A genomic catalogue of Earth's microbiomes

Nature Biotechnology [IF:36.558] 2020-11-09 Resource

DOI: https://doi.org/10.1038/s41587-020-0718-6

PDF：https://www.nature.com/articles/s41587-020-0718-6.pdf

第一作者：Stephen Nayfach

通讯作者：Emiley A. Eloe-Fadrosh

其它作者：Simon Roux, Rekha Seshadri, Daniel Udwary, Neha Varghese, Frederik Schulz, Dongying Wu, David Paez-Espino, I-Min Chen, Marcel Huntemann, Krishna Palaniappan, Joshua Ladau, Supratim Mukherjee, T. B. K. Reddy, Torben Nielsen, Edward Kirton, José P. Faria, Janaka N. Edirisinghe, Christopher S. Henry, Sean P. Jungbluth, Dylan Chivian, Paramvir Dehal, Elisha M. Wood-Charlson, Adam P. Arkin, Susannah G. Tringe, Axel Visel, IMG/M Data Consortium, Tanja Woyke, Nigel J. Mouncey, Natalia N. Ivanova, Nikos C. Kyrpides

作者单位：DOE联合基因组研究所，美国加利福尼亚州伯克利(DOE Joint Genome Institute, Berkeley, CA, USA)

文章翻译：NBT：5万个基因组和1.2万个新种的地球微生物基因组集 (qq.com)

摘要

构建了一个地球微生物基因组(GEM)目录，对不同生境的10,450个全球分布的宏基因组进行组装和分箱，recover 超过52,515 MAG，大大扩展了细菌和古细菌的已知系统发育多样性。本文还用此数据集用于**了解次级代谢物的生物合成潜力和解决成千上万的新宿主与未经培养的病毒的联系** 。

对鉴定phage-host的总结

- **GEM数据地址**：Index of /GEM (nersc.gov)
- **GEM把MAG和virus建立联系的具体方法**

    1. 方法一：**使用CRISPR-spacer匹配（≤1 SNP）和基因组序列匹配（> 90% identity over >500 bp）**两种方法结合。先使用两种方法结合，查看一致性，用到52,515个GEM和IMG/VR中的760,453病毒数据上，表示了良好的一致性，病毒预测宿主在family水平有95%的一致性

（Supplementary Note[1]）。最终预测了81,449个IMG/VR病毒和23,082个GEM之间的关联（图5a和附表7），但这些增加的病毒-宿主关联仍然只覆盖了IMG/VR 760,453个病毒基因组的10.7%和GEM中MAG的44.0%。

1) **CRISPR-spacer匹配（≤1 SNP）**：先把菌中的CRISPR array找出来，然后用blastn比对到病毒基因组，寻找760,453个IMG / VR病毒基因组的**Protospacers**（噬菌体入侵细菌的短片段 DNA）

   - **鉴定CRISPR-spacer**：使用**CRT**和**PILER-CR**的组合在MAG中长于10kb的contig上鉴定CRISPR array。为了最大程度地减少虚假的预测，我们丢弃了少于三个间隔子的阵列，不保守重复的序列（与一致重复的平均同一性<97%）或MAGs中包含少于四个CRISPR相关蛋白的array。最终在13,540个MAG中的23,851个CRISPR array中鉴定出长度超过25bp的567,316个CRISPR-spacer。
   - **鉴定Protospacers**：通过将CRISPR-spacer用blastn比对到760,453个IMG / VR病毒基因组，鉴定出近乎完美的匹配（最多一个不匹配≤1 SNP，覆盖至少95%的间隔子长度），以关联病毒-宿主关系。

2) **基因组序列匹配（> 90% identity over >500 bp）**：使用blastn将MAG contig与IMG / VR基因组比对以鉴定prophage(整合的噬菌体序列)。

   - **符合prophage的判断标准**：MAG contig在contig长度大于500 bp的范围上以> 90%的identity比对上病毒基因组，且其contig长度是IMG / VR基因组长度的> 1.5倍，则确定整合在MAG中。
   - **不符合prophage的判断标准**：比对上病毒基因组的contig但该病毒基因组长度1.5倍的contig被认为是“完整病毒序列”，由于缺乏宿主信息以及可能存在不正确的分箱而被丢弃

2. 方法二：为了最大化在MAG中识别出的噬菌体数量，清除病毒污染后，**使用VirSorter de novo预测在GEMs中的prophage**。这种方法使得比起方法一，多提供了10,410个病毒与7,805个GEM建立关系。

