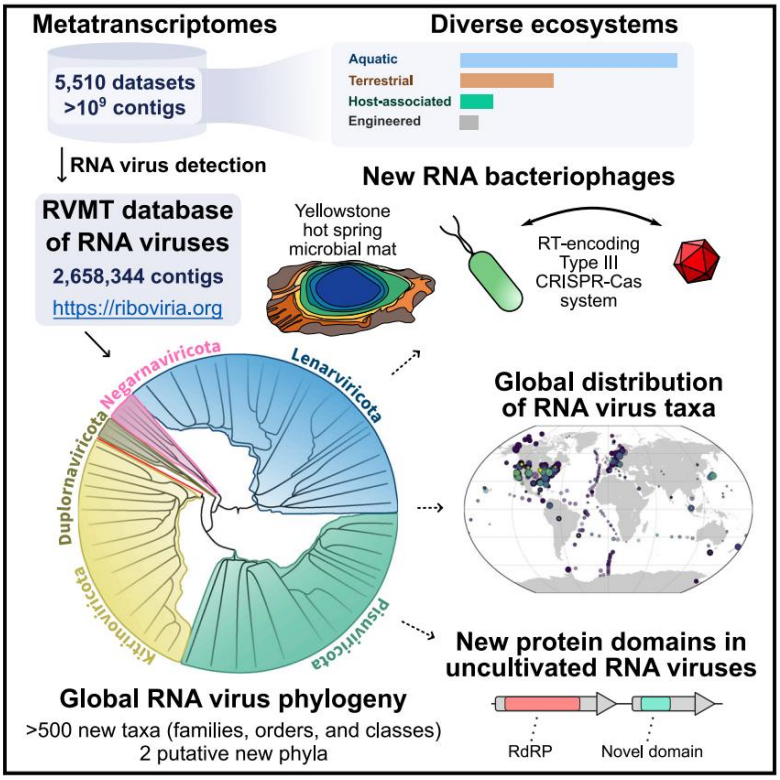
Cell

资源

全球RNA病毒组的扩展揭示了噬菌体的多样化分支

图形概要



亮点d宏转录

组挖掘揭示了 RNA 病毒多样性的重大扩展

d一个假定的新 RNA 噬菌体门编码
不同的裂解蛋白

d Partiti-like RNA 噬菌体被细菌 CRISPR 靶向
系统

d 与病毒-宿主相互作用有关的蛋白质结构域是
已确定

作者

乌里·内里 (Uri Neri)、尤里·I·沃尔夫 (Yuri I. Wolf)、西蒙·鲁克斯 (Simon Roux)、...、尼科斯·C·基皮德斯 (Nikos C. Kyrpides)、尤金·V·库宁 (Eugene V. Koonin)、乌里·戈夫纳 (Uri Gophna)

通讯地址

uri.neri@gmail.com (UN)、 doljav@oregonstate.edu (VVD)、 nckyrpides@lbl.gov (NCK)、 koonin@ncbi.nlm.nih.gov (EVK)、 urigo@tauex.tau.ac .il (UG)

简单来说

对来自数千个不同生态系统的病毒 RNA 基因组的分析极大地扩展了已知的 RNA 病毒多样性，并表明 RNA 噬菌体在全球 RNA 病毒组中所占的比例要大得多。

Neri 等人，2022 年，Cell 185, 4023–4037 2022
年 10 月 13 日a 2022 年作者。爱思唯尔公司出版https://doi.org/10.1016/j.cell.2022.08.023

Cell

CellPress
开放获取

资源

全球RNA病毒组的扩展揭示噬菌体的不同分支

Uri Neri,1,12,* Yuri I. Wolf,2 Simon Roux,3 Antonio Pedro Camargo,3 Benjamin Lee,2,4 Darius Kazlauskas,5 I. Min Chen,3 Natalia Ivanova,3 Lisa Zeigler Allen,6,7 David Paez-Espino,3 Donald A. Bryant,8 Devaki Bhaya,9 RNA 病毒发现联盟,Mart Krupovic,10 Valerian V. Dolja,2,11,* Nikos C. Kyrpides,3,* Eugene V. Koonin,2,*和 Uri Gophna1,* 1特拉维夫大学 Shmunis 生物医学和癌症研究学院,以色列特拉维夫 6997801

2国家生物技术信息中心、国家医学图书馆、国立卫生研究院,贝塞斯达,MD 20894,美国

3劳伦斯伯克利国家实验室能源部联合基因组研究所,伯克利,CA 94720,美国

4牛津大学纳菲尔德医学系,Oxford OX3 7BN,英国

5维尔纽斯大学生命科学中心生物技术研究所,Sauletekio av。 7, 维尔纽斯 10257, 立陶宛6微生物与环境基因组学, J. Craig Venter Institute, La Jolla, CA, USA

7斯克里普斯海洋学研究所海洋生物学研究部,美国加利福尼亚州拉霍亚

8宾夕法尼亚州立大学生物化学与分子生物学系,大学公园,PA 16802,美国

9卡内基科学研究所植物生物学系,斯坦福,CA 94305,美国

10 巴黎大学巴斯德研究所11 俄勒冈州立大学植 ，CNRS UMR 6047,古菌病毒学单位,75015 巴黎,法国

物学和植物病理学系,科瓦利斯,OR 97331,美国

12引线触点

*通讯地址： uri.neri@gmail.com （联合国）， doljav@oregonstate.edu (VVD), nckyrpides@lbl.gov （NCK）， koonin@ncbi.nlm.nih.gov （EVK）， urigo@tauex.tau.ac.il （UG）

https://doi.org/10.1016/j.cell.2022.08.023

概括

高通量 RNA 测序为探索地球 RNA 病毒组提供了广阔的机会。采矿 5,150
不同的宏转录组发现了超过 250 万个 RNA 病毒重叠群。分析 >330,000 个 RNA 依赖性
RNA 聚合酶 (RdRP) 表明,这种扩增相当于已知 RNA 病毒多样性增加了 5 倍。基因内容分析揭示了以前在 RNA 病毒中未发现的多个蛋白质结构域
并涉及病毒与宿主的相互作用。扩展的 RdRP 系统发育支持五个已建立的门的单系性,并揭示了两个假定的附加噬菌体门和许多假定的附加噬菌体门
课程和订单。由细菌和相关真核病毒组成的 Lenarviricota 门急剧扩大,目前占 RNA 病毒组的三分之一。 CRISPR 间隔区匹配的鉴定和

溶菌蛋白表明,先前与相关的小核糖核酸病毒和部分病毒的亚群
与真核生物一起感染原核宿主。

介绍

病毒是活生物体的专性细胞内寄生虫，
被认为是地球上数量最多的生物实体
（穆什吉安,2020）。从历史上看,只有病毒和模型细菌病毒才会引起人类、牲畜和农作物的疾病。
（噬菌体)已被详细研究。最近,由于基因组测序和宏基因组学的进步,以前未曾预料到的 DNA 病毒多样性已经被发现（Call 等人，

2021 年； Roux 等人,2021）。认识宏基因组学在病毒中的作用
发现,国际病毒分类委员会
（ICTV)批准在此基础上正式承认新病毒分类群
宏基因组序列分析（Simmonds et al., 2017）。
与DNA病毒相比,人们对RNA病毒在微生物生态系统中的多样性和作用知之甚少。
最近，
然而,宏转录组调查（大量 RNA 测序）
整个微生物群落)发现了大量的
以前未检测到的 RNA 病毒（Krishnamurthy 等,2016；
齐格勒·艾伦等人,2017;多利亚和库宁,2018）。尤其，
无脊椎动物转录组分析结果使
已知 RNA 病毒的数量（Shi et al., 2016）,然后是
通过RNA序列分析进一步扩增2倍
在元病毒组中（亚细胞大小部分的测序）
来自单个位点,这意味着一个巨大的、几乎没有采样的全球 RNA 病毒组(Wolf et al., 2020)。对 RNA 病毒组的其他尝试包括
真菌转录组分析（Sutela et al., 2020）、各类土壤的元转录组分析（Starr et al., 2019； Wu et al.,
2021）,以及水生环境中 RNA 噬菌体组的扩展（Callanan 等人，2020）。
除了 deltaviruses 之外,所有 RNA 病毒都有一个共同的特征
蛋白质,RNA 依赖性 RNA 聚合酶 (RdRP) (Koonin
等人,2020）。因此,RNA病毒多样性和进化的研究取决于RdRP的检测和分析。虽然到期
对于 RdRP 的极端序列分歧,置信度
系统发育树最深分支中的分支较低,确定了五个分离良好的主要分支（Wolf 等人,2018； Holmes

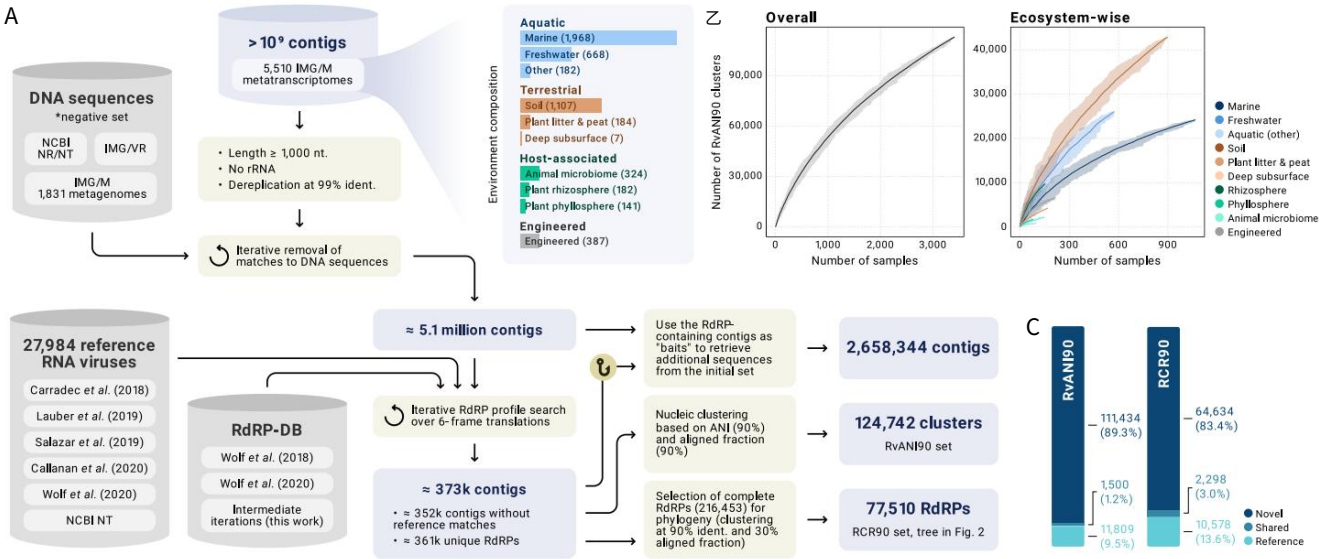


图 1. RNA 病毒发现流程(A) RNA 病毒发现流程。

(B) RvANI90 稀疏曲线:独特簇的累积作为分析样本数量的函数（黄金字段 - ITS.PID）。这些值是通过引导获得的;半不透明部分代表 25 个随机子采样中测量的独特 RvANI90 簇的范围。中心线代表 25 个随机样本的平均值。颜色表示环境类型（右图）。

(C) RCR90 簇（左)和 RvANI90（右)的数量,其成员要么完全是“参考”（仅来自“参考集”的重叠群）,要么是“新颖”（仅在分析中识别)元转录组）,或“共享”（包含每种类型的成员）。

另请参见图S1。

和 Duche ne,2019）,随后被认为是核病毒领域内Orthornavira界的门（国际病毒分类委员会执行委员会,2020； Koonin 等,2020）。

显然,对来自不同栖息地和宿主的 RNA 病毒基因组进行广泛普查对于了解 RNA 病毒进化至关重要。在这里,我们从不同环境中挖掘 5,150 个宏转录组,在种和属之间的粒度水平上将 RNA 病毒多样性从 13,282 个不同簇扩展至 124,873 个不同簇。我们确定了两个候选的附加门和许多暂定的纲、目和科。其中包括可能感染细菌的未报告谱系。此外,我们报告了多个意想不到的蛋白质结构域,其中一些可能会对抗抗病毒防御。

结果

多种RNA病毒的鉴定

在这里,我们设计了一个用于敏
感 RNA 病毒检测的计算管道,适用于分析数千个元转录组（图1;参见STAR 方法）。简而言之,该流程首先通过将宏转录组重叠群与一组不同的 DNA 基因组和宏基因组进行比较,过滤出可能由 DNA 实体编码的序列。随后,迭代地搜索大大减少的序列集（<初始集的 1%)中的 RdRP,并将置信匹配视为假定的 RNA 病毒（参见STAR 方法）。所查询的 5,150 个元转录组中有 3,598 个包含一个或多个编码 RdRP 的重叠群进一步分析的完整性（参见STAR 方法）。然后,我们使用 RdRP 编码重叠群作为诱饵来识别与 RdRP 编码重叠群（包括 RdRP 区域外部)具有高度核酸相似性的其他宏转录组重叠群。

总共鉴定了 2,658,344 个 RNA 病毒重叠群,并补充了来自已发表来源的 27,984 个序列（图1A）。其中,348,762 个重叠群代表长度为 R1 kbp 的去重复、非冗余序列集。这些被分为 124,743 个簇,平均核苷酸同一性为 90%（RNA 病毒 ANI90 簇 [以下简称 RvANI90]）,其中只有 13,308 个簇（10.7%)包含至少一个先前已知的序列,这相当于大约 9 倍的扩展全球 RNA 病毒组,多样性水平为 ANI90。

RNA病毒序列簇按大小呈幂律分布,以小簇为主,大簇长尾,最大的簇包括429个重叠群（图S1）。根据累积曲线,在RvANI90水平上评估的RNA病毒的全球多样性没有显示出饱和的迹象（图1B）,其中土壤环境中的丰富度特别高（图1B）。大约 5.8% 的 RdRP 编码重叠群显示出利用替代遗传密码的证据（图2）,大约 0.5% 显示出 RdRP 内保守基序的改组（结构域排列） （图2）。

RdRP 系统发育和 RNA 病毒多样性的主要扩展

为了构建全球 RNA 病毒系统发育史,我们首先收集了全长 RdRP 核心结构域序列,并以 90% 氨基酸同一性阈值对它们进行聚类,得到 77,510 个

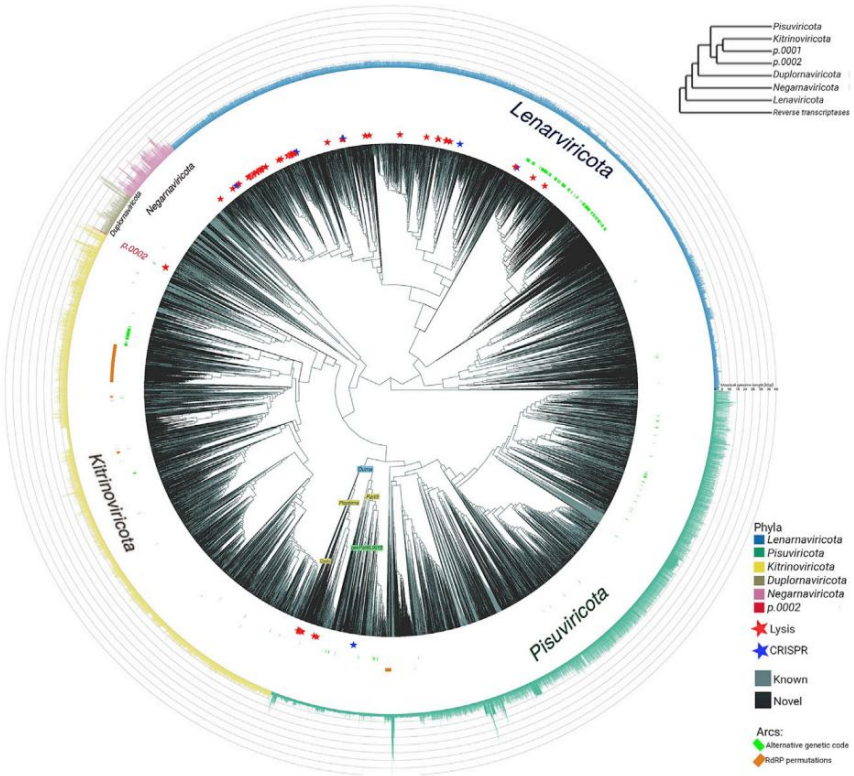


图 2. 全局 RNA 病毒圈的系统发育重建 使用逆转录酶作为外群的超计量 RdRP 树,并使用
ggtree 和 ggtreeExtra 进行可视化（Xu 等人,2021； Yu 等人,2018）。分支被涂成黑色,除非它们的任何后代包含至少一个来自“参考集”（青色)的序列。与星号对齐的尖端表明原核宿主的证据 蓝色为 CRISPR 间隔区匹配,红色为溶菌域。绿色弧线表示具有

R50% 的序列中存在替代遗传密码。
橙色弧线表示 RdRP 的 R50% 中具有基序排列的进化枝。 5 个已建立的门和建议的候选门 p.0002 在文本和最外环的条形图中进行了颜色编码,它代表了每个 RCR90 簇（即树尖)观察到的最大基因组长度。主要分类单元直接标记在树上。树的附加可视化可用

能够在项目的 Zenodo 存储库中（请参阅数据和代码可用性）。

另请参见图S2。

尽管共识树将 Picobir-naviridae和Cystoviridae自信地置于Pisuviricota内（见下文） ,但它们脱离了各自的门。

当子采样树减少到最低共同祖先时

代表（RCR90套）。即使减少到 RCR90 粒度,该集合仍然太大且多样化,无法直接使用先进的最大似然系统发育方法进行多序列比对和系统发育分析。因此,我们采用了迭代过程,其中使用序列簇比对的一致性比对来重建树（参见STAR 方法）。

由此产生的 RdRP 树包含 77,520 个代表性序列（77,510 个 RCR90 序列和 10 个逆转录酶 [RT],作为外群包括在内;图2）。尽管出现了如此巨大的扩张,之前建立的 5 个门（Wolf 等人,2018)在很大程度上仍然是单系的。此外,该树还包括 Kitrinoviri-cota门基部以下的两个类群,对此进行了详细分析（见下文）。

RdRP 树中主要分支的单系性,特别是 5 个门,通过二次抽样进行了验证。病毒科的代表被重复随机抽样,根据每个样本的多重比对重建系统发育,追踪门分支的位置,并计算其单系的定量测量（参见STAR方法）。在大多数样本中,5 个门基本上保持单系（图S2A）。

倾向于打破门级单系性的序列形成了一个严重偏向的子集,其中Flasuviricetes是最常见的“罪犯”。在这项工作中， Flasuviricetes被放置在Pisuviricota 内部,而在之前的分析中,它是Kitrinoviricota的基础分支。然而,黄病毒在二次采样树中的不一致位置表明它们的系统发育位置仍然不确定。呼肠孤病毒科、 Picobirna-viridae、囊病毒科和几个候选科也经常

对于五个门中的每一个,发现最深的分支顺序是稳健的， Pisuviricota和Kitrinoviricota在共有树中形成冠群,而Lenarviricota和Negarnaviri-cota占据基础位置（图2,右上插图）。正如之前的分析（Wolf 等人,2018)一样,当树被 RT 扎根时， Orthornavirae内最深的分支是Lenarvir-icotal门,其中包括 Leviviruses（正义 RNA 噬菌体； Allassoviricetes 类)及其表观真核病毒、线粒体病毒(Howeltoviricetes)、纳尔病毒(Amabilivirecetes)和肉毒病毒(Miaviricetes) 的直系后代。

尽管明确验证这种分支顺序可能不可行,但Lenarviricota的这一位置在生物学上是合理的,将Orthornavirae的起源置于细菌域中。相比之下， Negarnaviricota的深入放置是出乎意料的,因为 ssRNA 病毒几乎完全从动物和植物中分离出来。Negarnaviricota 的位置可能反映了一个古老的起源,但更可能的是,这是一种系统发育人工产物,可能是由 Negarnaviricota 基础的进化加速引起的。

当前 RdRP 系统发育树和之前报道的树（Wolf 等人,2020)的系统发育深度的比较反映了通过总分支长度（TBL)测量的全局 RNA 病毒组的大约 5 倍扩展。为了将 RdRP 系统发育转化为暂定的分类方案,我们开发了一种半定量方法,用于根据邻近的成熟分类群将分类等级分配给未分类的节点（参见STAR 方法）。分类群被指定为等级,并分别以p、 c、 o、 f 和g为前缀,分别表示门、纲、目、科和属,后跟拟议分类群的序号

表 1. 全局 RNA 病毒组的扩展

秩	数量 已知类群	更新数量 类群的	折叠 增加
RvANI90 簇	13,282	124,873	9.4
RCR90 簇	12,862	77,510	6.0
家庭	98	第489章	4.9
命令	26	121	4.7
班级	19	93	4.9
门	5	7	1.4

那个排名。与先前描述的分类单元相关的分类单元
以“base”结尾,例如， f.0127.base-Noda是第 127 个
RdRP 树中Nodaviridae的基础新科（表S1）。
与之前的结果相比,这种方法导致门以下所有等级的多样性扩大了大约 5 倍。

最新的 RNA 病毒组分析（Wolf 等人,2020;表1）。然而，
必须强调的是,这一估计是在没有
考虑到自进行该分析以来发表的两项大规模 RNA 病毒调查结果（参见限制

研究部分）（Edgar 等人,2022； Zayed 等人,2022）。
当按门分类时,最大的扩展
排名位于Lenarviricota 之内,其次是Kitrinoviricota和
Pisuviricota。相比之下,只有少数类群被添加到Duplor-naviricota和
Negarnaviricota中（图2;表S1）。
除了 RdRP 系统发育树中反映的扩展之外,一些 RNA 病毒（形成了 39,000 个重叠群）
在这项工作中通过基于 RdRP 的配置文件搜索识别出的 24,742 个 RvANI90 簇被从系统发育中丢弃
分析作为边界和核心的一些主题
无法可靠地识别 RdRP 域。

假定的附加门和类

由于目前 ICTV 没有关于 RNA 病毒门和类别形成的官方指导,我们选择了标准

类似于用于浅等级的方法（参见STAR 方法），
也就是说,要形成一个门或类,需要一个群体进行分支
在现有的门或类之外。这里确定的两个最不同的分支位于 RdRP 系统发育中Ki-trinoviricota基部下方,原则上可以

