微生物学报
Acta Microbiologica Sinica
2021, 61(9): 2921–2933
http://journals.im.ac.cn/actamicrocn
DOI: 10.13343/j.cnki.wsxb.20200761

Research Article 研究报告

组装断裂导致宏基因组来源的基因组污染度高估的评估与修正

李浩[1#]，杨东旭[1#]，温林冉[1]，郑伟[1]，郭峰[1,2,3*]

[1]厦门大学生命科学学院，福建 厦门 361102
[2]南方海洋科学与工程广东省实验室(珠海)，广东 珠海 510275
[3]资源微生物福建省高校重点实验室，福建 厦门 361102

摘要：【目的】识别并修正由断裂的标记基因引起的来自宏基因组测序组装的基因组污染度的高估。【方法】利用纯菌完整基因组构造的模拟数据来分析断裂基因对基因组质量评估的影响以及设定矫正参数，基于nr库的分类学注释结果来判定2个断裂标记基因(即断裂基因对)是否来自于同一标记基因，在剔除断裂冗余基因后重新计算污染度。【结果】基于纯菌完整基因组模拟打断数据的结果表明基因组片段化程度越高，基因组的污染度越高，并且该现象在分箱获得的微生物基因组草图中也有体现。我们设计的矫正流程能将纯菌模拟打断数据的污染度纠正到完整基因组的水平。在对760个肠道和土壤宏基因组来源的污染度大于0的基因组草图进行矫正后，接近半数基因组的污染度降低，其中43个基因组的污染度降至0。【结论】我们的流程可以在一定程度上矫正由断裂基因引起的基因组污染度的高估，提高分箱基因组草图的可利用率，并可应用于需求日益增加的宏基因组来源的基因组质量评估中。

关键词：宏基因组组装基因组，基因组质量，CheckM，污染度

随着近年来高通量测序技术的普及和宏基因组学的发展，从宏基因组测序数据中通过分箱(binning)获取海量具有中等或高质量的微生物基因组草图(metagenome-assembled genome，MAG)的技术已经日益成熟[1–4]。这些来源于不同环境或宿主相关样本的MAGs[5–7]，极大地丰富了生命之树的谱系，拓宽了人们对微生物代谢和功能多样性的认知，为挖掘那些尚未被人们所熟知的微生物资源宝库提供了便利[8–11]。然而，大量积累的MAGs数据也要求研究人员必须开发更加准确的

基金项目：国家自然科学基金(31670492，31500100)
[#]并列第一作者。
[*]**通信作者**。Tel/Fax：+86-592-2880330；E-mail：fguo.bio@xmu.edu.cn
收稿日期：2020-12-12；**修回日期**：2021-02-20；**网络出版日期**：2021-07-14

基因组质量评估方法[12]。相比于早期采用的基因组 N50 和覆盖度等指标来评估纯菌完整基因组的质量[13−14]，宏基因组来源的基因组草图中由于通常含有被错误划分的 contigs 或者 scaffolds 且绝大多数都是不完整的[4,15−16]，仅靠上述指标不足以描述基因组质量，需要有效评估 MAGs 的完整度和污染度[17]。其中，完整度用于描述基因组的完整程度，而污染度则是指基因组混杂其他来源序列的潜在程度。目前，相关评估主要是依据 MAGs 的必需单拷贝基因来实现的[5,15,18−19]。其中，污染度对基于基因组信息预测 MAGs 的相应功能时具有潜在假阳性的指示作用，因此往往在评价基因组质量时具有更大的权重。例如，对基因组质量评估中通常采用完整度减去 5 倍污染度作为参考质量值[20]。

Parks 等基于谱系特异的单拷贝保守标记基因的策略开发了 CheckM 软件[17]。简单来说，CheckM 基于 43 个与系统发生历史具有高度一致性的单拷贝基因对 5656 个高质量的参考基因组构建系统发生树，然后以系统发生树为依据，将每个分类学单元节点内在超过 95%的基因组中都出现的单拷贝基因定义为谱系特异(clade-specific)的标记基因。一方面，该程序根据基因组中相应的谱系特异标记基因的数量来计算其完整度和污染度，较大地提高了在较低分类阶元上的谱系标记基因数量；另一方面，该方法也充分考虑了在基因组中距离相近的多个单拷贝标记基因相对于距离较远的单拷贝标记基因的评价权重不同，设定了标记基因集(mark gene set)的概念。该方法具有计算速度较快、结果相对准确且适用范围较广的优点，因此其被应用于绝大多数微生物基因组质量评估中[21−22]。

然而，所有基于标记基因进行污染度分析理论上均可能由于基因组组装断裂引起单一基因被重复计算的情况，如图 1-A 所示。鉴于目前测序技术、组装策略和测序深度的限制，从宏基因组中获取的 MAGs 通常的断裂程度较高(一般为数十至上千个 contigs)[23]。这里我们将某个基因由于组装问题被分割到 2 个 contigs 上的现象称为基因断裂现象(某些情况下基因断裂后可能产生超过 2 个片段，但一般获得 MAGs 的过程中会过滤长度小于 1–2 kb 的 contig，因而出现超过 2 个片段的概率较低)，而这 2 个 contigs 上属于该基因的部分称为一对断裂基因。由于绝大多数基因组草图由较多 contigs 组成，原核生物基因组的高编码率决定了一般情况下 MAGs 中有较多的基因断裂现象。当被用于评估污染度的标记基因成为一对断裂基因时，则有可能会被重复计算而高估污染度。

在本研究中，我们首先采用纯菌的完整基因组数据确定了上述假设，即基因断裂现象会引起对基因组的污染度高估，并评估了这种高估与基因组的 contigs 数量、断裂基因数和标志基因数量的关联性。在此基础上，我们基于对断裂基因同源性的可靠判定，较为准确地实现了污染度的修正。相关分析方法将可以应用于对日益增加的海量 MAGs 的准确质量评估。

1 材料和方法

1.1 实验材料

本研究共涉及 3 个基因组数据集，其中数据集 I 是从 NCBI 中获取的 24 个来自纯培养菌株的完整原核微生物代表基因组(Taxon ID-物种：890402-*Bifidobacterium longum*，336982-