   - 具体方法：我们使用VirSorter（v1.0.3）进行了从头预测，保留了第4类和第5类的所有预测。之后进行筛选，筛选标准：

     - **排除可能decayed噬菌体**（**decayed phage是现在不活跃的整合病毒的基因组，随着突变会从宿主基因组中逐渐被删除？那为什么要排除呢**）
     - **排除了超过30%基因对Pfam有最佳hit的**噬菌体（阈值：hmmsearch得分≥50和E≤0.001）（**为啥**）

- **排除与方法一预测的IMG / VR中81,449个病毒基因组相似的contig**：相似标准>90% DNA identity over >500 bp

- **附录提到的两个指标**
  - 一致性：计算方法是对两种方法得到的virus-host关系对，先提取每个virus host在不同tax rank的most common taxonomy，再据此比较两种方法的一致性
    - **For viruses with hosts predicted by both methods, we observed agreement at the following ranks: phylum (91.9%), class (91.8%), order (88.5%), family (82.4%), genus (73.9%), species (53.2%)**.
    - **When only considering MAG contigs >1.5x the length of matched viruses, the agreement with CRISPR spacers was increased: phylum (96.9%), class (96.9%), order (94.9%), family (88.6%), genus (79.3%), species (56.3%)**.
    - when **only considering "confident" predictions by each method (>10 virus-host connections with >90% agreement within-method): phylum (99.2%), class (99.2%), order (99.0%), family (98.7%), genus (98.0%), species (98.6%)**.
  - purity指标：即一个phage预测出的不同宿主在不同taxonomic rank一致性如何
    - **For hosts predicted based on CRISPR-targeting**, we observed the following purity values: phylum (99.1%), class (99.1%), order (98.5%), family (95.7%), genus (91.1%), **species (84.9%)**.
    - **For hosts predicted based on genome sequence matches**, we observed the following purity values: phylum (99.6%), class (99.6%), order (99.0%), family (95.2%), genus (89.0%),**species (75.8%)**.
- 问题
  1. CRISPR spacer和Prophage两种方法的一致性是否过高，明明是两种方法，是否可以相互验证
  2. purity是否过高

关于鉴定phage-host相关的原文

结果：GEMs揭示了新病毒-宿主的联系

GEMs reveal thousands of new virus–host connections

除了组装微生物基因组外，最近的研究还强调了如何从宏基因组挖掘新病毒基因组。但是，大多数未经培养的病毒不能与微生物宿主相关，这对于了解它们在自然界中的作用和影响至关重要。我们认为，GEM集中的MAG可用于改善病毒基因组的宿主预测。为了解决这个问题，我们**使用CRISPR-spacer匹配（≤1 SNP）和基因组序列匹配（> 90% identity over >500 bp）**的组合确定了52,515个GEM和IMG/VR中的760,453病毒的关系，表现良好的一致性（Supplementary Note）。IMG/VR病毒匹配到一致的宿主分类单元（每个病毒产生的病毒-宿主关系对中95%与同一family的细菌建立链接/95% of linkages per virus to the same host family），并且超过 96%已关联的病毒和GEM来源于相似环境基于GOLD environmental ontology数据库的top level 。

