肠道噬菌体组生物信息学分析方法的研究进展

胡晴玥，李德志，刘 箐*
(上海理工大学 健康科学与工程学院，上海 200093)

摘 要 噬菌体是感染细菌的病毒，广泛存在于各类环境中。由于传统实验研究的局限性及噬菌体基因的特异性，导致对肠道噬菌体的研究很少。随着宏基因组测序技术的发展和各种生物信息分析软件的开发，可以通过噬菌体组学，加深对肠道噬菌体的认识。噬菌体组分析流程主要包括原始数据质量控制和预处理，病毒基因组序列的拼接组装，类病毒颗粒的筛选和系统分类注释以及进化分析和预测相应宿主细菌。本文对噬菌体组分析流程和其中所需要的常用生物信息分析工具和数据库进行详细的介绍，可以为肠道噬菌体研究以及相关的研究人员提供参考。

关键词 肠道噬菌体；生物信息学；噬菌体组；病毒组；机器学习软件；数据库
中图分类号 Q939.9 **文献标识码** A **文章编号** 1005－7021(2022)03－0089－11
doi: 10.3969/j.issn.1005－7021.2022.03.012

Advances in Bioinformatics Analysis of Intestinal Bacteriophageome

HU Qing-yue, LI De-zhi, LIU Qing*
(*Schl. of Health Sci. & Engin., Uni. of Shanghai for Sci. &Tech., Shanghai* 200093)

Abstract Bacteriophages are viruses that infect bacteria and widely exist in various environments. Due to the limitation of traditional experimental studies and the specificity of phage genes, have led people to study and know little about intestinal phages. With the development of macrogenome sequencing technology and various bioinformatics analysis software tools, people can have an exacerbation of understanding about intestinal phages through phageomics. The phageome analysis workflow mainly includes quality control and preprocessing of raw data, splice and assembly of the viral genome sequence, screening of virus-like particles and systematic classification annotation as well as evolutionary analysis and prediction of relevant host bacteria. In this roundup, the analysis workflow of phageome and common use tools for bioinformatics and database required in the phageome analysis process are introduced in detail, hopefully it could provide references for related researchers and promote the development of the intestinal phage research field.

Keywords intestinal phages; bioinformatics; phageome; virome; machine learning software; database

病毒是地球上数量最多的生物实体，分布在生物圈的每一个角落，其中噬菌体是感染细菌的病毒，据估计在地球上的数量约为 10^{31} 个，可能超过所有物种数量之和[1]。噬菌体广泛存在于身体的各个部位，如皮肤、口腔、肺部、肠道和尿道[2]，其中肠道中噬菌体数量尤为庞大，研究估计健康成人的每克粪便约有 $10^9 \sim 10^{10}$ 数量级的噬菌体[3-4]。一方面，肠道噬菌体通过与宿主细菌的相互作用，调节肠道菌群的结构，如消灭宿主细菌，传递新的细菌表型，调节细菌群落组成以及基因表达和进化，进而影响动物宿主的生理健康[5]；另一方面，噬菌体与哺乳动物细胞的病毒有一些共同特征，可被先天宿主受体(如 Toll 样受体家族)识别，调节免疫系统[6]。肠道噬菌体通过胞吞转

基金项目：国家自然科学基金项目(31871897)
作者简介：胡晴玥 女，硕士研究生。研究方向为生物信息学及肠道噬菌体。E-mail: 15043257496@163.com
* 通讯作者。男，教授，博士，博士生导师。研究方向为病原微生物及肠道菌群和噬菌体。E-mail: liuq@usst.edu.cn
收稿日期：2021-10-21

运的方式跨过肠道上皮细胞进入动物宿主体内血液、淋巴、器官甚至大脑，直接与体内的细胞及器官相互作用，进而直接影响哺乳动物免疫系统和生理健康[7]。早期研究人员从肠道微生物群出发，探索并奠定了肠道微生物对人体健康及功能的重要作用，提出了健康的肠道微生物群(Healthy gut microbiome)这一概念[8]。而肠道噬菌体对哺乳动物的生理健康同样有着不可忽视的重要影响，因此有学者受到启发，从噬菌体的角度出发，提出健康的肠道噬菌体群(Healthy gut phageome)，发现不同动物个体间噬菌体差异较大，但在健康个体间存在着一些共有噬菌体，组成健康肠道噬菌体组(Phageome)。虽然肠道内噬菌体对人体健康以及疾病治疗有着重要的作用，但是对噬菌体的了解相对于其宿主细菌而言只是些零星的数据[9-10]，主要原因: ①传统实验室培养方法的局限; ②噬菌体基因组的特殊特性; ③相较于细菌而言缺乏通用的基因标记; ④病毒组数据库非常不完善，超过 80% 的病毒缺少数据库信息，早期研究发现大多数已鉴定的肠道噬菌体与数据库中的序列没有同源性，不同研究之间变异性很高[10]; ⑤缺乏病毒分析标准。随着高通量测序和生物信息学以及机器学习技术的发展，许多以研究病毒组为目的的生物信息分析软件的开发与应用，使得研究人员可以深入探寻肠道噬菌体对人体的影响。Gregory 等[11]整合来自 32 个研究项目的 2 697 个肠道宏基因组，含有 33 242 个病毒群(长度大于 5 kb)，覆盖来自 16 个国家的 1 986 名个体，构建了人类肠道病毒基因组数据库(Human Gut Virome Database, GVD)。为了进一步推进对肠道噬菌体多样性、进化分支以及全球分布情况的分析，Camarillo-Guerrero 等[12]将来自 6 个大洲 28 个国家的 28 060 个人类肠道宏基因组，共 142 809 个非冗余噬菌体基因组和 2 898 个培养的肠道细菌基因组汇编入肠道噬菌体数据库(GPD)，揭示了噬菌体的多样性以及宿主肠道细菌范围，类似病毒数据库的建立和逐步完善帮助原始数据在下游分析过程中进行分类和功能注释，提高了分析的准确性。通过生物信息学技术，可以揭示噬菌体的多样性、进化分支、相应宿主细菌以及功能，了解肠道噬菌体组在疾病中的角色，进而为疾病的诊断与治疗提供新的策略。肠道宏噬菌体组学的分析流程包括以下步骤: ①原始数据质量控制和预处理，主要是过滤接头序列、低质量序列(quality scores < 30)，以及来自宿主动物或细菌等其他非病毒基因组序列; ②肠道宏噬菌体组的拼接组装; ③评估组装质量; ④类病毒颗粒的筛选以及系统分类和功能注释; ⑤进化分析和预测宿主细菌。对于肠道噬菌体组的相关研究学者，尤其是没有生物信息学背景的研究人员来说，选择合适的软件和分析方法成为一项挑战。本文将概述现阶段肠道噬菌体组数据分析主流的方法和思路，并对涉及到的软件工具和数据库进行详细介绍(图 1)。

