PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2023, 50(2): 358~370
www.pibb.ac.cn

洋山港宏病毒组分析揭示CRISPR-Cas系统病毒靶标序列的特异性*

李　聪[1)]　田敬孜[2)]　王永杰[1,2,3)**]

([1)] 上海海洋大学食品学院，上海 201306；[2)] 农业农村部水产品质量安全贮藏保鲜风险评估实验室（上海），上海 201306；
[3)] 青岛海洋国家实验室海洋生物学与生物技术功能实验室，青岛 266200）

摘要　目的　为了探究CRISPR-Cas系统的靶标特点。**方法**　运用了包括BALST在内的多种生物信息学分析技术和方法对洋山港宏病毒组数据进行了分析。**结果**　洋山港宏病毒组数据集中共有25 391条双链DNA病毒序列，其中238条序列上的265个开放读码框（ORF）与134条规律间隔短回文重复（CRISPR）间隔序列产生了315个匹配。经注释后获得了128个ORF和135个匹配的功能信息，占比前5位的依次为终止酶（terminase）、衣壳蛋白（capsid protein）、门蛋白（portal protein）、肽酶（peptidase）和DNA甲基化转移酶（DNA methyltransferase）。匹配多在病毒特定功能基因的保守域或关键结构域中。这表明，CRISPR-Cas系统发挥免疫功能时表现出针对病毒特定基因、功能和结构域的靶标特异性。此外，属于Class 1门类下的Type_I型系统的CRISPR-Cas间隔序列匹配数量远高于其他类型系统，占总数的89.0%。**结论**　本文的研究结果揭示了CRISPR-Cas系统的靶标偏好，增进了对其的认识，为更好地理解病毒-宿主免疫互作机制提供了新的证据和线索。

关键词　CRISPR，间隔序列，功能基因，DNA病毒
中图分类号　Q71，Q936
DOI：10.16476/j.pibb.2022.0025

海洋环境中，病毒感染和裂解造成了约40%原核细胞生物的死亡[1-2]。为应对病毒侵染而产生的生存压力，原核宿主进化出了多样的防御系统，例如，限制性修饰系统（restriction-modification）、流产感染（abortive infection）[3]以及规律间隔短回文重复（CRISPR）系统。在这些防御系统中，CRISPR-Cas系统是最具适应性和特异性的，在细菌和古菌中作为后天免疫系统对抗病毒，以及诸如质粒等其他外源性可移动遗传元件（mobile genetic elements）[4]。目前，CRISPR系统可针对特定DNA或RNA区域进行核酸酶切割的特点已成功应用于特异性编辑动植物基因组。CRISPR基因座有CRISPR相关（Cas）基因和一个或多个CRISPR阵列（array）组成，该序列阵列由不同的间隔序列（spacer）和高度保守的重复序列（repeat）组成。正是这些多变的间隔序列使得CRISPR系统具有适应性以及特异性免疫机制。间隔序列是噬菌体或病毒DNA同源序列的片段，在可移动遗传元件中被称为原间隔序列（protospacer），长度通常在26~70 bp。每个间隔序列两侧都有相同的重复序列，重复序列可形成发夹二级结构[5]。在完整的CRISPR阵列中，间隔序列的数量2~200不等。

CRISPR系统发挥免疫功能时包含了3个不同的作用阶段，适应阶段、表达阶段和干扰阶段[6-7]。在适应阶段，来自病毒或质粒等外源DNA片段作为新的间隔序列被整合到CRISPR阵列中。在表达阶段，CRISPR阵列先转录为长的RNA序列（pre-crRNA），再进一步加工产生成熟的短的RNA序列（crRNA）。进入干扰阶段后，crRNA与Cas蛋白形成CRISPR核糖核蛋白（crRNP）复合物，特异性靶向同源的病毒或质粒核酸序列并对其进行切割。与其他免疫机制类似，CRISPR系统具有

* 国家自然科学基金（41376153，31570112）资助项目。
** 通讯联系人。
Tel：021-61900505，E-mail：yjwang@shou.edu.cn
收稿日期：2022-01-18，接受日期：2022-04-26

Cas蛋白序列、基因组成和基因组位点结构的多样性。因此根据在效应阶段是否有多个Cas蛋白参与，CRISPR-Cas系统被分为Class 1和Class 2两个大的门类（class）、6个类型（type）、44个亚型（subtype）[8]。

Wu等[9]在对洋山港（Yangshan Harbor，YSH）水体进行CRISPR系统间隔序列靶标分析时发现，部分间隔序列靶定在病毒序列的DNA甲基化转移酶上，表现出一定的靶向偏好。为了进一步探索CRISPR间隔序列对某一特定病毒基因的靶标偏好是特例还是代表了一种普遍现象，本文首先建立了公开数据库中原核生物基因组中含有的CRISPR间隔序列数据库，并基于对洋山港水体宏病毒组原间隔序列的靶标分析，以探寻海洋环境中CRISPR间隔序列的靶标特异性和多样性。

1　材料与方法

本文利用原核细胞间隔序列集对洋山港宏病毒组进行CRISPR-Cas系统间隔序列靶标分析，各步骤处理分析流程见图1。

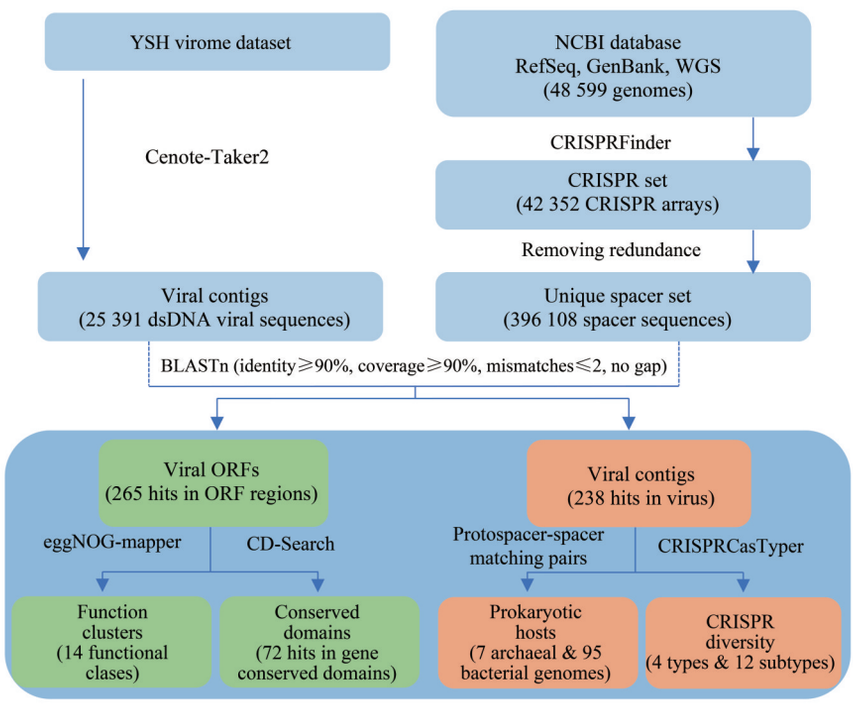


Fig. 1　Computational pipeline for Yangshan Harbor spacer targeting analysis

1.1　病毒序列的鉴定

本课题组前期工作已构建洋山港（海水样品采集点见图S1，采样点海水理化指标见表S1）宏病毒组数据库[9]。运用软件Cenote-Taker2[10]对拼接后长度大于1 000 bp的序列进行病毒序列的鉴定和提取。