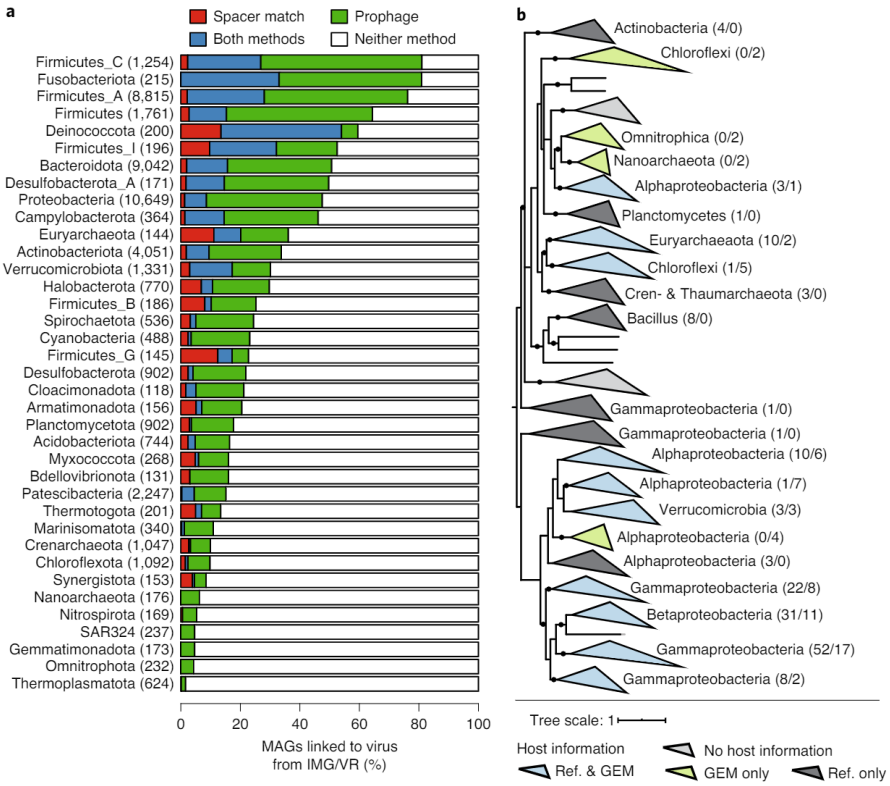

图5 MAG解决了宿主与病毒的关联问题

> a，来源于GEM catalog的细菌和古细菌门与病毒相关。条形图显示与包含100个或更多MAG的每个门的病毒链接的MAG的百分比。门的名称来自GTDB，右边的数字代表每个门的MAG数量 。条形颜色表示将病毒关联到宿主的比例。白色表示与任何病毒无关的MAG的百分比。
>
> b，DJR病毒的系统发育以及相关的宿主信息。对于与同一宿主组关联的三个或更多DJR序列的每个进化枝，该进化枝旁边显示宿主信息，以及将该DJR进化枝与该宿主组连接的序列数，首先是参考序列，然后是GEM集。参考序列获自Kauffman等人。进化枝会根据宿主信息的来源进行着色，而从GEM集中唯一识别的新宿主组将以粗体突出显示。所有支持> 50%的节点将显示为多分叉，而支持率> 80%的节点将以黑点突出显示。

使用这两种方法的组合，我们**预测**了81,449个IMG/VR病毒和23,082个GEM之间的关联（**图5a**和附表17），带有预测宿主的IMG/VR病毒总数增加了> 2.5倍（从36,976至92,872）。但是，这些增加的病毒-宿主关联仍然只覆盖了IMG/VR 760,453个病毒基因组的10.7%和GEM集中MAG的44.0%。对于某些门类（例如，Thermoplasmatota）来说，其中病毒仅与624个组装的MAG的1.6%相关。

为了解决这个限制，我们在仔细清除病毒污染后， **使用VirSorter从头预测在GEMs中的整合噬菌体** 。这种方法还提供了10,410个与7,805个GEM相关的病毒。这些新的源自MAG的病毒-宿主关联包括几类尚未充分研究的进化枝，如double jelly roll (DJR) 谱系，这是一种通常被忽视的非尾双链DNA病毒。DJR病毒多样性的最新研究表明，该组成员感染了生活的三个域的宿主，但研究也突出了没有已知宿主的亚种。在这里，我们在GEM集中确定了73个DJR序列，这些序列为另外四个DJR进化枝提供了宿主信息（ **图5b** ）。此外，这些进化枝中的两个通过GEM与尚未鉴定为假定DJR宿主的未培养细菌和古细菌群体（即Omnitrophica和Nanoarchaeota）相关联。除DJR组外，我们还鉴定了两个单链DNA病毒家族的假定宿主，包括四个Microviridae和28个Inoviridae（附图12和附表18）。这些不同的例子一起说明了MAG如何解决新型的病毒与宿主的联系。

Methods：将MAG与IMG/VR和VirSorter识别出的病毒建立联系

Connecting MAGs to viruses identified from IMG/VR and VirSorter

利用CRISPR-spacer匹配和病毒与MAGs之间的序列相似性，将MAGs用于预测IMG / VR的81,449个病毒基因组的宿主。使用**CRT**和**PILER-CR**的组合在MAG中长于10kb的重叠群上鉴定CRISPR array。为了最大程度地减少虚假的预测，我们丢弃了少于三个间隔子的阵列，不保守重复的序列（与一致重复的平均同一性<97%）或MAGs中包含少于四个CRISPR相关蛋白的系列。最终在13,540个MAG中的23,851个阵列中鉴定出长度超过25bp的567,316个CRISPR间隔子。通过将间隔子与blastn对齐到760,453个IMG / VR基因组，并鉴定出近乎完美的匹配（最多一个不匹配，覆盖至少95%的间隔子长度）来鉴定Protospacers。另外，使用blastn将MAG重叠群与IMG / VR基因组比对以鉴定整合的噬菌体序列。如果IMG / VR基因组在MAG以> 90%的identity比对上contig，contig长度大于500 bp且其contig长度是IMG / VR基因组长度的> 1.5倍，则确定它整合在MAG中。小于IMG / VR基因组长度1.5倍的contig被认为是“完整病毒序列”，由于缺乏宿主信息以及可能存在不正确的分箱而被丢弃（即基于病毒基因组特征进行分箱而不是宿主）