图 1 肠道噬菌体组生物信息分析流程

Fig. 1 Bioinformatics analysis workflow of gut phageome

1 原始数据质量控制和预处理

样品测序后得到的原始数据(raw data)中包含许多非噬菌体基因组序列，例如构建文库时所产生的接头序列、引物序列和在提取噬菌体过程中残留的宿主基因等其他非病毒基因，以及由于测序错误而产生的低质量序列，非目标基因序列的存在会影响后续的下游分析，为避免分析结果受到影响，需要将这些序列去除。最常见的质量分析评估软件为 FastQC，在 Windows、Linux 和 Mac 操作系统中均可使用，FastQC 是基于 Java 所编译的，输出的结果以网页 HTML 的形式呈现。通过质量评估得出的结果对数据进行相应的预处

理。Cutadapt[13] 是常用的数据预处理软件,可以去除接头序列和超出指定范围长度的序列以及低质量序列(序列中含有较多的“N”) ,支持 FASTA 和 FASTQ 格式的文件输入和输出。在研究过程中常使用 Minion[14] 预测原始数据中的未知接头序列,其默认输入格式为 FASTQ,需要将预测结果通过 NCBI 上的 BLASTN 与核酸数据库进行比对,确定其为生物学序列后再使用 Cutadapt 将其除去。AfterQC[15] 是用 Python 开发的质控软件,可以发现并纠正绝大部分的错误序列,并具备高度自动化的数据过滤功能,还提供了校正重叠区域错误碱基以及预估序列错误率的功能,AfterQC 可以自动检测和设置所有选项,大多数情况下的运行过程不需要添加参数和选项,方便研究人员使用。Trimmomatic[16] 是适用于 Illumina 测序数据的质量控制过滤工具,常作为功能性软件嵌合在 16S 扩增子测序分析流程中[17],也可用于宏基因组数据前期的质量过滤和序列修剪。Fastp[18] 是同时具备 FastQC、Cutadapt、Trimmomatic、AfterQC 四款软件中主要功能的数据前处理工具(表 1) ,区别于前面提到的使用 Java 和 Python 编写的软件,Fastp 使用 C/C + + 语言汇编而成,可以实现多线程运行,运行速度比同类软件快 2 ~5 倍。

由于样本中含有一定比例的动物宿主基因,所以需要去除宿主基因或其他污染序列,再进行后续分析。通常使用比对参考基因组的方式识别其他来源的基因组,使用 Bowtie2[19]、BWA[20]、BLAST[21] 将数据与宿主参考基因组比对,识别出的污染序列可使用 samtools[22-23] 工具去除。FastQ Screen[24] 同时包含质量控制和去除非病毒来源污染基因组的功能,可以同时比对多种不同来源的参考基因组,通过嵌套 Bowtie1、Bowtie2、BWA 三款序列比对软件包来识别其他污染序列,结果以文本和图形两种形式展现。MultiQC[25] 可将所有样本的输出信息汇集到一个表格及图形文件中,便于研究人员比较样本间数据质量差异,在肠道噬菌体的分析中常用于对数据分析流程处理结果的评估。Aozan[26] 通过嵌套 FastQC、MultiQC、FastQ Screen 三款工具,可自动处理原始数据(表 2) 。

表 1　质量控制软件的数据输入输出格式和特点

Table 1　Data input and output format and features of quality control software

软件名称	输入格式	输出格式	特征
FastQC	Bam, Sam or FASTA	HTML	支持多种输入格式
Cutadapt	FASTQ, FASTA	FASTQ, FASTA	主要用于去除接头序列,可与其他数据前处理软件组合使用
Trimmomatic	FASTQ	FASTQ	数据过滤后的结果以 FASTQ 格式输出,缺乏可视化呈现效果
AfterQC	FASTQ	FASTQ, HTML	根据质量高低将序列分类,并分别存储
Fastp	FASTQ	HTML, JSON	功能齐全,数据处理执行效率高

表 2　原始测序数据质量控制和数据前处理软件工具

Table 2　Quality control of raw sequencing data and preprocessing software tools

软件名称	软件描述	软件网址	参考文献
FastQC	原始数据的质量控制和评估	https: //www. bioinformatics. babraham. ac. uk/projects/fastqc	
Cutadapt	去除接头序列、低质量序列以及指定长度序列	https: //cutadapt. readthedocs. io/en/stable/	[13]
Minion	寻找和预测数据中的接头序列	https: //www. ebi. ac. uk/research/enright/software/kraken	[14]
AfterQC	过滤除去低质量或错误序列	https: //github. com/OpenGene/AfterQC	[15]
Trimmomatic	质量控制和修剪序列	http: //www. usadellab. org/cms/? page = trimmomatic	[16]
Fastp	去除接头序列、过滤低质量序列	https: //github. com/OpenGene/fastp	[18]
MultiQC	比较不同样本中的序列质量控制报告、过滤低质量序列	https: //multiqc. info/	[25]

续表 2

软件名称	软件描述	软件网址	参考文献
FastQ Screen	去除低质量序列、过滤宿主基因的污染	https://www.bioinformatics.babraham.ac.uk/projects/fastq_screen/	[24]
Aozan	结合了 FastQC、MultiQC 和 FastQ Screen	https://github.com/GenomicParisCentre/aozan	[26]
Bowtie2	序列比对	https://bowtie-bio.sourceforge.net/bowtie2/index.shtml	[19]
BWA	序列比对	https://github.com/lh3/bwa	[20]
samtools	转换数据格式、建立索引、过滤指定序列	https://github.com/samtools/samtools	[22-23]
bedtools	转换数据格式	https://github.com/arq5x/bedtools2	[27]

