分析
https://doi.org/10.1038/s41592-022-01431-4

完整作者单位列表见文末。

宏基因组解读的关键评估：第二轮挑战

Fernando Meyer 1,2,76, Adrian Fritz 1,2,3,76, 邓智卢(Zhi-Luo Deng)1,2,4, David Koslicki 5, Till Robin Lesker3,6, Alexey Gurevich 7, Gary Robertson1,2, Mohammed Alser8, Dmitry Antipov9, Francesco Beghini 10, Denis Bertrand11, Jaqueline J. Brito12, C. Titus Brown 13, Jan Buchmann 14, Aydin Bulu□15,16, Bo Chen15,16, Rayan Chikhi 17, Philip T. L. C. Clausen 18, Alexandru Cristian19,20, Piotr Wojciech Dabrowski 21,22, Aaron E. Darling 23, Rob Egan24,25, Eleazar Eskin26, Evangelos Georganas27, Eugene Goltsman24,25, Melissa A. Gray19,28, Lars Hestbjerg Hansen29, Steven Hofmeyr15,16, 黄品琴(Pingqin Huang)30, Luiz Irber13, 贾慧杰(Huijue Jia)31,32, Tue Sparholt J□rgensen 33,34, Silas D. Kieser 35,36, Terje Klemetsen37, Axel Kola38, Mikhail Kolmogorov39, Anton Korobeynikov 9,40, Jason Kwan41, Nathan LaPierre26, Claire Lemaitre 42, 陈浩(Chenhao Li)11, Antoine Limasset 43, Fabio Malcher-Miranda 44, Serghei Mangul12, Vanessa R. Marcelino45,46, Camille Marchet43, Pierre Marijon47, Dmitry Meleshko9, Daniel R. Mende48, Alessio Milanese 49,50, Niranjan Nagarajan 51,52, Jakob Nissen53, Sergey Nurk54, Leonid Oliker15,16, Lucas Paoli49, Pierre Peterlongo42, Vitor C. Piro 44, Jacob S. Porter55, Simon Rasmussen 56, Evan R. Rees 41, Knut Reinert57, Bernhard Renard 44,58, Espen Mikal Robertsen37, Gail L. Rosen19,28,59, Hans-Joachim Ruscheweyh49, Varuni Sarwal26, Nicola Segata 10, Enrico Seiler57, 石立振(Lizhen Shi)60, 冯柱(Fengzhu Sun)61, Shinichi Sunagawa 49, S□ren Johannes S□rensen 62, Ashleigh Thomas 24,63, 童成宣(Chengxuan Tong)11, Mirko Trajkovski 35,64, Julien Tremblay65, Gherman Uritskiy66, Riccardo Vicedomini 17, 王正阳(Zhengyang Wang) 30, 王紫叶(Ziye Wang)67, Zhong Wang68,69,70, Andrew Warren55, Nils Peder Willassen37, Katherine Yelick15,16, 尤荣辉(Ronghui You)30, Georg Zeller 50, 赵正桥(Zhengqiao Zhao) 19, 朱山风(Shanfeng Zhu)71,72, 朱杰(Jie Zhu) 31,32, Ruben Garrido-Oter 73, Petra Gastmeier38, Stephane Hacquard 73, Susanne H□u□ler 6, Ariane Khaledi6, Friederike Maechler38, Fantin Mesny 73, Simona Radutoiu74, Paul Schulze-Lefert 73, Nathiana Smit6, Till Strowig6, Andreas Bremges1,3, Alexander Sczyrba 75 and Alice Carolyn McHardy 1,2,3,4 □

评估宏基因组软件对于优化宏基因组解读至关重要，也是“宏基因组解读关键评估”(CAMI)倡议的核心。CAMI II挑战赛吸引了社区在真实且复杂的数据集上评估方法，这些数据集由约1,700个新基因组和已知基因组以及600个新质粒和病毒计算生成，包含长读段和短读段序列。本文分析了76个程序版本的5,002个结果。组装方面取得了显著进步，部分归功于长读段数据。相关菌株仍然是组装和通过分箱进行基因组恢复的难题，后者的组装质量也是如此。分类分析程序明显成熟，分类单元分析程序和分箱程序在较高细菌分类等级上表现优异，但在病毒和古菌方面表现不佳。临床病原体检测结果揭示了提高可重复性的必要性。运行时间和内存使用分析确定了高效的程序，包括在其他指标上表现最佳的程序。结果指出了挑战所在，并为研究人员选择分析方法提供了指导。

Nature Methods | 第19卷 | 4月 | 429–440页 | www.nature.com/naturemethods

引言

在过去二十年中，宏基因组学的进步极大地扩展了我们对微生物世界的认识，并加剧了数据分析技术的发展。这产生了对无偏和全面评估这些方法的需求，以识别该领域的最佳实践和开放性挑战。CAMI（宏基因组解读关键评估倡议）是一个社区驱动的努力，通过提供代表微生物组研究中常见实验设置、数据生成技术和环境的综合基准测试数据集来满足这一需求。除了其开放和协作的性质外，数据FAIR原则（可查找、可访问、可互操作和可重用）和可重复性也是其定义原则。

第一次CAMI挑战赛为宏基因组组装、基因组和分类学分箱以及分析程序在多个复杂基准数据集上的表现提供了洞察，包括具有不同进化趋异性和分类学分类不良的类群（如病毒）的未发表基因组。在缺乏菌株多样性的情况下，分箱程序所表现出的稳健性和高准确性支持了它们在各种环境的大规模数据中的应用，恢复了数千个宏基因组组装基因组，并加剧了推进菌株分辨率组装和分箱的努力。在此我们描述第二轮CAMI挑战赛的结果，我们在更大、更复杂的数据集上评估了程序性能和进展，包括长读段数据和更多的性能指标（如运行时间和内存使用）。

结果

我们创建了代表海洋环境、高菌株多样性环境（'菌株疯狂'）和植物相关环境（包括真菌基因组和宿主植物材料）的宏基因组基准数据集。数据集包括从1,680个微生物基因组和599个环状元件中采样的长读段和短读段（方法和补充表1）。其中，772个基因组和所有环状元件是新测序的，与公共基因组序列集合不同（新基因组），其余为高质量公共基因组。与任何其他基因组的平均核苷酸同一性（ANI）低于95%的基因组被归类为“独特基因组”，否则为“常见基因组”，与第一次挑战赛相同。总体而言，901个基因组为独特基因组（474个海洋、414个植物相关、13个菌株疯狂），779个为常见基因组（303个海洋、81个植物相关、395个菌株疯狂）。在这些数据上，提供了组装、基因组分箱、分类学分箱和分析方法的挑战赛，于2019年和2020年开放并允许数月的提交（方法）。此外，还提供了来自一名危重患者（患有未知感染）的临床宏基因组样本的病原体检测挑战赛。鼓励挑战赛参与者通过提供带有参数设置和所用参考数据库的可执行软件来提交可重复的结果。总体而言，收到了来自30个团队的76个程序的5,002个结果（补充表2）。

组装挑战赛

序列组装是宏基因组分析的关键，用于恢复基因组和分类单元分箱。对于进化趋异性较低的基因组，组装质量会下降，导致一致性或碎片化组装。由于其对理解微生物群落的重要性，我们评估了方法使用长读段和短读段数据组装菌株分辨率基因组的能力（方法）。

总体趋势

我们评估了20个组装器版本的155份提交，包括一些具有多种设置和数据预处理选项的提交（补充表2）。此外，我们如参考文献4、16所述创建了金标准共组装和单样本组装。短读段、长读段和混合海洋数据的金标准分别包含2.59、2.60和2.79千兆碱基（Gb）的组装序列，而菌株疯狂金标准各包含1.45 Gb。组装结果使用MetaQUAST v.5.1.0rc（参考文献17）进行评估，该版本已适配用于评估菌株分辨率组装（补充文本）。我们确定了菌株召回率和精确率，类似于参考文献18（方法和补充表3）。为了便于比较，我们根据关键指标（方法）对用不同版本和参数设置产生的方法组装进行排名，并选择排名最高的作为代表（图1、补充图1和补充表3–7）。

短读段组装器在菌株疯狂数据上实现了高达10.4%的基因组覆盖率，在海洋数据上达到41.1%，两者均由MEGAHIT实现。金标准分别报告了90.8%和76.9%（图1a和补充表3）。HipMer在所有指标和数据集上排名最佳，在海洋数据上表现突出，因为它产生了较少的不匹配，同时具有相对较高的基因组覆盖率和NGA50（表1）。在菌株疯狂数据上，GATB排名最佳，HipMer位列第二。在植物相关数据集上，HipMer再次排名最佳，其次是Flye v.2.8（参考文献23），后者在大多数指标上优于其他短读段组装器（补充图2）。

最佳混合组装器A-STAR在基因组覆盖率方面表现突出（海洋44.1%，菌株疯狂30.9%），但产生了更多的错误组装和不匹配（海洋数据每100 kb有773个不匹配）。HipMer在海洋数据上每100 kb的不匹配数最少（67），GATB在菌株疯狂数据上最少（98，图1b）。GATB在海洋数据集的混合组装器中引入的不匹配数最少（173）。ABySS在海洋数据上产生的错误组装最少，GATB在菌株疯狂数据上最少（图1c）。混合组装器OPERA-MS为海洋数据创建了最连续的组装（图1d），所有基因组的平均NGA50为28,244，而金标准为682,777。SPAdes混合提交的NGA50更高，为43,014，但不是排名最高的SPAdes提交。A-STAR在菌株疯狂数据上具有最高的连续性（13,008，而金标准为155,979）。对于短读段组装，MEGAHIT在海洋数据（NGA50 26,599）和菌株疯狂数据（NGA50 4,793）上具有最高的连续性。值得注意的是，Flye在植物相关

长读段数据上表现良好，但在海洋数据上的大多数指标上表现不如其他方法（补充图2），可能是由于不同的版本或参数设置（补充表2）。

对于几个组装器，使用读段质量修剪或纠错软件（如trimmomatic或DUK）进行预处理改善了组装质量（补充表2和3）。基因组覆盖也是一个关键因素（图1g）。虽然短读段和混合组装的金标准包括基因组覆盖率超过90%且覆盖度大于3.3倍的基因组组装，但SPAdes最佳组装了低覆盖度海洋基因组，从9.2倍开始。MEGAHIT、A-STAR、HipMer和Ray Meta分别需要10倍、13.2倍、13.9倍和19.5倍的覆盖度。几个组装器很好地重建了高拷贝环状元件，HipMer、MEGAHIT、SPAdes和A-STAR全部重建（图1g）。与第一次CAMI挑战赛中评估的软件相比，A-STAR在菌株疯狂数据上的基因组覆盖率提高了20%，几乎是MEGAHIT的三倍。HipMer在海洋数据上引入的不匹配数最少（每100 kb 67个不匹配）。这比Ray Meta（也参加CAMI 1的最佳方法）少30%。OPERA-MS在NGA50上比MEGAHIT提高了1,645（6%），尽管使用了双倍（长读段和短读段）数据。SPAdes在第一次挑战中未被评估，在大多数指标上名列前茅。

密切相关的基因组

第一次CAMI挑战赛揭示了独特菌株基因组和常见菌株基因组之间组装质量的显著差异。在所有指标、数据集和软件结果中，独特基因组组装再次表现更优，海洋基因组在菌株召回率上提高9.7%，基因组覆盖率提高19.3%，NGA50提高7倍，菌株精确率提高6.5%，导致更完整和更少碎片化的组装（图1和补充表4–7）。这在菌株疯狂数据集上更为明显，菌株召回率差异为79.1%，基因组覆盖率为75.9%，菌株精确率为20.6%，NGA50差异50倍。虽然独特基因组比常见基因组产生了更多的错误组装（海洋+1.5，菌株疯狂+5.4），这是由于前者的组装尺寸更大，这从相似的错误组装重叠群比例可以看出（独特基因组2.6%，常见基因组3.1%）。虽然独特基因组和常见基因组的重复率相似（海洋+0.01，菌株疯狂-0.08），但独特海洋基因组组装的不匹配数比常见基因组多12%（每100 kb 548个不匹配 vs 486个）。相比之下，独特菌株疯狂基因组组装的不匹配数比常见基因组少62%（每100 kb 199个不匹配 vs 511个不匹配），可能是由于菌株多样性较高。

