphRNA installation from Docker Hub

```
# pull latest PanGraphRNA Docker image from docker hub
$ docker pull malab/pangraphrna
```

- **Step 3**: Launch PanGraphRNA local server

```
$ docker run -it -p 880:8080 malab/pangraphrna bash
$ bash /home/galaxy/run.sh
```

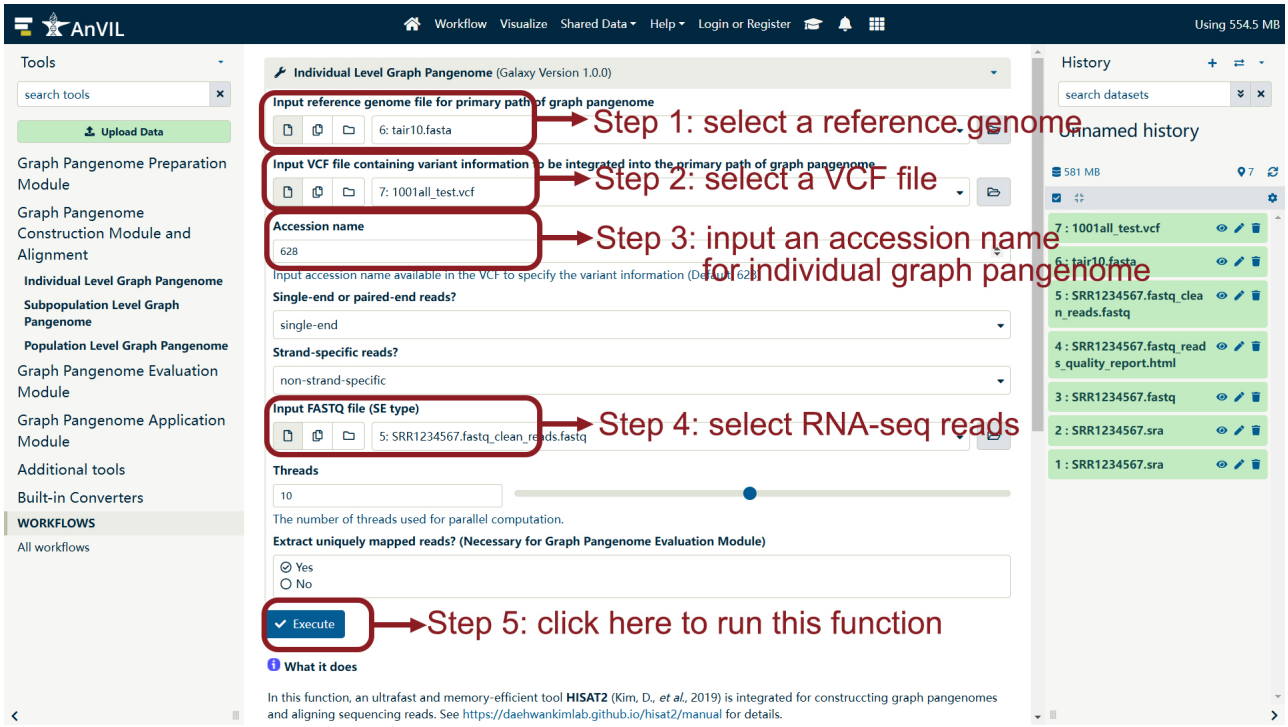
Then, PanGraphRNA local server can be accessed via http://localhost:8080



- **Step 4**: Upload RNA-seq, reference genome and variation data
  Specify the data format and upload a single-genome FASTA file reference genome to delineate primary paths, a GTF (general transfer format) file containing the gene annotations, a VCF (variant call format) file detailing genetic variations, a collection of RNA-seq FASTQ files and any of other file required in functions.
  Here, we have provided small test files in the `testdata.tar.gz` to help users quickly get started with PanGraphRNA.

- **Step 5**: Construct graph pangenome and perform read-genome alignment
  Then we can construct graph pangenomes at the individual, subpopulation, or population level. Subsequently, this module performs read-genome alignment.

