A genomic catalog of Earth’s microbiomes

地球微生物的基因组目录

**摘要：**通过鸟枪法重构细菌和古菌的基因组使人们对生态学以及与环境和宿主相关的微生物群落的进化有了深入的了解。在这里，我们将这种方法应用于从覆盖整个地球各大洲和海洋的不同栖息地收集的10,000多个基因组，其中包括来自人和动物宿主，工程环境以及自然和农业土壤的基因组，以获取现有的微生物，代谢和功能潜力。该综合目录包括52,515个由宏基因组组装的基因组，它们代表跨越135个门的12,556个新的候选种水平的OTU。该目录将已知细菌和古细菌的系统发育多样性扩展了44％，可广泛用于简化比较分析，交互式探索，代谢建模和批量下载。我们展示了此集合的实用性，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。该资源强调了以基因组为中心的方法在揭示影响生态系统过程的未培养微生物的基因组特性方面的价值。

**背景介绍：**

迄今为止，许多各种各样的微生物尚未被培养，并且仅通过不依赖培养的分子方法仍可得到。基因组解析（genome-resolved）的宏基因组学是一种能够从微生物种群中重建复合基因组的方法，它首先应用于低复杂性的酸性矿山排水微生物群落。随着计算方法和测序技术的进步，这种方法现在已经在更大范围内得到了应用。以及包括许多其他环境，例如全球海洋，牛瘤胃，人类微生物组，深层地下和含水层。这些研究引发了对未培养细菌和古菌的进化关系和代谢特性的深刻见解。

除了扩展和填充微生物的生命之树外，全面的未培养细菌和古细菌基因组目录还将为大规模的比较基因组学，挖掘感兴趣的基因和功能（例如CRISPR–Cas9变体），构建基因组规模的代谢模型为实现系统生物学方法提供机会。此外，最近将未培养细菌和古细菌的基因组重建对真核生物的进化轨迹和祖先微生物的性状产生了独特见解。

在这里，我们应用了大规模的基因组分解（genome-resolved）宏基因组学，重新获得了52,515个中等质量和高质量的宏基因组组装基因组（metagenome-assembled genomes，MAG），这些来自地球微生物组基因组（Genomes from Earth’s Microbiomes，GEM）目录。GEM目录是根据从各种微生物栖息地和地理位置取10,450个宏基因组样品构建的（图1）。这些基因组代表了12,556个新候选种水平的操作生物分类单位（operational taxonomic units，OTU），这意味着可得到更广泛的未经培养的细菌和古细菌的系统发育和功能多样性的资源。为了证明这种资源的价值，我们使用GEM目录在地球各个生物群落中进行了宏基因组阅读募集（metagenomic read recruitment），确定了新的生物合成能力，进行了代谢建模并预测了宿主与病毒的联系。

**结果：**

**1. Over 52,000 metagenome-assembled genomes recovered from environmentally diverse metagenomes.**

**从环境多样的宏基因组中重新获得了超过52,000个宏基因组组装基因组。**

我们对10,450个全球分布的宏基因组进行了宏基因组学组装和分箱，这些样品来自海洋和其他水生环境（3,345），人和动物宿主相关的环境（3,536）以及土壤和其他陆地环境（1,919），获得了52,515个MAG（图1a–c和附表1和2）。除了可公开获得的宏基因组外，这些宏基因组还包括由综合微生物基因组和微生物组（Integrated Microbial Genomes and Microbiomes，IMG/M）数据联盟提供的数千个未发表的数据（方法和附表1和2）。MAG的全球目录包含了来自全球大陆和海洋的样本，其中来自北美，欧洲和太平洋的样本特别有代表性（图1d和附图1）。GEM目录可随环境元数据（数据可用性和附表1）一起大量下载，并可通过IMG/M（https://img.jgi.doe.gov）或能源部（Department of Energy，DOE）系统生物学知识库（Kbase; https://kbase.us）门户网站进行交互式浏览，以简化比较分析和代谢建模。

来自GEM目录的MAG均达到或超过MIMAG标准的中等质量水平（平均完整性= 83％；平均污染度= 1.3％），并包括9,143（17.4％）个基于接近完整的高质量rRNA，tRNA和单拷贝蛋白质编码基因（图1a，b和附表2）。高质量GEM的基因组大小在0.63至11.28 Mb之间，多数小型MAG属于预期减少的基因组谱系，例如Nanoarchaeota或Mycoplasmatales，以及类似地，大型MAG属于Myxococcota和Planctomycetota。基因组大小和GC含量在寄主相关微生物组中最低（中位数：2.61 Mb； 46.9％），而在陆地微生物组中最高（中位数：3.77 Mb； 57.1％），这与在土壤环境中的泛基因组扩展一致。MAG的大小与相同物种的分离基因组（isolate genomes）一致，表明单个基因组中的基因含量没有重大遗失（附图2）。一个例外是*Sinorhizobium* *medicae*，其中来自根瘤的MAG几乎是分离基因组的两倍（MAG是11-12Mb，分离基因组6-7Mb; *S. medicae*：99％ANI和65％AF）。尽管分箱scaffolds的四核苷酸频率组成总体上显示出良好的一致性，但检测到许多SNP，这表明是同一物种的两个菌株的复合物。此外，我们还比较了由Parks为GEM样本子集进行独立的MAG组装，这提升了复合基因组bin的可重复性（附表3和补充说明）。