对于常见海洋基因组，HipMer在所有指标上排名最佳，GATB在常见菌株疯狂基因组上排名最佳。在独特基因组上，HipMer在海洋和菌株疯狂数据集上排名第一。HipMer在常见和独特海洋基因组上具有最高的菌株召回率和精确率（召回率4.5%和20.4%，精确率各为100%）。对于菌株疯狂数据集，A-STAR在常见菌株疯狂基因组上具有最高的菌株召回率（1.5%），但精确率较低（23.1%）。GATB、HipMer、MEGAHIT和OPERA-MS以100%的召回率和精确率组装了独特基因组。A-STAR在基因组覆盖率方面表现突出，在所有四个数据分区中排名第一，HipMer的不匹配数最少。HipMer还在常见和独特海洋基因组上的错误组装数最少，GATB在常见菌株疯狂基因组上错误组装最少，SPAdes在独特基因组上最少。常见海洋基因组上最高的NGA50由OPERA-MS实现，常见菌株疯狂基因组上由A-STAR实现，两个数据集的独特基因组上由SPAdes实现。

results, unique genome assemblies again were superior, for marine genomes by 9.7% in strain recall, 19.3% genome fraction, sevenfold NGA50 and 6.5% strain precision, resulting in more complete and less fragmented assemblies (Fig. 1 and Supplementary Tables 4–7). This was even more pronounced on the strain-madness dataset, with a 79.1% difference in strain recall, 75.9% genome fraction, 20.6% strain precision and 50-fold NGA50. Although there were more misassemblies for unique than for common genomes (+1.5 in

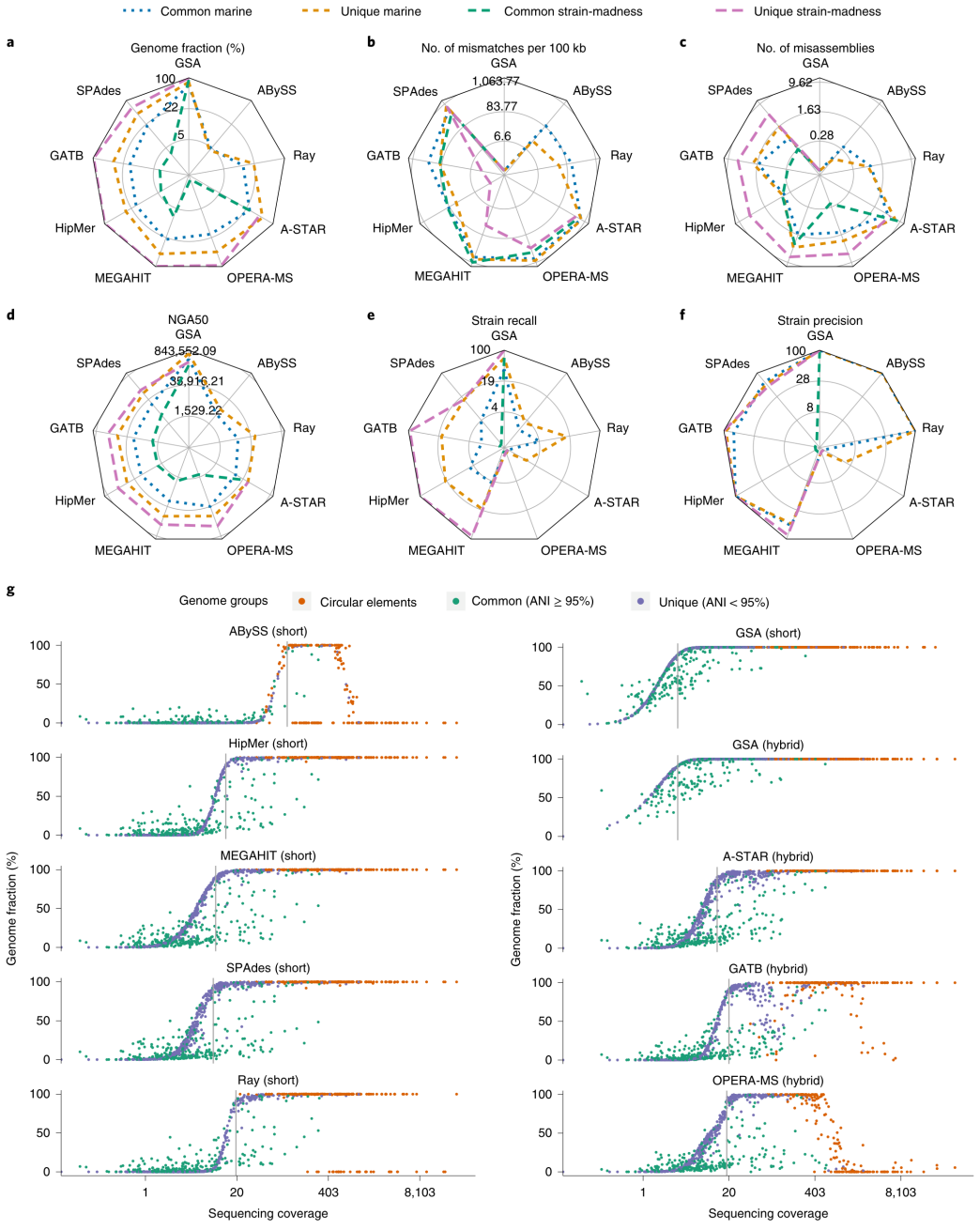

图1 | 宏基因组组装器在海洋和菌株疯狂数据集上的性能。a，基因组覆盖率的雷达图。b，每100千碱基(kb)的不匹配数。c，错误组装数。d，NGA50。e，菌株召回率。f，菌株精确率。对于有多个评估版本的方法，显示在海洋数据上排名最佳的版本（补充图1和补充表3）。指标的绝对值采用对数标度。线条表示分析的基因组子集，GSA的值表示指标的上界。指标显示为独特菌株基因组和常见菌株基因组。g，海洋数据集上基因组恢复比例与基因组测序深度（覆盖度）的关系。蓝色表示独特基因组（<95% ANI），绿色表示常见基因组（ANI≥95%），橙色表示高拷贝环状元件。灰色线表示第一个基因组以≥90%基因组覆盖率被恢复时的覆盖度。

难以组装的区域

我们评估了难以组装区域（如重复序列或保守元件，例如16S核糖体RNA基因）的组装性能，使用的是海洋数据中包含的高质量公共基因组。这些区域对基因组恢复很重要，但经常被遗漏。我们选择了50个具有注释16S序列的独特公共基因组，并在金标准组装（GSA）中以单一重叠群呈现。我们使用Minimap2（参考文献31）将组装提交映射到这些16S序列，并测量其完整性（基因组覆盖率%）和趋异性（补充图3a,b,e）。A-STAR部分恢复了131个16S序列中的102个（78%）。混合组装器GATB（平均完整性60.1%）和OPERA-MS（平均47.1%）恢复了最完整的16S序列。短读段组装的平均完整性从29.6%（HipMer）到36.9%（MEGAHIT）不等。ABySS和HipMer的组装非常准确（趋异性<1%）。混合组装器GATB和OPERA-MS产生了最长的与16S
rRNA基因比对的重叠群，中位长度分别为8,513和4,430个碱基对(bp)，而其他组装器的中位重叠群长度小于16S
rRNA基因的平均长度（1,503
bp）。对于所有组装器和16S序列，共有17个跨基因组嵌合体，由MetaQUAST报告为种间易位：MEGAHIT 10个，A-STAR 5个，HipMer和SPAdes各1个，而GATB、ABySS和OPERA-MS未产生嵌合序列。我们对使用不同方法在50个基因组中的30个中发现的CRISPR盒进行了相同的评估。CRISPR盒区域更容易组装，完整度更高（5-50%）且组装的携带CRISPR的重叠群更长（中位长度高达22倍），比16S rRNA基因更优（补充图3c,d,f）。在组装和方法中，公共基因组在基因组覆盖率和NGA50等关键指标上的平均组装质量优于新基因组（补充图4）。

单样本组装与共组装

对于多样本宏基因组数据集，常见的组装策略是合并样本（共组装）和单样本组装。我们使用以特定覆盖度添加到植物相关数据中的基因组（补充表8），评估了五种组装器结果（补充图5）的两种策略的组装质量。只有HipMer从合并样本中恢复了一个分布在16个样本中的独特基因组，而独特的单样本基因组被所有组装器用两种策略很好地重建。对于在合并样本中常见但独特的样本（LjRoot109, LjRoot170），HipMer在单样本上表现更好，而OPERA-MS在合并样本上表现更好（补充图5），其他组装器则在高基因组覆盖率与更多不匹配之间进行权衡。因此，对于OPERA-MS和短读段组装器在低覆盖度且预期无跨样本菌株多样性的基因组上，共组装通常可以改善组装。对于HipMer，如果覆盖度充足且预期存在密切相关的菌株，单样本组装可能更可取。

表1 | 四个类别中跨数据集、有无菌株多样性及计算需求方面排名最佳的软件

基因组分箱挑战赛

基因组分箱器将重叠群或读段分组以从宏基因组中恢复基因组。我们评估了短读段组装上18个分箱器版本的95个结果：菌株疯狂GSA 22个，菌株疯狂MEGAHIT组装（MA）17个，海洋MA 19个，海洋GSA 15个，植物相关GSA 12个，植物相关MA 10个（补充表9–15）。此外，还评估了植物相关混合组装上的7个结果。方法包括第一次CAMI挑战中表现良好的和流行的软件（补充表2）。对于GSA重叠群，已知真实基因组分配；对于MA，我们认为这是使用MetaQUAST v.5.0.2为重叠群识别的最佳匹配基因组。我们使用AMBER v.2.0.3（参考文献36）评估了平均分箱纯度和基因组完整性（及其F1分数汇总）、恢复的高质量基因组数量以及调整兰德指数（ARI）（方法）。ARI与分箱数据的比例一起量化了整个数据集的分箱性能。

基因组分箱器的性能在指标、软件版本、数据集和组装类型之间存在差异（图2），而参数对性能的影响通常小于3%。对于海洋GSA，平均分箱纯度为81.3±2.3%，基因组完整性为36.9±4.0%（图2a,b和补充表9）。对于海洋MA，平均纯度（78.3±2.6%）相似，而平均完整性仅为21.2±1.6%（图2a,c和补充表10），这是由于大多数分箱器未分箱的1.5-2 kb短重叠群（补充图6）。对于菌株疯狂GSA，平均纯度和完整性相对于海洋GSA分别下降了20.1至61.2±2.3%和18.7至18.2±2.2%（图2a,d和补充表11）。虽然菌株疯狂MA（65.3±4.0%）和GSA的平均纯度相似，但平均完整性进一步降至5.2±0.6%，同样是由于更大比例的未分箱短重叠群（图2a,e和补充表12）。对于植物相关GSA，纯度几乎与海洋相当（78.2±4.5%；图2a,f和补充表13），但分箱完整性相对于其他GSA下降（13.9±1.4%），这是由于低丰度大型真菌基因组的恢复不佳。值得注意的是，拟南芥宿主基因组（5.6倍覆盖度）以及覆盖度超过八倍的真菌以更高的完整性和纯度被分箱（补充图7）。混合组装的分箱进一步将平均纯度提高到85.1±6.3%，而完整性保持相似（11.9±2.1%，补充表14）。对于植物相关MA，平均纯度（83±3.3%）和完整性（12.4±1.5%）与GSA相似（图2a,g和补充表15）。

为了定量评估数据集的金标准和实际组装的分箱器，我们根据指标对提交进行了排名（补充表16–19和补充图8）（方法）。对于海洋和菌株疯狂数据，CONCOCT和MetaBinner在MA上具有最佳的权衡性能，UltraBinner在GSA上最佳，MetaBinner总体最佳。CONCOCT在植物相关组装上也表现最佳（表1）。UltraBinner在海洋GSA上具有最佳完整性，CONCOCT在菌株疯狂GSA和植物相关MA上最佳，MetaWRAP在海洋和菌株疯狂MA上最佳，MaxBin在植物相关GSA上最佳。Vamb始终具有最佳纯度，而UltraBinner在海洋GSA上具有最佳ARI，MetaWRAP在菌株疯狂GSA上最佳，MetaBAT在MA和植物相关组装上最佳。MetaWRAP和MetaBinner分别在海洋和植物相关组装上分配了最多的数据。许多方法分配了所有菌株疯狂重叠

群，尽管ARI较低（图2b–g）。UltraBinner从海洋GSA恢复了最多的高质量基因组，MetaWRAP从海洋MA，CONCOCT从菌株疯狂组装和植物相关GSA，MetaBinner从植物相关GSA和混合组装（图2和补充表20）。对于质粒和其他高拷贝环状元件，Vamb表现最佳，F1分数为70.8%，完整性54.8%，纯度100%，而第二好的方法MetaWRAP的F1分数仅为12.7%（补充表21）。

菌株多样性的影响

对于海洋和菌株疯狂GSA，独特菌株分箱明显优于常见菌株（补充图9和补充表9和11）。差异在菌株疯狂数据上更为明显，其中独特菌株分箱纯度特别高（97.9±0.4%）。UltraBinner在独特基因组的四个数据分区和总体上排名最佳，CONCOCT在常见菌株上排名最佳（补充表22）。UltraBinner在独特菌株上具有最高完整性，而CONCOCT在常见菌株和所有分区上排名最佳。Vamb始终按纯度排名第一，UltraBinner按ARI排名第一，MetaBinner按最多分配排名第一。由于海洋数据集中独特菌株占主导地位，菌株疯狂数据集中常见菌株占主导地位，各自数据和整个数据集的最佳分箱器相同（补充表9和11），大多数指标的性能相似。