包含在该门的扩展版本中。其中第一个
深分支， p.0001,仅包含 3 个 RCR90 簇,并且
因此没有进一步分析。第二个， p.0002，
拥有明显的特征,看起来与
候选门指定而不是Kitrino-viricota 的扩展。这个假定的门由来自 30 个的 234 个重叠群组成
RCR90簇,编码10个ORF的最完整簇
平均长度约为 12 kb。除 RdRP 外,只有其中一项
ORF（在两个暂定家族之一中保守）
p.0002)与已知的蛋白质结构域具有显着的相似性，
专门针对 M15 或 M35 锌金属肽酶家族
涉及细胞裂解（见下文）。 p.0002基因组中的 ORF
间隔紧密且前面有核糖体结合基序
(Shine Dalgarno [SD]) 参与原核翻译起始
（图3A）。总的来说， p.0002似乎包括
噬菌体,支持该组的门指定为
所有分离的Kitrinoviricota成员都会感染真核生物。
另一个高度分化的候选 RNA 噬菌体门是
RvANI90_0011770,系统发育工作中遗漏的病毒簇之一,因为它们扭曲了 RdRP 比对（因此,没有p
指定）。所有 RvANI90_0011770 成员均来自 27
不同的活性污泥样品,其中这 55 个 con-tig 中最大的长度为 10-12-kb,编码 7-9 个紧密排列的 ORF
没有保守的 SD 基序。与 p.0002 类似,唯一公认的
蛋白质结构域包括 RdRP 和预测的裂解酶
（见下文）。
班级水平多样性大幅增加（参见STAR
方法)在 5 个已建立的门中的 4 个中观察到,包括
Lenarviricota已知 14 类与 4 类， 4 类中已知 18 类
Pisuviricota已知有20 个类,而 Kitrinoviricota 已知有 3 个类，
Negarnaviricota已知有 18 个类,而已知有 6 个类。在Duplornavir-icota 中,仅发现了两个候选类水平进化枝
除了两个公认的类别之外。总体而言， Or-thornavirae的 5 个门包含 91 个类,而之前有 19 个类
较上年增加 489 个家庭
公认的 98 个（表1;表S1）。其中一些额外的候选分类单元包括先前报道的不同病毒，因此
远远避开了安置并且缺乏 ICTV 指定。

相关RNA病毒范围的重大扩展有细菌

迄今为止,大多数RNA病毒都与真核生物有关。
宿主,已知只有两个群体可以感染细菌,即利维病毒（Leviviricetes)和囊病毒（Vidaverviricetes）。直到最近，
左旋病毒,特别是囊病毒,包括少量
宿主范围窄的病毒。在这里,我们扩展了囊病毒科
从 8 个已发表的 RCR90 簇到 132 个 RCR90 簇的多样性。 Levivirus 多样性最近得到了扩展（Callanan
et al., 2020)到 1,940 个 RCR90 簇,进一步增加
这里还有 13,512 个 RCR90 集群。
扩展的Lenarviricota门现在占超过
第三个RNA病毒RCR90簇,包括四个最大的
科（图2;表S1），其中第一个和第四个分别是Steitz-viridae和Fiersviridae,是真正的Leviviricetes
噬菌体。第二大科Botourmiaviridae包括
真核病毒似乎是从
与Leviviricetes、无衣壳的纳纳病毒科和线粒体病毒科（RNA 病毒第三大家族)有共同祖先
作为中间体（Koonin 等人,2020）。除了主要的
Lenarviricota的扩展,汇聚的证据表明,以前认为仅感染真核生物的几组病毒重新分配给细菌宿主（图3B）。

现在,噬菌体似乎散布在Pisuviricota 内感染真核生物的噬菌体中。具体来说,囊病毒科,
在当前的 RdRP 系统发育中,从Duplornaviricota迁移到Pisuviricota ,形成了一个强有力的支持分支
小核糖核酸病毒和部分病毒（双链 RNA
[dsRNA] 家族嵌入 +ssRNA 病毒之中
[图2))。在这个杜纳病毒目中,几个分支显示
SD 基序在50 个非翻译区 (UTR) 中出乎意料地保守,表明这些病毒感染细菌（Ba-hiri Elitzur 等人,2021 年； Hockenberry 等人,2018 年）。这些假定的

噬菌体包括Picobirnaviridae 的成员,其存在

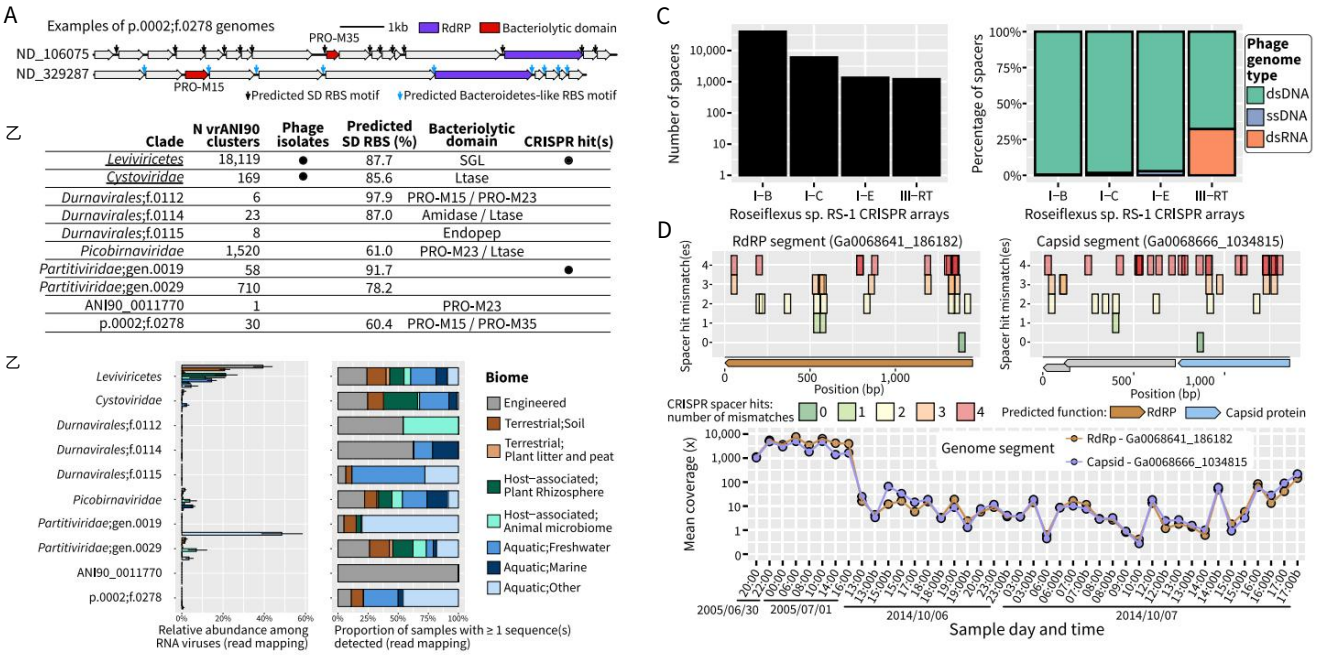


图 3. 原核 RNA 病毒的多样性和丰度(A) 来自拟定门 p.0002 的暂定科f.0278的病毒基因组图谱。 ORF 根据功能注释进行着色,预测的 SD 基序用彩色箭头表示。

(B) 已识别（下划线)和预测的原核 RNA 病毒概述。对于每个组,都指出了支持其与原核宿主关联的证据类型。在排除重叠群边缘的预测基因后,如果预测 ORF 的 R50% 与 SD 基序相关,则认为进化枝可能主要由噬菌体组成。 Ltase、裂解性转糖基酶、溶菌酶超家族折叠酶； SGL，“单基因裂解”（细胞壁合成抑制剂）； PRO-M15,Zn-DD-羧肽酶（sensu PF08291.13）； PRO-M35,M35家族锌金属内肽酶； PRO-M23,M23 家族金属肽酶;酰胺酶,N-乙酰胞壁酰-L-丙氨酸酰胺酶； Endopep,L-丙氨酰-D-谷氨酸内肽酶。

(C) Roseiflexus sp. 的 CRISPR 间隔区景观。黄石温泉中的 RS-1,包括与genPartiti.0019基因组匹配的间隔区。左侧面板显示每种 Roseiflexus sp 类型所识别的间隔物总数。 RS-1 CRISPR 阵列（见图 S3 ）。右图显示了噬菌体类型（dsDNA、ssDNA 或 RNA），针对每种 CRISPR 类型,鉴定了 CRISPR 间隔区的命中。

(D) 来自genPartiti.0019噬菌体的预测 RdRP 和衣壳编码片段对的示例。上图:CRISPR 间隔区匹配显示在每个片段的基因组图谱旁边。不匹配的数量显示在 y 轴上,命中的位置显示在 x 轴上。底部面板显示了宏转录组时间序列中两个片段的相对丰度。

(E) 生物群落中不同原核 RNA 病毒组的相对丰度。仅考虑由包含至少 10 个原核 RNA 病毒的原核序列（“P 主导”)主导的数据集。右图显示了每组生物群落分布的细分,该数据集是根据每个环境 50 个样本的随机子样本组成的平衡数据集计算得出的（随机子采样进行了 100 次,并绘制了平均值）。

另请参见图 S3 和 S4。

SD 基序之前已被注意到（Boros 等人,2018； Krishna-murthy 和 Wang，2018）,以及两个 Cysto 样家族（f.0114.base-Cysto和f.0112.base-Cysto)和另外两个Partitiviridae内的属（genPartiti.0029、genPartiti.0019.-base-Deltapartitivirus）（表 S2;图3B）。

某些细菌关联的另一个证据

已确定的病毒组是溶菌蛋白的保守出现（图3B）。许多 dsDNA 噬菌体和 dsRNA 胞囊病毒编码可降解细菌肽聚糖的裂解酶（内溶素）（Cahill 和 Young,2019）。相比之下，左旋病毒通过称为单基因裂解（Sgl)的小蛋白质抑制肽聚糖合成来诱导宿主裂解（Cahill 和 Young,2019）。 Leviviruses sgl通常重叠或嵌套在其他基因内（Chamakura 和 Young,2020）。

在这里,我们使用此类裂解域的集合来检测可能感染细菌的宏转录组病毒基因组（参见 STAR 方法）（图3B）。该搜索产生了 546 个与裂解蛋白谱显着匹配的结果,大部分来自 Leviviricetes (469) 和囊病毒科(17)。尽管已知的胞囊病毒编码溶菌酶超家族 (SF) 折叠的裂解性转糖基酶(Dessau et al., 2012),但本文鉴定的一些胞囊样家族编码其他肽聚糖消化酶。具体来说,一些f.0114.base-Cysto病毒编码 N-乙酰胞壁酰基-L-丙氨酸酰胺酶,而f.0112.base-Cysto病毒编码 M15 或 M23 家族的金属肽酶（表S2）,这两种酶都常见于 dsDNA 噬菌体中已知可以裂解交联肽的键（Oli-veira 等人,2013）。一些f.0112.base-Cysto病毒还编码可能进一步诱导宿主裂解的脂肪酶。最后， f.0115.base-Cysto病毒编码一种 L-丙氨酰-D-谷氨酸内肽酶,通常在 dsDNA 噬菌体中充当内溶素（Cahill 和 Young，2019； Oliveira 等,2013）。

囊病毒中内溶素的这种进化枝特异性分布表明,与 dsDNA 噬菌体一样,裂解基因会受到频繁的非同源替换,可能与宿主范围的变化有关。

另外两组 RNA 病毒被发现编码裂解蛋白,即小核糖核酸病毒和拟定门 p.0002 中的f.0278家族。六种小核糖核酸病毒编码裂解性转糖基酶或 M23 家族金属肽酶。 f.0278的成员编码 M15 或 M35 家族锌金属肽酶（表S2）。 M15 家族酶参与一些 dsDNA 噬菌体(Kutyshenko et al., 2021)和一些 ssDNA 噬菌体(Roux et al., 2012) 的宿主裂解,而 M35 家族酶此前并未与噬菌体出口相关。鉴于这两种酶在f.0278中是相互排斥的,并且相应的基因占据相同的位置,我们建议 M15 和 M35 家族蛋白充当内溶素。 f.0278中 M15 和 M35 蛋白的保守性强烈支持细菌宿主分配。最后,RvANI90_0011770（通过 RdRP 搜索鉴定的推定 RNA 噬菌体门,未包含在当前系统发育中)显示出 M23 家族金属肽酶的类似保守性。

原核宿主分配的最后证据是检测 RNA 病毒和 CRISPR 间隔区之间的匹配。尽管大多数已知的 CRISPR 系统以 DNA 模板为目标,但 III 型 CRISPR 系统的很大一部分编码 RT,可以保护细菌免受 RNA 噬菌体的侵害（Ma-karova 等人,2020; Silas 等人,2017）。我们将所有已识别的 RNA 病毒基因组与 R5000 万间隔区的 IMG 数据库进行了比较（参见STAR 方法）,检测来自 23 个 RvANI90 簇的 161 个 RNA 病毒的间隔区匹配,跨越两个进化枝： Leviviricetes和genPartiti.0019（图3B;表S2）。所有与Leviviricetes病毒的匹配均来自于 IMG 宏基因组衍生的短重叠群,没有可靠的分类信息或相邻的cas基因（表S3）。相比之下,与 genPartiti.0019病毒的匹配与Roseiflexus sp种群特别相关。 RS-1并进行了进一步分析。 Chloroflexi门的这种丝状不产氧光养细菌是蘑菇中微生物垫的主要成员。

room Spring (Davison et al., 2016),从中获得了genPar-titi.0019序列。 Roseiflexus sp.的基因组。 RS-1 包含四个 CRISPR 位点,其中一个亚型 III-B 编码与 Cas1 蛋白融合的 RT（见图S3） （van der Meer 等,2010）。跨 16 个宏基因组编译间隔区,每个 CRISPR 阵列都可以与 z1,000–40,000 个间隔区相关联,但在 RT 编码 III-B 阵列中检测到除了一个间隔区之外的所有间隔区与 genPartiti.0019匹配,这表明这些间隔区是从RNA 模板（图3C）。在跨越 9 年的样本中观察到这些 CRISPR 间隔区匹配,并显示出随时间变化的动态间隔区增益/丢失,表明病毒-宿主关联（图S3）。

由于所有genPartiti.0019重叠群均单独编码 RdRP,而相关的 Partitiviruses 具有分段基因组,其中衣壳和其他蛋白质在单独的片段中编码,因此我们在 Mushroom Spring 宏转录组中搜索了编码相应衣壳蛋白 (CP) 的重叠群。

将匹配与来自Roseiflexus sp . 的 RT 编码 III-B 型阵列的间隔区组合起来。 RS-1、Mushroom Springs DNA 宏基因组中缺乏相应序列,以及与至少一个gen-Partiti.0019 RdRP 编码序列的强相对丰度相关性 (>0.9),我们鉴定了 88 个潜在的衣壳编码重叠群（图3D;表S3),其中 86 个编码的蛋白质与已知的 HMM 图谱具有最佳比对部分病毒衣壳（图S3）。因此， genPartiti.0019成员很可能是感染Roseiflexus sp 的分段 RNA 噬菌体。 RS-1。

有趣的是,在由原核宿主（“P 主导”,见下文)主导的数据集中,大多数潜在的 RNA 噬菌体在广泛的生物群落中被检测到,其中Leviviricetes是迄今为止最丰富的原核 RNA 病毒组,除了一些黄石温泉以genPar-titi.0019为主（图3E）。

RNA 病毒在样本和栖息地之间的差异分布

我们的 RNA 病毒调查覆盖了全球,反映了 RNA 病毒在地球上的普遍性（图4A）。宏基因组研究表明,DNA 病毒的分布是由环境类型和宿主群落组成决定的（Gregory 等,2019； Martinez-Hernandez 等,2017； Roux 等,2016）,并且相同的因素可能决定RNA病毒分布。

对于宏转录组,样本处理方案可能是另一个因素,即是否对总 RNA 进行了测序,或者是否使用了任何特定的预处理（例如通过 Poly(A) 扩增富集 mRNA,或去除 rRNA） （Gann 等人,2021））。在这里,大多数分析的数据集都是 rRNA 耗尽的（67%,图S4）。虽然富含 Poly(A) 的数据集和总 RNA 数据集以真核生物序列为主,但 rRNA 耗尽的数据集主要由原核生物序列组成（图S4）。根据分类学,数据集分为三组：“真核生物 (E) 主导”(811)、“原核生物 (P) 主导”(2,706) 和“混合”(452)非病毒重叠群的组成。大多数 RNA 病毒类别在数据集类型和环境中显示出清晰的分布模式,可能反映了其主要宿主群体的分布（图4B 和 4E）。例如, Leviviricetes在来自工程、根际和土壤生境的以 P 为主的样品中持续富集（图4B）。这意味着 RNA 噬菌体的全球生态分布不均匀,支持了之前的发现（Callanan 等人,2020）。此外,在Lenarviricota 中,主要感染真菌、无脊椎动物和植物的 Miaviricetes与 E 主导和混合数据集相关,而Howeltoviricetes成员（包括线粒体病毒)在所有样本类型中都很常见,但优先在也富含真菌的植物相关数据集中发现。

尽管将特定的真核生物宿主分配给 RNA 病毒是一项具有挑战性的任务,在这项工作中没有解决,但我们怀疑许多检测到的病毒会感染不同的单细胞真核生物,因为它们利用替代遗传密码（见下文）。

假设病毒的广泛宿主分配（植物、动物或真菌)可以扩展到较小的序列差异（小于 10%）,我们仅鉴定了 1,038 个元转录组重叠群,它们与来自病毒的病毒属于同一 RvANI90 簇。 VirusHostDB (MIhara et al., 2016)分配给植物或动物宿主,表明在分析的数据集中感染这些宿主的病毒流行率较低。此外,还可以为 1,038 个元转录组重叠群（属于 6 个科： Tombusviridae、Virgaviridae、Betaflexiviridae、Alphaflexiviridae、Benyviridae和Mayoviri-dae)进行植物特定宿主分配,编码运动蛋白 (MP),使病毒能够通过胞间连丝。

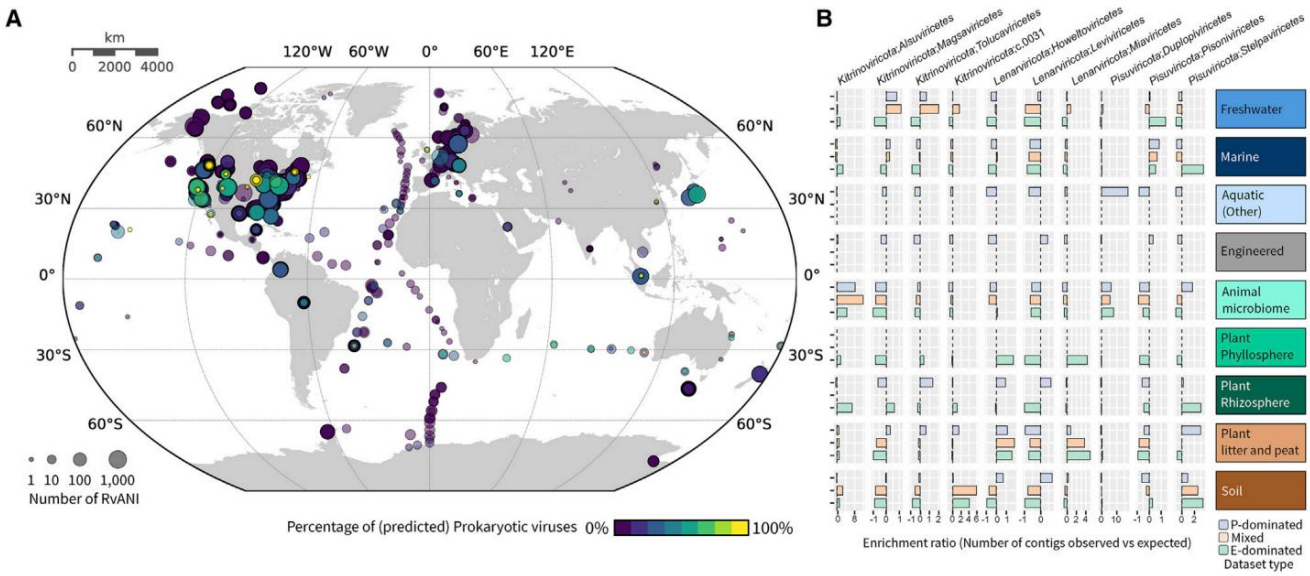


图4. RNA病毒的全球分布

(A) 含有 RNA 病毒的分析样品的位置。对于每个样本,圆圈大小反映了不同 RvANI90 的数量,圆圈颜色表示预测为噬菌体的序列的比例。

(B) 在生态系统类型（y 轴)中检测到的（建议的)RNA 病毒类别（x 轴)的相对比例。考虑基因组总数检测到的每个类别和每个生态系统类型的样本总数,计数表示为与预期相比的富集度假设所有生态系统中所有类别均匀分布的基因组数量。数据集被分为“E-主导”（主要由真核转录本）、“P 主导”（主要由原核转录本组成)和“混合”（见图S4）。仅显示组合的富集检测到至少 20 个带有 R1 RNA 病毒的宏转录组的生态系统和数据集类型（例如“海洋 P 主导”）。另请参见图S4。

RNA病毒基因组的模块化进化

在这里,我们对病毒基因组进行了比较分析
从相关的进化枝中,识别基因组模块化的实例,
例如基因组片段的融合、蛋白质的重排、
和多蛋白的分割。在小核糖核酸病毒目中观察到涉及结构模块的常见基因组重排,其中 CP 在下游或上游编码

基因组复制模块的一部分,作为相同多蛋白的一部分或作为单独的蛋白质（图S5,基因组图谱）。已知
Benyviridae、 Picobirnaviridae和Botourmiaviridae病毒
通常在不同的段上对 CP 和 RdRP 进行编码。
在这里,我们确定了这些家庭的成员,其中 RdRP
和CP在同一网段。

我们检测到多起结构基因模块被非同源对应物置换的情况。例如,虽然

马铃薯病毒科、本尼病毒科和马托病毒科的成员编码
3 个不相关的 CP 并分别形成螺旋丝状、杆状或包膜病毒颗粒,这些病毒附近的一些分支谱系编码预期的单果冻卷 (SJR) CP

