噬菌体分类学技术进展

黄士轩[1,2]，朱斌[1,2]，何嘉欣[1,2]，李滢[2]，薛亮[2]，周星佑[1]，伍君权[1]，陈谋通[1,2]，
张菊梅[1,2]，吴清平[2]，杨美艳[1,2*]

（1.华南农业大学农学院&食品学院，广东省食品质量安全重点实验室，广东广州 510642）（2.广东省科学院微生物研究所，华南应用微生物国家重点实验室，广东省微生物安全与健康重点实验室，国家卫健委微生物食品营养与安全科技创新平台，广东广州 510070）

摘要：噬菌体是一类细菌病毒，能特异性识别并杀灭细菌，具有替代抗生素的潜力，特别是在食品安全领域中，以解决“人－环－食”生态链中泛耐药病原微生物的威胁。噬菌体分类学是阐明噬菌体间物种进化和发展规律的重要学科。近年来，随着新一代测序、培养组学等技术发展，噬菌体分类出现了许多新的技术进展与挑战。该文通过回顾噬菌体分类学的发展历程，阐明目前常用分类技术的优缺点；同时，该文系统总结了噬菌体进化分类的具体方案，重点探讨了分类学标尺及系统发育构建的前沿进展；最后，该文归纳了目前噬菌体分类技术的新进展，重点推荐了纯培养噬菌体分类的可靠流程，并讨论了培养组学、宏病毒组及原噬菌体分析手段对噬菌体分类学技术的挑战与要求。该文为噬菌体分类学技术的进一步发展提供参考。

关键词：噬菌体；分类学；进化分类法；系统发育；宏病毒组

文章编号：1673-9078(2024)09-346-358 DOI: 10.13982/j.mfst.1673-9078.2024.9.1146

Advances in Bacteriophage Taxonomy Techniques

HUANG Shixuan[1,2], ZHU Bin[1,2], HE Jiaxin[1,2], LI Ying[2], XUE Liang[2], ZHOU Xingyou[1], WU Junquan[1], CHEN Moutong[1,2], ZHANG Jumei[1,2], WU Qingping[2], YANG Meiyan[1,2*]

(1.College of Agriculture, College of Food Science, South China Agricultural University, Guangdong Key Laboratory of Food Quality and Safety, Guangzhou 510642, China) (2.National Health Commission Science and Technology Innovation Platform for Nutrition and Safety of Microbial Food, Guangdong Provincial Key Laboratory of Microbial Safety and Health of Guangdong Academy of Sciences, State Key Laboratory of Applied Microbiology Southern China, Institute of Microbiology, Guangzhou 510070, China)

Abstract: Bacteriophages are a class of viruses that parasitize bacteria, specifically recognize, and kill them. Bacteriophages have the potential to replace antibiotics, especially in the field of food safety, addressing the threat of pan-drug-resistant pathogenic microorganisms in the era of "human-environment-food" ecological chain. Bacteriophage taxonomy is an important discipline to elucidate the evolution and development patterns among the species. In recent years, with the advancements of next-generation sequencing, culturomics and other technologies, many new technological advances

引文格式：

黄士轩,朱斌,何嘉欣,等.噬菌体分类学技术进展[J].现代食品科技,2024,40(9):346-358.

HUANG Shixuan, ZHU Bin, HE Jiaxin, et al. Advances in bacteriophage taxonomy techniques [J]. Modern Food Science and Technology, 2024, 40(9): 346-358.

收稿日期：2023-09-21

基金项目：国家自然科学基金项目（32202194）；国家重点研发计划项目（2022YFF1100700）；广州市重点研发计划项目（SL2022B03J01243）；广东省重点研发计划项目（2022B1111040002）；广东省科学院项目（2022GDASZH-2022020402-01）

作者简介：黄士轩（1999-），男，硕士研究生，研究方向：微生物生物信息学，E-mail：nhuhuganaxis@outlook.com

通讯作者：杨美艳（1984-），女，博士，讲师，研究方向：食品微生物安全，E-mail：ymy@scau.edu.cn

and challenges have emerged in bacteriophage taxonomy. By reviewing the development of bacteriophage taxonomy, this study identified the advantages and disadvantages of commonly used taxonomy techniques. Simultaneously, the paper systematically summarizes the specific scheme of bacteriophage evolutionary taxonomy, focusing on the frontier progress of taxonomic scale and phylogenetic construction. Finally, this study summarizes the recent progress of bacteriophage taxonomic technology, specifically recommending a reliable process of pure culture bacteriophage taxonomy, and discussing the challenges and requirements of culture omics, meta-virome and bacteriophage analysis techniques of bacteriophage taxonomy. This paper provides a reference for the further progress of phage taxonomic techniques.

Key words: bacteriophage; taxonomy; molecular evolutionary classification; phylogenetic; meta-virome

噬菌体（Bacteriophages, Phages）是一类寄生于细菌或古菌的病毒，由于部分噬菌体能引起宿主菌的裂解，故而得名。噬菌体是地球上生物数量最大的生物，约有 10^{31} 个[1]。1986 年，研究者们发现具裂解宿主作用的噬菌体可特异性识别并杀灭细菌，可作为替代抗生素的新型生物制剂[2]。随着抗生素的滥用，细菌在药物压力下不断进化并逐渐形成耐受，继而获得生存优势、在全球范围内流行，严重威胁食品安全以及人们生活健康[3]。区别于结构固定、不可进化的抗生素，噬菌体与其宿主之间存在“共同进化”，对耐药突变株仍可高效裂解[4]。因此，在“人－环－食”生态链均进入细菌泛耐药的年代，噬菌体成为了最有希望能防控致病菌的新型抗生素替代物，为食品安全提供了新思路。如今，对噬菌体认识的不足严重限制了其开发应用。噬菌体分类学（Phage Taxonomy）是研究噬菌体生物学分类的科学，以阐明噬菌体间的亲缘关系、基因遗传、物种进化过程和发展规律为目标[5]，从而认识和了解噬菌体。开展系统的噬菌体生物学分类研究有助于了解噬菌体的遗传与进化过程，为后续的噬菌体应用提供理论依据。近年来，随着新一代测序、培养组学等技术发展，噬菌体分类出现了许多新的技术进展与挑战。如今，噬菌体分类学以进化分类为主，其他分类法为辅。进化分类学离不开系统发育分析，随着科技的进步，众多系统发育分析软件的出现降低了分析的难度，但是当前的噬菌体数据库提供的参考较少。为解决数据库参考较少的瓶颈，利用当前的培养组学、宏病毒组及原噬菌体分析新技术完成对噬菌体数据库的扩充是合理的方法。本文从噬菌体分类学的历史与现状、当前分类方法以及新技术研究进展进行综述，系统总结了噬菌体分类学的常用技术及潜在发展方向，以期提供噬菌体分类学的详尽介绍及当前推荐的常用分类方法。

1 噬菌体分类学方法

噬菌体的分类法主要有三种：巴尔的摩分类法（Baltimore Classification）、形态学分类法（Morphology Classification）与进化分类法（Evolutionary Classification）（图 1）。国际病毒分类委员会（International Committee on Taxonomy of Viruses, ICTV）成立于 1966 年，是一个致力于对病毒进行生物学分类并制定相关标准的组织，其制定的分类标准给病毒研究者们提供了重要参考，现今使用最广泛的噬菌体分类依据均源于 ICTV 的报告[6]。以进化分类法为主的 ICTV 分类法是当前的主流方法。另外，也有研究者根据其他方法对噬菌体进行分类，如依据噬菌体的宿主、噬菌体的应用场景或噬菌体存活环境等对其分类的方法，但这些分类法与 ICTV 分类法有所冲突，较少被系统地应用于学术研究中。如今，单纯依据巴尔的摩分类法和形态学分类法已成为噬菌体分类学的历史。