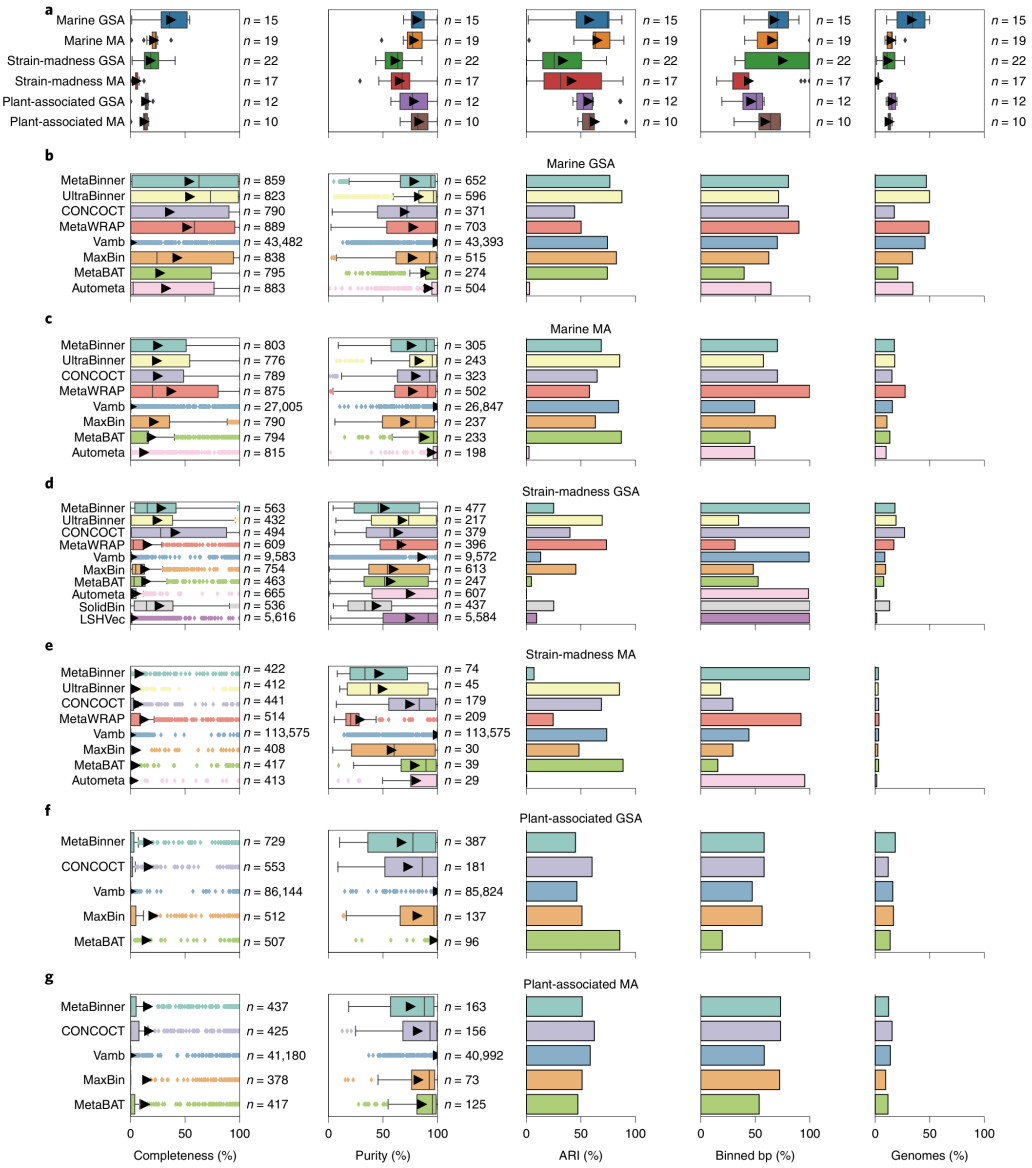

Fig. 2 | Performance of genome binners on short-read assemblies (GSA and MA, MEGAHIT) of the marine, strain-madness, and plant-associated data. a, Boxplots of average completeness, purity, ARI, percentage of binned bp and fraction of genomes recovered with moderate or higher quality (>50% completeness, <10% contamination) across methods from each dataset (Methods). Arrows indicate the average. **b**–**g**, Boxplots of completeness per genome and purity per bin, and bar charts of ARI, binned bp and moderate or higher quality genomes recovered, by method, for each dataset: marine GSA (**b**), marine MA (**c**), strain-madness GSA (**d**), strain-madness MA (**e**), plant-associated GSA (**f**) and plant-associated MA (**g**). The submission with the highest F1-score per method on a dataset is shown (Supplementary Tables 9–15). Boxes in boxplots indicate the interquartile range of *n* results, the center line the median and arrows the average. Whiskers extend to 1.5 × interquartile range or to the maximum and minimum if there is no outlier. Outliers are results represented as points outside 1.5 × interquartile range above the upper quartile and below the lower quartile.

图2 | 基因组分箱器在海洋、菌株疯狂和植物相关数据短读段组装（GSA和MA，MEGAHIT）上的性能。a，各数据集方法的平均完整性、纯度、ARI、分箱bp百分比和中等或更高质量（>50%完整性，<10%污染）恢复基因组比例的箱线图（方法）。箭头表示平均值。b–g，各数据集按方法分组的每个基因组的完整性箱线图和每个分箱的纯度箱线图，以及ARI、分箱bp和恢复的中等或更高质量基因组的条形图：海洋GSA(b)、海洋MA(c)、菌株疯狂GSA(d)、菌株疯狂MA(e)、植物相关GSA(f)和植物相关MA(g)。显示的是每个数据集上每个方法F1分数最高的提交（补充表9–15）。箱线图中的箱体表示n个结果的四分位距，中心线为中位数，箭头为平均值。须延伸至1.5倍四分位距，或在没有异常值时延伸至最大值和最小值。异常值为四分位距1.5倍以上或以下的点。

分类学分箱挑战赛

分类学分箱器将序列分组到带有分类学标识符的标记分箱中。我们评估了9种方法和版本的547个结果：海洋数据75个，菌株疯狂405个，植物相关67个，分别基于读段或GSA（补充表2）。我们使用提供给参与者的NCBI分类学版本评估了不同分类等级的分箱平均纯度和完整性以及每个样本的准确性（方法）。

在海洋数据上，跨分类等级的平均分类单元分箱完整性为63%，平均纯度40.3%，每样本bp准确性74.9%（图3a和补充表23）。在菌株疯狂数据上，准确性相似（76.9%，图3b和补充表24），而完整性高约10%，纯度相应降低。在植物相关数据上，纯度介于前两个数据集之间（35%），但完整性和准确性较低（分别为44.2%和50.8%；图3c和补充表25）。对于所有数据集，性能在较低分类等级下降，最显著的是从属到种级别，完整性平均下降22.2%，纯度下降9.7%，准确性下降18.5%。

在所有数据集中，MEGAN在重叠群上的表现在所有指标和分类等级上排名第一（补充表26），紧随其后的是Kraken v.2.0.8 beta（重叠群上）和Ganon（短读段上）。Kraken在重叠群上对属和种级别最佳，在海洋数据上所有指标以及完整性和准确性方面最佳（89.4%和96.9%，补充表23和27以及补充图10）。由于公共基因组的存在，Kraken在海洋数据上的完整性远高于第一次CAMI挑战，特别是在种和属级别（平均分别为84.6%和91.5%，而之前为50%和5%），而纯度保持相似。MEGAN在重叠群上在海洋和植物相关数据的分类单元分箱纯度上排名最高（分别为90.7%和87.1%，补充表23、25、27和28）。PhyloPythiaS+在菌株疯狂数据上所有指标以及跨等级的完整性（90.5%）和纯度（75.8%）上排名最佳（补充表24和29）。DIAMOND在重叠群上在植物相关数据的完整性（67.6%）上排名最佳，Ganon在短读段上在准确性（77.1%）上排名最佳。

过滤每个分类等级1%最小的预测分箱是一种流行的后处理方法。在所有数据集中，过滤将平均纯度提高到71%以上，并降低了完整性，在海洋和菌株疯狂数据上降至约24%，在植物相关数据上降至13.4%（补充表23–25）。准确性受影响不大，因为大型分箱对此指标贡献更多。Kraken在重叠群上仍在过滤后准确性中排名第一，MEGAN在所有过滤后指标中排名第一（补充表26）。MEGAN在重叠群上和Ganon在短读段上从过滤中获益最多，分别在所有数据集和分类等级的过滤后完整性和纯度中排名第一。

趋异基因组的分类学分箱

为了研究查询序列与参考序列之间趋异性增加对分类学分箱器的影响，我们根据基因组与公共基因组的距离对基因组进行了分类（补充图11和补充表30和31）。已知海洋菌株的序列在种级别上被Kraken特别好地分配（准确性、完整性和过滤后纯度均超过93%），MEGAN也很好（91%纯度，33%完整性和准确性）。Kraken也最佳地在种级别对新菌株序列进行了分类，尽管海洋数据的完整性和准确性较低（分别为68%和80%）。它在所有等级上具有最佳准确性和完整性，但未过滤纯度较低。对于菌株疯狂数据，PhyloPythiaS+在属级别以下表现同样出色，并在属级别上最佳地分配了新物种（93%准确性和完整性，75%过滤后纯度）。只有DIAMOND正确分类了病毒重叠群，尽管纯度较低（50%），完整性和准确性也低（均为3%）。

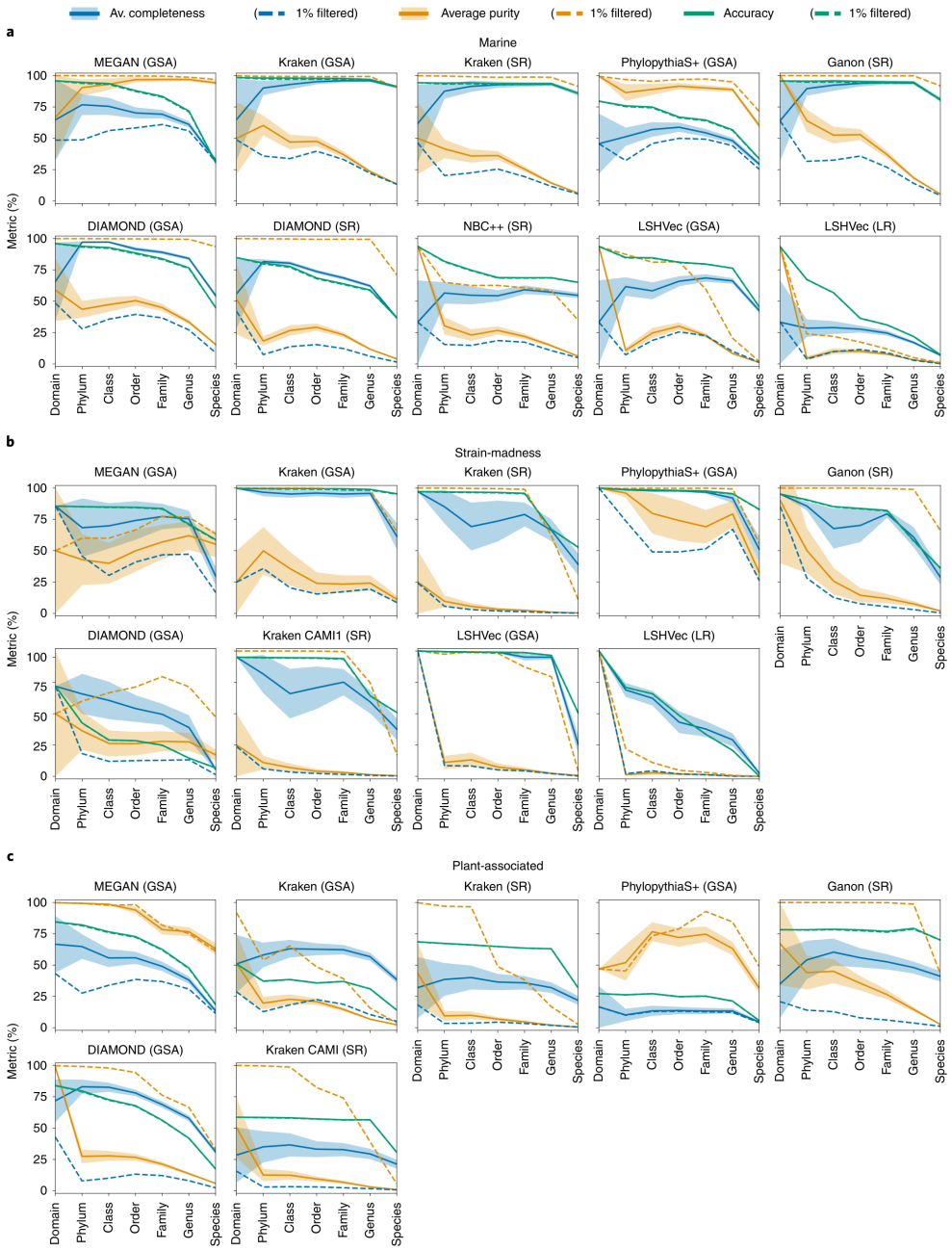

Fig. 3 | Taxonomic binning performance across ranks per dataset. a, Marine. **b**, Strain-madness. **c**, Plant-associated. Metrics were computed over unfiltered (solid lines) and 1%-filtered (that is, without the 1% smallest bins in bp, dashed lines) predicted bins of short reads (SR), long reads (LR) and contigs of the GSA. Shaded bands show the standard error across bins.

