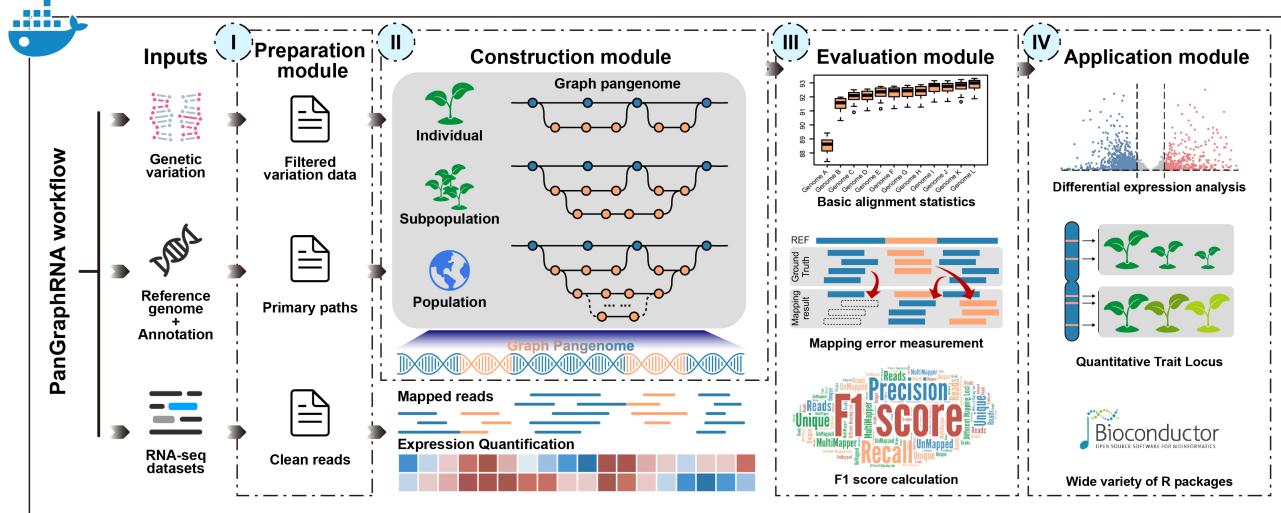


PanGraphRNA: an efficient and flexible bioinformatics platform for graph pangenome-based RNA-seq data analysis

Docker image ready docker pulls 58 Source codes support

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 Professional 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>

Welcome to PanGraphRNA
An Efficient and Flexible Bioinformatics Platform for Graph Pangenome-based RNA-seq Data Analysis

About

- 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 functional modules: Graph Pangenome Preparation Module, Construction Module, Evaluation Module, and Application Moudule.
- PanGraphRNA was powered with an advanced packaging technology, which enables compatibility and portability.
- PanGraphRNA project is hosted on <https://github.com/cma2015/PanGraphRNA>.
- PanGraphRNA docker image is available at <https://hub.docker.com/r/malab/pangraphna>.

- **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](#))

AnVIL

Workflow Visualize Shared Data Help Login or Register

Using 554.5 MB

Tools

Upload Data

Graph Pangenome Preparation Module

Graph Pangenome Construction Module and Alignment

Individual Level Graph Pangenome Subpopulation Level Graph Pangenome

Population Level Graph Pangenome

Graph Pangenome Evaluation Module

Graph Pangenome Application Module

Additional tools

Built-in Converters

WORKFLOWS

All workflows

Individual Level Graph Pangenome (Galaxy Version 1.0.0)

Step 1: select a reference genome
Input reference genome file for primary path of graph pangenome
6: tair10.fasta

Step 2: select a VCF file
Input VCF file containing variant information to be integrated into the primary path of graph pangenome
7: 1001all_test.vcf

Step 3: input an accession name
Accession name
628

Step 4: select RNA-seq reads
Input FASTQ file (SE type)
5: SRR1234567.fastq_clean_reads.fastq

Threads
10

Extract uniquely mapped reads? (Necessary for Graph Pangenome Evaluation Module)
Yes
No

Step 5: click here to run this function
Execute

In this function, an ultrafast and memory-efficient tool **HISAT2** (Kim, D., et al., 2019) is integrated for constructing graph pangenomes and aligning sequencing reads. See <https://daehwankimlab.github.io/hisat2/manual> for details.