巴尔的摩病毒分类法由 1971 年诺贝尔生理学或医学奖得主戴维·巴尔的摩（David Baltimore）提出[7]。依据中心法则，病毒可以根据基因组产生 mRNA 的方式分为七大类（双链 DNA 病毒、有缺口的双链 DNA 病毒、单链 DNA 病毒、双链 RNA 病毒、单链 RNA(+) 病毒、单链 RNA(-) 病毒、逆转录病毒），如双链 DNA 噬菌体 T4、单链 RNA(+) 噬菌体 MS2 等[8]。然而随着噬菌体研究的深入发展，学者发现大多数噬菌体为双链 DNA 病毒，少数为单链 DNA 病毒，极少数是 RNA 病毒，几乎未发现逆转录病毒[8]。巴尔的摩分类法虽然对于所有病毒的初步分类非常有效，但由于其无法良好地展现噬菌体的多样性，因此，沿用了将近 50 年的巴尔的摩分类法没有进一步分类噬菌体。

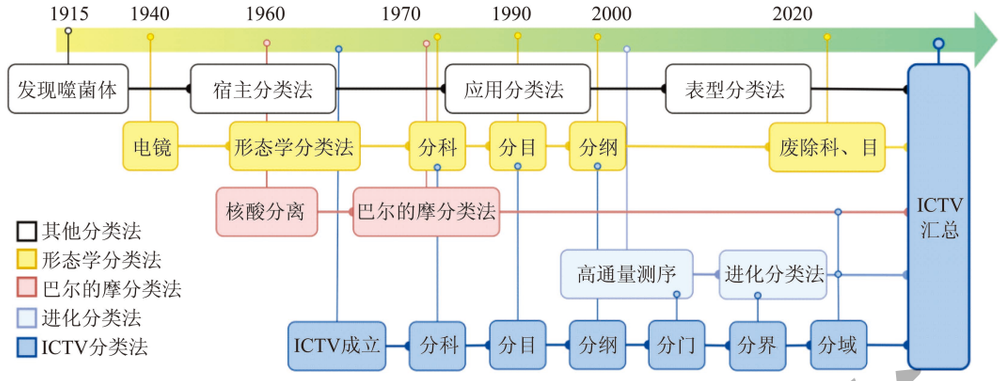

图 1 噬菌体分类学历史进程图

Fig.1 Historical progression of phage taxonomy

1940~2022 年通用的形态学分类法是采用形态特征对噬菌体进行分类的方法，也是最直观的分类方法[9]。随着透射电子显微镜（Transmission Electron Microscope, TEM）的普遍使用，各类噬菌体的形态被逐一捕捉，为形态学分类提供了参考基准。在形态学分类中，噬菌体依据其是否拥有尾部可初步分类为有尾噬菌体与无尾噬菌体；而有尾噬菌体又再依据其尾部的长短以及是否携带尾鞘进一步分类为长尾噬菌体（*Siphoviridae*）、短尾噬菌体（*Podoviridae*）和肌尾噬菌体（*Myoviridae*）；无尾噬菌体则可根据其头部特征分为复层噬菌体（*Tectiviridae*）、覆盖噬菌体（*Corticoviridae*）、囊状噬菌体（*Cystoviridae*）、光滑噬菌体（*Leviviridae*）和丝状噬菌体（*Inoviridae*）等[9]。随着被观察到的噬菌体逐渐增多，研究者发现同类噬菌体中不同个体间形态差异巨大，且形态的差异难以关联其进化关系及功能表征[1]。同时，在高通量测序技术的快速发展下，对于不可培养的噬菌体来说，由于无法捕捉病毒颗粒的形态特征，亦无法对其进行分类[10]。因此，形态学分类法虽然较巴尔的摩法进一步加深了人们对噬菌体的认识，但仅基于形态学特征难以全面开展噬菌体的细分分类与进化分析。

自 2005 年第二代测序技术（高通量测序技术）出现，人们对微生物的研究逐渐从实体逐步深入至基因组阶段[11]。噬菌体全基因组挖掘技术的创新促使噬菌体分类学方法从表型迈向基因组层面，其中，噬菌体核苷酸序列、氨基酸序列以及蛋白结构等的差异等均可以作为噬菌体的分类的依据。进化分类学是依据核苷酸突变速率计算物种之间进化距离的分类方法，其良好地体现了物种基因组的进化情况并能够依据基因组相似度划分类群[12]；对于相对远源的物种，则可采用氨基酸序列聚类方法代替核苷酸进化模型辅助分类[13]；对于遗传距离特别远的大分子 DNA 序列，则使用同源的衣壳蛋白结构差异进行分类[14]。因此，进化分类学能很好地在生物学分类的各个层级中完成分类工作。至今，噬菌体基因组测序数据增长速度远超表型验证实验的进展，基于核苷酸序列的进化分类学逐渐占据主导地位。

在 2021 年 ICTV 的第九次大会中，专家们提议放弃传统形态学分类法中分类目与科的方法，即删除长尾病毒科（Siphoviridae）、短尾病毒科（Podoviridae）、肌尾病毒科（Myoviridae）以及有尾病毒目（Caudovirales）的分类，以有尾病毒为名的分类中仅保留了有尾噬菌体纲（Caudoviricetes）的描述[15]。至 2022 年 10 月，ICTV 最终决议各种噬菌体形态仅用作噬菌体描述，而不直接作为分类依据[16]，标志着沿用了将近 80 年的形态学分类成为历史，噬菌体分类进入以进化分类为基础、其他分类法为辅的新时代。目前，基于基因组测序分析的 ICTV 分类法已成为应用范围最广的噬菌体分类方法，本文将重点分析。

由于进化分类学方法应用广泛，因此 ICTV 对该方法的实施提出了系列规范。ICTV 在 2023 年 2 月的报告中指出，噬菌体在 15 个分类等级中应采用四种不同的进化分类方法，而不同方法的适用范围分别为：从域到纲水平使用模体（Motif）或折叠区分析，从亚门到科水平使用核心基因分析，从目到属水平使用氨基酸序列相似度分析，从亚科到种水平使用核酸序列相似度分析，而且这次报告提出了四大分类原则[17]：（1）病毒分类法应反映病毒的进化历史及病毒的特性；（2）病毒分类应该可以指

导等级的分配，使其效用最大化；(3) 进化分类法只是对病毒进行分类的多种可能手段之一；(4) 从宏基因组序列推断得出的病毒分类学分析需执行严格的序列质量控制。这一报告为目前噬菌体进化分类学研究中存在的问题提出了系统可行的解决方案以及研究规范。

对比噬菌体序列在各水平未获得分类地位的比例与 ICTV 提出的分类规范，可知噬菌体在目、科水平仍存在很大的认识不足，而这些不足也是目前分类学研究者们最关注的研究领域。虽然根据 ICTV 规范，噬菌体在目与科水平的分类应采用基因组成分析、同源基因相似度分析及氨基酸序列相似度分析的方法开展，但考虑到分类应兼顾对物种进化的描述，因此围绕这两个水平现存的分类问题，相关研究均以系统发育分析为主、聚类分析为辅的方法以明确噬菌体更细致的分类学地位。