2 噬菌体基因组序列的拼接组装

原始序列经过质量控制及预处理后得到的短序列需要拼接组装成较长的序列(contigs)才能进行后续的物种注释等下游分析。基因的组装算法主要分为 *Comparative* 组装和 *De novo* 组装两类，*Comparative* 组装是利用参考数据库中的同源序列来指导新基因组的构建；*De novo* 组装是将大量(短或长) DNA 片段重新组装的方法，无需预先知道这些 DNA 片段的正确序列或顺序，在没有参考序列的情况下对未知基因组进行序列拼接[28]。由于目前数据库对自然界病毒种类的覆盖率不到 20%，许多噬菌体在参考数据库中很难找到同源物，所以针对肠道噬菌体宏基因组的研究主要是通过 *De novo* 组装来进行后续的下游分析。*De novo* 组装软件根据不同的策略主要分为三类：①基于 *Greedy* 策略的组装，这种方法在早期的基因组研究中比较常见，代表软件有 TIGR[29]、VCAKE[30]，此类组装策略的缺点是在合并 reads 或 contigs 的过程中过度注重局部序列最优化的选择而忽视序列之间的整体关系，从而导致在重复序列的组装中会出现错误；②基于 *Overlap-Layout-Consensus* 策略的组装，适用于较长 reads 的测序数据(Sanger、PacBio、Nanopore)，常用的代表软件有 Celera Assembler[31]、Canu[32]、Falcon[33]、Arachne[34]、MIRA[35]。Celera Assembler 的出现极大推动了基因组学的研究，多细胞生物的第一个全基因组霰弹枪测序序列[36]和人类个体第一个二倍体序列[37]就是由 Celera Assembler 组装完成的。Canu 是 Celera Assembler 的一个分支，适用于高噪声的单分子测序(PacBio RSII 和 Oxford Nanopore)。Falcon 适用于复杂基因组的组装。MIRA 和 Arachne 支持拼接组装全基因组霰弹枪序列；③基于 *De Bruijn Graph* 策略的组装，适用于序列长度较短的 Illumina 等二代测序数据，是分析肠道噬菌体组中最常用的一种组装策略，常用的软件有 IDBA-UD[38]、SPAdes[39]、Megahit[40]、SOAPdenovo2[41]、SKESA[42]。IDBA-UD 适用于组装测序深度不均匀的短读长序列，使用多个深度相对阈值来过滤低深度和高深度区域的错误 k-mers，以及采用双端局部组装技术解决低深度短重复区域的分支问题，准确性要高于其他同类的短序列组装软件，但资源消耗较高。Megahit 采用了简洁版的 *De Bruijn graph* 算法，在运行时间和内存需求方面相对于其他同类组装软件具有优势，适用于组装大型复杂的宏基因组数据，在大量样本混合组装方面优势明显，运行速度很快，对硬件设备资源的消耗少(需要运行内存约 30 G)。SOAPdenovo 常用来组装许多大型真核生物基因组，也可用于细菌和病毒基因组的组装，SOAPdenovo2 在 SOAPdenovo 的基础上改进了纠错算法，减少 *De Bruijn Graph* 构造过程中的内存消耗，在 contigs 组装中能够解析较长的重复区域，并且组装得到的 contigs 数量较多，增加了组装长度和 scaffolding 的覆盖范围，适用于较大基因组的短序列组装，组装速度快，但是错误率较高。SPAdes 是应用最广泛、各项指标参数最突出的组装工具，提供了很多样品类型的分析技术，当前版本适用于 Illumina 和 IonTorrent 测序数据，并且支持 PacBio、Nanopore 和 Sanger 测序数据的混合组装，SPAdes 由八种

不同的组装流程包构成,用于宏基因组和宏转录组的组装,可以从宏基因组数据集中组装质粒和生物合成基因簇,其中内嵌的 metaSPAdes[43] 包是目前宏基因组领域组装指标较好的软件,组装得到的 contigs 数较多,不足之处是拼接时间较长,运行过程设备的资源消耗较高,需要较大的运行内存(约 250 G) ,适用于对肠道噬菌体组进行拼接。SKESA 是近几年新开发的组装软件,适用于 Illumina 测序数据的组装,可以过滤污染序列,得到的 contigs 具有较高的质量和连续性,其运行速度相较于 SPAdes、MEGAHIT 有明显的提升,组装得到的 N 50 平均长度高于上述常用的两种拼接软件,并且组装错误率较低,现阶段主要用于拼接 SRA 数据库中的微生物基因组,并且嵌套于 Pathogen Detection Project(PDP) 分析流程中。序列组装完成后通常需要评估组装质量,常用的工具有 QUAST[44] 、MetaQUAST、CheckV。QUAST 支持 FASTA 格式的 contigs 和参考序列以及 FASTQ、SAM 和 BAM 格式[45] 的数据,QUAST 融合了现有软件(Plantagora、GAGE、GeneMark. hmm[46] 、GlimmerHMM[47]) 的质量度量方法,并对其进行了扩展,既可以通过比对参考基因组来评估已知物种的组装质量,也可以计算评估缺少参考基因组的未知物种,评估结果以图表的形式输出。MetaQUAST[48] 是在 QUAST 基础上衍生出的更加先进的宏基因组组装评估软件,可以同时比对多个参考基因,并制作多个基因组的组装质量评估图表,在分析评估常见的未知物种时,会自动检测并从 NCBI 数据库中下载相近的同源参考序列以提高评估准确性,可以检测嵌合序列并报告“种间装配错误”。CheckV[49] 可以自动评估宏病毒组组装的完整性和病毒组 single-contig 中的宿主污染情况,通常将 CheckV 与 MetaQUAST 结合使用以准确评估噬菌体组的组装质量(表 3) 。

表 3　病毒基因组的拼接组装和评估序列组装质量工具

Table 3　Viral genome assembly and quality assessment of assembly software tools

软件名称	软件描述	软件网址	参考文献
VCAKE	利用简单 K-mer 算法组装短读长基因组	https: //bioinformaticshome. com/tools/wga/descriptions/VCAKE. html	[30]
Celera Assembler	*De novo* 全基因组 DNA 组装软件	http: //wgs-assembler. sourceforge. net/wiki/index. php?title = Main_Page	[31]
Canu	适用于高噪声单分子测序数据的组装	https: //canu. readthedocs. io/en/latest/	[32]
Falcon	适用于复杂基因组组装	https: //github. com/PacificBiosciences/pb-assembly	[33]
MIRA	全面的组装序列软件	https: //sourceforge. net/projects/mira-assembler/	[34]
IDBA-UD	*De Bruijn Graph* 二代测序数据组装软件	https: //github. com/loneknightpy/idba	[38]
SPAdes	内含多种组装流程包的序列组装软件	https: //cab. spbu. ru/files/release3. 15. 2/manual. html	[39]
MEGAHIT	运行速度快,内存高效的二代测序数据组装软件	https: //github. com/voutcn/megahit	[40]
SOAPdenovo2	短读长序列组装软件	https: //github. com/aquaskyline/SOAPdenovo2	[41]
SKESA	微生物基因组的 *De novo* 组装软件	https: //github. com/ncbi/SKESA	[42]
QUAST	通过计算组装序列的各种指标以评估组装质量	https: //github. com/ablab/quast	[44]
MetaQUAST	评估宏基因组序列的组装质量	http: //bioinf. spbau. ru/metaquast	[48]
CheckV	评估病毒宏基因组的组装质量	https: //bitbucket. org/berkeleylab/checkv/src/master/	[49]