History

search datasets

Unnamed history

581 MB

7 : 1001all_test.vcf
6 : tair10.fasta
5 : SRR1234567.fastq_clean_reads.fastq
4 : SRR1234567.fastq_reads_quality_report.html
3 : SRR1234567.fastq
2 : SRR1234567.sra
1 : SRR1234567.sra

- 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

AnVIL

Workflow Visualize Shared Data Help Login or Register

Using 571.0 MB

Tools

Upload Data

Graph Pangenome Preparation Module

Graph Pangenome Construction Module and Alignment

Individual Level Graph Pangenome Subpopulation Level Graph Pangenome

Population Level Graph Pangenome

Graph Pangenome Evaluation Module

Graph Pangenome Application Module

Additional tools

Send to cloud

Export datasets to remote files source

Built-in Converters

WORKFLOWS

All workflows

Subpopulation Level Graph Pangenome (Galaxy Version 1.0.0)

Step 1: select a reference genome
Input reference genome file for primary path of graph pangenome
6: tair10.fasta

Step 2: select a VCF file
Input VCF file containing variant information to be integrated into the primary path of graph pangenome
7: 1001all_test.vcf

Step 3: input an accession list
Input accession name list (TXT file) available in the VCF file to specify the variant data
10: accession_list

Step 4: choose single-end or paired-end
Single-end or paired-end reads?
single-end

Step 5: select clean reads
Input FASTQ file (SE type)
5: SRR1234567.fastq_clean_reads.fastq

Threads
10

The number of threads used for parallel computation (Default: 10)

Extract uniquely mapped reads? (Necessary for Graph Pangenome Evaluation Module)
Yes
No

Step 6: click here to run this function
Execute

In this function, an ultrafast and memory-efficient tool **HISAT2** (Kim, D., et al., 2019) is integrated for constructing graph pangenomes and aligning sequencing reads. See <https://daehwankimlab.github.io/hisat2/manual> for details.

History

search datasets

Unnamed history

599 MB

10 : accession_list
9 : hisat2_alignment_indiv dual.bam (on SRR1234567. fastq_clean_reads.fastq)
8 : hisat2_alignment_indiv dual_summary.txt (on SRR1234567. fastq_clean_reads.fastq)
7 : 1001all_test.vcf
6 : tair10.fasta
5 : SRR1234567.fastq_clean_reads.fastq
4 : SRR1234567.fastq_reads_quality_report.html
3 : SRR1234567.fastq
2 : SRR1234567.sra
1 : SRR1234567.sra

- 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

The screenshot shows the 'Population Level Graph Pangenome' tool in Galaxy Version 1.0.0. The interface includes a sidebar with various tools and a main panel for input parameters. The main panel has five highlighted steps:

- Step 1: select a reference genome (input tair10.fasta)
- Step 2: select a VCF file (input 1001all_test.vcf)
- Step 3: choose single-end or paired-end (single-end selected)
- Step 4: select clean reads (input SRR1234567.fastq_clean_reads.fastq)
- Step 5: click here to run this function (Execute button)

The right side of the interface shows a history list of datasets used in the workflow.

- **Step 6:** Assess the performance of graph pangenomes against the SLR strategy

Obtain the basic alignment information among multiple read-genome alignment results (Alignment information matrix in TXT format).

