•综述•

基因组学技术在病毒鉴定与宿主溯源中的应用

韩本凤 周 欣 张 雪*

(中国农业大学植物保护学院昆虫学系, 北京 100193)

摘要: 基因组学技术, 特别是宏基因组测序在未知病毒的鉴定与溯源中起到了重要作用。相较于传统的病毒分离培养方法, 宏基因组技术可以从混合样本中获得病毒的核酸序列, 极大加速了未知病毒的鉴定与溯源, 在针对高流行性、高致病性的病毒研究中发挥了重要作用。基于宏基因组技术对未知病毒进行鉴定和溯源, 其准确性很大程度上依赖于取样及已知宿主的病毒库的完整性。然而, 当前病毒多样性的基础研究相对薄弱, 病毒的宿主信息则更加匮乏。野生动物和畜禽是人畜共患病致病病毒的重要中间宿主, 构建广泛的动物-病毒关联数据库对于准确、快速地鉴定和预防致病性病毒具有重要意义。本综述以SARS-CoV-2为例, 总结了基因组学技术在病毒的鉴定与溯源上的应用, 并针对当前动物病毒库完整性低的现状, 对构建野生和家养动物携带病毒的关联数据库的可行性提出依据与建议。

关键词: 新冠病毒; 高通量测序; 病毒多样性; 宿主溯源; 病毒进化

Verification of virus identity and host association using genomics technology

Benfeng Han, Xin Zhou, Xue Zhang*

Department of Entomology, College of Plant Protection, China Agricultural University, Beijing 100193

Abstract: Genomics technology, especially metagenomic sequencing, has played an important role in identifying and tracing unknown viruses. While classical methods in virus taxonomy rely on phenotypic traits, the metagenomics pipeline assembles new virus genomes from short nucleotide fragments without the need for any *a priori* reference sequences. This new technology increases the efficiency in identifying viruses and hosts associated with those viruses. This is particularly useful in identifying viruses that can cause epidemics. One current challenge in accomplishing this, is the ability to trace the original and intermediate viral hosts. To do this, a comprehensive virus sequence library characterized by definite host information is needed. Unfortunately, such information is still limited. As wild and stock animals are main sources for pathogenic viruses, an extensive survey of the global virome is vitally important to help identify and prevent zoonotic epidemics. This review summarizes the application of genomics technologies in the identification of viruses and the hosts associated with those viruses, using the outbreak of SARS-CoV-2 as an example. We also address intrinsic drawbacks of current methodologies as well as the incompleteness of available virus libraries. We propose the necessity and feasibility in constructing a comprehensive virus database with host association that emphasizes the diversity of viruses and their interactions with other organisms.

Key words: SARS-CoV-2; high-throughput sequencing; virus diversity; host association; virus evolution

2019年12月, 在新型冠状病毒暴发之初, 通过对患者的肺泡灌洗液样品进行宏基因组测序, 获得了病毒株全基因组序列, 并确认该病毒为冠状病毒

科新型RNA病毒(Zhou et al, 2020)。基于系统发育关系的结果, 国际病毒分类学委员会的冠状病毒科研究小组(CSG)确认该病毒与SARS-CoVs形成姊妹支,

收稿日期: 2020-04-16; 接受日期: 2020-06-09

基金项目: 国家自然科学基金(31772493)、科技部科技基础资源调查专项(2018FY100403)和中国农业大学北京食品营养与人类健康高精尖创新中心基金

* 通讯作者 Author for correspondence. E-mail: zhangxue0215@gmail.com

并依据分类学和已有惯例将其命名为“SARS-CoV-2” (Gorbalenya et al, 2020), 新型冠状病毒感染的疾病被正式命名为“COVID-19”。虽然新型冠状病毒与SARS-CoV病毒近缘, 但并非同种, 且临床症状具有差异, 所以学界对新型冠状病毒使用“SARS-CoV-2”的命名依然存在争议。

经典的病毒鉴定主要依赖表型特性, 具体操作流程包括病毒培养、电镜观察、血清学反应和聚合酶链式反应(polymerase chain reaction, PCR)等, 操作繁琐, 费时费力, 且鉴定效率较低(Tang & Chiu, 2010)。而在本次疫情暴发期间, 从发现第一例患者到获得SARS-CoV-2全基因组序列, 用时仅20余天。对SARS-CoV-2的快速鉴定得益于宏基因组技术的迅速发展。

目前, 基于高通量测序技术的宏基因组方法被广泛应用于病毒组(virome)的研究中, 在发现和鉴定新病毒及检测病毒流行性上具有显著优势。基于宏基因组和宏转录组技术对环境样品的调查揭示了生物圈中存在着丰富的病毒种类(Simmonds et al, 2017), 其遗传物质、基因序列、编码蛋白、复制机制以及与细胞宿主的相互作用等方面具有丰富的多样性(Roossinck, 2015)。虽然基于测序的病毒研究方法无法直接获得经典的病毒表型特征, 但该方法有助于快速构建全球性的病毒多样性数据库(Simmonds et al, 2017)。而整合了病毒多样性和宿主关联信息的数据库对深入开展病毒学研究至关重要, 如: 病毒鉴定、病毒进化、宿主溯源以及病毒与宿主的协同进化等。

近年来, 利用基因组学技术开展区域性或全球性流行病的研究已经迅速普及。例如, 严重威胁人类健康的严重急性呼吸系统综合征(SARS) (Ge et al, 2013)、中东呼吸综合征(MERS) (Hemida et al, 2014)、埃博拉(Ebola, 出血性烈性传染病) (Quick et al, 2016)等疫病病毒的鉴定与宿主溯源都依赖于宏基因组技术的应用。本文以新型冠状病毒等为例, 介绍宏基因组技术在病毒鉴定、病毒宿主溯源等方面的应用现状和局限性, 并就如何完善广泛的动植物病毒基因组学数据库提出建议, 讨论其可行性。