1.1 噬菌体分类学现状

当前，传统的巴尔的摩法及形态学分类法均出现了诸多不足，其中最大的局限是必须获得噬菌体纯培养物。而得益于基因组挖掘手段的创新，近年来新分离的可培养噬菌体十分有限，但基因组数量却迅猛上涨[18]。截至 2023 年 3 月，虽然可搜索到的纯培养噬菌体及其全基因组测序的相关论文仅 3 125 篇，然而美国国家生物技术信息中心（National Center for Biotechnology Information, NCBI）数据库中却有高达 19 625 条噬菌体全基因组记录，是分离获得噬菌体的 6.28 倍。

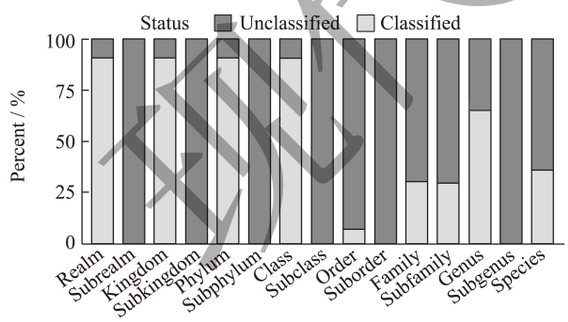

图 2 噬菌体各分类等级的占比图

Fig.2 The proportion of bacteriophages in each taxonomic grade

注：深灰色为暂时无分类，浅灰色为已分类。总计纳入 20 592 条噬菌体序列信息，信息来源均来自 NCBITaxonomy 数据库。

噬菌体是病毒的一种，进化分类法再次证实，噬菌体在病毒中数量最多，同时分类也较多[1]。截至 2023 年 3 月，ICTV 将已知的噬菌体分类为 5 域、7 界、7 门、9 纲、13 目、56 科、1 623 属、5 204 种，结果同步在 NCBI Taxonomy 数据库中[19]。在病毒的六域分类中，噬菌体暂时缺乏 *Ribozyviria* 域的相关分类信息，而且 15 个分类等级中，许多噬菌体在域、界、门、纲下暂无目、科、属级别以及各种亚级别分类[20]（图 1）。这些分类地位的缺失将在后续的分类学研究中不断被完善补充。

1.2 系统发育分析分类噬菌体

系统发育分析可以反映物种的进化历史，是进化分类法的重要分析方法，而依据进化历程获得的物种分类关系是最准确可信的系统发育分析方法[21]。在噬菌体的分类研究中，研究者常使用噬菌体的某些核心蛋白、核心功能域或者核心序列作为进化标尺开展系统发育分析。通过核苷酸替代模型计算获得的系统发育树能较好反映噬菌体某个基因的进化历程，同时也能解释噬菌体的系统发育历程[22]。开展噬菌体系统发育分析，一般按照 4 大流程进行：多序列比对（Multiple Sequence Alignment, MSA）、进化模型选择、系统发育树构建及数据可视化，这 4 大流程的常用工具及应用特征如表 1 所示。

一个可靠的系统发育分析流程是：通过大数据库（如 NR）比对使用的进化标尺，筛选合理的结果，获取距离合适的参考序列使用 MAFFT 进行 MSA，MSA 结果通过 IQ-tree 构建系统发育树，最后，使用 iTOL 对系统发育树的可视化。

目前，常用于遗传分析的进化模型主要有最大似然法（Maximum Likelihood, ML）和邻接法（Neighbor-Joining, NJ）。对于遗传距离较近的物种，一般使用 ML 的计算模型进行系统发育树的构建，而对于遗传距离较远的情况，可采用 NJ 法构建系统发育树[23]。

对于噬菌体的系统发育树的构建，较多研究使用噬菌体的标志基因（*hallmark*）作为进化标尺进行分析[24-26]。常见的噬菌体分类标尺基因包括末端酶大亚基（用于距离较近的 DNA 噬菌体）、衣壳蛋白基因（用于距离较远的 DNA 噬菌体）、门蛋白（用于有尾噬菌体）以及 RdRp 功能域（用于 RNA 噬菌体）。这些标尺为噬菌体在不同分类水平提供

了重要参考，但目前仅采用单一标尺难以在噬菌体的目水平进行有效的分类。针对这一局限，近年来有学者提出，采用基于单核苷酸多态性（Single Nucleotide Polymerase, SNP）的 ML 法构建的系统发育树[27]。SNP 是基因进化的基本单元，其记录了基因组的进化历史，因此基于全基因组 SNP 估计的进化距离较单一基因序列估计的进化距离构建进化树更为合理[28]。但也有研究指出，基于 SNP 计算距离的应用范围较为狭窄，因此目前只推荐该方法用于噬菌体种内的进化、分型分析[29]。目前，开展 SNP 进化分析的工具主要有两类，一类是基于基因组序列获取 SNP，常用软件为 Harvest 的 ParSNP[30]，另外一类是基于全基因组变异检测获取 SNP，常用的软件是 VCF2Dis（https://github.com/hewm2008/VCF2Dis）或 Phylip 工具包（https://evolution.genetics.washington.edu/phylip. html）。

1.3 基于新标尺的系统发育分析

由于可作用标识的噬菌体标志基因突变有限，因此难以采用单一的标尺对所有噬菌体进行分类分析。为解决这一局限，学者们提出对于不同层级的噬菌体分类，应采用不同的分子进化标尺开展系统发育分析[17]。研究发现，不同噬菌体中的 DNA 复制相关基因及结构相关基因具有很强的保守性以及高变异率[25]，因此这些基因常被用作为噬菌体的分类标尺。

表 1 噬菌体ML法系统发育分析流程及应用软件

Table 1 Phage ML phylogenetic analysis process and application software

分类流程	软件名称	客户端	进化模型	年份/年
多序列比对	Clustal W[31]	本地	/	1994
	T-Coffee[32]	本地	/	2000
	MAFFT[33]	本地	/	2002
	Muscle[34]	本地	/	2004
	Kalign[35]	网页、本地	/	2005
	Clustal Omega[36]	网页、本地	/	2011
模型选择	ModelTest[37]	本地	核酸替换模型	1998
	ProtTest[38]	本地	蛋白替换模型	2005
	jModelTest[39]	本地，Java	核酸替换模型	2008
	ModelTest-NG[40]	本地	jModelTest + ProtTest	2008
	PALM[41]	本地	序列替换模型	2009
	ProtTest3[42]	本地	蛋白替换模型	2011
	ModelFinder[43]	本地	序列替换模型	2017
	SMS[44]	本地	序列替换模型	2017
系统发育树构建	PhyML[45]	网页	手动选择模型	2005
	PAML4[46]	本地	手动选择模型	2007
	RAxML8[47]	本地	自动选择模型	2014
	FastTree2[48]	本地	GTR+ Γ_4	2010
	IQ-tree[49]	本地	ModelFinder 选择模型	2015
	IQ-tree2[50]	本地	ModelFinder 选择模型	2020
数据可视化	Unipro UGENE[51]	本地	/	2012
	Bio.Phylo[52]	本地，Python	/	2012
	ggtree[53]	本地，R	/	2018
	MEGA X[54]	本地	/	2018
	Evolview v3[55]	网页	/	2019
	iTOL5[56]	网页	/	2021

