中国科学: 生命科学 2023 年 第 53 卷 第 5 期: 647 ~ 659
SCIENTIA SINICA Vitae lifecn.scichina.com
《中国科学》杂志社 SCIENCE CHINA PRESS
评 述 中国科学院学部 科学与技术前沿论坛 微生物群系与大健康专辑

肠道病毒组学数据挖掘与分析方法的进展及挑战

江小青, 李墨, 尹衡闯, 郭倩, 谭洁, 吴姝芳, 王春晖, 朱怀球*

北京大学未来技术学院生物医学工程系, 北京大学定量生物学中心, 北京 100871
* 联系人, E-mail: hqzhu@pku.edu.cn

收稿日期: 2022-12-27; 接受日期: 2023-02-08; 网络版发表日期: 2023-04-26
国家自然科学基金(批准号: 32070667, 31671366)和国家重点研发计划重点专项(批准号: 2021YFC2300300, 2017YFC1200205)资助

摘要 肠道病毒对肠道微生物群系的种群结构、细菌性状乃至人体健康都有十分重要的影响, 但相比肠道细菌, 人们对其的研究和了解仍然很缺乏. 高通量测序技术以及机器学习、深度学习等方法的快速发展, 为从组学途径深入研究肠道病毒提供了契机. 本文针对当前肠道病毒组学领域以噬菌体、真核病毒等为对象的高通量数据, 总结并分析了近年来数据挖掘和分析的共性方法和技术的发展, 梳理了一系列相关的生物信息学方法和技术, 其中大多适用于基于宏基因组或宏病毒组两种策略的病毒组学分析. 同时, 针对目前实际生物学问题和临床问题的复杂性, 人工智能方法在生物信息学领域的广泛运用, 以及未来三代测序技术可能的广泛使用, 讨论了病毒组学数据挖掘与数据分析方法面临的问题和挑战.

关键词 肠道病毒组, 宏基因组, 组学分析, 数据挖掘

随着高通量测序技术的发展, 微生物群系对人类健康和环境的重要性已经得到越来越多的关注. 其中, 肠道微生物群系尤为重要, 肠道菌群失衡与多种疾病息息相关, 加上细菌在其中占绝对主要的比例(例如, 细菌的遗传物质占人肠道微生物总和的93%左右[1]), 因此在过去相当长的时期, 肠道微生物群系的研究主要以细菌为主. 实际上, 细菌具有如16S rRNA等可用于系统分类的生物标记物, 其物种组成、丰度以及动力学行为[2]等分析亦比较方便操作, 相关研究结果也获得广泛的关注. 随着对微生物群系复杂性越来越深入的了解, 人们逐渐意识到, 除细菌外, 微生物群系中的其他成员同样非常重要, 相当数量的来自非染色体元件的遗传物质, 如噬菌体、真核病毒、质粒等, 它们对整个微生物群系(尤其是人肠道微生物)的种群结构、细菌性状等产生非常重要的影响.

作为地球上数量最丰富的生命体, 病毒几乎无处不在, 它们广泛存在于海洋、深海热泉、土壤到人体等各种环境[3], 总数量约达10^{31}个[4]. 据估计, 自然界中约存在百万数量级的病毒物种[5], 但迄今为止仅有几万种收录于美国国家生物技术信息中心(National Center for Biotechnology Information, NCBI)的Refseq数据库(https://www. ncbi.nlm.nih.gov/refseq/). 人肠道被认

引用格式: 江小青, 李墨, 尹衡闯, 等. 肠道病毒组学数据挖掘与分析方法的进展及挑战. 中国科学: 生命科学, 2023, 53: 647–659
Jiang X Q, Li M, Yin H C, et al. Data mining and analysis techniques for gut virome: the prospects and challenges (in Chinese). Sci Sin Vitae, 2023, 53: 647–659, doi: 10.1360/SSV-2022-0330

为携带了数量最多、组成最丰富的与人体共同栖息的微生物，这类微生物群系包含了原核生物(细菌、古菌)、真核生物(真菌、原生动物)、病毒等[1]. 对于肠道微生物群系，其中所有病毒的集合，被认为包括感染细菌的噬菌体以及感染人、动物、植物、原生生物等真核生物的真核病毒；对它们进行高通量测序所获得的基因组序列，即肠道病毒组(gut virome)，也由此成为肠道微生物群系的组学研究对象之一. 据估计, 每克粪便中至少有10^9个病毒颗粒[6]. 尽管病毒在肠道中的含量不可忽略，人们对病毒的了解仍十分有限. 2018年，全球病毒组计划(The Global Virome Project)正式启动, 该项目旨在鉴定地球上大部分未知病毒, 并制止可能出现的病毒大流行. 与此同时, 高通量组学技术及宏基因组学(以及宏转录组、宏蛋白质组、宏代谢组等)的快速发展使对肠道病毒组学的深入研究成为可能. 近年来, 通过对人肠道病毒组的组成、结构与稳态变化进行分析，人们发现肠道病毒组不仅具有很强的个体特异性[7]，且大多由未曾收录过的新病毒组成[8]. 此外，诸如炎症性肠病[9]、营养不良[10]与II型糖尿病[11]等不良状态也被证明与肠道病毒组分变化有关. 从PubMed检索可以看出(图1), 病毒组学以及肠道病毒组学的研究论文数量近年来呈现稳定增长的趋势.

本文围绕当前以噬菌体、真核病毒为对象的基因组高通量测序数据的肠道病毒组学研究，试图对组学数据挖掘和分析的共性技术进行系统梳理和归纳. 本文既考虑通过病毒颗粒富集获得宏病毒组(metavirome)的途径，也讨论包含病毒核苷酸序列在内的微生物群落的全基因组测序(即宏基因组或元基因组，metagenome)的途径，由此主要产生了病毒(尤其是噬菌体)的序列拼接、基因注释、序列鉴定、宿主预测、系统分类，以及从病毒序列出发的基因组学分析等一系列方法和技术的共性问题. 本文力图系统地介绍在这些共性问题上近年来发展的方法和工具，其中也包括本课题组在这一问题上的若干工作. 同时，针对目前实际生物学问题和临床问题的复杂性，以及深度学习等人工智能方法在生物信息学领域的广泛运用、未来三代测序技术可能的广泛使用，本文也致力于讨论这些方法和技术面临的困难与挑战.

1 肠道病毒组分析

1.1 肠道病毒组分析的两种策略

微生物群系的组学分析流程包括采样、测序、数

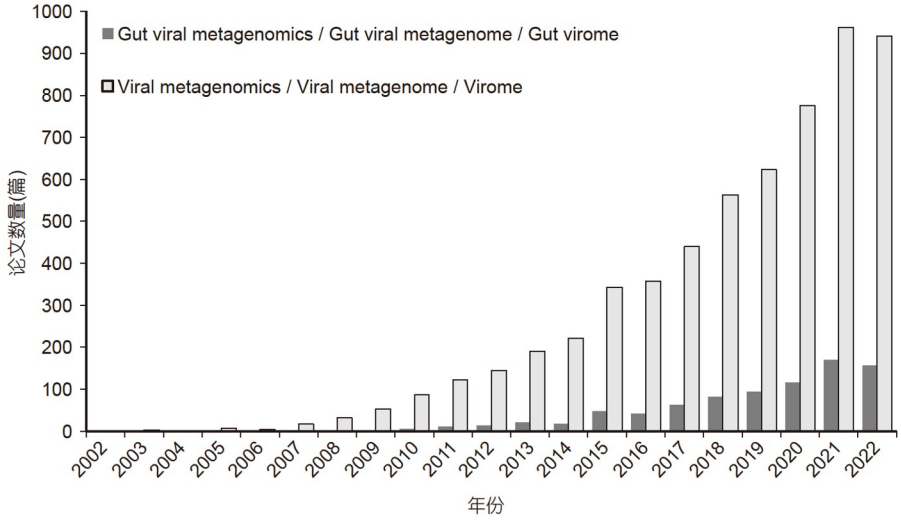

图 1 病毒组和肠道病毒组研究论文数量快速增长. 病毒组研究正在成为近年来的热点, 与此同时, 肠道病毒组学研究的热度增长值得期待. PubMed网站上以关键词(viral metagenomics) OR (viral metagenome) OR (virome)和(gut viral metagenomics) OR (gut viral metagenome) OR (gut virome)进行搜索, 统计截至2022年12月31日