Mycobacterium tuberculosis，272559-*Bacteroides fragilis*，553174-*Prevotella melaninogenica*，761193-*Runella slithyformis*，290315-*Chlorobium limicola*，765962-*Helicobacter pylori*，383372-*Roseiflexus castenholzii*，63737-*Nostoc punctiforme*，167555-*Prochlorococcus marinus*，762633-*Thermus thermophilus*，243231-*Geobacter sulfurreducens*，289380-*Clostridium perfringens*，655816-*Bacillus marinus*，272635-*Mycoplasma pulmonis*，1278073-*Myxococcus stipitatus*，174633-*Kuenenia stuttgartiensis*，224911-*Bradyrhizobium diazoefficiens*，1105095-*Rickettsia prowazekii*，392499-*Rhizorhabdus wittichii*，574521-*Escherichia coli*，381754-*Pseudomonas aeruginosa*，521045-*Kosmotoga olearia*，452637-*Opitutus terrae*)；数据集 II 来自 Almeida 等从人肠道微生物宏基因组数据集中获得的 1949 个代表 MAGs[24]。基因组数据集 III 来自本实验室从土壤样本的宏基因组中分箱获得的共 310 个 MAGs。其中来自数据集 II 和 III 的基因组均达到中等质量(完整度>50%，污染度<10%)水平[25]。

1.2 基因组片段化及质量评估

为了模拟MAGs的断裂情况并评估组装断裂对基因组污染度的影响，我们将数据集 I 来源的完整基因组进行随机打断，要求输出的断裂基因组最大片段长度不超过该基因组长度的 1/10，最小片段大于 1000 nt，最终生成 contigs 数量分别为 10、20、50、100、200 和 500 的模拟断裂基因组，每个基因组在每个数量下重复随机打断 10 次。该过程在 Python 中基于 Biopython v1.78 实现[26]。本文所有的基因组质量评估均由 CheckM v1.0.11 软件来实现[17]，根据需要采用 CheckM 软件的 lineage 和 taxonomy 两种工作模式：在 lineage 工作模式下，基因组首先会根据 43 个单拷贝基因建树的结果被分配到相应的谱系，再提取出单拷贝基因与该谱系对应的标记基因进行比对来计算完整度和污染度。在 taxonomy 工作模式下，基因组则需要通过人为指定而非通过建树来确定其分类学阶元，后续评估流程与 lineage 模式一致。本研究在 taxonomy 模式下的基因组质量评估均指定到细菌域。基因组质量值的计算方法为完整度减去 5 倍的污染度[20]。

1.3 矫正标记基因断裂对基因组草图污染度的影响

如图 1-B 所示，矫正基因断裂导致的基因组污染度的高估，其本质是识别断裂基因对，而后在统计污染度时避免重复计算。针对识别来源于同一基因的断裂基因对，我们按照以下原则进行。

1.3.1 断裂基因的判定原则：首先，我们将每个 MAG 中被 CheckM 软件鉴定为同一个单拷贝标记基因的多个污染基因提取出来，对于其中任意 2 个基因，如果是由于组装断裂而导致的断裂基因对，那么它们需满足以下条件：(a) 该基因对首先应均是不完整的基因，且应位于 contig 末端(即组装断裂)，这一点由 Prodigal 软件的输出信息进行判定[27]；(b) 由于断裂基因对的来源相同，该基因对编码的氨基酸序列可以通过 BLASTp 比对到 nr 数据库(下载于 2018 年 12 月)中同一条参考序列上[28]，且它们与同一条参考序列的比对相似度差异应较小；(c) 在 nr 数据库参考序列比对结果中，该基因对的 2 条基因与数据库参考序列的比对区域不能存在较大片段的重叠。这里允许小片段的重叠，原因在于基于

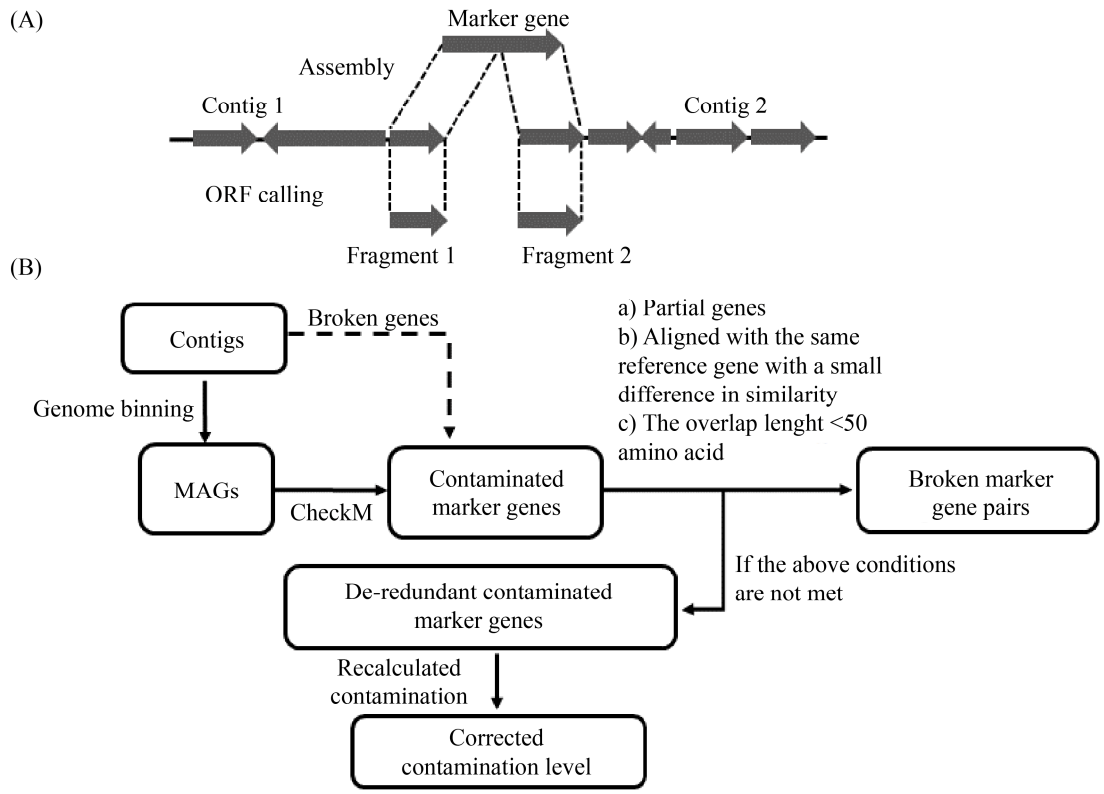

图 1.　断裂基因对产生原因(A)及识别并矫正由断裂基因引起的污染度高估的工作流程示意图(B)

Figure 1.　Causes of gene fragmentation (A) and the workflow of identification and correction the overestimated contamination caused by gene fragments (B).