1 基于宏基因组方法鉴定病毒种类

核酸测序技术已广泛应用于分子分类学研究。其中, 二代测序技术成本较低, 相关生物信息学的

分析方法比较完善, 可对不同物种来源的海量核酸片段进行分拣、组装和比对。宏基因组技术不需要对样品进行纯化, 对核酸质量要求相对较低, 而且可以在混合样本中分离出病毒与宿主序列, 并组装成更长的片段甚至是完整的基因组序列。将拼接出的序列与已知病毒库进行比对, 根据序列相似度及系统发育关系可以确定未知病毒所属的类群。由于直接从头拼接(*de novo assembly*)病毒序列的过程可以不依赖于同源参考序列, 宏基因组方法极大促进了未知病毒病原物的检测和鉴定(Hadidi et al, 2016)。高通量测序方法在动物和植物源病毒的检测方面都有较广泛的应用, 如: 与脑膜炎密切相关的沙粒病毒(Palacios et al, 2008)、与南非出血热疾病相关的病原体Lujo病毒(沙粒病毒科) (Briese et al, 2009)、新甲型H1N1流感病毒和季节性甲型H3N2流感病毒(Hu et al, 2011)、番茄新病毒—蛇鞭菊斑斑病毒(Gayfeather mildmottle virus, GMMV) (Adams et al, 2009)等。

由于SARS-CoV-2与SARS-CoV具有极高的系统发育相关性和较高的序列相似度, 其序列结果一经公布就引起了公共卫生系统的高度警惕, 并迅速采取紧急防范与隔离措施, 展开救治, 充分体现出宏基因组技术在重大疫情的鉴定和预警方面的时效性。国家生物信息中心2019新型冠状病毒信息库(2019nCoV) (<https://bigd.big.ac.cn/ncov/>)以及全球流感共享数据库(Global Initiative on Sharing All Influenza Data, GISAID) (<https://www.gisaid.org/>)及时汇总和更新SARS-CoV-2病毒序列信息, 至2020年5月5日已收录SARS-CoV-2序列超过16,000条, 分别来自英国、美国、澳大利亚、中国等200多个国家和地区, 为基于全基因组的SARS-CoV-2病毒的进化和扩散进程研究提供了有力的数据支撑。

虽然宏基因组技术在序列的组装环节可以不依赖于参考序列, 但其后对待检病毒序列的鉴定仍然需要与已知病毒库进行比对, 利用序列相似性和系统发育关系对未知病毒进行鉴定。然而, 基于经典病毒学方法建立起来的已知病毒数据库的完整性较低, 目前在病毒病原体数据库和分析资源(Virus Pathogen Database and Analysis Resource, ViPR) (<https://www.viprbrc.org/brc/home.spg?decorator=vipr>)中收录的正式命名、有较详细描述病毒种类仅6,264多种(截至2020年5月5日)。而近期基于高通量

测序技术和宏基因组方法的普查都提示自然生态系统中可能存在大量的未知病毒序列(Simmonds et al, 2017), 自然界中实际病毒种类可能高达百万。宏基因组方法不仅为加速完善病毒数据库提供了快速通道, 也为基于序列比对和构建进化关系来完成病毒溯源提供了可能性。

2 构建基因组系统发育关系, 分析病毒进化历史并推测其宿主来源

在获得病毒基因组序列的基础上, 可以使用物种间具有同源关系的基因片段构建病毒的系统发育关系, 解析病毒的进化历史。常用的方法包括构建系统发育树(phylogenetic tree)、重建系统发育中的祖先状态(reconstructing ancestral state in phylogenies, RASP)、系统发育网络分析(phylogenetic network analysis)、单倍型网络分析(haplotype network analysis)、贝叶斯进化分析(Bayesian evolutionary analyses)、重组分析(recombination analysis)等(表1)。

系统发育树可用于分析不同生物之间的相关性, 通过树状分支可视化生物之间的亲缘关系并推测其进化历史, 常用软件有BEAST、IQ-TREE、RAxML、phyML等。用于构建系统发育树的方法很

多, 其中最大似然法(maximum likelihood, ML)被广泛应用, 比如用于对结核分支杆菌(*Mycobacterium tuberculosis*)的进化分析(Luo et al, 2015)以及对近期暴发的SARS-CoV-2病毒的进化关系研究(Wu et al, 2020)。ML的原理是将每个位置可能出现的所有残基替换概率进行累加, 产生特定定位点的似然值, 并对所有可能的系统发育树都计算似然函数, 似然函数值最大的树即为最可能的系统发育树; 其特点是需要确定序列进化的模型, 如Jukes-Cantor模型等。在进化模型选择合理的情况下, ML是与进化事实吻合最好的建树算法。而贝叶斯推断法(Bayesian inference)是一种新的利用贝叶斯演绎法预测种系发生史的系统进化分析方法, 相较于最大似然法其在计算速度与样本量上更具优势, 已被应用于乙型肝炎病毒(hepatitis B virus)和乳头状瘤病毒(papillomaviruses)等的进化分析(Frias-De-Diego et al, 2019; Yuen et al, 2019)。如果用于构建系统关系的已知病毒序列带有明确的宿主关联信息, 那么将有助于通过未知病毒在系统树上的相对位置推测其可能宿主。系统发育树为重建祖先状态提供依据, 比如使用RASP等方法, 结合系统发育树和病毒地理分布来推断病毒的历史传播路径(Yu et al, 2015)。通过构建系统发育树方法来分析病毒进化关系, 其