Figure 1 Number of publications in virome and gut virome research is growing rapidly. Virome is becoming a hotspot in recent years, and gut virome also develops fast. We searched PubMed using "(viral metagenomics) OR (viral metagenome) OR (virome)" and "(gut viral metagenomics) OR (gut viral metagenome) OR (gut virome)" as the query terms, respectively. Only publications published before Dec 31, 2022 are included

据处理分析等主要步骤，这些步骤也同样适用于病毒组研究. 由于病毒在许多样品中的生物量比较小，因此需要在采样及保存过程中非常注意环境的污染. 在采样过后, 样品应当立即置于−80℃的环境冷冻. 对于肠道微生物，DNA病毒占肠道病毒的绝大多数，加上RNA病毒在样本中的稳定性不如DNA病毒，使得宏基因组测序难以识别相应的病毒，所以主要通过提取粪便样品中的DNA来研究病毒. 目前，通过高通量测序手段来研究病毒组通常有两种策略: 一种是常规的宏基因组测序，即将一个微生物群落中所有的双链DNA都提取出来，包括细菌、古菌、病毒等微生物. 测序后，通过生物信息学方法和工具来鉴定其中的病毒序列. 但因此也给相应的生物信息学分析方法和工具带来困难和挑战. 另一种方式是宏病毒组测序，即先对样品中的病毒颗粒进行富集，然后进行遗传物质的提取和测序. 对于常规的宏基因组测序，首先需要从样品中提取遗传物质，许多方便且高效的试剂盒可用于提取样品中的DNA. 而对于病毒颗粒富集的宏病毒组测序，需要经过离心、过滤等操作以去除尺寸较大的宿主细胞核及其他微生物细胞，然后使用乙二醇沉淀法来富集病毒颗粒，最后提取其中的DNA. 值得注意的是，这种对病毒颗粒进行富集和分离的宏病毒组研究手段只能研究群落中活跃的病毒颗粒，以溶原状态整合在宿主基因组中的病毒则跟随宿主细胞一起被去除，而这部分溶原的病毒可能占整个群落中病毒的很大一部分. 研究人员可根据研究目的选择病毒组的研究策略, 这两种策略的优缺点总结见表1.

上述两种病毒组学研究策略都存在进一步的生物信息学分析的共性技术需求和问题. 具体地说, 提取样品中遗传物质后, 下一步是建库和上机测序. 目前, 主流的二代Illumina测序平台所测得的原始序列一般为150或300 bp长度的短读序(short reads). 在得到宏基因组或宏病毒组的原始测序序列后，首先要进行质量控制，包括去除环境宿主的序列，以及去除低质量的序列. 去除环境宿主序列的方法主要是通过将短读序比对到宿主基因组上后去除. 由于病毒等微生物的基因组序列一般与哺乳动物宿主的基因组序列差异较大, 这种方法能够高效地去除环境宿主的序列. 接下来就可以对干净的序列进行后续分析. 由于序列库中包含了来自多种不同物种的基因组序列，对测序数据进行分析是宏基因组及宏病毒组研究的最重要的挑战之一. 近年来, 围绕这一需求, 人们发展了一系列的生物信息学方法和技术(表2)，其中大多数可适用于基于宏基因组或宏病毒组两种策略的病毒组学分析，表中也包括本课题组多年来完成的若干工作.

1.2　肠道病毒组数据的拼接、注释与分析

一般来说，对宏基因组和宏病毒组测序数据分析最基本的目标，是要得到样本中的所有微生物或病毒的物种及功能基因的组成. 这通常可以采取两种办法: 第一种是直接将短读序比对到已知的基因组或蛋白序列上进行物种注释；第二种是从头拼接而得到更长的基因组片段序列，然后进一步对长序列或预测到的基因进行物种或功能注释. 根据不同的研究目的，可以采取不同的办法. 直接比对方法适用于已知病毒和微生物的鉴定，其优点在于简单快速，缺点是对于数据库中没有记录的物种的序列只能被遗弃，从而损失相当一部分的测序数据. 而一般的微生物群落测序样本中都含有大量数据库中没有记录的微生物和病毒序列. 从头拼接方法可以获得更长的基因组序列, 也能进

表 1　宏基因组测序和病毒富集的宏病毒组测序两种策略在病毒组研究中的比较

Table 1　Comparison of strategies based on viral metagenomes and metaviromes in virome studies

比较	病毒宏基因组测序	宏病毒组测序
优点	全部双链DNA物质, 包括整合在细菌基因组上的病毒序列 能够同时对细菌和病毒进行鉴定, 便于研究细菌和病毒之间的互作关系 样本携带了病毒及其感染的细菌宿主关联的信息	所测的是有活性的病毒颗粒序列 使用多重置换扩增(multiple displacement amplification, MDA)进行建库, 可以测到单链DNA病毒 便于直接分析和讨论样本群落中病毒的组成和丰度等信息
缺点	遗漏单链DNA病毒及RNA病毒 缺乏准确、高效的病毒、细菌等序列鉴定方法和工具	MDA扩增对单链DNA病毒具有偏好性 不能同时估计病毒及群落中其他微生物的含量, 不能构建病毒与其他微生物的互作网络 遗漏整合在细菌基因组上的噬菌体

表 2 病毒宏基因组及宏病毒组研究中常用的生物信息学方法和工具

Table 2 Bioinformatics methods and tools commonly used in viral metagenomes and metaviromes