1.2　间隔序列集的构建

从NCBI数据库（下载地址：ftp://ftp.ncbi.nlm.nih.gov/genomes/all/）中原核生物所携带的CRISPR-Cas系统中获得了720 391条CRISPR间隔序列。对这些间隔序列进行长度筛选（小于100 bp）及去除冗余后，构建包含396 108条序列的间隔序列数据集。

1.3　CRISPR间隔序列的靶标分析

以获取的非冗余间隔序列集作为BLAST查询序列，对洋山港宏病毒组数据库进行BLAST（identify≥90%，coverage≥90%）[11]扫描获得匹配（hits）以确定间隔序列所靶标的病毒序列及开放读码框（ORF）。

1.4　病毒基因的功能注释与分类

运用Batch CD-Search Tool（默认参数）[12]对间隔序列匹配到的病毒蛋白序列进行功能注释，而后对所注释的蛋白质功能进行归纳，汇总间隔序列靶标病毒基因的主要功能类群。

1.5　原核宿主预测及CRISPR系统的分型

基于BLAST获得的间隔序列与原间隔序列的匹配信息，确定细菌（或古菌）的基因组序列，并

与公开数据库进行比对得到原核宿主的生物学分类地位。同时，运用在线软件CRISPRCasTyper（默认参数）[13]对所有包含了可与病毒序列产生匹配的间隔序列宿主基因组内CRISPR-Cas系统进行分型分析。本文所用分析软件及参数详见表S2。

2　结　　果

2.1　病毒序列鉴定

经Cenote-Taker2鉴定，有25 391条序列被注释为双链DNA病毒。这些病毒分为11个纲、14个目、21个科。其中，被注释为“有尾噬菌体目（*Caudovirales*）”的序列有21 480条，约占所有可分类序列的84.6%；余下包括“藻类病毒目（*Algavirales*）”在内的感染真核宿主病毒序列及其他暂无分类学地位病毒序列共3 911条（图2）。选取所有原核生物病毒序列进行后续分析。

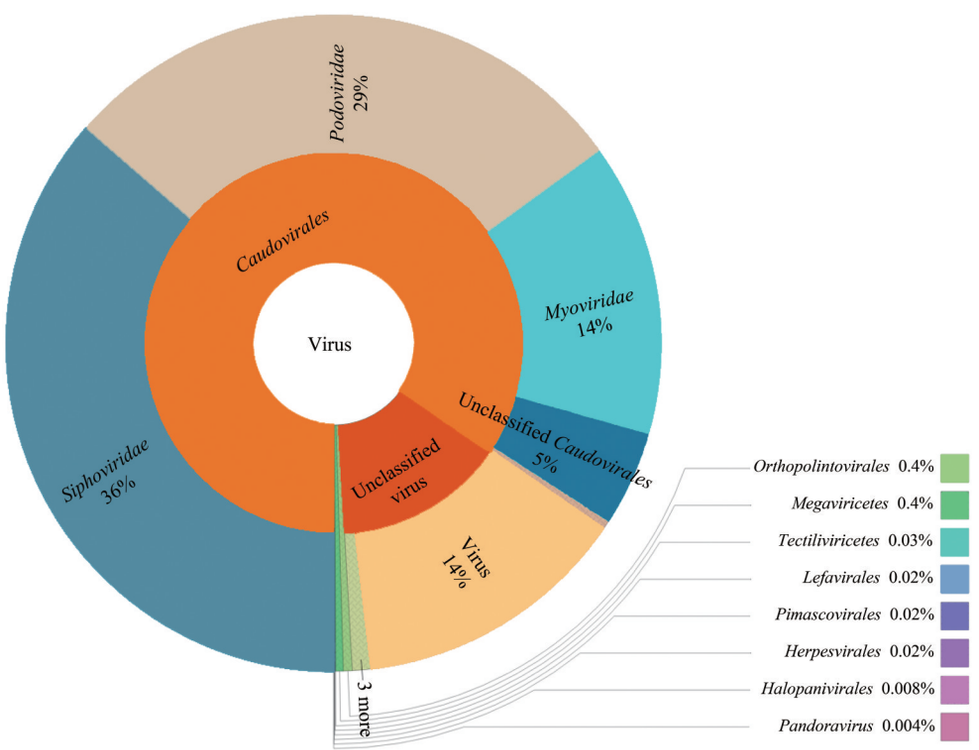


Fig. 2　The taxonomic composition of dsDNA viral contigs in YSH virome displayed by Krona [14]

2.2　间隔序列靶标分析

经BLAST扫描后，共有134个间隔序列与238条病毒序列产生了315个“间隔序列-原间隔序列”间的匹配（表S3），并显示出“一对多”的靶标特点（图3）。从图3a可以看出共有60个间隔序列呈现“一对多”的匹配特点，编号为352826的间隔序列与16个来自不同病毒序列的ORF均产生了匹配（图3b）。

此外，靶标分析显示在宿主与其噬菌体间也存在着“一对多”的特点（图4）。源自宿主*Methylomicrobium agile*（ATCC35068）基因组CRISPR阵列上的第2个、第35个、第37个以及第52个间隔序列分别与编号为k_141_215877、k_141_252123、k_141_543810和k_141_446164的4条病毒序列产生了匹配。这意味着该宿主近期对来自属于长尾病毒科的某一噬菌体建立了免疫，并在其进化早期曾连续遭受来自属短尾病毒科病毒和暂未分类地位病毒的侵染。

这些匹配上的间隔序列来自细菌宿主的有127个，古菌宿主的则为7个。间隔序列靶标的序列中，共有238条序列被注释为双链DNA病毒，包括107条长尾噬菌体科（*Siphoviridae*）病毒序列、63条短尾噬菌体科（*Podoviridae*）病毒序列、28条肌尾噬菌体科（*Myoviridae*）病毒序列、3条埃凯曼病毒科（*Ackermannviridae*）病毒序列、9条未分类有尾噬菌体目（unclassified *Caudovirales*）病毒序列以及28条未分类双链DNA病毒（unclassified dsDNA virus）序列（表S4）。

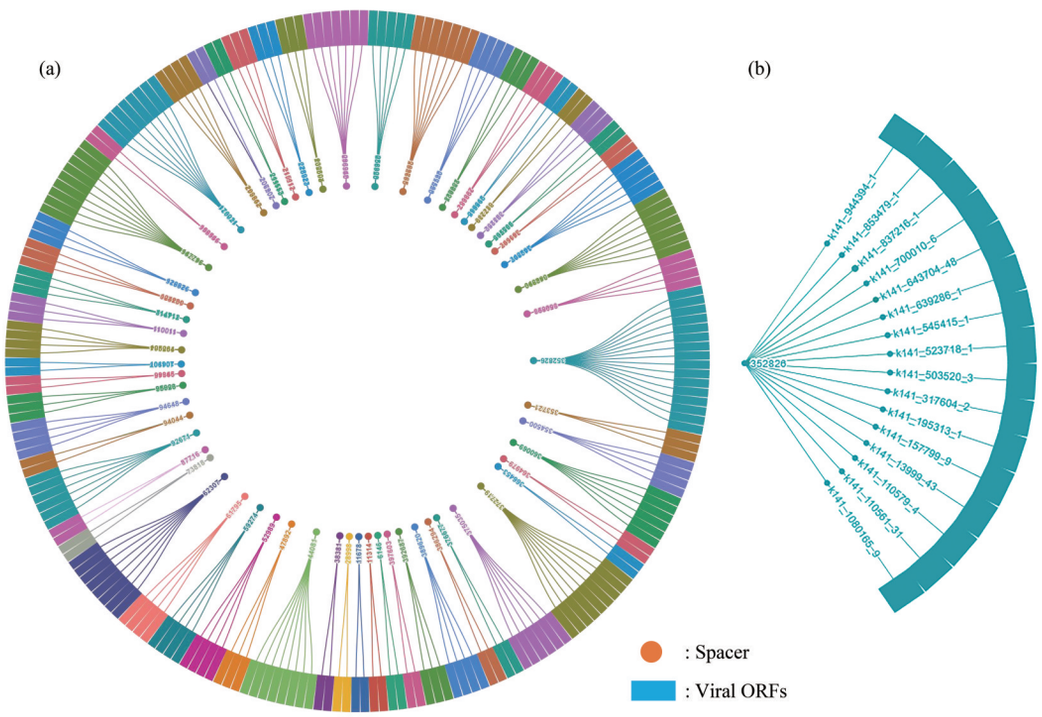


Fig. 3　Spacers matched virome ORFs in one-to-many models

The same colored links denote spacer and viral ORFs matching pairs assigned by BLSATn.