表1 用于病毒系统发育分析的常用方法
Table 1 Methods used in the phylogenetic analysis of viruses

分析方法 Method	作用 Application	常用软件 Software	参考文献 Reference
系统发育树 Phylogenetic trees	分析不同生物间相关性, 通过树状分支可视化生物之间的亲缘关系并推测进化历史 Analyzing the correlation between different organisms, visualizing the relationship between organisms through tree branches and speculating on the evolutionary history	RAxML, phyML, IQ-TREE	Zhou et al, 2020 Wu et al, 2020 Frias-De-Diego et al, 2019 Yuen et al, 2019
重建系统发育中的祖先状态 Reconstructing ancestral state in phylogenies (RASP)	重建祖先在系统发生树上的地理分布, 推断历史生物地理学信息 Inferring historical biogeography through reconstructing ancestral geographic distributions on phylogenetic trees	RASP	Luo et al, 2015 Frias-De-Diego et al, 2019 Yuen et al, 2019
系统发育网络分析 Phylogenetic network analysis	综合一系列系统发育树的可视化结果, 较直观地展示重组等性状冲突事件 Enabling the visualization of a multitude of optimal trees, displaying reorganization and other trait conflict events	Network, SplitsTree	Yu et al, 2020
单倍型网络分析 Haplotype network analysis	用于在群体水平上显示个体基因型之间的关系 An intuitive method used in visualizing relationships between individual genotypes in a population level	PopART	Tang et al, 2020 Leigh & Bryant, 2015
贝叶斯进化分析 Bayesian evolutionary analyses	根据时间进化树建模推断种群发生分歧的时间 Inferring the time when the population diverged based on time evolutionary tree modeling by BEAST	BEAST	Luo et al, 2015 Bouckaert et al, 2014 Suchard et al, 2018
重组分析 Recombination analysis	检验可能存在的重组信号, 揭示重组在基因进化中的作用 Identifying possible recombination signals and revealing the role of recombination in gene evolution	RDP4 Simplot phyML	Wu et al, 2020 Lam et al, 2020

准确性会受到基因组重组等因素的干扰,而系统发育网络分析方法通过综合一系列系统发育树的可视化结果,较直观地展示重组等性状冲突事件,可以弥补进化树方法的不足(Gusfield et al, 2004; Huson & Kloepper, 2005),在微生物、植物和昆虫的进化分析中被广泛使用(Kunin et al, 2005; Mavarez et al, 2006; McBreen & Lockhart, 2006)。重组分析通过检测重组信号发现重组事件,揭示重组在进化中的作用,广泛应用于病毒进化分析(Wu et al, 2020; Lam et al, 2020)。另外,还有其他一些应用于种群进化分析的方法,比如单倍型网络,可以用于种内水平的系谱关系构建,依据样品的地理学特征推断种群的传播路径(Leigh & Bryant, 2015)。基于网状树的种群分析(population analysis with reticulate trees, PopART)软件包,是单倍型网络构建的重要工具,可以对种内序列间的遗传关系进行推理和可视化(Leigh & Bryant, 2015)。贝叶斯进化分析利用高效的马尔科夫链蒙特卡洛(Markov Chain Monte Carlo)统计方法,将分子系统发育关系重建与复杂的离散或连续的性状演变、分歧时间、人口统计模型相结合,用于推断群体的系统动力学(Suchard et al, 2018)。贝叶斯进化分析的常用工具,即基于抽样树的贝叶斯进化分析(Bayesian Evolutionary Analysis by Sampling Trees, BEAST)软件包,可以通过遗传序列数据对样品进行贝叶斯系统发育(phylogenetic)与系统动力学(phylogenetic)推断,该软件依赖于时间尺度树(time-scaled trees)的进化分析方法,使其在共同祖先推断、种群演化历史分析等方面具有重要地位(Suchard et al, 2018)。

系统发育分析在SARS-CoV-2的溯源研究中起到了重要作用。已有研究发现,蝙蝠携带丰富的SARS相关冠状病毒(SARS-related coronaviruses, SARSr-CoVs) (Hu et al, 2017),被认为是天然的冠状病毒基因库,因此也被认为是SARS-CoV-2的可能宿主来源。这一假定在一定程度上得到近期病毒系统关系研究的支持: SARS-CoV-2与来自中华菊头蝠(*Rhinolophus sinicus*)的1株冠状病毒RaTG13具有最近的系统关系,用于建树的同源基因序列在两种病毒间的序列相似性达到96.2% (Zhou et al, 2020)。病毒系统树的最终拓扑关系与基因片段的选取和建树方法直接相关。与前述使用全基因组水平同源基因的方法不同,部分研究倾向于选择具有重

要功能的蛋白进行建树和进化分析。对于冠状病毒而言, S基因编码的刺突蛋白(spike protein)具有病毒侵染人体的重要标志受体结构域(receptor-binding domain, RBD),通过结合人体细胞表面受体血管紧张素转化酶II (angiotensin converting enzyme II, ACE2)等受体完成侵入(Zhou et al, 2020)。而基于RBD蛋白序列的系统发育分析并不支持蝙蝠冠状病毒RaTG13是SARS-CoV-2最相近的病毒。Lam等(2020)研究发现,在穿山甲(*Manis javanica*)体内发现的1株冠状病毒与SARS-CoV-2在RBD序列上具有更高的相似度,由此推测穿山甲可能为中间宿主,促进SARS-CoV-2的产生与传播。在野生动物中发现与SARS-CoV-2高度相似的RBD,以及SARS-CoV-2中不含用于人工病毒合成的任何病毒基因骨架(Almazan et al, 2014; Cui et al, 2019),证明SARS-CoV-2并不是通过人工基因编辑合成的病毒(Andersen et al, 2020)。必须指出的是,基于全基因组同源基因序列,只能对病毒的可能来源进行推测,其准确性极大依赖于样本选取的完整性、基因的选择和进化树构建的方法。而分子系统发育研究中所面临的一系列经典问题,如:核酸组成偏向性、突变速率的异质性、长支吸引、进化模型的选择等,都可能影响病毒进化和溯源的结果。

