MetaSVs:一个结合长、短reads用于宏基因组结构变异分析和可视化的管道

李月娟[1,2]，曹佳宝[1]，王军[1,2]

[1]中国科学院微生物研究所病原微生物与免疫学重点实验室
[2]中国科学院大学

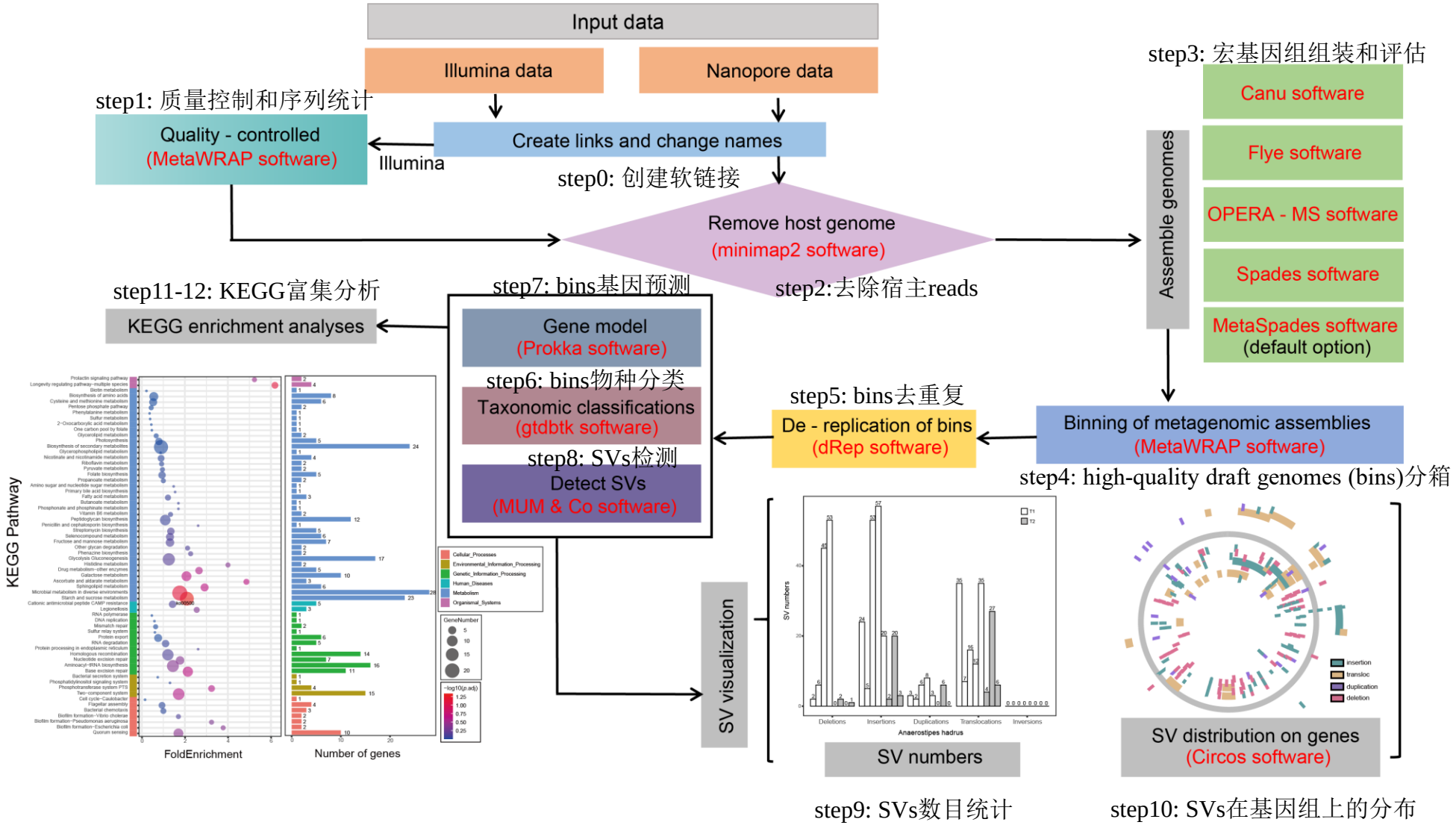
Yuejuan Li, Jiabao Cao, and Jun Wang. 2023. MetaSVs: A Pipeline Combining Long and Short Reads for Analysis and Visualization of Structural Variants in Metagenomes. ***iMeta*** e139. https://doi.org/10.1002/imt2.139