图3 | 各数据集跨分类等级的分类学分箱性能。a，海洋。b，菌株疯狂。c，植物相关。指标在短读段(SR)、长读段(LR)和GSA重叠群的未过滤（实线）和1%过滤（即去除按bp计最小的1%分箱，虚线）的预测分箱上计算。阴影带表示分箱间的标准误。

分类学谱分析挑战赛

分类学分析器从宏基因组样本中量化微生物群落的存在和相对分类单元丰度。这与分类学序列分类不同，后者将分类标签分配给单个序列，产生分类单元特异性的序列分箱（和序列丰度谱）。我们评估了来自22个方法版本的4,195个

分析结果（292个海洋、2,603个菌株疯狂和1,300个植物相关数据集）（补充表2），大多数结果为短读段样本，少数为长读段样本、组装或跨样本平均值。性能使用OPAL v.1.0.10（参考文献42）进行评估（方法）。预测分类单元谱的质量基于与潜在真实情况的比较，确定了各分类等级的已识别分类单元的完整性和纯度，而分类单元丰度估计使用各等级的L1范数和跨等级的加权UniFrac误差进行评估。Alpha多样性估计的准确性使用Shannon均匀度指数测量（方法）。总体而言，mOTUs v.2.5.1和MetaPhlAn v.2.9.22在海洋和植物相关数据集上跨分类等级和指标排名最佳，mOTUs v.cami1和MetaPhlAn v.2.9.22在菌株疯狂数据集上排名最佳（表1、补充表33、35和37以及补充图12）。

分类单元鉴定

方法在属级别以下表现良好（海洋平均纯度70.4%，菌株疯狂52.1%，植物相关62.9%；海洋平均完整性63.3%，菌株疯狂80.5%，植物相关42.1%；图4a,c、补充图13和补充表32、34和36），在种级别有显著下降。mOTUs v.2.5.1（参考文献43）在海洋数据上的属和种级别的完整性和纯度均超过80%，Centrifuge v.1.0.4 beta（参考文献44）和MetaPhlAn v.2.9.22（参考文献45,46）在属级别（图4a）。Bracken（参考文献47）和NBC++（参考文献48）在任一级别的完整性超过80%，CCMetagen（参考文献49）、DUDes v.0.08（参考文献50）、LSHVec v.gsa（参考文献51）、Metalign（参考文献52）、MetaPalette（参考文献53）和MetaPhlAn v.cami1的纯度超过80%。过滤每个等级最稀有的（1%）预测分类单元使完整性降低约22%，同时精确度提高约11%。

在菌株疯狂数据的属级别上，MetaPhlAn v.2.9.22（89.2%完整性，92.8%纯度）、MetaPhyler v.1.25（参考文献54）（92.3%完整性，79.2%纯度）和mOTUs v.cami1（92.9%完整性，69.1%纯度）表现最佳，但没有方法在种级别上表现突出。DUDes v.0.08和LSHVec v.gsa具有高纯度，而Centrifuge v.1.0.4 beta、DUDes v.cami1、TIPP v.4.3.10（参考文献55）和TIPP v.cami1具有高完整性。在植物相关数据的属级别上，sourmash_gather v.3.3.2_k31_sr（参考文献56）总体最佳（53.3%完整性，89.5%纯度）。Sourmash_gather v.3.3.2_k31在PacBio读段上和MetaPhlAn v.3.0.7在属（98.5%, 95.5%）和种级别（64.4%, 68.8%）上具有最高纯度，sourmash_gather v.3.3.2_k21_sr具有最高完整性（属61.9%，种53.8%）。

相对丰度

各等级和提交的丰度预测在菌株疯狂数据上平均优于海洋数据，后者在菌株级别以上的复杂性较低，L1范数从0.44改善到0.3，平均加权UniFrac误差从4.65改善到3.79（补充表32、34和36）。这些加权UniFrac值远高于生物学重复样本的值（0.22，方法）。植物相关数据上的丰度预测不如其他数据集，L1范数平均为0.57，加权UniFrac为5.16。在海洋数据上，mOTUs v.2.5.1在几乎所有等级上具有最低的L1范数，平均为0.12，属级别0.13，种级别0.34。其次是MetaPhlAn v.2.9.22（平均0.22，属0.32，种0.39）。两种方法也具有最低的加权UniFrac误差，其次是DUDEs v.0.08。在菌株疯狂数据上，mOTUs v.cami1在跨等级的L1范数上表现最佳（平均0.05），属和种分别为0.1和0.15，其次是MetaPhlAn v.2.9.22（平均0.09，属0.12，种0.23）。后者也具有最低的加权UniFrac误差，其次是TIPP v.cami1和mOTUs v.2.0.1。在植物相关数据上，Bracken v.2.6在跨等级的L1范数上表现最佳，平均0.36，属级别0.34。Sourmash_gather v.3.3.2_k31在短读段上在种级别上具有最低值（0.55）。两种方法也在此数据集上具有最低的UniFrac误差。几种方法使用Shannon均匀度准确重建了样本的alpha多样性；在海洋数据上跨等级最佳（与金标准的绝对差异≤0.03）的是：mOTUs v.2.5.1、DUDes v.0.08和v.cami1以及MetaPhlAn v.2.9.22和v.cami1；在菌株疯狂数据上：DUDes v.cami1、mOTUs v.cami1和MetaPhlAn v.2.9.22。在植物相关数据上，mOTU v.cami1和Bracken v.2.6在此指标上表现最佳（0.08和0.09）。

难易分类单元

对于所有方法，病毒、质粒和古菌在海洋数据中难以检测（补充图14和补充表38）。虽然几种方法检测到了许多古菌分类单元，但其他如Candidatus Nanohaloarchaeota则完全未被检测到。只有Bracken和Metalign检测到了病毒。相比之下，Terrabacteria群中的细菌分类单元以及拟杆菌门和变形菌门始终被正确检测到。基于提交的分类单元精确率和召回率，使用相似信息的方法倾向于聚类（补充图15）。

临床病原体预测：概念挑战赛

从宏基因组数据进行临床病原体诊断是一个高度相关的转化问题，需要计算处理。为提高认识，我们提供了一个概念挑战赛（方法）：提供了一份来自出血热患者血液样本的短读段宏基因组数据集，供参与者识别病原体并指出那些可能导致病例报告中所述症状的病原体。收到了10个经过人工筛选的结果，因此不可重复（补充表39）。每个结果中识别的分类单元数量差异很大（补充图16）。三份提交使用分类学分析器MetaPhlAn v.2.2、Bracken v.2.5和CCMetagen v.1.1.3（参考文献49）正确识别了致病病原体——克里米亚-刚果出血热正内罗病毒（CCHFV）。另一份使用Bracken v.2.2的提交正确识别了正内罗病毒，但未将其识别为致病病原体。

计算需求

我们测量了提交方法在海洋和菌株疯狂数据上的运行时间和内存使用（图5、补充表40和方法）。每个方法类别中都有能够在几分钟到几小时内处理整个数据集的高效方法，包括一些在其他指标上排名靠前的技术。各类别内甚至版本之间差异显著，从可在标准台式机上执行的方法到需要大量硬件和重度并行化的方法不等。MetaHipMer是最快的组装器，处理海洋短读段需要2.1小时，比第二快的组装器MEGAHIT少3.3倍。然而，MetaHipMer使用的内存最多（1,961吉字节(GB)）。MEGAHIT使用的内存最少（42 GB），其次是GATB（56.6 GB）。在海洋组装上，基因组分箱器平均所需时间约为较小的菌株疯狂组装的三分之一（29.2 vs 86.1小时），但使用的内存几乎是四倍（69.9 vs 18.5 GB）。MetaBAT v.2.13.33在两个数据集上都是最快（1.07和0.05小时）和内存效率最高（最大内存使用2.66和1.5 GB）的分箱器。它比第二快的方法Vamb v.fa045c0分别快约5倍和635倍，比LSHVec v.1dfe822在海洋数据上快约6倍，比SolidBin v.1.3在菌株疯狂数据上快765倍；比排名第二的MaxBin v.2.0.2和CONCOCT v.1.1.0在海洋数据上的内存效率分别高约2倍和5倍。MetaBAT和CONCOCT都比其CAMI 1版本分别快约11倍和4倍。与基因组分箱器一样，分类学分箱器在海洋组装上比菌株疯狂组装运行时间更长，例如PhyloPythiaS+分别为287.3小时和36小时，但内存使用相似或略高。

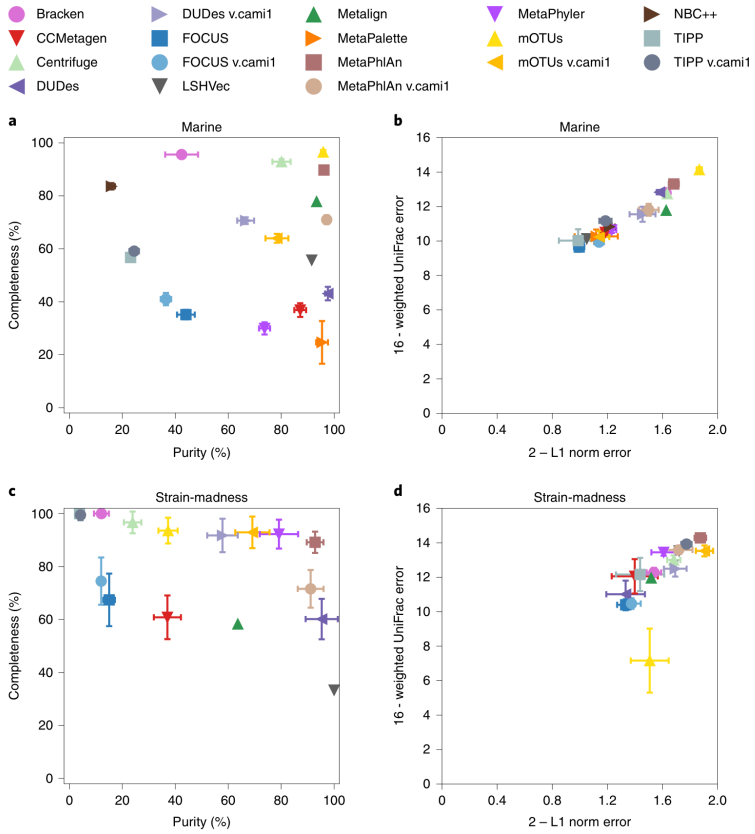

Fig. 4 | Taxonomic profiling results for the marine and strain-madness datasets at genus level. a,**b**, Marine datasets. **c**,**d**, Strain-madness datasets. Results are shown for the overall best ranked submission per software version (Supplementary Tables 33 and 35, and Supplementary Fig. 12). **a**,**c**, Purity versus completeness. **b**,**d**, Upper bound of L1 norm (2) minus actual L1 norm versus upper bound of weighted UniFrac error (16) minus actual weighted UniFrac error. Symbols indicate the mean over ten marine and 100 strain-madness samples, respectively, and error bars the standard deviation. Metrics were determined using OPAL with default settings.

received (Supplementary Table 39). The number of identified taxa per result varied considerably (Supplementary Fig. 16). Three submissions correctly identified the causal pathogen, Crimean–Congo hemorrhagic fever orthonairovirus (CCHFV), using the taxonomic profilers MetaPhlAn v.2.2, Bracken v.2.5 and CCMetagen v.1.1.3 (ref. [49]). Another submission using Bracken v.2.2 correctly identified orthonairovirus, but not as the causal pathogen.