为了最大化在MAG中识别出的噬菌体数量，我们使用VirSorter（v1.0.3）进行了从头预测，保留了第4类和第5类的所有预测。为了排除可能decayed噬菌体，即现在不活跃的整合的病毒基因组从宿主基因组中逐渐删被除，排除了30%或更多的基因对Pfam有最佳hit的预测（阈值：hmmsearch得分≥50和E≤0.001）。对大于500bp的contig 对IMG / VR先前检测到的81,449个病毒基因组任何一个contig显示有90%DNA同源性contig进行过滤

附录：Benchmarking host-prediction method

- 总结
  - **使用CRISPR-spacer匹配（≤1 SNP）和基因组序列匹配（> 90% identity over >500 bp）**两种方法结合，评估一致性来确定可信的预测（10 virus-host connections with >90% agreement within-method），一致性水平非常高
  - **定义purity指标**，即一个phage预测出的不同宿主在不同taxonomic rank一致性如何
- 问题：CRISPR spacer和prophage鉴定的方式一致性是否真的这么高。

We used MAGs from the GEM catalogue to predict hosts for IMG/VR viruses using a combination of CRISPR targeting and prophages (see Methods[2] in main text for details). Each MAG was taxonomically annotated by the GTDB, enabling us to link each virus to a host at a variety of taxonomic ranks (phylum, class, order, family, genus, and species). The predicted host taxonomy for viruses was then validated using two separate approaches.

We first compared the host taxonomy of viruses determined by genome sequences matches (i.e. prophages) versus predictions based on spacer matches. For each method, and at each taxonomic rank, the predicted host was determined based on the most commonly observed taxon. These predicted hosts were then compared between methods. **For viruses with hosts predicted by both methods, we observed agreement at the following ranks: phylum (91.9%), class (91.8%), order (88.5%), family (82.4%), genus (73.9%), species (53.2%)**. We found that viruses matching MAG contigs 'end-to-end' often had discordant predictions with CRISPR spacers. We reasoned that these contigs of entirely viral origin may be incorrectly assigned to a MAG during the binning process, likely because of differences in GC content and codon usage between viral and host genome, and/or because of differences in genome copy number (i.e. coverage) for viruses actively replicating. **When only considering MAG contigs >1.5x the length of matched viruses, the agreement with CRISPR spacers was increased: phylum (96.9%), class (96.9%), order (94.9%), family (88.6%), genus (79.3%), species (56.3%)**. Finally, we found that agreement between methods further increased when **only considering "confident" predictions by each method (>10 virus-host connections with >90% agreement within-method): phylum (99.2%), class (99.2%), order (99.0%), family (98.7%), genus (98.0%), species (98.6%)**.

Second, we evaluated the **"purity"** of predicted hosts for each virus at different taxonomic ranks. This was performed for each prediction method. Purity was defined at each taxonomic rank by (1) identifying the most common host taxon for a virus, and (2) determining the percent of predictions matching this taxon. For example, at the phylum-rank, a virus matching 9 Proteobacteria and 1 Bacteroidetes would have a purity of 90%. **For hosts predicted based on CRISPR-targeting**, we observed the following purity values: phylum (99.1%), class (99.1%), order (98.5%), family (95.7%), genus (91.1%), **species (84.9%)**. **For hosts predicted based on genome sequence matches**, we observed the following purity values: phylum (99.6%), class (99.6%), order (99.0%), family (95.2%), genus (89.0%),**species (75.8%)**.

---

1. ## 附录：Benchmarking host-prediction method

   - 总结

     - **使用CRISPR-spacer匹配（≤1 SNP）和基因组序列匹配（> 90% identity over >500 bp）**两种方法结合，评估一致性来确定可信的预测（10 virus-host connections with >90% agreement within-method），一致性水平非常高
     - **定义purity指标**，即一个phage预测出的不同宿主在不同taxonomic rank一致性如何

   - 问题：CRISPR spacer和prophage鉴定的方式一致性是否真的这么高。

   We used MAGs from the GEM catalogue to predict hosts for IMG/VR viruses using a combination of CRISPR targeting and prophages (see Methods：将MAG与IMG/VR和VirSorter识别出的病毒建立联系[2] in main text for details). Each MAG was taxonomically annotated by the GTDB, enabling us to link each virus to a host at a variety of taxonomic ranks (phylum, class, order, family, genus, and species). The predicted host taxonomy for viruses was then validated using two separate approaches.