德布鲁因图策略的组装在形成岔路必须断裂时，重叠的末端序列可以保留于不同 contig 中。我们规定两条断裂基因的重合度不超过 50 个氨基酸(即典型二代测序的 150 nt 序列读长)。

1.3.2　判定参数的优化：针对判定原则(b)的 3 个相关参数，即有效断裂基因长度阈值、断裂基因与 nr 数据库参考序列的相似度阈值，以及断裂基因对与 nr 数据库参考序列相似度差值的阈值，我们利用数据集 I 的随机打断模拟数据来优化矫正参数。这里与 nr 库的比对输出结果中我们排除了数据集 I 本身作为输出参考序列的情况。前 2 个参数的优化思路如下：在不同的参数设定下，如果由于打断造成的断裂基因对被正确判定为来源同一个基因，则被视为正确结果，如果未被正确判定则被视作错误结果。进行一系列的参数调试，选择出错误结果比例最低的合理参数。最终选定的有效断裂基因长度阈值为 20 个氨基酸，断裂基因与数据库参考基因的相似度阈值为 50% (同时要求覆盖度>80%)。对于第 3 个参数，即相似度差值的阈值的设定，我们统计了所有被正确检测为断裂基因的 2 个片段，与数据库比对上的同一条参考序列(相似度大于 50%，覆盖度大于 80%)的相似度差值，计算其均值和标准差，将其均值加 3 倍标准差的结果作为相似度差异的阈值，实际选定的阈值为 18%。

1.4 数据统计及可视化

本文的数据分析均在 R3.6.3 中实现[29]，数据可视化基于 ggplot2 包实现[30]。

1.5 代码及数据共享

本文的矫正流程已上传至 https://github.com/yang-dongxu/MAG_contam_corrector。

2 结果和分析

2.1 基因断裂对基因组污染度的高估

为了评估基因断裂对基因组质量的影响，我们首先将数据集 I 中 24 个纯菌完整基因组随机打断成 10、20、50、100、200 和 500 个 contigs 来构造模拟数据，并且在 CheckM 的 lineage 和 taxonomy 的工作模式下进行质量评估。结果显示无论在哪种工作模式下，基因打断都会导致基因组的完整度轻微下降(<1%，结果未显示)。对污染度的计算结果与我们的假设一致，即基因组断裂后的确会导致 CheckM 对污染度的高估，且污染度伴随着打断片段数量增加而上升，特别是 contig 数量大于 100 后上升趋势明显(图 2)。在 lineage 工作模式下，纯菌完整基因组的污染度均值为 0.52%，被打断为 200 和 500 个 contigs 后，污染度均值分别上升至 1.48%和 2.82%。在 taxonomy 模式下也有一致的结果，即污染度从完整基因组的 0.466%分别上升到 200 段和 500 段的 1.18%和 2.17%。其中，污染度提升最大的基因组是 *Mycoplasma pulmonis*，污染度从完整基因组的 1.92%，上升到 200 和 500 段的 5.15%和 7.91%，推测可能与该基因组的大小有关(<1 M，所有检测基因组中最小)。此外，图 2 还给出了随着 contig 数量增加检出的冗余基因数量增加的情况，其趋势基本与污染度上升相吻合。综合以上结果，我们认为标记基因的断裂会导致 CheckM 高估基因组的污染度，并且污染度随基因组片段化程度的升高而增加。

在评估过程中我们观察到大多数基因组在 lineage 模式下的污染度上升趋势普遍快于 taxonomy 模式(图 2)。值得指出二者的区别是：在 taxonomy 模式中我们只指定在细菌域的水平下进行的评估，而在 lineage 模式下基因组可以被鉴定到较低的分类阶元(属或者种水平)。较低的分类阶元往往包含较多的谱系特异标记基因，相应的标记基因被打断的概率也会增大。taxonomy 模式下指定细菌域的标记基因数量是 104 个，而 24 个基因组在各自的 lineage 模式下的标记基因数量为 104–1200 (中位数 462)。另外，需要注意的是有较多完整基因组(12 个，lineage 模式)的污染度不为 0，表明了这些基因组中本身就含有冗余的标记基因，暗示了 CheckM 对其标记基因的设定是不完善的，这可能是由 CheckM 在特定分类单元的参考数据集不全引起的。

2.2 MAGs 污染度的影响因素

为了探究 CheckM 对基因组污染度估计的主要影响因素，我们针对基因组数据集 II 计算了 MAGs 污染度与 contigs 数量、标记基因数量、标记基因集和分类阶元的相关性。结果显示 MAGs 的污染度与 contigs 的数量间存在显著的正相关关系(图 3-A，皮尔森相关系数=0.51，P=5.31e^{-186})，即存在基因组碎片化程度越高其污染度越高的趋势。而污染度与标记基因数目和标志基因集数目之间未观察到相关关系(图 3-B、C)。对 lineage 模式下 CheckM 对 MAGs 鉴定到不同分类阶元对污染度的影响进行分析，可以发现可鉴定到种水平(较低的分类阶元)基因组的污染度显著地高于