分类学上定义的参考基因组通常用于从宏基因组中推断出微生物的丰度，但无法收集到人类微生物组以外的大多数reads。为了探索GEM目录中的MAG是否可以解决此问题，我们将来自3170个元基因组的高质量reads与52515个GEM和NCBI RefSeq的所有分离基因组的可用read数据进行了比对。这表明每个样本将平均为30.5％（IQR= 5.9–49.3％）和14.6％（IQR = 0.9–15.8％）的宏基因组reads分别分配给一个或多个GEM或分离的基因组（附图3和附表4）。在所有样品中，GEM引起mapped reads数增加3.6倍，这在某些环境中尤其明显，例如生物反应器或无脊椎动物宿主（附图3）。尽管有些改进，但仍然有近70％的reads未mapped到任何MAG或分离基因组。这在土壤群落中尤其明显（例如，高于95％的reads未mapped到55％的样品中的任何基因组），这些高度复杂且组装困难。与此结果一致，k-mer多样性最高的宏基因组往往具有最低的mapping rates（Spearman's r = -0.68；P value = 0）。这些群落可能包含密切相关的生物体，这对宏基因组学的组装和分箱是主要问题。低的mapping rates也可能反映了病毒，质粒和真核生物微生物的存在，但本研究使用的pipeline无法重新获取。

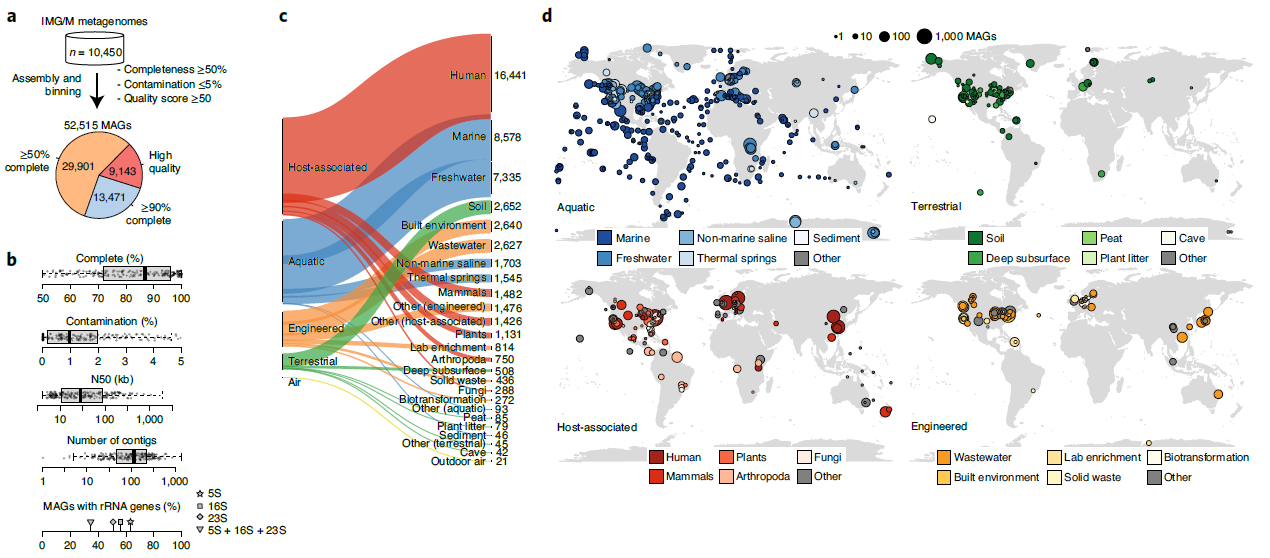


图1 宏基因组组装的环境和地理分布。a，在IMG/M从地理和环境多样的宏基因组中获得52,515个MAG。使用最新的组装pipeline（附表1）对大部分（10,450个中的6,380个；61％）的元基因组进行了重新组装。这些基因组构成了GEM目录。所有MAG≥50％完整性，受污染≤5％，且质量得分（完整性− 5×污染）≥50。b，跨MAG的质量指标分布。在每个箱线图上覆盖大约200个随机选择的数据点，显示最小值，第一四分位数，中位数，第三四分位数和最大值。有关所有MAG的质量统计信息，请参见附表2。c，基于Genomes OnLine Database（GOLD； https：//gold.jgi-psf.org）中的环境元数据，MAG分布于整个生物群落和亚生物群落中。与每个亚基生物群落相关的MAG数在图旁边显示。d，每个生物群落内MAG的地理分布。