3　类病毒颗粒的筛选和系统分类注释

组装后的基因组需要确定其是否为噬菌体序列,因此需要注释和筛选出病毒基因组,对噬菌体组进行系统分类和注释是功能分析的关键步骤,也是研究肠道噬菌体的重要问题[50] 。系统分类和注释的方法根据原理不同可分为两类,一类是基于将序列或组装得到的 contigs 与参考数据库

进行比对,常用 BLAST 中的 tBLASTx、BLASTn、BLASTx[51] 工具对序列进行比对注释,或将组装得到的 scaffolds 与 NCBI 中的 Refseq virus 数据库进行比对(ftp: //ftp. ncbi. nlm. nih. gov/refseq/release/viral) 。国际病毒分类委员会(ICTV) 存储了病毒分类、分类单元名称和相关的宏病毒组数据,包括每个已命名物种的示例病毒信息,ICTV 官方网站(https: //talk. ictvonline. org/) 内含病毒分类数据库,对每种经过定义的病毒链进行了完整的描述[52]。pVOGs[53] (Prokaryotic Virus Orthologous Groups) 数据库包含近 3 000 个完整的原核宿主病毒基因组(超过 97% 为噬菌体) 和 9 518 个直系同源组,该数据库可以应用于分析已知噬菌体的进化分类、了解病毒蛋白家族的历史、噬菌体基因组的重建以及帮助特征不佳的基因组注释同源基因。IMG/VR[54] (The Integrated Microbial Genome/Virus) 是目前最大的致力于研究病毒组学的公共数据管理和分析平台,最新版本的 IMG/VR 包含 18 373 个已培养和 2 314 329 个未培养的病毒基因组,可以进行基因组的注释以及预测宿主细菌分类,支持用户根据基因组特征或序列相似性高效地浏览、搜索和选择未培养的病毒基因组。PHASTER[55] 是用于对细菌和质粒中的前噬菌体进行快速鉴定和注释的网页服务器,输入文件支持 FASTA 格式的原始核酸序列或 GenBank 格式的已注释基因组数据,也可以通过数据库中的序列编号对相应参考序列进行分析,分析结果以图表的形式直观展现,PHASTER 提供用户友好型的图形交互界面便于研究人员使用,不足之处是单次上传的文件大小不能超过 40 M,需要使用脚本将 FASTA 文件分为多个小文件,并且网页服务器运行不稳定。

由于噬菌体基因组系统发育受到类群内广泛的水平基因转移和基因组模块化的损害,导致环境样本中大量的噬菌体颗粒复杂化增加,使得在参考数据库中查找噬菌体的同源序列变得非常困难[56],并且相关数据库包含的信息有限,采用同源序列比对的方法会有大量序列被标记为“未知”[57]。为解决这些难题,许多分析工具使用机器学习或深度学习的算法来进行病毒的系统分类和注释,有效解决了基于参考数据库和同源序列比对的方法不能从宏基因组数据中识别未知病毒或短病毒序列的问题。Prophage Hunter[58] 提供一站式的网络服务,从细菌中提取原噬菌体基因组并评估其活性,识别系统遗传学相关的噬菌体,并注释噬菌体蛋白功能,Prophage Hunter 在建库的过程中采用了基于序列相似性的搜索和噬菌体遗传特征的机器学习分类算法,能够识别参考数据库之外的未知噬菌体。更多的宏基因组功能注释方法可参考 Prakash 等[50] 的文章。MARVEL[59] 采用随机森林机器学习的方法来预测筛选宏基因组数据中的双链 DNA 噬菌体,筛选结果具有较高的准确性。RNN-VirSeeker[60] 是基于长短期记忆网络模型构建的病毒识别软件,模型通过数据库进行训练,软件可以自动查询基因组的高级特征,并根据 softmax 层的评分来预测病毒序列,在识别筛选较短的病毒序列(<500 bp) 和人体肠道宏噬菌体组时具有较高的准确性。VirSorter[61] 是基于隐马尔科夫模型(Hidden Markov model,HMM) 建立的,能够以较高的准确性检测多种类型微生物的较短(3 kb) 组装序列(contigs) 中的病毒信号,使用 RefSeqABVir 或 Viromes[62] 作为参考数据库。VIBRANT[63] 使用混合机器学习和蛋白质相似性方法,从宏基因组组装序列中表征病毒群落功能,突出病毒辅助代谢基因和代谢途径,利用蛋白质特征的神经网络和新开发的 v-score 度量标准,可以最大限度地识别裂解噬菌体基因组,包括高度多样化的噬菌体组。病毒和宿主有明显不同的 k-mer 特征,通过基因组 k-mer 频率分布来发掘病毒基因是常用的筛选肠道噬菌体组的分类鉴定方法,这类方法的代表软件是 VirFinder[64],它是第一个使用 k-mer 频率的机器学习方法来筛选病毒序列。ViraMiner[65] 包含了卷积神经网络(Convolutional Neural Networks,CNN) 的两个分支,用于检测原始宏基因组 contigs 中病毒模式和模式频率,可以从不同人类样本的原始宏基因组序列中检测噬菌体序列。DeepVirFinder[66] 同样采用卷积神经网络的深度学习方法来自动学习病毒基因组特征,并同时基于这些基因组特征建立预测模型来判断序列是否源自噬菌体基因组,在研究分析中,通常将 MARVEL 与 DeepVirFinder 组合使用的准确性和覆盖率较高(表 4) 。

