MetaSVs: A pipeline combining long and short reads for analysis and visualization of structural variants in metagenomes

Yuejuan Li[1,2], Jiabao Cao[1], Jun Wang[1,2]

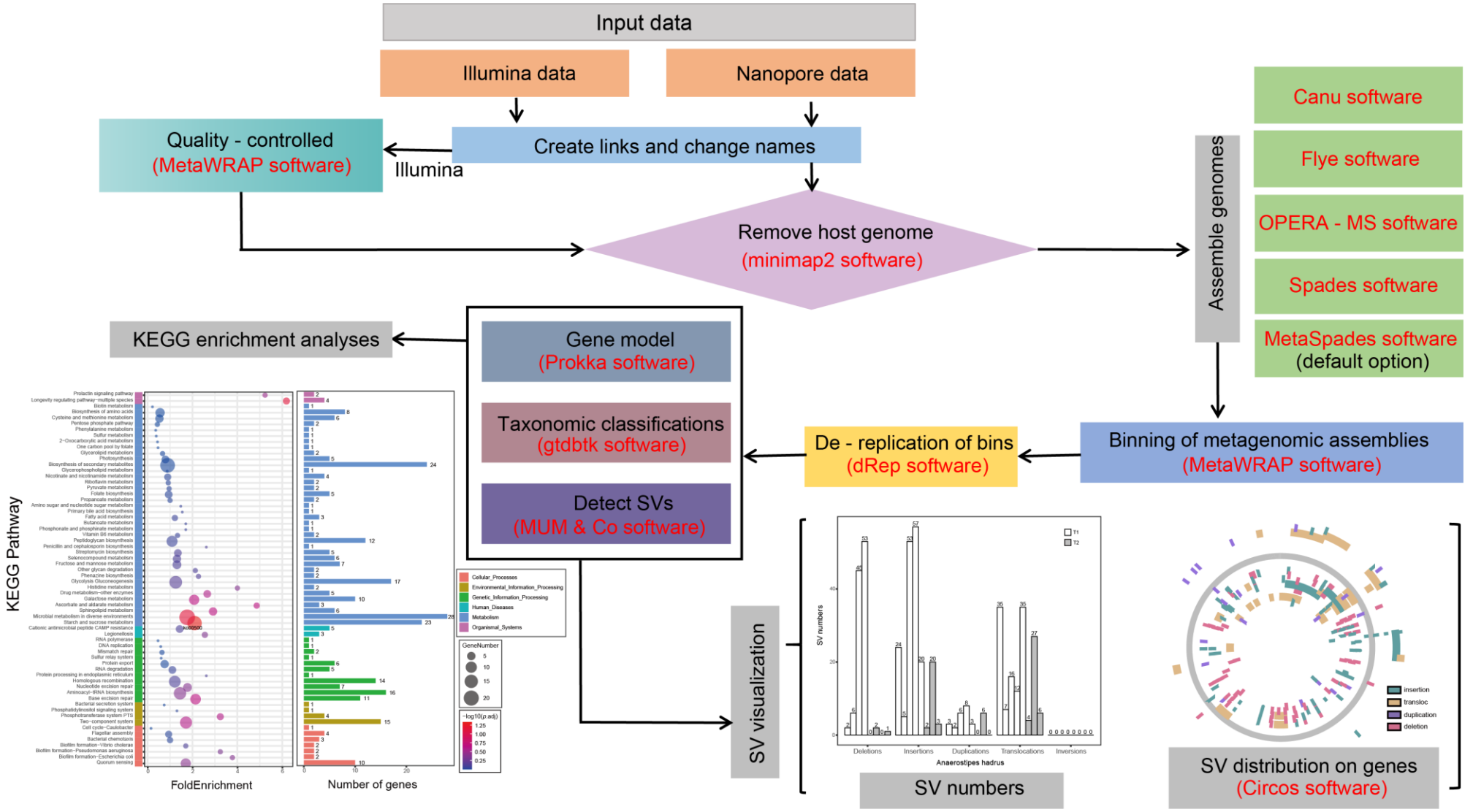
[1]CAS Key Laboratory of Pathogenic Microbiology and Immunology, Institute of Microbiology, Chinese Academy of Sciences, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China

Yuejuan Li, Jiabao Cao, and Jun Wang. 2023. MetaSVs: A Pipeline Combining Long and Short Reads for Analysis and Visualization of Structural Variants in Metagenomes. ***iMeta*** e139. https://doi.org/10.1002/imt2.139