形成无包膜的二十面体病毒体。鉴于这个血统
从基础位置来看,SJR CP 可能是所有三种病毒的祖先
组。在f.0226.base-Beny组中,几种病毒编码
SJR 和烟草花叶病毒 (TMV) 样 CP 都可以
预测分别形成二十面体和螺旋衣壳
（图S5）,表明这些病毒可能获得了第二个CP,但保留了祖先的CP。其中一项的扩展
正如之前对长线病毒的描述,CP 似乎很可能
（Dolja 等人,2006 年）。非同源 CP 也被鉴定为
披膜病毒科的基础基因（f.0271.base-Toga和f.0273.base-Toga）,其中典型的
披膜病毒科二十面体形成 CP
被类似 TMV 的 CP 取代,可能形成杆状螺旋
病毒体,表明类似 TMV 的 CP 出现在一个共同的祖先中
属于Hepelivirales和Martellivirales 目。相反,在两个确定的
Virgaviridae重叠群（ND_191857 和 ND_019381）,TMV 样
CP被Kitaviridae的结构蛋白取代。在
f.0268.base-Toga,典型的披膜病毒科结构模块
（包括 CP 和 II 类融合蛋白 [CIIF] 蛋白的基因)被巢病毒的 I 类融合蛋白和 M 蛋白取代
（ND_164660;图5）。膜融合的类似替代
糖蛋白也在Xinmoviridae重叠群中被发现,其中
CIIF蛋白取代了典型的III类融合蛋白,但保留了典型的单阴性病毒核衣壳蛋白。

我们鉴定了几个亚病毒科（无衣壳真菌病毒)的基础病毒群,其编码的 CP 与柔性病毒同源。
螺旋病毒（f.0066.base-Hypo)或二十面体病毒的SJR CP （f.0067.base-Hypo、 f.0068.base-Hypo、 f.0069.base-Hypo）,
表明这些家族的祖先可能是衣壳编码的。类似地,我们鉴定了编码 SJR CP 的
Deltaflexiviridae近缘种
（来自f.0215.base-Deltaflexi 的 ND_196199 和 ND_246366)类似
与 tymoviruses 相比,表明Deltaflexiviridae是进化而来的
是由Tymoviridae的成员转变为真菌宿主后产生的。
SJR CP 在基础谱系中反复出现
几组结构不同的病毒是兼容的
据推测,大多数真核生物 RNA 病毒起源于
编码 RdRP 和 SJR CP 的简单祖先（Koonin
等人,2020）。

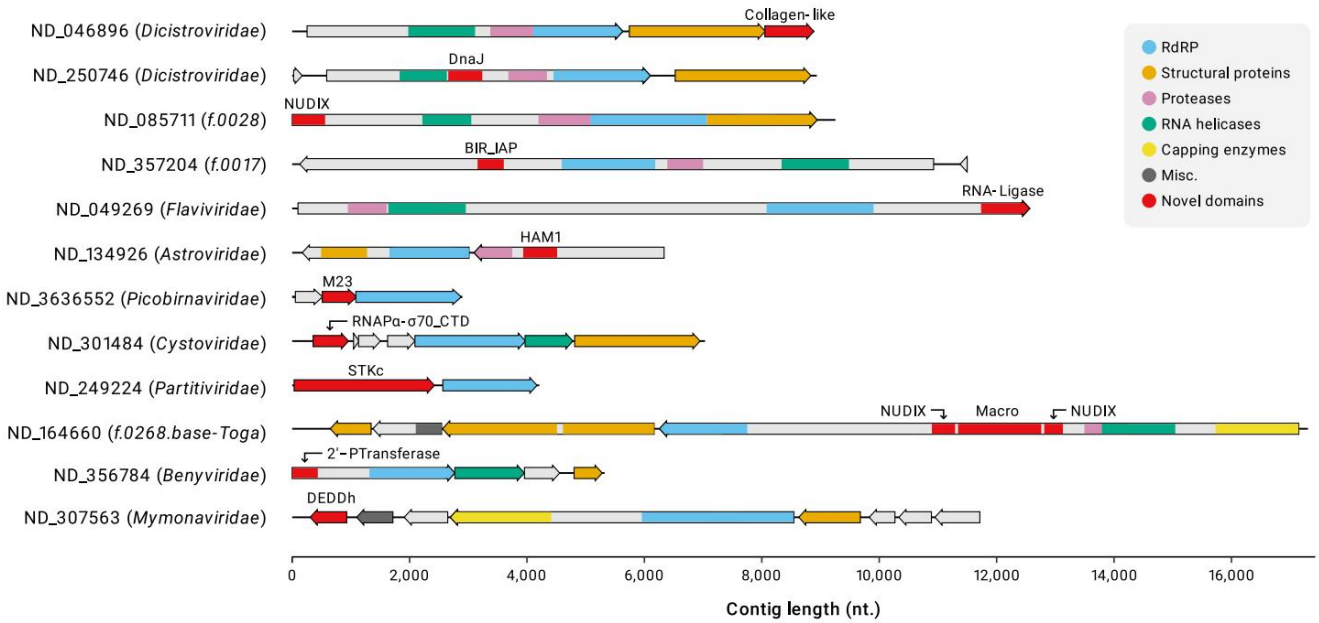


图 5. RNA 病毒中蛋白质结构域的多样性同源结构域显示为相同颜色的框
（参见右侧的图标）。 RNA 病毒中不常见的结构域以红色显示,并标记在相应的框上方。病毒分类群和重叠群标识符标注在每个病毒基因组的左侧。底部的刻度表示核苷酸的长度。缩写:NUDIX，核苷二磷酸-X 水解酶； BIR_IAP,凋亡抑制剂（IAP)的杆状病毒IAP重复（BIR)结构域； HAM1,肌苷三磷酸焦磷酸酶； M15、M23和M34,分别是MEROPS家族M15、M23和M34的肽聚糖消化肽酶； RNAPa-s70_CTD,细菌DNA依赖性RNA聚合酶a亚基的C端结构域与s70因子C端结构域的融合； STKc,丝氨酸/苏氨酸蛋白激酶； 20 -P转移酶,tRNA 20-磷酸转移酶； DEDDh,DEDDh-超家族30 -50核酸外切酶； RdRP,RNA依赖性RNA pol聚合酶;杂项,杂项。

另请参见图S5、S6 和 S7。

最初在Permutotetraviridae和Birnaviridae 中描述,一种独特的重排（称为“结构域排列”)发生在 RdRP 结构域内,其中基序（A、B、C.)的顺序与规范形式不同。此处,RCR90 RdRP 集 (2,241) 的 2.9% 被确定为已排列。我们的分析表明，基序交换是两个类的祖先（图2）， Pisuviricota中的候选类c.0017 （包括 Permutotetraviridae、 Birnaviridae和其他 14 个暂定科[f.0088-f.0101])和候选类c。 Kitrinoviricota中的0032 （涵盖 8 个假定的科[f.0167–f.0174],包括来自阳山组合的许多病毒[Wolf 等人,2020])。除了Pisuviri-cota和Kitrinoviricota之外,我们仅在Botourmiaviridae (Lenarviricota)内检测到一个由 2 个排列的 RCR90 RdRP 组成的小进化枝。

预计干线可以调节病毒与宿主的相互作用并抑制宿主的抗病毒反应。

主要感染脊椎动物的几个Tobaniviridae成员编码细胞因子受体相关的 Janus 激酶 (JAK) TYK2 的同源物 (HHpred p = 100%) (Zimmermann et al., 2018),一旦激活,就会触发宿主免疫反应（Haan 等人,2006）。这些病毒 JAK 缺乏典型 TYK2 的 FERM 和 SH2 结构域,可能通过其假激酶结构域充当细胞 JAK 的显性负性抑制剂。预计编码丝氨酸/苏氨酸激酶的唯一其他 RNA 病毒是部分病毒（图5），尽管该激酶与 JAK 无关。 f.0059.base-Poty 和f.0167家族的成员编码肿瘤坏死因子受体 SF 的细胞因子受体同源物,已知参与细胞凋亡和炎症（Gravestein 和 Borst,1998）。病毒同源物可能充当宿主对应物的诱饵,隔离细胞因子。

RNA病毒蛋白质结构域库的扩展

在这里,我们通过对蛋白质结构域的广泛搜索来注释已识别的病毒（参见STAR 方法和图S3）。与之前的研究一致（Wolf 等人,2020）,检测到的域的频率遵循幂律分布,其中大多数域仅出现在特定的病毒组中（图S7）。

在 RNA 病毒树中广泛存在的少数标志性结构域中,最普遍的是 RdRP,其次是不同类型的 CP（CP_SJR、CP_levi）、RNA 解旋酶（SF1、SF2、SF3)和丝氨酸/半胱氨酸蛋白酶（图S7）。除了前面提到的裂解域之外,我们还确定了几个 do-

一些双顺反子病毒科成员,以及索林病毒科(f.0024.base-Solinvi、f.0014.base-Solinvi、 f.0017.base-Solinvi、 f.0018.base-Solinvi)和多环病毒科(f.0008.base -Solinvi)的几个基础谱系。 base-Polycipi），包含杆状病毒 IAP 重复（BIR)结构域（杆状病毒凋亡抑制剂)的同源物,已知在细胞周期控制和死亡中发挥作用（Clem,2015）。

核苷二磷酸-X 水解酶 (NUDIX) SF 水解酶在生命的所有领域和 dsDNA 病毒中都很常见（Vasude-van 和 Ryoo,2015）。在这里,我们在 13 个不同的 RNA 病毒科（黄病毒科、诺达病毒科、

囊病毒科和几个候选科）。除了感染囊病毒科的细菌外，我们怀疑这些 RNA 病毒编码的 NUDIX 水解酶的功能与 dsDNA 病毒类似，充当促进宿主蛋白质合成关闭的脱帽酶（Kago 和 Parrish，2021）。

在来自Kitrinoviricota和Pisuviricota门的 11 个不同的 RNA 病毒家族中，我们鉴定了 J 结构域，即 DnaJ (Hsp40) 共伴侣的活性部分(Laudenbach et al., 2021)。在这些病毒中，J结构域是病毒多蛋白的一部分，可能促进多蛋白折叠和加工和/或病毒体组装。

我们还鉴定了与 RNA 修复和代谢有关的几个酶结构域，包括 RtcB 样30-磷酸 RNA 连接酶(Hughes et al., 2020)、 HAM1 样焦磷酸酶(Simone et al., 2013)、 DEDD-SF 30 -50可能参与免疫抑制的核酸外切酶，如沙粒病毒(Hastie et al., 2011)，以及参与 tRNA 剪接的tRNA 20 -磷酸转移酶 (Sawaya et al., 2005)。在细胞生物中，后一种酶通常由 NAD 和 ADP-核糖 (NADAR) 结构域蛋白编码，在 RNA 加工的背景下参与 NAD 代谢（de Souza 和 Aravind，2012）。

NADAR 结构域最初在Roniviridae （+ssRNA 病毒)和巨型 dsDNA 病毒中检测到（de Souza 和 Ara-vind，2012）。我们在 12 个科的 RNA 病毒中鉴定出了 NADAR 结构域，强调了该结构域对于 RNA 病毒复制的潜在重要性。

在某些囊病毒中，我们检测到一种蛋白质，其 N 端结构域与 sigma70 因子（细菌 RNA 聚合酶全酶的一个亚基，将 RNA 聚合酶引导至特定的启动子)的 C 端结构域 (CTD) 同源；佩吉特和赫尔曼，2003）。该囊病毒蛋白的 CTD 与细菌 RNA 聚合酶 α 亚基的 C 末端区域相似。已知 sigma70 和 RNA 聚合酶 α 的 CTD 会相互作用（Chen 等人，2003），这表明这种囊病毒蛋白重建了这种相互作用界面，并可能参与感染期间的转录接管，从而有可能克服宿主的抗病毒防御。

RNA 病毒中这些不同结构域的鉴定
一个或多个谱系的存在意味着病毒与宿主相互作用的多种机制，特别是反防御，这仍有待研究。

RNA 病毒中的替代遗传密码之前的调查发现了几个利用非标准
遗传密码的 RNA 病毒群，表明它们用匹配的密码感染宿主，例如纤毛虫（Wolf 等人，2020）。在此，在 77,510 名 RCR90 代表中，5,843 名（7.5%)显示出替代遗传密码的证据，这通过 RdRP 核心域编码区内存在规范终止密码子来表明（参见STAR 方法）。虽然在大多数情况下，不可能确定具体的替代代码，但在可行的情况下，最常见的代码是 6（UAA 和 UAG 编码 Gln)和 14（UAA 和 UGA 编码 Tyr 和 Trp，分别在纤毛虫（Ring 和 Cavalcanti，2008 年)和扁虫线粒体（Ross 等人，2016 年)中鉴定出的三个有义密码子的重新编码。与许多使用替代遗传密码的 DNA 病毒不同，这些病毒会主动重新编程宿主细胞的翻译机制，以实现其目标。

受益（Ivanova 等人，2014； Yutin 等人，2021），在 RNA 病毒中，此类代码可能代表对宿主翻译机制的适应。这种现象对于多种分离的线粒体病毒来说是众所周知的，它们使用线粒体遗传密码（UGA从stop到Trp重新编码)并在线粒体内复制（Ni-bert，2017），事实上，51%的病毒在大大扩展的线粒体内复制（5,006 个 RCR90 中的 2,553 个)线粒体病毒科使用真菌线粒体中常见的代码 4。除了线粒体病毒外，在大多数大型 RNA 病毒组中都检测到了具有替代遗传密码的重叠群，频率通常为几个百分点。

我们在 RdRP 的系统发育树中鉴定出此类代码富集（>50%)的病毒谱系（表 S5，图2中的绿色弧线）。在Duplornaviricota和Negarnoviricota中没有检测到替代密码的连贯系统发育信号。

相比之下，我们检测到Pisuviricota 的 19 个科，通常包含一两个小分支（8-30 个 RCR90 成员），具有明显的原生生物代码（氨基酸的 UAA 和/或 UAG 代码）。双顺反子病毒科（单部分 + ssRNA 节肢动物病毒)有 12 个这样的分支，表明其中一些双顺反子病毒可能是感染原生生物，可能与节肢动物相关。最后，在Kitrinoviricota，我们观察到替代代码的惊人分布：7 个家族包括具有替代代码的小分支，而其他 7 个家族完全由（f.0150，f.0177–f.181)或主要（f.0176)组成使用替代类原生生物代码的病毒。与之前的发现一致（Wolf 等人，2020），目前的分析表明Kitrinoviricota包含大量以前未曾怀疑的原生病毒多样性。

讨论

宏基因组和宏转录组已分别成为 DNA 和 RNA 病毒发现的主要来源（Call et al., 2021； Simmonds et al., 2017）。在这里，我们分析了从 3,598 个不同的宏转录组中回收的超过 250 万个 RNA 病毒重叠群。宏转录组分析很容易出现伪影，尤其是源自嵌合 RNA 组装的伪影。因此，需要强调的是，这项工作的所有结论都是基于对 RNA 病毒序列的进化保守组的分析，而不是单例分析，假设不同组装体中出现相同嵌合体的可能性极小；实施了其他几种针对嵌合组装的保护措施（参见STAR 方法）。

我们的分析导致 90% RvANI 簇的数量（种和属等级之间)增加了 9 倍，总系统发育深度增加了 5 倍，代表性 RdRP 的数量增加了近 6 倍序列（RCR90），并且从科到纲的水平上假定的分类群数量增加了 5 倍。相比之下，在门水平上，RNA病毒分类学基本保持稳定，只是在之前建立的5个门的基础上增加了两个候选门。

尽管有明显的例外，但之前大多数 RNA 病毒科的门分配仍然保持稳定。因此，囊病毒科扩大了一个数量级，并从Kitrinoviricota迁移到Pisuviricota，现在它与其他 dsRNA 病毒、picobirnaviruses、

和部分病毒。鉴于系统发育的可靠性更高
与大家庭的分析和合理性
这三组 dsRNA 病毒的单系性,目前
囊病毒科的立场可能是有效的。但是,那
其他几个科,特别是黄病毒科的安置
不稳定。虽然这家人也是从Kitrinoviricota搬来的
到Pisuviricota,在这种情况下,实际的隶属关系仍然是
不确定。

Orthornavirae界的门分类出现
稳健,但 RdRP 系统发育的分辨率
靠近根可能不足以破译这种关系
门之间。先前提出的起源场景
来自正义 RNA 病毒内的 dsRNA 病毒
来自Duplornaviricota的负义 RNA 病毒（Negarnaviricota)的多个独立事件（Wolf 等人，
2018)在生物学上仍然是合理的。然而,系统发育
对扩展的 RdRP 集的分析未能证明这一点
尽管支持 dsRNA 病毒的多个起源,但仍支持完整的情况。 Negarnaviricota的基础位置
在这里观察到,尽管对执行的测试很稳健,但很可能,
是深度系统发育分析的产物。相比之下,基础
Lenarviricota在以 RT 为根的树中的位置可能反映了
其余RNA病毒的起源来自一个共同的祖先
该门属于细菌域。考虑到大规模扩张,这种情况似乎特别合理。

这项工作中的细菌 RNA 病毒组。考虑到尺寸和
由于分析数据集的多样性,RdRP 序列中包含的信息似乎确实不足以

解决RNA病毒之间最深层的关系。这
一旦 RdRP 具有足够的多样性,这个问题就值得重新审视
结构的积累可能提供更好的系统发育解析。

本分析消除了长期存在的偏见
RNA 病毒组针对真核生物感染病毒（Koonin 等人，
2015）。除了类维病毒多样性的大幅扩展之外,我们还获得了多种额外的迹象

病毒群感染细菌,特别是小核糖核酸病毒
和一些部分病毒分支。支持这种可能性的关键证据是大量 CRISPR 的发现

针对 RNA 病毒（均为Leviviricetes成员)的间隔区
以及部分病毒内的一组候选 RNA 噬菌体。
目前的结果强烈表明宿主的剧烈转变,
称为水平病毒转移（HVT）,
相关宿主,甚至跨越原核生物与真核生物的鸿沟,
是RNA病毒进化的主要途径（Dolja和Koonin，
2018）。 HVT 事件可能发生在多个独立的
RNA病毒的不同门、纲甚至目中的情况。在这方面,一小群病毒,

检测到多个 CRISPR 间隔区匹配并且
值得注意的是,因此暂时将玫瑰弯杆菌作为宿主。这个狭窄的病毒群来自

独特的栖息地,可能是一个属,深深地存在于部分病毒中,其中许多已知会感染真菌、植物和无脊椎动物（Cross 等人,2020 年； Shi 等人,2016 年；Vainio 等人,2016 年）。 ,2018）。
除了全球RNA病毒组的大幅扩张之外,
这项工作还大大扩展了 RNA 病毒基因组中编码的蛋白质结构域的目录。共同的主题
在这些域中,每个域都以窄线表示-

RNA病毒时代似乎通过多种方式进行反防御
分子机制。这些发现表明,尽管
通常基因组较小,RNA病毒更类似于
DNA病毒在宿主基因的外展方面比
以前曾受到赞赏（Koonin等人， 2022）。
总之,结果极大地扩展了
Orthornavira王国,特别是与细菌相关的 RNA 病毒,同时引入相对较小的变化

纳入最新的分类方案,支持其总体稳健性。此外,还预测了多种蛋白质功能

在RNA病毒中。大量的序列和导数
这项工作中生成的数据可通过同伴获得
网站（riboviria.org)或通过 Zenodo 存款。我们期待
该资源使研究人员能够获得有意义且
描述新 RNA 病毒时的综合背景
未来的研究,例如,通过提供对特定病毒谱系的生态分布的见解或通过特定进化枝

蛋白质结构域注释。此外,该资源可以帮助
研究人员确定了需要表征的关键RNA病毒基因组
实验性地。

研究的局限性

我们检测 RNA 病毒的方法很大程度上依赖于
通过个人资料搜索可能会错过的 RdRP 的存在
具有改变的规范序列的极其遥远的同源物
主题。此外,一些RNA病毒具有“分裂”能力
RdRP,其中基序编码在不同的 ORF 中或
甚至基因组片段（Sutela et al., 2020； Chiba et al.,
2021）。我们基于 RdRP 的发现的另一个缺点是
缺乏对分段RNA病毒的系统鉴定工作
基因组（因为非 RdRP 编码片段将不会被报告）。目前,除了编码基因组片段之外

RdRP 仅通过共现分析来识别
一组感染细菌的类帕蒂蒂病毒
CRISPR。分段 RNA 病毒基因组的综合检测是未来分析的一项任务,分配

彼此/特定病毒基因组的不同片段。
与这项工作同时进行的两项研究产生
相关见解。 Serratus 团队已发表了一项大规模的 RNA 测序档案调查,报告了许多新型 RNA 病毒（Edgar 等人,2022）,并且大规模的

海洋RNA病毒的宏转录组分析
由Tara Oceans 项目发布（Zayed 等人,2022 年）。 A
三个研究结果的综合比较
在许多方法论方面都有所不同,包括
分析的宏转录组的范围,仍然是一个主要的
未来的任务。然而,为了量化之间的重叠
三个项目的结果并相应评估新颖性
对于每个聚类,我们对使用两个聚类阈值（细粒度为 0.9)获得的 RdRP 聚类进行了自动比较
粗粒分类为 0.5（参见STAR 方法）。
该比较的结果（表 S8 “簇交叉点”)检测到共享的簇数量相对较少。

所有三个项目都表明有数千个集群
每个都是独一无二的。在细粒度（阈值为 0.9)时,Serratus 数据中识别出最大数量的独特簇,如下所示
鉴于该项目包含的数据集比其他两个项目大得多,这是可以预期的。然而,在粗粒

（阈值为 0.5），我们目前的结果比其他两项研究的总和包含更多独特的簇，这表明我们的工作涵盖了更大的 RNA 病毒系统发育深度。这一比较支持了我们的结论，即当前全球 RNA 病毒组的采样远未达到饱和。因此，这三项研究是互补的，将结果纳入单一系统发育框架并综合结论应该会大大增进我们对 RNA 病毒圈的了解。