表 4　类病毒基因组的筛选、系统分类注释软件以及常用的病毒数据库

Table 4　Identify viruses, system classification or annotation software tools and commonly used virus databases

软件名称	软件描述	软件网址	参考文献
PHASTER	快速识别和注释细菌基因组和质粒中的前噬菌体序列	http: //phaster. ca/	[55]
Prophage Hunter	自定义噬菌体部件库中的相似性匹配和噬菌体遗传特征的机器学习	https: //prohunter. genomics. cn/index. php/Home/hunter/hunter. html	[58]
MARVEL	预测宏基因组组装序列中的双链 DNA 噬菌体序列	https: //github. com/LaboratorioBioinformatica/MARVEL	[59]
RNN-VirSeeker	从宏基因组数据的 FASTA 文件中识别病毒 contigs	https: //github. com/crazyinter/RNN-VirSeeker	[60]
VirSorter2	筛选多种 DNA 和 RNA 病毒基因组	https: //github. com/jiarong/VirSorter2	[61]
VIBRANT	从宏基因组组装序列中筛选和注释病毒基因	https: //github. com/AnantharamanLab/VIBRANT	[63]
VirFinder	筛选 FASTA 文件中的病毒序列	https: //github. com/jessieren/VirFinder	[64]
ViraMiner	在多种人体生物标本中识别病毒基因	https: //github. com/NeuroCSUT/ViraMiner	[65]
DeepVirFinder	使用深度学习的方法预测病毒序列	https: //github. com/jessieren/DeepVirFinder	[66]
pVOGs	原核病毒同源簇数据库	http: //dmk-brain. ecn. uiowa. edu/pVOGs/home. html	[53]
ICTV	国际病毒分类委员会	https: //talk. ictvonline. org/	[52]
Refseq virus	提供病毒基因组序列数据和相关信息	https: //www. ncbi. nlm. nih. gov/genome/viruses/	
Viromes	探索宏基因组中的病毒信息资源	http: //virome. dbi. udel. edu/	[62]
IMG/VR	病毒基因组参考资源,衍生于宏基因组样本	https: //img. jgi. doe. gov/cgi-bin/vr/main. cgi	[54]

4　进化分析和预测相应宿主细菌

肠道噬菌体对相应宿主细菌群落的调节和生化作用会直接或间接的影响哺乳动物生理健康,了解噬菌体-宿主的感染性,对于从微生物群系出发理解噬菌体对细胞生命的影响以及它们作为肠道生态系统的重要组成部分至关重要。研究人员发现整合到宿主基因组的溶原噬菌体拥有与宿主细菌 tRNA 基因完全匹配的附着位点[67],因此 Bellas 等[68]将噬菌体基因组 scaffolds 通过 BLASTN 比对到 tRNADB-CE 数据库中,用匹配 tRNA 的方法在门或纲的水平上推测宿主细菌。Boeckaerts 等[69]通过构建受体结合蛋白(RBP) 序列数据库,使用机器学习的方法在种水平上预测噬菌体的宿主细菌。Young 等[70]结合噬菌体基因组的 k-mer 组成和蛋白结构域,开发了一种新的预测宿主细菌的计算框架,通过研究噬菌体基因组的核苷酸、氨基酸特性和蛋白质结构域等特征,并结合这些互补特征,提高宿主预测的准确性,认为感染同一宿主的噬菌体之间进化关系所产生的系统发育信号也可以被预测,这是由于在共同进化过程中尽管发生了频繁的宿主切换,但病毒和宿主系统发育树倾向于一致。Villarroel 等[71]开发了 HostPhinder 工具,该工具通过比较 k-mers,根据基因组相似性预测噬菌体的细菌宿主,HostPhinder 支持交互式网站服务。Ahlgren 等[72]利用基于病毒和宿主寡核苷酸频率模式来预测给定病毒的宿主细菌,并且提供了 VirHostMatcher 程序用于计算寡核苷酸频率(ONF) 分数以及结果的可视化。Galiez 等[73]开发的 WIsH 软件使用齐次马尔可夫模型(Homogeneous Markov Model) 预测噬菌体 contigs 中的细菌宿主,WIsH 可以快速准确的预测短噬菌体序列的宿主。Leite 等[74-75]基于噬菌体和宿主细菌结构域相互作用的得分情况和蛋白质一级结构信息,应用机器学习的方法预测相应宿主细菌。Lu 等[76]推出了原核病毒宿主预测器(Prokaryotic Virus Host Predictor, PHP) ,利用病毒和宿主基因组序列之间的k-mer频率差异作为特征来预测原核病毒,PHP 是使用高斯模型构建的交互式网页服务器,输入文件支持 FASTA 格式的纯核酸序列。相关机器学习的方法在识别噬菌体宿主细菌方面的应用研究请参考 Nami 等[77]的文章。

也有学者通过单细胞病毒荧光标记的实验方法来预测和鉴别人类肠道中噬菌体相应的宿主细菌[78]，而生物信息软件利用噬菌体和宿主细菌之间共有的生物特征或生物大分子模式通过机器学习的方法来预测宿主细菌，相较于实验方法可以在短时间内大批量分析鉴别多个样本中的噬菌体宿主细菌，提高研究效率，避免实验操作中出现的误差。

系统发育分析用于研究基因或物种之间的历史关系，并以分支图的形式描述这些关系，称为系统发育树(Phylogenetic analysis)。构建系统发育树常用的软件有 MEGA[79]、ggtree[80]、FastTree[81]、Cytoscape[82]，数理统计分析及可视化展现通常使用 R 语言来完成，常用的 R 包有 ggplot2、ggplot、ggiraph、ggfortify，有关使用 R 语言进行数据分析的详细内容可参考 Chan[83] 的文章。IMP[84](Integrated Meta-Omics Pipeline) 是基于 *De novo* 组装的开源生物信息分析流程，可对多组学数据集进行标准化、自动化、可重复的大规模集成分析，IMP 嵌套了多种生物信息分析工具用于实现流程一体化运行，包含的功能有原始数据的预处理、宏基因组或宏转录组数据的迭代共组装、微生物群落结构和功能的分析、自动装箱分类以及基于基因组特征的可视化展现(表 5)。

表 5 预测相应宿主细菌和构建系统发育树的软件工具

Table 5 Software tools for predicting the corresponding host bacteria and building phylogenetic trees