Introduction

- ✓ Structural variants (SVs) are highly variable segments of microbial genomes, that exhibit significant between-individual variability versus high within-individual stability in the metagenomic genomes, and can therefore serve as a personalized “microbiome fingerprint” to distinguish metagenomic samples belonging to the same or different individuals.
- ✓ Identification of metagenomic SVs is mostly performed by short-read-based metagenomic profiling, which limits their detection accuracy. Rapid advancements in long-read sequencing technology, however, make it possible to produce reads of several thousand base pairs, even reaching up to 2 Mbp in length for Nanopore Technology.
- ✓ In recent years, a growing number of microbiome researchers have combined Nanopore long reads and Illumina short reads, resulting in more complete microbial genomes, resolving complex regions such as tandem repeats and large SVs, and providing unprecedented potential for SV profiling.
- ✓ The profiling of the metagenomic SVs regularly requires a series of complex analyses that are both time-consuming and parameter-heavy, plus millions of sequencing data being used as the source input, demanding a workflow that is high efficiency, traceable, and flexible.

Results


Input data
Illumina data
Nanopore data
Create links and change names
Illumina
Quality - controlled
(MetaWRAP software)
Remove host genome
(minimap2 software)
Assemble genomes
Canu software
Flye software
OPERA - MS software
Spades software
MetaSpades software
(default option)
Binning of metagenomic assemblies
(MetaWRAP software)
De - replication of bins
(dRep software)
Gene model
(Prokka software)
Taxonomic classifications
(gtdbtk software)
Detect SVs
(MUM & Co software)
KEGG enrichment analyses
KEGG Pathway
FoldEnrichment
Number of genes
SV visualization
SV numbers
Deletions
Insertions
Duplications
Translocations
Inversions
Anaerostipes hadrus
SV distribution on genes
(Circos software)
insertion
transloc
duplication
deletion

Results

(A)

```
$ cat /SV_procedure/test/config.ini

[database]
checkmdb = /opt/project/database/CheckM_db
gtdbtk = /opt/project/database/gtdbtk_db
sep_pep = /opt/project/database/kobas_db/seq_pep
sqlite3 = /opt/project/database/kobas_db/sqlite3

[fastq]
# file=XXXsamples.fastq.gz / XXXsamples_1.fastq.gz / XXXsamples_R2.fq.gz
# the table of sample information, including 4 rows
# (sample_IDs NGS_rawnames ONT_rawnames Groups)
datapath=/opt/project/data/
infotable=/opt/project/sample_info.txt

[par]
outdir=/opt/project/
multiprocessing=5

[faqc]
# metawrap (read_qc): version 1.3.2

[filter_host]
# minimap2: version 2.1
genome=none
bowtie2_threads=20
[assembly]
# SPAdes: version 3.13.0; Flye: version 2.9; OPERA-MS; Canu
# the options of method: Spades, Flye, OPERA-MS, MetaSpades
method=MetaSpades
# k-mer must be odd and less than 128
# default: MetaSpades 'auto'; OPERA-MS 60;
k-mer=default
assembly_threads=20

[binning]
# metawrap: version 1.3.2
# step1: metawrap binning (bin contigs with metabat2, maxbin2 and concoct)
# minimum contig length to bin (default=1000bp). Note: metabat2 will default to 1500bp minimum
# step2: metawrap bin_refinement
# step3: metawrap reassemble_bins (Reassemble bins using metagenomic reads)
mini_completion=70
max_contamination=10
threads_bin=20

[dereplicated_bin]
# dRep: version 3.4.3
primary_ANI=0.9
secondary_ANI=0.95
Min_overlap=0.30
coverage_method=larger
# choose total | larger
bins_samp=bins_samp

[bin_taxonomy]
# gtdbtk: version 1.7.0
[gene_model]
# prokka: version 1.13
[sv]
# mumandco: version 2.4.2
# the SV events within 10 bp of the start/end point of contigs in MAGs are not considered.
sv_loc=10
[circos]
# the name of control group which draw on the inside of the circos diagram
group_name=T1
#This parameter threshold is called only if the number of samples is greater than 10
threshold=30%
```

(A) The only input argument (the "config.ini" file) of the main program. The red fonts indicated the parameters that need to be modified for different projects, while annotation information, marked by '#' at the beginning of the line, is indicated in green fonts.