财团

RNA 病毒发现联盟成员是 Adrienne B.
Narrowe、亚历山大 J. 普罗布斯特、亚历山大 Sczyrba、Annegret Kohler、Armand Se´ guin、Ashley Shade、Barbara J. Campbell、Bjorn D. Lindahl、Brandi Kiel Reese、Breanna M. Roque、Chris DeRito、Colin Averill、Daniel Cullen , 大卫·AC·贝克, 大卫·A.
沃尔什、大卫·M·沃德、吴东英、艾米莉·埃洛-法德罗什、Eoin L. Brodie、Erica B. Young、Erik A. Lilleskov、Federico J. Cast-tillo、Francis M. Martin、Gary R. LeCleir、Graeme T. Attwood、Hinsby Cadillo-Quiroz、Holly M. Simon、Ian Hewson、Igor V. Gri-goriev、James M. Tiedje、Janet K. Jansson、Janey Lee、Jean S.
VanderGheynst、Jeff Dangl、Jeff S. Bowman、Jeffrey L. Blanc-chard、Jennifer L. Bowen、Jiangbing Xu、Jillian F. Banfield、Jody W. Deming、Joel E. Kostka、John M. Gladden、Josephine Z. Rapp、约书亚·夏普、凯瑟琳·D·麦克马洪、凯瑟琳·K.
Treseder、Kay D. Bidle、Kelly C. Wrighton、Kimberlee Thamatra-koln、Klaus Nusslein、Laura K. Meredith、Lucia Ramirez、Marc Buee、Marcel Huntemann、Marina G. Kalyuzhnaya、Mark P. Wal-drop、Matthew B. Sullivan马修·O·施伦克、马蒂亚斯·赫斯、迈克尔·A·维加、米歇尔·A·奥马利、莫妮卡·梅迪纳、内奥米·E·
吉尔伯特、Nathalie Delherbe、Olivia U. Mason、Paul Dijkstra、Peter F. Chuckran、Petr Baldrian、Philippe Constant、Ramunas Stepa-nauskas、Rebecca A. Daly、Regina Lamendella、Robert J. Grun-inger、Robert M. McKay、Samuel海兰德、莎拉·L·勒贝斯、莎拉·P·埃瑟、西尔维娅·G·阿西纳斯、史蒂文·S·威廉、史蒂文·W.
辛格、苏珊娜·S·特林奇、塔尼娅·沃伊克、TBK·雷迪、特伦斯·H·贝尔、托马斯·莫克、蒂姆·麦卡利斯特、维拉·泰尔、文森特·J·德内夫、刘文作、威尔姆·马滕斯-哈贝纳、刘晓军扎卡里·库珀和王忠

明星+方法

本文的在线版本提供了详细的方法，包括以下内容：

补充信息

补充信息可在线找到： https://doi.org/10.1016/j.cell。 2022年8月23日

致谢

作者要感谢 Shai Zilberzwige-Tal、David Burstein、Adi Stern、Leah Reshef 和 Omry Lieber 进行的有益讨论。 UG 和 UN 得到欧洲研究理事会 (ERC-AdG 787514) 的支持。联合国得到特拉维夫大学埃德蒙·J·萨夫拉生物信息学中心的奖学金支持。 YIW 和 EVK 得到美国国立卫生研究院（国家医学图书馆）校内研究计划的支持。 VVD 得到了 NIH/NLM/NCBI 访问科学家奖学金的部分支持。美国能源部联合基因组研究所（SR、APC、IMC、NI、DP-E.、NCK 和所有 JGI 共同作者）（DOE 科学用户设施办公室）的工作得到了该办公室的支持美国能源部科学部根据合同号。 DE-AC02-05CH11231。 MK 得到了 l Agence Nationale de la Recherche 资助 ANR-20-CE20-009-02 和 ANR-21-CE11-0001-01 的支持。 DK 由欧洲社会基金资助，资金编号为：09.3.3-LMT-K-712-14-0027。 DAB 得到了 NASA 外星生物学计划 NNX16SJ62G 的资助，以及美国能源部基础能源科学办公室化学科学、地球科学和生物科学部 (CSGB) 光合系统计划 DE-FG02-94ER20137 的资助。我们衷心感谢许多科学家和主要研究人员的贡献，作为能源部联合基因组研究所社区科学计划的一部分，他们发送了分离基因组、环境宏基因组和宏转录组的提取遗传材料或测序结果。使我们能够将这些公开数据集中检测到的 RNA 病毒序列纳入我们的研究中，无论发布状态如何。

作者贡献

UG、EVK、VVD 和 NCK 构思并监督了该项目。
UN、UG、YIW 和 SR 设计了发现管道。 YIW 进行了系统发育分析。 UN、YIW 和 SR 进行了主机分配预测。 SR 进行了栖息地和生态分布分析。 APC、UN、DK 和 MK 进行了蛋白质结构域分析。

UN、SR、YIW 和 APC 执行序列聚类。 SR、DAB 和 DB 分析了黄石温泉组件和 Roseiflexus 样本。 UN 和 BL 构建了配套网站。

IMC、NI、DP-E. 和 NCK 为 IMG 数据库中的数据和元数据收集和管理做出了贡献。
LZA为生态和环境做出了贡献

原生生物分析。 UN、SR、VVD、NCK、UG、YIW、APC、EVK 和 MK 撰写了该手稿，并经过所有作者的编辑和批准。

利益声明

作者声明没有竞争利益。

收件日期：2022 年 2 月 15 日
修订日期：2022 年 5 月 16 日
接受日期：2022 年 8 月 24 日
发布日期：2022 年 9 月 28 日

参考

Altschul, SF、Madden, TL、Schaffer, AA、Zhang, J.、Zhang, Z.、Miller, W. 和 Lipman, DJ (1997)。 Gapped BLAST 和 PSI-BLAST：新一代蛋白质数据库搜索程序。核酸研究。 25、3389–3402 。 https://doi.org/10.1093/nar/25.17.3389。

Andreeva, A.、Howorth, D.、Chothia, C.、Kulesha, E. 和 Murzin, AG (2014)。 SCOP2原型：蛋白质结构挖掘的新方法。核酸研究。 42、 D310–D314。 https://doi.org/10.1093/nar/gkt1242。

Andreeva, A.、Kulesha, E.、Gough, J. 和 Murzin, AG (2020)。 2020年SCOP数据库：已知蛋白质结构的代表性家族和超家族域的扩展分类。核酸研究。 48， D376–D382。 https://doi.org/10.1093/nar/gkz1064。

Arroyo Muëhr, LS、Lagheden, C.、Hassan, SS、Kleppe, SN、Hultin, E. 和 Dillner, J. (2020)。从头序列组装需要对嵌合序列进行生物信息学检查。 PLoS One 15， e0237455。 https://doi.org/10.1371/journal.pone.0237455 。

Attwood, TK、Coletta, A.、Muirhead, G.、Pavlopoulou, A.、Philippou, PB、Po-pov, I.、Roma´-Mateo, C.、Theodosiou, A. 和 Mitchell, AL (2012) 。 PRINTS 数据库：细粒度蛋白质序列注释和分析资源 - 2012 年状态。数据库（牛津） 2012 年， bas019。 https://doi. org/10.1093/database/bas019.

Bahiri Elitzur, S.、Cohen-Kupiec, R.、Yacobi, D.、Fine, L.、Boaz, A.、Diament, A. 和 Tuller, T. (2021)。 原核 rRNA-mRNA 相互作用参与所有翻译步骤并塑造细菌转录本。 RNA 生物学18, 684–698。 https://doi.org/10.1080/15476286.2021.1978767。

Bickhart, DM、Kolmogorov, M.、Tseng, E.、Portik, DM、Korobeynikov, A.、Tolstoganov, I.、Uritskiy, G.、Liachko, I.、Sullivan, ST、Shin, SB 等。 （2022）。从复杂的微生物群落中生成谱系解析的、完整的宏基因组组装的基因组。国家生物技术40, 711–719。 https://doi.org/10.1038/s41587-021-01130-z。

Boros, A´.Polga´r, B.、Pankovics, P.、Fenyvesi, H.、Engelmann, P.、Phan, TG、Delwart, E. 和 Reuter, G. (2018)。腹泻鸡泄殖腔样本中存在多种具有功能性原核 Shine-Dalgarno 核糖体结合位点的不同小核糖核酸病毒。病毒学525, 62–72。 https://doi.org/10.1016/j.virol.2018.09.008 。

Buchfink, B.、Xie, C. 和 Huson, DH (2015)。使用 DIAMOND 进行快速、灵敏的蛋白质比对。纳特。方法 12、59–60 。 https://doi.org/10.1038/nmeth.3176 。

布什内尔，B.（2014）。 BBTools 软件包。 https://sourceforge.net/projects/bbmap/ 。

卡希尔，J. 和杨，R. (2019)。噬菌体裂解：多个基因用于多个屏障。 副词。病毒研究。 103、33-70 。 https://doi.org/10.1016/bs.aivir.2018.09.003。

Call, L.、Nayfach, S. 和 Kyrpides, NC (2021)。通过全球宏基因组学照亮病毒圈。安努。生物医学牛顿。数据科学。 4、369-391。 https://doi.org/10.1146/annurev-biodatasci-012221-095114。

Callanan, J.、Stockdale, SR、Shkoporov, A.、Draper, LA、Ross, RP 和 Hill, C. (2020)。已知 ssRNA 噬菌体基因组的扩展：从数十个到一千多个。科学。副词。 6、eaay5981。 https://doi.org/10.1126/sciadv.aay5981。

Camacho, C.、Coulouris, G.、Avagyan, V.、Ma, N.、Papadopoulos, J.、Bealer, K. 和 Madden, TL (2009)。 BLAST+：架构和应用程序。 BMC 生物信息学10, 421。https ://doi.org/10.1186/1471-2105-10-421。

Carradec, Q.、Pelletier, E.、Da Silva, C.、Alberti, A.、Seeleuthner, Y.、Blanc-Ma-thieu, R.、Lima-Mendez, G.、Rocha, F.、Tirichine, L .，拉巴迪，K.，等人。 （2018）。真核生物基因的全球海洋图谱。纳特。交流。 9、373。https: //doi.org/10.1038/s41467-017-02342-1 。

Chamakura, KR、Tran, JS、O Leary, C.、Lisciandro, HG、Antillon, SF、Garza, KD、Tran, E.、Min, L. 和 Young, R. (2020)。单链RNA噬菌体中裂解基因的快速从头进化。纳特。交流。 11、6009。https : //doi.org/10.1038/s41467-020-19860-0。

Chamakura, KR 和 Young, R. (2020)。宏基因组时代的单基因裂解。电流。意见。微生物。 56、109–117 。 https://doi.org/10.1016/j.mib.2020。 09.015。

Chan, PP、Lin, BY、Mak, AJ 和 Lowe, TM (2021)。 TRNAscan-SE 2.0：改进了转移 RNA 基因的检测和功能分类。核酸研究。 49、9077–9096 。 https://doi.org/10.1093/nar/gkab688。

Chen, H.、Tang, H. 和 Ebright, RH (2003)。 RNA 聚合酶 α 亚基 C 端结构域和 sigma70 在 UP 元件和激活剂依赖性转录中的功能相互作用。摩尔。第 11 号牢房， 1621–1633。 https://doi.org/10.1016/s1097-2765(03)00201-6。

Chen, I.-MA、Chu, K.、Palaniappan, K.、Ratner, A.、Huang, J.、Huntemann, M.、Hajek, P.、Ritter, S.、Varghese, N.、Seshadri, R .，等人。 （2021）。 IMG/M 数据管理和分析系统 v.6.0：新工具和高级功能。 核酸研究。 49、 D751–D763。 https://doi.org/10.1093/nar/gkaa939。

Cheng, H.、Liao, Y.、Schaeffer, RD 和 Grishin, NV (2015)。 ECOD 数据库中的手动分类策略。蛋白质 83, 1238–1251。 https://doi. org/10.1002/prot.24818。

Chiba, Y.、Oiki, S.、Yaguchi, T.、Urayama, SI 和 Hagiwara, D. (2021)。真菌病毒中分裂的 RdRp 序列和迄今为止未知的基因组复杂性的发现。病毒进化。 7、veaa101。 https://doi.org/10.1093/ve/veaa101。

克莱姆，RJ (2015)。病毒式 IAP，过去和现在。塞明。细胞开发。生物。 39、72–79 。 https://doi.org/10.1016/j.semcdb.2015.01.011。

Clum, A.、Huntemann, M.、Bushnell, B.、Foster, B.、Foster, B.、Roux, S.、Hajek, PP、Varghese, N.、Mukherjee, S.、Reddy, TBK 等。 （2021）。 DOE JGI 元基因组工作流程。 mSystems 6， e00804-20。 https://doi.org/10.1128/mSystems.00804-20 。

Cross, ST、Maertens, BL、Dunham, TJ、Rodgers, CP、Brehm, AL、Miller, MR、Williams, AM、Foy, BD 和 Stenglein, MD (2020)。感染果蝇和埃及伊蚊的部分病毒表现出有效的双亲垂直传播。 J.维罗尔。 94， e01070-20。 https://doi.org/10.1128/JVI。 01070-20。

Csardi, G. 和 Nepusz, T. (2006)。用于复杂网络研究的 igraph 软件包。 Int.J.复杂系统。 1695 年， 1-9。

戴维森，M.，特雷根，TJ，科伦，S.，波普，M.，和巴亚，D.（2016）。通过对病毒组、内溶素和 CRISPR 间隔区的分析揭示了多微生物群落的多样性。 PLoS One 11， e0160574。 https://doi.org/10.1371/journal。电话.0160574。

de Souza, RF 和 Aravind, L. (2012)。鉴定 NAD 利用代谢途径的新成分并预测其生化功能。摩尔。百奥系统。 8, 1661–1677。 https://doi.org/10.1039/c2mb05487f。

Dessau, M.、Goldhill, D.、McBride, RL、Turner, PE 和 Modis, Y. (2012)。选择压力导致 RNA 病毒牺牲繁殖适应性来换取病毒酶结构和热稳定性的提高。公共图书馆基因。 8、e1003102。 https://doi.org/10.1371/journal.pgen.1003102。

Dolja，VV 和 Koonin，EV (2018)。宏基因组学通过揭示广泛的水平病毒转移重塑了 RNA 病毒进化的概念。病毒研究。 244、36-52 。 https://doi.org/10.1016/j.virusres.2017.10.020。

Dolja, VV、Kreuze, JF 和 Valkonen, JPT (2006)。线形病毒的比较和功能基因组学。病毒研究。 117、38-51 。 https://doi.org/10。 1016/j.virusres.2006.02.002。

埃德加，RC（2021）。 MUSCLE v5通过集成引导（生物信息学)改进了系统发育树置信度的估计。

Edgar, RC、Taylor, J.、Lin, V.、Altman, T.、Barbera, P.、Meleshko, D.、Lohr, D.、Novakovsky, G.、Buchfink, B.、Al-Shayeb, B. ，等人。 （2022）。 PB级

序列比对催化病毒发现。自然602, 142–147。 https://doi.org/10.1038/s41586-021-04332-2 。

Enright, AJ、Van Dongen, S. 和 Ouzounis, CA (2002)。一种用于大规模检测蛋白质家族的有效算法。核酸研究。 30, 1575–1584。 https://doi.org/10.1093/nar/30.7.1575。

Finn, RD、Clements, J. 和 Eddy, SR (2011)。 HMMER Web 服务器:交互式序列相似性搜索。核酸研究。 39、 W29–W37。 https://doi.org/10.1093/nar/gkr367 。

付丽、牛本、朱志、吴胜、李文 (2012)。 CD-HIT:加速下一代测序数据的聚类。生物信息。牛津。英语。 28、3150-3152。https : //doi.org/10.1093/bioinformatics/bts565。

Galperin, MY、Wolf, YI、Makarova, KS、Vera Alvarez, R.、Landsman, D. 和 Koonin, EV (2021)。 COG 数据库更新:重点关注微生物多样性、模式生物和广泛传播的病原体。核酸研究。 49， D274–D281。 https://doi.org/10.1093/nar/gkaa1018。

Gann, ER、Kang, Y.、Dyhrman, ST、Gobler, CJ 和 Wilhelm, SW (2021)。
宏转录组文库制备影响褐潮水华期间病毒群落活动的分析。正面。微生物。 12、664189。https :// doi 。 org/10.3389/fmicb.2021.664189。

Gravestein, LA 和 Borst, J. (1998)。免疫系统中的肿瘤坏死因子受体家族成员。塞明。免疫学。 10、423-434 。 https://doi. org/10.1006/smim.1998.0144。

格雷戈里，AC，扎耶德，AA，康塞克·奥内托，N.，坦普顿，B.，博尔杜克，B.，阿尔贝蒂，A.，阿迪纳，M.，阿尔希波娃，K.，卡迈克尔，M.，克鲁德，C.，等人。
（2019）。从极地到极地的海洋 DNA 病毒宏观和微观多样性。单元格177, 1109–1123.e14。 https://doi.org/10.1016/j.cell.2019.03.040。

Haan, C.、Kreis, S.、Margue, C. 和 Behrmann, I. (2006)。 Jaks 和细胞因子受体 亲密关系。生物化学。药理学。 72, 1538–1546。 https://doi.org/10.1016/j.bcp.2006.04.013。

Hastie, KM、Kimberlin, CR、Zandonatti, MA、MacRae, IJ 和 Saphire, EO (2011)。拉沙病毒核蛋白的结构揭示了 dsRNA 特异性30至50核酸外切酶活性，这对于免疫抑制至关重要。过程。国家。

阿卡德。科学。美国108，2396–2401 。 https://doi.org/10.1073/pnas.1016404108。

豪瑟，M.，斯坦内格，M.，和索丁，J. (2016)。 MMseqs 软件套件，用于快速深度聚类和搜索大型蛋白质序列集。生物信息学32, 1323–1330。 https://doi.org/10.1093/bioinformatics/btw006。

Hockenberry, AJ、Jewett, MC、Amaral, LA 和 Wilke, CO (2018)。基因内 Shine-Dalgarno 序列未针对功能进行选择。分子生物学进化35, 2487–2498。 https://doi.org/10.1093/molbev/msy150。

Holmes, EC 和 Duche ne, S. (2019)。系统发育测序能否安全地推断出整体病毒组的起源？ mBio 10。e00289-00219 。 https://doi.org/10。 1128/mBio.00289-19。

Hughes, KJ、Chen, X.、Burroughs, AM、Aravind, L. 和 Wolin, SL (2020)。
受受损 tRNA 调节的 RNA 修复操纵子。细胞报告33, 108527。https ://doi.org/10.1016/j.celrep.2020.108527。

亨特，JD（2007）。 Matplotlib:2D 图形环境。计算。科学。工程师。 9、90-95。 https://doi.org/10.1109/MCSE.2007.55。

Hyatt, D.、Chen, G.-L.、Locascio, PF、Land, ML、Larimer, FW 和 Hauser, LJ (2010)。浪子:原核基因识别和翻译起始位点识别。 BMC 生物信息学11, 119。https ://doi.org/10.1186/1471-2105-11-119 。

国际病毒分类委员会执行委员会（2020）。
病毒分类学的新范围:将病毒圈划分为 15 个层级。纳特。微生物。 5、668–674。 https://doi.org/10.1038/s41564-020-0709-x 。

Ivanova, NN、Schwientek, P.、Tripp, HJ、Rinke, C.、Pati, A.、Huntemann, M.、Visel, A.、Woyke, T.、Kyrpides, NC 和 Rubin, EM (2014)。停止野外密码子重新分配。科学344, 909–913。 https://doi.org/10.1126/science。 1250691。

Johnson, M.、Zaretskaya, I.、Raytselis, Y.、Merezhuk, Y.、McGinnis, S. 和 Madden, TL (2008)。 NCBI BLAST:更好的网络界面。核酸研究。 36， W5–W9。 https://doi.org/10.1093/nar/gkn201。

Jones, P.、Binns, D.、Chang, HY、Fraser, M.、Li, W.、McAnulla, C.、McWilliam, H.、Maslen, J.、Mitchell, A.、Nuka, G. 等等人。 （2014）。 InterProScan 5:基因组规模蛋白质功能分类。生物信息学30, 1236–1240。 https://doi.org/10.1093/bioinformatics/btu031 。

Kago, G. 和 Parrish, S. (2021)。 Mimivirus L375 Nudix 酶可水解50 mRNA 帽。 PloS One 16， e0245820。 https://doi.org/10.1371/journal。电话.0245820。

Kall, L.、Krogh, A. 和 Sonnhammer, ELL (2004)。一种组合的跨膜拓扑和信号肽预测方法。 J.莫尔。生物。 338、1027–1036 。 https://doi.org/10.1016/j.jmb.2004.03.016。

Kall, L.、Krogh, A. 和 Sonnhammer, ELL (2007)。跨膜拓扑和信号肽预测相结合的优点 Phobius 网络服务器。核酸研究。 35、 W429–W432。 https://doi.org/10.1093/nar/gkm256 。

Katoh, K. 和 Standley, DM (2013)。 MAFFT 多序列比对软件版本 7:性能和可用性方面的改进。摩尔。生物。
进化。 30、772–780 。 https://doi.org/10.1093/molbev/mst010。

Koonin, EV、Dolja, VV 和 Krupovic, M. (2015)。真核生物病毒的起源和进化:终极模块化。病毒学 479–480, 2–25。 https://doi.org/10.1016/j.virol.2015.02.039 。

Koonin, EV、Dolja, VV 和 Krupovic, M. (2022)。病毒进化的逻辑。
细胞宿主微生物30, 917–929。 https://doi.org/10.1016/j.chom.2022.06.008。

Koonin, EV、Dolja, VV、Krupovic, M.、Varsani, A.、Wolf, YI、Yutin, N.、Zerbini, FM 和 Kuhn, JH (2020)。全球组织和提议的病毒世界巨分类法。微生物。摩尔。生物。修订版84， e00061-19。 https://doi. org/10.1128/MMBR.00061-19。

Krishnamurthy, SR、Janowski, AB、Zhao, G.、Barouch, D. 和 Wang, D.
（2016）。 RNA 噬菌体多样性的过度扩张。公共科学图书馆生物学。 14、 e1002409。 https://doi.org/10.1371/journal.pbio.1002409。

Krishnamurthy, SR 和 Wang, D. (2018)。已知和新型小核糖核酸病毒中原核核糖体结合位点的广泛保守。病毒学516， 108-114。

Krogh, A.、Larsson, B.、von Heijne, G. 和 Sonnhammer, ELL (2001)。使用隐马尔可夫模型预测跨膜蛋白拓扑:应用于完整基因组。 J.莫尔。生物。 305、567–580 。 https://doi.org/10.1006/jmbi.2000.4315 。

Kutyshenko，副总裁，Prokhorov，DA，Mikoulinskaia，GV，Molochkov，NV，Ye-gorov，AY，Paskevich，SI 和 Uversky，VN（2021 年）。埃希氏菌裂解噬菌体 T5、RB43 和 RB49 的直系同源内溶素活性位点的比较分析。国际。 J.Biol。大分子。 166、1096–1105 。 https://doi.org/10.1016/j.ijbiomac.2020.10.264。

Lauber, C.、Seifert, M.、Bartenschlager, R. 和 Seitz, S. (2019)。与植物相关的星状病毒的高度分化谱系的发现揭示了马铃薯病毒的出现。病毒研究。 260, 38–48。 https://doi.org/10。 1016/j.virusres.2018.11.009。

Laudenbach, BT、Krey, K.、Emslander, Q.、Andersen, LL、Reim, A.、Scaturro, P.、Mundigl, S.、Dauchert, C.、Manske, K.、Moser, M.、等人。 （2021）。 NUDT2 通过去除50磷酸启动病毒 RNA 降解。纳特。交流。 12、6918。https : //doi.org/10.1038/s41467-021-27239-y。