- Individual Level Graph Pangenome
  Input reference genome file: Input reference genome file for primary path of graph pangenome in FASTA format (e.g. `Col0_test.fasta`)
  Input VCF file: Input VCF file containing variant information to be integrated into the primary path of graph pangenome in VCF format (e.g. `10015.vcf`, accession name `10015`)
  Input FASTQ file: Cleaned single-end or paired-end RNA-seq reads in FASTQ format (e.g. `test.fastq`)



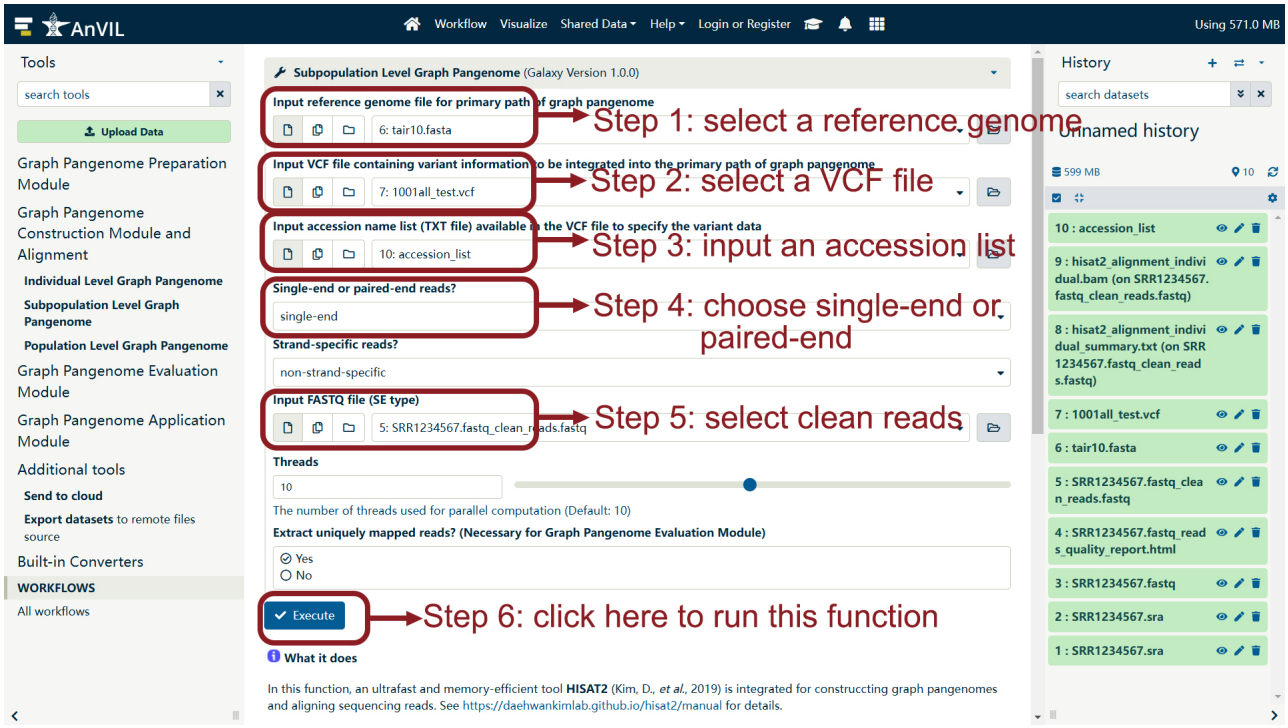
- Subpopulation Level Graph Pangenome

  Input reference genome file: Input reference genome file for primary path of graph pangenome in FASTA format