**2. The GEM catalog expands genomic diversity across the tree of life.**

**GEM目录延申了生命之树的基因组多样性。**

为了发现新的种水平多样性，我们基于95％的全基因组ANI对GEM进行了聚类，揭示了18,028个物种水平的OTU（图2a，b，附图4和附表5）。尽管原核生物的物种概念存在争议，但该操作定义是常用的并且被认为是黄金标准。根据来自基因组分类数据库（Genome Taxonomy Database，GTDB）的分类注释，我们发现GEM涵盖137个已知门，305个已知纲和787个已知目。绝大多数非单一OTU仅包含来自单一环境或多重紧密相关环境的GEM（例如生物反应器和废水环境；附图5），这表明很少有物种具有较宽的生境范围，然而接近40％的物种在多个采样点发现（图2c）。MAG的积累曲线显示种水平的OTU没有平稳期（附图6），表明在整个生物群落中仍有待发现的其他物种，从mapped reads的低百分比也能反映出这个结论。

接下来，我们将18,028个OTU与524,046个参考基因组的广泛数据库进行了比较，其中包括之前研究的> 300,000个MAG，> 200,000个从纯培养物中分离的生物体基因组（包括所有RefSeq）和> 2,000个单扩增基因组（single-amplified genomes，SAG; 图2a）。其中包括在人类微生物组，全球海洋，含水层系统，永久冻土层解冻梯度，牛瘤胃，高盐湖沉积物和热液沉积物中进行的大量MAG研究，以及多项大型分离基因组测序研究，例如细菌和古细菌的基因组百科全书（Genomic Encyclopedia of Bacteria and Archaea，GEBA）和人类微生物组计划（Human Microbiome Project，HMP），尽管当前发表了几项研究，但并没有包括在内。所有参考基因组均遵循与我们应用于GEM数据集相同的质量标准（≥50％的完整性，≤5％的污染和≥50的质量得分）

值得注意的是，来自GEM目录的12,556个OTU（代表23,095个MAG）与95％ANI的参考基因组不同，因此代表了新的候选物种。同时，所有ANI> 95％参比基因组中有70％的被归类到GEM目录中，这表明它对现有基因组具有良好的覆盖率。在326个研究中发现了新的OTU，平均每个研究有40个。微生物暗物质（Microbial Dark Matter，MDM）第二阶段研究是GEBA-MDM项目的延伸，它贡献了最新的790种OTU，它们来自80个宏基因组中的1,124个MAG。

为了支持它们的新颖性，在12556个新的OTU中，绝大多数与参考基因组远亲或几乎没有亲缘关系（与参考相比，ANI小于90％或AF小于10％的OTU占93.7％），> 99％的物种未在 GTDB中注释到。但是，来自新OTU的MAG趋向于不太完整（新OTU平均为81.0％相对于旧OTU平均为84.6％），稍高的污染（新OTU平均为1.5％相对于旧OTU平均为1.1％），并且通常以singletons形式被发现（图2d，附表6和补充说明）。这些观察结果可能由多个因素引起，包括未培养lineages（翻译为物种，是否恰当？）的基因组减少，16S rRNA基因座的组装问题和稀有生物圈成员恢复。

我们将未收集的参考基因组聚集成了另外的27,571个OTU，从而形成了45,599个种水平OTU的组合数据集（图2a，b）。这表明，尽管GEM目录包含较少的基因组，但与以前发表的任何研究相比，它代表的多样性是其3.8倍（图2e）。例如，Parks等人对NCBI Sequence Read Archive中可用的所有环境元基因组进行了大规模的组装和分箱，以前所未有的努力扩展未培养lineages的基因组。根据当前研究中进行的聚类和质量控制，这10,728个MAG代表5,200个OTU，仅占GEM目录中OTU的12％（附表7）。

接下来，我们基于30个串联的标记基因构建了45599个OTU的系统发育树（图3a，附表8和方法）。对系统发育树分析表明，GEM目录是迄今为止发布的最多样化的数据集（图2f）。总体而言，GEM目录使整个细菌和古细菌系统发育树多样性提高了44％，基于累积分支长度，目前占所有已知多样性的31％。

在不同类群中，系统发育多样性的增加相对一致，但对于某些大进化分支的物种，系统发育多样性尤其高，其中包括Planctomycetota（增长了79％），Verrucomicrobiota（增长了68％）和Patescibacteria（也被称为“Candidate Phyla Radiation”，增长了60％）（图3b和附表9）。GEM目录在整个环境中产生了更多的可变增加（附表10），尽管在最近MAG研究的先前分析的人类相关样本（图3b）中几乎没有发现新的多样性。值得注意的是，这些分析还表明，分类微生物多样性的系统发育多样性中的75％仅由未培养的基因组（即MAG或SAG）代表。为了确定GEM目录是否包含较高分类学等级的新lineages，我们使用相对进化差异（relative evolutionary divergence，RED）将所有45,599个OTU聚类为单系，包括singletons，分别代表16,062属，5,165个科，1,928个目，368个纲和129个门（附表11-13，附图7和方法）。在门的水平上，我们确定了16个进化支只代表GEM（1个进化枝属于细菌，5个进化枝属于古细菌），这可能表明有新的门。但是，这些进化枝仅由29个GEM支持，它们通过GTDB-Tk工具（28/29）很大程度上分配给已知门。在较低的分类学等级上，发现了更多的新物种，包括456个新目，1,525个新科和5,463个新属。我们得出的结论是，与早期的元基因组分箱研究相比，发现了很多新lineages，大多数深度分支的lineages由当前的基因组序列代表。