最近研究指出，相较使用完整的基因作为标尺，使用同源基因的功能结构域作为分子进化标尺将有望构建更为精准的系统发育树[57]。一个经典的例子是 *Crucivirus* 属的噬菌体可以采用最佳的复制相关基因（DNA 聚合酶）进行系统发育树构建。*Crucivirus* 属 DNA 聚合酶中部的 S-domain 是其高变功能域，使用这个功能域作为标尺的发育树较采用整个 DNA 聚合酶基因来说更为准确[58]。因此，相比于目前常用的基于基因的标尺，采用基因的功能域标尺能更精准反映噬菌体特定功能的进化关系。

随着分离技术的发展，研究者们将对噬菌体基因的功能有更为深入的理解。在此基础上，基于大规模噬菌体的泛基因组或同源基因组分析将能协助研究者们批量筛选获得更多的分子进化标尺，为系统发育分析提供更多可能。而对于出现多个分子进化标尺难以择优的情况，可以参考细菌基因组的 cgMLST 分型方式，通过使用多个管家基因序列串联形成新的分子进化标尺，进而开展分类[59]。

1.4 聚类分析辅助噬菌体分类

聚类分析是指将基因的集合分组为类似对象组成的多个类的分析方法，该方法常用于噬菌体的科、亚科、属水平分类[17]。目前，噬菌体基础聚类主要采用全蛋白质组的隐性马尔可夫模型（Hidden Markov Model, HMM）开展，这一方法常通过 vConTACT[60] 软件开展分析。而对于系统发育分析难以解决的分类，ICTV 建议开展基因序列相似性聚类进行分析，可用的工具包括 DEmARC[61]、VIRIDIC[62] 等。然而，通过聚类分析获得的遗传距离只是真实进化关联性的近似值，因此聚类分析仅可作为系统发育分析方法的补充，而不应单独依靠聚类分析判断噬菌体的进化分类。

1.5 深度学习辅助噬菌体分类

机器学习（Machine Learning）是一种依据一定的策略对已知关联的模型进行训练，最后运用训练好的模型对未知数据进行分析预测的方法。对于一些结构繁复的分类标尺，目前采用传统算法难以完成准确分类的噬菌体系统发育分析，而机器学习则为此类分类学局限提供新的解决思路。利用机器学习的噬菌体分类模型，目前较为流行的有蛋白空间结构模型、核苷酸比对模型、氨基酸比对模型和单核苷酸多态性（Single-nucleotide Polymorphism, SNP）进化模型等。

主要衣壳蛋白（Major Capsid Protein, MCP）是噬菌体重要的基因组复制元件，MCP 作为分类标尺可在域、纲水平对噬菌体进行精准的分类学分析。然而，MCP 结构复杂，除了其一级结构外，该蛋白的三维结构也可以作为噬菌体的分类依据。研究指出，噬菌体 MCP 主要有三种折叠方式（单果冻卷、双果冻卷及 HK97 折叠）[63]，而这三种折叠方式是噬菌体在域至纲级别分类的主要方法。对比传统应用冷冻电镜解析 MCP 三维结构的方法，目前基于机器学习的 AlphaFold[64] 可以通过计算氨基酸分子间的相互作用力获得其排布规律，无需进行实验则可直接计算获得 MCP 的三维结构，进而开展分类分析。

图卷积神经网络（Graph Convolutional Networks, GCN）模型可以用来解释聚类分类结果。近年来，利用 GCN 开展的聚类分类研究最具代表性的是 PhaGCN[65]，其原理是基于氨基酸序列相似度聚类分析的深度学习构建 GCN 模型，继而对图像聚类结果的解释，即通过蛋白相似度较高的噬菌体分类来判断未知的噬菌体分类。一般来说，GCN 的分析结果基本为科、属级别的分类。

基于 SNP 构建的模型可以准确阐述噬菌体种内进化关系，常用于溯源分析中。目前最常用于构建 SNP 模型的软件包括 BEAST[66]、Phycas[67] 等，其中 BEAST 通过分子钟模型和蒙特卡洛取样的马尔可夫链构建时间、地域与 SNP 的关系模型，因此，在 SNP 构建的系统发育树基础上进行分类学分析，有助于研究人员根据噬菌体的分类地位推测出进化的时间与地点信息。

在生物数据量大、生物结构复杂的大数据时代，机器学习能帮助研究者分析繁复的数据，开展分类学分析[68]。近年出现的这些机器学习研究很大程度上推进了噬菌体分类学的发展，特别是在以前难以分类的门、纲、科、属水平上开拓了一条新的道路，其他水平的分类也会在未来逐步完善，机器学习也是一个非常有效的技术方法。

1.6 通过机器学习辅助噬菌体分类学展望

除了 GCN 模型外，利用类似于 MMSeqs2 比对原理的分类模型也是种内噬菌体分类的新工具[69]，该模型依据每个种构建相应的 k-mer 为特征作为分类模型，随着数量变多，可能通过模型中得分高的特征作为鉴定种水平分类的依据。

如今，利用深度学习与人工智能开展噬菌体研究是分类学研究的新热点，如已有基于深度学习的

DeePVP[70] 开始应用于噬菌体研究、DeephageTP[71] 通过神经网络识别病毒基因，进而提供更多的噬菌体分类标尺。现在这类研究的主要局限在于是可供机器进行学习的噬菌体样本量较小，未来需要更大量的数据集才能通过各种方法训练出更为合理的模型以提供更系统的分类技术。

2 噬菌体分类学技术的新进展

随着研究的不断深入，未来将开展更为深入的噬菌体研究，而这就需要现在提出更完善的噬菌体分类方案。目前，基因组检测与分析领域涌现了大批技术突破，如宏基因组测序、机器学习分析策略等。这些新的技术进展将进一步突破目前噬菌体分类方法的局限，填补对噬菌体分类的认识空缺，从而为噬菌体在食品安全的应用中提供更多的参考。

目前，基于分离纯化的病毒开展分析仍是噬菌体分类的金标准，因此，基于培养组学的噬菌体分离及培养则成为噬菌体分类学研究的一个重要方向[72]。同时，许多研究使用基于混合菌群的宏基因组测序数据开展噬菌体基因组分类及功能研究（宏病毒组），该方法具有快速、高通量的应用特性，且规避了噬菌体分离的技术瓶颈[73]，是目前噬菌体分类学研究的热点前沿。采用类似分析手段，另外也有研究者通过挖掘宿主细菌基因组内的原噬菌体基因组信息，继而开展进一步的分类学研究，这一策略也开拓了噬菌体分类方法的又一新方向。

2.1 纯培养的噬菌体分类

分离纯化得到的纯培养噬菌体是研究的标准范式，而通过对分离纯化的菌体开展全基因组的测序分析及进化分类，是现今噬菌体分类的主要方法。同时，ICTV 认为研究者们应开展系列实验以获取噬菌体更多的表型（如宿主谱、一步生长曲线、裂解量、环境耐受、电镜下形态、最佳感染复数等），以作为种水平之下的进一步分类依据[20]。噬菌体纯培养物的获得基本依赖于研究者使用双层平板法进行多轮分离纯化及扩大培养。噬菌体基因组测序常使用酚仿法提取噬菌体遗传物质后使用高通量测序仪完成全基因组测序。获得纯培养的噬菌体基因组后，可通过完成噬菌体的进化分类。一个标准的分析流程是：1）比对核苷酸数据库。2）查看比对相似度，比对相似度≥95% 时，建议采用 2.1 所述的系统发育方法分类至种；比对相似度≥70% 则采用 2.3 所述的聚类方法辅助分类至亚科；若相似度＜70% 则需要参考 2.4 所述的基于机器学习的更大尺度的分类方法。3）获取合适的邻近基因组佐证分类地位。