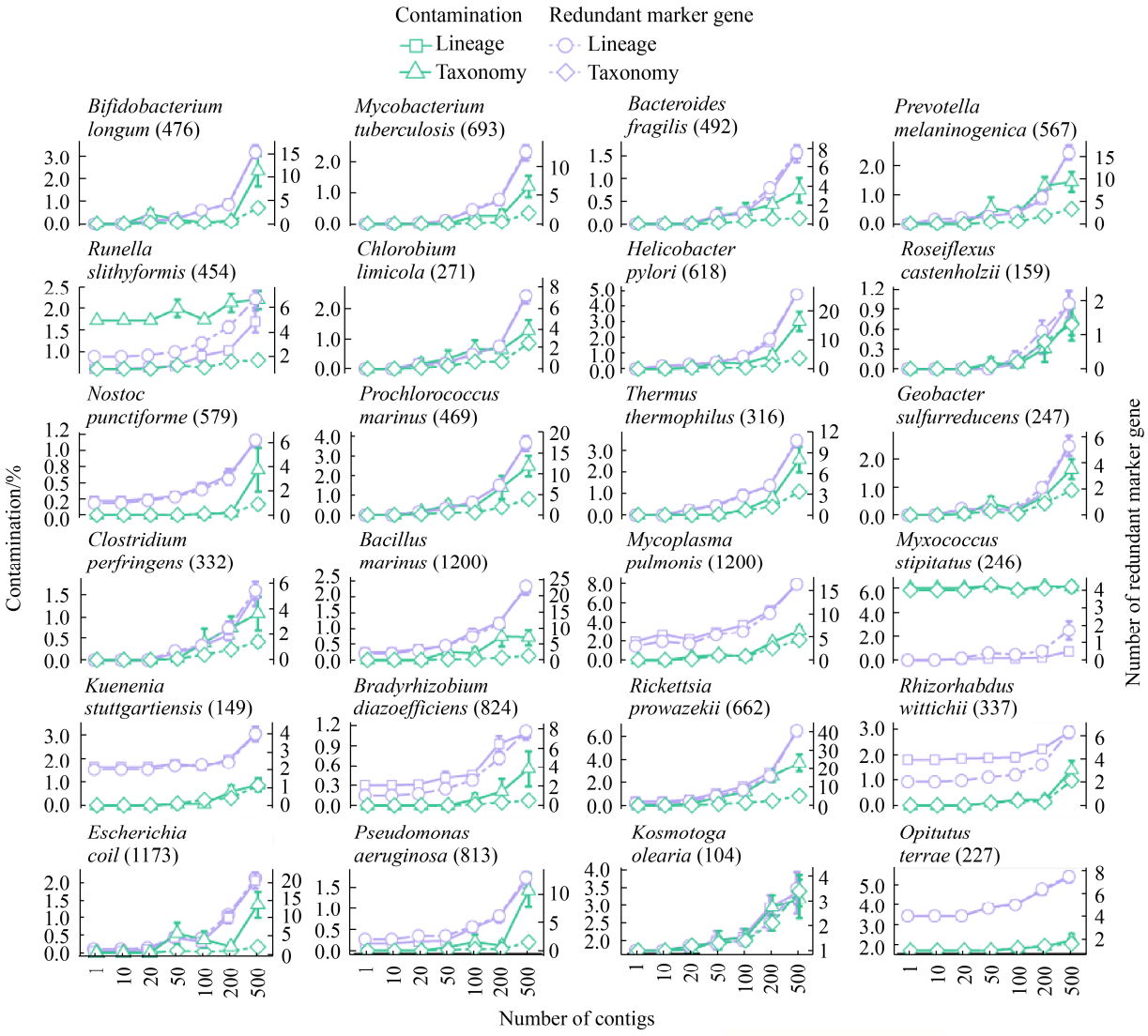

图 2. 基于模拟随机打断纯菌完整基因组的污染度及冗余标记基因数目分析

Figure 2. Contamination level and number of redundant marker genes of the simulated genomes. The number in brackets represents the number of marker genes under the lineage workflow of CheckM.

其他分类阶元(图 3-D)。这组结果表明对于真实的 MAGs 而言，其污染度与 contigs 数量也体现出了正相关，暗示真实 MAG 的质量值评估也广泛存在由断裂基因引起的高估。此外，尽管未发现MAGs的污染度与标记基因或标记基因集数量间存在显著正相关，但联系图 2 和图 3-D，我们推测当用于计算污染的标记基因的数量较高时可能会引起 MAG 的污染度高估。

2.3 对纯菌完整基因组模拟数据的修正

前面分析明确了断裂基因可导致基因组污染度被高估，我们构建了相应的流程来识别断裂基因并矫正由其引起的污染度高估(详细流程见材料和方法)。我们将该流程对基因组数据集 I 中被打断为 200 和 500 个片段的基因组(n=480)进行了测试，结果见图 4。经矫正后，所有基因组的污染度都有下降，其中 87.30%的基因组污染度回到

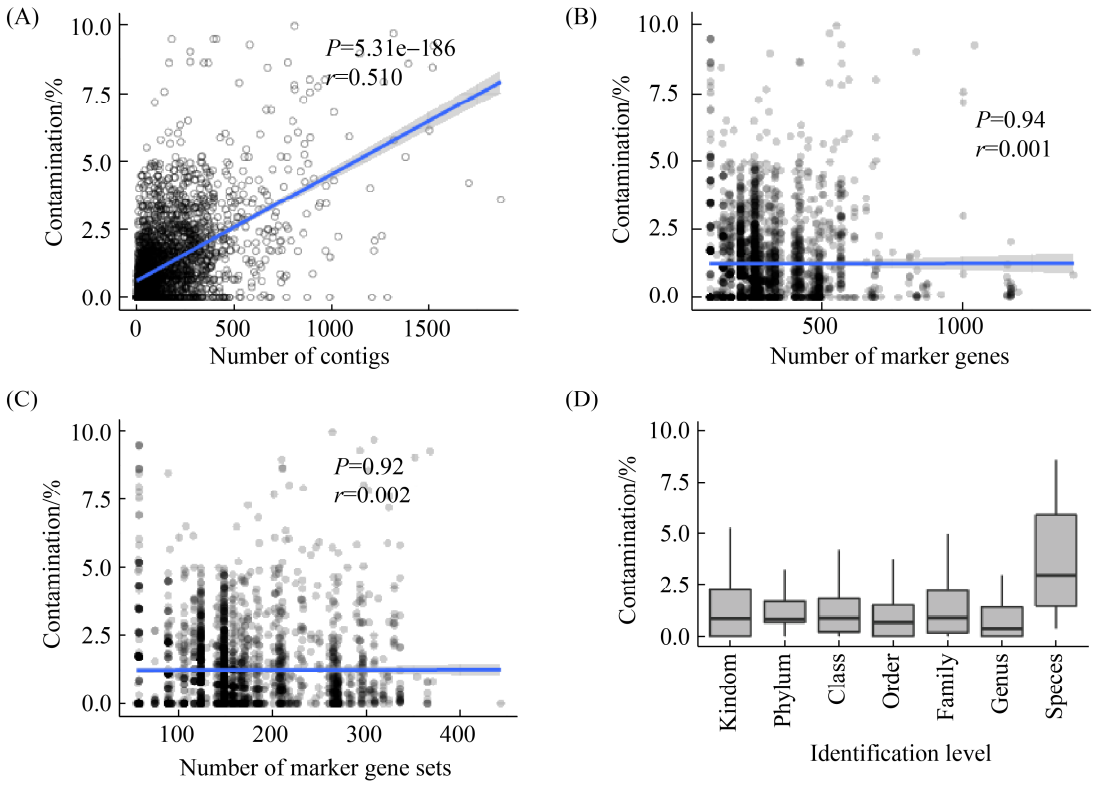

图 3. MAGs 污染度的影响因素

Figure 3. Factors affecting the contamination of MAGs. The Pearson correlation between MAGs contamination and the number of contigs (A), marker genes (B) and marker gene sets (C); D: The comparison of contamination at different identification level.