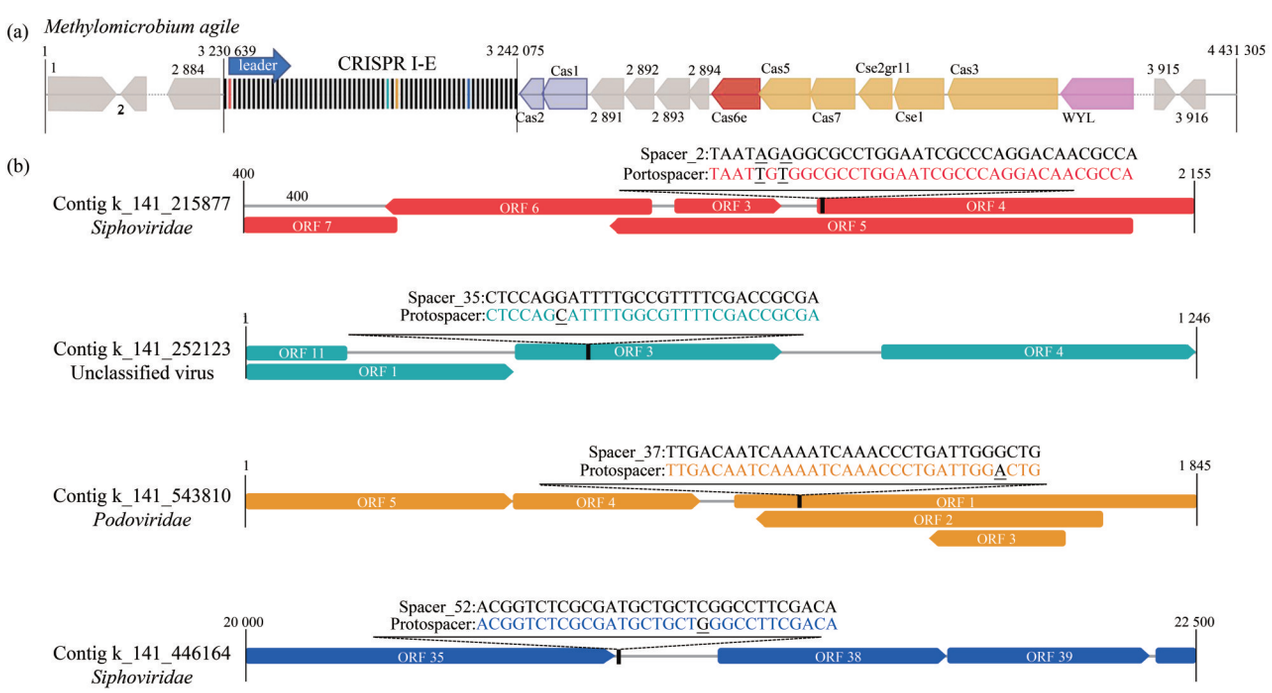


Fig. 4　Prokaryotic genome matched viral contigs in one-to-many models

(a) Genome map of CRISPR-Cas system in prokaryotic host. The sticks in CRISPR array represented the different spacer sequences, and the spacers that matched with the viral sequence in (b) were represented by the same color. (b) Partial genome maps of viral contigs. The black parts were where spacers matched with protospacers, and bases in black were mismatch sites.

所有的被间隔序列靶标到的265个病毒ORF中，共有128个ORF可被注释，约占总数的48.3%。经eggNOG[15-16]比对后，112个ORF被分到14个功能类群中（图5，表S4）。在该分类中，类群“X（mobilome）”数量最多，占总数的44.6%（50/112），类群“L（replication，recombination and repair）”和类群“S（function unknown）”数量相同，各约占总数的16.1%（18/112），类群“M（cell wall/membrane/envelope biogenesis）”次之，约占总数的8.9%（10/112）；余下功能类群总数占比均小于2%。结果表明，间隔序列特异性靶标病毒特定功能类群的基因以发挥CRISPR-Cas系统的免疫功能。

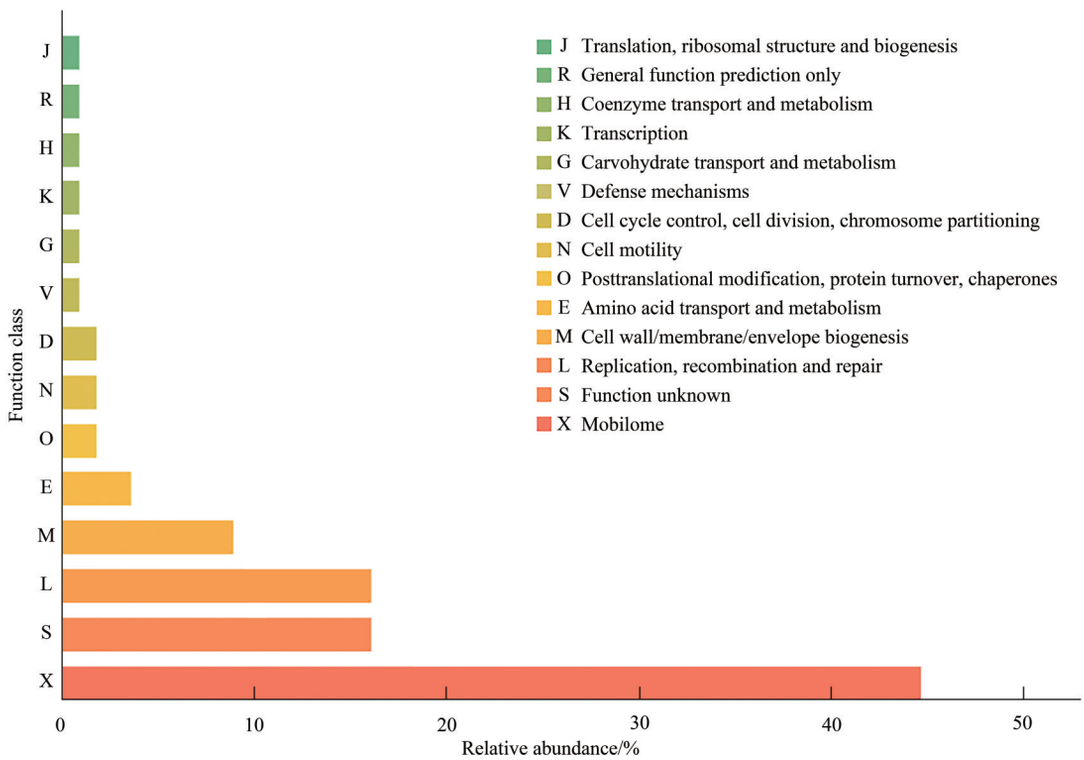


Fig. 5　Function classes of the viral ORFs targeted by spacer sequences

在所有的“间隔序列-原间隔序列”匹配对中，共有135个匹配可被注释，占总数的42.9%（图6）。同时，这些匹配中的ORF又可被概括为两个大的功能群。第一类为参与病毒DNA复制（replication）、转录（transcription）、修饰（modification）的酶，包括DNA解旋酶（DNA helicase）、DNA聚合酶（DNA polymerase）、DNA甲基化转移酶（DNA methyltransferase）等；第二类则与病毒颗粒组装（packing）和熟化（maturation）有关，例如，终止酶（phage terminase）、门蛋白（portal protein）、衣壳蛋白（phage capsid protein）、尾蛋白（phage tail protein）及少数其他功能蛋白共计73种（表S5、S6）。

