

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 Preparation Module

This module prepares input files required for subsequent graph pangenome-based analysis.

Tools	Description	Input	Output	Time (test data)	Reference
Upload File	Upload input files required for all modules	Files or links	/	Depends on the file size	Galaxy
Download File	Directly fetch RNA-seq reads from NCBI's SRA database or other databases	SRR accession or HTTP/FTP link	Sequencing reads in SRA format	Depends on the network speed	SRA Toolkit
Sequencing Data Preparation	Convert RNA-seq reads from SRA to FASTQ format	RNA-seq reads in SRA format	RNA-seq reads in FASTQ format	~2 mins	SRA Toolkit

Tools	Description	Input	Output	Time (test data)	Reference
Quality Control for Sequencing Data	Check RNA-seq reads quality and obtain high-quality reads	RNA-seq reads in FASTQ format and adapter sequences in FASTA format	RNA-seq reads in FASTQ format	~2 mins	fastp

Upload File

This function is designed to upload input files required for all modules.

Input

- **Input file and data format:** 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.
- **An HTTP/FTP link:** An HTTP/FTP link specifying the path of the file to be downloaded, e.g. ftp://download.big.ac.cn/gwh/Genome/Plants/Arabidopsis_thaliana/Athaliana_167_TAIR10/TAIR10_genomic.fna.gz

Download File

This function is designed to download RNA-seq reads from NCBI SRA (Short Read Archive) database or from an user-specified HTTP/FTP link automatically. For the former, the **prefetch** function implemented in [SRA Toolkit](#) is wrapped to enable users to download sequencing data from NCBI SRA database; For the latter, **wget** command line is used to download the file according to an user-specified HTTP/FTP link.

Input

- For **Download sequencing data from Short Read Archive:**
 - **Accession:** An SRA accession ID (start with SRR, DRR or ERR, e.g. SRR1508371)
- For **Download sequencing data from an HTTP/FTP link:**
 - **An HTTP/FTP link:** An HTTP/FTP link specifying the path of the file to be downloaded, e.g.

ftp://download.big.ac.cn/gwh/Genome/Plants/Arabidopsis_thaliana/Athaliana_167_TAIR10/TAIR10_genomic.fna.gz

- **Data format:** Specify the data format, in the current version, the supported format include: txt, gff, gtf, tsv, gz, tar, vcf, fasta, html and pdf
- **Prefix:** A string specifying the prefix of the file to be downloaded

Output

- For **Download sequencing data from Short Read Archive:**
 - The compressed sequencing data in SRA format
- For **Download sequencing data from an HTTP/FTP link:**
 - The downloaded file according to the provided HTTP/FTP link

The screenshot displays the AnVIL web interface for the 'Download file from other databases (Galaxy Version 1.0.0)' workflow. The workflow is configured to 'Download sequencing data from Short Read Archive'. The 'Accession' input field contains 'SRR1234567'. A red box highlights the 'Execute' button. Red annotations indicate the steps: 'Step 1: input an accession number' pointing to the 'Accession' field, and 'Step 2: click here to run this function' pointing to the 'Execute' button. The 'What it does' section explains the function's purpose. The 'Inputs' section lists parameters: 'Accession' (An SRA accession ID), 'HTTP/FTP link' (An HTTP/FTP link), 'Data format' (Supported formats: txt, gff, gtf, tsv, gz, tar, vcf, fasta, html and pdf), and 'Prefix' (A string specifying the prefix). The 'Outputs' section shows 'The compressed sequencing data in SRA format'. The right sidebar shows a 'History' panel with a search bar and a message that the history is empty.

Sequencing Data Preparation

This function wrapped **fastq-dump** function implemented in **SRA Toolkit**.

See <http://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software> for details.

Input

- **Input Sequencing Data:** The sequencing reads in SRA format. Users can upload their local SRA file or download SRA by function **Sequencing Data Preparation**.

Parameters

- **Minimum reads length:** An integer specifying the minimum reads length to be retained
- **Single-end or paired-end reads:** Select if the SRA file is single-end or paired-end

Output

- **Sequencing dataset in FASTQ format**

The screenshot shows the AnVIL web interface for the 'Sequencing Data Preparation' tool (Galaxy Version 1.0.0). The interface is divided into three main panels: Tools, the main tool configuration area, and History.