目前，许多条件苛刻的宿主菌很难分离，即使分离了宿主菌，相应的噬菌体也很难分离。近年来中科院先进院采用了肠道噬菌体分离收集（Gut Phage Isolate Collection, GPIC）方法，采用全自动工作站对单份样品实施高达 42 种人类肠道常见宿主细菌特异性噬菌体的高通量筛选，从而实现人类粪便和废水样品中噬菌体颗粒的高效富集，通过该方法，研究团队从 20 名志愿者的粪便与污水的混合样品中获取了 209 种噬菌体，分属于 45 个病毒簇，其中包含 2 个新病毒科、34 个未鉴定的属[72]。因此，随着培养组学的发展，联合更广阔的宿主谱筛选以及全自动工作站的分离模式，可能是未来噬菌体分离培养的发展新方向。而通过分离培养获得的新型噬菌体，将为噬菌体分类研究提供更多的比对标准及分类依据。

2.2 基于宏病毒组分析的噬菌体分类

宏基因（Metagenome）是特定环境下所有生物遗传物质的总和，包含了可培养的和未培养的微生物的基因。一般从环境样品中提取基因组 DNA 进行高通量测序，从而分析微生物多样性、种群结构、功能信息、与环境之间的关系等[74]。病毒宏基因组（宏病毒组）是使用宏基因组方法对环境样本的病毒进行分析的方法，相较于传统的分离纯化来说，宏病毒组分析可以一次性获取大量的病毒，且可以获取不可培养的病毒，因此，使用宏病毒组分析噬菌体是未来的趋势[75]。宏病毒组分析目前广泛应用在环境中噬菌体的检测，如人体口腔、人体肠道、海洋、土壤等。获得宏病毒组数据后，研究者们可通过基于比对、深度学习等方法的软件查询或预测得到噬菌体序列，进而推测这些噬菌体的进化历程及潜在功能。目前宏病毒组分析软件及其应用特性如表 2 所示。

保证噬菌体完整的情况下，研究者可从宏病毒组挖掘的序列中利用噬菌体的功能域充当分子进化标尺，进而开展这些噬菌体各种层级的分类分析。其中，由于烈性噬菌体应用较广，而末端酶大亚基（TerL）是其包装基因组的重要基因功能域[76]，因此 TerL 基因功能域在宏病毒组分析中常被用作为科至亚域水平的分类标尺。

表 2 病毒基因组序列挖掘软件应用特性比较

Table 2 Comparison of application characteristics of viral genome sequence mining software

软件	客户端	最小长度限制/bp	溶原预测	宏病毒组	分类鉴定	算法	年份/年
Metavir[77]	网页	/	否	是	是	比对	2011
VIROME[78]	网页	1 000	否	是	否	比对	2012
VirSorter[79]	网页、本地	500	是	是	否	比对	2015
ViromeScan[80]	本地	60	否	是	否	比对	2016
MetaPhinder[81]	本地	5 000	是	是	是	比对	2016
VirFinder[82]	本地	500	是	是	是	深度学习	2017
MARVEL[83]	本地	2 000	否	是	否	比对	2018
ViraMiner[84]	本地	300	否	是	是	深度学习	2019
VIBRANT[85]	本地	1 000	是	是	是	深度学习	2020
DeepVirFinder[86]	本地	1 000	否	是	否	深度学习	2021
VirSorter2[87]	网页、本地	3 000	是	是	否	比对	2021
MetaPhage[88]	本地	3 000	否	是	是	比对、深度学习	2022
Virtifier[89]	本地	/	否	是	是	深度学习	2022
VirHunter[90]	本地	750	否	是	是	深度学习	2022
RNN-VirSeeker[91]	本地	100	否	是	否	深度学习	2022
Phanta[92]	本地	1 000	是	是	是	比对	2023
Phage_Finder[93]	网页	/	是	否	是	比对	2006
Prohinder[94]	网页	/	是	否	否	比对	2008
PHAST[95]	网页	/	是	否	否	比对	2011
PhiSpy[96]	本地	/	是	否	否	比对	2012
PHASTER[97]	网页	2 000	是	否	否	比对	2016
Prophage-Hunter[98]	网页	/	是	否	否	比对	2019
PHASTEST[99]	网页	2 000	是	否	否	比对	2023

2.3 原噬菌体的分类

原噬菌体是溶原于细菌染色体或质粒上的噬菌体。相较于可培养的噬菌体来说，更多不可培养的噬菌体以原噬菌体的状态存于细菌细胞中，这些噬菌体含有整合酶、转座酶、DNA 切割酶、溶原酶等，并随着细菌的复制而进行溶原循环[100]。在环境恶劣下，原噬菌体出于生存压力可直接进入裂解循环、逃离细菌。通过原噬菌体这一特征，研究者可采用丝裂霉素 C 等诱导剂使宿主处于恶劣环境，诱导细菌体内的原噬菌体脱离宿主，进而获得分离纯化的噬菌体[101]。不同于利用药物诱导分离原噬菌体，现在研究者也可以通过软件对宿主细菌的全基因组进行比对查找，无需分离则可高效认识原噬菌体的基因组信息。而在获得完整的原噬菌体基因组后，研究者可采用进化分类学的方法，对这些不可分离的噬菌体进行系统发育分析。

2.4 噬菌体分类学的技术瓶颈

如今，噬菌体分类学方向基本从巴尔的摩分类法、形态学分类法转向进化分类法。但进化分类法最大的瓶颈在于用作分类参考的噬菌体数量不足，进而新的噬菌体因为无法比对到相近的参考物种而缺乏分类学地位。在最新的研究进展中出现了通过分离纯化测序、环境宏病毒组挖掘、细菌基因组挖掘噬菌体的新方法，为解决噬菌体数量不足的瓶颈提供了解决方案。

相较于分离纯化的方法获得的噬菌体来说，宏病毒组分析具有效率高的优势，但其准确性与可信度仍有所欠缺。对于这一局限，有学者提出因采用 CheckV[102] 等软件对宏病毒组挖掘的序列进行控制分析，以提高获取序列的准确性。除此之外，宏病毒组结果中基因组序列较多以碎片化的形式存在，基因组的不完整导致宏病毒组来源的噬菌体序列仅 10% 能被明确一定的分类学地位[103]。这些技术的瓶颈仍有待研究者们通过提升高通量测序技术、优化噬菌体序列识别等方法逐一突破。