在将“间隔序列-原间隔序列”匹配对与病毒功能基因的保守域（conserved domain）[12，17]比较后发现72个匹配的匹配位点在其对应病毒功能基因的保守域内，约占可注释匹配的53.3%。这一趋势在某些病毒功能基因中则更为明显，例如，在14个靶标至“衣壳蛋白”的匹配中，有12个靶标位点位于其保守域内；在19个靶标为“终止酶”的匹配中，有12个靶标位点位于其保守域内（图6）。

针对间隔序列靶标最高的终止酶，本文对其匹配位点又进行了更为细致的分析。结果发现，所有被靶标的终止酶分为6个超家族（superfamily）即“Terminase_1”、“Terminase_2”、“Terminase_3”、“Terminase_6”、“Terminase_GpA”和“17_Superfamily”，并以“Terminase_3”和“Terminase_6”为主（图7）。

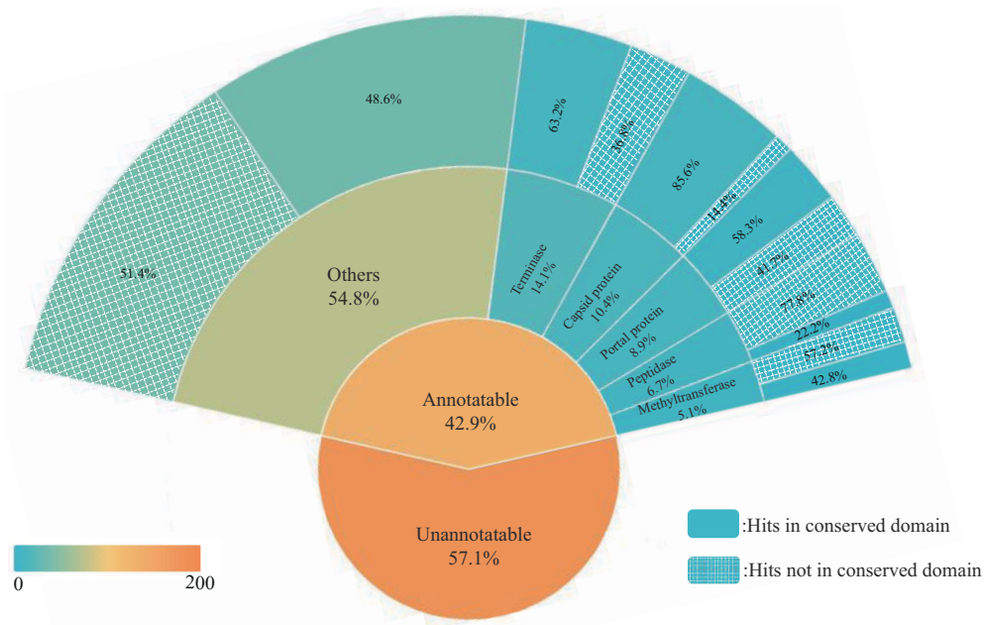


Fig. 6　Specific viral genes and their conserved domains preferentially targeted by spacer sequences

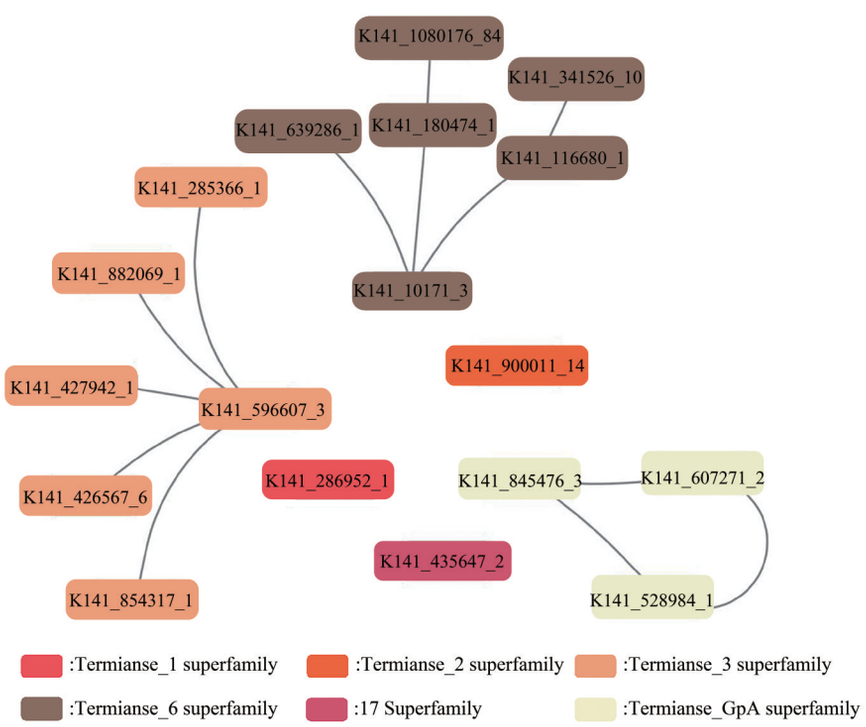


Fig. 7　Gene functional domain clusters of viral terminase superfamilies，targeted by spacer sequences

同时，值得注意的是，除编号为k141_180474_1的ORF因长度较短无法鉴别外，其余17个终止酶的匹配位点均出现在了终止酶大亚基（TerL）中（表1）。

2.3　宿主及CRISPR分型分析

315个间隔序列-原间隔序列间的匹配中，共有来自102个原核CRISPR阵列的134个间隔序列，其中7个来自古菌宿主，95个来自细菌宿主。在这

Table 1　Targeting sites of viral terminase

Taxonomy	Viral ORF	Superfamily	Subunit
Siphoviridae	k141_596607_3	Terminase_3	TerL
Siphoviridae	k141_286952_1	Terminase_1	TerL
Siphoviridae	k141_528984_1	Terminase_GpA	TerL
Siphoviridae	k141_639286_1	Terminase_6	TerL
Siphoviridae	k141_426567_6	Terminase_3	TerL
Siphoviridae	k141_845476_3	Terminase_GpA	TerL
Siphoviridae	k141_607271_2	Terminase_GpA	TerL
Siphoviridae	k141_10171_3	Terminase_6	TerL
Siphoviridae	k141_116680_1	Terminase_6	TerL
Podoviridae	k141_900011_14	Terminase_2	TerL
Podoviridae	k141_854317_6	Terminase_3	TerL
Podoviridae	k141_882069_1	Terminase_3	TerL
Podoviridae	k141_285366_1	Terminase_3	TerL
Podoviridae	k141_341526_10	Terminase_6	TerL
Podoviridae	k141_180474_1	Terminase_6	—
Podoviridae	k141_427942_1	Terminase_3	TerL
Myoviridae	k141_435647_2	17_Superfamily	TerL
Myoviridae	k141_1080176_84	Terminase_6	TerL