软件	功能	应用
metaSPAdes (http://cab.spbu.ru/software/metaspades/)[12]	序列拼接	metaSPAdes是一个基于de Bruijn图算法, 并在其基础上进行优化, 使其适用于宏基因组数据的从头拼接软件. 目前, 它在宏基因组拼接领域各项评估指标都比较靠前, 尤其对宏基因组中病毒的拼接效果显著[13]. 该工具也能高精度地拼接菌株水平的基因组
MetaviralSPAdes (https://github.com/ablab/spades/tree/metaviral_publication)[14]	序列拼接	MetaviralSPAdes是一个专门针对宏基因组数据中病毒的拼接软件. 该方法在metaSPAdes的算法基础上, 考虑了病毒和细菌之间测序深度和覆盖度之间的差异, 从而更有效地针对病毒序列进行拼接
IDBA-UD (https://github.com/loneknight-py/idba)[15]	序列拼接	IDBA-UD是一个基于de Bruijn图算法, 用于测序深度非常不均匀的单细胞和宏基因组测序数据的拼接软件
InteMAP (http://cqb2.pku.edu.cn/ZhuLab/InteMAP/)[16]	序列拼接	InteMAP是一个整合了ABySS, IDBA-UD, CABOG三个拼接软件的宏基因组从头拼接软件. 对这三个拼接软件的整合使它们的缺点可以互补, 从而达到更好的拼接效果
MAP (http://cqb2.pku.edu.cn/ZhuLab/MAP/)[17]	序列拼接	MAP针对Sanger和454测序技术下的宏基因组DNA序列, 有效利用mate pair的测序信息, 考虑多基因组混合的宏基因组测序特点进行序列拼接
MetaGeneMark (http://exon.gatech.edu/meta_gmhmmp.cgi)[18]	基因预测	MetaGeneMark开发了通过推导不同来源基因组序列的三周期的二阶隐马尔可夫模型(hidden Markov model, HMM)的参数, 来对宏基因组片段进行基因预测. 模型中包含病毒基因组信息, 所以也可以用于病毒基因的预测
MetaGeneAnnotator (http://metagene.nig.ac.jp/)[19]	基因预测	MetaGeneAnnotator是一个整合了原噬菌体、细菌、古菌基因的统计模型的宏基因组基因预测软件, 通过对输入序列建立自学习模型进行预测, 使其不仅可以预测典型的基因, 也可以预测非典型的基因
MetaGUN (http://cqb2.pku.edu.cn/ZhuLab/MetaGUN/)[20]	基因预测	MetaGUN是一个基于支持向量机模型的宏基因组基因预测软件. 该算法首先根据序列的k-mer对序列进行无监督分类, 然后将密码子使用偏好熵密度谱、翻译起始位点得分和开放阅读框(open reading frame, ORF)长度作为特征输入支持向量机, 最后用MetaTISA[21]对翻译起始位点进行校正
PlasGUN (http://cqb2.pku.edu.cn/ZhuLab/PlasGUN/)[22]	基因预测	PlasGUN是一个针对质粒宏基因组数据进行基因预测的工具, 其将每一个可能的ORF密码子序列、起始密码子上下游序列、ORF长度与完整性等特征输入卷积神经网络来对基因进行预测
NCBI BLAST+ (https://blast.ncbi.nlm.nih.gov/Blast.cgi) [23]	序列比对	BLAST可用于核酸或蛋白序列比对, 通过自定义的参考库进行物种或功能注释
DIAMOND (https://ab.inf.uni-tuebingen.de/software/diamond/) [24]	序列比对	DIAMOND可用于核酸或蛋白序列比对, 但其所比对的库只能是蛋白库. 该工具运行速度是BLAST的几千到上万倍, 常用于宏基因组序列的功能注释
HMMER (http://hmmer.org/)[25]	序列比对	HMMER是基于HMM进行序列比对的工具, 与BLAST和DIAMOND的区别在于其比对的数据库不是直接的序列, 而是预先建好的HMM模型, 一般通过同一个蛋白家族的多序列比对结果用hmmbuild程序得到. HMMER与BLAST和DIAMOND相比可以得到更多同源序列的比对结果, 常用于未知病毒序列的鉴定
Kaiju (https://kaiju.binf.ku.dk/)[26]	物种注释	Kaiju是一个基于蛋白比对的宏基因组物种注释工具. 该工具预先下载指定数据库中所有的蛋白序列, 然后利用Borrows-Wheeler转换和启发式贪婪搜索以加速比对过程, 以及允许搜索过程中任意的氨基酸替代. 在比对完之后, Kaiju可以方便地通过最近公共祖先(lowest common ancestor, LCA)方法生成各个分类水平上的丰度矩阵
Kraken2 (https://ccb.jhu.edu/software/kraken2/)[27]	物种注释	Kraken2是一个基于k-mer分类的宏基因组物种注释软件
Clark (https://clark.cs.ucr.edu/)[28]	物种注释	Clark是一个基于k-mer分类的宏基因组物种注释软件
Metavir2 (http://metavir-meb.univ-bpclermont.fr/)[29]	分析流程	Metavir2是一个用于病毒宏基因组测序数据的分析软件, 输入拼接后的contig, 它可以对序列进行基因预测, 比对病毒参考基因组序列数据库及蛋白结构域数据库, 并且对不同的样本进行比较分析
VIROME (http://virome.dbi.udel.edu/)[30]	分析流程	VIROME是一个提供对病毒宏基因组数据进行标准分析流程的工具, 包括预测ORF, 基于序列同源性比对的物种和功能注释等
MEGAN (https://github.com/husonlab/megan-ce/)[31]	分析流程	MEGAN是一个综合的分析工具, 能对宏基因组、宏转录组、宏蛋白组及rRNA数据进行分析, 包括基于NCBI的物种注释和基于KEGG数据库的功能注释

一步提高对物种和基因的鉴定和注释准确性，从而更好地研究微生物群系的特征.

拼接是根据序列之间的重叠将短读序拼接成基因组上连续的更长的序列，直到形成重叠群(contig)或脚手架(scaffold). 从头拼接软件通常利用de Bruijn图或作为OLC(overlap-layout-consensus)方法的一部分的序列重叠图来进行拼接. 与单基因组相比, 宏基因组或宏病毒组的序列拼接对算法提出了更高的要求，其主要挑战在于[16,17,32,33]: (i) 样本中的基因组数量多且由于丰度不一致引起测序深度的不统一，由此导致宏基因组或宏病毒组中每个基因组的覆盖度常小于单一测序的基因组，且基于深度进行测序错误校正十分困难; (ii) 一个微生物群系中的不同菌株或物种间有很多高度保守的区域, 即物种间的重复序列, 由此可能造成跨物种的错误拼接; (iii) 一个物种的细菌或病毒通常是很多个菌株或病毒株的混合体. 针对这些挑战, 人们开发了许多成熟的拼接软件，以专门用于宏基因组的序列拼接，如metaSPAdes[12]，IDBA-UD[15]，InteMAP[16]，MAP[17]等. 这些软件通常也被用于宏病毒组的拼接. 值得关注的是，metaSPAdes的开发团队最近发布了针对宏基因组中病毒序列进行拼接的版本——MetaviralSPAdes[14]. 该方法在metaSPAdes的算法基础上进行了改进，考虑了病毒和细菌之间测序深度和覆盖度之间的差异，从而更有效地针对病毒序列进行拼接，可以很好地在种水平拼接出病毒的全基因组而非片段化的重叠群.

在得到拼接后的重叠群后，可以对序列进行基因预测. 宏基因组学研究中通常采用从头开始(*ab initio*)的方法进行基因预测[20]，即不同于数据库搜索的方式，从头开始的基因预测根据开放阅读框(open reading frame，ORF)长度、密码子使用偏好、GC含量、熵密度分布和翻译起始信号打分等参数构建模型，使用机器学习的方式进行基因预测. 目前已有许多宏基因组的基因预测软件，但这些软件并非都能用于病毒基因的预测，需要软件针对病毒基因构建模型. 常见的能用于病毒基因预测软件有MetaGeneMark[18]，MetaGeneAnnotator[19]，MetaGUN[20]等. 此外，本课题组[22]开发的PlasGUN是针对质粒宏基因组短读序数据的基因预测工具, 也可用于噬菌体的基因预测. 该方法使用了一个多输入的卷积神经网络的深度学习模型，从多维度提取每个质粒候选ORF的序列特征，包括ORF的密码子序列、起始密码子上下游序列、ORF长度与完整性等. 因为卷积核会在ORF序列中逐个密码子进行扫描，与使用全局统计量的工具相比可以更有效地检测序列局部特征. 通过模拟的基准数据集的测试表明, 与传统的基因预测工具相比，PlasGUN在质粒短读序数据中的预测精度有很大提升，并且在可能包含新基因的序列上的预测表现尤其明显.