了完整基因组的水平，这表明我们的流程能很好地识别断裂基因并矫正其引起的污染度的高估。此外，有 10.83%的基因组的污染度并未被矫正到完整基因组水平，分析后发现这是由断裂基因被错误地注释为其他标记基因进而被错误统计造成的。有趣的是另有 9 个(1.88%)基因组污染度出现矫正后比完整基因组更低的情况，分析发现该现象集中于 1 个基因组中(8 次)，其原因为在原完整基因组中的冗余标记基因打断后由于片段过短不能被 CheckM 识别所致。总之，基于完整基因组模拟数据的结果，我们认为这一矫正方法效果良好，可尝试将其应用到宏基因组来源的数据集上。

2.4 宏基因组来源基因组污染度的矫正

我们从数据集 II 中随机抽取了约三分之一的样本，并结合数据集III共 937 个 MAGs 进行了断裂基因检测及矫正，得到的结果如图 5 所示。经矫正后，MAGs 的污染度与 contig 片段数的拟合直线的斜率下降，且其皮尔森相关性系数从 0.46 下降至 0.39 (图 5-A)，表明由断裂基因引起的基因组污染度的高估在一定程度上被修正，尽管修正率未知。具体而言，在 760 个污染度大于 0 的 MAGs 中，有 42.76%的 MAGs 的污染度发生了修正后下降，尽管大多数 MAGs 污染度的下降值都在 1%以内(图 5-B)，但有 43 个 MAGs 的污染

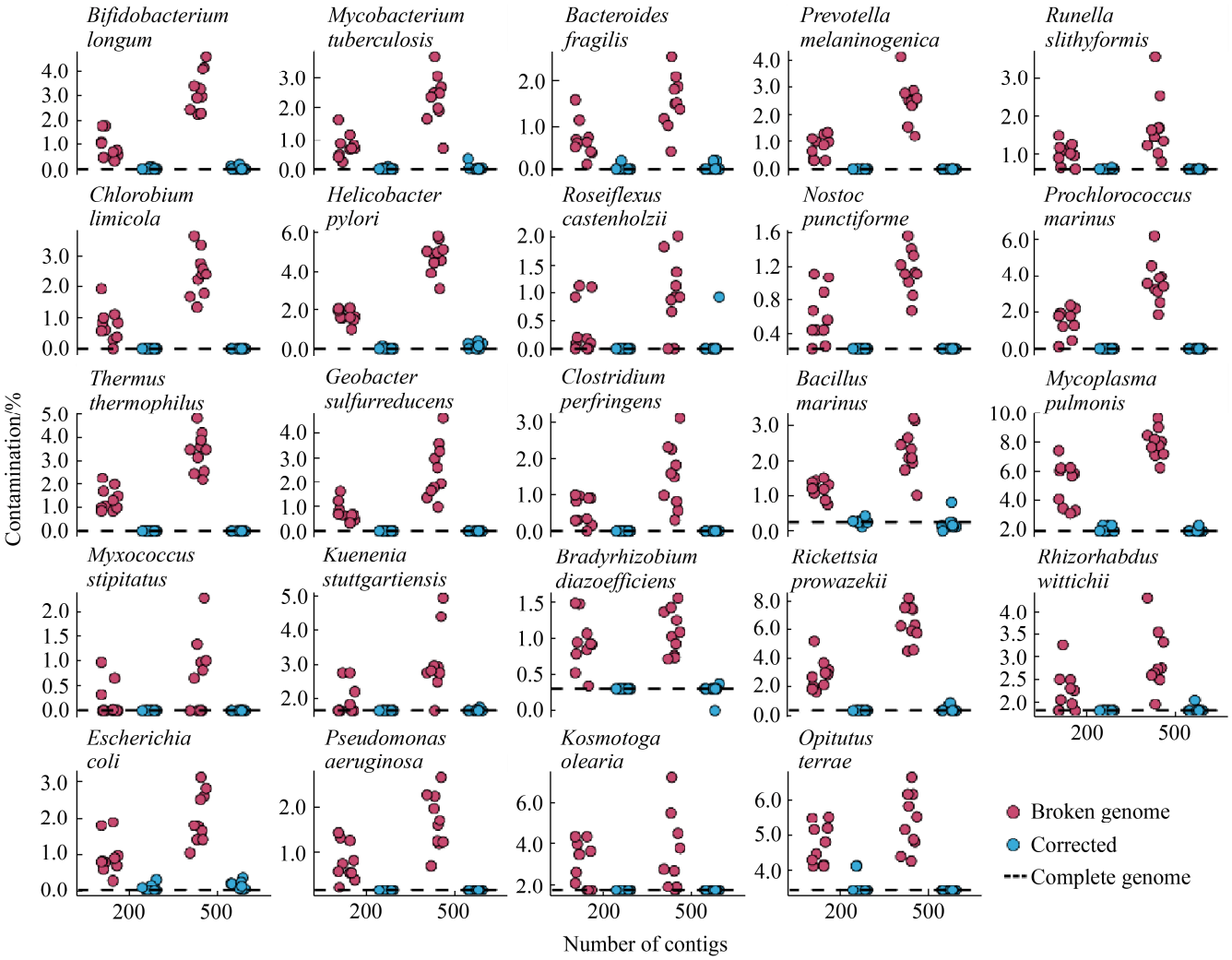

图 4. 基于纯菌完整基因组模拟数据的污染度矫正结果评估

Figure 4. Evaluation of corrected contamination based on the simulated genomes. The genome quality is assessed under the lineage workflow of CheckM.