7个古菌中有6个为泉古菌门（Crenarchaeota）下的热变形菌（Thermoprotei）以及一个广古菌门（Euryarchaeota）下的甲烷杆菌（Methanobacteria）；而在细菌中占比最多的门类为变形菌门（Proteobacteria），所占比例约为49.5%（47/95），余下主要细菌门类及其占比分别为：放线菌门（Actinobacteria）约占26.3%（25/95）、厚壁菌门（Firmicutes）约占13.7%（13/95）、梭杆菌门（Fusobacteria）约占4.2%（4/95）以及少数其他门类细菌约占6.3%（6/95）。

基于Cas及其同源蛋白和重复序列的比对，对102个原核宿主中的CRISPR-Cas基因座进行分析。除了20个原核宿主因基因组不完整未能找到Cas基因簇外，其余82个原核宿主基因组内CRISPR-Cas系统均被确认并对其进行了分型分析（表S7）。分型结果表明，CRISPR-Cas系统门类、类型和亚型的分布趋势明显（图8）。值得注意的是，属于Class 1门类下的Type_I型系统的数量要远高于其他类型系统，占总数的89.0%（73/82）。在这些Type_I型系统中，Type_I-E亚型多达39个，且分布最为广泛。在细菌和古菌群中，Class 1类系统均比Class 2类系统更丰富。

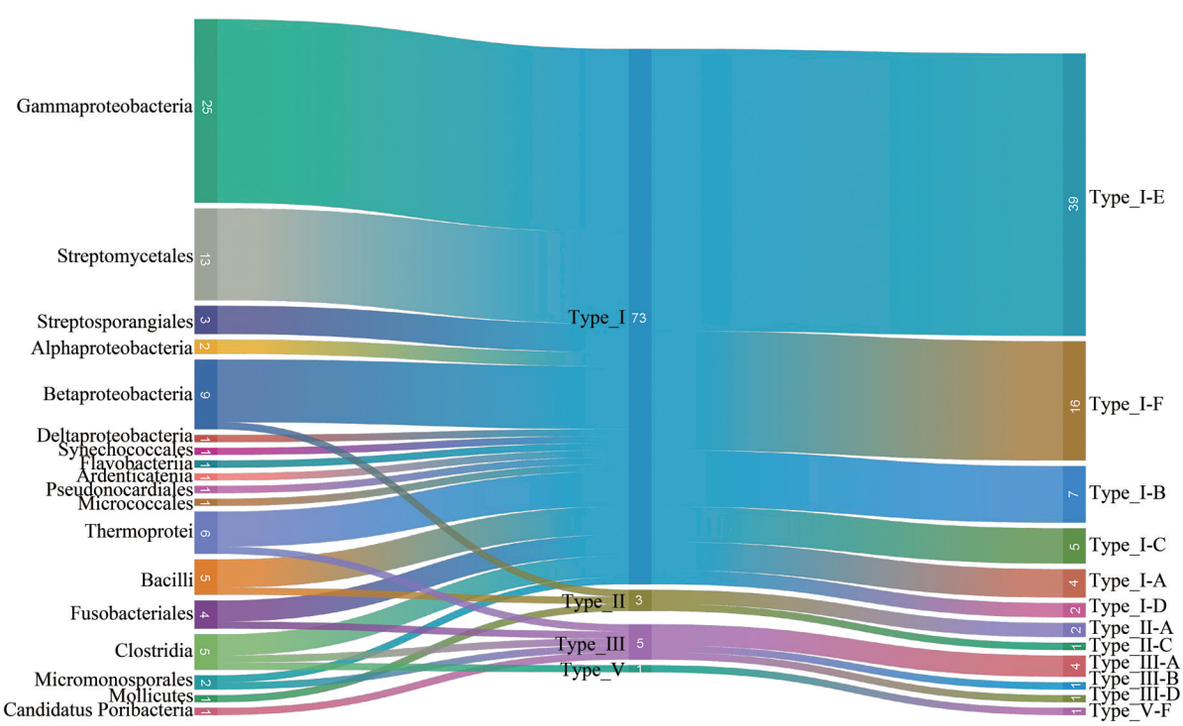


Fig. 8　Diversity of the types and subtype of CRISPR−Cas system identified in the archaeal and bacterial phyla with protospacer matching

3　分析与讨论

3.1　方法

CRISPR-Cas系统是原核宿主在抵抗外源性遗传物质中发挥重要作用的获得性免疫系统，在其系统基因座中所包含的病毒片段是系统抵抗外源性遗传物质反复侵入的“免疫记忆”[18]。本研究基于原核生物基因组间隔序列集反向搜索洋山港水域表层水中的病毒，这一方法对于病毒宿主进行预测可在宿主科（family）水平达到97%的准确性[11]。通过该方法为238条病毒序列确定了其对应的原核宿主，这些宿主绝大多数为海洋环境中常见的细菌，仅有少量的海洋古菌。推其原因，这可能与公共数据库中关于古菌基因组特别是全基因组数量较少有关。在今后类似的研究工作中应及时更新间隔序列数据集，特别是来自古菌宿主中的间隔序列集以扩大对病毒-宿主间互作关系的认识。同时，仅在唯一一条真核藻细胞病毒（*Mimiviridae*）序列上发现其存在与间隔序列产生潜在匹配的片段。这表明借助间隔序列与原间隔序列匹配可真实反映自然环境下病毒与其原核宿主的侵染关系。

3.2　CRISPR靶标特异性

原间隔序列是病毒基因上的一小段核酸片段，作为间隔序列被整合到CRISRP阵列中。在Type_I型和Type_II型CRISPR-Cas系统中，被称为原间隔序列相邻基序（protospacer adjacent motif，PAM）的序列对间隔序列的获取极为重要[5, 19]。然而，考虑到PAM长度及序列多样性，在一个病毒的基因组中可能存在着成百上千个潜在的PAM位点。因此，虽然PAM在确定CRISPR-Cas系统靶标基因的选择上发挥了作用，但具体是什么确定了具体基因的选择及在CRISPR阵列上的保留仍然是不确定的。借助所发现到的对于病毒宏基因组中特定功能基因的选择，可以成为原核宿主在面对生存压力时有效免疫最简单的解释。在原核宿主基因组中发现的CRISPR间隔序列成功降低了病毒感染和裂解对细菌和古菌种群的影响。本研究中的相关数据也提供了那些对病毒侵染和裂解原核宿主过程中至关重要基因的证据。

间隔序列与病毒ORF能够以一对多或多对一的方式产生匹配，即部分CRISPR-Cas系统针对多个不同的病毒产生免疫或多个不同的CRISPR-Cas系统针对单一病毒个体的特定基因发挥免疫功能。第一种情况下间隔序列针对多个不同的基因所共有的片段进行靶标，这一方式提升了CRISPR-Cas系统的免疫范围。而在后一种情况下，某一特定功能类群的病毒基因似乎被CRISPR-Cas系统高度针对，这些被“过度”靶标的ORF也表明了它们对病毒复制至关重要，因此成为了原核宿主CRISPR免疫的特异性靶点。

在CRISPR阵列上处于不同位置的间隔序列代表着宿主第一次识别并切割获取原间隔序列的先后时间[18]。在目前已知的所有类型CRISPR中，最新获得的间隔序列总是被插入整合至CRISPR阵列的最前端，即紧邻Leader序列的第一个间隔序列位点[20]。这样的排列方式保证了针对近期侵染原核宿主噬菌体的间隔序列将被优先转录成为Pre-crRNA，在被加工后参与Cas蛋白形成CRISPR核糖核蛋白聚合物参与免疫活动。由于CRISPR-Cas系统的可遗传性[5, 18]，处在CRISPR阵列上远离Leader端的那些间隔序列可能“继承”其母细胞（mother cell）。连续多次识别并获得源自不同病毒的原间隔序列并将其整合在CRISPR阵列之中意味着原核宿主曾在较短的时间内接连遭受侵染并成功建立获得性免疫。