得到病毒基因后，进一步可对其进行物种注释或功能注释，通常使用BLAST[23]和DIAMOND[24]等比对工具将查询序列比对到数据库中已有注释信息的序列上, 对样本中的序列进行注释. 常用的物种注释数据库有GenBank[34]，RefSeq[35]等. 除BLAST外, 有些工具会将病毒序列的分析流程进行一定的整合，如VIROME[30]，这是一种集成了多个序列和功能数据库搜索结果的全面工具; MEGAN[33]提供了普遍适用的宏基因组分类器，它通过BLAST的结果推断给定序列的最低共同祖先，并通过图形界面提供功能分析; MetaVir2[29]提供网页版应用，将用户输入的数据集与已发布的病毒序列进行比较. 此外，也有工具为了加速直接比对, 采用比对k-mer的方法, 如Kraken2[27]，Clark[28]等.

基因功能注释一般通过将序列比对到参考的蛋白库来实现. COG(Clusters of Orthologous Groups of proteins)和KEGG(Kyoto Encyclopedia of Genes and Genomes)是两个经典的蛋白功能注释数据库，其中COG是原核生物同源蛋白簇数据库[36]，KEGG是包含真核基因和原核基因以及提供详细功能注释的数据库[37]. Grazziotin等人[38]构建了噬菌体的蛋白数据库pVOG(Prokaryotic Virus Orthologous Groups), 其包含3000多个可用的NCBI病毒基因组的基因序列和蛋白注释. ACLAME数据库则是一个整合了NCBI，SwissProt，SCOP等蛋白库数据的、全面的、可移动元件的蛋白库[39]，包括噬菌体蛋白、质粒蛋白和其他病毒的蛋白序列. 这些数据库收集了大量病毒蛋白序列，然而，它们对病毒蛋白的注释信息和分类信息还不够全面. 本课题组[40]构建的VirGenFunD是一个有详细注释和功能分类信息的疾病相关的人肠道病毒基因数据库(https://yjiang724.github.io/VirGenFunD/)，该数据库目前收录了438个疾病和健康人群的肠道宏基因组种所鉴定的病毒基因序列，可望为肠道病毒基因的功能注释提供有帮助的参考库.

2 肠道病毒组数据挖掘与生物信息学分析的重要问题

2.1 宏基因组/宏转录组数据中病毒序列的鉴定

上述的病毒宏基因组数据序列注释过程目前依赖于蛋白数据库, 然而宏基因组数据中含有大量与已知病毒并不相似的病毒序列. 据估计, 肠道宏基因组中仅有约15%的病毒与已知病毒相似[9]. 从混杂着大量短片段的宏基因组或宏转录组数据中准确、高效鉴定出新病毒序列, 是当前十分重要的一项技术需求.

目前已有一些工具可以从细菌基因组中识别病毒序列, 包括Prophinder[41], Phage_Finder[42], PhiSpy[43], PHAST(其增强版PHASTER)[44,45], ProphET[46]. 这些工具主要采用一个扫描窗口在细菌基因组上滑动, 并通过对病毒数据库的相似性搜索来检测病毒序列. 这些工具通常要求扫描窗口的长度能够覆盖若干个基因, 而宏基因组序列一般较短, 常常无法包含完整的基因, 因此这些工具不太适用于宏基因组序列. 此外, 这些工具也无法检测到未整合到细菌基因组上的噬菌体序列, 如烈性噬菌体和部分温和噬菌体序列.

近年来也有一些适用于从宏基因组序列中识别病毒序列的工具, 它们大都是不依赖于序列比对(alignment-free)的方法(如VirFinder[47], DeepVirFinder[48], MARVEL[49], PPR-meta[50], VirMC[51]), 以及少数基于序列比对的方法(如VirSorter[52]及其改进版VirSorter2[53]). VirFinder是一个基于k-mer特征的病毒预测软件, 通过逻辑回归模型对原核病毒及其宿主进行区分. DeepVirFinder是一个基于深度学习的方法, 其准确率高于VirFinder. MARVEL则使用随机森林模型对宏基因组序列中的双链DNA噬菌体序列进行鉴定, 但是MARVEL需要输入较长的序列来实现较高的准确率. 本课题组最近开发的PPR-Meta(https://github.com/zhenchengfang/PPR-Meta)采用基于双通道卷积神经网络的深度学习算法, 可以将宏基因组序列区分为来自噬菌体、染色体或质粒的序列. VirMC是一种基于马尔可夫链的方法, 其优势在于可以在含有真核序列的污染宏基因组样本中识别病毒序列. VirSorter可以从宏基因组序列中鉴定新病毒, 其缺点在于鉴定噬菌体的灵敏性较低. VirSorter2是一个针对双链DNA噬菌体、单链DNA病毒、RNA病毒、巨型病毒和噬病毒体的五分类器, 其优势在于对真核病毒的预测和未知病毒预测的能力. 这些方法大多采用机器学习或深度学习等算法进行病毒序列的预测, 但受限于需要较长的序列(如拼接好的重叠群)来达到较好的预测效果. 由于噬菌体的序列拼接效果往往不如宿主染色体DNA的序列拼接效果[54], 因此, 宏基因组数据中病毒序列的鉴定问题仍然是一个方法研究上的挑战.

2.2 噬菌体类型(烈性、温和)鉴定

噬菌体作为肠道中最主要的病毒, 在微生物群系中的重要作用越来越受到关注. 研究结果表明, 肠道中的噬菌体包含大量的遗传信息, 并与其宿主细菌共同影响着肠道环境, 对人类健康有十分重要的影响. 按生活模式分, 噬菌体可分为烈性噬菌体与温和噬菌体两类. 温和型噬菌体首先会以原噬菌体的形式将自身的基因组插入宿主细菌的基因组上, 或者作为游离的病毒粒子与宿主细菌形成稳定的共存关系. 在合适的环境下, 温和噬菌体从细菌基因组上游离出来, 然后通过裂解的方式杀死宿主细菌[55]. 而烈性噬菌体感染宿主细菌后会立即产生大量后代, 导致宿主细菌裂解死亡. 噬菌体不同的生活模式会对宿主、环境产生不同的影响, 因此有必要对噬菌体的类型进行区分. 传统的分类方法是通过分离培养直接观察噬菌体的生活模式, 从而鉴定噬菌体类型. 但这种方法会消耗大量的时间和资源, 同时噬菌体在传统的培养基上很难分离培养[56], 使得传统的鉴定方法受到极大的阻碍.

基于现有的测序技术, 主要是二代测序技术, 人们近年来发展了一些基于测序分析的噬菌体类型鉴定方法和工具. Emerson等人[57]发现了一些温和噬菌体的标记基因, 例如整合酶和切除酶基因. McNair等人[58]使用随机森林算法作为分类器, 开发了PHACTS工具, 利用噬菌体序列的蛋白信息对噬菌体类型进行鉴定. 若从噬菌体基因组中获得至少25个蛋白信息, PHACTS的预测效果较好. 但是, 当获得蛋白信息的个数少于5时, PHACTS的预测准确率仅有约65%, 效果很不理想. 此外, Ahmed等人[59]使用寡核苷酸相似性评分来衡量噬菌体与其宿主序列之间的相似性, 结果显示, 相比于烈性噬菌体, 温和噬菌体序列和宿主序列更相似. Deschavanne等人[60]通过分析所有有注释信息的噬菌体基因组, 发现可以利用四核苷酸频率作为特征来衡量噬菌体与其宿主之间的基因组距离, 进而预测噬菌体的类型. 尽管上述方法在某些条件下有较

好的预测效果，但由于宏基因组序列片段较短、拼接效果较差、与当前数据库同源性较低，上述大多方法并不适用于宏基因组以及宏病毒组数据的噬菌体类型鉴定问题. 发展不依赖于充分蛋白质信息且能对宏基因组数据中的短DNA片段直接做出判断的噬菌体类型鉴定工具至关重要.