Computational requirements. We measured the runtimes and memory usage for submitted methods across the marine and strain-madness data (Fig. 5, Supplementary Table 40 and Methods). Efficient methods capable of processing the entire datasets within minutes to a few hours were available in every method category, including some top ranked techniques with other metrics. Substantial differences were seen within categories and even between versions, ranging from methods executable on standard desktop machines to those requiring extensive hardware and heavy parallelization. MetaHipMer was the fastest assembler and required 2.1 h to process marine short reads, 3.3× less than the second fastest assembler, MEGAHIT. However, MetaHipMer used the most memory (1,961 gigabytes (GB)). MEGAHIT used the least memory (42 GB), followed by GATB (56.6 GB). On marine assemblies, genome binners on average required roughly three times less time than for the smaller strain-madness assemblies (29.2 versus 86.1 h), but used almost 4× more memory (69.9 versus 18.5 GB). MetaBAT v.2.13.33 was the fastest (1.07 and 0.05 h) and most memory efficient binner (maximum memory usage 2.66 and 1.5 GB) on both datasets. It was roughly 5× and 635× faster than the second fastest method, Vamb v.fa045c0, roughly 6× faster than LSHVec v.1dfe822 on marine and 765× faster than SolidBin v.1.3 on strain-madness data; roughly twice and 5× more memory efficient than the next ranking MaxBin v.2.0.2 and CONCOCT v.1.1.0 on marine data, respectively. Both MetaBAT and CONCOCT were substantially (roughly 11× and 4×) faster than their CAMI 1 versions. Like genome binners, taxonomic binners ran longer on the marine than the strain-madness assemblies, for example PhyloPythiaS+ with 287.3 versus 36 h, respectively, but had a similar or slightly

图4 | 海洋和菌株疯狂数据集属级别的分类学谱分析结果。a,b，海洋数据集。c,d，菌株疯狂数据集。结果显示每个软件版本总体排名最佳的提交（补充表33和35，以及补充图12）。a,c，纯度与完整性的关系。b,d，L1范数上界(2)减去实际L1范数与加权UniFrac误差上界(16)减去实际加权UniFrac误差的关系。符号表示10个海洋样本和100个菌株疯狂样本的平均值，误差条为标准差。指标使用OPAL默认设置确定。

在海洋读段数据上，分类单元分析器平均快近4倍（16.1 vs 60.8小时），比大十倍的菌株疯狂读段数据集使用的内存更多（38.1 vs 25 GB）。最快和内存效率最高的分类学分箱器是Kraken，分别仅需0.05和0.02小时，两个数据集约37

GB内存，适用于读段或重叠群。其次是DIAMOND，在海洋和菌株疯狂GSA上分别运行约500倍和910倍的时间。FOCUS v.1.5（参考文献58）和Bracken v.2.2分别是海洋（0.51，0.66小时）和菌株疯狂（1.89，3.45小时）数据上最快的分析器。FOCUS v.1.5还需要的内存最少（海洋0.16 GB，菌株疯狂0.17 GB），其次是mOTUs v.1.1.1和MetaPhlAn v.2.2.0。

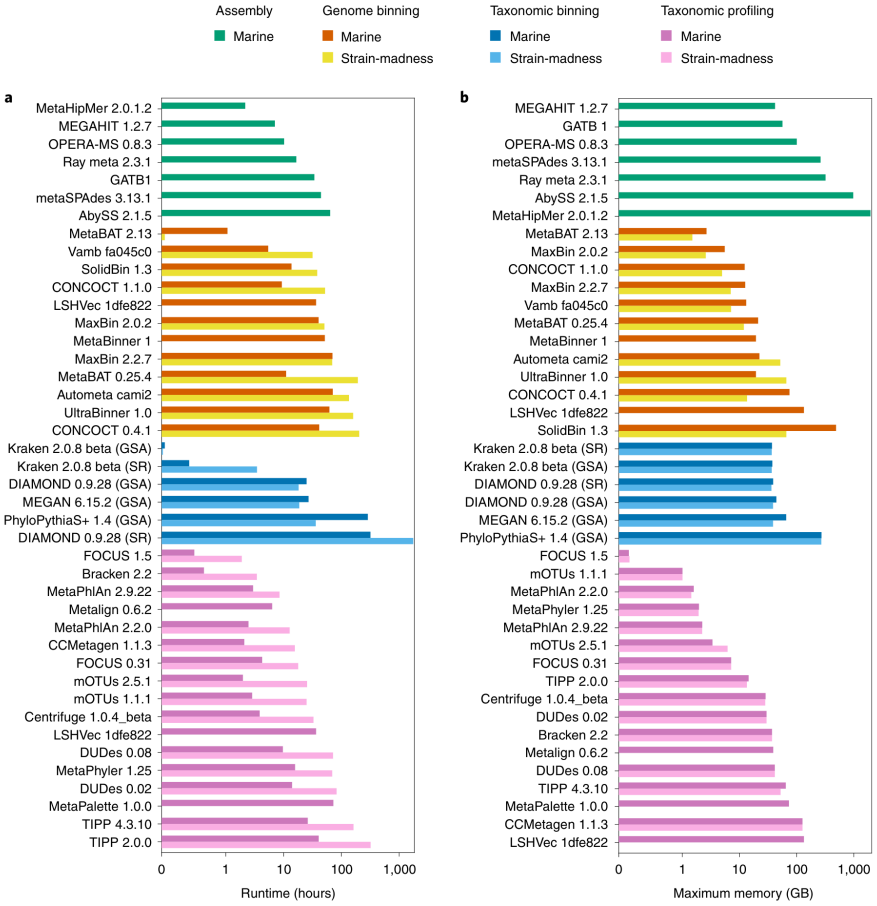

Fig. 5 | Computational requirements of software from all categories. a, Runtime. **b**, Maximum memory usage. Results are reported for the marine and strain-madness read data or GSAs (Supplementary Table 40). The *x* axes are log scaled and the numbers given are the software version numbers.

higher memory usage. On the marine read data, taxon profilers, however, were almost 4× faster on average (16.1 versus 60.8 h) than on the ten times larger strain-madness read dataset, but used more memory (38.1 versus 25 GB). The fastest and most memory efficient taxonomic binner was Kraken, requiring only 0.05 and 0.02 h, respectively, and roughly 37 GB memory on both datasets, for reads or contigs. It was followed by DIAMOND, which ran roughly 500× and 910× as long on the marine and strain-madness GSAs, respectively. FOCUS v.1.5 (ref. [58]) and Bracken v.2.2 were the fastest profilers on the marine (0.51, 0.66 h, respectively) and strain-madness (1.89, 3.45 h) data. FOCUS v.1.5 also required the least memory (0.16 GB for marine, 0.17 GB for strain-madness), followed by mOTUs v.1.1.1 and MetaPhlAn v.2.2.0.

Discussion

Assessing metagenomic analysis software thoroughly, comprehensively and with little bias is key for optimizing data processing strategies and tackling open challenges in the field. In its second round, CAMI offered a diverse set of benchmarking challenges across a comprehensive data collection reflecting recent technical developments. Overall, we analyzed 5,002 results of 76 program versions with different parameter settings across 131 long- and short-read metagenome samples from four datasets (marine, plant-associated, strain-madness, clinical pathogen challenge). This effort increased the number of results 22× and the number of benchmarked software versions 3× relative to the first challenge, delivering extensive new insights into software performances across a range of conditions. By systematically assessing runtime and memory requirements, we added two more key performance dimensions to the benchmark, which are important to consider given the ever-increasing dataset sizes.

In comparison to software assessed in the first challenges, assembler performances rose by up to 30%. Still, in the presence of closely related strains, assembly contiguity, genome fractions and strain recall decreased, suggesting that most assemblers, sometimes intentionally[19,26], did not resolve strain variation, resulting in more fragmented, less strain-specific assemblies. In addition, genome coverage, parameter settings and data preprocessing impacted assembly quality, while performances were similar across software

图5 | 所有类别软件的计算需求。a，运行时间。b，最大内存使用。结果报告海洋和菌株疯狂读段数据或GSA（补充表40）。x轴为对数标度，给出的数字为软件版本号。

讨论

对宏基因组分析软件进行彻底、全面且少偏见的评估是优化数据处理策略和解决该领域开放性挑战的关键。在第二轮中，CAMI提供了跨综合数据集合的多样化基准测试挑战赛，反映了最新的技术发展。总体而言，我们分析了来自四个数据集（海洋、植物相关、菌株疯狂、临床病原体挑战）的131个长读段和短读段宏基因组样本上76个程序版本不同参数设置的5,002个结果。这一努力使结果数量相对于第一次挑战增加了22倍，基准测试的软件版本数量增加了3倍，为各种条件下的软件性能提供了广泛的新见解。通过系统评估运行时间和内存需求，我们为基准测试增加了两个关键性能维度，考虑到不断增长的数据集大小，这一点很重要。

与第一次挑战中评估的软件相比，组装器性能提高了多达30%。尽管如此，在存在密切相关菌株的情况下，组装连续性、基因组覆盖率和菌株召回率下降，表明大多数组装器有时是有意地不解决菌株变异，导致更碎片化、更少菌株特异性的组装。此外，基因组覆盖度、参数设置和数据预处理影响组装质量，而不同软件版本之间的性能相似。大多数提交的宏基因组组装仅使用短读段，长读段和混合组装没有更高的总体质量。然而，混合组装在难以组装的区域（如16S rRNA基因）表现更好，比大多数短读段提交恢复了更完整的基因。混合组装器在合并样本中也较少受到密切相关菌株的影响，表明长读段有助于区分菌株。

与第一次CAMI挑战相比，集成分箱器展示了一项发展，在大多数指标上比单个方法有显著改进。总体而言，基因组分箱器在指标和数据集类型之间表现出不同的性能，菌株多样性和较低的组装质量是显著降低性能的挑战，即使对于菌株疯狂数据集的大量样本也是如此。对于植物相关数据中具有足够覆盖度的植物宿主和55个真菌基因组，也获得了高质量分箱。

对于分类学分箱器和分析器，有高性能且计算效率高的软件可用，在各种条件和指标下表现良好。特别是分析器自第一次挑战以来已经成熟，在分类单元鉴定、丰度和多样性估计方面顶级表现者之间的差异较小。性能在属及以上级别较高，在细菌种级别有显著下降。由于第二次挑战数据包含高质量公共基因组，数据与公开可用数据的趋异性低于第一次挑战（后者的方法性能已经从科到属级别下降）。古菌和病毒的性能也较低，表明开发者需要扩展其参考序列集合和模型开发。另一个令人鼓舞的结果是，在临床病原体挑战中，几份提交识别了致病病原体。然而，由于人工筛选，没有一个是可重复的，表明这些方法仍需改进，以及在大数据集合上进行评估。尽管临床宏基因组学在病原体诊断和表征方面具有巨大潜力，但多个挑战仍阻碍其在常规诊断中的应用。

在其第二次挑战中，CAMI确定了常见宏基因组软件类别的关键进展以及当前挑战。随着方法和数据生成技术的进步，不断重新评估这些问题将很重要。此外，其他微生物组数据模式和多功能组学数据整合的计算方法可以被联合评估。最重要的是，CAMI是一个社区驱动的努力，我们鼓励所有对微生物组研究基准测试感兴趣的人加入我们。

在线内容：任何方法、补充参考文献、Nature Research报告摘要、源数据、扩展数据、补充信息、致谢、同行评审信息；作者贡献和利益冲突的详情；以及数据和代码可用性声明可在https://doi.org/10.1038/s41592-022-01431-4获取。

收稿日期：2021年7月22日；接受日期：2022年2月14日；在线发表日期：2022年4月8日

参考文献

1. Ghurye, J. S., Cepeda-Espinoza, V. & Pop, M. Metagenomic assembly: overview, challenges and applications. Yale J. Biol. Med. 89, 353-362 (2016).
2. Breitwieser, F. P., Lu, J. & Salzberg, S. L. A review of methods and databases for metagenomic classification and assembly. Brief. Bioinform. 20, 1125-1136 (2019).
3. Sangwan, N., Xia, F. & Gilbert, J. A. Recovering complete and draft population genomes from metagenome datasets. Microbiome 4, 8 (2016).
4. Sczyrba, A. et al. Critical Assessment of Metagenome Interpretation: a benchmark of metagenomics software. Nat. Methods 14, 1063-1071 (2017).
5. McIntyre, A. B. R. et al. Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. Genome Biol. 18, 182 (2017).
6. Van Den Bossche, T. et al. Critical Assessment of Metaproteome Investigation (CAMPI): a multi-lab comparison of established workflows. Nat. Commun. 12, 7305 (2021).
7. Commichaux, S. et al. A critical assessment of gene catalogs for metagenomic analysis. Bioinformatics 37, 2848-2857 (2021).

8. Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3, 160018 (2016).

9. Pasolli, E. et al. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. Cell 176, 649-662.e20 (2019).

10. Almeida, A. et al. A new genomic blueprint of the human gut microbiota. Nature 568, 499-504 (2019).

11. Bremges, A. & McHardy, A. C. Critical assessment of metagenome interpretation enters the second round. mSystems 3, e00103-e00118 (2018).