噬菌体终止酶是一类多功能的寡聚蛋白，广泛存在于各种双链DNA病毒中[21]。在噬菌体裂解周期中，DNA的包装是一个极为重要的过程，末端酶通过切断DNA连接体以启动DNA包装过程，并为整个过程提供大量动力ATP以驱动DNA压缩至狭小的头部衣壳蛋白中[22-23]。TerL亚基作为终止酶的关键亚基具有ATP酶、核酸内切酶和DNA解旋酶活性，DNA包装驱动需要核酸酶活性将噬菌体基因组进行剪切，其切割活性的发挥高度依赖于ATP酶供能，核酸酶或ATP酶任意缺失或突变均不能表现功能活性。门蛋白同样也参与到DNA包装过程中，不仅影响了包装效率同时也决定了进入衣壳蛋白中病毒基因组的大小，也能保证包装极性，防止DNA从衣壳蛋白中逸出[23]。衣壳蛋白作为外壳，对压缩进入其中的核酸起保护作用。因此，针对参与病毒包装过程重要阶段的相关功能基因的特异性靶标，保证了对噬菌体复制的“致命”打击。同时，TerL广泛分布于各类有尾噬菌体中[21]，特异性的靶标这一特定亚基意味着可为宿主提供高效广泛且长效的免疫保护。

为何CRISPR-Cas系统的间隔序列倾向于靶标在此阶段发挥作用的病毒功能基因？我们认为这可能与宿主在对于容忍病毒侵染和抵抗病毒入侵之间

选择有关。宿主和病毒之间的竞争在无时无刻地进行着[24]，在这场“竞赛”中宿主需要不断提升自己的抵抗力，而病毒则需要不断提升入侵的能力。除了常见的排斥型抵抗——对外源性DNA进行靶向裂解（例如CRISPR-Cas系统、R-M系统），也存在着另一种抵抗机制，“接纳性免疫”[25]。原噬菌体整合到宿主的基因组中，此时宿主进入溶原态；溶原体抑制体内噬菌体DNA的转录，进而使得噬菌体DNA无法表达，从而产生针对特定噬菌体的同源病毒重复感染的抵抗力。噬菌体在将DNA整合到宿主基因组上时可能会造成宿主基因的突变。而这种突变方式是宿主的一种进化动力，以此丰富宿主基因的多样性。无论是温和噬菌体还是烈性噬菌体，它们的最终目的是要在宿主体内完成DNA的复制，以及衣壳蛋白的合成。而对于噬菌体而言，一旦装配阶段受阻，则无法形成具有侵染能力的噬菌体，也意味着无法行使生物功能，进而被宿主“囚禁”在其体内，最终被分解成单个的核苷酸以及氨基酸被宿主重新利用。这样的方式对宿主而言是最“经济”的，被噬菌体所掠夺的参与生命活动的“原材料”最终都留在了自己体内。因此，宿主采取这样的方式进行免疫的同时也在利用入侵的噬菌体。同时，不释放病毒意味着对同一种群中的其他宿主细胞提供了保护。

DNA甲基化转移酶作为原核生物限制性修饰系统的重要组成部分也出现在了间隔序列靶标功能基因的高频匹配中。对于病毒DNA甲基化转移酶的特异性靶标可以看作一个典型的原核宿主“协同免疫”。在同时拥有限制性修饰系统和CRISPR-Cas系统的宿主中，当携带编码DNA甲基化转移酶的病毒入侵时，病毒通过对自身基因的甲基化来躲避限制性修饰系统的免疫[26]；但是，若在病毒在向宿主体内注射DNA时，CRISPR-Cas系统将迅速对病毒基因组上的DNA甲基化酶基因进行免疫识别并切割，使病毒无法对自身基因组甲基化修饰，宿主的限制性修饰系统将彻底清除残余的病毒DNA片段，以保证免疫效果。

此外，CRISPR间隔序列所靶标的ORF中超过40%已确定其功能，这表明与非靶向基因相比，被CRISPR所靶标的基因更有可能具有确定的功能作用。余下未知功能CRISPR的靶标基因一直以来被认为是病毒遗传的“暗物质（dark matter）”[27]，但可以肯定的是这一类被CRISPR所靶标的未知病毒暗物质基因可能在病毒侵染和裂解过程中发挥重要作用。

同时本文发现，在所有产生匹配的238条双链DNA病毒序列中有210条序列属于有尾噬菌体病毒，但是仅有5个间隔序列靶标在了编码噬菌体尾纤维蛋白的基因上，这一频率低于预期。尾纤维蛋白结构相对简单，在编码水平上选择压力小，也意味着编码尾纤维蛋白的基因具有较高的多样性。事实上，噬菌体尾纤维基因不仅是多变的，同时还可以通过逆转录因子进行靶向突变，以扩大病毒宿主范围[28-29]。也许正是这样的原因使得CIRPSPR-Cas系统放弃了以此类基因为主要靶标对象，转而选择其他基因。

值得一提的是，与非保守域相比某些病毒功能基因的保守域更容易被CRISPR间隔序列所靶标，对于这一现象本文认为，相较于非保守域，病毒功能基因的保守域不易发生基因突变，使间隔序列可为原核宿主提供长效的保护[19]。同时，病毒功能基因的保守域广泛存在于不同的病毒种群中，例如终止酶大亚基中的核酸内切酶结构域和ATP酶结构域，针对病毒功能基因保守域特异性靶标可为宿主提供较大的免疫范围。无论是扩大免疫范围还是延长免疫的时长，CRISPR-Cas系统针对病毒功能基因保守域的特异性免疫对于原核宿主而言是一种十分“实惠”的选择，这样的选择将大大降低因频频获取新的间隔序列来保证免疫效果所带来的负担，即因新的间隔序列增加而导致CRISPR阵列长度的增加。

3.3 CRISPR−Cas系统在宿主中的分布

在通过CRISPR靶标而确定的102个原核宿主中有20个由于基因组信息不完整未能对其所携带的CRISPR-Cas系统进行分型，余下82个宿主的CRISPR-Cas系统均得以分型确认[8]。结果表明，CRISPR-Cas系统在细菌及古菌中的分布是不均匀的，特定的CRISPR-Cas系统门类、类型和亚型在分布上呈现出了明显的趋势。例如，Class 2门类下的Type_II型、Type_V型只存在于细菌宿主中（本次研究结果中并未发现Type_VI型系统）。出现这一现象的原因可归因于RNaIII酶在古菌中的缺失，RNaIII是一种广泛存在于细菌中的核糖核酸酶，负责Type_II型和Type_V型系统及其亚型的前crRNA（pre-crRNA，加工成熟后引导Cas蛋白定位至免疫靶点）处理。而Type_I型系统在宿主中的分布最为广泛，尤其是对Type_I-E亚型的分析表明，相较于其他CRISPR-Cas系统，Type_I-E亚