基于病毒的关键蛋白序列建树所得出的物种间的拓扑关系会受到该序列重组等因素的影响。例如,蝙蝠携带的沙贝病毒亚属(*Sarbecovirus*, 属于冠状病毒属*Betacoronavirus*)存在广泛的基因重组(Hon et al, 2008),而SARS-CoV的直接祖先有可能是由蝙蝠体内的SARSr-CoVs经频繁重组产生的(Hu et al, 2017)。据此推测,可以利用人类ACE2受体的各种SARSr-CoV可能仍在蝙蝠中广泛传播,而类似SARS的疾病风险也可能长期存在(Hu et al, 2017)。在蝙蝠中与SARS-CoV-2亲缘关系非常近的两株冠状病毒ZC45和ZC21很有可能是重组病毒,各自携带SARS-CoV和SARS-CoV-2的部分基因片段(Lam et al, 2020)。使用不同基因片段进行的进化分析获得的SARS-CoV-2在沙贝病毒亚属内的拓扑位置不同,表明这类病毒中曾经发生过某些基因的重组(Wu et al, 2020)。因此,如果仅基于单个蛋白基因序列进行分子系统发育分析,并不能对含有重组蛋白基因的病毒进行准确的关系构建,需结合多维度的证据对病毒起源进行推导。

如果将基因组技术与系统发育网络方法结合起来进行系统发育关系分析则会大大降低病毒基因重组的干扰, 提供较精细的进化分析, 用于寻找疾病暴发点。Yu等(2020)将SARS-CoV-2基因组中鉴定出的单倍型分为5组, 以bat-RaTG13-CoV为外群, 通过SplitsTree进行系统发育网络分析, 发现H13和H38单倍型可能是祖先单倍型, 而H1则来自中间单倍型H3。Tang等(2020)使用SARS-CoV-2基因组中的所有单倍型(基于DnaSP生成的单倍型), 通过PopART重建单倍型网络。发现SARS-CoV-2分为“L”与“S”两种主要类型, 并推测“S”型可能为“L”型的祖先。Forster等(2020)首次将系统发育网络方法用于研究SARS-CoV-2的传播途径, 通过大量的人群个体取样, 追踪SARS-CoV-2在人群中的暴发与传播轨迹, 发现了SARS-CoV-2中普遍存在的3个氨基酸的变异, 并指出变异之一为其余二者的祖先病毒。然而, 有研究者对该研究提出了质疑, 认为其选用的根类群(蝙蝠冠状病毒RaTG13)与SARS-CoV-2进化距离较远, 无法准确构建内群的进化关系。RaTG13与SARS-CoV-2的差异远远大于SARS-CoV-2之间的差异, 不能为内群特征的极向(即对祖征/衍征的判断)提供直接支持。可见, 建立和完善全球病毒库有利于更合理地选择外群病毒, 提高病毒进化关系分析的准确性。

综上所述, 基因组学技术可以通过比较未知病毒基因组序列和已知宿主的病毒库, 根据病毒间的系统发育关系来推测未知病毒的可能宿主来源。但是, 以构建系统发育关系为核心的病毒进化研究仍然面临一系列固有的方法学局限和挑战。其中, 用于建树的病毒样本的选择及已知病毒的宿主信息等都对数据库的完整性和准确性提出了很高的要求。如果待检病毒与已知病毒存在一定的差异, 系统发育关系的研究方法仅能做出可能性的推测。因此, 构建一个相对完善的宿主-病毒库对于病毒溯源具有重要意义。

3 利用基因组学方法构建完善的病毒基因数据库及“宿主-病毒”关联信息库

测序技术的发展促进了大量病毒序列信息的获取。在国家生物技术信息中心——病毒数据库(National Center for Biotechnology Information—Virus, NCBI—Virus)中已收录病毒全核酸序列

(all-nucleotides) 300多万条, 全基因组序列(complete-nucleotides) 140多万条, 其中可侵染人类的病毒全基因组序列3.6万多条。相比于自然界广泛的病毒种类, 这只是很少一部分。

病毒-宿主数据库(Virus-Host Database, Virus-Host DB)和病毒病原体数据库和分析资源(Virus Pathogen Database and Analysis Resource, ViPR)整合RefSeq、GenBank、UniProt、ViralZone等多个数据库, 对病毒信息进行了分类汇总。Virus-Host DB针对的是已知宿主信息的病毒种类, 约11,000多种(序列14,000多条), 根据宿主种类进行分类(包括动物、植物、其他真核生物以及细菌、古生菌), 提供病毒全基因组与其宿主信息的分类链接, 其中人类病毒1,300多种。ViPR则是对人类病毒病原体及其相关病毒进行汇总, 包含20多个病毒科, 对每个病毒科的基因组数量以及蛋白数量、菌株数量和未知宿主比例等进行整理与统计(表2)。根据ViPR的统计, 人类病毒病原体及其相关病毒中存在相当比例的未知宿主信息的病毒种类。除上述病毒库外, 其他很多病毒数据库大都限于某个子集, 比如流感病毒数据库、RNA病毒数据库、LANL出血热病毒数据库、乙型肝炎病毒数据库等(Dong et al, 2017)。虽然Virus-Host DB与ViPR从病毒与宿主关联性的角度对病毒的宿主信息进行汇总, 但是动物源病毒多样性高, 病毒与宿主的互作关系较为复杂, 可能存在多重中间宿主或宿主转移, 为确定病毒与宿主复杂的关联关系带来了巨大挑战。在此背景下, 对病毒多样性及其与宿主的关联信息进行广泛的本底调查势在必行。

近年来, 大规模的物种宏基因组或宏转录组研究取得了一系列重要进展。相关研究针对特定的生境类型或物种类型开展了广谱的高通量测序工作。例如, 通过宏转录组方法对脊椎动物(包括爬行动物、两栖动物等)体内的病毒进行鉴定与系统发育分析, 发现RNA病毒与其脊椎动物宿主之间存在多样化的关联关系(Shi et al, 2018)。部分病毒可侵染广泛的宿主类型, 包括哺乳动物、鸟类、两栖类、爬行动物及鱼类, 如: 黄病毒科(Flaviviridae) (Shi et al, 2016)、肝病病毒(Hepadnaviridae) (Nie et al, 2018)、汉坦病毒(Hantaviruses) (Holmes & Zhang, 2015)等。同时, 也有部分病毒与其宿主存在较为一致的系统发生关系, 提示病毒与宿主间可能存在广泛的协同进