12. Turnbaugh, P. J. et al. The human microbiome project. Nature 449, 804-810 (2007).

13. Meyer, F. et al. Tutorial: assessing metagenomics software with the CAMI benchmarking toolkit. Nat. Protoc. 16, 1785-1801 (2021).

14. Nawy, T. Microbiology: the strain in metagenomics. Nat. Methods 12, 1005 (2015).

15. Segata, N. On the road to strain-resolved comparative metagenomics. mSystems 3, e00190-17 (2018).

16. Fritz, A. et al. CAMISIM: simulating metagenomes and microbial communities. Microbiome 7, 17 (2019).

17. Mikheenko, A., Saveliev, V. & Gurevich, A. MetaQUAST: evaluation of metagenome assemblies. Bioinformatics 32, 1088-1090 (2016).

18. Fritz, A. et al. Haploflow: strain-resolved de novo assembly of viral genomes. Genome Biol. 22, 212 (2021).

19. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics 31, 1674-1676 (2015).

20. Hofmeyr, S. et al. Terabase-scale metagenome coassembly with MetaHipMer. Sci. Rep. 10, 10689 (2020).

21. Drezen, E. et al. GATB: genome assembly & analysis tool box. Bioinformatics 30, 2959-2961 (2014).

22. Chikhi, R. & Rizk, G. Space-efficient and exact de Bruijn graph representation based on a Bloom filter. Algorithms Mol. Biol. 8, 22 (2013).

23. Kolmogorov, M. et al. metaFlye: scalable long-read metagenome assembly using repeat graphs. Nat. Methods 17, 1103-1110 (2020).

24. Simpson, J. T. et al. ABySS: a parallel assembler for short read sequence data. Genome Res. 19, 1117-1123 (2009).

25. Bertrand, D. et al. Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes. Nat. Biotechnol. 37, 937-944 (2019).

26. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. Genome Res. 27, 824-834 (2017).

27. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114-2120 (2014).

28. Li, M., Copeland, A. & Han, J. DUK - A Fast and Efficient Kmer Based Sequence Matching Tool, Lawrence Berkeley National Laboratory. LBNL Report #: LBNL-4516E-Poster (2011).

29. Boisvert, S., Raymond, F., Godzaridis, E., Laviolette, F. & Corbeil, J. Ray Meta: scalable de novo metagenome assembly and profiling. Genome Biol. 13, R122 (2012).

30. Maguire, F. et al. Metagenome-assembled genome binning methods with short reads disproportionately fail for plasmids and genomic Islands. Micro. Genom. 6, mgen000436 (2020).

31. Li, H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34, 3094-3100 (2018).

32. Bland, C. et al. CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. BMC Bioinf. 8, 209 (2007).

33. Couvin, D. et al. CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. Nucleic Acids Res. 46, W246-W251 (2018).

34. Mreches, R. et al. GenomeNet/deepG: DeepG pre-release version. Zenodo https://doi.org/10.5281/zenodo.5561229 (2021).

35. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. Nature 486, 207-214 (2012).

36. Meyer, F. et al. AMBER: assessment of metagenome BinnERs. Gigascience 7, giy069 (2018).

37. Alneberg, J. et al. Binning metagenomic contigs by coverage and composition. Nat. Methods 11, 1144-1146 (2014).

38. Wu, Y.-W., Simmons, B. A. & Singer, S. W. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics 32, 605-607 (2016).

39. Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ 3, e1165 (2015).

40. Kang, D. D. et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ 7, e7359 (2019).

41. Sun, Z. et al. Challenges in benchmarking metagenomic profilers. Nat. Methods 18, 618-626 (2021).

42. Meyer, F. et al. Assessing taxonomic metagenome profilers with OPAL. Genome Biol. 20, 51 (2019).

43. Milanese, A. et al. Microbial abundance, activity and population genomic profiling with mOTUs2. Nat. Commun. 10, 1014 (2019).

44. Kim, D., Song, L., Breitwieser, F. P. & Salzberg, S. L. Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Res. 26, 1721-1729 (2016).

45. Segata, N. et al. Metagenomic microbial community profiling using unique clade-specific marker genes. Nat. Methods 9, 811-814 (2012).

46. Beghini, F. et al. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. eLife 10, e65088 (2021).

47. Lu, J., Breitwieser, F. P., Thielen, P. & Salzberg, S. L. Bracken: estimating species abundance in metagenomics data. PeerJ Computer Sci. 3, e104 (2017).

48. Zhao, Z., Cristian, A. & Rosen, G. Keeping up with the genomes: efficient learning of our increasing knowledge of the tree of life. BMC Bioinf. 21, 412 (2020).

49. Marcelino, V. R. et al. CCMetagen: comprehensive and accurate identification of eukaryotes and prokaryotes in metagenomic data. Genome Biol. 21, 103 (2020).

50. Piro, V. C., Lindner, M. S. & Renard, B. Y. DUDes: a top-down taxonomic profiler for metagenomics. Bioinformatics 32, 2272-2280 (2016).

51. Shi, L. & Chen, B. LSHvec: a vector representation of DNA sequences using locality sensitive hashing and FastText word embeddings. In Proc. 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics 1-10 (Association for Computing Machinery, 2021).

52. LaPierre, N., Alser, M., Eskin, E., Koslicki, D. & Mangul, S. Metalign: efficient alignment-based metagenomic profiling via containment min hash. Genome Biol. 21, 242 (2020).

53. Koslicki, D. & Falush, D. MetaPalette: a k-mer painting approach for metagenomic taxonomic profiling and quantification of novel strain variation. mSystems 1, e00020-16 (2016).

54. Liu, B., Gibbons, T., Ghodsi, M., Treangen, T. & Pop, M. Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. BMC Genomics 12, S4 (2011).

55. Shah, N., Molloy, E. K., Pop, M. & Warnow, T. TIPP2: metagenomic taxonomic profiling using phylogenetic markers. Bioinformatics 37, 1839-1845 (2021).

56. Pierce, N. T., Irber, L., Reiter, T., Brooks, P. & Brown, C. T. Large-scale sequence comparisons with sourmash. F1000Res. 8, 1006 (2019).

57. Chiu, C. Y. & Miller, S. A. Clinical metagenomics. Nat. Rev. Genet. 20, 341-355 (2019).

58. Silva, G. G. Z., Cuevas, D. A., Dutilh, B. E. & Edwards, R. A. FOCUS: an alignment-free model to identify organisms in metagenomes using non-negative least squares. PeerJ 2, e425 (2014).

59. Dulanto Chiang, A. & Dekker, J. P. From the pipeline to the bedside: advances and challenges in clinical metagenomics. J. Infect. Dis. 221, S331-S340 (2020).

方法

社区参与

我们在公开研讨会和黑客马拉松中收集了社区对基准测试挑战赛和数据集的性质和实施原则的意见（https://www.microbiome-cosi.org/cami/participate/schedule）。在与挑战赛参与者和评估软件开发者的公开研讨会中讨论了性能评估和数据解释的最相关指标，其中以匿名方式展示了第一次挑战的结果。de.NBI云为挑战赛参与者提供了计算支持。

标准化和可重复性

为确保可重复性并评估用于创建挑战提交的软件的计算行为（运行时间和内存消耗），我们根据提交规范重新生成和重新评估了结果（补充表2，https://data.cami-challenge.org/）。对于宏基因组组装器，计算需求在配备Intel Xeon处理器（2.6 GHz）的机器上评估，虚拟化为56核（使用50核）和2,755 GB主内存；对于分箱器和分析器，在配备Intel Xeon E5-4650 v4 CPU（虚拟化为16个CPU核心，每核一线程）和512 GB主内存的机器上评估。方法逐个在每台硬件上独占执行。我们还更新了实现一系列常用性能指标的Docker BioContainers，以包含本次评估中使用的所有指标（MetaQUAST17: https://quay.io/repository/biocontainers/quast，AMBER36和https://quay.io/repository/biocontainers/cami-amber，OPAL42: https://quay.io/repository/biocontainers/cami-opal）。

基因组测序和组装

796个新测序基因组的Illumina双端读段数据，其中224个来自拟南芥根环境，176个来自海洋环境，384个来自临床肺炎链球菌菌株，12个来自小鼠肠道环境，使用SPAdes宏基因组组装器（v.3.12）的流水线进行组装。去除了小于1 kb的重叠群，以及由CheckM v.1.011确定的污染≥5%且完整性≤90%的基因组组装。新组装和数据库基因组使用CAMITAX进行分类学分类，并用作CAMISIM微生物群落和宏基因组数据模拟的输入，海洋和植物相关数据集使用from_profile模式，菌株疯狂数据集使用de novo模式。所有脚本和参数见补充材料和GitHub（https://github.com/CAMI-challenge/second_challenge_evaluation/tree/master/scripts/data_generation）。

对于质粒数据集，使用了来自丹麦西兰岛污水处理厂的进水来生成类似参考文献64的质粒样本。在NextSeq 500上使用Nextera测序文库（Illumina）进行测序。参考文献65中描述的生物信息流程用于识别数据集中大于1 kb的完整环状质粒。

挑战赛数据集

对于挑战赛，参与者获得了分别代表海洋和植物相关环境的两个宏基因组数据集的长读段和短读段序列，以及一个具有极高菌株多样性的'菌株疯狂'数据集。此外，还提供了一名危重患者的短读段临床宏基因组数据集。

十样本100 GB海洋数据集使用CAMISIM从深海环境的BIOM文件创建，使用了155个来自该环境的新测序海洋分离株基因组和622个来自MarRef（一个手工管理的完整测序海洋基因组数据库）的匹配分类来源基因组。其中303个（39%）——204个数据库基因组（31.9%）和99个新基因组（72.3%）——具有ANI≥95%的密切相关菌株。此外，还添加了200个新测序的环状元件，包括质粒和病毒。每个样本创建5千兆碱基(Gb)的双端Illumina短读段和Pacific Biosciences长读段（补充文本）。

100样本400 GB菌株疯狂数据集包含408个新测序基因组，其中97%（395个）有密切相关菌株。每个样本使用CAMISIM生成2 Gb双端短读段和长读段序列，使用与CAMI 1相同的参数和错误谱（补充文本）。

21样本315 GB植物相关数据集包含894个基因组。其中224个来自proGenomes陆地代表性基因组，216个来自拟南芥根际的新测序基因组，55个是与根际相关的真菌基因组，398个是质粒或环状元件，1个是拟南芥基因组。其中15.3%（137个）至少有一个密切相关的基因组。每个样本生成5 Gb双端短读段序列，以及2×5 Gb分别模拟Pacific Biosciences和Oxford Nanopore测序数据的长读段序列。注意，90%的宏基因组序列数据来自细菌基因组，9%来自真菌基因组序列，1%来自拟南芥。为评估单样本与共组装策略的组装质量，从八个密切相关基因组簇中选择了23个新基因组，并以预定丰度添加到某些样本中。对于所有三个数据集，我们为每个宏基因组样本和合并样本分别生成了金标准，包括短读段、长读段和混合读段的组装、基因组分箱和分类单元分箱分配以及分类学谱。

最后，提供了一份来自出血热患者血液样本的688 MB双端MiSeq宏基因组测序数据集。之前的样本分析揭示了与CCHFV基因组匹配的序列（NCBI taxid 1980519），随后通过PCR确认了病毒基因组的存在（循环阈值27.4）。由于原始样本的来源，CCHFV的致病性无法在临床上证明，且CCHFV previously已被证明可引起亚临床感染。然而，样本中未发现其他可能导致出血热的病原体的证据，使CCHFV致病成为症状最合理的解释。为创建真实的数据集和案例同时保护患者身份，临床案例描述来源于真实病史并以与致病因子一致的方式进行了修改。此外，映射到人类基因组的读段被从1000基因组数据集中相同基因组区域随机抽取的序列替换。要求挑战赛参与者识别致病病原体以及样本中存在的所有其他病原体。

挑战赛组织

第二轮CAMI挑战赛评估了宏基因组组装、基因组分箱、分类学分箱、分类学谱分析和诊断病原体预测的软件。与之前一样，从公共基因组创建了两个宏基因组“练习”基准数据集，并在挑战赛之前与真实情况一起提供，使参赛者熟悉数据类型和格式。这些包括一个49样本数据集（模拟人类微生物组数据）和一个64样本数据集（模拟小鼠肠道样本的分类学组成），分别有5 Gb长读段（Pacific Biosciences，可变长度平均3,000 bp）和5 Gb短读段（Illumina HiSeq2000，150 bp）双端读段序列。读段谱（读段长度和错误率）根据MBARC-26数据集的测序运行创建。2019年1月8日的NCBI RefSeq、nr/nt和分类学参考数据集合提供给参与者，用于挑战赛中基于参考的方法。为减少分类学分箱器最终使用预编译参考数据库导致的分类学差异，在评估期间使用NCBI的merged.dmp文件映射同义分类单元。