针对以上问题，Song[61]基于马尔可夫模型开发了PhagePred，该工具以k-mer频率作为序列特征衡量待预测噬菌体序列和已知类型的噬菌体序列的距离，从而推断待预测噬菌体序列的类型. 但k-mer频率在提取短序列特征时存在噪声，对短序列特征提取的效果不如长序列. 本课题组[62]开发基于宏病毒组以及宏基因组数据的噬菌体类型鉴定算法DeePhage(https://github.com/shufangwu/DeePhage). DeePhage采用碱基独热矩阵编码短噬菌体序列，并采用卷积神经网络提取序列特征. 测试表明，DeePhage在测试集和真实数据集上的预测性能和计算时长上都优于PhagePred和PHACTS. 通过应用DeePhage对人肠道宏基因组和宏病毒组的数据进行整合分析，还能提出一种新的探究噬菌体生存模式转换的策略，有助于相关疾病发生的机理研究[62].

2.3　病毒的宿主预测

根据所侵染的宿主类型，肠道病毒常常需要分为原核病毒与真核病毒两类. 其中，真核病毒通常远远少于原核病毒即噬菌体，一般不超过病毒的10%. 尽管比例相差很大，但二者对人体的健康的影响都不可忽视. 对于病毒组的序列分析而言，从高通量测序出发，如何运用生物信息学手段确定病毒(通常是大量的短读序，它们也代表了相应的病毒全基因组)的宿主信息, 无疑具有十分重要的意义.

病毒的宿主范围取决于病毒与宿主细胞之间的分子相互作用，包括受体识别、对宿主细胞机制的适应和对宿主先天免疫识别的逃逸[63]. 其中，受体识别促进病毒附着到宿主细胞上，是感染宿主的首要步骤. 因此，病毒用来识别宿主受体的糖蛋白以及全基因组序列被广泛用于识别病毒的潜在宿主[64]. 为了鉴定新病毒的潜在宿主和致病性，传统的计算方法运用新病毒和其他病毒的基因组组成的相似性，以及宿主受体之间的相似性. 这两种策略都基于系统发育相关性能反映宿主关联性的假设，即系统发育关系近的病毒宿主具有一致性，以及具有相似受体蛋白的宿主可以被同一病毒感染. 尽管系统发育树上相邻病毒的宿主之间具有密切关联, Babayan等人[65]发现, 基于病毒系统发育相关性预测宿主关联的算法只能准确识别58.1%±0.07%(标准差)的病毒天然宿主. 而且, 公共数据库中有限的病毒(特别是真核病毒)参考基因组使基于相似性的方法对新病毒宿主的有效推断受到极大的限制.

为打破基于序列相似性方法的局限性，人们近年来发展了一些基于机器学习的算法和工具，如ViralHostPredictor[65], HostPhinder[66], WIsH[67], Host Taxon Predictor[68]和VIDHOP[69]. HostPhinder将待预测噬菌体的宿主物种指定为与其基因组相似度最高的参考噬菌体的宿主物种. 由于微生物群落中含有大量与已知噬菌体相似度较低的新噬菌体，这种依赖现有数据库的方法显然难以胜任新型噬菌体的宿主预测任务. WIsH使用马尔可夫链模型进行噬菌体的宿主预测, 并在短噬菌体片段上获得了较好的性能. 但是HostPhinder和WIsH仅用于预测噬菌体宿主，不适合非噬菌体病毒. ViralHostPredictor基于选定的进化基因组特征和系统发育信息预测感染人的RNA病毒的宿主和节肢动物载体, 该方法无法预测病毒是否能感染人类. 另外，由于依赖病毒和其天然宿主长期共进化后演化出的基因组特征，ViralHostPredictor缺乏预测病毒中间宿主的能力. 以上机器学习类的方法基于手工提取的特征，如密码子对、k-mer频率和密码子偏好性，往往忽略病毒基因组中的其他重要信息. VIDHOP是一种基于深度学习的工具，可在仅使用少量病毒基因组序列的情况下获得病毒的高精度分类，但该方法仅限于甲型流感病毒、狂犬病溶血酶病毒和轮状病毒A这三类病毒. 以上大部分工具通过使用病毒序列或基于病毒-宿主关系相关的病毒基因组特征进行宿主预测. 虽然这些工具在某些条件下表现良好，但它们普遍不适用于对宿主范围未知的情况.

本课题组[70]最近基于深度学习方法和马尔可夫链模型，发展了预测宏基因组或宏病毒组数据中噬菌体短序列宿主的算法HoPhage(http://cqb2.pku.edu.cn/ZhuLab/HoPhage/data/). 该算法由HoPhage-G和HoPhage-S两个模块组成. HoPhage-G模块在属水平上基于深度学习构建噬菌体序列片段和原核生物的配对关系，将宿主鉴定问题从复杂的多类预测问题转变为判

断配对的噬菌体序列和原核生物之间是否存在侵染关系的二分类任务; HoPhage-S模块在菌株水平对每个候选宿主基因组编码区上的密码子序列构建马尔可夫链模型, 然后计算待预测噬菌体片段编码区上的密码子序列在各个马尔可夫链模型上的似然度得分. 最终, 通过计算各个候选宿主属在两个模块的加权平均分数来整合HoPhage-G与HoPhage-S的结果. 测试结果表明, 针对宏基因组数据中的噬菌体短序列在属水平上的宿主预测问题, HoPhage在更广的候选宿主范围上表现更优的性能; 运用真实的肠道宏基因组数据测试, 表明HoPhage可在预测新噬菌体宿主方面发挥作用, 并帮助研究人员探索噬菌体在微生物群落中的潜在影响.

为应对预测包含真核病毒在内的新病毒(特别是与现有病毒相似度较低的病毒)的潜在宿主这一挑战, 本课题组[71]近年基于双通道卷积神经网络的深度学习方法, 发展了病毒宿主预测算法DeepHoF(https://github.com/PKUbioinfo-ZhuLab/DeepHoF). 该算法最终输出病毒以五大类宿主为宿主的可能性, 包括植物、微生物、无脊椎动物、非人脊椎动物(除人类以外的脊椎动物)和人类, 宿主范围涵盖所有活生物体. DeepHoF弥补了现有方法对新病毒的宿主预测的不足, 并显著优于基于BLAST的宿主预测方法, 分类效果的AUC值可高达0.975. 早在2020年1月, 本课题组[71]使用DeepHoF对2019年12月发布的最早的SARS-CoV-2(severe acute respiratory syndrome coronavirus 2)分离株进行了预测, 在疫情的最早阶段报告了宿主预测结果, 为新冠病毒的早期防控提供了重要信息. DeepHoF评估了SARS-CoV-2感染人类和非人类脊椎动物的可能性, 并用宿主得分谱对分离株进行了表征. 本课题组进一步对DeepHoF的预测结果进行深入分析, 以获取SARS-CoV-2的更详细的宿主信息. 最终预测水貂、蝙蝠、狗和猫可能是SARS-CoV-2的宿主, 其中, 水貂是最有可能、最值得注意的中间宿主. 这一预测得到了疫情期间全球范围SARS-CoV-2的动物感染流调数据及实验的有力支持. 对疫情中公共数据库中收录的大规模基因组的分析进一步验证了DeepHoF预测结果的长期有效性, 并揭示了人和水貂之间的SARS-CoV-2双向传播的关联. DeepHoF有能力为新病毒提供可靠的宿主信息, 从而有望缩短新病毒的发现与早期预防之间的时间差.