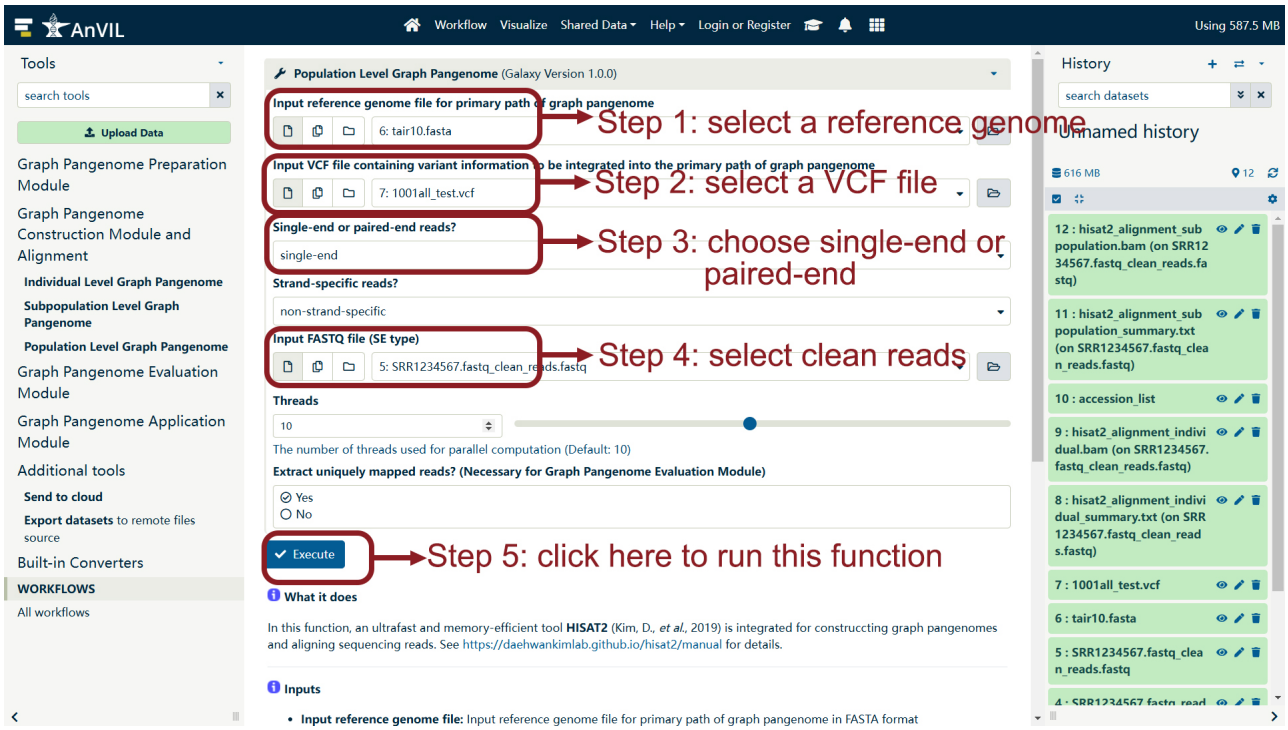
  Input VCF file: Input VCF file containing variant information to be integrated into the primary path of graph pangenome in VCF format

  Input accession name list: Input accession name list (TXT file) available in the VCF file to specify the variant data