表2 ViPR数据库已收录的病毒信息(数据来源于<https://www.viprbrc.org/brc/home.spg?decorator=vipr>)Table 2 Virus sequences deposited in the ViPR database (Data are from <https://www.viprbrc.org/brc/home.spg?decorator=vipr>)

科 Families	种 Species	基因组 Genomes	病毒株 Strains	蛋白 Proteins	缺乏宿主信息比例 Percent without host identification
Arenaviridae	177	4,532	3,012	6,949	23.23%
Caliciviridae	225	52,189	49,073	96,673	28.16%
Coronaviridae	1,043	34,864	28,823	119,573	22.80%
Filoviridae	16	3,577	3,390	22,038	22.53%
Flaviviridae	367	345,546	261,780	877,286	39.36%
Hantaviridae	304	10,189	6,867	10,603	14.02%
Hepeviridae	33	17,838	15,022	19,203	17.95%
Herpesviridae	782	58,371	45,281	300,180	60.18%
Nairoviridae	38	3,669	1,931	3,553	9.92%
Paramyxoviridae	574	50,355	44,898	67,728	24.22%
Peribunyaviridae	183	5,031	2,434	6,367	22.80%
Phasmaviridae	16	1,106	340	1,114	0.27%
Phenuiviridae	215	5,189	2,934	7,133	13.37%
Picornaviridae	1,038	127,336	116,298	346,846	26.91%
Pneumoviridae	17	37,289	33,516	60,235	30.36%
Poxviridae	283	10,444	7,487	125,948	40.87%
Reoviridae	363	107,566	39,985	108,677	15.93%
Rhabdoviridae	530	33,347	26,510	46,761	22.12%
Togaviridae	60	12,800	10,854	46,764	36.25%
总计 Total	6,264	921,238	700,435	2273,631	—

化关系(Shi et al, 2018)。稳定的宿主关系甚至体现在组织特异性上,例如肝病病毒属的病毒多发于爬行动物、两栖动物、鱼类和哺乳动物的肝脏组织(Shi et al, 2018)。这些结果表明,即使伴随着病毒的快速进化和频繁的宿主转换,脊椎动物中的RNA病毒依然保持着与宿主高度匹配的进化历史。这种进化关系很可能始于海洋,并延续了数亿年之久(Shi et al, 2018)。

病毒与宿主的协同关系不仅仅局限于脊椎动物,近期在昆虫中也揭示了相似的结果。基于大规模昆虫转录组分析的千种昆虫转录组演化(1KITE)研究,从昆虫表达的RNA中分离出其共生病毒的表达序列,其中包含至少20个可能的新病毒属(Kafer et al, 2019)。这一结果不仅加深了对昆虫病毒多样性的认知,还为病毒与昆虫宿主的趋同进化(codivergence)提供了证据。病毒与宿主的协同进化关系提示:宿主的进化信息可为病毒进化及溯源提供重要参考。

因此,构建宿主与病毒关联性的数据库不仅必

要而且可行。从时间和资源的优先级上考虑,可以首先从已知的携带重要人畜共患病致病病毒的动物类群及与人类密切接触的动物类群入手,开展较为全面深入的宏基因组摸底调查,如蝙蝠(Zhou et al, 2018; Cui et al, 2019)等。同时,加强对已有数据的病毒信息的挖掘。当前已经完成或正在展开的大规模基因组、宏基因组、宏转录组方面的研究,涵盖多类群物种,并区分动物的不同组织类型。使用成熟的生物信息学方法,可以从中分离得到宿主与病毒的相关遗传信息,用以补充宿主-病毒基因库。

宏基因组学的方法可直接获取宿主携带的多种病毒信息。然而,在实际鉴定过程中易受取样部位、取样时间和环境传播等多种因素的影响,而不能准确鉴定病毒宿主,进而干扰宿主-病毒关联数据库的准确性。所以,建立常态监测网络,对数据进行长期跟踪,将宿主赋予更广泛的定义,比如从“动物”,进一步外延涵盖“植物”,甚至是“环境”,例如食品、饮用水等。通过试点摸底,建立区域环境监测热点,统计病毒鉴定阳性频率信息,最终形成

全国监测网络,便于在第一时间发现各种致病源。

基于病毒的基因组序列信息可对其可能的宿主进行预测,从而完善宿主-病毒关联数据库。例如基于病毒具有与其宿主相似的密码子使用偏好性的特点,针对病毒基因组序列和各种宿主的基因组序列,通过“机器学习”的方法建立通用模型,可直接预测病毒的动物宿主(Babayan et al, 2018)。然而,传统的“机器学习”方法通常需要宿主基因组信息,其应用受已知宿主物种的参考基因组数量的限制。“深度学习”是“机器学习”的一种新技术,深层神经网络(deep neural network)的方法是“深度学习”的一种,在应用于病毒的宿主预测时,该方法克服了传统“机器学习”方法的弊端,通过分析给定病毒基因组的全部或部分序列来预测病毒宿主,无需依赖于宿主的序列信息(Mock et al, 2020)。病毒宿主预测工具VIDHOP (virus deep learning host prediction)便是基于深层神经网络的方法开发的(Mock et al, 2020)。

另外,基于病毒基因组序列建立其关键蛋白质三维结构模型,使用蛋白功能预测工具(比如Swiss-Model)与已知宿主病毒的同源蛋白进行比较,也可以推测未知病毒的宿主来源(Lu et al, 2020)。比如,对SARS-CoV-2刺突蛋白受体结合域建模,发现其与SARS-CoV非常相似,推测SARS-CoV-2与SARS-CoV均使用ACE2为受体进入细胞,根据相同的致病机制可推测共同的宿主来源(Lu et al, 2020)。