软件名称	软件描述	软件网址	参考文献
HostPhinderv	识别噬菌体基因组的宿主细菌	https://cge.cbs.dtu.dk/services/HostPhinder/	[71]
WIsH	识别病毒宏基因组中的宿主细菌	https://github.com/soedinglab/WIsH	[73]
PHP	预测原核病毒的宿主	http://www.computationalbiology.cn/phageHostPredictor/home.html	[76]
MEGA	分子进化遗传学分析	https://www.megasoftware.net/resources	[79]
ggtree	系统发育树的可视化呈现和注释	https://github.com/YuLab-SMU/ggtree	[80]
FastTree	构建系统发育树	http://www.microbesonline.org/fasttree/	[81]
Cytoscape	复杂网络可视化	https://cytoscape.org/	[82]
IMP	宏基因组和宏转录组数据处理和分析软件	https://imp.pages.uni.lu/web/	[84]
MG-RAST	宏基因组分析软件	https://www.mg-rast.org/	[85]

5 展 望

肠道内存在大量噬菌体，它们在调节肠道微生物群落动态平衡，动物宿主的生理机能与免疫系统中发挥着至关重要的作用。随着高通量测序技术的发展以及相应生物信息分析软件的开发与应用，使许多肠道病毒基因组数据可以被获取，相关数据库可以覆盖种类更加丰富的病毒基因组，如 GPD、ICTV、pVOGs、MG/VR、Viromes 等数据库的出现，让人们对肠道噬菌体以及其对肠道菌群和宿主生理健康的影响有了更广泛的认识。本文主要对肠道噬菌体宏基因组的分析流程和所需的相关工具软件以及数据库进行综述。机器学习和数理统计算法的兴起，使得肠道噬菌体组的分析不再仅局限于使用参考数据库和序列比对的方法，但是对噬菌体功能注释和筛选的准确性要低于序列比对得出的结果。并且宏基因组的数据量一般较大，在一些分析步骤中(如序列拼接组装和序列比对)会占用较大的内存和主存(RAM)空间，硬件配置较低的计算机在运行过程中会因此报错，且运行时间较长。病毒基因变异性强，只依据机器学习算法筛选噬菌体基因组或对其进行功能分析会有较大的误差。因为许多肠道细菌不能通过传统的体外培养技术存活，所以这些肠道细菌对应的噬菌体无法有效分离纯化，进而造成运用传统的分离-纯化-测序的研究思路分析噬菌体存在一定的局限性。肠道噬菌体组学另辟蹊径，避开分离纯化这个步骤，直接分析肠道噬菌体的结构和组成，具有一定的先进性，但是需要将生物信息分析与实验观察分析相结合，以提高结果的

准确性。目前大多数噬菌体的功能特性仍然无法得到解释,测序方法得到的噬菌体组序列大部分是新序列,噬菌体组的分类和功能注释依赖于病毒序列数据库,而数据库中记录的病毒信息是通过实验培养,质谱分析和分子生物学研究等得出的结论,其宿主范围,详细的生物功能和形态记录只能通过实验培养噬菌体来明确[86],体外培养噬菌体技术的限制导致现有的数据库无法对许多新噬菌体组序列进行分类和功能注释,这也是噬菌体组研究目前面临的瓶颈,并且需要开发通过更加精简的算法和脚本构建而成的生物信息分析工具,以减少软件在运行过程中对硬件的消耗。

参考文献:

[1] Hendrix RW, Smith MC, Burns RN, et al. Evolutionary relationships among diverse bacteriophages and prophages: All the world's a phage[J]. Proc. Natl. Acad. Sci. USA, 1999, 96(5): 2192-2197.

[2] Barr JJ. A bacteriophages journey through the human body[J]. Immunological Reviews, 2017, 279(1): 106-122.

[3] Guerin E, Hill C. Shining Light on Human Gut Bacteriophages[J]. Frontiers in Cellular and Infection Microbiology, 2020(10): 481.

[4] Shkoporov AN, Hill C. Bacteriophages of the human gut: The "known unknown" of the microbiome[J]. Cell Host & Microbe, 2019, 25(2): 195-209.

[5] Keen EC, Dantas G. Close Encounters of three kinds: Bacteriophages, commensal bacteria, and host immunity[J]. Trends in Microbiology, 2018, 26(11): 943-954.

[6] Federici S, Nobs SP, Elinav E. Phages and their potential to modulate the microbiome and immunity[J]. Cellular & Molecular Immunology, 2021, 18(4): 889-904.

[7] Nguyen S, Baker K, Padman BS, et al. Bacteriophage transcytosis provides a mechanism to cross epithelial cell layers[J]. mBio, 2017, 8(6): e01874-17.

[8] Qin J, Li R, Raes J, et al. A human gut microbial gene catalog established by metagenomic sequencing[J]. Nature, 2010, 464(7285): 59-65.

[9] Mirzaei MK, Maurice CF. Ménage à trois in the human gut: interactions between host, bacteria and phages[J]. Microbiology, 2017, 15(7): 397-408.

[10] Unterer M, Khan Mirzaei M, Deng L. Gut phage database: phage mining in the cave of wonders[J]. Signal Transduction and Targeted Therapy, 2021, 6(1): 193.

[11] Gregory AC, Zablocki O, Zayed AA, et al. The gut virome database reveals age-dependent patterns of virome diversity in the human gut[J]. Cell Host & Microbe, 2020, 28(5): 724-740.

[12] Camarillo-Guerrero LF, Almeida A, Rangel-Pineros G, et al. Massive expansion of human gut bacteriophage diversity[J]. Cell, 2021, 184(4): 1098-1109.

[13] Martin M. Cutadapt removes adapter sequences from high-through put sequencing reads[J]. EMBnet. journal, 2011, 17(1): 10-12.

[14] Davis MP, van Dongen S, Abreu-Goodger C, et al. Kraken: A set of tools for quality control and analysis of high-throughput sequence data[J]. Methods, 2013, 63(1): 41-49.

[15] Chen S, Huang T, Zhou Y, et al. AfterQC: Automatic filtering, trimming, error removing and quality control for fastq data[J]. BMC Bioinformatics, 2017, 18(3): 92-100.

[16] Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data[J]. Genome analysis, 2014, 30(15): 2114-2120.

[17] Xiang Y, Xin T, Yan T, et al. Transcriptome and 16S rRNA analyses revealed differences in the responses of largemouth bass (*Micropterus salmoides*) to early *Aeromonas hydrophila* infection and immunization[J]. Aquaculture, 2021(541): 736-759.