正文

- ✓ 结构变异(structural variants, SVs，包括大片段的插入、缺失、倒置和易位)是微生物基因组中高度可变的片段，表现出显著的个体间变异性和高度的个体内稳定性，因此可以作为个性化的“微生物组指纹”来区分属于相同或不同个体的宏基因组样本。
- ✓ 目前对SVs的鉴定主要是通过基于短reads的宏基因组分析来完成的，这限制了它们检测的准确性。然而，长reads测序技术的快速发展使得产生数千个碱基对的reads成为可能，Nanopore平台甚至可以获得2 Mbp长度的reads。
- ✓ 近年来，越来越多的微生物组研究人员将Nanopore长reads和Illumina短reads结合起来，得到了更完整的微生物基因组，解决了串联重复序列和SVs等复杂区域，为SVs分析提供了前所未有的潜力。
- ✓ 宏基因组SVs的分析通常需要一系列复杂的分析，既耗时又参数繁重，再加上数以百万计的测序数据被用作源输入，急需一个高效，可追溯而且灵活的工作流程。

正文

结合Illumina短reads和Nanopore长reads，建立了一个集成的宏基因组SVs分析管道MetaSVs。

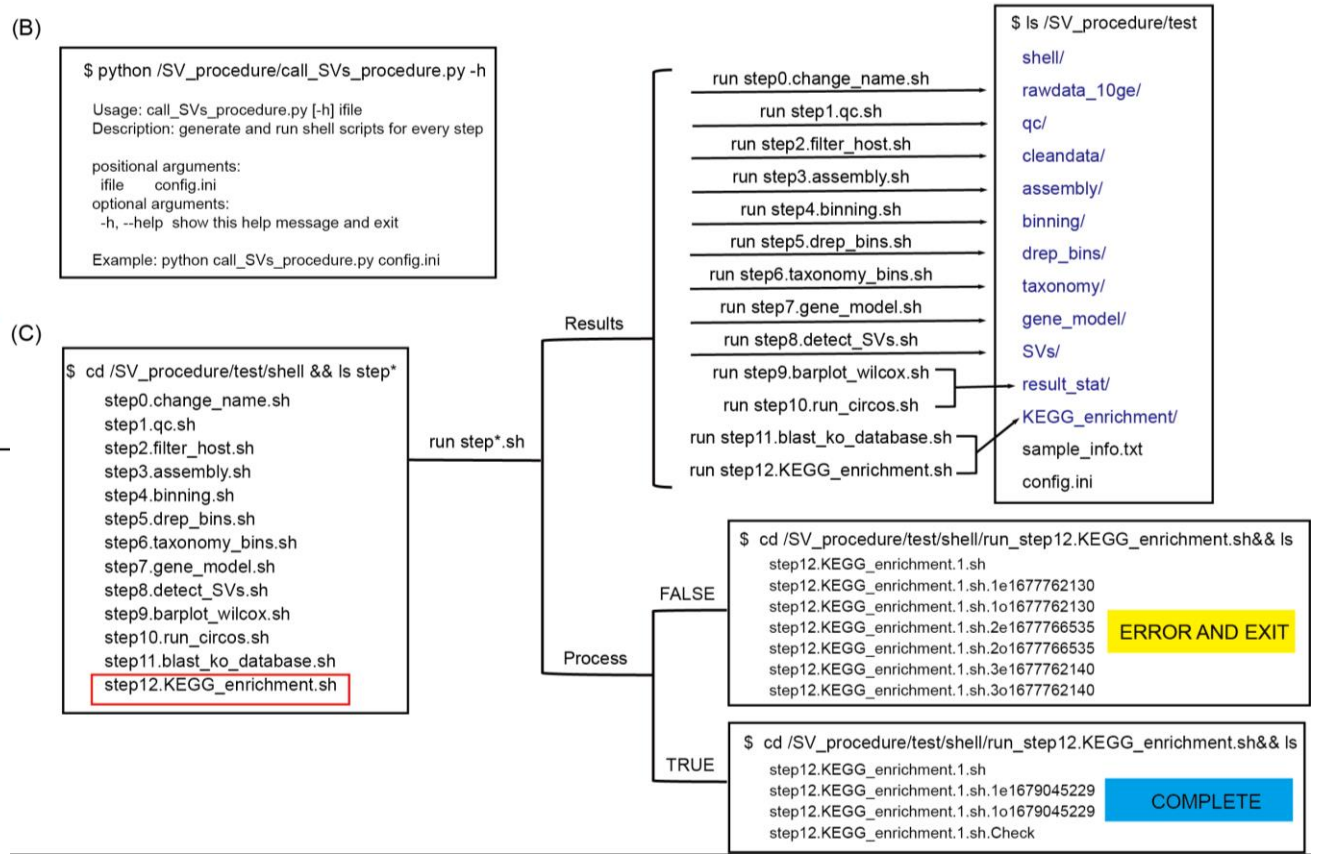


正文

(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)主程序的唯一输入参数(config.ini文件)。红色字体表示需要针对不同项目修改的参数，绿色字体表示行首用#标记的注释信息；



- ✓ 高灵活性和可扩展性
- ✓ 运行高效性
- ✓ 可追溯性和透明性

(B)主程序(call_SVs_procedure.py)；(C)主程序生成的shell子脚本，运行过程以及子脚本运行结果。以step12 (KEGG富集分析)为例，用红色框标记。