度降至 0。我们观察到原质量值(完整度–5×污染度)小于 50 的 34 个 MAGs 中有 26.47%被提高到超过 50 的水平(图 5-C)，表明该流程可以有效提高 MAG 数据利用率。此外，我们还分析了 contigs 数量、标记基因数量与标记基因集数量对污染度矫正值(即下降数值)的贡献，在剔除没有矫正效果的 MAGs 后发现，contigs 数量与矫正值大小显著相关(图 5-D)。与图 3 结果类似的是，我们并未发现矫正值与标记基因数量和标记基因集的数量间显著的相关关系(图 5-E、F)。这组结果说明我们设计的方法可以在一定程度上矫正由基因断裂引起的 MAGs 污染度的高估，矫正值大小与 contigs 数量有关，而与标记基因和标记基因集的数量相关性不强。

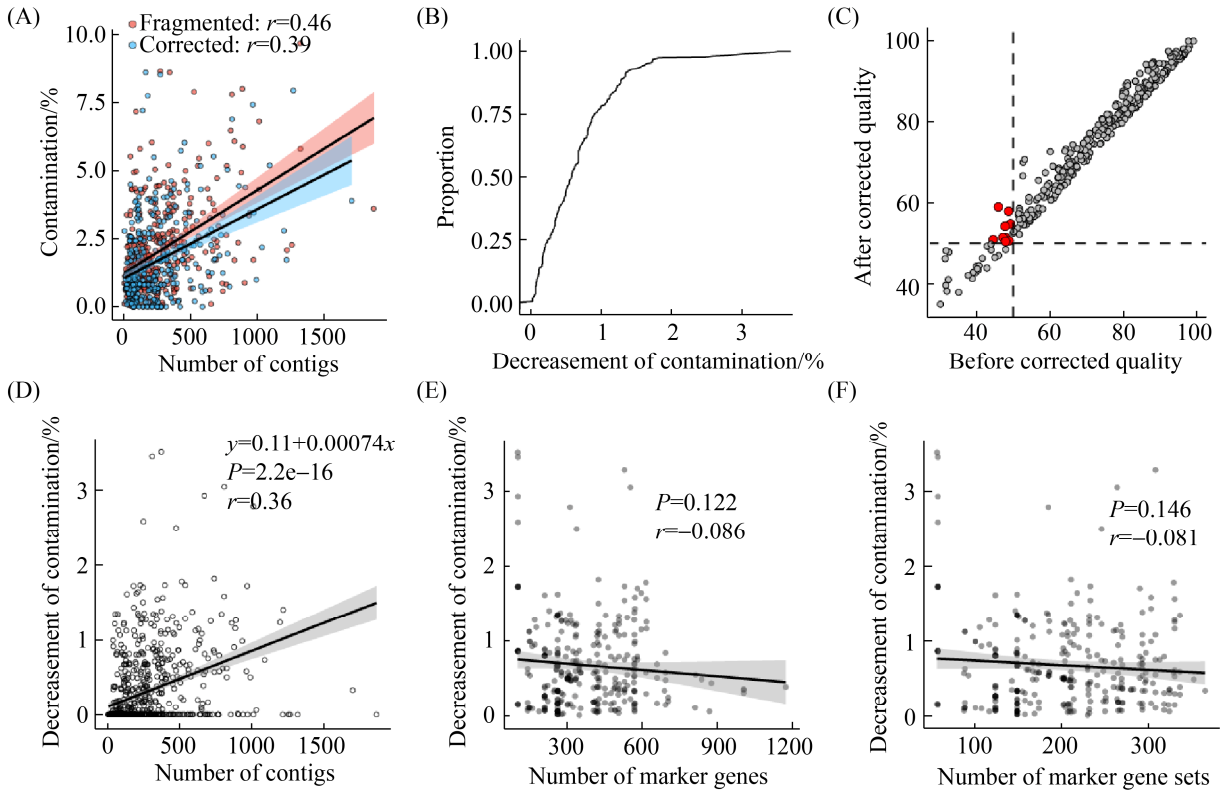

图 5. 基于 MAGs 数据的污染度修正结果分析

Figure 5. Assessment of contamination correction for MAGs. The genome quality is assessed under the lineage workflow of CheckM. A: the Pearson correlation between contamination and the number of contig before and after correction; B: the distribution of corrected contamination; C: changes of quality score before and after correction. The red dots are indicated the genomes with original quality score<50 that are corrected larger than 50. The Pearson correlation between contamination changes and number of contigs (D), marker genes (E) and marker gene sets (F).

3 讨论

基因组质量评估是基因组学的一个重要组成部分，可靠的评估是 MAG 下游分析的前提，而其中污染度一般具有更为重要的参考价值[20]。不够精准的基因组分箱过程是MAGs污染的主要来源[31]，采用 DAS Tool 和 Binning_refiner 等流程化工具可以避免 contig 在不同 MAGs 间的重复分配[32−33]，基于对参考基因组比对和序列组成特征分析的 ProDeGe 工具可自动识别 MAGs 中污染的 contigs[34]，从而降低潜在的污染度。与上述算法工具直接去除污染序列不同，本研究所针对的问题是污染度的数值高估，并不去除污染序列。由断裂标记基因引发的污染度数值高估引起的主要问题是对 MAGs 的质量误判，以及由于误判造成的数据损失。我们基于对模拟数据和真实 MAGs 的分析结果证实了由组装导致的基因断裂会引起 CheckM 对基因组污染度的高估，并且被高估的程度与 contigs 数量有显著正相关。在此基础上，我们采用 nr 库对污染基因进行分类学注释

并识别潜在的断裂基因，实现了对污染度高估的矫正，该方法在模拟打断数据和 MAG 数据中均有良好的表现。值得指出的是，不仅限于 CheckM，理论上所有基于标记基因进行的基因组完整度和污染度评价均会受到断裂基因的影响。

本研究发现 contig 数量与 MAGs 污染度以及其污染度被矫正程度均呈现显著正相关(图 2，3-A，5-A)，符合理论推测，也提示我们改进组装减少 contig 数量是降低污染度高估的直接途径。同时，我们发现标记基因(集)数量与 MAGs 污染度以及其污染度被矫正程度无相关性(图 3-B，3-C，5-E，5-F)。这一现象可被解释为，尽管标记基因数量与断裂基因出现的概率理论上相关，标记基因数量作为计算污染度的分母，其增加抵消了作为分子的断裂标记基因增加引起的污染度高估。另外，采用 Lineage 模式以及能够在 Lineage 模式下鉴定到物种水平的 MAGs，统计上体现其污染度往往更容易被高估(图-2，3-D)，其原因尚未可知，我们推测其原因之一是对进化相对保守的标记基因判定，随着分类阶元的下降(从域到种)，可参考的基因组越少，趋向于更多误判，即有更多可以是多拷贝的基因被认定为单拷贝，从而引起了污染度高估。这也提示我们应根据情况考虑 CheckM 的工作模式。