型的系统进化十分缓慢，其CRISPR阵列可在10^3~10^5年内保持不变[30]。这一研究结果将帮助人们更好地理解为何CRISPR-Cas系统的间隔序列更多的靶标在了病毒功能基因的保守域中。即缓慢的进化意味着更少的机会获得新的间隔序列，这就要求现有的间隔序列应尽可能的在较长的时间内发挥免疫作用，因此在不易产生突变的病毒功能基因保守域似乎是个很好的选择。本研究发现，I型系统的数量远高于其他型系统，除病毒种类及其相对丰度在一定程度上可能会影响其差异外，其他特点还有以下几点。a. I型系统是一种较为活跃的CRISPR-Cas系统。该型系统具有较高的获取新的间隔序列的频率（包括直接识别MGE上的原间隔序列和在已失效的原间隔序列上识别新的原间隔序列），保证可对宿主提供及时有效的免疫保护[19]。b. I型系统对于间隔序列与原间隔序列间的错配容忍度较高。与其他型系统，特别是II型系统相比，I型系统对于原间隔序列及间隔序列间发生错配后仍可免疫的容忍度较高，只需保证原间隔序列上的“种子片段”（seed sequence）未发生突变即可完成对原间隔序列的识别及引导Cas蛋白对MGE进行免疫作用[31-32]。c. I型系统是一种由早期自适应免疫系统（ancestral adaptive immunity system）进化而来且高度分化的CRISPR-Cas系统。I型系统的进化源头为早期自适应免疫系统，经过长期演化，逐渐形成了I型以及III型系统；同时，由于结构简单（相较于III型系统而言，I型系统*Cas*基因簇较短）个别Cas蛋白的加入就可使I型系统分化成为不同亚型灵活地为原核宿主提供免疫保护[33]。d. I型系统的类转座子结构帮助其水平转移。在I型系统*Cas1*以及*Cas2*串联基因的两端有TIRs（terminal inverted repeats）结构，这种典型的转座子结构帮助其在原核宿主间可广泛地进行水平转移[25]。显然，灵活多变的免疫机制，具有可传播的特性和长期以来的进化选择，使得I型系统在原核宿主中得以广泛分布。这一结果与Makarova等[8]关于现有数据库中CIRSPR-Cas系统分布的研究结果相一致。

4　结　　论

本研究基于公共数据库中CRISPR间隔序列集，对洋山港港区表层水宏病毒组进行间隔序列-原间隔序列的匹配分析，结果表明特定的病毒功能基因被CRISPR-Cas系统特异性的靶标。这些被靶标的基因在宿主与病毒间的“军备竞争”中表现出了病毒群体的遗传“缺陷”。还发现，间隔序列可以识别自然环境中病毒群体中最为重要的基因及其最为重要的片段（保守域）。这些基因或功能域可用于进一步探索CRISPR-Cas系统以及病毒与宿主的共进化机制。此外，基于anti-CRISPR蛋白的结构及功能特点，推测anti-CRISPR是伴随着CIRSPR-Cas系统的形成而形成。同时随着CRISPR-Cas进化分化，anti-CRISPR也朝着亚型特异性的方向进行分化。

附件　请见本文网络版（http://www.pibb.ac.cn或http://www.cnki.net）：

PIBB_20220025_FigS1.jpg
PIBB_20220025_TableS1.xlsx
PIBB_20220025_TableS2.xlsx
PIBB_20220025_TableS3.xlsx
PIBB_20220025_TableS4.xlsx
PIBB_20220025_TableS5.xlsx
PIBB_20220025_TableS6.xlsx
PIBB_20220025_TableS7.xlsx

参　考　文　献

[1] Weitz J S, Stock C A, Wilhelm S W, *et al*. A multitrophic model to quantify the effects of marine viruses on microbial food webs and ecosystem processes. ISME J, 2015, **9**(6): 1352-1364

[2] Poorvin L, Rinta-Kanto J M, Hutchins D A, *et al*. Viral release of iron and its bioavailability to marine plankton. Limnol Oceanogr, 2004, **49**(5): 1734-1741

[3] Dorman C J. H-NS: a universal regulator for a dynamic genome. Nat Rev Microbiol, 2004, **2**(5): 391-400

[4] Sorek R, Kunin V, Hugenholtz P. CRISPR - a widespread system that provides acquired resistance against phages in bacteria and archaea. Nat Rev Microbiol, 2008, **6**(3): 181-186

[5] Mcginn J, Marraffini L A. Molecular mechanisms of CRISPR-Cas spacer acquisition. Nat Rev Microbiol, 2019, **17**(1): 7-12

[6] Wiedenheft B, Sternberg S H, Doudna J A. RNA-guided genetic silencing systems in bacteria and archaea. Nature, 2012, **482**(7385): 331-338

[7] Marraffini L A. CRISPR-Cas immunity in prokaryotes. Nature, 2015, **526**(7571): 55-61

[8] Makarova K S, Wolf Y I, Iranzo J, *et al*. Evolutionary classification of CRISPR-Cas systems: a burst of class 2 and derived variants. Nat Rev Microbiol, 2020, **18**(2): 67-83

[9] Wu S, Zhou L, Zhou Y, *et al*. Diverse and unique viruses discovered in the surface water of the East China Sea. BMC Genomics, 2020, **21**(1): 441

[10] Tisza M J, Belford A K, Dominguez-Huerta G, *et al*. Cenote-Taker 2 democratizes virus discovery and sequence annotation. Virus

Evol, 2021, **7**(1): veaa100

[11] Shmakov S A, Sitnik V, Makarova K S, *et al*. The CRISPR spacer space is dominated by sequences from species-specific mobilomes. mBio, 2017, **8**(5): e01397-17

[12] Marchler-Bauer A, Bo Y, Han L, *et al*. CDD/SPARCLE: functional classification of proteins *via* subfamily domain architectures. Nucleic Acids Res, 2017, **45**(D1): D200-D203

[13] Russel J, Pinilla-Redondo R, Mayo-Munoz D, *et al*. CRISPRCasTyper: automated identification, annotation, and classification of CRISPR-Cas loci. CRISPT J, 2020, **3**(6): 462-469

[14] Ondov B D, Bergman N H, Phillippy A M. Interactive metagenomic visualization in a Web browser. BMC Bioinformatics, 2011, **12**: 385

[15] Huerta-Cepas J, Forslund K, Coelho L P, *et al*. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. Mol Biol Evol, 2017, **34**(8): 2115-2122

[16] Huerta-Cepas J, Szklarczyk D, Forslund K, *et al*. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res, 2016, **44**(D1): D286-D293

[17] Lu S, Wang J, Chitsaz F, *et al*. CDD/SPARCLE: the conserved domain database in 2020. Nucleic Acids Res, 2020, **48**(D1): D265-D268

[18] Amitai G, Sorek R. CRISPR-Cas adaptation: insights into the mechanism of action. Nat Rev Microbiol, 2016, **14**(2): 67-76

[19] Mosterd C, Rousseau G M, Moineau S. A short overview of the CRISPR-Cas adaptation stage. Can J Microbiol, 2021, **67**(1): 1-12

[20] Alkhnbashi O S, Shah S A, Garrett R A, *et al*. Characterizing leader sequences of CRISPR loci. Bioinformatics, 2016, **32**(17): 576-585

[21] 宋少云,贾艳华,祝心舟,等. 末端酶及其在噬菌体包装中的作用. 生物技术通报, 2013, **11**: 40-45
Song S Y, Jia Y H, Zhu X Z, *et al*. Biotechnol Bull, 2013, **11**: 40-45

[22] Nadal M, Mas P J, Blanco A G, *et al*. Structure and inhibition of herpesvirus DNA packaging terminase nuclease domain. Proc Natl Acad Sci USA, 2010, **107**(37): 16078-16083

[23] Isidro A, Henriques A O, Tavares P. The portal protein plays essential roles at different steps of the SPP1 DNA packaging process. Virology, 2004, **322**(2): 253-263

[24] Borges A L, Zhang J Y, Rollins M F, *et al*. Bacteriophage cooperation suppresses CRISPR-Cas3 and Cas9 immunity. Cell, 2018, **174**(4): 917-925

[25] Koonin E V, Makarova K S, Wolf Y I. Evolutionary genomics of defense systems in archaea and bacteria. Annu Rev Microbiol, 2017, **71**: 233-261