2.4 病毒的系统分类

病毒的系统分类也是病毒组研究关注的一个重要问题. 如前所述, 人肠道中的病毒仅有少部分为人们所认识和了解, 目前仍有大量病毒未被分类、鉴定出, 有人称之为病毒的暗物质(viral dark matter). 由于病毒缺少相应的标记基因, 当前病毒的系统分类主要依赖于与参考数据库进行序列比对. 正如前面所指出的, 大量病毒与已知数据库序列相似度较低, 依靠比对的算法实际上难以满足预测新病毒的需求.

近年来, 人们为此发展了一系列的基于机器学习的鉴定算法. Shang等人[72]开发的PhaGCN对噬菌体的重叠群进行科水平的分类. 该方法使用半监督学习的方式, 根据两个重叠群的碱基序列在蛋白水平的相似度来确定是否来源于同一类. 但PhaGCN在一些噬菌体的部分序列不足的分类中表现不佳, 而且作者没有评估模型在短序列(<1 kb)的表现. ViBE是最新发表的一个工具[73], 它基于预训练模型BERT对宏基因组数据进行分类, 可以重新训练对真核RNA病毒和DNA病毒进行目水平的分类, 但是在短序列水平表现不如Karken2.

病毒组的物种鉴定能够帮助人们缩小病毒组“暗物质”的范围, 对于理解病毒与宿主表型关系有十分重要的作用[74]. 需要指出的是, 目前专门用于病毒组进行病毒有效分类的工具还较少, 因此需要更多的努力来推动相关算法和工具的发展. 目前看来, 这一努力一方面可能应该建立在病毒与宿主表型相关蛋白质分子的序列或结构特征上, 另一方面有必要针对这些特征更加深入地运用诸如自然语言处理(natural language processing, NLP)等人工智能方法来发展相应的数据挖掘和新型算法.

3 展望

为了进一步挖掘肠道病毒组中的未知组分, 探究病毒如何调节肠道微生物群落动态变化以及这种变化与人类健康之间的关系, 本文认为今后需要结合多组学测序并求助于新兴技术. 其中, 转录组测序可以获取肠道中的RNA病毒, 蛋白质组测序可以提供病毒衣壳结构信息, 以三代测序为基础的单病毒测序技术(single-virus genomics)能够得到大量的较为完整的新

病毒基因组，为肠道病毒组学研究带来了新的机遇和挑战.

单病毒测序作为一种新兴技术，是对宏基因组学方法的一种补充. 与单细胞测序相似，单病毒测序的基本流程包括样本收集、荧光染色、单病毒分离、衣壳裂解、全基因组扩增及测序. 目前，单病毒测序的发展还处于初期阶段，面临着许多技术问题. 在单病毒分离阶段, 由于病毒的大小与许多污染物(如细胞囊泡)相近，流式细胞术分选病毒的效果较差，且目前没有专门针对不同类型病毒开发的荧光染料，这也限制了流式细胞术的表现. 在序列扩增阶段，缺少商业化、能灵敏扩增RNA病毒基因组的酶. 为了解决这些问题，人们正在研究能够代替流式细胞术进行单病毒分离的技术[75]. 尽管单病毒测序还有许多技术难题亟待解决，它在揭示病毒序列多样性方面有着不可替代的作用, 能够有力地促进病毒学的发展. 目前, 单病毒测序相关工作还较少，但已经在海洋与人体等环境中应用并发现了多种重要的新病毒[76]. 一些研究还将非基因组学技术与单病毒测序相结合，以探索病毒的结构以及病毒与其生境的互作. 比如使用生物正交非标准氨基酸标记(bioorthogonal noncanonical amino acid tagging, BONCAT)以寻找环境中新生成并释放的活跃病毒[77]，使用蛋白质组学、高分辨率成像、质谱和拉曼光谱等技术得到有关病毒结构、形态、化学成分和结构的信息. 此外，单细胞病毒RNA测序还能够获得病毒感染后细胞的即时反应信息，与病毒大小相似的胞内囊泡的研究也可以借助单病毒测序技术发展迎来新机遇.

同样值得期待的是，对肠道病毒组进行数据挖掘与分析，有助于人们更好地认识噬菌体，包括对其进行鉴定、注释以及应用. 2021年Camarillo-Guerrero等人[78]通过分析28060例人肠道宏基因组数据，获得了包含142809个肠道噬菌体基因组数据库GVD(human gut virome database), 大大拓展了人们对于肠道噬菌体的认知. 抗生素耐药性是21世纪最大的公共卫生安全威胁之一，而噬菌体对抗生素抗性基因在菌群中传播的作用已经得到了很多研究的证实. Shousha等人[79]发现，在243种大肠杆菌噬菌体中，24.7%的噬菌体能够将一种或多种抗生素基因转导至实验室大肠杆菌ATCC 13706菌株中. 在多种临床病原菌中也观察到噬菌体介导的抗生素抗性基因的获得[80,81]. 目前, 肠道病毒组中抗生素抗性基因的存在与功能尚不明确. Yan等人[82]在人肠道病毒组中鉴定出31种抗生素抗性基因. Wu等人[83]发现猪肠道病毒组波动很大，大多数病毒相关的抗生素抗性基因来自温和噬菌体. 显然, 肠道病毒组的定量分析将有助于推进噬菌体在抗生素抗性基因方面的研究，为揭示抗生素耐药性传播提供新策略与新启示. 此外，噬菌体可作为一种有前途的新型抗细菌制剂. 目前已有噬菌体疗法在临床上成功治疗铜绿假单胞菌、鲍曼不动杆菌、绿脓杆菌等菌引起的感染的报道[84~87]，但是相关临床试验还未取得显著成效[88]. 噬菌体作为抗细菌制剂的优势在于特异性强，非目标菌群受到的干扰小，噬菌体在自然界中资源丰富，且与细菌共进化[89]，可满足大部分临床需求分离到针对特定细菌的噬菌体. 但是相应地，这也是噬菌体疗法面临的一大挑战，即噬菌体的宿主范围需要十分明确. 本文前述的肠道病毒组相关的病毒系统分类以及宿主预测算法为明确噬菌体宿主范围提供了新策略. 除噬菌体治疗外, 还有研究者尝试使用粪便病毒组移植方法治疗代谢相关疾病，比如Ⅱ型糖尿病与肥胖[90], 但相关研究还处于初期阶段. 因此, 肠道病毒组的定量分析也可望为粪便病毒组移植技术的发展助一臂之力[91]. 综上所述，肠道病毒组的数据挖掘与分析技术的发展，尤其是对噬菌体在耐药、治疗方面的定量分析研究，可望为未来精准医疗提供有力的帮助和支持.

参考文献

1 Shkoporov A N, Hill C. Bacteriophages of the human gut: the “known unknown” of the microbiome. Cell Host Microbe, 2019, 25: 195–209

2 Jiang X, Li X, Yang L, et al. How microbes shape their communities? A microbial community model based on functional genes. Genomics Proteomics Bioinformatics, 2019, 17: 91–105

3 Garmaeva S, Sinha T, Kurilshikov A, et al. Studying the gut virome in the metagenomic era: challenges and perspectives. BMC Biol, 2019, 17: 84