基于模型推测宿主来源需要与先进的实验验证相结合。Letko等(2020)开发了人工病毒颗粒系统用于病毒的宿主确认,将来自不同病毒的受体结合域序列与病毒颗粒系统融合,若该病毒可以与细胞受体结合,那么进入细胞的病毒核酸片段会激发荧光反应。通过荧光反应来判断该受体结合域所对应的病毒是否具有侵入细胞的能力,以及病毒侵入细胞所需的细胞受体。

4 问题与展望

基因组学技术极大丰富了人类认知病毒多样性和宿主关系的方法。野生动物是人类致病病毒的重要中间宿主,它们携带的病毒具有丰富的遗传多样性和宿主关联性。然而,人们对这两个重要方面的认知都极其有限,为病毒宿主溯源带来了巨大挑

战。针对频繁出现的新型病毒,现有的数据积累和研究方法还不能保障迅速、准确地开展宿主溯源。对大部分病毒的宿主判断依然基于推测,COVID-19疫情还在全球持续升级,而SARS-CoV-2的直接宿主尚未确认。除此之外,全球性气候变暖可能带来新的病毒威胁,冻土与冰川的融化增加了古老病毒被释放到环境中的风险。青藏高原冰川中已检测出大量新型病毒(Zhong et al, 2020);而西伯利亚冻土层中复活的巨型病毒仍能感染单细胞的阿米巴原虫(变形虫)(Legendre et al, 2015)。古老病毒的新增威胁进一步提高了对病毒多样性认知的迫切性。

针对目前病毒基因组学信息仍然匮乏的现状,基因组学技术应当被更广泛地运用到病毒基因库建立的基础研究中,宿主-病毒数据库的建立将为新病毒种的鉴定和进化研究提供依据。宿主-病毒数据库的构建必然存在一系列的挑战。比如,目前对病毒与宿主的关联性确认方法学手段仍然很匮乏;另外,对类群选择的合理性和完整性的把握、样本量巨大、样本的分类信息复杂、数据挖掘不充分、分析方法不完善等问题都为数据库的构建增加难度。相关问题的解决将可能成为病毒基因组学研究的重点。

参考文献

- Adams IP, Glover RH, Monger WA, Mumford R, Jackeviciene E, Navalinskiene M, Samuitiene M, Boonham N (2009) Next-generation sequencing and metagenomic analysis: A universal diagnostic tool in plant virology. *Molecular Plant Pathology*, 10, 537–545.
- Almazan F, Sola I, Zuniga S, Marquez-Jurado S, Morales L, Becares M, Enjuanes L (2014) Coronavirus reverse genetic systems: Infectious clones and replicons. *Virus Research*, 189, 262–270.
- Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF (2020) The proximal origin of SARS-CoV-2. *Nature Medicine*, 26, 450–452.
- Babayan SA, Orton RJ, Streicker DG (2018) Predicting reservoir hosts and arthropod vectors from evolutionary signatures in RNA virus genomes. *Science*, 362, 577–580.
- Bouckaert R, Heled J, Kuhnert D, Vaughan T, Wu CH, Xie D, Suchard MA, Rambaut A, Drummond AJ (2014) BEAST 2: A software platform for Bayesian evolutionary analysis. *PLoS Computational Biology*, 10, e1003537.
- Briese T, Paweska JT, McMullan LK, Hutchison SK, Street C, Palacios G, Khristova ML, Weyer J, Swanepoel R, Egholm M, Nichol ST, Lipkin WI (2009) Genetic detection and characterization of Lujo virus, a new hemorrhagic