[26] Dimitriu T, Szczelkun M D, Westra E R. Evolutionary ecology and interplay of prokaryotic innate and adaptive immune systems. Curr Biol, 2020, **30**(19): R1189-R1202

[27] Roux S, Hallam S J, Woyke T, *et al*. Viral dark matter and virus-host interactions resolved from publicly available microbial genomes. Elife, 2015, **4**: e08490

[28] Minot S, Grunberg S, Wu G D, *et al*. Hypervariable loci in the human gut virome. Proc Natl Acad Sci USA, 2012, **109**(10): 3962-3966

[29] Doulatov S, Hodes A, Dai L X, *et al*. Tropism switching in Bordetella bacteriophage defines a family of diversity-generating retroelements. Nature, 2004, **431**(7007): 476-481

[30] Westra E R, Buckling A, Fineran P C. CRISPR-Cas systems: beyond adaptive immunity. Nat Rev Microbiol, 2014, **12**(5): 317-326

[31] Steens J A, Zhu Y, Taylor D W, *et al*. SCOPE enables type III CRISPR-Cas diagnostics using flexible targeting and stringent CARF ribonuclease activation. Nat Commun, 2021, **12**(1): 107838

[32] Nussenzweig P M, Mcginn J, Marraffini L A. Cas9 cleavage of viral genomes primes the acquisition of new immunological memories. Cell Host Microbe, 2019, **26**(4): 515-526 e516

[33] Koonin E V, Makarova K S. Origins and evolution of CRISPR-Cas systems. Philos T R Soc B, 2019, **374**(1772): 20180087

Yangshan Harbor Virome Analysis Reveals CRISPR Spacer Targeting Specificity*

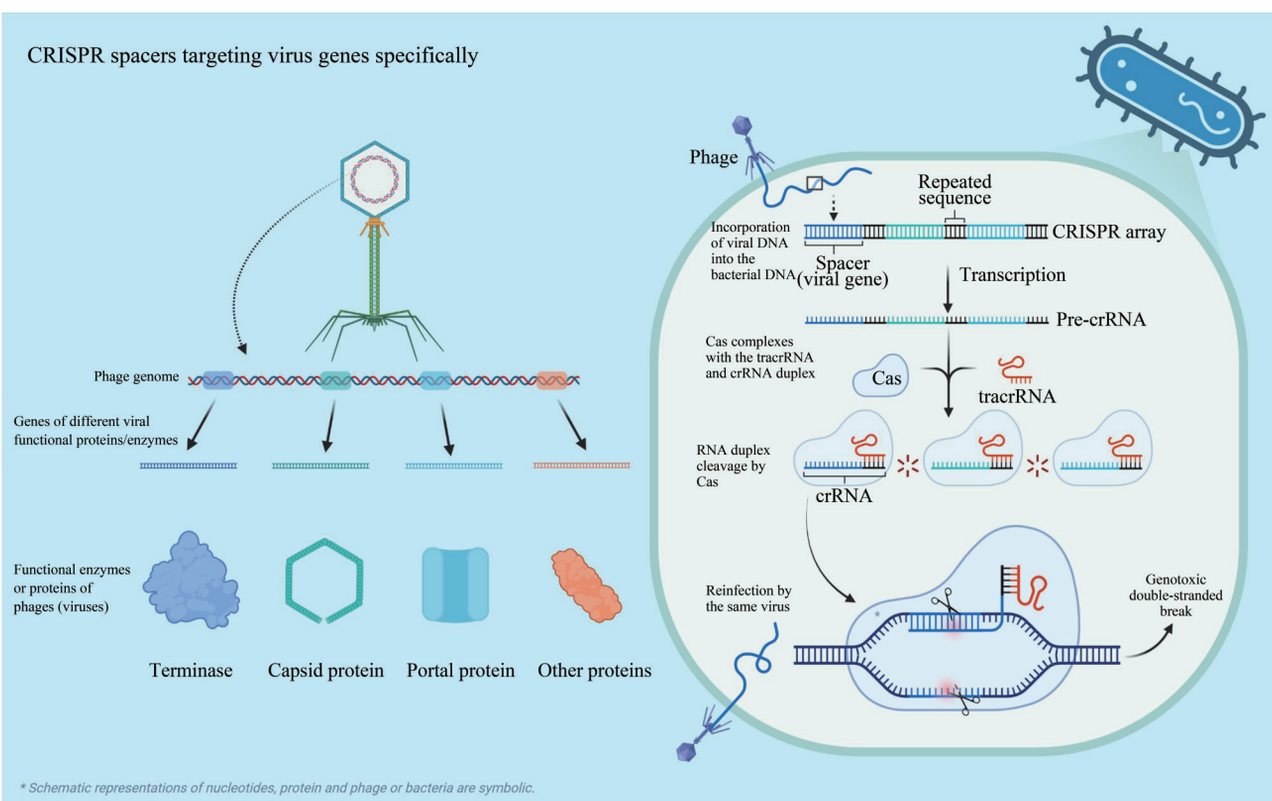
LI Cong[1)], TIAN Jin-Zi[2)], WANG Yong-Jie[1,2,3)**]

([1)]*College of Food Science and Technology, Shanghai Ocean University, Shanghai* 201306, China;

[2)]*Laboratory of Quality and Safety Risk Assessment for Aquatic Products on Storage and Preservation (Shanghai), Ministry of Agriculture and Rural Affairs, Shanghai* 201306, China;

[3)]*Laboratory for Marine Biology and Biotechnology, Qingdao National Laboratory for Marine Science and Technology, Qingdao* 266200, China)

Graphical abstract



Abstract　Objective　Since the first discovery that CRISPR-Cas system provides adaptive immunity of prokaryotic hosts to virus and other mobile genetic elements (MGEs). Numerous studies yielded vital insights into the immune mechanisms, and CRISPR-Cas system has been wildly utilized in gene editing and related research efforts. In the three major immune stages—adaptation, expression and maturation, interference spacer sequences play important roles separately. Although PAM (protospacer adjacent motif) determines the identification of

* This work was supported by grants from The National Natural Science Foundation of China（41376135，31570112）.

** Corresponding author.

Tel：86-21-61900505，E-mail：yjwang@shou.edu.cn

Received：January 18，2022　Accepted：April 26，2022

targeted genes by CRISPR-Cas system, what drives the selection of specific genes and reservation on the CRISPR array remains uncertain. To explore the targeting characteristics of CRISPR-Cas systems, the virome of Yangshan harbor surface water and the CRISPR-Cas spacers available in public datasets were subjected to analysis. **Methods** Based on BLAST searching, viral sequence identification, gene function prediction, and gene conserved domain annotation, the final analysis results were obtained. **Results** As a result, 25 391 double-stranded DNA viral sequences were identified in the virome; 265 open reading frames (ORFs) were predicted for 238 sequences, and 134 CRISPR-Cas spacer sequences yielded 315 viral hits. 128 viral ORFs and 135 hits were functionally annotated, and the top 5 hits including terminase, capsid protein, portal protein, peptidase, and DNA methyltransferase. The matching of spacer (host)-protospacer (virus) often occurred in conserved domains or key structural domains of viral functional genes. Meanwhile, the number of CRISPR-Cas spacer sequence matches for Type_I systems under the Class 1 was much higher than that for other types of systems, accounting for 89.0% of the total. The results show that the CRISPR system will specifically identify and act on key functional genes of the virus. **Conclusion** The results of this study reveal the targeting specificity of the CRISPR-Cas system, showing new insights and providing new evidence for a better understanding of the mechanisms of virus-host immune interactions.

Key words CRISPR, spacer sequence, functional gene, DNA virus
DOI：10.16476/j.pibb.2022.0025