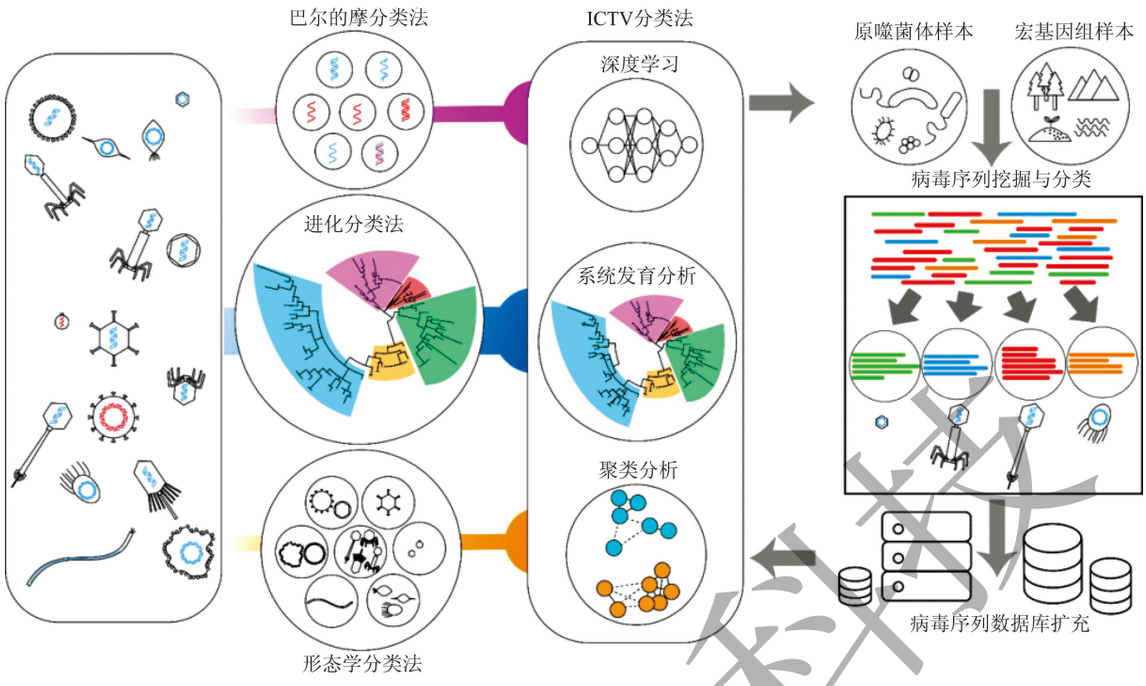

图 3 噬菌体分类学研究进展及当前瓶颈解决方案

Fig.3 Research progress of phage taxonomy and current bottleneck solution

注：遗传物质颜色，红色为 RNA，蓝色为 DNA。

3 总结与展望

为了加深对噬菌体的认识，从而安全地应用噬菌体控制食品中的耐药细菌污染，不断完善可靠的噬菌体分类技术是首要任务。目前，噬菌体分类学已经进入基因组分析的年代。ICTV 主持的进化分类学方法突破了巴尔的摩分类法和形态学分类法的局限，实现了在域、界、门、纲、属水平较为精准的分类。但是目前噬菌体在目和各个亚分类水平仍存在很多有待突破的空间。系统发育分析、聚类分析与深度学习，结合机器学习寻找新的分类标尺、采用机器学习辅助分类是目前噬菌体分类学发展的新方向之一。另一方面，采用基因挖掘手段可以直接从复杂样品中提取噬菌体基因组，规避了目前噬菌体分离培养效率低的技术瓶颈，该策略从另外一个角度完善了人们对噬菌体的认识。随着未来分离获得的噬菌体增多，将对噬菌体有更多的认识，而这些新认识又将推动噬菌体分类新标尺的选择与分类分析方法的进一步优化。系统准确的噬菌体分类体系将协助研究者们开展更为高效的噬菌体研究，推动噬菌体在食品安全等多个领域的应用发展。

参考文献

[1] DION M B, OECHSLIN F, MOINEAU S. Phage diversity, genomics and phylogeny [J]. Nature Reviews Microbiology, 2020, 18(3): 125-138.

[2] HANKIN E H. L'action bactericide des eaux de la Jumna et du Gange sur le vibrion du cholera [J]. Annales de l'Institut Pasteur, 1896, 10(11): 511.

[3] WALSH C. Where will new antibiotics come from? [J]. Nature Reviews Microbiology, 2003, 1(1): 65-70.

[4] SAMSON J E, MAGADÁN A H, SABRI M, et al. Revenge of the phages: defeating bacterial defences [J]. Nature Reviews Microbiology, 2013, 11(10): 675-687.

[5] LAUBER C, GORBALENYA A E. Toward genetics-based virus taxonomy: comparative analysis of a genetics-based classification and the taxonomy of picornaviruses [J]. Journal of Virology, 2012, 86(7): 3905-3915.

[6] VAN REGENMORTEL M H V, MAHY B W J. Emerging issues in virus taxonomy [J]. Emerging Infectious Diseases, 2004, 10(1): 8-13.

[7] BALTIMORE D. Expression of animal virus genomes [J]. Bacteriological Reviews, 1971, 35(3): 235-241.

[8] KOONIN E V, KRUPOVIC M, AGOL V I. The baltimore classification of viruses 50 years later: how does it stand in the light of virus evolution? [J]. Microbiology and Molecular Biology Reviews, 2021, 85(3): e5321.

[9] ACKERMANN H. (2012). Chapter 1-Bacteriophage Electron Microscopy [M]. Academic Press, 1-32.

[10] NELSON D. Phage taxonomy: we agree to disagree [J]. Journal of Bacteriology, 2004, 186(21): 7029-7031.

[11] LLOYD-PRICE J, ABU-ALI G, HUTTENHOWER C. The healthy human microbiome [J]. Genome Medicine, 2016, 8(1): 51.

[12] CHIBANI C M, FARR A, KLAMA S, et al. Classifying the unclassified: a phage classification method [J]. Viruses, 2019, 11(2): 195.

[13] BIN J H, BOLDUC B, ZABLOCKI O, et al. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks [J]. Nature Biotechnology, 2019, 37(6): 632-639.

[14] EVSEEV P, GUTNIK D, SHNEIDER M, et al. Use of an integrated approach involving alphafold predictions for the evolutionary taxonomy of duplodnaviria viruses [J]. Biomolecules, 2023, 13(1): 110.

[15] ADRIAENSSENS E M, SULLIVAN M B, KNEZEVIC P, et al. Taxonomy of prokaryotic viruses: 2018-2019 update from the ICTV bacterial and archaeal viruses subcommittee [J]. Archives of Virology, 2020, 165(5): 1253-1260.

[16] TURNER D, SHKOPOROV A N, LOOD C, et al. Abolishment of morphology-based taxa and change to binomial species names: 2022 taxonomy update of the ICTV bacterial viruses subcommittee [J]. Archives of Virology, 2023, 168(2): 74.

[17] SIMMONDS P, ADRIAENSSENS E M, ZERBINI F M, et al. Four principles to establish a universal virus taxonomy [J]. PLoS Biology, 2023, 21(2): e3001922.

[18] SIMMONDS P, AIEWSAKUN P. Virus classification - where do you draw the line? [J]. Archives of Virology, 2018, 163(8): 2037-2046.

[19] SCHOCH C L, CIUFO S, DOMRACHEV M, et al. NCBI Taxonomy: a comprehensive update on curation, resources and tools [J]. Database, 2020, 2020:1-21.

[20] GORBALENYA A E, KRUPOVIC M, MUSHEGIAN A, et al. The new scope of virus taxonomy: partitioning the virosphere into 15 hierarchical ranks [J]. Nature Microbiology, 2020, 5(5): 668-674.

[21] YANG Z, RANNALA B. Molecular phylogenetics: principles and practice [J]. Nature Reviews Genetics, 2012, 13(5): 303-314.

[22] KAPLI P, YANG Z, TELFORD M J. Phylogenetic tree building in the genomic age [J]. Nature Reviews Genetics, 2020, 21(7): 428-444.

