

PanGraphRNA User Manual

(version 1.0)

- 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/pangraphrna>

Graph Pangenome Evaluation Module

This module assesses the performance of graph pangenomes against the SLR strategy with six read mapping-relevant measurements: *Unique alignment rate, Multiple alignment rate, Error mapping rate, F1 score, Recall and Precision.*

Tools	Description	Input	Output	Time (test data)	Reference
Basic Alignment Statistics	Obtain the basic alignment information among multiple read-genome alignment results	HISAT2 alignment report in TXT format and graph pangenome files	Basic alignment statistics matrix in TXT format	~1 mins	/

Tools	Description	Input	Output	Time (test data)	Reference
Mapping Error Measurement	Perform RNA-seq reads simulation and measure mapping errors	Ground truth files and graph pangenome files	Measurement of mapping errors in CSV format	~10 mins	Flux and HISAT2
F1 Score Calculation	Perform RNA-seq reads simulation and calculate F1 score	Graph pangenome files	Calculation of F1 score in CSV format	~10 mins	Flux and HISAT2

Basic Alignment Statistics

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

Input

- **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

Output

- **Alignment information matrix in TXT format**

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 pang genome in FASTA format

6: tair10.fasta

Execute

Step 1: select an alignment report

Step 2: choose to add more

Step 3: choose the reference genome used in previous functions

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 pang genome 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 HISAT2 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_subpopulation.bam (on SRR1234567.fastq_clean_reads.fastq)

11: hisat2_alignment_subpopulation_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)

Mapping Error Measurement

Mapping error measurement is designed to identify incorrectly mapped reads and record their error types in 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 mapping errors

The alignment results for different reference genomes were compared with the ground truth, and the five error types were defined according to the read alignment type and position.

- Five **mapping error** types:

Mapping error type	Groundtruth alignment result (read type)	Alignment result (read type)
Mapped to UnMapped	Mapped read	Unmapped read
UnMapped to Mapped	Unmapped read	Mapped read

Mapping error type	Groundtruth alignment result (read type)	Alignment result (read type)
Unique to MultiMapper	Uniquely mapped read	Multiply mapped read
MultiMapper to Unique	Multiply mapped read	Uniquely mapped read
Different Mapping Loci	Mapped to A region	Mapped to B region

Input

- **Ground Truth Setting**
- **Input ground truth genome file:** Input ground truth reference genome in FASTA format
- **Input chain file:** Input chain file of ground truth genome (GroundTruth to ReferenceGenome) in CHAIN format
- **Graph Pangenome Information**
- **Input reference genome file:** Input reference genome file that has been used for graph pangenome construction in FASTA format
- **Input VCF file:** Input vcf file that has been used for graph pangenome construction in VCF format
- **Input exon annotation file:** Input exon annotation file (GTF file) of reference genome file used for constructing the primary path of graph pangenome in GTF format
- **RNA-Seq Reads Simulation**
- **Input VCF file:** Input VCF file containing variation information for pseudo genome construction in VCF format

Parameters

- **Generation of simulated RNA-seq datasets:** Input number of how many sets of simulated reads you need to generate (Default: 15)
- **Simulated RNA-seq read number:** Input simreads number to generate for each dataset (Default: 1000000)
- **Threads:** The number of threads used for parallel computation (Default: 10)

Output

- **Measurement of mapping errors in CSV format**

The screenshot shows the AnVIL Galaxy interface with the 'Mapping Error Measurement' workflow selected. The workflow is divided into several sections: 'Ground Truth Setting', 'Graph Pangenome Information', and 'RNA-Seq Reads Simulation'. Eight steps are overlaid on the interface with red arrows pointing to the corresponding input fields:

- Step 1: input a ground truth genome (points to 'Input ground truth genome file' with value '16: Sha.fasta')
- Step 2: input a chain file (points to 'Input chain file (GroundTruth to ReferenceGenome)' with value '18: ShaTotal10.chain')
- Step 3: select a graph pangenome (points to 'Graph Pangenome Information')
- Step 4: input accession for read simulation (points to 'Input accession name for pseudo genome construction' with value '997')
- Step 5: input a VCF file (points to 'Input VCF file containing variation information for pseudo genome construction' with value '7: 1001all_test.vcf')
- Step 6: input the number of read simulation sets (points to 'Generation of simulated RNA-seq datasets' with value '15')
- Step 7: input the number of reads in each simulation set (points to 'Simulated RNA-seq read number' with value '1000000')
- Step 8: click here to run this function (points to the 'Execute' button)

The right sidebar shows a 'History' panel with a list of datasets, including '18: ShaTotal10.chain', '17: araport_exon.gtf', '16: Sha.fasta', '15: Total_mapping_rate_in', '14: hisat2_alignment_population.bam', '13: hisat2_alignment_population_summary.txt', '12: hisat2_alignment_subpopulation.bam', '11: hisat2_alignment_subpopulation_summary.txt', and '10: accession_list'.

F1 Score Calculation

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 F1 score

- **Recall** = $100 \times \text{Num.correct} / \text{Num.reads}$
- **Precision** = $100 \times \text{Num.correct} / \text{Num.unique}$
- **F1 score** = $(2 \times \text{Recall} \times \text{Precision}) / (\text{Recall} + \text{Precision})$
- *Num.reads* represent the total number of reads in the simulated RNA-seq data; *Num.correct* represent the number of correctly aligned reads, matching the coordinates of the simulated read generation position; and *Num.unique* represent the number of uniquely mapped reads, with the uniquely mapped tag in the HISAT2 alignment result being "NH:i:1".

Input

- **Graph Pangenome Information**
- **Input FASTA files:** Input the reference genome file that have been used for graph pangenome construction in FASTA format
- **Input VCF files:** Input the vcf file that have been used for graph pangenome construction in VCF format
- **Input GTF files:** Input the exon annotation file of reference genome that have been used for graph pangenome construction in GTF format
- **RNA-Seq Reads Simulation**
- **Input VCF files:** Input the vcf file containing the accession name for pseudo genome construction in VCF format

Parameters

- **Generation of simulated RNA-seq datasets:** Input number of how many sets of simulated reads you need to generate (Default: 15)
- **Simulated RNA-seq read number:** Input simreads number to generate for each dataset (Default: 1000000)
- **Threads:** The number of threads used for parallel computation (Default: 10)

Output

- **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

F1 Score Calculation (Galaxy Version 1.0.0)

Graph Pangenome Information

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

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 read number to generate for each dataset (Default: 1000000).

Threads

10

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

Execute

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_andongGenomes.csv

18 : ShaTotalr10.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 (on SRR1234567.fastq_clean_reads.fastq)