  Input FASTQ file: Cleaned single-end or paired-end RNA-seq reads in FASTQ format

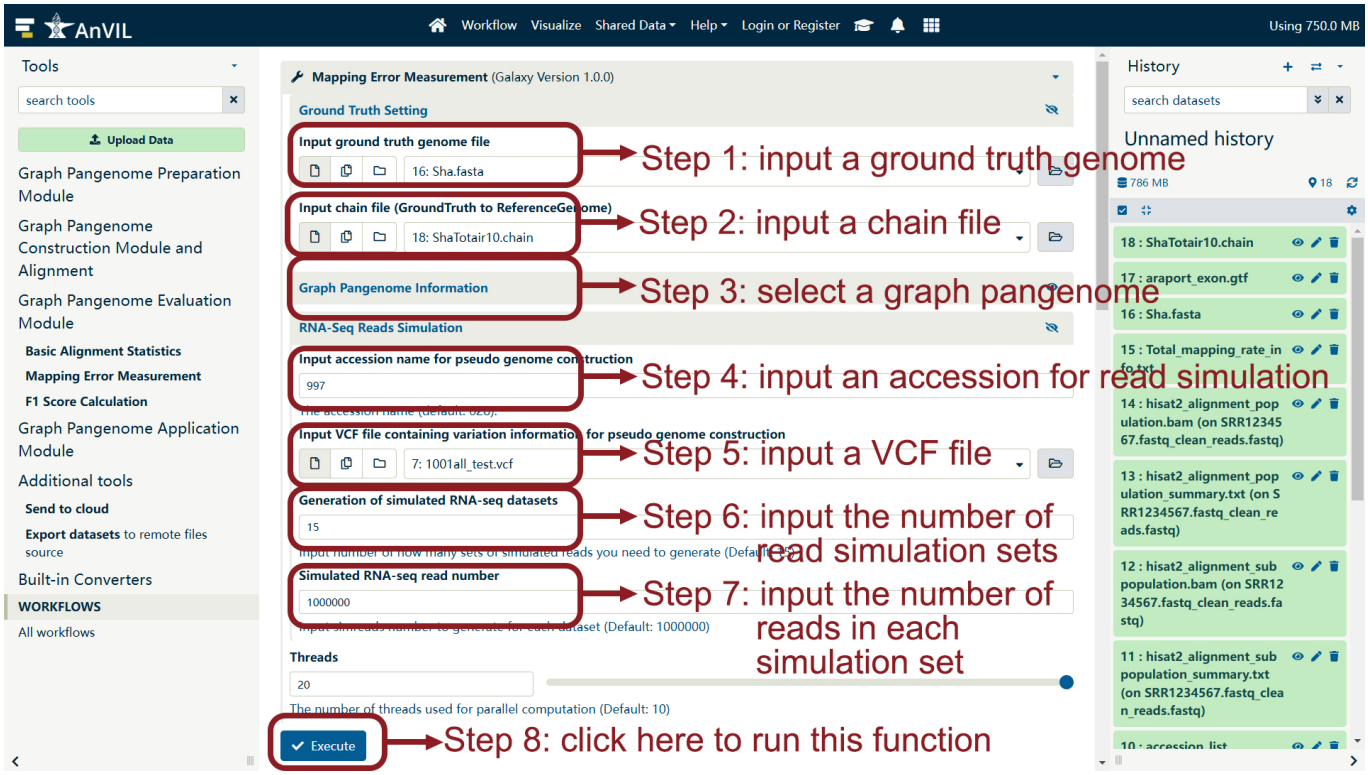


- Population Level Graph
  Input reference genome file: Input reference genome file for primary path of graph pangenome in FASTA format
  Input VCF file: Input VCF file containing variant information to be integrated into the primary path of graph pangenome in VCF format
  Input FASTQ file: Cleaned single-end or paired-end RNA-seq reads in FASTQ format



- **Step 6**: Assesse the performance of graph pangenomes against the SLR strategy
  Obtain the bascic alignment information among multiple read-genome alignment results (Alignment information matrix in TXT format).

Perform RNA-seq reads simulation and measure mapping errors (Measurement of mapping errors in CSV format).



Perform RNA-seq reads simulation and calculate F1 score (Calculation of F1 score in CSV format).


AnVIL
Workflow
Visualize
Shared Data
Help
Login or Register
Using 750.0 MB
Tools
search tools
Upload Data
Graph Pangenome Preparation Module
Graph Pangenome Construction Module and Alignment
Graph Pangenome Evaluation Module
Basic Alignment Statistics
Mapping Error Measurement
F1 Score Calculation
Graph Pangenome Application Module
Additional tools
Built-in Converters
WORKFLOWS
All workflows
Graph Pangenome Information
RNA-Seq Reads Simulation
Input accession name for pseudo genome construction
628
Input VCF file containing variation information for pseudo genome construction
7: 1001all_test.vcf
Generation of simulated RNA-seq datasets
15
Simulated RNA-seq read number
1000000
Threads
10
The number of threads used for parallel computation (Default: 10)
Execute
Step 1: select a graph pangenome
Step 2: input an accession for read simulation
Step 3: input a VCF file
Step 4: input the number of read simulation sets
Step 5: input the number of reads in each simulation set
Step 6: click here to run this function
What it does
Mapping error measurement is designed to calculate Recall, Precision and F1 score based on RNA-seq read-genome alignments. This function also enables Flux features (Griebel, T., et al., 2012) for generating simulated reads and performs read-genome alignment using HISAT2 (Kim, D., et al., 2019). For more details, see https://bio.tools/the_flux_simulator and https://daehwankimlab.github.io/hisat2/manual for details.
Measuring F1score
Recall = 100 X Num.correct / Num.reads
History
search datasets
Unnamed history
786 MB
19
12
31 : compare_readtype_amongGenomes.csv
18 : ShaTotair10.chain
17 : araport_exon.gtf
16 : Sha.fasta
15 : Total_mapping_rate_info.txt
14 : hisat2_alignment_population.bam (on SRR1234567.fastq_clean_reads.fastq)
13 : hisat2_alignment_population_summary.txt (on SRR1234567.fastq_clean_reads.fastq)
12 : hisat2_alignment_subpopulation.bam (on SRR1234567.fastq_clean_reads.fastq)
11 : hisat2_alignment_subpopulation_summary.txt