Li, D.、Luo, R.、Liu, CM、Leung, CM、Ting, HF、Sadakane, K.、Yamashita, H.和 Lam, TW (2016)。 MEGAHIT v1.0:由先进方法和社区实践驱动的快速且可扩展的宏基因组组装器。

方法102、3-11 。

Lu, S.、Wang, J.、Chitsaz, F.、Derbyshire, MK、Geer, RC、Gonzales, NR、Gwadz, M.、Hurwitz, DI、Marchler, GH、Song, JS 等。 （2020）。 CDD/ SPARCLE:2020 年保守域数据库。核酸研究。 48， D265–D268。 https://doi.org/10.1093/nar/gkz991。

Makarova, KS、Wolf, YI、Iranzo, J.、Shmakov, SA、Alkhnbashi, OS、Brons, SJJ、Charpentier, E.、Cheng, D.、Haft, DH、Horvath, P. 等。
（2020）。 CRISPR-Cas 系统的进化分类:2 类及其衍生变体的爆发。纳特。微生物学牧师。 18、67-83 。 https://doi.org/10。 1038/s41579-019-0299-x。

马丁内斯·埃尔南德斯 (F.)、福纳斯 (O.)、卢埃斯马·戈麦斯 (M.)、博尔杜克 (B.)、德拉克鲁斯·佩纳 (MJ)、马丁内斯 (JM)、安东 (J.)、加索尔 (JM)、罗塞利 (R) .,Rodri-guez-Valera,F.,等人。 （2017）。单病毒基因组学揭示了隐藏的世界性和丰富的病毒。纳特。交流。 8、15892。https: //doi.org/10.1038/ncomms15892 。

Mihara, T.、Nishimura, Y.、Shimizu, Y.、Nishiyama, H.、Yoshikawa, G.、Uehara, H.、Hingamp, P.、Goto, S. 和 Ogata, H. (2016)。将病毒基因组与宿主分类学联系起来。病毒8、66。https: //www.mdpi.com/1999-4915/8/3/66。

Mistry, J.、Chuguransky, S.、Williams, L.、Qureshi, M.、Salazar, GA、Sonnham-mer, ELL、Tosatto, SCE、Paladin, L.、Raj, S.、Richardson, LJ 等。 （2021）。
Pfam：2021 年蛋白质家族数据库。核酸研究。 49、 D412-D419。 https://doi.org/10.1093/nar/gkaa913。

Morgulis, A.、Gertz, EM、Schaffer, AA 和 Agarwala, R. (2006)。一种快速、对称的 DUST 实现，用于掩盖低复杂性 DNA 序列。
J. 计算机。生物。 13, 1028–1040。 https://doi.org/10.1089/cmb.2006.13.1028。

Mukherjee, S.、Stamatis, D.、Bertsch, J.、Ovchinnikova, G.、Sundaramurthi, JC、Lee, J.、Kandimalla, M.、Chen, I.-MA、Kyrpides, NC 和 Reddy, TBK（2021）。 Genomes OnLine 数据库 (GOLD) v.8：概述和更新。
核酸研究。 49、 D723–D733。 https://doi.org/10.1093/nar/gkaa983。

阿肯色州穆什吉安 (2020)。地球上的病毒颗粒是10 31 个还是更多还是更少？ J.细菌。 202，e00052-20。 https://doi.org/10.1128/JB.00052-20。

Nayfach, S.、Camargo, AP、Schulz, F.、Eloe-Fadrosh, E.、Roux, S. 和 Kyrpi-des, NC (2021)。CheckV 评估元基因组组装的病毒基因组的质量和完整性。纳特。生物技术。 39、578–585 。 https://doi. org/10.1038/s41587-020-00774-7。

NCBI 资源协调员（2018 年）。国家生物技术信息中心的数据库资源。核酸研究。 46， D8–D13。 https://doi. org/10.1093/nar/gkx1095。

Nguyen, L.-T.、Schmidt, HA、von Haeseler, A. 和 Minh, BQ(2015)。 IQ-TREE：一种快速有效的随机算法，用于估计最大似然系统发育。摩尔。生物。进化。 32、268-274 。 https://doi.org/10.1093/molbev/msu300。

尼伯特，ML（2017）。线粒体病毒 UGA(Trp) 密码子使用与宿主线粒体相似。病毒学507, 96–100。 https://doi.org/10.1016/j.virol.2017。 04.010。

Oliveira, H.、Melo, LDR、Santos, SB、No´ brega, FL、Ferreira, EC、Cerca, N.、Azeredo, J. 和 Kluskens, LD (2013)。噬菌体内溶素的分子方面和比较基因组学。 J.维罗尔。 87、4558–4570 。 https://doi. org/10.1128/JVI.03277-12。

佩吉特，MSB 和赫尔曼，JD (2003)。 sigma 因子的 sigma70 家族。
基因组生物学。 4、203。https: //doi.org/10.1186/gb-2003-4-1-203。

Potenza, E.、Di Domenico, T.、Walsh, I. 和 Tosatto, SCE (2015)。 MobiDB 2.0：一个改进的本质无序和可移动蛋白质的数据库。核酸研究。 43、 D315–D320。 https://doi.org/10.1093/nar/gku982。

Potter, SC、Luciani, A.、Eddy, SR、Park, Y.、Lopez, R. 和 Finn, RD (2018)。
HMMER 网络服务器：2018 更新。核酸研究46， W200–W204。 https://doi.org/10.1093/nar/gky448。

Price, MN、Dehal, PS 和 Arkin, AP (2010)。 FastTree 2 – 用于大型比对的近似最大似然树。PLoS One 5，e9490。 https://doi.org/10.1371/journal.pone.0009490 。

Quast, C.、Pruesse, E.、Yilmaz, P.、Gerken, J.、Schweer, T.、Yarza, P.、Peplies, J. 和 Glo´ckner, FO (2013)。 SILVA 核糖体 RNA 基因数据库项目：改进的数据处理和基于网络的工具。核酸研究。 41、 D590–D596。

昆兰，阿肯色州（2014）。 BEDTools：瑞士军队用于基因组特征分析的工具。电流。协议。生物信息。 47、11.12.1–11.12.34 。 https://doi.org/10.1002/0471250953.bi1112s47 。

Rice, P.、Longden, I. 和 Bleasby, A. (2000)。 EMBOSS：欧洲分子生物学开放软件套件。趋势基因。 16、276-277 。 https://doi.org/10.1016/s0168-9525(00)02024-2 。

Richter, M. 和 Rossello´-Mo´ ra, R. (2009)。改变原核物种定义的基因组金标准。过程。国家。阿卡德。科学。美国106，19126–19131。https : //doi.org/10.1073/pnas.0906412106。

Ring, KL 和 Cavalcanti, ARO (2008)。终止密码子重新分配对具有替代遗传密码的纤毛虫蛋白质进化的影响。摩尔。
生物。进化。 25、179-186 。 https://doi.org/10.1093/molbev/msm237。

Ross, E.、Blair, D.、Guerrero-Herna´ ndez, C. 和 Sa´nchez Alvarado, A. (2016)。比较和转录组分析揭示了扁虫线粒体基因组中编码RNA和长非编码RNA的关键方面。 G3（贝塞斯达） 6，1191–1200。 https://doi.org/10.1534/g3.116.028175。

Roux, S.、Brum, JR、Dutilh, BE、砂川 S.、Duhaime, MB、Loy, A.、Pou-los, BT、Solonenko, N.、Lara, E.、Poulain, J. 等。 （2016）。全球丰富的海洋病毒的生态基因组学和潜在的生物地球化学影响。自然537, 689–693。 https://doi.org/10.1038/nature19366。

Roux, S.、Krupovic, M.、Poulet, A.、Debroas, D. 和 Enault, F. (2012)。通过从病毒组读数组装的 81 个新完整基因组的集合，了解微病毒科病毒家族的进化和多样性。 PloS One 7，e40418。 https://doi.org/10.1371/journal.pone.0040418。

Roux, S.、Pa´ ez-Espino, D.、Chen, I.-MA、Palaniappan, K.、Ratner, A.、Chu, K.、Reddy, TBK、Nayfach, S.、Schulz, F.、卡尔，L.，等人。 （2021）。 IMG/VR v3：用于询问未培养病毒基因组的集成生态和进化框架。核酸研究。 49、 D764–D775。 https://doi.org/10。 1093/nar/gkaa946。

Salazar, G.、Paoli, L.、Alberti, A.、Huerta-Cepas, J.、Ruscheweyh, HJ、Cuenca, M.、Field, CM、Coelho, LP、Cruaud, C.、Engelen, S. 等等人。 （2019）。基因表达变化和群落更替对全球海洋宏转录组产生了不同的影响。单元格179, 1068–1083.e21。 https://doi.org/10。 1016/j.cell.2019.10.014。

Sawaya, R.、Schwer, B. 和 Shuman, S. (2005)。酵母 NAD+依赖性 tRNA 20 -磷酸转移酶 Tpt1 的结构功能分析。 《RNA 纽约》 11, 107–113。 https://doi.org/10.1261/rna.7193705。

Schulz, F.、Roux, S.、Paez-Espino, D.、Jungbluth, S.、Walsh, DA、Denef, VJ、McMahon, KD、Konstantinidis, KT、Eloe-Fadrosh, EA、Kyrpides, NC 等。
（2020）。通过全球宏基因组学实现巨大的病毒多样性和宿主相互作用。自然578, 432–436。 https://doi.org/10.1038/s41586-020-1957-x。

石明、林X.-D.、田J.-H.、陈L.-J.、陈X.、李C.-X.、秦X.-C.、 Li,J.,Cao,J.-P.,Eden,J.-S.,等。 （2016）。重新定义无脊椎动物 RNA 病毒球。自然540, 539–543。 https://doi.org/10.1038/nature20167。

Silas, S.、Makarova, KS、Shmakov, S.、Pa´ ez-Espino, D.、Mohr, G.、Liu, Y.、Da-vison, M.、Roux, S.、Krishnamurthy, SR、Fu 、BXH 等人。 （2017）。关于使用逆转录酶的 CRISPR-Cas 系统的起源及其高度多样化、神秘的间隔序列。 mBio 8。e00897–e00817。 https://doi.org/10。 1128/mBio.00897-17。

Sillitoe, I.、Bordin, N.、Dawson, N.、Waman, VP、Ashford, P.、Scholes, HM、Pang, CSM、Woodridge, L.、Rauer, C.、Sen, N. 等人。 （2021）。 CATH：增加功能空间的结构覆盖。核酸研究。 49， D266–D273。 https://doi.org/10.1093/nar/gkaa1079。

Simmonds, P.、Adams, MJ、Benko, M.、Breitbart, M.、Brister, JR、Carstens, EB、Davison, AJ、Delwart, E.、Gorbalenya, AE、Harrach, B. 等。 （2017）。
共识声明：宏基因组学时代的病毒分类学。纳特。
微生物学牧师。 15、161-168 。 https://doi.org/10.1038/nrmicro.2016.177。

Simone, PD、Pavlov, YI 和 Borgstahl, GEO (2013)。 ITPA（肌苷三磷酸焦磷酸酶）：从核苷酸库监测到人类疾病和药物遗传学。变异。资源。 753、131-146 。 https://doi.org/10。 1016/j.mrrev.2013.08.001。

Skennerton, CT、Imelfort, M. 和 Tyson, GW (2013)。 Crass：从未组装的宏基因组数据中识别和重建 CRISPR。核酸研究。 41、 e105。 https://doi.org/10.1093/nar/gkt183。

索丁，J. (2005)。通过HMM-HMM比较检测蛋白质同源性。
生物信息。牛津。英语。 21、951–960 。 https://doi.org/10.1093/bioinformatics/bti125 。

Starr, EP、Nuccio, EE、Pett-Ridge, J.、Banfield, JF 和 Firestone, MK
（2019）。宏转录组重建揭示了 RNA 病毒具有影响土壤碳循环的潜力。过程。国家。阿卡德。科学。美国116, 25900–25908。https: //doi.org/10.1073/pnas.1908291116。

Steinegger, M.、Meier, M.、Mirdita, M.、Vohringer, H.、Haunsberger, SJ 和
索丁,J.（2019）。 HH-suite3 用于快速远程同源性检测和深度蛋白质注释。 BMC 生物信息学20, 473。https ://
doi.org/10.1186/s12859-
019-3019-7。

Steinegger, M. 和 So ding, J. (2017)。 MMseqs2 使敏感蛋白质成为可能
序列搜索用于分析海量数据集。纳特。生物技术。
35、1026–1028 。 https://doi.org/10.1038/nbt.3988。

Sutela, S.、Forgia, M.、Vainio, EJ、Chiapello, M.、Daghino, S.、Vallino, M.、Martino, E.、Giranda, M.、Perotto,
S. 和 Turina, M. .（2020）。病毒组来自
内生菌根真菌的收集揭示了前所未有的新病毒类群
基因组组织。病毒进化。 6、veaa076。 https://doi.org/10.1093/ve/
veaa076。

Vainio, EJ、Chiba, S.、Ghabrial, SA、Maiss, E.、Roossinck, M.、Sabanadzovic,
S.、铃木 N.、谢 J. 和尼伯特 M.； ICTV 报告联盟 (2018)。 ICTV 病毒分类学概况:Partitiviridae。维罗尔将军。
99、17-18 。 https://doi.org/10。
1099/jgv.0.000985。

van der Meer, MTJ, Klatt, CG, Wood, J., Bryant, DA, Bateson, MM, Lam-merts, L., Schouten, S., Damste
´, JSS, Madigan, MT 和 Ward, DM
（2010）。与居住在黄石温泉微生物垫中的主要原位种群密切相关的玫瑰弯属菌株的培养以及基因组、营养和脂质
生物标志物表征。 J.细菌。 192、3033–3042 。

https://doi.org/10.1128/JB.01610-09。

Vasudevan, D. 和 Ryoo, HD (2015)。 IAPs 调节细胞死亡
他们的对手。电流。顶部。开发。生物。 114、185–208 。 https://doi.org/10。
1016/bs.ctdb.2015.07.026。

Wheeler, TJ 和 Eddy, SR (2013)。 nhmmer:使用配置文件 HMM 进行 DNA 同源性搜索。生物信息学29, 2487–
2489。 https://doi.org/10.1093/bioinformatics/btt403 。

Wolf, YI、Kazlauskas, D.、Iranzo, J.、Luc´a-Sanz, A.、Kuhn, JH、Krupovic, M.、
Dolja,VV 和 Koonin,EV (2018)。全球RNA的起源和进化
病毒组。 mBio 9。e02329–e02318。 https://doi.org/10.1128/mBio.02329-18。

Wolf,YI,Silas,S.,Wang,Y.,Wu,S.,Bocek,M.,Kazlauskas,D.,Krupovic,M.,
Fire, A.、Dolja, VV 和 Koonin, EV (2020)。已知一组 RNA 的加倍
通过水生病毒组的宏基因组分析来识别病毒。纳特。微生物。 5、
1262–1270。 https://doi.org/10.1038/s41564-020-0755-4。

Wu, R.、Davison, MR、Gao, Y.、Nicora, CD、Mcdermott, JE、Burnum-John-son, KE、Hofmockel, KS 和
Jansson, JK (2021)。水分调节土壤
活性 DNA 和 RNA 病毒的储存库。交流。生物。 4、992。https ://doi。
org/10.1038/s42003-021-02514-2。

徐S.,戴Z.,郭平,付X.,刘S.,周L.,唐文,冯T.,陈明,
詹,L.,等人。 （2021）。 ggtreeExtra:丰富注释的紧凑可视化
系统发育数据。摩尔。生物。进化。 38、4039–4042 。 https://doi.org/10.1093/
莫尔贝夫/msab166。

于 G.、林 TT-Y.、朱 H. 和关 Y. (2018)。两种映射方法
并使用 ggtree 可视化系统发育的相关数据。摩尔。生物。进化。 35、
3041–3043。 https://doi.org/10.1093/molbev/msy194。

尤廷,N.,本勒,S.,什马科夫,SA,沃尔夫,YI,托尔斯泰,I.,雷科,M.,安蒂波夫,
D.、Pevzner, PA 和 Koonin, EV (2021)。对人类肠道宏基因组组装病毒基因组的分析揭示了多种假定的 CrAss 样
病毒基因组
具有独特基因组特征的噬菌体。纳特。交流。 12, 1044。https ://doi。
org/10.1038/s41467-021-21350-w。

Zayed, AA、Wainaina, JM、Dominguez-Huerta, G.、Pelletier, E.、Guo, J.、
Mohssen, M.、Tian, F.、Pratama, AA、Bolduc, B.、Zablocki, O. 等。 （2022）。
地球进化起源中隐藏且丰富的海洋病毒
RNA病毒组。科学376, 156–162。 https://doi.org/10.1126/science。
abm5847。

Zeigler Allen, L.、McCrow, JP、Ininbergs, K.、Dupont, CL、Badger, JH、Hoff-man, JM、Ekman, M.、Allen, AE、
Bergman, B. 和 Venter, JC (2017) 。这
波罗的海病毒组:DNA 和 RNA 病毒的多样性和转录活性。 mSystems 2,e00125-16。 https://doi.org/10.1128/
mSystems.00125-16。

Zimmermann, L.、Stephens, A.、Nam, S.-Z.、Rau, D.、Kubler, J.、Lozajic, M.、Ga-bler, F.、So ding, J.、Lupas 、
AN 和 Alva, V. (2018)。完全重新实现的 MPI 生物信息学工具包,其核心是新的 HHpred 服务器。

J.莫尔。生物。 430、2237–2243 。 https://doi.org/10.1016/j.jmb.2017.12.007。

明星+方法

关键资源表

试剂或资源	来源	识别码
存入数据		
所有原始数据和代码均产生于这项工作	这张纸	https://doi.org/10.5281/zenodo.6553771
本研究产生的原始代码	这张纸	https://github.com/UriNeri/RVMT
附带的交互式门户网站	这张纸	https://riboviria.org
软件和算法		
MMseqs2	斯坦格和索丁,2017	https://github.com/soedinglab/MMseqs2
NCBI BLAST+ 套件	Altschul 等人,1997;约翰逊等人,2008	https://ftp.ncbi.nlm.nih.gov/blast/可执行文件/blast+/最新/
钻石	Buchfink 等人,2015	https://github.com/bbuchfink/diamond
bbmap v38.81	Bushnell,2014	https://sourceforge.net/projects/bbmap/
肌肉 v.5	埃德加,2021	https://www.drive5.com/muscle/下载.htm
玛夫特 v7.407	Katoh 和 Standley,2013 Steinegger	https://mafft.cbrc.jp/alignment/software/
HH-套房	等人,2019	https://github.com/soedinglab/hh-suite
HMMER	索丁,2005； Potter 等人,2018 Fu 等人,2012	http://hmmer.org/
CD-HIT	Enright 等人,2002	https://github.com/weizhongli/cdhit
中层细胞层		https://micans.org/mcl/index.html
格树	余等人,2018	https://bioconductor.org/packages/发布/bioc/html/ggtree.html
ggtree额外	徐等人,2021	https://bioconductor.org/packages/发布/bioc/html/ggtreeExtra.html
智商树	阮等人,2015	http://www.iqtree.org/
防尘罩（v1.0.0）	莫尔古利斯等人,2006	https://www.ncbi.nlm.nih.gov/IEB/工具箱/CPP_DOC/lxr/source/src/app/防尘罩/
双人双人舞 (v6.6.0.0)	赖斯等人,2000	http://emboss.open-bio.org/rel/rel6/apps/etandem.html
右	用于统计计算的 R 项目	https://cran.r-project.org/
Python	Python 软件基金会Csardi 和 Nepusz，	https://www.python.org
Igraph	2006 Hyatt 等人,2010 Chan 等人，	https://igraph.org/
Prodigal (v2.6.3)	2021	https://github.com/hyattpd/Prodigal
tRNA扫描ME2		https://github.com/UCSC-LoweLab/tRNA扫描-SE

资源可用性

铅接触

更多信息以及对资源和附加数据的请求应直接发送给主要联系人,并由其满足，尿里@mail.tau.ac.il （联合国）。

材料可用性

这项研究没有产生新的独特试剂、物理样本或特定的生物材料。作为一个计算项目，这项研究的输入是公开的,详见下文“元转录组获取”。本研究的所有结果和输出下面的“数据和代码可用性”部分中进行了描述。

数据和代码可用性本工作中产生的所有
原始数据和代码均可通过多个场所免费且完全获得（DOI 也列在关键资源表中）： d本项目过程中产生的所有数据、代码和结果，以及随附的交互式门户网站的最新版本(https://riboviria.org)，可通
过 CERN 的 Zenodo 存储库 (https://doi.org/10.5281/zenodo.6553771) 获取。该项目旨在作为社区范围的资源。因此，Zenodo 存储库包括附加信息和各种中间结果和二次分析，例如预测的编码序列、宿主分配、系统发育和分类隶属关系、原始域隐马尔可夫模型 (HMM) 搜索匹配、在此生成的附加域配置文件数据库。工作（例如比对、HMM、原始种子序列和预测功能)以及扩展（2.6M 宏转录组衍生)重叠群集和手动合并的“参考集”的核酸序列（参见STAR 方法）。

d如上所述，Zenodo 存款包括本研究中生成的原始代码，该代码对应于该项目 GitHub 存储库的最新版本，该存储库可根据开源 MIT 许可证在https://github.com/UriNeri/RVMT。 d重新分析本文报告的数据所需的任何其他信息可根据要求向主要联系人提供。

方法详情

元转录组采集对总共 5,150 个公开可用的预组
装元转录组进行了 RNA 病毒的鉴定，这些元转录组于 2020 年 1 月从 IMG/M 检索（Chen 等人，2021 年； Mukherjee 等人，2021 年）。如前所述，其中大部分是使用 MEGAHIT 进行组装的（Li 等人，2016） （有关不同样本中使用的组装程序的信息，请参阅表 S4 ，如果可用，请参考最初发布样本的研究）。

初级过滤过程为了方便起见，我们在表
S6 - 发现管道搜索和过滤阈值中总结了初级和二级过滤过程的最终工具和截止点。

我们对从 IMG/M 门户获取的重叠群的初始标准丢弃了短于 1,000 nt 的序列或编码 rRNA 基因
（剩余的重叠群通过 mmseqs easy-linclust 以 99% 的序列同一性进行去复制） （Steinegger 和 So ding，2017）。
为了过滤掉极不可能代表 RNA 病毒的序列，我们将获得的宏转录组重叠群与由 1,831 个宏基因组构建的 DNA 序列概要进行了比较，这些宏基因组源自与 1,306 个宏转录组相同的研究。我们选择的元基因组与基因组在线数据库 (GOLD) 门户（Mukherjee 等人，2021)中的 5,510 个元转录组共享“Study_ID”元数据属性，因为这些 DNA 组件将覆盖与分析了宏转录组。使用多种序列搜索工具（特别是 MMseqs2（核酸 - 核酸（搜索类型 3））