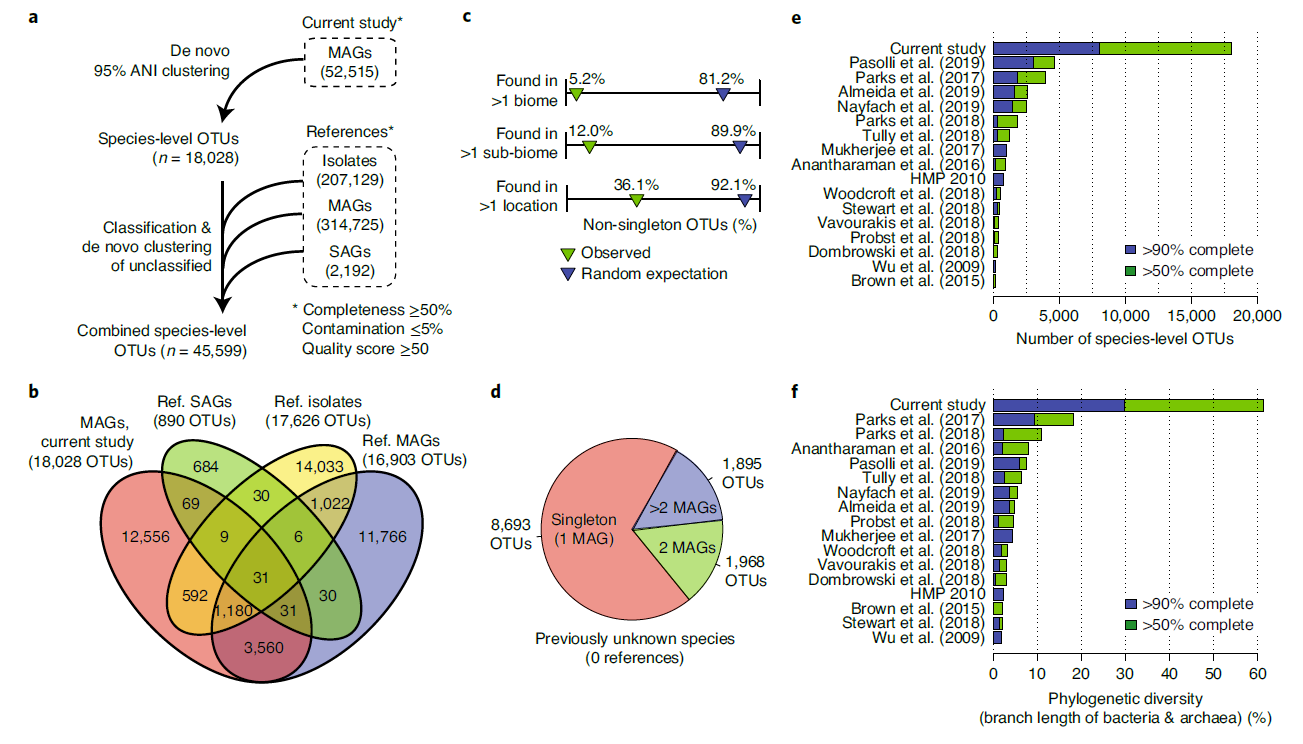
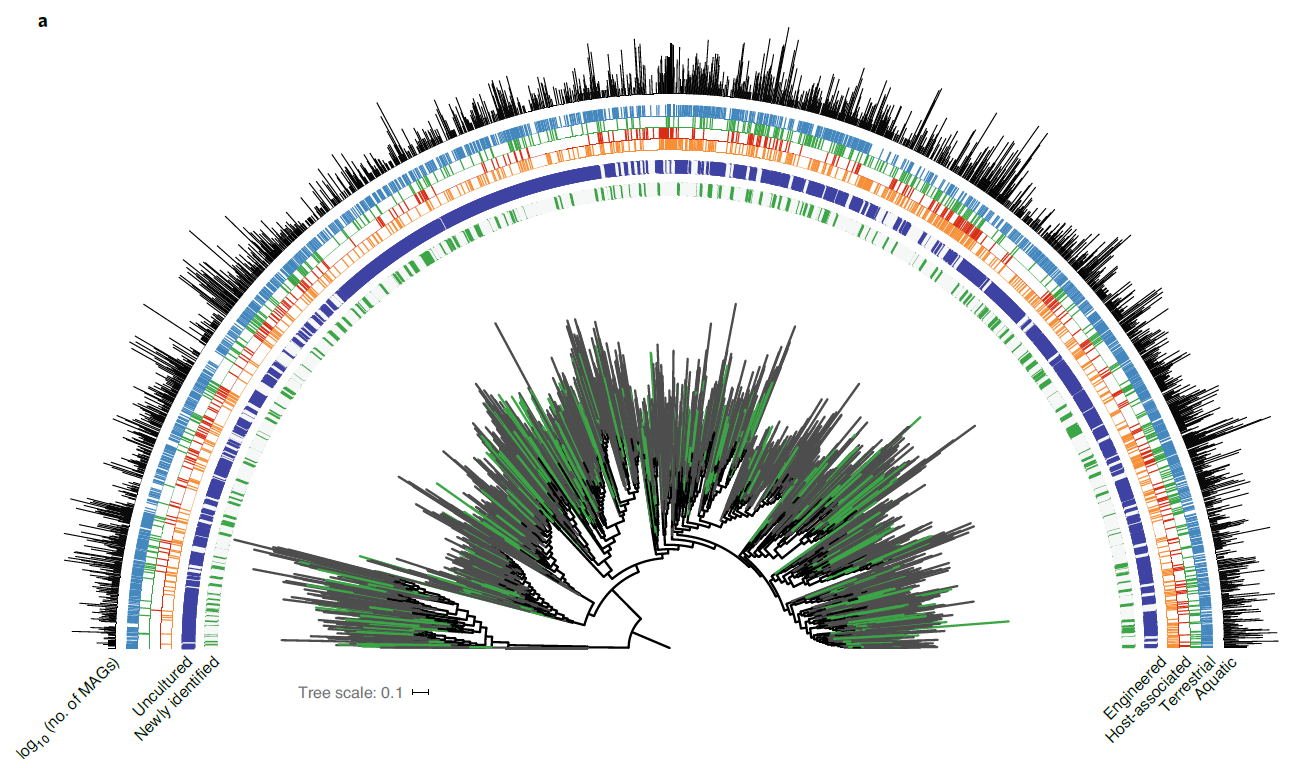