[23] GASCUEL O. Neighbor-joining revealed [J]. Molecular Biology and Evolution, 2006, 23(11): 1997-2000.

[24] WOLF Y I, KAZLAUSKAS D, IRANZO J, et al. Origins and evolution of the global rna virome [J]. mBio, 2018, 9(6): e02329-18.

[25] KOONIN E V, DOLJA V V, KRUPOVIC M, et al. Global organization and proposed megataxonomy of the virus world [J]. Microbiology and Molecular Biology Reviews, 2020, 84(2): e00061-19.

[26] BENLER S, YUTIN N, ANTIPOV D, et al. Thousands of previously unknown phages discovered in whole-community human gut metagenomes [J]. Microbiome, 2021, 9(1): 78.

[27] GHORBANI A, SAMARFARD S, ESKANDARZADE N, et al. Comparative phylogenetic analysis of SARS-CoV-2 spike protein-possibility effect on virus spillover [J]. Briefings in Bioinformatics, 2021, 22(5): 1-9.

[28] STONEKING M, KRAUSE J. Learning about human population history from ancient and modern genomes [J]. Nature Reviews Genetics, 2011, 12(9): 603-614.

[29] PANG S, OCTAVIA S, REEVES P R, et al. Genetic relationships of phage types and single nucleotide polymorphism typing of *Salmonella enterica* serovar typhimurium [J]. Journal of Clinical Microbiology, 2012, 50(3): 727-734.

[30] TREANGEN T J, ONDOV B D, KOREN S, et al. The harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes [J]. Genome Biology, 2014, 15(11): 524.

[31] THOMPSON J D, HIGGINS D G, GIBSON T J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice [J]. Nucleic Acids Research, 1994, 22(22): 4673-4680.

[32] NOTREDAME C, HIGGINS D G, HERINGA J. T-Coffee: A novel method for fast and accurate multiple sequence alignment [J]. Journal of Molecular Biology, 2000, 302(1): 205-217.

[33] ROZEWICKI J, LI S, AMADA K M, et al. MAFFT-DASH: integrated protein sequence and structural alignment [J]. Nucleic Acids Research, 2019, 47(W1): W5-W10.

[34] EDGAR R C. MUSCLE: multiple sequence alignment with high accuracy and high throughput [J]. Nucleic Acids Research, 2004, 32(5): 1792-1797.

[35] LASSMANN T, SONNHAMMER E L. Kalign-an accurate and fast multiple sequence alignment algorithm [J]. BMC Bioinformatics, 2005, 6(1): 298.

[36] SIEVERS F, WILM A, DINEEN D, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega [J]. Molecular Systems Biology, 2011, 7(1): 539.

[37] POSADA D, CRANDALL K A. MODELTEST: testing the model of DNA substitution [J]. Bioinformatics, 1998, 14(9):

817-818.

[38] ABASCAL F, ZARDOYA R, POSADA D. ProtTest: selection of best-fit models of protein evolution [J]. Bioinformatics, 2005, 21(9): 2104-2105.

[39] POSADA D. jModelTest: phylogenetic model averaging [J]. Molecular Biology and Evolution, 2008, 25(7): 1253-1256.

[40] DARRIBA D, POSADA D, KOZLOV A M, et al. Model test-NG: A new and scalable tool for the selection of DNA and protein evolutionary models [J]. Molecular Biology and Evolution, 2020, 37(1): 291-294.

[41] CHEN S, SU S, LO C, et al. PALM: A paralleled and Integrated Framework for Phylogenetic Inference with Automatic Likelihood Model Selectors [J]. PLoS One, 2009, 4(12): e8116.

[42] DARRIBA D, TABOADA G L, DOALLO R, et al. ProtTest 3: fast selection of best-fit models of protein evolution [J]. Bioinformatics, 2011, 27(8): 1164-1165.

[43] KALYAANAMOORTHY S, MINH B Q, WONG T K F, et al. ModelFinder: fast model selection for accurate phylogenetic estimates [J]. Nature Methods, 2017, 14(6): 587-589.

[44] LEFORT V, LONGUEVILLE J, GASCUEL O. SMS: Smart model selection in PhyML [J]. Molecular Biology and Evolution, 2017, 34(9): 2422-2424.

[45] GUINDON S, LETHIEC F, DUROUX P, et al. PHYML Online-a web server for fast maximum likelihood-based phylogenetic inference [J]. Nucleic Acids Research, 2005, 33(Web Server): W557-W559.

[46] YANG Z. PAML 4: Phylogenetic analysis by maximum likelihood [J]. Molecular Biology and Evolution, 2007, 24(8): 1586-1591.

[47] STAMATAKIS A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies [J]. Bioinformatics, 2014, 30(9): 1312-1313.

[48] PRICE M N, DEHAL P S, ARKIN A P. Fast tree 2-approximately maximum-likelihood trees for large alignments [J]. PLoS One, 2010, 5(3): e9490.

[49] NGUYEN L, SCHMIDT H A, von HAESELER A, et al. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies [J]. Molecular Biology and Evolution, 2015, 32(1): 268-274.

[50] MINH B Q, SCHMIDT H A, CHERNOMOR O, et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era [J]. Molecular Biology and Evolution, 2020, 37(5): 1530-1534.

[51] OKONECHNIKOV K, GOLOSOVA O, FURSOV M. Unipro UGENE: a unified bioinformatics toolkit [J]. Bioinformatics, 2012, 28(8): 1166-1167.

[52] TALEVICH E, INVERGO B M, COCK P J, et al. Bio. Phylo: A unified toolkit for processing, analyzing and visualizing phylogenetic trees in Biopython [J]. BMC Bioinformatics, 2012, 13(1): 209.

[53] YU G, LAM T T, ZHU H, et al. Two methods for mapping and visualizing associated data on phylogeny using ggtree [J]. Molecular Biology and Evolution, 2018, 35(12): 3041-3043.

[54] KUMAR S, STECHER G, LI M, et al. MEGA X: molecular evolutionary genetics analysis across computing Platforms [J]. Molecular Biology and Evolution, 2018, 35(6): 1547-1549.

[55] SUBRAMANIAN B, GAO S, LERCHER M J, et al. Evolview v3: a webserver for visualization, annotation, and management of phylogenetic trees [J]. Nucleic Acids Research, 2019, 47(W1): W270-W275.

[56] LETUNIC I, BORK P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation [J]. Nucleic Acids Research, 2021, 49(W1): W293-W296.

[57] TOMBA NGANGAS S, BISSEUX M, JUGIE G, et al. Coxsackievirus A6 recombinant subclades D3/A and D3/H were predominant in hand-foot-and-mouth disease outbreaks in the paediatric population, france, 2010–2018 [J]. Viruses, 2022, 14(5): 1078.

[58] I DE LA HIGUERA, KASUN G W, TORRANCE E L, et al. Unveiling crucivirus diversity by mining metagenomic data [J]. mBio, 2020, 11(5): e01410-e01420.

[59] STANTON C R, RICE D T F, BEER M, et al. Isolation and characterisation of the bundooravirus genus and phylogenetic investigation of the salasmaviridae bacteriophages [J]. Viruses, 2021, 13(8): 1557.

[60] BOLDUC B, JANG H B, DOULCIER G, et al. vConTACT: an iVirus tool to classify double-stranded DNA viruses that infect Archaea and Bacteria [J]. PeerJ, 2017, 5: e3243.