(B)

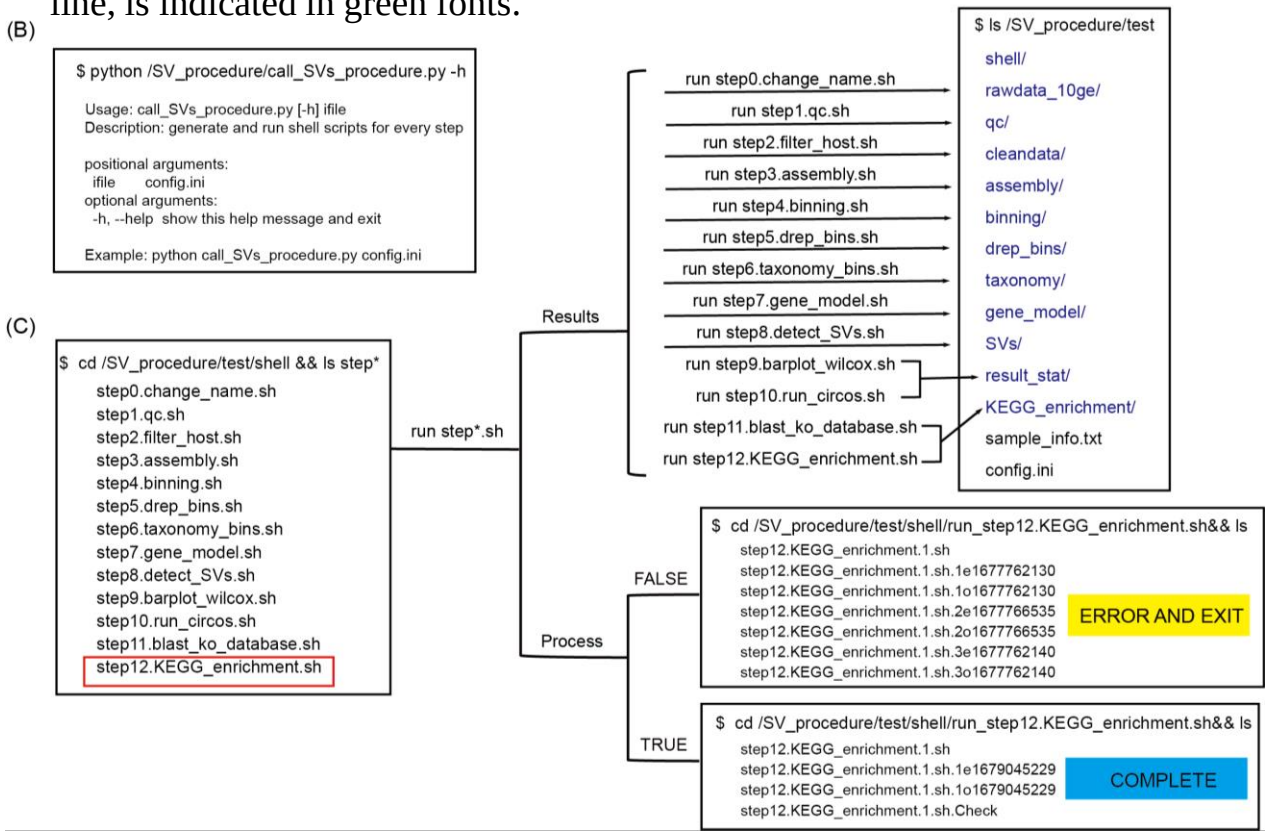
```
$ python /SV_procedure/call_SVs_procedure.py -h

Usage: call_SVs_procedure.py [-h] ifile
Description: generate and run shell scripts for every step

positional arguments:
  ifile       config.ini
optional arguments:
  -h, --help  show this help message and exit

Example: python call_SVs_procedure.py config.ini
```

(C)