第二轮挑战赛于2019年1月16日开始（https://www.microbiome-cosi.org/cami/cami/cami2）。参与者注册下载挑战赛数据集，从那时到2021年1月共有332个团队注册。为可重复性，参与者可以提交包含完整工作流程的Docker容器、bioconda脚本或带有详细安装说明、指定所有参数设置和所用参考数据库的软件库。组装结果可针对短读段数据、长读段数据或两种数据类型组合提交。对于无法提交整个数据集跨样本组装的方法，可以提交数据集前十个样本的跨样本组装。参与者也可以为数据集的前五个样本分别提交单样本组装。菌株感知组装的性能标准规格见补充材料。组装挑战赛于2019年5月17日关闭。紧接着，提供了两个数据集的金标准和MEGAHIT组装。GSA包括合并宏基因组数据集中被一条短读段覆盖的所有参考基因组和环状元件序列。GSA分箱的分析使我们能够独立于组装质量评估分箱性能。我们通过与MEGAHIT组装上的分箱结果比较来评估组装质量的贡献。分析结果分别针对所有单个样本和整个数据集提交。分箱结果包括数据集每个样本的分析组装或读段的基因组或分类单元分箱分配。病原体检测挑战赛的结果包括所有病原体的预测和负责临床病例描述中所述症状的致病病原体。CAMI II挑战赛于2019年10月25日结束。随后，从2020年2月14日开始提供了另一轮植物相关数据的挑战赛（“CAMI II b”）。组装提交于2020年9月29日关闭，基因组和分类学分箱以及分析于2021年1月31日关闭。

四个挑战赛数据集共收到来自30个外部团队和CAMI开发者的76个程序的5,002份提交（补充表2）。用于生成基准数据集的所有基因组数据及其元数据在挑战赛期间保密，之后发布（10.4126/FRL01-006421672）。为支持无偏见评估，程序提交在门户网站中以匿名名称表示（仅提交者知晓），并在评估研讨会中使用另一组匿名名称进行评估和讨论，使得除数据分析团队（F.Meyer，Z.-L.D.，A.F.，A.S.）外的所有人都不知晓身份，程序身份仅在达成初步共识后揭示。

评估指标

组装

组装结果使用metaQUAST v.5.1.0rc和--unique-mapping标志进行评估。此标志允许每个重叠_contig仅映射到单个参考基因组位置。我们关注常用的组装指标，如基因组覆盖率、每100 kb不匹配数、重复率、NGA50和错误组装数。基因组覆盖率指定通过基于相似性映射后被组装重叠群覆盖的参考碱基百分比。每100 kb不匹配数指定重叠群-参考比对中不匹配碱基的数量。重复率定义为组装的比对碱基总数除以参考基因组的比对碱基总数。NGA50是衡量组装连续性的指标。对于每个参考基因组，所有比对的重叠_contig按大小排序。该基因组的NGA50定义为累积超过50%基因组覆盖率时的重叠_contig长度。如果基因组覆盖率未达到50%，则NGA50未定义。由于我们报告所有基因组的平均NGA50，对于基因组覆盖率低于50%的基因组将其设为0。最后，错误组装数描述包含超过1 kb间隙、包含超过1 kb插入或比对到两个或更多不同基因组的重叠_contig数量。除这些指标外，类似于参考文献18，我们确定了菌株召回率和菌株精确率以量化高质量菌株分辨率组装的存在。菌株召回率定义为所有真实基因组中恢复的高质量（>90%基因组覆盖率且<特定每100 kb不匹配数）基因组组装的比例。菌株精确率指定所有高基因组覆盖率（>90%）组装中低不匹配和高基因组覆盖率组装的比例。对于菌株疯狂数据集，所需基因组覆盖率设为75%，允许不匹配<0.5%，因为总体组装质量较低。

基因组分箱

对于基因组分箱，对于每个预测基因组分箱b，真阳性TPb是b中最丰富基因组g的碱基对数，假阳性FPb是b中属于g以外基因组的碱基对数，假阴性FNb是属于g但不在b中的碱基对数。纯度定义为：纯度b = TPb / (TPb + FPb)。平均纯度是所有预测基因组分箱集合B中分箱b纯度的简单平均值。完整性基于基因组g到其最丰富的基因组分箱b的映射定义：完整性gb = TPgb / (TPgb + FNgb)。平均完整性在样本中所有基因组上定义，包括那些在任何预测基因组分箱中都不是最丰富的基因组。设X为这类基因组的集合，则平均完整性 = Σ(b∈B)完整性gb / (|B| + |X|)。

作为另一个指标，我们考虑满足特定质量标准的预测基因组分箱数量。完整性>50%且污染<10%的分箱称为'中等或更高'质量分箱，完整性>90%且污染<5%的分箱称为高质量基因组分箱，类似于CheckM。

ARI的定义如参考文献36。兰德指数比较同一组项目的两种聚类。假设项目是不同序列的碱基对，属于同一基因组且被分到同一基因组分箱的碱基对被视为真阳性，属于不同基因组且被放入不同基因组分箱的碱基对被视为真阴性。兰德指数是真阳性和真阴性之和除以碱基对总数。ARI考虑了兰德指数可能偶然大于0的情况，标准化后结果范围在1（最佳，表示聚类完美匹配）和0（最差，表示不比偶然好）之间。由于分箱方法可能留下部分数据未分箱，而ARI不适用于仅部分分配的数据集，它仅对分箱部分计算，并与数据集的分箱碱基对百分比一起解释。

分类学分箱

对于分类学分箱，指标从超界或域到种的主要分类等级分别计算。每个分类分箱b（即分配给同一分类单元的序列组及其中的碱基对）的纯度和完整性通过将TPb设为分配给b的真实分类单元t的碱基对数，FPb为分配给b但属于其他分类单元的碱基对数，FNb为属于t但未分配给b的碱基对数来计算。某一分类等级的平均纯度是该等级所有预测分类分箱纯度的简单平均值。某一分类等级的平均完整性是所有预测分类分箱的完整性之和除以该等级金标准中的分类单元数|GS|。某一分类等级的准确性定义为：准确性 = Σ(b∈B)TPb / n，其中B是该等级的预测分类分箱集合，n是该等级金标准中的碱基对总数。

平均纯度、完整性和准确性也在每个分类等级去除1%最小分箱的过滤子集Bf上计算，分别记为平均纯度f、平均完整性f和准确性f。Bf通过将B中所有分箱按碱基对大小递增排序，过滤掉累计大小之和小于或等于B中所有分箱总大小1%的前几个分箱获得。

分类学谱分析

对于分类学谱分析，我们确定了分类单元鉴定中的纯度和完整性、L1范数和加权UniFrac作为丰度指标，以及使用Shannon均匀度指数的alpha多样性估计。

分类学谱的纯度和完整性衡量方法在某一分类等级确定样本中分类单元存在和不存在的能力，不考虑其相对丰度。设真阳性TP和假阳性FP为正确和错误检测到的分类单元数（即在某一特定样本和等级的金标准谱中存在或不存在的分类单元），假阴性FN为金标准谱中存在但方法未能检测到的分类单元数。纯度、完整性和F1分数的定义如上。

L1范数误差、Bray-Curtis距离和加权UniFrac误差衡量方法确定样本中分类单元相对丰度的能力。除UniFrac指标（与等级无关）外，这些在各个分类等级定义。设xt和x*t分别为某一等级样本中分类单元t的真实和预测相对丰度。L1范数给出xt和x*t之间的总误差，范围在0到2之间。Bray-Curtis距离是L1范数误差除以各等级所有丰度之和。Bray-Curtis距离范围在0到1之间。由于金标准通常包含100%数据的丰度，如果分析器也对100%数据做出预测，则等于L1范数误差的一半，否则更高。

加权UniFrac指标使用预测和实际丰度之间的差异，按分类树中的距离加权。范围在0（最佳）到16（最差）之间。值'16'的存在是因为NCBI分类学有八个主要分类等级（界、门、纲等）。因此，使用单位分支长度时，最差可能的UniFrac值为16：一个样本100%的丰度在不同于另一个样本的界中的情况，需要分别向上和向下遍历分类树八个等级。我们使用UniFrac距离的EMDUnifrac实现。在不同条件下保存的生物重复样本对之间可以找到0.22的平均加权UniFrac值（标准差0.16，最小0.01，最大0.43，中位数0.14），来自参考文献76中使用的数据，可在Qiita中获取，研究ID 10394。这些值作为该指标良好（0.22）到优秀（0.01）谱分析预测的基线。

Shannon均匀度指数对每个等级定义为：Shannon均匀度指数 = Σ(t) x*t × ln(x*t) / ln(m)，其中m是分类单元t的总数。指数范围从0到1，1表示完全均匀。由于多样性估计仅从预测谱计算，我们评估其与金标准指数的绝对差异进行比较。

汇总统计（所有软件类别）

对于汇总统计的计算，我们首先按每个类别（即组装、基因组分箱、分类学分箱和分类学谱分析）中每个数据集上每个指标的表现对所有软件结果提交进行评分。每个结果根据排名获得分数（所有方法中第一名得0分，第二名得1分，

以此类推）。一个软件提交对数据集多个样本的指标结果取平均进行排名。分类学分箱器和分析器按分类等级（从域到种）排名，分数为各分类等级排名之和。在所有指标上，这些分数之和作为软件结果提交在数据集上的总体汇总统计量（补充图1、8、10和12）。为进一步探索特定问题的加权指标组合，交互式HTML页面（补充结果）允许用户为各个指标选择自定义权重并可视化结果。

报告摘要

有关研究设计的更多信息可在本文链接的Nature Research报告摘要中获取。

数据可用性

基准测试挑战赛和示范数据集（供开发者提前熟悉数据类型和格式）可在PUBLISSO获取，DOI: https://doi.org/10.4126/FRL01-006425521（海洋、菌株疯狂、植物相关），https://doi.org/10.4126/FRL01-006421672（小鼠肠道）和10.4126/FRL01-006425518（人类），以及CAMI数据门户（https://data.cami-challenge.org/participate）。数据集包括金标准、基准数据创建基础上的组装基因组、NCBI分类学版本和NCBI RefSeq、nt和nr的参考序列集合（2019/01/08状态）。基准测试的软件输出可在Zenodo获取（https://zenodo.org/communities/cami/）。新测序和之前未发表基因组的原始测序数据可通过BioProject编号PRJEB50270、PRJEB50297、PRJEB50298、PRJEB50299、PRJEB43117和PRJEB37696获取。源数据随本文提供。

代码可用性

用于数据分析和图1-5的软件和脚本以及汇总结果可在https://github.com/CAMI-challenge/second_challenge_evaluation获取。补充表2指定了评估的程序、使用的参数和安装选项，包括软件库、Bioconda包配方、Docker镜像、Bioboxes和BioContainers。

方法参考文献

60. Nguyen, T. T. & Landfald, B. Polar front associated variation in prokaryotic community structure in Arctic shelf seafloor. Front. Microbiol. 6, 17 (2015).

61. Bankevich, A. et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19, 455-477 (2012).

62. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 25, 1043-1055 (2015).

63. Bremges, A., Fritz, A. & McHardy, A. C. CAMITAX: Taxon labels for microbial genomes. Gigascience 9, giz154 (2020).

64. Browne, P. D., Kot, W., J□rgensen, T. S. & Hansen, L. H. The mobilome: metagenomic analysis of circular plasmids, viruses, and other extrachromosomal elements. Methods Mol. Biol. 2075, 253-264 (2020).

65. Alanin, K. W. S. et al. An improved direct metamobilome approach increases the detection of larger-sized circular elements across kingdoms. Plasmid 115, 102576 (2021).

66. Klemetsen, T. et al. The MAR databases: development and implementation of databases specific for marine metagenomics. Nucleic Acids Res. 46, D692-D699 (2018).

67. Mende, D. R. et al. proGenomes2: an improved database for accurate and consistent habitat, taxonomic and functional annotations of prokaryotic genomes. Nucleic Acids Res. 48, D621-D625 (2020).

68. Durán, P. et al. Microbial interkingdom interactions in roots promote Arabidopsis survival. Cell 175, 973-983.e14 (2018).

69. Bodur, H., Akinci, E., Ascioglu, S., □ngürü, P. & Uyar, Y. Subclinical infections with Crimean-Congo hemorrhagic fever virus, Turkey. Emerg. Infect. Dis. 18, 640-642 (2012).

70. 1000 Genomes Project Consortium et al. A global reference for human genetic variation. Nature 526, 68-74 (2015).

71. Roy, U. et al. Distinct microbial communities trigger colitis development upon intestinal barrier damage via innate or adaptive immune cells. Cell Rep. 21, 994-1008 (2017).

72. Fritz, A., Lesker, T., Bremges, A., McHardy, A. CAMI 2 - Multisample Benchmark Dataset of Mouse Gut (PUBLISSO, 2020).