4 Mushegian A R. Are there 10^{31} virus particles on earth, or more, or fewer? J Bacteriol, 2020, 202: e00052-20,
5 Breitbart M, Rohwer F. Here a virus, there a virus, everywhere the same virus? Trends Microbiol, 2005, 13: 278–284
6 Mokili J L, Rohwer F, Dutilh B E. Metagenomics and future perspectives in virus discovery. Curr Opin Virol, 2012, 2: 63–77
7 Manrique P, Bolduc B, Walk S T, et al. Healthy human gut phageome. Proc Natl Acad Sci USA, 2016, 113: 10400–10405
8 Reyes A, Haynes M, Hanson N, et al. Viruses in the faecal microbiota of monozygotic twins and their mothers. Nature, 2010, 466: 334–338
9 Norman J M, Handley S A, Baldridge M T, et al. Disease-specific alterations in the enteric virome in inflammatory bowel disease. Cell, 2015, 160: 447–460
10 Reyes A, Blanton L V, Cao S, et al. Gut DNA viromes of Malawian twins discordant for severe acute malnutrition. Proc Natl Acad Sci USA, 2015, 112: 11941–11946
11 Yang K, Niu J, Zuo T, et al. Alterations in the gut virome in obesity and type 2 diabetes mellitus. Gastroenterology, 2021, 161: 1257–1269.e13
12 Nurk S, Meleshko D, Korobeynikov A, et al. metaSPAdes: a new versatile metagenomic assembler. Genome Res, 2017, 27: 824–834
13 Sutton T D S, Clooney A G, Ryan F J, et al. Choice of assembly software has a critical impact on virome characterisation. Microbiome, 2019, 7: 12
14 Antipov D, Raiko M, Lapidus A, et al. Metaviral SPAdes: assembly of viruses from metagenomic data. Bioinformatics, 2020, 36: 4126–4129
15 Peng Y, Leung H C M, Yiu S M, et al. IDBA-UD: a *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics, 2012, 28: 1420–1428
16 Lai B, Wang F, Wang X, et al. InteMAP: integrated metagenomic assembly pipeline for NGS short reads. BMC Bioinformatics, 2015, 16: 244
17 Lai B, Ding R, Li Y, et al. A *de novo* metagenomic assembly program for shotgun DNA reads. Bioinformatics, 2012, 28: 1455–1462
18 Zhu W, Lomsadze A, Borodovsky M. *Ab initio* gene identification in metagenomic sequences. Nucleic Acids Res, 2010, 38: e132
19 Noguchi H, Taniguchi T, Itoh T. MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. DNA Res, 2008, 15: 387–396
20 Liu Y, Guo J, Hu G, et al. Gene prediction in metagenomic fragments based on the SVM algorithm. BMC Bioinformatics, 2013, 14: S12
21 Hu G Q, Guo J T, Liu Y C, et al. MetaTISA: metagenomic translation initiation site annotator for improving gene start prediction. Bioinformatics, 2009, 25: 1843–1845
22 Fang Z, Tan J, Wu S, et al. PlasGUN: gene prediction in plasmid metagenomic short reads using deep learning. Bioinformatics, 2020, 36: 3239–3241
23 Mount D W. Using the Basic Local Alignment Search Tool (BLAST). Cold Spring Harb Protoc, 2007, 2007: pdb.top17
24 Buchfink B, Xie C, Huson D H. Fast and sensitive protein alignment using DIAMOND. Nat Methods, 2015, 12: 59–60
25 Eddy S R. A new generation of homology search tools based on probabilistic inference. Genome Inform, 2009, 23: 205–211
26 Menzel P, Ng K L, Krogh A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nat Commun, 2016, 7: 11257
27 Wood D E, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol, 2019, 20: 257
28 Ounit R, Wanamaker S, Close T J, et al. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. BMC Genomics, 2015, 16: 236
29 Roux S, Tournayre J, Mahul A, et al. Metavir 2: new tools for viral metagenome comparison and assembled virome analysis. BMC Bioinformatics, 2014, 15: 76
30 Wommack K E, Bhavsar J, Polson S W, et al. VIROME: a standard operating procedure for analysis of viral metagenome sequences. Stand Genomic Sci, 2012, 6: 427–439
31 Huson D H, Mitra S, Ruscheweyh H J, et al. Integrative analysis of environmental sequences using MEGAN4. Genome Res, 2011, 21: 1552–1560
32 Kashtan N, Roggensack S E, Rodrigue S, et al. Single-cell genomics reveals hundreds of coexisting subpopulations in wild *Prochlorococcus*. Science, 2014, 344: 416–420
33 Sharon I, Kertesz M, Hug L A, et al. Accurate, multi-kb reads resolve complex populations and detect rare microorganisms. Genome Res, 2015, 25: 534–543
34 Wheeler D L, Barrett T, Benson D A, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res, 2007, 36: D13–D21
35 O'Leary N A, Wright M W, Brister J R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and

functional annotation. Nucleic Acids Res, 2016, 44: D733–D745
36 Tatusov R L. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res, 2000, 28: 33–36
37 Kanehisa M, Sato Y, Kawashima M, et al. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res, 2016, 44: D457–D462
38 Grazziotin A L, Koonin E V, Kristensen D M. Prokaryotic Virus Orthologous Groups (pVOGs): a resource for comparative genomics and protein family annotation. Nucleic Acids Res, 2017, 45: D491–D498
39 Leplae R, Lima-Mendez G, Toussaint A. ACLAME: a CLAssification of mobile genetic elements, update 2010. Nucleic Acids Res, 2010, 38: D57–D61
40 Li M, Wang C, Guo Q, et al. More positive or more negative? Metagenomic analysis reveals roles of virome in human disease-related gut microbiome. Front Cell Infect Microbiol, 2022, 12: 846063
41 Lima-Mendez G, Van Helden J, Toussaint A, et al. Prophinder: a computational tool for prophage prediction in prokaryotic genomes. Bioinformatics, 2008, 24: 863–865
42 Fouts D E. Phage_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences. Nucleic Acids Res, 2006, 34: 5839–5851
43 Akhter S, Aziz R K, Edwards R A. PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. Nucleic Acids Res, 2012, 40: e126
44 Arndt D, Grant J R, Marcu A, et al. PHASTER: a better, faster version of the PHAST phage search tool. Nucleic Acids Res, 2016, 44: W16–W21
45 Zhou Y, Liang Y, Lynch K H, et al. PHAST: a fast phage search tool. Nucleic Acids Res, 2011, 39: W347–W352
46 Reis-Cunha J L, Bartholomeu D C, Manson A L, et al. ProphET, prophage estimation tool: a stand-alone prophage sequence prediction tool with self-updating reference database. PLoS One, 2019, 14: e0223364
47 Ren J, Ahlgren N A, Lu Y Y, et al. VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. Microbiome, 2017, 5: 69
48 Ren J, Song K, Deng C, et al. Identifying viruses from metagenomic data using deep learning. Quant Biol, 2020, 8: 64–77
49 Amgarten D, Braga L P P, da Silva A M, et al. MARVEL, a tool for prediction of bacteriophage sequences in metagenomic bins. Front Genet, 2018, 9: 304
50 Fang Z, Tan J, Wu S, et al. PPR-Meta: a tool for identifying phages and plasmids from metagenomic fragments using deep learning. Gigascience, 2019, 8: giz066
51 Song K. Reads binning improves the assembly of viral genome sequences from metagenomic samples. Front Microbiol, 2021, 12: 1266
52 Roux S, Enault F, Hurwitz B L, et al. VirSorter: mining viral signal from microbial genomic data. PeerJ, 2015, 3: e985
53 Guo J, Bolduc B, Zayed A A, et al. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. Microbiome, 2021, 9: 37
54 Rozov R, Brown Kav A, Bogumil D, et al. Recycler: an algorithm for detecting plasmids from *de novo* assembly graphs. Bioinformatics, 2017, 33: 475–482
55 Erez Z, Steinberger-Levy I, Shamir M, et al. Communication between viruses guides lysis-lysogeny decisions. Nature, 2017, 541: 488–493
56 Riley P A. Bacteriophages in autoimmune disease and other inflammatory conditions. Med Hypotheses, 2004, 62: 493–498
57 Emerson J B, Thomas B C, Andrade K, et al. Dynamic viral populations in hypersaline systems as revealed by metagenomic assembly. Appl Environ Microbiol, 2012, 78: 6309–6320
58 McNair K, Bailey B A, Edwards R A. PHACTS, a computational approach to classifying the lifestyle of phages. Bioinformatics, 2012, 28: 614–618
59 Ahmed S, Saito A, Suzuki M, et al. Host-parasite relations of bacteria and phages can be unveiled by *Oligostickiness*, a measure of relaxed sequence similarity. Bioinformatics, 2009, 25: 563–570
60 Deschavanne P, DuBow M S, Regeard C. The use of genomic signature distance between bacteriophages and their hosts displays evolutionary relationships and phage growth cycle determination. Virol J, 2010, 7: 163
61 Song K. Classifying the lifestyle of metagenomically-derived phages sequences using alignment-free methods. Front Microbiol, 2020, 11: 2865
62 Wu S, Fang Z, Tan J, et al. DeePhage: distinguishing virulent and temperate phage-derived sequences in metavirome data with a deep learning approach. Gigascience, 2021, 10: giab056