[18] Chen S, Zhou Y, Chen Y, et al. fastp: An ultra-fast all-in-one FASTQ preprocessor[J]. Bioinformatics, 2018, 34(17): i884-i890.

[19] Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2[J]. Nature Methods, 2012, 9(4): 357-359.

[20] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform[J]. Bioinformatics, 2009, 25(14): 1754-1760.

[21] National Center for Biotechnology Information (US), Bethesda (MD). BLAST Command Line Applications User Manual[Z]. 2021-06-28.

[22] Li H, Handsaker B, Wysoker A, et al. The Sequence alignment/Map format and SAMtools[J]. Bioinformatics, 2009, 25(16): 2078-2079.

[23] Danecek P, Bonfield JK, Liddle J, et al. Twelve years of SAMtools and BCFtools[J]. GigaScience, 2021, 10(2): giab008.

[24] Wingett SW, Andrews S. FastQ Screen: A tool for multi-genome mapping and quality control[J]. F1000Research, 2018, 7(2018): 1338.

[25] Ewels P, Magnusson M, Lundin S, et al. MultiQC: Summarize analysis results for multiple tools and samples in a single report[J]. Bioinformatics, 2016, 32(19): 3047-3048.

[26] Perrin S, Firmo C, Lemoine S, et al. Aozan: An automated post-sequencing data-processing pipeline[J]. Bioinformatics, 2017, 33(14): 2212-2213.

[27] Quinlan AR, Hall IM. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics[J]. Bioinformatics, 2010, 26(6): 841-842.

[28] Ghurye JS, Cepeda-Espinoza V, Pop M. Metagenomic assem-

bly: Overview, challenges and applications [J]. Yale Journal of Biology & Medicine, 2016, 89(3): 353-362.

[29] Sutton GG, White O, MD Adams, et al. TIGR assembler: A new tool for assembling large shotgun sequencing projects [J]. Genome Science & Technology, 2009, 1(1): 9-19.

[30] Jeck WR, Reinhardt JA, Baltrus DA, et al. Extending assembly of short DNA sequences to handle error [J]. Bioinformatics, 2007, 23(21): 2942-2944.

[31] Denisov G, Walenz B, Halpern AL, et al. Consensus generation and variant detection by celera assembler [J]. Bioinformatics, 2008, 24(8): 1035-1040.

[32] Koren S, Walenz BP, Berlin K, et al. Canu: Scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation [J]. Genome Research, 2017, 27(5): 722-736.

[33] Chin CS, Peluso P, Sedlazeck FJ, et al. Phased diploid genome assembly with single molecule real-time sequencing [J]. Nat Methods, 2017, 13(12): 1050-1054.

[34] Batzoglou S, Jaffe DB, Stanley K, et al. Arachne: A whole-genome shotgun assembler [J]. Genome Research, 2002, 12(1): 177-189.

[35] Chevreux B, Wetter T, Suhai S. Genome sequence assembly using trace signals and additional sequence information [J]. J Computer Sci Syst Biol, 1999, 99(1): 45-56.

[36] Myers EW, Sutton GG, Delcher AL, et al. A whole-genome assembly of drosophila [J]. Science, 2000, 287(5461): 2196-2204.

[37] Levy S, Sutton G, Ng PC, et al. The diploid genome sequence of an individual human [J]. PLoS Biol, 2007, 5(10): e254.

[38] Peng Y, Leung HC, Yiu SM, et al. IDBA-UD: A *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth [J]. Bioinformatics, 2012, 28(11): 1420-1428.

[39] Prjibelski A, Antipov D, Meleshko D, et al. Using SPAdes *De novo* assembler [J]. Current Protocols in Bioinformatics, 2020, 70(1): e102.

[40] Li D, Luo R, Liu CM, et al. MEGAHIT v1.0: A fast and scalable metagenome assembler driven byadvanced methodologies and community practices [J]. Methods, 2016(102): 3-11.

[41] Luo R, Liu B, Xie Y, et al. SOAPdenovo2: An empirically improved memory-efficient short-read *de novo* assembler [J]. GigaScience, 2012, 1(1): 18.

[42] Souvorov A, Agarwala R, Lipman DJ. SKESA: strategic k-mer extension for scrupulous assemblies [J]. Genome Biology, 2018, 19(1): 153.

[43] Nurk S, Meleshko D, Korobeynikov A, et al. metaSPAdes: A new versatile metagenomic assembler [J]. Genome Research, 2017, 27(5): 824-834.

[44] Mikheenko A, Prjibelski A, Saveliev V, et al. Versatile genome assembly evaluation with QUAST-LG [J]. Bioinformatics, 2018, 34(13): i142-i150.

[45] The sam/bam format specification working group. Sequence Alignment/Map Format Specification [Z]. 2021.

[46] Lukashin AV, Borodovsky M. GeneMark.hmm: New solutions for gene finding [J]. Nucleic Acids Res, 1998, 26(4): 1107-1115.

[47] Majoros WH, Pertea M, Salzberg SL. TigrScan and glimmer-HMM: Two open source ab initio eukaryotic gene-finders [J]. Bioinformatics, 2004, 20(16): 2878-2819.

[48] Mikheenko A, Saveliev V, Gurevich A. MetaQUAST: Evaluation of metagenome assemblies [J]. Bioinformatics, 2016, 32(7): 1088-1090.

[49] Nayfach S, Camargo AP, Schulz F, et al. CheckV assesses the quality and completeness of metagenome-assembled viral genomes [J]. Nature Biotechnology, 2021, 39(5): 578-585.

[50] Prakash T, Taylor TD. Functional assignment of metagenomic data: challenges and applications [J]. Brief Bioinform, 2012, 13(6): 711-727.

[51] Altschul SF, Madden TL, Schäffer AA, et al. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs [J]. Nucleic Acids Research, 1997, 25(17): 3389-3402.

[52] Lefkowitz EJ, Dempsey DM, Hendrickson RC, et al. Virus taxonomy: The database of the international committee on taxonomy of viruses (ICTV) [J]. Nucleic Acids Research, 2018, 46(D1): D708-D717.

[53] Grazziotin AL, Koonin EV, Kristensen DM. Prokaryotic virus orthologous groups (pVOGs): are source for comparative genomics and protein family annotation [J]. Nucleic Acids Research, 2017, 45(D1): D491-D498.