73. Singer, E. et al. Next generation sequencing data of a defined microbial mock community. Sci. Data 3, 160081 (2016).

74. Lozupone, C. & Knight, R. UniFrac: a new phylogenetic method for comparing microbial communities. Appl. Environ. Microbiol. 71, 8228-8235 (2005).

75. McClelland, J. & Koslicki, D. EMDUniFrac: exact linear time computation of the UniFrac metric and identification of differentially abundant organisms. J. Math. Biol. 77, 935-949 (2018).

76. Marotz, C. et al. Evaluation of the effect of storage methods on fecal, saliva, and skin microbiome composition. mSystems 6, e01329-20 (2021).

77. Gonzalez, A. et al. Qiita: rapid, web-enabled microbiome meta-analysis. Nat. Methods 15, 796-798 (2018).

致谢

我们感谢宏基因组学社区所有在公开研讨会上为项目提供意见和反馈的成员，并衷心感谢以下机构的资助：DZIF（项目编号TI 12.002_00；F.Meyer）、德国卓越集群RESIST（EXC 2155项目编号390874280；Z.-L.D.）和NFDI4Microbiota（项目编号460129525）。D.K.部分由美国国家科学基金会资助（资助号1664803）；A.G.由圣彼得堡州立大学资助（资助ID PURE 73023672）；D.A.、A.Korobeynikov、D.M.和S.N.由俄罗斯科学基金会资助（资助号19-14-00172）；C.T.B.和L.I.部分由Gordon and Betty Moore基金会数据驱动发现倡议资助（资助号GBMF4551）；R.C.和R.V.由ANR Inception（ANR-16-CONV-0005）和PRAIRIE（ANR-19-P3IA-0001）资助；S.D.K.由欧洲研究理事会(ERC)在欧洲地平线2020研究与创新计划下资助（ERC-COG-2018）；J.K.和E.R.R.由美国国家科学基金会资助（资助号1845890）；S.M.部分由美国国家科学基金会资助（资助号2041984）；V.R.M.由Tony Basten Fellowship和Sydney Medical School Foundation资助。G.L.R.和Z.Z.部分由美国国家科学基金会资助（资助号1936791和1919691）；M.T.由ERC在欧洲地平线2020研究与创新计划下资助（ERC-COG-2018）；S.Z.由上海市科学技术委员会（资助号2018SHZDZX01）和111项目（资助号B18015）资助；S. Hacquard由德国研究基金会(DFG)通过'2125 DECRyPT'优先计划资助；R.E.、E.Goltsman、Zho.W.和A.T.由美国能源部(DOE)生物和环境研究办公室在合同号DE-AC02-05CH11231下资助；S.S.由瑞士国家科学基金会资助（NCCR Microbiomes - 51NF40_180575）。本研究使用了国家能源研究科学计算中心的资源，该中心由美国能源部科学办公室在合同号DE-AC02-05CH11231下支持。美国DOE联合基因组研究所（DOE科学办公室用户设施）开展的工作在合同号DE-AC02-05CH11231下支持。

作者贡献

F. Meyer、A.F.、Z.-L.D.、D.K.、M.A.、D.A.、F.B.、D.B.、J.J.B.、C.T.B.、J.B.、A. Bulu□、B.C.、R.C.、P.T.L.C.C.、A.C.、R.E.、E.E.、E. Georganas、E. Goltsman、S. Hofmeyr、P.H.、L.I.、H.J.、S.D.K.、M.K.、A. Korobeynikov、J.K.、N.L.、C. Lemaitre、C.Li、A.L.、F.M.-M.、S.M.、V.R.M.、C.M.、P.M.、D.M.、D.R.M.、A.M.、N.N.、J.N.、S.N.、L.O.、L.P.、P.P.、V.C.P.、J.S.P.、S. Rasmussen、E.R.R.、K.R.、B.R.、G.L.R.、H.-J.R.、V.S.、N. Segata、E.S.、L.S.、F.S.、S.S.、A.T.、C.T.、M.T.、J.T.、G.U.、Zho Wang、Zi Wang、Zhe Wang、A.W.、K.Y.、R.Y.、G.Z.、Z.Z.、S.Z.、J.Z.和A.S.参与了挑战赛并创建了结果。P.W.D.、L.H.H.、T.S.J.、T.K.、A. Kola、E.M.R.、S.J.S.、N.P.W.、R.G.-O.、P.G.、S. Hacquard、S. H□u□ler、A. Khaledi、T.R.L.、F. Maechler、F. Mesney、S. Radutoiu、P.S.-L.、N. Smit和T.S.生成并贡献了数据。A.F.、A. Bremges、T.R.L.、A.S.和A.C.M.生成了基准数据集。F. Meyer、D.K.、A.G.、M.A.G.、L.I.、G.L.R.、Z.Z.和A.C.M.实现了基准测试指标。F. Meyer、A.F.、D.K.、A.S.、Z.-L.D.和A.C.M.进行了评估并在许多作者的意见下解释了结果。F. Meyer、A.F.、Z.-L.D.、D.K.、T.R.L.、A.G.、G.R.、F.B.、R.C.、P.W.D.、A.E.D.、R.E.、D.R.M.、A.M.、E.R.R.、B.R.、G.L.R.、H.-J.R.、S.S.、R.V.、Z.Z.、A. Bremges、A.S.和A.C.M.对挑战赛设计或评估提出了概念性意见。F. Meyer、A.C.M.、A.F.、D.K.和Z.-L.D.在许多作者的意见下撰写了论文。A.S.和A.C.M.在许多作者的意见下构思了本研究。

资助

开放获取资金由Helmholtz-Zentrum für Infektionsforschung GmbH (HZI)提供。

利益冲突

A.E.D.共同创立了Longas Technologies Pty Ltd（一家致力于合成长读段测序技术开发的公司），并受雇于Illumina Australia Pty Ltd。A.C.受雇于Google

LLC。L.I.受雇于10X Genomics。E.R.R.在Empress Therapeutics进行了实习。E. Georganas受雇于Intel Corporation。G.U.受雇于Amazon.com, Inc。其余作者声明无利益冲突。

补充信息

补充信息：在线版本包含可在https://doi.org/10.1038/s41592-022-01431-4获取的补充材料。通信和材料请求应发送给Alice Carolyn McHardy。

同行评审信息：Nature Methods感谢Nikos Kyrpides和其他匿名审稿人对本工作同行评审的贡献。主要处理编辑：Lin Tang，与Nature Methods团队协作。

作者单位

1. 感染研究计算生物学，赫尔姆霍兹感染研究中心，布伦瑞克，德国。
2. 布伦瑞克系统生物学综合中心(BRICS)，布伦瑞克工业大学，布伦瑞克，德国。
3. 德国感染研究中心(DZIF)，汉诺威-布伦瑞克分部，布伦瑞克，德国。
4. 卓越集群RESIST (EXC 2155)，汉诺威医学院，汉诺威，德国。
5. 宾夕法尼亚州立大学，州学院，宾夕法尼亚州，美国。
6. 赫尔姆霍兹感染研究中心，布伦瑞克，德国。
7. 圣彼得堡州立大学，圣彼得堡，俄罗斯。
8. 信息技术与电气工程系，苏黎世联邦理工学院，苏黎世，瑞士。
9. 算法生物技术中心，圣彼得堡州立大学，圣彼得堡，俄罗斯。
10. CIBIO系，特伦托大学，特伦托，意大利。
11. 新加坡基因组研究所，新加坡。
12. 南加州大学，洛杉矶，加利福尼亚州，美国。
13. 加州大学戴维斯分校，戴维斯，加利福尼亚州，美国。
14. 生物数据科学研究所，海因里希·海涅大学，杜塞尔多夫，德国。
15. 劳伦斯伯克利国家实验室，伯克利，加利福尼亚州，美国。
16. 加州大学伯克利分校，伯克利，加利福尼亚州，美国。
17. 巴斯德研究所，巴黎，法国。
18. 国家食品研究所，全球监测司，丹麦技术大学，灵比，丹麦。
19. 德雷塞尔大学，费城，宾夕法尼亚州，美国。
20. Google Inc.，费城，宾夕法尼亚州，美国。
21. 罗伯特·科赫研究所，柏林，德国。
22. 柏林技术与经济学院，柏林，德国。
23. 悉尼科技大学，悉尼，澳大利亚。
24. DOE联合基因组研究所，伯克利，加利福尼亚州，美国。
25. 劳伦斯伯克利国家实验室，伯克利，加利福尼亚州，美国。
26. 加州大学洛杉矶分校，洛杉矶，加利福尼亚州，美国。
27. Intel Corporation，圣克拉拉，加利福尼亚州，美国。
28. 生态与进化信号处理与信息学实验室，费城，宾夕法尼亚州，美国。
29. 哥本哈根大学，植物与环境科学系，腓特烈斯贝，丹麦。
30. 计算机科学学院，复旦大学，上海，中国。
31. BGI-深圳，深圳，中国。
32. 深圳人体共生微生物与健康研究重点实验室，BGI-深圳，深圳，中国。
33. 丹麦技术大学，诺和诺德基金会生物可持续发展中心，灵比，丹麦。
34. 奥胡斯大学，环境科学系，罗斯基勒，丹麦。
35. 细胞生理学与代谢系，医学院，日内瓦大学，日内瓦，瑞士。
36. 瑞士生物信息学研究所，日内瓦，瑞士。
37. 挪威北极大学，特罗姆瑟，挪威。
38. 夏里特医学院，柏林，德国。
39. 计算机科学与工程系，加州大学圣地亚哥分校，圣地亚哥，加利福尼亚州，美国。
40. 统计建模系，圣彼得堡州立大学，圣彼得堡，俄罗斯。
41. 威斯康星大学麦迪逊分校，麦迪逊，威斯康星州，美国。
42. 雷恩大学，Inria，CNRS，IRISA，雷恩，法国。
43. 里尔大学，CNRS，CRIStAL，里尔，法国。
44. 哈索·普拉特纳研究所，数字工程学院，波茨坦大学，波茨坦，德国。
45. 悉尼医学院，悉尼大学，悉尼，澳大利亚。
46. 先天免疫与感染性疾病中心，哈德逊医学研究所，克莱顿，澳大利亚。

47. 计算机科学系，Inria，里尔大学，CNRS，里尔，法国。
48. 阿姆斯特丹大学医学中心，阿姆斯特丹，荷兰。
49. 生物学系，微生物研究所和瑞士生物信息学研究所，苏黎世联邦理工学院，苏黎世，瑞士。
50. 结构与计算生物学单元，EMBL，海德堡，德国。
51. 新加坡基因组研究所，A*STAR，新加坡。
52. 新加坡国立大学，新加坡。
53. DTU Health Tech，Kongens，灵比，丹麦。
54. 基因信息学部，计算与统计基因组学分支，国家人类基因组研究所，国立卫生研究院，贝塞斯达，马里兰州，美国。
55. 弗吉尼亚大学，夏洛茨维尔，弗吉尼亚州，美国。
56. 诺和诺德基金会蛋白质研究中心，健康与医学科学学院，哥本哈根大学，哥本哈根，丹麦。
57. 生物信息学研究所，柏林自由大学，柏林，德国。
58. 生物信息学单元(MF1)，罗伯特·科赫研究所，柏林，德国。
59. 大数据生物发现中心，费城，宾夕法尼亚州，美国。
60. 佛罗里达理工大学，莱克兰，佛罗里达州，美国。
61. 定量与计算生物学系，南加州大学，洛杉矶，加利福尼亚州，美国。
62. 哥本哈根大学，哥本哈根，丹麦。
63. 不列颠哥伦比亚大学，温哥华，不列颠哥伦比亚省，加拿大。
64. 糖尿病中心，医学院，日内瓦大学，日内瓦，瑞士。
65. 能源、矿业与环境，加拿大国家研究委员会，蒙特利尔，魁北克省，加拿大。
66. Phase Genomics，西雅图，华盛顿州，美国。
67. 数学科学学院，复旦大学，上海，中国。
68. 能源部联合基因组研究所，伯克利，加利福尼亚州，美国。
69. 环境基因组学与系统生物学部，劳伦斯伯克利国家实验室，伯克利，加利福尼亚州，美国。
70. 自然科学学院，加州大学默塞德分校，默塞德，加利福尼亚州，美国。
71. 类脑智能科学与技术研究所，复旦大学，上海，中国。
72. 计算神经科学与类脑智能重点实验室(复旦大学)，教育部，上海，中国。
73. 马克斯·普朗克植物育种研究所，科隆，德国。
74. 奥胡斯大学，奥胡斯，丹麦。
75. 生物技术中心(CeBiTec)，比勒费尔德大学，比勒费尔德，德国。
76. 这些作者同等贡献：F. Meyer, A. Fritz。□电子邮件：amc14@helmholtz-hzi.de