   We first compared the host taxonomy of viruses determined by genome sequences matches (i.e. prophages) versus predictions based on spacer matches. For each method, and at each taxonomic rank, the predicted host was determined based on the most commonly observed taxon. These predicted hosts were then compared between methods. **For viruses with hosts predicted by both methods, we observed**

agreement at the following ranks: phylum (91.9%), class (91.8%), order (88.5%), family (82.4%), genus (73.9%), species (53.2%). We found that viruses matching MAG contigs 'end-to-end' often had discordant predictions with CRISPR spacers. We reasoned that these contigs of entirely viral origin may be incorrectly assigned to a MAG during the binning process, likely because of differences in GC content and codon usage between viral and host genome, and/or because of differences in genome copy number (i.e. coverage) for viruses actively replicating. **When only considering MAG contigs >1.5x the length of matched viruses, the agreement with CRISPR spacers was increased: phylum (96.9%), class (96.9%), order (94.9%), family (88.6%), genus (79.3%), species (56.3%)**. Finally, we found that agreement between methods further increased when **only considering "confident" predictions by each method (>10 virus-host connections with >90% agreement within-method): phylum (99.2%), class (99.2%), order (99.0%), family (98.7%), genus (98.0%), species (98.6%)**.

Second, we evaluated the **"purity"** of predicted hosts for each virus at different taxonomic ranks. This was performed for each prediction method. Purity was defined at each taxonomic rank by (1) identifying the most common host taxon for a virus, and (2) determining the percent of predictions matching this taxon. For example, at the phylum-rank, a virus matching 9 Proteobacteria and 1 Bacteroidetes would have a purity of 90%. **For hosts predicted based on CRISPR-targeting**, we observed the following purity values: phylum (99.1%), class (99.1%), order (98.5%), family (95.7%), genus (91.1%), **species (84.9%)**. **For hosts predicted based on genome sequence matches**, we observed the following purity values: phylum (99.6%), class (99.6%), order (99.0%), family (95.2%), genus (89.0%),**species (75.8%)**.

2. Methods：将MAG与IMG/VR和VirSorter识别出的病毒建立联系

Connecting MAGs to viruses identified from IMG/VR and VirSorter

利用CRISPR-spacer匹配和病毒与MAGs之间的序列相似性，将MAGs用于预测IMG / VR的81,449个病毒基因组的宿主。使用**CRT**和**PILER-CR**的组合在MAG中长于10kb的重叠群上鉴定CRISPR array。为了最大程度地减少虚假的预测，我们丢弃了少于三个间隔子的阵列，不保守重复的序列（与一致重复的平均同一性<97%）或MAGs中包含少于四个CRISPR相关蛋白的系列。最终在13,540个MAG中的23,851个阵列中鉴定出长度超过25bp的567,316个CRISPR

间隔子。通过将间隔子与blastn对齐到760,453个IMG / VR基因组，并鉴定出近乎完美的匹配（最多一个不匹配，覆盖至少95％的间隔子长度）来鉴定Protospacers。另外，使用blastn将MAG重叠群与IMG / VR基因组比对以鉴定整合的噬菌体序列。如果IMG / VR基因组在MAG以> 90％的identity比对上contig，contig长度大于500 bp且其contig长度是IMG / VR基因组长度的> 1.5倍，则确定它整合在MAG中。小于IMG / VR基因组长度1.5倍的contig被认为是“完整病毒序列”，由于缺乏宿主信息以及可能存在不正确的分箱而被丢弃（即基于病毒基因组特征进行分箱而不是宿主）

为了最大化在MAG中识别出的噬菌体数量，我们使用VirSorter（v1.0.3）进行了从头预测，保留了第4类和第5类的所有预测。为了排除可能decayed噬菌体，即现在不活跃的整合的病毒基因组从宿主基因组中逐渐删被除，排除了30％或更多的基因对Pfam有最佳hit的预测（阈值：hmmsearch得分≥50和E≤0.001）。对大于500bp的contig 对IMG / VR先前检测到的81,449个病毒基因组任何一个contig显示有90%DNA同源性contig进行过滤