（Hauser 等人，2016； Steinegger 和 So ding，2017）、 DIAMOND（翻译核苷酸与 IMG 来源的 DNA 宏基因组预测 ORF（diamondblastx）） （Buchfink 等人，2015）和 NCBI BLAST（核 - 核 - blastn） （Altschul 等人，1997； Camacho 等人，2009； Johnson 等人，2008）)以迭代方式，我们识别并排除了与 DNA 序列数据集中的序列匹配的元转录组重叠群（图1A） ，基于 RNA 病毒不会存在于 DNA 组装体中的假设 DNA 组装体将由细胞有机体、基于 DNA 的移动元件和整合的逆转录病毒组成。进行迭代搜索，使得每次迭代逐渐增加搜索灵敏度（例如，通过减少字长（BLASTn)和更高的灵敏度值（MMseqs2 -灵敏度 ）），同时丢弃元转录组集合中的所有序列在将过滤后的输出推进到下一次迭代之前，与“DNAome”中的序列产生可靠的匹配。

这个过程总共重复了五次迭代，但我们应该注意到初始迭代主要是探索性的（用于过程的粗略调整）。

二次过滤过程为了进一步过滤重叠群集，我
们用参考数据库中的 5,954 个 RNA 病毒序列补充了上述过滤过程的输出，并使用公共数据库（NCBI NT/NR 和 IMG/VR)作为 DNA 集执行了额外的迭代过滤过程。为了防止排除真正的 RNA 病毒序列，我们屏蔽了与后续迭代中的参考 RNA 病毒匹配的公共数据库条目。所有丢弃的重叠群均被聚合并补充有手动识别的 DNA 编码重叠群，创建“误报”数据库，该数据库用于通过排除序列来进一步过滤元转录组数据集，并产生与“误报”的合格匹配放。收集废弃的火柴以进一步完善工作集的过程重复了三次。

中间集中 DNA 残留物的估计为了通过过滤过程评估工作集中 DNA 序列的
残留物，我们通过 (1) 计算 RdRP 与逆转录酶结构域的比率作为 RNA 病毒与 DNA 的代理来常规分析随机重叠群子集-编码重叠群； (2)手动检查最常见的非RNA病毒相关域的存在。值得注意的是，几个特定领域

在此性能评估期间频繁重复出现,并且手动检查显示这些是已知重复的域。
大多数情况下,这些重叠群完全填充了与此类重复域的匹配,并且这些重叠群在公共数据库中具有细胞匹配,其比对值略低于我们的报告或接受标准。因此,如果这些重叠群完全编码多个重复,我们决定丢弃它们,因为它们没有足够的编码空间来编码可识别的 RdRP。

在下面的 RdRP 识别步骤（在下面的部分中描述)之后,大约 130 种逆转录酶已经通过了各种过滤过程并被手动去除。 MMseqs2、PFamA 数据库（Mistry 等人,2021)以及 Wolf 等人的 RdRP 和 RT 集合。 (2018),用于本次评估中执行的所有个人资料搜索。

RdRP鉴定

之前发表的 RdRP 和反转录酶的多序列比对（Wolf 等人,2018,2020)被格式化为特定于工具的主题数据库,并用作搜索由 6 帧端到端组成的序列数据库的查询使用 PSI-BLAST、hmmsearch、DIAMOND 和 MMseqs2 对通过上述过滤过程的重叠群进行翻译。

为了估计所需的搜索截止值,我们用可能产生错误匹配的非 RdRP 序列（称为“真负”集)补充了查询集,其构造如下：(1)使用大量 RdRP 作为查询针对 PDB70 数据库 (2019) 的 hhsearch（来自 HH-Suite）,收集非来自 RNA 病毒且与至少 2 个 RdRP 对齐的 bitcore R 20 的所有匹配项； (2) 获取与 70% 身份聚类的 PDB 条目（通过ftp://resources.rcsb.org/sequence/clusters/bc-70.out）； (3) 获取与所得 PDB ID 相关的 Pfam 条目以及链接到 Pfam 条目的序列； (4) 将高度相似的序列折叠为单个代表（MMseqs2最小覆盖率:100%,最小同一性:90%）。能够与“真阴性”组中的任何序列进行比对的受试者 RdRP 配置文件被丢弃。否则,RdRP 配置文件搜索的接受标准为:配置文件覆盖率 R 50%、E 值 % 1e-10 和评分 R 70。然后对这些严格的参数进行微调,以代表非 RdRP 序列能够达到的最佳可能值。达到。

随后,可靠的 RdRP 匹配被修剪到近似的核心域,我们在操作上将其定义为基序 A-D（参见下面的“基序 A-D 识别”）。提取的 RdRP 核心序列进行预聚类（CD-HIT,覆盖率 R 75%,% ID R 90） （Fu 等人,2012）,传递到全部对全部（DIAMOND BLASTp)运行,格式化以供使用MCL 使用 mcxload (–stream-mirror –stream-neg-log10 -stream-tf ceil(200))、集群（MCL,通货膨胀值在 3.6 和 2.8 之间）、对齐 (MUSCLE),并按照描述格式化为配置文件数据库（ Altschul 等人,1997 年； Buchfink 等人,2015 年； Edgar,2021 年； Enright 等人,2002 年； Steinegger 和 So ding,2017 年）。这个过程重复两次。随后,具有推定 RdRP 的重叠群被用于从整个宏转录组集合中恢复额外的重叠群,这些重叠群高度相似但比初始搜索长度标准短（详细信息请参见下面的“综合识别”）。在所得集合中,覆盖 RdRP 图谱 R 75% 或具有可识别基序 A-D 的序列被认为对于下游系统发育分析来说足够完整。

RdRP 催化基序的识别 A–D通过对 “RdRP 识别”部分中提到的先前发

布的 RdRP MSA 进行半手动分区来构建自定义基序库（可在 Zenodo 项目存档中找到,请参阅数据和代码可用性） 。为了识别沿各个 RdRP 序列的基序,对全长 RdRP 结构域进行了与上述类似的迭代搜索。

假定移码的校正一组 1,656 个重叠群在多个框架上

包含清晰的 RdRP 结构域特征,通常间隔 < 20 个核苷酸 (n=1,118)。为了避免因简单不完整而遗漏这些签名,我们以两种方式解决这些问题：(1)如果任何一个签名覆盖了主题 RdRP 图谱的 R75%,或者编码了所需的催化基序 A-C,将使用该签名;或 (2)通过将两个标记串联成单个氨基酸序列。

已发表的基因组的重叠群集增强为了评估我们的发现在新预测的病毒基因组

的数量和多样性方面的新颖性,并且为了避免排除可能在环境宏转录组中代表性不足的已建立的病毒谱系,我们汇总并编译了“先前发表的”病毒基因组的集合,称为“参考集”。其中包括在 NCBI 的 NT 数据库（NCBI 资源协调员,2018)中识别出的 RdRP 携带序列,以及在此类公共数据库中未索引的序列（在撰写本文时）,这些序列是在之前的几次大规模和著名的 RNA 病毒调查中识别出的和转录组图谱。

我们添加这些补充序列的标准要求它们来自同行评审的出版物,并且所有基础序列都是完全公开的,没有任何限制。 NCBI NT 序列通过与上述程序类似的 RdRP 扫描程序进行鉴定（参见RdRP 鉴定）。之前发布的集合是由Callanan 等人描述的广泛的Leviviricetes集合组成的。 （2020）、 “阳山组合”和Wolf 等人描述的其他。 (2020),以及Lauber 等人描述的拟议的质体病毒组。 (2019),以及在海洋基因图谱中发现的几个 RdRP （Carradec 等人,2018;Salazar 等人,2019）。聚合后,这些序列经历了与本工作中鉴定的宏转录组序列描述的类似程序（即长度过滤、聚类和 RdRP 核心域提取）。最终的序列集被标记为“已知”（即不是新颖的）,并在本工作生成的数据中如此标注（例如图1中的分支颜色）。处理后的“补充序列集”被合并到主序列集中（那些

本研究中确定的)和组合集（称为“VR1507”）用于所有下游分析（系统发育重建、域分析等）。

跨宏转录组的 RNA 病毒重叠群的全面鉴定由于宏转录组组装通常会产生不完整的基因组，无法满足从头RdRP 检测的标准
（见上文），因此我们使用“VR1507”重叠群集（见上文），我们启动了二次“清扫”扫描，从非聚类、非过滤（长度、DNA 相似性、RdRP 存在）“批量组”元转录组重叠群中寻找其他 RNA 病毒重叠群（图1）。为此，“VR1507”被用作“批量组”中高度相似的重叠群的诱饵，使用非敏感 mmseqs 搜索（mmseqs search –search-type 3 –min-aln-len 120 –min-seq-id 0.66 -s 1 -c 0.85 –cov-mode 1），然后严格过滤恢复的匹配项（E 值 < 1e-9，同一性 > 95%，目标覆盖率 R 95%）。选择这些标准作为质量保证措施，以便恢复的重叠群将大部分包含在“VR1507”重叠群对应物中（这个大型扩展数据集可在项目的 Zenodo 存储库中找到，请参阅数据和代码可用性）。添加此包封标准是为了避免捕获延伸到“VR1507”查询的嵌合或其他不确定的核酸区域。过滤后的批量重叠群集与“VR1507”组合，由 2,658,344 个重叠群组成（称为“Add1507”）。为了确定该程序在避免捕获假阳性方面足够严格，我们验证了如果我们对含有 DNA 宏基因组的非 RNA 病毒进行该程序，则不会捕获任何重叠群。为此，我们使用了最近发布的高质量牛（瘤胃）DNA 宏基因组（即长读、HiFi 组件）（Bickhart 等人，2022），之所以选择它，是因为它不是本研究中使用的 DNA 序列集的一部分。发现管道中使用的初级和二级过滤步骤，使其成为可靠的基准。在此搜索中，没有一个重叠群通过了 95% 同一性的比对阈值（单个重叠群产生了 72% ID 的短比对）。

系统发育重建我们通过对包含完整或接近完
整 RdRP 的序列子集执行初步 MMseqs2 聚类运行（参见表S6，表“聚类信息”），选择了一组不同的代表性 RdRP 进行系统发育分析。

这些代表被称为 RCR90，并经历了多次聚类（序列同一性阈值为 0.5 的 MMseqs2）、比对（MUSCLE5）（Edgar，2021）和配置文件-配置文件比较（HHsearch）（Steinegger 等人，2019）的迭代，如如下面所描述的。鉴定出“置换的”RdRP（具有转置基序 C 的序列，遵循 CABD 配置）并“去置换”（即，从序列中切下包含基序 C 的环并重新插入基序 B 的下游）。

一旦所有已识别的具有转置基序的序列都被纳入规范的 ABCD 配置中，则采用以下程序来生成由所有 RCR90 集组成的多序列比对：

d使用 MMseqs2 对序列进行聚类，序列同一性阈值为 0.3；使用 MUSCLE5 比对所得 4,514 个簇中的序列；使用 HHSEARCH 对聚类比对进行轮廓-轮廓比较，生成 4,514x4,514 距离矩阵（距离估计为dAB = -ln(SAB/min(SAA, SBB))，其中SAB是用于比较的 HHSEARCH 分数型材 A 和 B）；使用 R 函数 hclust() 从距离矩阵生成最大链接树；

d树在深度阈值 1.5 处被砍伐，产生 1,360 个子树； d每个子树都用作使用 HHALIGN 进行相应配置文件
的分层对齐的指南，生成
1,360 个对齐；
d从这些比对中提取 1,360 个共有序列（排除具有超过 2/3 间隙字符的位点）
使用 MUSCLE5 对齐；
d共有序列比对中的每个位置都扩展到原始比对的相应列，产生 77,510 RdRp 的比对（其中原始 RdRp 序列被简化为一组位置，匹配其局部共有序列）；

d具有 >90% 间隙字符的位点已从该比对中删除；得到的对齐方式与对齐方式对齐
使用 HHALIGN 的 10 个 RT（5 个 II 组内含子序列和 5 个非 LTR 逆转录转座子序列）。

RdRps 和 RT 的对齐用于使用 FastTree (V.2.1.4
SSE3， Price 等，2010）程序（WAG 进化模型，伽马分布站点速率）并植根于 RT 和 RdRps 之间。

进化枝的分类隶属关系

通过映射识别具有现有分类信息的树叶（MEGA-BLAST，E值<1e-30，查询覆盖率R 95%，主题覆盖率R 95%，对齐长度> 200，身份R 98%，（对齐长度）/查询长度> 0.95) VR1507 序列设置为分析时的最新 ICTV 数据（2021 年 7 月 20 日发布的病毒元数据存储库 (VMR) 文件，对应于 MSL36，可从https://talk.ictvonline.org/taxonomy获取/vmr/m/vmr-文件存储库/13175）。总体而言，映射了 2,765 个重叠群，并且根据最高分数将 ICTV 分类信息克隆到 VR1507 查询。为了提醒 VR1507 重叠群，我们使用 NCBI 的 NR 数据库执行了类似的程序（这些重叠群总共有 6,878 个映射的重叠群，尽管其中有不可忽略的数量缺乏分类信息或匹配废除的分类名）。

建立树上内部节点（即进化枝）的分类从属关系的过程依赖于上述参考树叶的分类分配，并且全部基于两个原则：

d所有来自参考叶最后共同祖先的序列,分配给分类单元T,也属于分类单元T;从较深的树节点下降的序列不属于分类单元T ,因此应分配给同一等级的新分类单元(taxa);

d树分支分裂成给定等级的分类单元的深度由相同等级的现有分类单元定义,并且是局部性的-
依赖性（例如,不同门的科的特征深度可能不同）;

这些原则的应用假设现有的分类法与树不矛盾,即分配给分类群的参考序列形成在同一等级内不重叠和不嵌套的单系进化枝（例如,科进化枝可以不会融入另一个家庭）。对参考叶的分类从属关系的检查表明,这一假设虽然通常得到满足,但在多个地方都被违反了。这就需要首先理清相互冲突的关系。为此,以下程序适用于给定等级的所有分类单元（即分别针对门、纲等）:

d树被修剪为仅包含定义了此等级的叶子（例如,没有族分配的所有叶子都被剥离）;叶子
权重(wi) 来自修剪后的树;
d对于树中存在的每个分类单元T,计算该分类单元中叶子的总重量（WT =叶子上的Swi ,as-
签署给T）;
d对于树中的任意树分支,计算该分支中叶子的总重量（WC = 跨叶子的 Swi,属于C）; d对于进化枝C和分类单元T 的每个组合,计算进化枝-分类单元权重（WCT = 跨叶子的 Swi ,属于C并分配给T）;然后可以计算类似精确度和类似召回的度量（PCT = WCT / WC和RCT = WCT / WT）,并将其组合成质量指数QCT = PCT * RCT。

d对于树中存在的每个分类单元T,进化枝CT = argmax QCT被确定为分类单元T的“原生”位置（该进化枝,其中分类单元T的最大权重集中,其他干扰最小类群）;属于分支CT但不属于T 的叶子,以及属于T 但不属于分支CT 的叶子,分别标记为“入侵”或“外围”;

所有与树不兼容的分类分配都经过检查和解决。在大多数情况下,使用最不可知的方法来解决冲突（即从相应的叶子中剥离分类标签）。在一个案例中,发现Lenarviricota的Timlo-virales目中的大多数科都嵌套在Blumeviridae 的一个非常深分支的科内。为了这项工作的目的,我们保留了Timlovirales最大分支上的Blumeviridae标签,该分支没有冲突的科分配,并从Timlovirales的其余部分中删除了Blumeviridae标签。在其他一些情况下,小科完全嵌套在较大的科中（例如,在大型副粘病毒科分支内分类为太阳病毒科的单叶）,为了后续分析的目的,嵌入的科标签被删除并事后恢复。一旦所有叶子的分类标签与树兼容,就执行以下过程,为每个分类等级的未标记叶子分别分配新的分类标签:

d树的所有节点都分配了深度,定义为跨所有叶子的最长节点到叶子路径,从
这个节点;
d在 77,510 个叶子的完整树中,确定了每个分类单元的最后一个共同祖先节点;记录分类单元的深度,定义为 LCA 节点的深度加上传入树边缘的长度;所有未标记的叶子,来自分类单元
LCA,被分配给这个分类单元;
d现有类群之外的所有进化枝均被隔离;对于每个这样的分支,所有现有姐妹类群的深度都已确定;如果一个分支只有一个姐妹分类单元,则对最近亲缘关系的搜索将扩展到根部,直到至少识别出另一个相关分类单元;阈值深度计算为相关类群集合的平均值;

d在阈值深度解剖现有类群之外的进化枝;每个产生的（亚)进化枝被分配给一个新的分类单元
给定的等级;
d具有单个现有分类单元作为姐妹的新分类单元被标记为与该分类单元相关。

新的分类单元被命名,表明等级（即前缀为p、 c、 o、f和g ,分别表示门、纲、目、科和属）,后面是该等级新分类单元的序号,并且可选地,以与先前描述的分类单元相关的分类单元的标签终止（例如f.0127.base-Noda是RdRP树中Nodaviridae的基础的第127个新科）。

深层系统发育的稳健性

为了评估深层系统发育重建的稳健性,执行了以下程序:

d收集至少有 20 个 RCR90 序列的 201 个家族的列表d从 RT 集中抽取每个家族的随机代表d
从主比对中提取样本的 202 个序列的子比对d重建系统发育树使用 IQ-Tree 程序(Nguyen et al., 2015)并自动选择最佳

拟合模型
按以下方式对 100 个独立样本进行分析:

首先,为五个已知的门中的每一个确定了具有最高质量指数（QI,在进化枝的分类从属关系部分中描述)的进化枝;质量指数值被用作子采样下门单系的度量。

记录了参与打破各自门的单系性的家族（注意,一片叶子既可以是其自身门的异常值,也可以是另一个门的入侵者）。

其次,子采样树被折叠到门级别;排除了 15 棵（共 100 棵)具有并系门的树（其中,例如Pisuviricota的最高质量进化枝嵌入Kitrinoviricota 的最高质量进化枝中）。使用 IQ-Tree 程序为其余 85 棵（大部分)单系门树构建了扩展的多数规则共识树;分支支持值乘以 0.85（此类树在整个样本中所占的比例）。

将单个重叠群分配给 RCR90 簇一旦上述 RCR90 大树的新区域完全被主
要分类等级（门/属)填充,我们就开始从更大的 VR1507 集中附属重叠群（见上文 - 重叠群集）。通过分为以下 4 个级别,逐步进行重叠群隶属关系:

A 级是对用于创建树的 RdRP 进行编码的重叠群。 B 级由编码 RdRP 的重叠群组成,与 A 级的 RdRP 具有极高的氨基酸同一性（通过最佳 BLASTp 匹配,同一性 R90%、查询覆盖率 R75% 和 E 值 < 1e-3）。 Level.C 由来自与水平 {A, B} 的重叠群相同的 RvANI90 簇（参见下面的定义)的重叠群组成,水平 D. 由与来自水平 {A - C} 的重叠群共享高度核酸相似性的重叠群组成,（通过最佳dc-MEGABLAST 达到身份 R90%、查询覆盖率 R75% 或 Nident R 900nt 且 E 值 < 1e-3)。根据 ICTV 标记的 RdRPS 在上述级别的分布,我们估计以这种方式附属的大多数重叠群将大致共享相同的分类等级,直至属级别。

值得注意的是,对于 C 级,我们设计了定制测量单位 RvANI,它是标准平均核酸同一性 (ANI) 聚类的扩展,旨在适应宏转录组组装体的碎片性质,从而避免由于相对的差异而导致的新颖性高估。相关序列的成对覆盖率低。简而言之,RvANI 的计算如下:最初,mmseqs 用于计算重叠群集中的所有成对序列比对,然后将其用于传统的 ANI 和比对分数 (AF) 计算,其中:

ANI = ð%ID 3比对长度ÞOMinðlengthofcontigm:;连续长度:Þ

AF = MinðAlignmentcoverageofcontigm:;重叠区域的对齐覆盖范围:Þ

给定所有 ANI 和 AF 对（对于原核生物,95-96% ANI 是普遍接受的物种边界,某些病毒具有类似的细粒度定义（Nayfach 等人,2021； Richter 和 Rossello´-Mo´ra,2009）,簇为定义为核酸相似性图中的连接成分,修剪为与 ANI R90% 和 AF R90% 的成对比对。RvANI 通过将特定的成对比对重新插入到修剪后的核酸相似性图中,纠正元转录组中基因组覆盖不均匀的情况,即使它们的 AF 低于要求的值截止,只要基础成对比对满足以下标准:%ID R 99,比对长度 R 150 [bp],并且比对发生在重叠群的边缘之间,即比对覆盖每个重叠群的 5 或 3 末端重叠群）。

随后,我们将 RvANI90 簇定义为核酸模拟中的不同连接组件（使用 R-igraph 包）
清晰度图按上述方式处理（Csardi 和 Nepusz,2006）。

鉴定可靠的 CRISPR 间隔区命中将 RNA 病毒序列与预测的细菌
和古细菌 CRISPR 间隔区序列进行比较,以 (i) 鉴定哪些病毒可能感染原核宿主,以及 (ii) 可能预测这些病毒的特定宿主分类群。首先,使用带有选项“-dust no -word_size 7”的blastn v2.9.0将非冗余RNA病毒序列与IMG数据库（Chen等人,2021)中细菌和古细菌全基因组预测的1,568,535个CRISPR间隔区进行比较 。为了最大限度地减少由于低复杂性和/或重复序列而导致的假阳性命中数,如果 (i) CRISPR 间隔区编码在包含 2 个或更少间隔区的预测 CRISPR 阵列中,则 CRISPR 间隔区被排除在本次分析之外,(ii) 它们是% 20bp,或 (iii) 它们包含由 Dustmasker (v1.0.0) 检测到的低复杂性或重复序列（Morgulis 等人,2006） （选项“-window 20 -level 10”)或直接重复序列使用 etandem (v6.6.0.0) 检测 R 4bp （Rice 等人,2000） （选项“-minrepeat 4 -maxrepeat 15 -threshold 2”）。为了将RNA病毒连接到CRISPR间隔区,仅考虑在整个间隔区长度上具有0或1个错配的blastn命中。进一步检查间隔区和命中的阵列,以检查 (i) 间隔区在整个阵列中的长度是否一致,以及 (ii) 是否在假定的宿主基因组中发现了 Cas 和/或 RT 基因,如果是,则这些基因是否存在与命中的 CRISPR 阵列相邻。为了将 CRISPR 链接的搜索范围扩大到有基因组草案的细菌和古细菌之外,我们接下来使用相同的方法将非冗余 RNA 病毒序列与从 IMG 数据库中可用的宏基因组组件预测的 53,372,161 个 CRISPR 间隔区进行比较。使用与基因组衍生的 CRISPR 阵列相同的方法（见上文)过滤掉虚假间隔区,并且仅对来自同一生态系统（如 GOLD 数据库中定义)的 RNA 病毒和 CRISPR 间隔区的匹配进行筛选。保留。由于 CRISPR 间隔阵列通常组装在没有任何其他基因的短重叠群上,因此我们使用阵列的重复序列将它们连接到假定的宿主。来自宏基因组衍生的 CRISPR 阵列的重复序列至少有 1 次命中