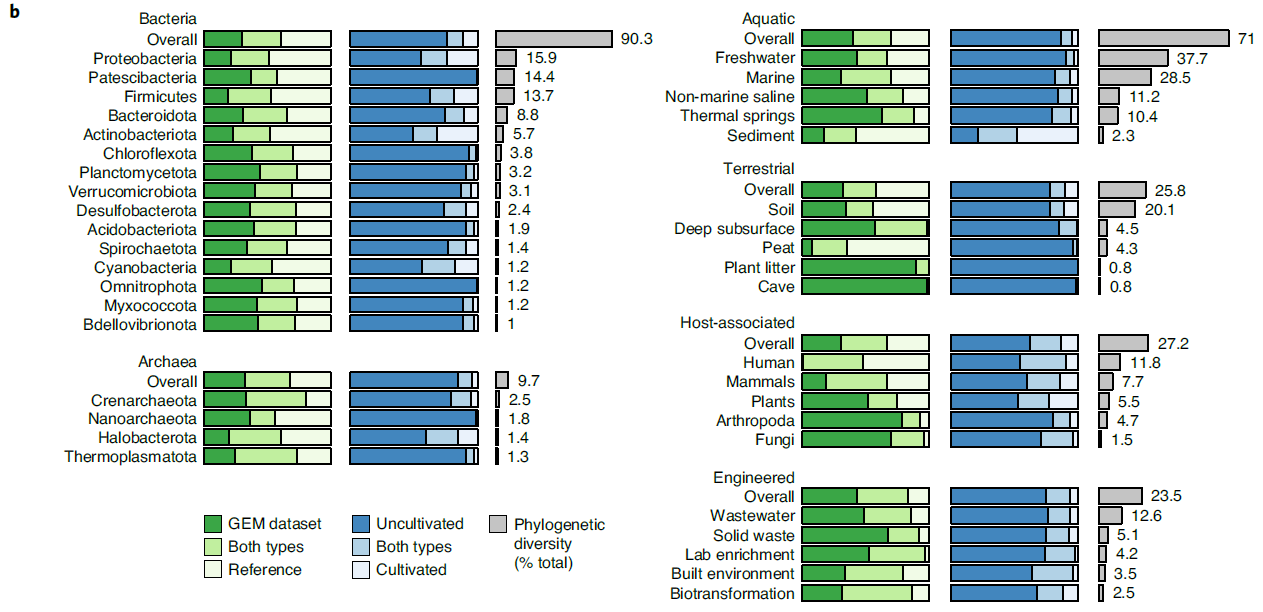