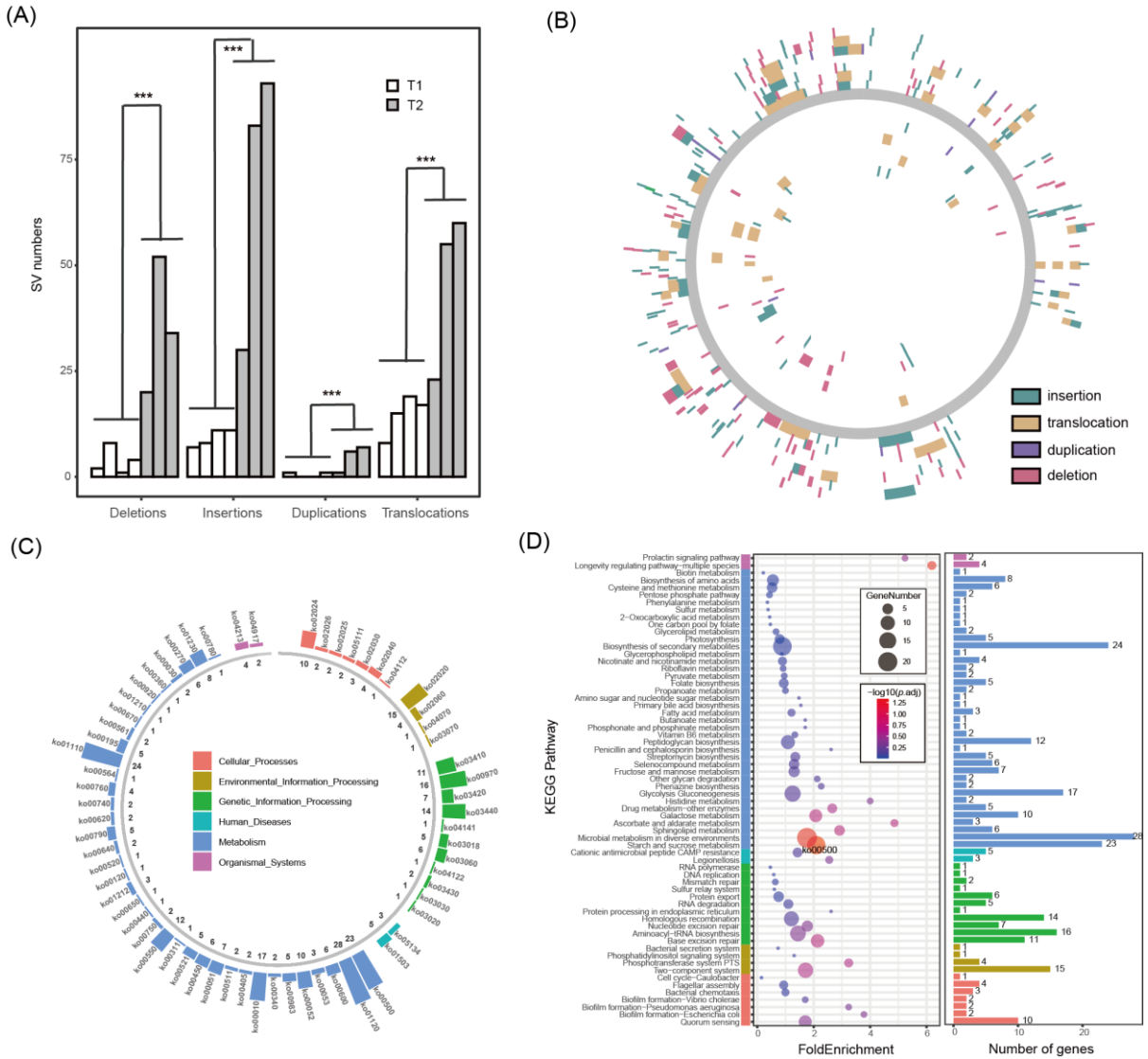
正文

一个人类肠道微生物组的简化示例

(A) *Anaerobutyricum hallii* 的T1和T2组中SVs数量(包括插入、缺失、重复和易位)。Wilcoxon检验，*** $p < 0.05$

(B)SVs在*A. hallii*的参考基因组上的分布。灰色圆圈为参考基因组, T1组在圈内，T2组在圈外；

发生SVs的基因进行KEGG功能富集分析的结果，包括映射到每个功能通路的基因数量(C)和对应的ko ID (D)。



总结

- ✓ 提供了一个结合Nanopore长reads和Illumina短reads的生物信息学管道MetaSVs，能够进行宏基因组结构变异(structural variant, SV)分析；
- ✓ MetaSVs管道的每个步骤的详细描述通过一个人类肠道微生物组的简化示例来说明；
- ✓ 这个管道将帮助那些对宏基因组SVs感兴趣但缺乏复杂生物信息学知识的研究人员实现相关分析
- ✓ GitHub链接：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: 整合宏组学重新认识生命和环境科学

iMeta WILEY

"***iMeta***" 是由威立、肠菌分会和本领域数百位华人科学家合作出版的开放获取期刊，主编由中科院微生物所刘双江研究员和荷兰格罗宁根大学傅静远教授共同担任。目的是发表原创研究、方法和综述以促进宏基因组学、微生物组和生物信息学发展。目标是发表前10%(IF > 15)的高影响力论文。期刊特色包括视频投稿、可重复分析、图片打磨、青年编委、前3年免出版费、50万用户的社交媒体宣传等。2022年的三月、六月和九月期已正式在线出版发行!

主页: http://www.imeta.science
出版社: https://wileyonlinelibrary.com/journal/imeta

投稿: https://mc.manuscriptcentral.com/imeta

office@imeta.science

iMeta 宣传片