使用带有选项“-perc_identity 90 -dust no -word_size 7”的blastn（v2.9.0)将RNA病毒序列与所有IMG细菌和古细菌基因组进行比较。然后检查这些命中在假定宿主基因组中的位置是否存在预测的 CRISPR 间隔阵列、Cas 基因和 RT 基因。当单个 RNA 病毒序列或间隔区被推定连接到多个宿主基因组时，根据以下标准对它们进行优先排序：(i)在 RT 编码 CRISPR 阵列旁边识别间隔区阵列，(ii)RT 编码 CRISPR 阵列在基因组的其他地方鉴定出，(iii) 在 III 型 CRISPR 阵列旁边鉴定出间隔阵列，(iv) 在基因组的其他地方鉴定出 III 型 CRISPR 阵列，(v) 在基因组的其他地方鉴定出另一种类型的 CRISPR 阵列。 (vi) 在基因组中无法识别出可识别的 Cas 基因。

由 Roseiflexus sp. 编码的 CRISPR 阵列的间隔区内容。蘑菇泉中的RS-1进一步研究如下。首先，使用专用工具 Crass v1.0.1 使用默认参数专门组装从蘑菇泉微生物丛中采样的 17 个宏基因组（表 S3)的 CRISPR 阵列（Skennerton 等人，2013）。接下来，所有基于重复的阵列均对应于Roseiflexus sp 中已知的 CRISPR 阵列。如前所述，鉴定了RS-1 （表 S3），并收集和过滤了相应的间隔物。将 IMG/VR v3 数据库（Roux 等人，2021)中的 RNA 病毒序列以及 DNA 病毒序列与 Roseiflexus sp. 数据库进行了比较。 RS-1间隔阵列使用blastn (v2.9.0) 和选项“-dust no -word_size 7”。来自感染 Roseiflexus sp. 的推定 RNA 噬菌体的序列。 RS-1最初是根据对 R 1 RS-1 间隔区的命中来识别的，整个间隔区长度上有 % 1 不匹配。对于这些选定的噬菌体，然后收集间隔区长度上最多有 4 个错配的命中，以便能够检测更远的病毒间隔区命中。

Roseiflexus sp . 的候选衣壳片段。 RS-1 clade genPartiti.0019 病毒的鉴定基于 3 个标准：间隔区与 RNA 靶向 CRISPR 阵列匹配、没有相应的 DNA 序列以及在宏转录组时间序列中与 R 1 RdRP 重叠群的高覆盖相关性。首先，使用与粗组装间隔区类似的blastn比较（blastn带有选项“-dust no -word_size 7”和允许%1不匹配)来识别假定的衣壳编码重叠群，即排除编码RdRP或a的所有重叠群CRISPR 阵列，位于 Roseiflexus sp. 靶向的相同元转录组中。 RS-1 III 型-RT CRISPR 阵列 (n=3,958)。接下来，将具有 R 1 间隔区匹配的候选者与来自 Mushroom Spring DNA 宏基因组的所有重叠群（blastn (v2.9.0)，带有选项 -task megablast -max_target_seqs 500 -perc_identity 90)以及具有匹配 DNA 重叠群的所有候选者进行比较（R 90% 同一性)被认为可能是 DNA 噬菌体并被排除（n=3,650）。最后，使用读取映射获得所有genPartiti.0019 RdRP 重叠群和所有候选衣壳片段的覆盖率，如下所述（bbmap.sh (v.38.90)，带有选项 vslow minid=0 indelfilter=2 inslenfilter=3 dellenfilter=3 ），并且 42 个 Mushroom Spring 宏转录组中 Pearson 相关性为 R 0.9 的候选者被保留为可能的衣壳片段 (n=88)。为了评估这些衣壳片段的基因内容，使用 Prodigal (v2.6.3) (Hyatt et al., 2010) （选项“-p meta”)从头预测 cd ，并使用标准 blast-mcl 管道进行聚类（blastp (v2.9.0) 使用默认选项，根据分数 R 50 选择命中，MCL 聚类 (v.14-137)，膨胀值为 2）。对于三个最大的蛋白质簇，使用 MAFFT v7.407 （Katoh 和 Standley，2013)构建序列比对，并用作针对以病毒为中心的 uniprot 公共数据库 (uniprot_sprot_vir70) 的 hhsearch 的输入，以及由已知的partitiviruses和picobirnaviruses的衣壳（可在项目的Zenodo存储库中找到，请参阅数据和代码可用性“Partiti_Picob_CP.tar.gz”和PC1_PROMALS3D_new.hhr）。

栖息地分布和相对丰度估计

出于可视化目的，从 IMG 和 GOLD 数据库获得了每个宏转录组的位置、生态和分类信息。具体来说，GPS 坐标和生态系统分类是从 GOLD 获得的，生态系统信息进一步分组为自定义类别（表S4）。为了粗略估计每个元转录组中存在的宿主多样性，在域级别查询了 IMG 注释管道（Clum 等人，2021)预测的所有重叠群的分类信息，即细菌、古细菌、真核生物和病毒。然后使用分配给细菌和古细菌的重叠群数量与分配给真核生物的重叠群数量之间的比率作为代理来确定“真核生物主导”数据集中的“原核生物主导”。具体来说，真核生物与原核生物重叠群的比例为 % 0.3 或 R0.7 的数据集分别被视为“原核生物主导”或“真核生物主导”，而其他数据集则被视为“原核生物主导”或“真核生物主导”。 “混合”。该地图是使用 python 3.8.5 的包 matplotlib v3.3.4 和底图 v1.2.2 绘制的（Hunter，2007）。

为了进行读映射，建立了一组去复制的 RNA 病毒序列（95% ANI 超过 95% AF，使用 CheckV anicalc.py 和 aniclust.py 脚本建立； Roux 等人，2021），下文称为“NR 映射”数据集。然后，来自 3,998 个元转录组（表 S4)的质量修剪读数（sensu； Clum 等人，2021)被映射到该数据集，如下所示。首先，使用blastn v2.9.0+（E值%0.01)将每个元转录组的重叠群与NR映射数据集进行比较。所有累积爆炸命中 R 90% 平均核苷酸同一性覆盖 R 80% 最短序列的重叠群被认为是推定的 RNA 病毒。映射到被识别为推定 RNA 病毒的重叠群的所有读段以及所有未映射的读段都是从现有的 IMG 读段映射信息中提取的，并使用 bbmap v38.81 (Bushnell, 2014)和以下选项从头映射到 NR 映射数据集上： vslow minid=0 indelfilter=2 inslenfilter=3 dellenfilter=3 。完成此步骤是为了通过排除映射到非病毒宏转录组重叠群的所有读取来减少计算时间和假阳性映射的风险。然后使用 FilterBam 过滤生成的 bam 文件（https://github.com/nextgenusfs/augustus/tree/master/auxprogs/filterBam)仅保留 R50% 同一性和 R50% 覆盖率的映射，并使用 bedtools v2.30.0 （Quinlan，2014)中的基因组cov 来计算每个样本中每个重叠群的平均覆盖深度。然后将分类单元的相对比例计算为分类单元成员的累积覆盖度除以该数据集中所有预测的 RNA 病毒重叠群的总累积覆盖度。

遗传密码分配和 ORF 调用目前,为各种宏基因组数据设计的 ORF
识别软件仅限于标准遗传密码 (11) 或霉菌线粒体遗传密码 (4)（当预测的 ORF 异常短时选择）。为了识别可能使用替代遗传密码的进化枝,我们提取了 RdRp 核心足迹并扫描它们以获取框内标准终止密码子。

我们首先将所有 RdRP 编码重叠群分为两个子集：“标准”和“非标准”（如果任何规范终止密码子出现在 RdRP 核心的狭窄坐标内）。然后,使用默认参数（通过 Prodigal 的 (v2.6.3) 宏基因组模式（“匿名”））对“标准”集进行 Metaprodigal CDS 预测（Hyatt 等人,2010）。在“非标准”子集中,终止密码子使用模式在与每个树叶相关的重叠群中聚合,并分类为“线粒体”（使用 UGA 作为有义密码子)和“原生生物” （其他模式）。计算内部树节点的模式流行度（后代叶子之间的相对频率）;注意到并调查了流行率高的进化枝。出于实际目的，“非标准”子集的 ORF 预测是通过使用第一个遗传密码来执行的,该遗传密码使整个 RdRP 核心能够被翻译。没有任何可用遗传密码能够实现 RdRP 核心不间断翻译的病例被指定为一般“非标准”值,并使用线粒体遗传密码进行预测 (4)。

为了排除这些预测的 RNA 病毒主动重新编码 tRNA 的可能性,VR1507 集使用“全局”标志（针对非特定域)进行了一次 tRNAscanME2 （Chan 等,2021）。生命 tRNA 预测）。在任何预计使用替代遗传密码的病毒重叠群上都没有发现 tRNA,这表明这些病毒很可能是对其宿主的适应,而不是病毒与宿主军备竞赛的一个因素,如一些 dsDNA 噬菌体中所见（Ivanova 等人） .,2014）。

RBS 识别和定量使用 VR1507 作为输入,按照 Schulz 等
人所述进行 RBS 定量。 （2020）。简而言之,Prodigal (v2.6.3) 如上所述运行（参见“遗传密码分配”） （Hyatt 等人,2010； Schulz 等人,2020）,然后我们从以下来源获取“rbs_motif”字段Prodigal 的 GFF 输出文件,并将不同的50 个UTR 序列分类为“SD”（与 AGGAGG 相似的基序,规范的 Shine-Dalgarno）、“无”和“其他”（有关详细信息,请参阅数据和代码可用性，

RBS_Motif2Type.tsv ）。然后,对于每个重叠群,我们将“%SD”定义为所有“SD”ORF 与具有真实值的所有 ORF 之间的比率
开始（即不被重叠群的边缘截断,字段“start_type”与“Edge”不同）。

领域注释

为了对包含 RdRP 的重叠群编码的蛋白质进行初始域注释,我们使用 hmmsearch（来自 HMMER V3.3.2 套件） （Finn 等人,2011； Wheeler 和 Eddy,2013)将这些蛋白质与隐马尔可夫模型进行匹配(HMM) 使用最大 E 值 0.001 从多个蛋白质谱数据库收集（PFam 34、COG 2020 版本、CDD v.3.19、CATH/Gene3D v4.3、RNAVirDB2020、ECOD 2020.07.17 版本、SCOPe v.1.75） （Andreeva 等人,2014、2020； Cheng 等人,2015； Galperin 等人,2021； Lu 等人,2020； Mistry 等人,2021； Sillitoe 等人,2021； Wolf 等人,2020）。我们用具有溶菌功能的自定义配置文件集合补充了这组 HMM（称为“LysDB” - 可在项目的 Zenodo 存储库中找到,请参阅数据和代码可用性）。 LysDB 的构建基础是 (1) 来自公共数据库的手动审查的配置文件条目,我们可以将其链接到与病毒细胞裂解或病毒从宿主细胞退出相关的 GO 术语,以及 (2) “Sgl”蛋白质的自定义配置文件,其中 Chamakura 等人通过实验证明可以诱导细胞裂解（Chamakura 等人,2020）。此外,我们使用 InterProScan (v.5.52-86.0) 使用 MobiDBLite (v2.0)、Phobius (v.1.01)、PRINTS (v. 42.0)、TMHMM (v.2.0c) 扫描蛋白质序列（Attwood 等人） ., 2012; Jones 等人, 2014; Kall 等人, 2004; Kall 等人, 2007; Krogh 等人, 2001; Potenza 等人, 2015)。

由于用于初始注释的公共蛋白质谱数据库可能包含代表多蛋白的 HMM,这些蛋白质跨越多个功能域,因此我们开发并采用了一种程序来识别在后续注释过程中被屏蔽的此类谱。在此过程中,我们首先使用 hmmemit 命令将 HMMER 配置文件转换为多个序列比对,然后将其用作使用 HH-Suite 执行的全部与全部配置文件比较的输入。接下来,通过标记包含至少两个其他非重叠配置文件的配置文件（“get_poly Proteins.ipynb”脚本,请参阅数据和代码可用性)来识别推定的多蛋白配置文件。提取多蛋白结构域之间的不匹配区域以创建一组保守但未知的结构域,称为“域间”。此外,使用 HHpred 手动检查具有超过 1000 个匹配状态的配置文件（定义为间隙小于 50% 的列）。几个已识别的多蛋白谱被分成它们的组成域。随后,聚合所有 hmmsearch 结果,并根据其分类级别（未整理的配置文件或未知功能（例如“DUF”）的配置文件被取消优先级)及其相对比对统计数据对配置文件匹配进行优先级排序。为了提高域配置文件的功能注释的质量并将功能分配给未注释的配置文件,我们识别了相似配置文件的集群（以下简称氏族）。首先,从原始数据库中提取在初始注释过程中至少有一次命中的配置文件,将其重新格式化为 HH-Suite 的 HHM（如上所述） ,并用于额外的全部对全部步骤。然后,将此配置文件比较的输出用作使用 Leiden 算法（“get_clan_membership.ipynb”脚本,请参阅数据和代码可用性)的基于图形的聚类过程的输入,该算法将氏族识别为高度相似领域的社区。然后,通过将注释从功能注释配置文件转移到其他氏族成员,氏族成员身份被用来提高功能注释的覆盖范围。简而言之,该过程遵循基于共识的标签分配。

例如,一个拥有 12 个标记为“RdRP”的个人资料和 2 个“未分类”个人资料的部落被设置为“RdRP”部落,而 2 个未分类的成员被重新分类为“RdRP” 。冲突案件要么悬而未决,要么选择最小分母。
例如,一个拥有 4 个“未分类”个人资料的部落,还有 12 个标记为“超级家族 2 Helicase”的成员个人资料,以及一个

另外 10 个标记为“超级家族 1 解旋酶”的成员资料被设置为“解旋酶不确定”，并且该标签扩展到这 4 个“未分类”成员。

所有后续配置文件都匹配通过预定义的截止值（E 值 % e-7,得分 R 9,比对长度 R 8[AA]）。用于生成新的自定义配置文件数据库,其过程类似于 RdRP（见上文）。仅具有 R 10 序列、共享相同功能分类的簇被用于生成 HMM。然后,该配置文件集由上述 RNAVirDB2020 数据库中的大多数配置文件以及其他数据库中的几十个精选配置文件进行了补充（这个名为“NVPC”的最终配置文件数据库可通过 Zenodo 存储库项目获得），请参阅数据和代码可用性）。

最后,我们使用新的配置文件数据库,如上所述使用 hmmsearch 查询 330k 重叠群集的六帧翻译。（图 S3 - 注释管道）。随后,我们使用 GGGenomes (https://github.com/thackl/gggenomes)为 400 多个已识别家族（新的和已建立的)中的每一个生成了 z4-20 代表性重叠群的暂定基因组图谱,然后进行手动检查,以识别新的域以及不常见的域融合和分割。

宏转录组组装的质量控制和可靠性宏基因组组装容易出现各种类型的伪影,这些伪影可能导致组装中出现明显的重叠群,这些重叠群不代表原始生物样品中任何现有的核酸分子（Arroyo Muëhr 等人， 2020）。众所周知,嵌合体（由至少两个不同的核酸分子错误组装而成的重叠群)可能是新颖性声明的主要挫折,并且可能难以识别并与真实的遗传实体分离。我们通过实施几个严格的程序来解决这个问题,以避免可能因分析潜在嵌合重叠群而产生的任何误解：

1. 首先,本工作中没有任何声明是基于单例的。相反,我们仅报告基于对进化上保守的序列干组（理想情况下来自多个组装的两个或多个可对齐重叠群)或来自粗略系统发育水平（科水平及以上)保守特征的分析的观察结果。嵌合体在多个组装体中重复出现的可能性似乎可以忽略不计。

2. 其次,当出现意外的观察结果时,例如基因组重排、基因裂变和基因融合,我们在读段水平上手动检查每个案例,即追踪原始测序运行并绘制图谱（通过上述步骤）“栖息地分布和相对丰度估计”部分)对相关重叠群进行原始 Illumina 短读段,并检查沿着组装的重叠群的读段分布,检查重叠群（而不仅仅是 RdRP 编码区域）覆盖得很好。某些部分表现出异常低覆盖率或 GC% 含量倾斜的重叠群被视为不可靠并被丢弃。

3. 我们观察并从我们通过将已发表的来源聚合为可能的嵌合体（大部分是部分Levivirus,部分rRNA)而构建的集合中删除了几十个重叠群。受此观察的启发,我们根据 SILVA rRNA 数据库搜索了整个 VR1507 重叠群集（针对 SILVA SSU 和 LSU Ref NR99 的 BLASTn,默认参数）（Quast 等人,2013）,并手动检查了编码在数据库中鉴定的核糖体蛋白的 40 个重叠群。“结构域注释”部分,公共数据库中的核糖体蛋白概况（例如核糖体蛋白 L3 PF00297.24)总的来说,我们标记了这些类型的 75 个潜在嵌合体（其中 23 个来自先前发布的来源,见表S6,表“rRNA_summary”了解详细信息）。仅这些可疑重叠群的 RdRP 用于下游分析,而重叠群的其余部分则被忽略。

4. 我们进行的 DNA 消减极大地减少了嵌合体的丰度,嵌合体部分由 RNA 病毒序列组成,部分由 DNA 编码的序列组成,无论是 rRNA 还是 mRNA。然而,显然,该程序不能消除由不同 RNA 病毒基因组部分组成的嵌合体。由于此类嵌合体很难与真正的重组病毒基因组区分开来,因此我们采用启发式方法来识别这些嵌合体,使用域注释来检测具有重复的全长 RdRP 足迹的重叠群。这些被认为是嵌合的,因为 RNA 病毒通常编码单个（全长)RdRP。我们发现了一个这样的病例,ND_250651，一种嵌合体,部分是利维病毒,部分是囊病毒。

与最近发布的 RNA 病毒发现工作进行定量比较从https://datacommons.cyverse.org/browse/iplant/home/shared/iVirus/ ZayedWainainaDominguez-Huerta_RNAevolution_Dec2021下载 Tara 项目的 44,779 个 RdRP 。 Serratus 项目 RdRP 由 296,623 个独特的 PalmDB 序列表示,这些序列是从https://github.com/rcedgar/palmdb下载的存储库。 Serratus 序列代表严格定义的 RdRP 核心（仅包含基序 A、B 和 C），中位长度为 107（与 RCR90 组的 453 个氨基酸相比）。值得注意的是,我们的研究、Tara 项目和 Serratus 项目各自对 RdRP 的哪些区域可用于 MSA 和随后的系统发育分析进行了不同的定义。因此,我们将比较限制在最接近研究之间最低公分母的 RdRP 区域,该区域是 palmDB 共享和定义的区域。我们通过使用本研究中的所有 329,202 个唯一 RdRP 序列和 Tara 项目中的 44,779 个 RdRP,针对 PalmDB 集进行 BLASTP 搜索（e 值 0.0001），使用最佳命中来修剪查询（具体来说,使用针对长度L的主题和 q1..q2 的命中足迹查询长度 K 和p1..p2的命中足迹,查询被修剪为max(1,p1-q1-1)..min(K, p2 L-q2)以说明主题的缺失部分）。对 PalmDB 没有造成重大影响的查询不会被修剪。将全套序列汇集在一起,并使用 MMseqs2 进行聚类,序列同一性阈值为 0.9 和 0.5 (–min-seq-id 0.5/0.9 -c 0.333 -e 0.1 –cov-mode 1 –clus-ter-mode 2)。所有序列被分为四类：i）“已知”（GenBank 和当前已发布的其他来源）

+

数据集（参见STAR 方法“已发布基因组的重叠群集增强”），ii)“RVMT”（来自当前数据集的 RNA 病毒元转录组），iii) Serrarus 和 iv) Tara。检查簇中是否存在来自四个集合中的每一个的成员，并且簇集交集列于表 S8 中。

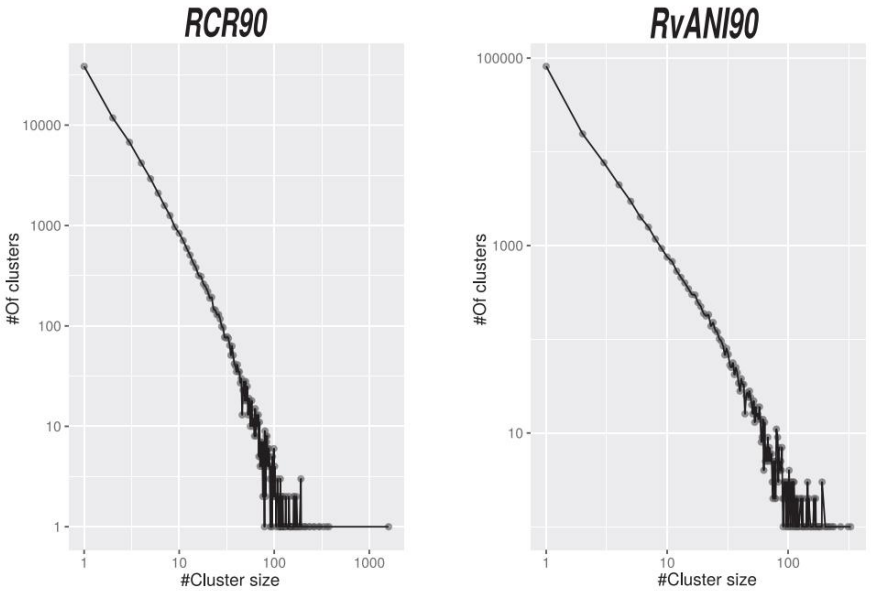
量化和统计分析

相关正文或方法细节以及表 S6中提供了源自序列搜索或比对程序（例如结构域预测、CRISPR 间隔区匹配等）的所有分析的精确阈值，包括期望值（E 值）（“过滤阈值”表用于 DNA 过滤过程中使用的 E 值，“聚类信息”表用于聚类阈值和相关量化）。