图2 具有> 500,000个参考基因组GEM目录下的种水平聚类。a，将当前研究的MAGs与IMG/M和NCBI中发现的524,046个可公开获得的参考基因组进行了比较。所有参考基因组均符合与GEM目录相同的最低质量标准。基于95％ANI和30％AF，所有MAG和参考基因组被聚集成45,599个种水平OTU。b，基因组之间的OTU重叠。来自当前研究的MAGs首次揭示了12,556个物种基因组。c，GEM目录中具有超过1个基因组的OTU绝大多数都局限于单个生物群落和亚生物群落，尽管在多个地理位置发现了超过三分之一 d，在新发现的12,556种物种中，很大一部分仅由一个基因组代表。e，f，基于种水平多样性选择的当前数据集与16个最大的先前发表的基因组研究的比较。研究标识符（Study identifiers）来自NCBI BioProject或GOLD。Wu，HMP（2010）和Mukherjee等人的研究包含了发表后产生的其他基因组。使用与GEM数据集相同的质量标准过滤了其他研究的所有MAG（图1a和方法）。与以前发表的任何研究相比，当前研究的基因组物种代表多样性高出三倍。

 图3 | GEM目录填补了生命之树中的空白。 a，基于30个普遍分布的单拷贝基因的串联比对，为45,599个OTU中的43,979个构建了系统发育树。完全比对包含4,689个氨基酸位点，每个OTU包含至少30％位点的数据。基于系统发生距离，将物种水平的OTU进一步聚类为1,928个近似等级的进化枝。绿色分支表示仅由GEM目录代表的新谱系。内部的条形图指示是新识别的目（绿色；代表GEM）还是先前已知的订单（浅灰色；代表参考基因组）。下一个带状图表示目是未培养的（蓝色；代表MAGs / SAG）还是已培养的（灰色；代表至少一个分离基因组）。接下来的四个带状图显示了订单的环境分布；最后一个图表示从每个目中回收的GEM目录中的MAG数量。尽管大多数新谱系散布在现有谱系之间，但GEM目录的复合基因组广泛分布在生命树上，包括许多新的有序进化枝。树的广阔区域仅由未培养的基因组代表。 b，针对由GEM目录/参考基因组（绿色）或培养/未培养的基因组（蓝色）代表的子树计算了系统发育多样性。灰色条表示由每个分类组（左）或环境（右）代表的总系统发育多样性的百分比。 GEM目录不断扩展细菌和古细菌中以及不同环境下不同门的系统发育多样性。人类微生物组是一个例外，GEM目录在其中几乎没有贡献新的多样性。将GEM目录与其他未培养的基因组结合起来，很明显，未培养的基因组支配着大多数门和环境内的多样性，特别是对于Patecibacteria（Candidate Phyla Radiation）和Nanoarchaeota等群体。

**3．Encoded functional potential in the GEMs.**

**GEM的编码功能的潜力。**

为了提供系统级的代谢潜力快照（snapshot），我们为KBase中每个环境（n = 3,255）的40个以上代表的非冗余，高质量GEM建立了基因组规模/尺度（genome-scale）的代谢模型（附图8和9，附表 14和补充说明）。除了已知的代谢途径外，我们假设GEM目录中的MAG包含新功能的库。为了解决这个问题，我们编制了一个5794,145个蛋白质簇（PC）的目录，代表了111,428,992个全长基因，其中51.7％的PC包含至少两个序列。与TIGRFAM或KEGG Orthology数据库相比，绝大多数PC均未进行功能注释，并且大多数PC甚至都缺少单个Pfam域（TIGRFAM，KEGG和Pfam未注释到的分别为95.2％，88.9％和74.5％）。相比之下，通过IMG/M获得76,000个参考细菌和古细菌基因组的2.7亿个基因的目录，这些基因在TIGRFAM，KEGG和Pfam未注释到的分别约为70％，50％和20％。大约70%的PC在三个平台中，功能都没注释到，并且47%与UniRef（一个数据量大且定期更新的蛋白质资源数据库https://www.uniprot.org）没有显著相似性。虽然最大的PC之前就知道，但大而复杂的PC却没有任何注释，包括至少具有1,000个成员的356个簇组成的蛋白簇和至少具有100个成员的28,869个簇组成的蛋白簇。

尽管系统地解释所有GEM的功能能力超出了本研究的范围，但在此我们提供一些说明性的插图。首先，我们发现，由于新古细菌门如Halobacterota，Hadesarchaea（包括Archaeoglobi和Syntrophoarchaeia）和Crenarchaeota在内的种群（例如Thermoprotei，Korarchaeia 和 Bathyarchaeia）的存在，GEM概述了最近对产甲烷作用范围延申的最新研究（附图10）。在较低的分类学等级上，我们确定了一种新属Coxiella的GEM，其中包括与大量健康和经济负担相关的B级生物危险菌*Coxiella burnetii*，这提供了一个获得新的见解的机会，可以深入研究宿主-病原体相互作用。在GEM中发现了几种毒力因子，包括用于将效应蛋白传递到宿主细胞质中的Dot/Icm IV型分泌系统（附图7）。但缺乏典型的*C. burnetii* T4SS效应蛋白。因此，GEM可以预测最高和最低分类学等级的潜在功能。