[61] LAUBER C, GORBALENYA A E. Partitioning the genetic diversity of a virus family: approach and evaluation through a case study of picornaviruses [J]. Journal of Virology, 2012, 86(7): 3890-3904.

[62] MORARU C, VARSANI A, KROPINSKI A M. VIRIDIC-A novel tool to calculate the intergenomic similarities of prokaryote-infecting viruses [J]. Viruses, 2020, 12(11): 1268.

[63] KOONIN E V, DOLJA V V, KRUPOVIC M. The logic of virus evolution [J]. Cell Host & Microbe, 2022, 30(7): 917-929.

[64] JUMPER J, EVANS R, PRITZEL A, et al. Highly accurate protein structure prediction with AlphaFold [J]. Nature, 2021, 596(7873): 583-589.

[65] JIANG J, YUAN W, SHANG J, et al. Virus classification for viral genomic fragments using PhaGCN2 [J]. Briefings in Bioinformatics, 2023, 24(1): 1-9.

[66] BAELE G, LEMEY P, RAMBAUT A, et al. Adaptive MCMC in Bayesian phylogenetics: an application to analyzing partitioned data in BEAST [J]. Bioinformatics, 2017, 33(12):

1798-1805.

[67] LEWIS P O, HOLDER M T, SWOFFORD D L. Phycas: Software for Bayesian Phylogenetic Analysis [J]. Systematic Biology, 2015, 64(3): 525-531.

[68] GREENER J G, KANDATHIL S M, MOFFAT L, et al. A guide to machine learning for biologists [J]. Nature Reviews Molecular Cell Biology, 2022, 23(1): 40-55.

[69] STEINEGGER M, SÖDING J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets [J]. Nature Biotechnology, 2017, 35(11): 1026-1028.

[70] FANG Z, FENG T, ZHOU H, et al. DeePVP: Identification and classification of phage virion proteins using deep learning [J]. GigaScience, 2022, 11: 1-10.

[71] CHU Y, GUO S, CUI D, et al. DeephageTP: a convolutional neural network framework for identifying phage-specific proteins from metagenomic sequencing data [J]. PeerJ, 2022, 10: e13404.

[72] SHEN J, ZHANG J, MO L, et al. Large-scale phage cultivation for commensal human gut bacteria [J]. Cell Host & Microbe, 2023, 31(4): 665-677.

[73] ZHANG Y, SHI M, HOLMES E C. Using metagenomics to characterize an expanding virosphere [J]. Cell, 2018, 172(6): 1168-1172.

[74] QUINCE C, WALKER A W, SIMPSON J T, et al. Shotgun metagenomics, from sampling to analysis [J]. Nature Biotechnology, 2017, 35(9): 833-844.

[75] EDWARDS R A, ROHWER F. Viral metagenomics [J]. Nature Reviews Microbiology, 2005, 3(6): 504-510.

[76] RAO V B, FEISS M. Mechanisms of DNA packaging by large double-stranded DNA viruses [J]. Annual Review of Virology, 2015, 2(1): 351-378.

[77] ROUX S, FAUBLADIER M, MAHUL A, et al. Metavir: a web server dedicated to virome analysis [J]. Bioinformatics, 2011, 27(21): 3074-3075.

[78] WOMMACK K E, BHAVSAR J, POLSON S W, et al. VIROME: a standard operating procedure for analysis of viral metagenome sequences [J]. Standards in Genomic Sciences, 2012, 6(3): 427-439.

[79] ROUX S, ENAULT F, HURWITZ B L, et al. VirSorter: mining viral signal from microbial genomic data [J]. PeerJ, 2015, 3: e985.

[80] RAMPELLI S, SOVERINI M, TURRONI S, et al. ViromeScan: a new tool for metagenomic viral community profiling [J]. BMC Genomics, 2016, 17: 165.

[81] JURTZ V I, VILLARROEL J, LUND O, et al. Meta phinder -identifying bacteriophage sequences in metagenomic data sets [J]. PLoS One, 2016, 11(9): e163111.

[82] 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 [J]. Microbiome, 2017, 5(1): 69

[83] AMGARTEN D, BRAGA L P P, DA SILVA A M, et al. MARVEL, a tool for prediction of bacteriophage sequences in metagenomic bins [J]. Frontiers in Genetics, 2018, 9: 304.

[84] 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): e222271.

[85] KIEFT K, ZHOU Z, ANANTHARAMAN K. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences [J]. Microbiome, 2020, 8(1): 90.

[86] REN J, SONG K, DENG C, et al. Identifying viruses from metagenomic data using deep learning [J]. Quantitative Biology, 2020, 8(1): 64-77.

[87] GUO J, BOLDUC B, ZAYED A A, et al. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses [J]. Microbiome, 2021, 9(1): 37.

[88] PANDOLFO M, TELATIN A, LAZZARI G, et al. MetaPhage: an automated pipeline for analyzing, annotating, and classifying bacteriophages in metagenomics sequencing data [J]. mSystems, 2022, 7(5): e74122.

[89] MIAO Y, LIU F, HOU T, et al. Virtifier: a deep learning-based identifier for viral sequences from metagenomes [J]. Bioinformatics, 2022, 38(5): 1216-1222.

[90] SUKHORUKOV G, KHALILI M, GASCUEL O, et al. VirHunter: A deep learning-based method for detection of novel rna viruses in plant sequencing data [J]. Frontiers in Bioinformatics, 2022, 2: 867111.

[91] 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.

[92] PINTO Y, CHAKRABORTY M, JAIN N, et al. Phage-inclusive profiling of human gut microbiomes with Phanta [J]. Nature Biotechnology, 2024, 42: 651-662.

[93] FOUTS D E. Phage_Finder: Automated identification and classification of prophage regions in complete bacterial genome sequences [J]. Nucleic Acids Research, 2006, 34(20): 5839-5851.

[94] LIMA-MENDEZ G, Van HELDEN J, TOUSSAINT A, et al. Prophinder: a computational tool for prophage prediction in prokaryotic genomes [J]. Bioinformatics, 2008, 24(6): 863-865.

[95] ZHOU Y, LIANG Y, LYNCH K H, et al. PHAST: A fast phage search tool [J]. Nucleic Acids Research, 2011, 39(suppl): W347-W352.

[96] 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 [J]. Nucleic Acids Research, 2012, 40(16): e126.

[97] ARNDT D, GRANT J R, MARCU A, et al. PHASTER: a better, faster version of the PHAST phage search tool [J]. Nucleic Acids Research, 2016, 44(W1): W16-W21.

[98] SONG W, SUN H, ZHANG C, et al. Prophage Hunter: an integrative hunting tool for active prophages [J]. Nucleic Acids Research, 2019, 47(W1): W74-W80.

[99] WISHART D S, HAN S, SAHA S, et al. PHASTEST: faster than PHASTER, better than PHAST [J]. Nucleic Acids Research, 2023, 51(W1): W443-W450.

[100] CANCHAYA C, PROUX C, FOURNOUS G, et al. Prophage Genomics [J]. Microbiology and Molecular Biology Reviews, 2003, 67(2): 238-276.

[101] SILPE J E, WONG J W H, OWEN S V, et al. The bacterial toxin colibactin triggers prophage induction [J]. Nature, 2022, 603(7900): 315-320.

[102] NAYFACH S, CAMARGO A P, SCHULZ F, et al. CheckV assesses the quality and completeness of metagenome-assembled viral genomes [J]. Nature Biotechnology, 2021, 39(5): 578-585.

[103] 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.