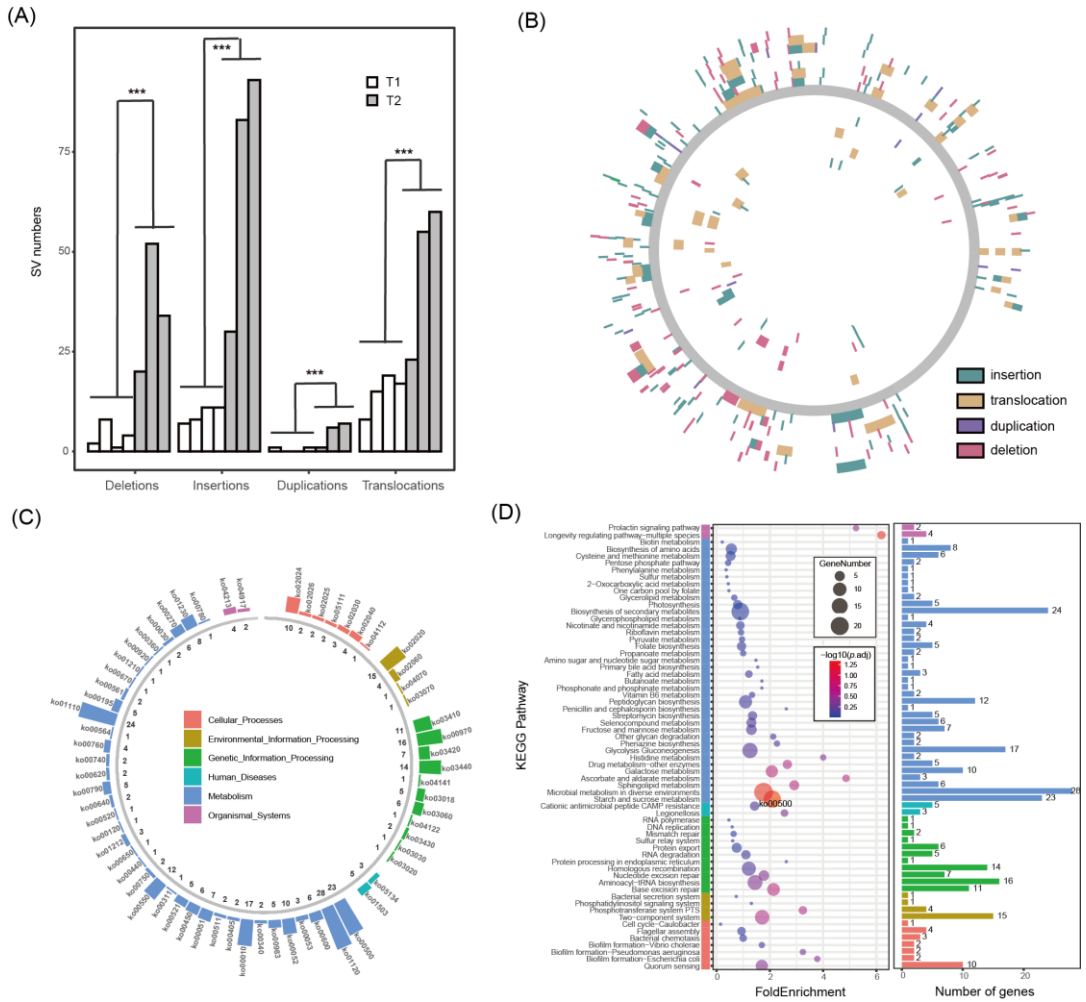


(B) The main program (call_SVs_procedure.py). (C) The main program generates shell scripts, executes the running process, and produces the resulting scripts. An illustrative example is provided for step 12 (KEGG enrichment analysis), highlighted with red boxes. KEGG stands for Kyoto Encyclopedia of Genes and Genomes, while SV represents structural variant.

- ✓ High flexibility and expansibility
- ✓ Efficient execution
- ✓ Traceability and transparency

Results

A simplified example of the human gut microbiome.



(A) Number of SVs (including insertions, deletions, duplications, and translocations) in *Anaerobutyricum hallii* of the T1 and T2 groups. ***Wilcoxon test, $p < 0.05$.

(B) Distribution of SVs on genes in reference genome of *A. hallii*. The gray circle denotes the reference genome, with the T1 group inside the circle and the T2 group outside.

The result of functional enrichment of SV-affected genes, including the number of genes (C) and the corresponding ko ID (D) mapped to each functional pathway based on KEGG; metabolism-related pathways account for six of them; *p* values were from Fisher's test.

Summary

- ✓ A bioinformatic pipeline MetaSVs to integrate Nanopore long reads and Illumina short reads is provided, capable of conducting metagenomic structural variant (SV) analyses.
- ✓ The detailed description of each step of the microbial SV pipeline is illustrated by a simplified example of the human gut microbiome.
- ✓ This pipeline will help beginners learn how to conduct microbial SV analyses and enable experienced scientists to improve their efficiency.
- ✓ GitHub link：https://github.com/Wlab518/SV_procedure

Yuejuan Li, Jiabao Cao, and Jun Wang. 2023. MetaSVs: A Pipeline Combining Long and Short Reads for Analysis and Visualization of Structural Variants in Metagenomes. ***iMeta*** e139. https://doi.org/10.1002/imt2.139

iMeta: Integrated meta-omics to change the understanding of the biology and environment

WILEY

"***iMeta***" is an open-access Wiley partner journal launched by scientists of the Chinese Academy of Sciences. iMeta aims to promote metagenomics, microbiome, and bioinformatics research by publishing original research, methods, or protocols, and reviews. The goal is to publish high-quality papers (Top 10%, IF > 15) targeting a broad audience. Unique features include video submission, reproducible analysis, figure polishing, APC waiver, and promotion by social media with 500,000 followers. Three issues were released in March, June , and September 2022.

Society: http://www.imeta.science
Publisher: https://wileyonlinelibrary.com/journal/imeta
Submission: https://mc.manuscriptcentral.com/imeta

office@imeta.science
iMeta

iMetaScience
iMetaScience