- Tools Panel (Left):** Contains a search bar and a list of tool categories including 'Graph Pangenome Preparation Module', 'Upload File from your computer', 'Download file from other databases', 'Sequencing Data Preparation', 'Quality Control for Sequencing Data', 'Graph Pangenome Construction Module and Alignment', 'Graph Pangenome Evaluation Module', 'Graph Pangenome Application Module', 'Additional tools', and 'Built-in Converters'. The 'Workflow' (流程) section is highlighted.
- Main Tool Configuration Area (Center):**
 - Input Sequencing Data (SRA file):** A red box highlights the input field containing '2: SRR1234567.sra'. An arrow points to it with the text 'Step 1: select a SRA file from history panel'.
 - Minimum reads length:** A text input field with the value '20'.
 - Filter by sequence length:** A dropdown menu.
 - Single-end or paired-end reads?:** A red box highlights the 'single-end' selection. An arrow points to it with the text 'Step 2: choose single-end or paired-end'.
 - Execute:** A blue button with a checkmark. A red box highlights it. An arrow points to it with the text 'Step 3: click here to run this function'.
 - What it does:** A section describing the function: 'This function wrapped **fastq-dump** function implemented in **SRA Toolkit**. See <http://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software> for details.'
 - Input:** A section stating: 'Input Sequencing Data: The sequencing reads in SRA format. Users can upload their local SRA file or download SRA by function Sequencing Data Preparation.'
 - Parameters:** A section listing:
 - **Minimum reads length:** An integer specifying the minimum reads length to be retained
 - **Single-end or paired-end reads:** Select if the SRA file is single-end or paired-end
 - Output:** A section stating: 'Sequencing dataset in FASTQ format'
- History Panel (Right):** Shows a list of datasets. The first entry is '2: SRR1234567.sra' with a size of 226 MB. It is highlighted with a green background.

Quality Control for Sequencing Data

This function is designed to check RNA-seq reads quality and obtain high-quality reads.

Input

- **Input FASTQ file:** Single-end or paired-end RNA-seq reads in FASTQ format
- **Adapter sequences:** Optional, adapter sequences in FASTA format

Parameters

- **Minimum read length:** Reads shorter than this value will be discarded, default is 15 (-l)
- **The quality value that a base is qualified**

Output

- **Clean reads in FASTQ format**
- **Clean reads fastp report in HTML format**

AnVIL

Workflow

Visualize

Shared Data

Help

Login or Register

Using 215.4 MB

Tools

search tools

Upload Data

Graph Pangenome Preparation Module

Upload File from your computer

Download file from other databases

Sequencing Data Preparation

Quality Control for Sequencing Data

Graph Pangenome Construction Module and Alignment

Graph Pangenome Evaluation Module

Graph Pangenome Application Module

Additional tools

Built-in Converters

WORKFLOWS

All workflows

Quality Control for Sequencing Data (Galaxy Version 1.0.0)

Single-end or paired-end reads?

Single-end

Input FASTQ file (SE type)

3: SRR1234567.fastq

Minimum read length

15

Reads shorter than this value will be discarded, default is 15. (-l)

The quality value that a base is qualified

15

Default 15 means phred quality >=Q15 is qualified. (-q)

Adapter sequences

☒ Auto detect

☐ Adapter sequences for single-end reads

☐ Adapter sequences for paired-end reads

Threads

10

The number of threads used for parallel computation.

Execute

What it does

In this function, one existing NGS tool **fastp** (Chen *et al.*, 2018) is integrated to check sequencing reads quality and obtain high-quality reads, respectively.

Inputs

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

Adapter sequences: Optional, adapter sequences in FASTA format

History

search datasets

Unnamed history

226 MB

3

3: SRR1234567.fastq

2: SRR1234567.sra

1: SRR1234567.sra

Step 1: choose single-end or paired-end

Step 2: select a FASTQ file from history panel

Step 3: click here to run this function