本文的分析方法也存在一定的局限性和改进空间。一方面，本文的分析是建立在组装导致基因断裂的基础上，但可能存在将来自近缘物种或同种不同菌株的污染基因判定为断裂基因的情况[35]。我们并未对这一情况的影响进行评估，原因在于一般认为近缘物种或同种不同菌株在 MAG 的分箱中成功率低[15]，但是该问题仍需要被进一步证实。另一方面，我们所采用的矫正方案需要使用 BLASTp 对 nr 库进行注释，该流程需要耗费较大的计算资源。基于我们分析平台的配置，对 480 个基因组模拟数据的矫正需要在 48 线程的计算机上运行约 84 h。我们后期考虑采用 DIAMOND 来代替 BLASTp 以提高计算速度[36]。此外，还可以通过从 nr 库中抽取相应标记基因集进行数据库缩减，以达到减少计算消耗的目的。不过，当矫正分类地位较为新颖的基因组时，可能存在数据库中没有合适的参考基因的情况，进而影响对断裂基因的判定。总之，我们将进一步完善这套流程对 MAGs 的质量评估。

参 考 文 献

[1] Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, Schulz F, Jarett J, Rivers AR, Eloe-Fadrosh EA, Tringe SG, Ivanova NN, Copeland A, Clum A, Becraft ED, Malmstrom RR, Birren B, Podar M, Bork P, Weinstock GM, Garrity GM, Dodsworth JA, Yooseph S, Sutton G, Glöckner FO, Gilbert JA, Nelson WC, Hallam SJ, Jungbluth SP, Ettema TJG, Tighe S, Konstantinidis KT, Liu WT, Baker BJ, Rattei T, Eisen JA, Hedlund B, McMahon KD, Fierer N, Knight R, Finn R, Cochrane G, Karsch-Mizrachi I, Tyson GW, Rinke C, Lapidus A, Meyer F, Yilmaz P, Parks DH, Murat Eren A, Schriml L, Banfield JF, Hugenholtz P, Woyke T. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and Archaea. *Nature Biotechnology*, 2017, 35(8): 725–731.

[2] Liu YX, Qin Y, Guo XX, Bai Y. Methods and applications for microbiome data analysis. *Hereditas*, 2019, 41(9): 845–862. (in Chinese)
刘永鑫, 秦媛, 郭晓璇, 白洋. 微生物组数据分析方法与应用. 遗传, 2019, 41(9): 845–862.

[3] Xu YK, Ma Y, Hu XQ, Wang J. Analysis of prospective microbiology research using third-generation sequencing

technology. *Biodiversity Science*, 2019, 27(5): 534–542. (in Chinese)
许亚昆, 马越, 胡小茜, 王军. 基于三代测序技术的微生物组学研究进展. 生物多样性, 2019, 27(5): 534–542.

[4] Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology*, 2017, 35(9): 833–844.

[5] Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics: Oxford, England, 2015, 31(19): 3210–3212.

[6] Zou YQ, Xue WB, Luo GW, Deng ZQ, Qin PP, Guo RJ, Sun HP, Xia Y, Liang SS, Dai Y, Wan DW, Jiang RR, Su LL, Feng Q, Jie ZY, Guo TK, Xia ZK, Liu C, Yu JH, Lin YX, Tang SM, Huo GC, Xu X, Hou Y, Liu X, Wang J, Yang HM, Kristiansen K, Li JH, Jia HJ, Xiao L. 1, 520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. *Nature Biotechnology*, 2019, 37(2): 179–185.

[7] Pasolli E, Asnicar F, Manara S, Zolfo M, Karcher N, Armanini F, Beghini F, Manghi P, Tett A, Ghensi P, Collado MC, Rice BL, DuLong C, Morgan XC, Golden CD, Quince C, Huttenhower C, Segata N. Extensive unexplored human microbiome diversity revealed by over 150, 000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell*, 2019, 176(3): 649–662.e20.

[8] Manara S, Asnicar F, Beghini F, Bazzani D, Cumbo F, Zolfo M, Nigro E, Karcher N, Manghi P, Metzger MI, Pasolli E, Segata N. Microbial genomes from non-human primate gut metagenomes expand the primate-associated bacterial tree of life with over 1000 novel species. *Genome Biology*, 2019, 20(1): 299.

[9] Nayfach S, Roux S, Seshadri R, Udwary D, Varghese N, Schulz F, Wu D, Paez-Espino D, Chen I-M, Huntemann M. A genomic catalog of Earth's microbiomes. *Nature Biotechnology*, 2020: 1–11.

[10] Xie JP, Han YB, Liu G, Bai LQ. Research advances on microbial genetics in China in 2015. *Hereditas*, 2016, 38(9): 765–790. (in Chinese)
谢建平, 韩玉波, 刘钢, 白林泉. 2015 年中国微生物遗传学研究领域若干重要进展. 遗传, 2016, 38(9): 765–790.

[11] Ma JC, Zhao FQ, Su XQ, Xu J, Wu LH. Strategies on establishment of China's microbiome data center. *Bulletin of Chinese Academy of Sciences*, 2017, 32(3): 290–296. (in Chinese)
马俊才, 赵方庆, 苏晓泉, 徐健, 吴林寰. 关于中国微生物组数据中心建设的思考. 中国科学院院刊, 2017, 32(3): 290–296.

[12] Parrello B, Butler R, Chlenski P, Olson R, Overbeek J, Pusch GD, Vonstein V, Overbeek R. A machine learning-based service for estimating quality of genomes using PATRIC. *BMC Bioinformatics*, 2019, 20(1): 486.

[13] Salzberg SL, Phillippy AM, Zimin A, Puiu D, Magoc T, Koren S, Treangen TJ, Schatz MC, Delcher AL, Roberts M, Marçais G, Pop M, Yorke JA. GAGE: a critical evaluation of genome assemblies and assembly algorithms. *Genome Research*, 2012, 22(3): 557–567.

[14] Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, 2013, 29(8): 1072–1075.

[15] Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, Nielsen PH. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nature Biotechnology*, 2013, 31(6): 533–538.

[16] Bjørn Nielsen H, Almeida M, Juncker AS, Rasmussen S, Li JH, Sunagawa S, Plichta DR, Gautier L, Pedersen AG, le Chatelier E, Pelletier E, Bonde I, Nielsen T, Manichanh C, Arumugam M, Batto JM, dos Santos MBQ, Blom N, Borruel N, Burgdorf KS, Boumezbeur F, Casellas F, Doré J, Dworzynski P, Guarner F, Hansen T, Hildebrand F, Kaas RS, Kennedy S, Kristiansen K, Kultima JR, Léonard P, Levenez F, Lund O, Moumen B, le Paslier D, Pons N, Pedersen O, Prifti E, Qin JJ, Raes J, Sørensen S, Tap J, Tims S, Ussery DW, Yamada T, Renault P, Sicheritz-Ponten T, Bork P, Wang J, Brunak S, Dusko Ehrlich S. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nature Biotechnology*, 2014, 32(8): 822–828.