**4．Broad and diverse secondary-metabolite biosynthetic potential.**

**广泛和多样的次生代谢物生物合成的潜力**

大多数次级代谢产物已从少部分可培养细菌中分离出来，这些细菌群包括*Streptomycetes*, *Pseudomonas*, *Bacillus* 和 *Streptococcus*。最近，宏基因组数据的挖掘已从土壤中扩展到Acidobacteria, Verrucomicobia, Gemmatimonadetes和candidate phylum Rokubacteria。GEM目录提供了一个独特的机会去探索，在分类学和生物地理学上多样化的基因组内，编码次级代谢生物合成基因簇（biosynthetic gene clusters，BGCs）的特点。我们使用AntiSMASH（v5.1）从52,515个GEM中确定了104,211个假定的BGC区域（附表15）。为了进行比较，这意味着IMG/ABC中的BGC（Atlas of BGCs）增加了31％，是手动创建的MIBiG数据集的54倍。大约66％的GEM BGC与一个或多个重叠群边界相交，表明大多数可能是不完整的（附图12），这与以前基于碎片恢复（fragmented recovery）的观察结果一致。我们将每个BGC合成的次级代谢物对应于GEM目录并进行分类（图4a）。从104种门中鉴定出总共44,835个基因簇或包含非核糖体肽合成酶（/或聚酮化合物合酶（PKS）的簇片段），从79种门中鉴定出23738萜烯簇，从76种门中鉴定出了12360种核糖体加工肽（RiPP）簇。尽管碎片可能以不可控的方式使簇数量产生偏差，但据我们观察这种情况可能是正常的。例如，Firmicutes的RiPP数量异常高（其BGC的一半以上是RiPP簇），而Thermoplasmatota和Verrucomicrobiota的萜烯簇数量相对较高（分别为BGC的68％和50％）。BGC的环境趋势分析不太清楚，没有环境来源组显示相对BGC家族含量有明显的偏差（图4a）。如果准确的话，这意味着特定的化学物质不受环境的限制或放大，并且大多数类别的次级代谢产物几乎可以在任何地方找到。

为了评估BGC的新颖性，我们将每个BGC序列与NCBI核苷酸序列对比。用75%阈值去比对长度超80％的序列，我们确定87,187（83％）个是编码新化学过程的新BGC（附表16）。尽管许多模块簇是零散的，但我们鉴定了3,000个BGC区域，且长度> 50 kb，超过17,000个BGC区域> 30 kb。

同时，GEM目录拥有预测新BGC的丰富资源潜力，并提供了充分的机会探索已知进化枝之外的生物合成潜力。如上所述，*Myxococcus*显示较好的生物合成潜力，在232个MAG上有1,751个区域，并且由antiSMASH定义的BGC家族种类繁多。单个最大的BGC区域位于Acidobacteria和UBA5704属的土壤细菌中，编码数量较多的62个PKS或NRPS模块，同时具有三个清晰的共线性模块链（图4b）。尽管一些Acidobacteria已知包含PKS和NRPS簇，但该MAG包含额外的66个BGC区域，说明Acidobacteria的生物合成潜力水平之前可能被低估。

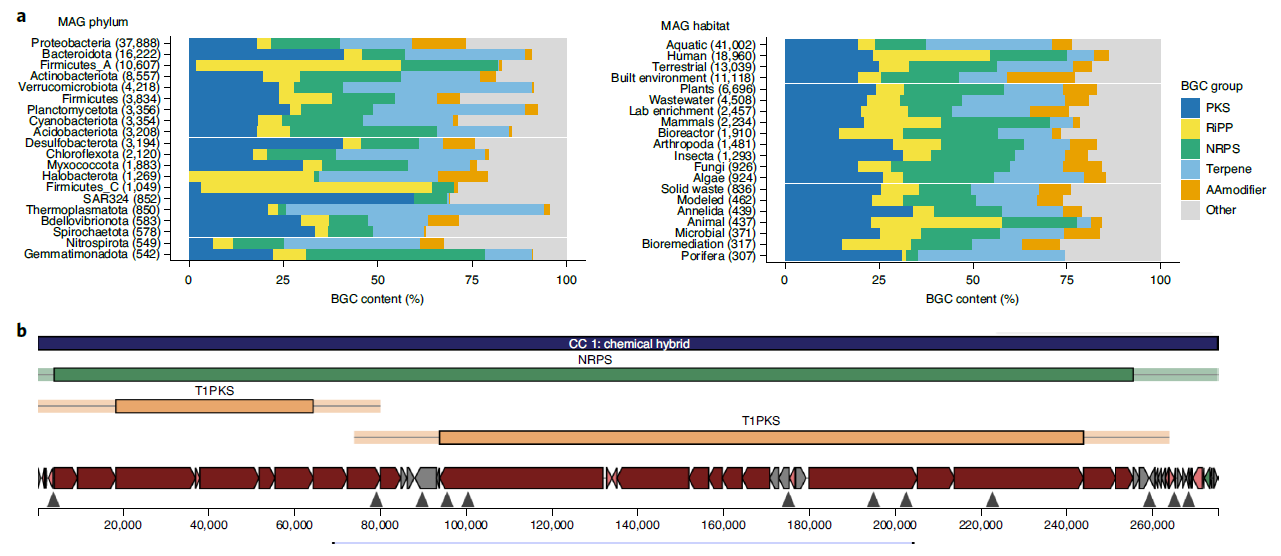


图4 | 从GEMs数据集中获得生物合成基因簇。 a，在优势门（左）和栖息地（右）中BGC类型的相对频率。 BGC类型在整个门系中变化很大，但在生境之间相对稳定。 AAmodifier，氨基酸修饰系统。 b，最大的BGC区域，存在于Acidobacteria和UBA5704属的土壤细菌中。 BGC使用三个共线性模块链对62个PKS或NRPS模块进行编码。