[54] Roux S, Páez-Espino D, Chen IA, et al. IMG/VR v3: An integrated ecological and evolutionary framework for interrogating genomes of uncultivated viruses [J]. Nucleic Acids Research, 2021, 49(D1): D764-D775.

[55] Arndt D, Grant JR, Marcu A, et al. PHASTER: a better, faster version of the PHAST phage search tool [J]. Nucleic Acids Research, 2016, 44(W1): W16-21.

[56] Vázquez-Castellanos JF, García-López R, Pérez-Brocal V, et al. Comparison of different assembly and annotation tools on analysis of simulated viral metagenomic communities in the gut [J]. BMC Genomics, 2014, 15(1): 37.

[57] Bzhalava Z, Tampuu A, Bała P, et al. Machine learning for detection of viral sequences in human metagenomic data set [J]. BMC Bioinformatics, 2018, 19(1): 336.

[58] Song W, Sun HX, Zhang C, et al. Prophage hunter: an integrative hunting tool for active prophages [J]. Nucleic Acids Research, 2019, 47(W1): W74-W80.

[59] Amgarten D, Braga LPP, da Silva AM, et al. MARVEL, a tool for prediction of bacteriophage sequences in metagenomic

bins[J]. Frontiers in Genetics, 2018(9): 304.

[60] Liu F, Miao Y, Liu Y, et al. RNN-VirSeeker: A deep learning method for identification of short viral sequences from metagenomes[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2022, 19(3):1840-1849.

[61] Guo J, Bolduc B, Zayed AA, et al. Sullivan. VirSorter2: A multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses[J]. Microbiome, 2021, 9(1):37.

[62] Wommack KE, Bhavsar J, Polson SW, et al. VIROME: A standard operating procedure for analysis of viral metagenome sequences[J]. Standards in Genomic Sciences, 2012, 6(3): 427-439.

[63] Kieft K, Zhou Z, Anantharaman K. VIBRANT: Automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequence[J]. Microbiome, 2020, 8(1): 90.

[64] Ren J, Ahlgren NA, Lu YY, et al. VirFinder: A novel k-mer based tool for identifying viral sequences from assembled metagenomic data[J]. Microbiome, 2017, 5(1):69.

[65] Tampuu A, Bzhalava Z, Dillner J, et al. ViraMiner: Deep learning on raw DNA sequences for identifying viral genomes in human samples[J]. Plos One, 2019, 14(9): e0222271.

[66] Ren J, Song K, Deng C, et al. Identifying viruses from metagenomic data using deep learning[J]. Quantitative Biology, 2020, 8(1): 64-77.

[67] Mizuno CM, Rodriguez-Valera F, Kimes NE, et al. Expanding the marine virosphere using metagenomics[J]. Plos Genetics, 2013, 9(12): e1003987.

[68] Bellas CM, Anesio AM, Barker G. Analysis of virus genomes from glacial environments reveals novel virus groups with unusual host interactions[J]. Frontiers in Microbiology, 2015(6):656.

[69] Boeckaerts D, Stock M, Criel B, et al. Predicting bacteriophage hosts based on sequences of annotated receptor-binding proteins[J]. Scientific Reports, 2021, 11(1):1467.

[70] Young F, Rogers S, Robertson DL. Predicting host taxonomic information from viral genomes: A comparison of feature representations[J]. PLOS Computational Biology, 2020, 16(5): e1007894.

[71] Villarroel J, Kleinheinz KA, Jurtz VI, et al. HostPhinder: A phage host prediction Tool[J]. Viruses, 2016, 8(5):116.

[72] Ahlgren NA, Ren J, Lu YY, et al. Alignment-free d_2^* oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences[J]. Nucleic Acids Research, 2016, 45(1): 39-53.

[73] Galiez C, Siebert M, Enault F, et al. WIsH: Who is the host? predicting prokaryotic hosts from metagenomic phage contigs[J]. Bioinformatics, 2017, 33(19): 3113-3114.

[74] Leite DMC, Lopez JF, Brochet X, et al. Exploration of multiclass and one-class learning methods for prediction of phage-bacteria interaction at strain level[J]. Ieee, 2018, pp. 1818-1825.

[75] Leite DMC, Brochet X, Resch G, et al. Computational prediction of inter-species relationships through omics data analysis and machine learning[J]. BMC Bioinformatics, 2018, 19(14):152-159.

[76] Lu C, Zhang Z, Cai Z, et al. Prokaryotic virus host predictor: A Gaussian model for host prediction of prokaryotic viruses in metagenomics[J]. BMC Biology, 2021, 19(1):5.

[77] Nami Y, Imeni N, Panahi B. Application of machine learning in bacteriophage research[J]. BMC Microbiology, 2021, 21(1):193.

[78] DKžunková M, Low SJ, Daly JN, et al. Defining the human gut host-phage network through single-cell viral tagging[J]. Nature Microbiology, 2019, 4(12): 2192-2203

[79] Kumar S, Stecher G, Li M, et al. MEGA X: Molecular evolutionary genetics analysis across computing platforms[J]. Molecular Biology & Evolution, 2018, 35(6): 1547-1549.

[80] Yu G. Using ggtree to Visualize Data on tree-like structure[J]. Current Protocols in Bioinformatics, 2020, 69(1): e96.

[81] Price MN, Dehal PS, Arkin AP. FastTree: Computing large minimum evolution trees with profiles instead of a distance matrix[J]. Molecular Biology & Evolution, 2009, 26(7): 1641-1650.

[82] Shannon P, Markiel A, Ozier O, et al. Cytoscape: A software environment for integrated models of biomolecular interaction networks[J]. Genome Research, 2020, 13(11):2498-504.

[83] Chan BKC. Data analysis using R programming[J]. Adv Exp Med Biol, 2018(1082): 47-122.

[84] Narayanasamy S, Jarosz Y, Muller EE, et al. IMP: A pipeline for reproducible reference-independent integrated metagenomic and metatranscriptomic analyses[J]. Genome Biology, 2016, 17(1):260.

[85] Meyer F, Bagchi S, Chaterji S, et al. MG-RAST version 4-lessons learned from a decade of low-budget ultra-high-throughput metagenome analysis[J]. Brief Bioinform, 2019, 20(4): 1151-1159.

[86] Moon K, Cho JC. Metaviromics coupled with phage-host identification to open the viral 'black box'[J]. Journal of Microbiology, 2021, 59(3):311-323.