- fever-associated arenavirus from southern Africa. *PLoS Pathogens*, 5, e1000455.
- Cui J, Li F, Shi ZL (2019) Origin and evolution of pathogenic coronaviruses. *Nature Reviews Microbiology*, 17, 181–192.
- Dong R, Zheng H, Tian K, Yau SC, Mao WG, Yu WP, Yin CC, Yu CL, He RL, Yang J, Yau SS (2017) Virus database and online inquiry system based on natural vectors. *Evolutionary Bioinformatics*, 13, 1–7.
- Forster P, Forster L, Renfrew C, Forster M (2020) Phylogenetic network analysis of SARS-CoV-2 genomes. *Proceedings of the National Academy of Sciences, USA*, 117, 9241–9243.
- Frias-De-Diego A, Jara M, Escobar LE (2019) Papillomavirus in wildlife. *Frontiers in Ecology and Evolution*, 7, 406.
- Ge XY, Li JL, Yang XL, Chmura AA, Zhu G, Epstein JH, Mazet JK, Hu B, Zhang W, Peng C, Zhang YJ, Luo CM, Tan B, Wang N, Zhu Y, Crameri G, Zhang SY, Wang LF, Daszak P, Shi ZL (2013) Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. *Nature*, 503, 535–538.
- Gorbalenya AE, Gulyaeva AA, Lauber C, Sidorov IA, Leontovich AM, Penzar D, Samborskiy DV, Baker SC, Baric RS, de Groot RJ, Drosten C, Haagmans BL, Neuman BW, Perlman S, Poon LLM, Sola I, Ziebuhr J (2020) The species severe acute respiratory syndrome-related coronavirus: Classifying 2019-nCoV and naming it SARS-CoV-2. *Nature Microbiology*, 5, 536–544.
- Gusfield D, Eddhu S, Langley C (2004) Optimal efficient reconstruction of phylogenetic networks with constrained recombination. *Journal of Bioinformatics and Computational Biology*, 2, 173–213.
- Hadidi A, Flores R, Candresse T, Barba M (2016) Next-generation sequencing and genome editing in plant virology. *Frontiers in Microbiology*, 7, 1325.
- Hemida MG, Chu DK, Poon LL, Perera RA, Alhammedi MA, Ng HY, Siu LY, Guan Y, Alnaeem A, Peiris M (2014) MERS coronavirus in dromedary camel herd, Saudi Arabia. *Emerging Infectious Diseases*, 20, 1231–1234.
- Holmes EC, Zhang YZ (2015) The evolution and emergence of hantaviruses. *Current Opinion in Virology*, 10, 27–33.
- Hon CC, Lam TY, Shi ZL, Drummond AJ, Yip CW, Zeng F, Lam PY, Leung FC (2008) Evidence of the recombinant origin of a bat severe acute respiratory syndrome (SARS)-like coronavirus and its implications on the direct ancestor of SARS coronavirus. *Journal of Virology*, 82, 1819–1826.
- Hu B, Zeng LP, Yang XL, Ge XY, Zhang W, Li B, Xie JZ, Shen XR, Zhang YZ, Wang N, Luo DS, Zheng XS, Wang MN, Daszak P, Wang LF, Cui J, Shi ZL (2017) Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus. *PLoS Pathogens*, 13, e1006698.
- Hu YF, Yang F, Dong J, Yang J, Zhang T, Sun LL, Jin Q (2011) Direct pathogen detection from swab samples using a new high-throughput sequencing technology. *Clinical Microbiology and Infection*, 17, 241–244.
- Huson DH, Kloepper TH (2005) Computing recombination networks from binary sequences. *Bioinformatics*, 21(Suppl. 2), 159–165.
- Kafer S, Paraskevopoulou S, Zirkel F, Wieseke N, Donath A, Petersen M, Jones TC, Liu S, Zhou X, Middendorf M, Junglen S, Misof B, Drosten C (2019) Re-assessing the diversity of negative strand RNA viruses in insects. *PLoS Pathogens*, 15, e1008224.
- Kunin V, Goldovsky L, Darzentas N, Ouzounis CA (2005) The net of life: Reconstructing the microbial phylogenetic network. *Genome Research*, 15, 954–959.
- Lam TT, Shum MH, Zhu HC, Tong YG, Ni XB, Liao YS, Wei W, Cheung WY, Li WJ, Li LF, Leung GM, Holmes EC, Hu YL, Guan Y (2020) Identification of 2019-nCoV related coronaviruses in Malayan pangolins in southern China. *bioRxiv*, doi: <https://doi.org/10.1101/2020.02.13.945485>.
- Legendre M, Lartigue A, Bertaux L, Jeudy S, Bartoli J, Lescot M, Alempic JM, Ramus C, Bruley C, Labadie K, Shmakova L, Rivkina E, Coute Y, Abergel C, Claverie JM (2015) In-depth study of *Mollivirus sibericum*, a new 30,000-y-old giant virus infecting *Acanthamoeba*. *Proceedings of the National Academy of Sciences, USA*, 112, e5327–e5335.
- Leigh J, Bryant D (2015) PopART: Full-feature software for haplotype network construction. *Methods in Ecology and Evolution*, 6, 1110–1116.
- Letko M, Marzi A, Munster V (2020) Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses. *Nature Microbiology*, 5, 562–569.
- Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, Wang W, Song H, Huang B, Zhu N, Bi Y, Ma X, Zhan F, Wang L, Hu T, Zhou H, Hu Z, Zhou W, Zhao L, Chen J, Meng Y, Wang J, Lin Y, Yuan J, Xie Z, Ma J, Liu WJ, Wang D, Xu W, Holmes EC, Gao GF, Wu G, Chen W, Shi W, Tan W (2020) Genomic characterisation and epidemiology of 2019 novel coronavirus: Implications for virus origins and receptor binding. *Lancet*, 395, 565–574.
- Luo T, Comas I, Luo D, Lu B, Wu J, Wei L, Yang C, Liu Q, Gan M, Sun G, Shen X, Liu F, Gagneux S, Mei J, Lan R, Wan K, Gao Q (2015) Southern East Asian origin and coexpansion of *Mycobacterium tuberculosis* Beijing family with Han Chinese. *Proceedings of the National Academy of Sciences, USA*, 112, 8136–8141.
- Mavarez J, Salazar CA, Bermingham E, Salcedo C, Jiggins CD, Linares M (2006) Speciation by hybridization in *Heliconius* butterflies. *Nature*, 441, 868–871.
- McBreen K, Lockhart PJ (2006) Reconstructing reticulate evolutionary histories of plants. *Trends in Plant Science*, 11, 398–404.
- Mock F, Viehweger A, Barth E, Marz M (2020) VIDHOP, viral host prediction with deep learning. *bioRxiv*, doi: <https://doi.org/10.1101/575571>.
- Nie FY, Lin XD, Hao ZY, Chen XN, Wang ZX, Wang MR,