AnVIL

Workflow Visualize Shared Data Help Login or Register Using 604.0 MB

Tools search tools

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

Send to cloud

Export datasets to remote files source

Built-in Converters

WORKFLOWS All workflows

Basic Alignment Statistics (Galaxy Version 1.0.0)

Input HISAT2 alignment report file

1: Input HISAT2 alignment report file
Alignment summary file generated by HISAT2
8: hisat2_alignment_individual_summary.txt (on SRR1234567.fastq_clean_reads.fastq)

+ Insert Input HISAT2 alignment report file
input reference genome file for primary path of graph pangenome in FASTA format
6: tair10.fasta

Step 4: click here to run this function

What it does This function is designed to count uniquely mapped reads, multiply mapped reads and calculate alignment rates among multiple read-genome alignment results.

Inputs

- Input HISAT2 alignment report file: Input alignment summary files generated by HISAT2 in TXT format
- Input reference genome file: Input reference genome file for primary path of graph pangenome in FASTA format

Outputs

- Alignment information matrix in TXT format

Citations: - Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT and HISAT-genotype. *Nature Biotechnology*, 37(8), 907–915. <https://doi.org/10.1038/s41587-019-0201-4>

History search datasets Unnamed history

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_sub_population.bam (on SRR1234567.fastq_clean_reads.fastq)
11 : hisat2_alignment_sub_population_summary.txt (on SRR1234567.fastq_clean_reads.fastq)
10 : accession_list
9 : hisat2_alignment_individual.bam (on SRR1234567.fastq_clean_reads.fastq)
8 : hisat2_alignment_individual_summary.txt (on SRR1234567.fastq_clean_reads.fastq)

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

AnVIL

Workflow Visualize Shared Data Help Login or Register Using 750.0 MB

Tools search tools

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

Send to cloud

Export datasets to remote files source

Built-in Converters

WORKFLOWS All workflows

Mapping Error Measurement (Galaxy Version 1.0.0)

Ground Truth Setting

Input ground truth genome file
16: Sha.fasta

Input chain file (GroundTruth to ReferenceGenome)
18: ShaTotair10.chain

Graph Pangenome Information

RNA-Seq Reads Simulation

Input accession name for pseudo genome construction
997
The accession name (default: 626).

Input VCF file containing variation information for pseudo genome construction
7: 1001all_test.vcf

Generation of simulated RNA-seq datasets
15
Input number of how many sets of simulated reads you need to generate (Default: 15).

Simulated RNA-seq read number
1000000
Input simulated number to generate for each dataset (Default: 1000000)

Threads
20
The number of threads used for parallel computation (Default: 10)

Step 8: click here to run this function

History search datasets Unnamed history

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_sub_population.bam (on SRR1234567.fastq_clean_reads.fastq)
11 : hisat2_alignment_sub_population_summary.txt (on SRR1234567.fastq_clean_reads.fastq)
10 : accession_list

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

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

F1 Score Calculation (Galaxy Version 1.0.0)

Step 1: select a graph pangenome

Graph Pangenome Information

Step 2: input an accession for read simulation

RNA-Seq Reads Simulation

Input accession name for pseudo genome construction
628

The accession name (Default: 628).

Input VCF file containing variation information for pseudo genome construction
7: 1001all_test.vcf

Step 3: input a VCF file

Generation of simulated RNA-seq datasets
15

Input number of how many sets of simulated reads you need to generate (Default: 15).

Simulated RNA-seq read number
1000000

Input simmed number to generate for each dataset (Default: 1000000)

Step 4: input the number of read simulation sets

Threads
10

The number of threads used for parallel computation (Default: 10)

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 = $100 \times \text{Num.correct} / \text{Num.reads}$
- Precision = $100 \times \text{Num.correct} / \text{Num.unique}$

History search datasets

Unnamed history

786 MB 19 12

- 31 : compare_readtype_annotationGenomes.csv
- 18 : ShaTotal10.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_sub_population.bam (on SRR1234567.fastq_clean_reads.fastq)
- 11 : hisat2_alignment_sub_population_summary.txt (on SRR1234567.fastq_clean_reads.fastq)