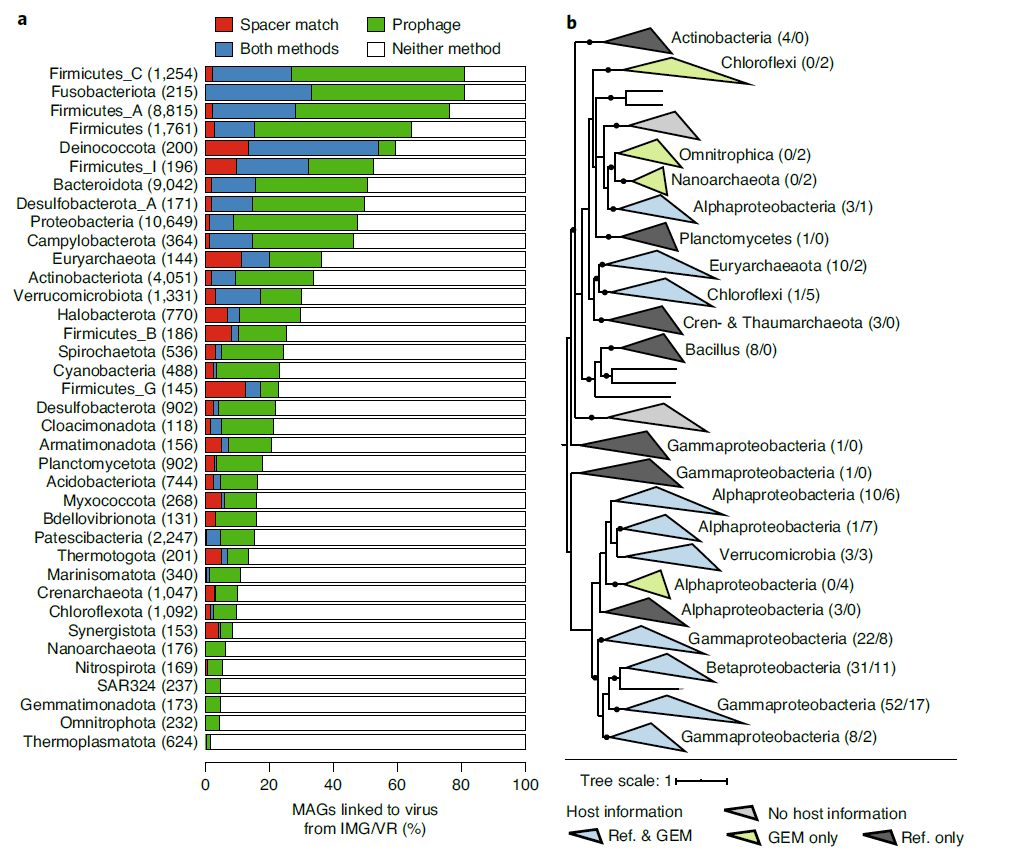
**5. GEMs reveal thousands of new virus–host connections.**

**GEMs揭示了新病毒-宿主的联系。**

除了装配微生物基因组外，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目录中的MAG可用于改善病毒基因组的宿主预测。为了解决这个问题，我们将IMG/VR56中的52,515个GEM和760,453病毒与CRISPR-spacer匹配（≤1 SNP）和基因组序列匹配（> 500 bp时> 90％的同一性）比对，都具有较好的一致性（补充说明）。IMG/VR病毒已匹配到一致的宿主分类单元（每种病毒与同一宿主达到95％的匹配），并且超过96％病毒和GEM的已关联到与GOLD environmental ontology数据库最相似环境中。

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目录中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 de novo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) lineage，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，但他们也突出了没有已知宿主的亚种。在这里，我们在GEM目录中确定了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）。这些不同的例子加在一起说明了MAG如何解决新型的病毒与宿主的联系。



**图5 | MAG解决了主机与病毒的关联问题。 a，GEM目录中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。根茎的名称来自GTDB，右边的数字代表每个门的MAG。条形颜色表示将病毒关联到宿主的方法。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目录。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目录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支，而支持> 80％的节点将以黑点突出显示。**

**讨论**

拥有52,515个中高品质MAG的资源代表了目前最大的成就，为了获得整个地球生物群系中细菌和古细菌基因组多样性的广度。GEM目录大大扩展了细菌和古细菌的已知系统发育多样性，增加了宏基因组测序reads，具有丰富的生物合成潜力，并改善了未培养病毒的宿主联系。尽管细菌和古细菌的系统发育多样性总体增加了44％，但我们发现几乎没有证据表明新的深度分支能产生新的门类，与最近对微生物多样性的研究一致。同样，尽管宏基因组reads的收集增加了3.6倍，但超过三分之二的宏基因组reads仍缺乏可mapping的参考基因组。因此，继续努力获得新物种和菌株的基因组将会进一步提高宏基因组学的实用性和适用性。

大规模的基因组数据为更广泛的研究提供了重要的资源。也就是说，GEM目录中的MAG就像迄今为止生成的其他MAG一样，有一些局限性供用户需要注意，包括未检测到的污染，低邻接性和不完整性。尽管这些MAG是许多新候选物种的重要占位符，但我们希望将来许多新的物种将被更高质量的MAG或最终被分离出的菌株基因组序列所取代。正如我们以新的次生代谢物BGCs和假定病毒-宿主关联所示，我们预计GEM目录将成为未来代谢和以基因组为中心的数据挖掘和实验验证的宝贵资源。