[17] Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, 2015, 25(7): 1043–1055.

[18] Creevey CJ, Doerks T, Fitzpatrick DA, Raes J, Bork P. Universally distributed single-copy genes indicate a constant rate of horizontal transfer. *PLoS ONE*, 2011, 6(8): e22099.

[19] Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng JF, Darling A, Malfatti S, Swan BK, Gies EA, Dodsworth JA, Hedlund BP, Tsiamis G, Sievert SM, Liu WT, Eisen JA, Hallam SJ, Kyrpides NC, Stepanauskas R, Rubin EM, Hugenholtz P, Woyke T. Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, 2013, 499(7459): 431–437.

[20] Parks DH, Rinke C, Chuvochina M, Chaumeil PA, Woodcroft BJ, Evans PN, Hugenholtz P, Tyson GW. Recovery of nearly 8, 000 metagenome-assembled genomes substantially expands the tree of life. *Nature Microbiology*, 2017, 2(11): 1533–1542.

[21] Mineeva O, Rojas-Carulla M, Ley RE, Schölkopf B, Youngblut ND. DeepMAsED: evaluating the quality of metagenomic assemblies. *Bioinformatics*, 2020, 36(10): 3011–3017.

[22] Saary P, Mitchell AL, Finn RD. Estimating the quality of eukaryotic genomes recovered from metagenomic analysis with EukCC. *Genome Biology*, 2020, 21(1): 244.

[23] Zhan XJ, Yao DJ, Zhu HQ. Bioinformatics methods for high-throughput DNA sequencing data. *Big Data Research*, 2016, 2(2): 76–87. (in Chinese)
詹晓娟, 姚登举, 朱怀球. 高通量 DNA 测序数据的生物信息学方法. 大数据, 2016, 2(2): 76–87.

[24] Almeida A, Mitchell AL, Boland M, Forster SC, Gloor GB, Tarkowska A, Lawley TD, Finn RD. A new genomic blueprint of the human gut microbiota. *Nature*, 2019, 568(7753): 499–504.

[25] Uritskiy GV, DiRuggiero J, Taylor J. MetaWRAP-a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome*, 2018, 6(1): 158.

[26] Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, de Hoon MJL. Biopython: freely available *Python* tools for computational molecular biology and bioinformatics. *Bioinformatics*, 2009, 25(11): 1422–1423.

[27] Hyatt D, Chen GL, Locascio PF, Land ML, Hauser LJ. Prodigal prokaryotic dynamic programming genefinding algorithm. *BMC Bioinformatics*, 2010, 11: 119

[28] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *Journal of Molecular Biology*, 1990, 215(3): 403–410.

[29] Ihaka R, Gentleman R. R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 1996, 5(3): 299–314.

[30] Wickham H. ggplot2: Elegant graphics for data analysis. New York: Springer, 2016.

[31] Chen LX, Anantharaman K, Shaiber A, Eren AM, Banfield JF. Accurate and complete genomes from metagenomes. *Genome Research*, 2020, 30(3): 315–333.

[32] Song WZ, Thomas T. Binning_refiner: improving genome bins through the combination of different binning programs. *Bioinformatics*, 2017, 33(12): 1873–1875.

[33] Sieber CMK, Probst AJ, Sharrar A, Thomas BC, Hess M, Tringe SG, Banfield JF. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nature Microbiology*, 2018, 3(7): 836–843.

[34] Tennessen K, Andersen E, Clingenpeel S, Rinke C, Lundberg DS, Han J, Dangl JL, Ivanova N, Woyke T, Kyrpides N, Pati A. ProDeGe: a computational protocol for fully automated decontamination of genomes. *The ISME Journal*, 2016, 10(1): 269–272.

[35] Olson ND, Treangen TJ, Hill CM, Cepeda-Espinoza V, Ghurye J, Koren S, Pop M. Metagenomic assembly through the lens of validation: recent advances in assessing and improving the quality of genomes assembled from metagenomes. *Briefings in Bioinformatics*, 2019, 20(4): 1140–1150.

[36] Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, 2015, 12(1): 59–60.

Marker gene broken caused overestimation on the contamination of metagenome-assembled genomes and its correction

Hao Li[1#], Dongxu Yang[1#], Linran Wen[1], Wei Zheng[1], Feng Guo[1,2,3*]

[1] School of Life Sciences, Xiamen University, Xiamen 361102, Fujian Province, China

[2] Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), Zhuhai 510275, Guangdong Province, China

[3] Key Laboratory of Microbial Resource (Fujian), Xiamen 361102, Fujian Province, China

Abstract: [Objective] Identifying and correcting the overestimation on contamination of metagenome-assembly genomes (MAGs) caused by the broken marker genes. **[Methods]** The impact of broken genes on quality assessment of genome was first analyzed using the simulated genomes from randomly fragmented the complete genome of isolates. We designed a corrected pipeline that identifying the broken genes pairs from the same "source" gene according to the taxonomic annotation against the nr database. Then the genome contamination was corrected by removing the redundant marker genes. **[Results]** The phenomenon that the genome contamination is positively correlated with the genome fragmentation degree was observed in both simulated genomes and MAGs obtained by genome binning. We designed a corrected pipeline based on the idea of identifying broken genes from the same "source" gene and the results based on the simulated genomes showed the contamination can be adjusted to complete genome level. Testing on 760 MAGs with contamination from gut and soil samples, we observed a reduction in contamination for nearly half of the MAGs, with 43 of them dropping to 0. **[Conclusion]** Our pipeline can correct the overestimated contamination of genome caused by broken genes to some extent and improve the availability of MAGs. The pipeline is expected to apply to the genome quality assessment of the increasing number of MAGs.

Keywords: metagenome assembled genome, genome quality, CheckM, contamination

(本文责编：李磊)

Supported by the National Natural Science Foundation of China (31670492, 31500100)

[#]These authors contributed equally to this work.

[*]Corresponding author. Tel/Fax: +86-592-2880330; E-mail: fguo.bio@xmu.edu.cn

Received: 12 December 2020; Revised: 20 February 2021; Published online: 14 July 2021