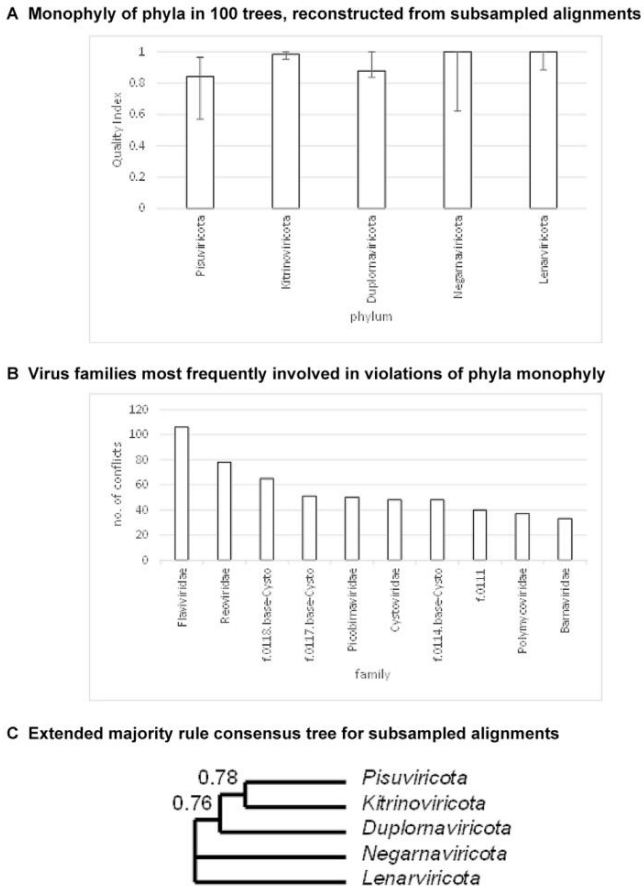
其他资源

为了提供持久的社区资源，我们创建了一个附带的交互式门户网站(riboviria.org)允许用户根据系统发育和数据类型下载本工作中生成的部分数据（例如，属于某个家族的所有重叠群的域注释的子集）。通过门户网站支持对数据的编程和图形访问。该网站的代码也可根据 MIT 许可证在github.com/Benjamin-Lee/riboviria.org 上获取。对于所有分类水平，该平台包括原始核酸序列、系统发育树、元数据和注释。

补充数字



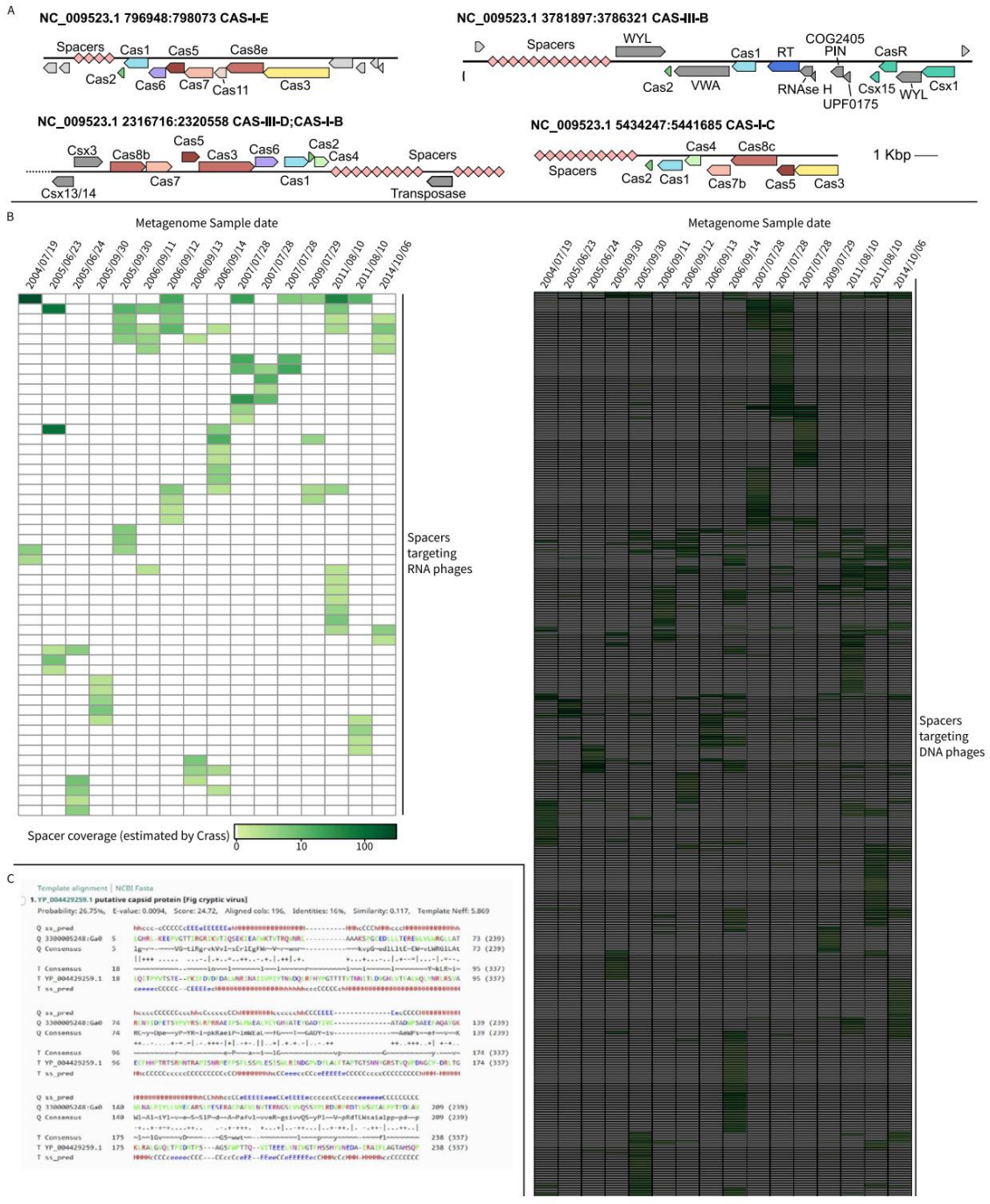
图S1。 RCR90/RvANI 簇中重叠群的分布，与图1B和 1C 以及表 1有关垂直轴（对数刻度）上的成员重叠群）。



图S2。深层系统发育重建的稳健性，与图 2 (A) 质量指数（形成单系进化枝的门成员分数和该进化枝中其他门成员分数的乘积）相关。条形图显示至少有 20 名成员的家庭中一名成员的 100 个独立样本的中值；胡须表示 5% 和 95% 的百分位数。

(B) 病毒家族，最常涉及单系违规（其中一片叶子要么在其门的进化枝之外，要么在另一门的进化枝之内）。显示违规数量。

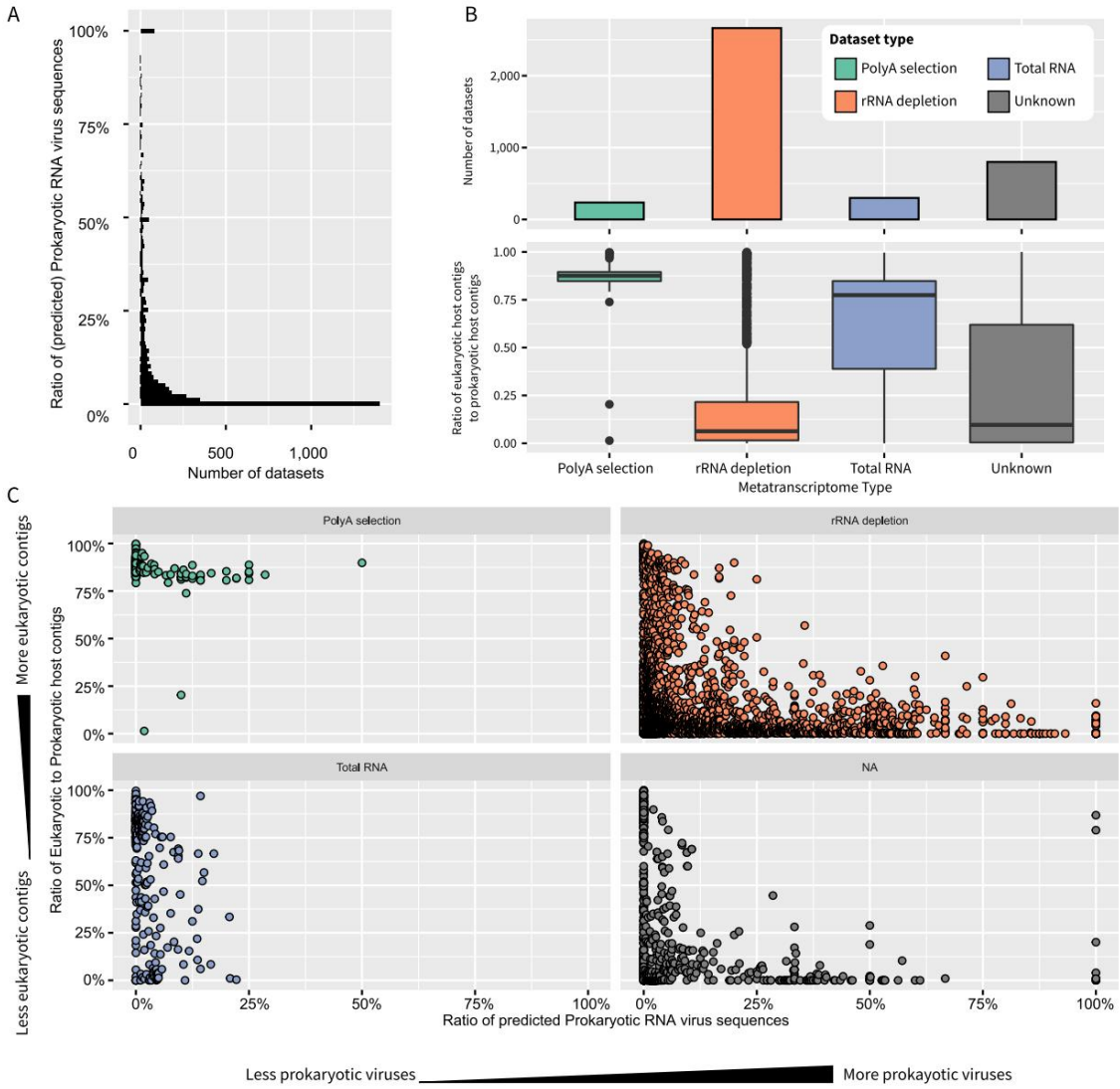
(C) 先前已知的五个门的扩展多数共识树。从 85 个（共 100 个）具有非嵌入单系门的样本中恢复了共有树，并将支持值乘以 0.85。



图S3。玫瑰鸾属 sp. RS-1 CRISPR 阵列和相关病毒，与图 3 (A) Roseiflexus sp . 4 个 CRISPR-Cas 区域图谱相关。 RS-1 (NC_009523.1) 包括预测的 CRISPR 阵列（红色菱形）和 Cas 基因（彩色基因）。

(B) Mushroom Spring 和 Octopus Spring 宏基因组的覆盖热图，用于与Roseiflexus sp 相关的间隔区。 RS-1（参见表S3 ）。与预测的RNA噬菌体匹配的间隔区显示在左侧，与DNA噬菌体匹配的间隔区显示在右侧以供参考。

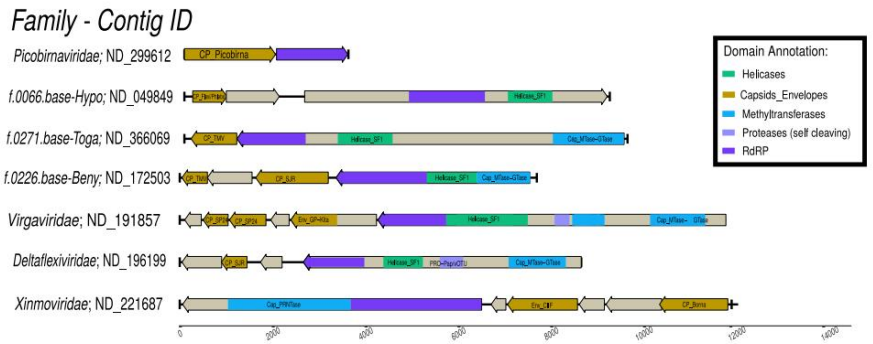
(C) 使用 hhpred 对来自感染Roseiflexus sp 的预测新型 RNA 噬菌体的推定衣壳蛋白进行比对的示例。 RS-1 和最接近的公开同源物：无花果隐病毒衣壳蛋白。



图S4。不同宏转录组类型和相关病毒类型的鉴定，与图 3和4相关（A)预测感染原核宿主的病毒在各个样本中的比例分布。

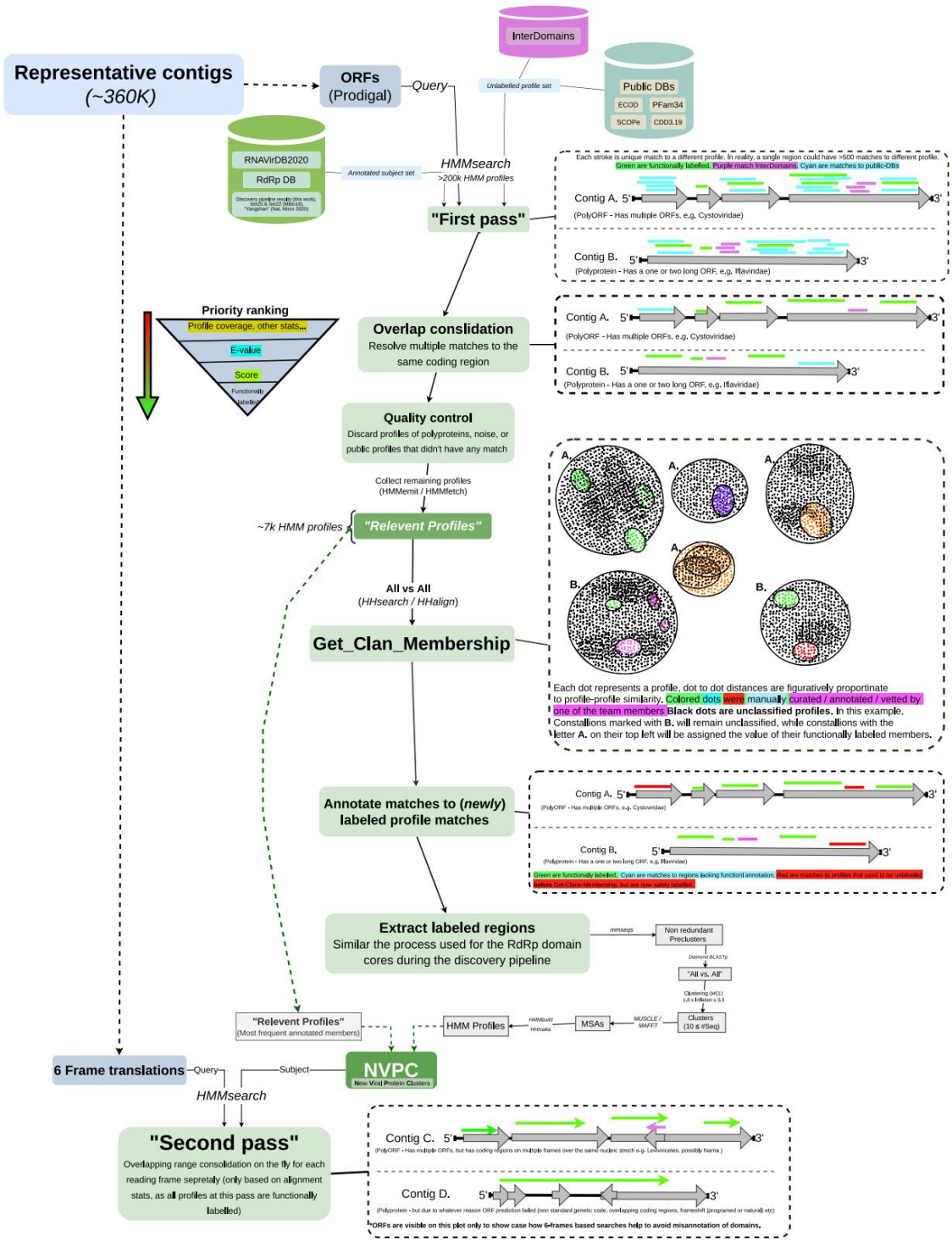
(B) 作为真核生物或原核生物（宿主)的非病毒重叠群在样本中的分布，根据用于生成宏转录组的方案进行分离。方案信息从 Gold 获得，总结如下：“poly(A) 选择”：基于 poly(A) 尾部的转录物富集，“rRNA 去除”：使用试剂盒和/或用于消除 rRNA 模板的方案，“总 RNA”：从提取的 RNA 中制备的 cDNA 文库，无需进行 Poly(A) 选择或 rRNA 消除步骤，“未知”：无可用信息。

(C) 真核生物/原核生物 RNA 病毒的比率（x 轴)与真核生物/原核生物宿主重叠群的比率（y 轴)之间的关系。每个数据集类型都显示在单独的面板中。

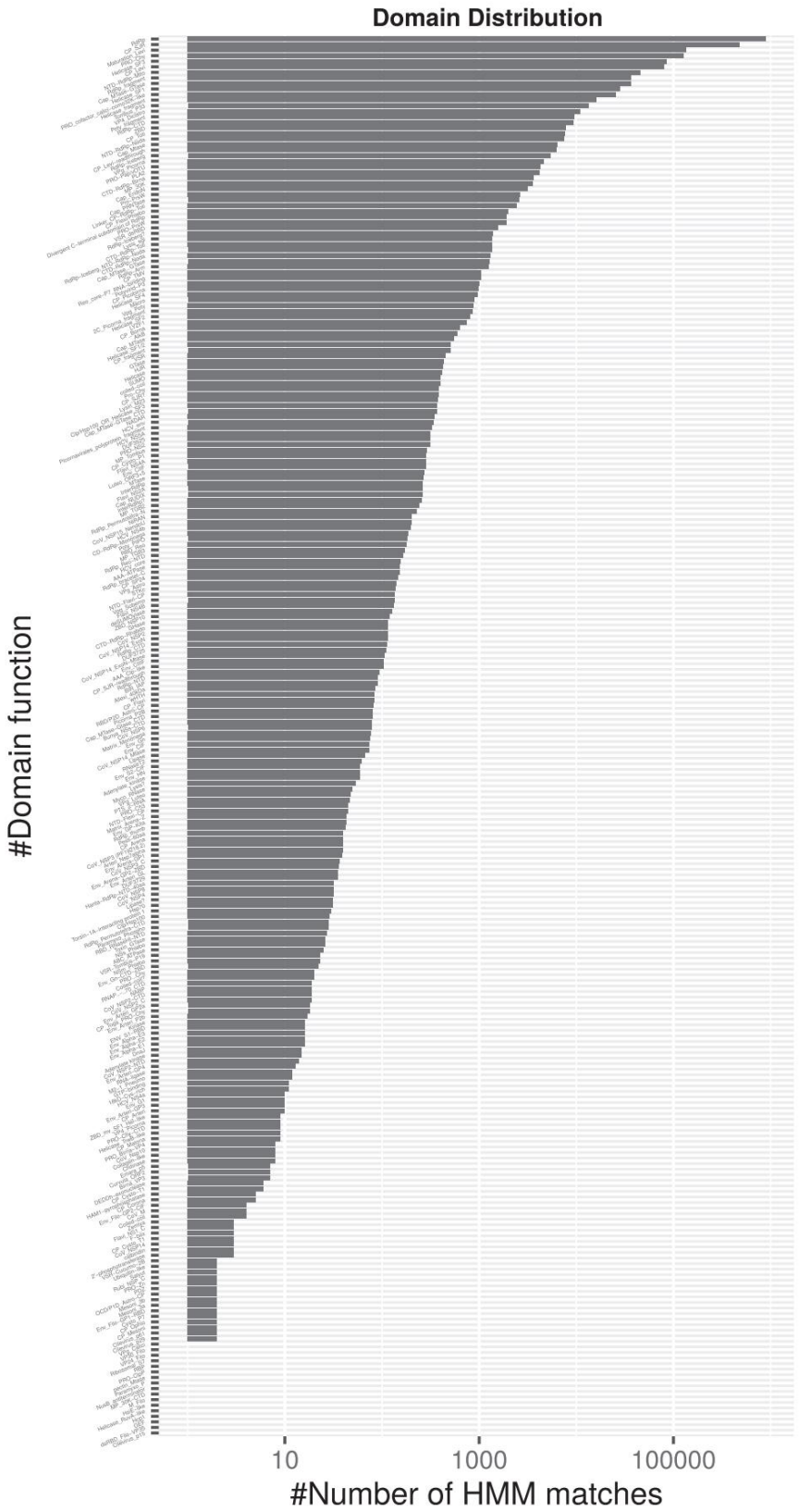


图S5。 RNA病毒中结构模块的获取和替换,与图5 “Picobirnaviridae”相关； ND_299612 和 f.0226.base-Beny； ND_172503 举例说明了编码衣壳蛋白（CP)和RdRP的基因组片段的融合,其在先前描述的小核糖核酸病毒和苯病毒中的单独片段上编码。 f.0066.base-Hypo; ND_049849 和 Deltaflexiviridae； ND_196199”分别编码 Flexi/Phlebo 样 CP 和单果冻卷 (SJR) CP,尽管各个家族的其他成员包含无衣壳病毒。 f.0271.base-Toga; ND_366069 和 Virgaviridae； ND_191857 代表具有 CP 基因非同源替换的基因组。在“新动病毒科”中； ND_221687，“新莫病毒”典型的 III 类融合糖蛋白基因已被编码 II 类融合糖蛋白 (CIIF) 的基因取代。

缩写:Env,包膜蛋白； GP,糖蛋白； PRO-Pap/vOTU,类木瓜蛋白酶； SF1,超家族1； Cap_MTase-GTase,具有甲基转移酶-鸟苷基转移酶活性的加帽酶。



图S6。扩展注释管道，与图 5流程图相关，该流程图可视化了项目的域识别和功能注释部分中使用的过程。



（下页的图例）

图S7。已识别的域分布，与图 5相关。最终域命中的预测病毒功能或结构（纵
轴、倾斜文本标签）与观察到的可靠 HMM 搜索匹配总数（横轴、对数刻度）相比。