- Wu J, Wang HW, Zhao G, Ma RZ, Holmes EC, Zhang YZ (2018) Extensive diversity and evolution of hepadnaviruses in bats in China. *Virology*, 514, 88–97.
- Palacios G, Druce J, Du L, Tran T, Birch C, Briese T, Conlan S, Quan PL, Hui J, Marshall J, Simons JF, Egholm M, Paddock CD, Shieh WJ, Goldsmith CS, Zaki SR, Catton M, Lipkin WI (2008) A new Arenavirus in a cluster of fatal transplant-associated diseases. *New England Journal of Medicine*, 358, 991–998.
- Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, Bore JA, Koundouno R, Dudas G, Mikhail A, Ouedraogo N, Afrough B, Bah A, Baum JH, Becker-Ziaja B, Boettcher JP, Cabeza-Cabrero M, Camino-Sanchez A, Carter LL, Doerrbecker J, Enkirch T, Dorival IGG, Hetzelt N, Hinzmann J, Holm T, Kafetzopoulou LE, Koropogui M, Kosgey A, Kuisma E, Logue CH, Mazzarelli A, Meisel S, Mertens M, Michel J, Ngabo D, Nitzsche K, Pallash E, Patrono LV, Portmann J, Repits JG, Rickett NY, Sachse A, Singethan K, Vitoriano I, Yemanaberhan RL, Zekeng EG, Trina R, Bello A, Sall AA, Faye O, Faye O, Magassouba N, Williams CV, Amburgey V, Winona L, Davis E, Gerlach J, Washington F, Monteil V, Jourdain M, Bererd M, Camara A, Somlare H, Camara A, Gerard M, Bado G, Baillet B, Delaune D, Nebie KY, Diarra A, Savane Y, Pallawo RB, Gutierrez GJ, Milhano N, Roger I, Williams CJ, Yattara F, Lewandowski K, Taylor J, Rachwal P, Turner D, Pollakis G, Hiscox JA, Matthews DA, O'Shea MK, Johnston AM, Wilson D, Hutley E, Smit E, Di Caro A, Woelfel R, Stoecker K, Fleischmann E, Gabriel M, Weller SA, Koivogui L, Diallo B, Keita S, Rambaut A, Formenty P, Gunther S, Carroll MW (2016) Real-time, portable genome sequencing for Ebola surveillance. *Nature*, 530, 228–232.
- Roossinck MJ (2015) Move over, bacteria! Viruses make their mark as mutualistic microbial symbionts. *Journal of Virology*, 89, 6532–6535.
- Shi M, Lin XD, Chen X, Tian JH, Chen LJ, Li K, Wang W, Eden JS, Shen JJ, Liu L, Holmes EC, Zhang YZ (2018) The evolutionary history of vertebrate RNA viruses. *Nature*, 556, 197–202.
- Shi M, Lin XD, Vasilakis N, Tian JH, Li CX, Chen LJ, Eastwood G, Diao XN, Chen MH, Chen X, Qin XC, Widen SG, Wood TG, Tesh RB, Xu J, Holmes EC, Zhang YZ (2016) Divergent viruses discovered in arthropods and vertebrates revise the evolutionary history of the Flaviviridae and related viruses. *Journal of Virology*, 90, 659–669.
- Simmonds P, Adams MJ, Benko M, Breitbart M, Brister JR, Carstens EB, Davison AJ, Delwart E, Gorbalenya AE, Harrach B, Hull R, King AM, Koonin EV, Krupovic M, Kuhn JH, Lefkowitz EJ, Nibert ML, Orton R, Roossinck MJ, Sabanadzovic S, Sullivan MB, Suttle CA, Tesh RB, Van Der Vlugt RA, Varsani A, Zerbini FM (2017) Consensus statement: Virus taxonomy in the age of metagenomics. *Nature Reviews Microbiology*, 15, 161–168.
- Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A (2018) Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution*, 4, vey016.
- Tang P, Chiu C (2010) Metagenomics for the discovery of novel human viruses. *Future Microbiology*, 5, 177–189.
- Tang XL, Wu CC, Li X, Song YH, Yao XM, Wu XK, Duan YZ, Zhang H, Wang YR, Qian ZH, Cui J, Lu J (2020) On the origin and continuing evolution of SARS-CoV-2. *National Science Review*, 7, 1012–1023.
- Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, Hu Y, Tao ZW, Tian JH, Pei YY, Yuan ML, Zhang YL, Dai FH, Liu Y, Wang QM, Zheng JJ, Xu L, Holmes EC, Zhang YZ (2020) A new coronavirus associated with human respiratory disease in China. *Nature*, 579, 265–269.
- Yu WB, Tang GD, Zhang L, Corlett RT (2020) Decoding the evolution and transmissions of the novel pneumonia coronavirus (SARS-CoV-2/HCoV-19) using whole genomic data. *Zoological Research*, 41, 247–257.
- Yu Y, Harris AJ, Blair C, He X (2015) RASP (reconstruct ancestral state in phylogenies): A tool for historical biogeography. *Molecular Phylogenetics and Evolution*, 87, 46–49.
- Yuen LKW, Littlejohn M, Duchêne S, Edwards R, Bukulatjpi S, Binks P, Jackson K, Davies J, Davis JS, Tong SYC, Locarnini S, Townsend J (2019) Tracing ancient human migrations into Sahul using hepatitis B virus genomes. *Molecular Biology and Evolution*, 36, 942–954.
- Zhong ZP, Solonenko NE, Li YF, Gazitua MC, Roux S, Davis ME, Van Etten JL, Mosley-Thompson E, Rich VI, Sullivan MB, Thompson LG (2020) Glacier ice archives fifteen-thousand-year-old viruses. *bioRxiv*, doi: <https://doi.org/10.1101/2020.01.03.894675>.
- Zhou P, Fan H, Lan T, Yang XL, Shi WF, Zhang W, Zhu Y, Zhang YW, Xie QM, Mani S, Zheng XS, Li B, Li JM, Guo H, Pei GQ, An XP, Chen JW, Zhou L, Mai KJ, Wu ZX, Li D, Anderson DE, Zhang LB, Li SY, Mi ZQ, He TT, Cong F, Guo PJ, Huang R, Luo Y, Liu XL, Chen J, Huang Y, Sun Q, Zhang XL, Wang YY, Xing SZ, Chen YS, Sun Y, Li J, Daszak P, Wang LF, Shi ZL, Tong YG, Ma JY (2018) Fatal swine acute diarrhoea syndrome caused by an HKU2-related coronavirus of bat origin. *Nature*, 556, 255–258.
- Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, Si HR, Zhu Y, Li B, Huang CL, Chen HD, Chen J, Luo Y, Guo H, Jiang RD, Liu MQ, Chen Y, Shen XR, Wang X, Zheng XS, Zhao K, Chen QJ, Deng F, Liu LL, Yan B, Zhan FX, Wang YY, Xiao GF, Shi ZL (2020) A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, 579, 270–273.

(责任编辑: 吴永杰 责任编辑: 周玉荣)