63 Rothenburg S, Brennan G. Species-specific host-virus interactions: implications for viral host range and virulence. Trends Microbiol, 2020, 28: 46–56
64 Zhou P, Yang X L, Wang X G, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature, 2020, 579: 270–273
65 Babayan S A, Orton R J, Streicker D G. Predicting reservoir hosts and arthropod vectors from evolutionary signatures in RNA virus genomes. Science, 2018, 362: 577–580
66 Villarroel J, Kleinheinz K, Jurtz V, et al. HostPhinder: a phage host prediction tool. Viruses, 2016, 8: 116
67 Galiez C, Siebert M, Enault F, et al. WIsH: who is the host? Predicting prokaryotic hosts from metagenomic phage contigs. Bioinformatics, 2017, 33: 3113–3114
68 Gałan W, Bąk M, Jakubowska M. Host taxon predictor—a tool for predicting taxon of the host of a newly discovered virus. Sci Rep, 2019, 9: 3436
69 Mock F, Viehweger A, Barth E, et al. VIDHOP, viral host prediction with deep learning. Bioinformatics, 2021, 37: 318–325
70 Tan J, Fang Z, Wu S, et al. HoPhage: an *ab initio* tool for identifying hosts of phage fragments from metaviromes. Bioinformatics, 2021, 38: 543–545
71 Guo Q, Li M, Wang C, et al. Predicting hosts based on early SARS-CoV-2 samples and analyzing the 2020 pandemic. Sci Rep, 2021, 11: 17422
72 Shang J, Jiang J, Sun Y. Bacteriophage classification for assembled contigs using graph convolutional network. Bioinformatics, 2021, 37: i25–i33
73 Gwak H J, Rho M. ViBE: a hierarchical BERT model to identify eukaryotic viruses using metagenome sequencing data. Brief Bioinform, 2022, 23: bbac204
74 Virgin H W. The virome in mammalian physiology and disease. Cell, 2014, 157: 142–150
75 Schmidt H, Hawkins A R. Single-virus analysis through chip-based optical detection. Bioanalysis, 2016, 8: 867–870
76 Garcia-Heredia I, Bhattacharjee A S, Fornas O, et al. Benchmarking of single-virus genomics: a new tool for uncovering the virosphere. Environ Microbiol, 2021, 23: 1584–1593
77 Pasulka A L, Thamatrakoln K, Kopf S H, et al. Interrogating marine virus-host interactions and elemental transfer with BONCAT and nanoSIMS-based methods. Environ Microbiol, 2018, 20: 671–692
78 Camarillo-Guerrero L F, Almeida A, Rangel-Pineros G, et al. Massive expansion of human gut bacteriophage diversity. Cell, 2021, 184: 1098–1109.e9
79 Shousha A, Awaiwanont N, Sofka D, et al. Bacteriophages isolated from chicken meat and the horizontal transfer of antimicrobial resistance genes. Appl Environ Microbiol, 2015, 81: 4600–4606
80 Varga M, Kuntová L, Pantůček R, et al. Efficient transfer of antibiotic resistance plasmids by transduction within methicillin-resistant Staphylococcus aureus USA300 clone. FEMS Microbiol Lett, 2012, 332: 146–152
81 Goh S, Hussain H, Chang B J, et al. Phage ϕC2 mediates transduction of Tn6215, encoding erythromycin resistance, between *Clostridium difficile* strains. mBio, 2013, 4: e00840-13
82 Yan Q, Wang Y, Chen X, et al. Characterization of the gut DNA and RNA viromes in a cohort of Chinese residents and visiting Pakistanis. Virus Evol, 2021, 7: veab022
83 Wu R, Cao Z, Jiang Y, et al. Early life dynamics of ARG and MGE associated with intestinal virome in neonatal piglets. Vet Microbiol, 2022, 274: 109575
84 Chan B K, Turner P E, Kim S, et al. Phage treatment of an aortic graft infected with *Pseudomonas aeruginosa*. Evol Med Public Health, 2018, 2018(1): 60–66
85 Duplessis C, Biswas B, Hanisch B, et al. Refractory pseudomonas bacteremia in a 2-year-old sterilized by bacteriophage therapy. J Pediatr Infect Dis Soc, 2018, 7: 253–256
86 Khawaldeh A, Morales S, Dillon B, et al. Bacteriophage therapy for refractory *Pseudomonas aeruginosa* urinary tract infection. J Med Microbiol, 2011, 60: 1697–1700
87 Schooley R T, Biswas B, Gill J J, et al. Development and use of personalized bacteriophage-based therapeutic cocktails to treat a patient with a disseminated resistant *Acinetobacter baumannii* infection. Antimicrob Agents Chemother, 2017, 61: e00954–17
88 Kortright K E, Chan B K, Koff J L, et al. Phage therapy: a renewed approach to combat antibiotic-resistant bacteria. Cell Host Microbe, 2019, 25: 219–232
89 Simmonds P, Aiewsakun P, Katzourakis A. Prisoners of war—host adaptation and its constraints on virus evolution. Nat Rev Microbiol, 2019,

17: 321–328

90 Rasmussen T S, Mentzel C M J, Kot W, et al. Faecal virome transplantation decreases symptoms of type 2 diabetes and obesity in a murine model. Gut, 2020, 69: 2122–2130

91 Rasmussen T S, Koefoed A K, Jakobsen R R, et al. Bacteriophage-mediated manipulation of the gut microbiome—promises and presents limitations. FEMS Microbiol Rev, 2020, 44: 507–521

Data mining and analysis techniques for gut virome: the prospects and challenges

JIANG XiaoQing, LI Mo, YIN HengChuang, GUO Qian, TAN Jie, WU ShuFang, WANG ChunHui & ZHU HuaiQiu*

Department of Biomedical Engineering, College of Future Technology, and Center for Quantitative Biology Peking University, Beijing 100871, China

The gut virome plays a very important role in the microbial community structure, the bacterial traits, and even the human health. However, it is still poorly understood compared to the bacterial metagenome among the gut microbiome. The rapid development of high-throughput sequencing technologies, machine learning, deep learning, and other methods provides an opportunity for in-depth study of gut virome. With a special focus on the high-throughput genomic data of bacteriophages and eukaryotic viruses, this paper reviewed those general character key technologies of data mining and data analysis in current gut virome research, most of which are applied in both viral metagenomes and metaviromes. In view of the complexity of biological problems and clinical trials, as well as the application of third-generation sequencing and artificial intelligence methods, we also discussed the challenges and opportunities for these tools and techniques in gut virome.

gut virome, metagenome, omics analysis, data mining

doi: 10.1360/SSV-2022-0330