微生物学通报
Microbiology China
tongbao@im.ac.cn

May 20, 2014, 41(5): 924−933

DOI: 10.13344/j.microbiol.china.130345

专论与综述

微生物完整基因组测定中的 Gap closure 策略

尤晓颜[1] 张彬[1] 郑华军[2] 姜成英[3*]

(1. 河南科技大学 食品与生物工程学院 河南 洛阳 471023)

(2. 上海人类基因组研究中心 上海 201203)

(3. 中国科学院微生物研究所 微生物资源前期开发国家重点实验室 北京 100101)

摘 要：微生物基因组空缺区域(Gap)中可能存在重要的生物学信息，如果无法补齐所有 Gap，不仅不能获得完整的基因组图谱，还会给后续的基因组信息解读造成很大困难。而基因组空缺区域填充(Gap closure)是获得微生物基因组完成图的关键，本文结合作者以及借鉴上海人类基因组研究中心在微生物基因组 Gap closure 中的经验，针对微生物基因组 Gap closure 常用的 6 种策略：参考序列比对、多引物 PCR、基因组步移、基因组文库克隆末端测序、末端配对(Paired-End)以及基因组光学图谱技术进行综述。

关键词：微生物，基因组，Gap closure

Strategies of gap closure in complete microbial genome sequencing

YOU Xiao-Yan[1] ZHANG Bin[1] ZHENG Hua-Jun[2] JIANG Cheng-Ying[3*]

(1. *College of Food and Bioengineering*, *Henan University of Science and Technology*, *Luoyang*, *Henan* 471023, *China*)

(2. *Chinese National Human Genome Center at Shanghai* (*CHGC*), *Shanghai* 201203, *China*)

(3. *State Key Laboratory of Microbial Resources*, *Institute of Microbiology*, *Chinese Academy of Sciences*, *Beijing* 100101, *China*)

Abstract: Gaps in microbial genome sequences may lead to loss of important biological information and cause trouble for further interpretation of genetic information. Gap closure is thus a critical step in completely genome finishing of microorganisms. In this review, we critically summarize six major strategies for gap filling of microbial genomes, including reference alignment, multiplex PCR, genome walking, genome library clone-end sequence, paired-end and whole genome mapping.

Keywords: Microbe, Genome, Gap closure

微生物是地球上种类最多、数量最大、分布最广的生物群，与人类、动植物和环境有着密切的相互作用，同时也是工业生物技术的核心及重要的国际竞争战略资源。随着测序技术的不断进步以及测序成本的不断降低，越来越多的微生物基因组序列得到测定，其中包括一些重要的病原微生物、工业微生物、极端微生物以及生物学研究中有重要意义的一些模式微生物[1-5]。世界各国对微生物基因组

基金项目：国家自然科学基金项目(No. 31200035，31171234)；河南科技大学博士启动基金项目(No. 090061608)

*通讯作者：Tel：86-10-64807581；✉：jiangcy@mail.im.ac.cn

收稿日期：2013-05-20；接受日期：2013-07-03；优先数字出版日期(www.cnki.net)：2013-10-12

测序这场生命科学的新技术革命也寄予极大的期望，相继投入巨额资金进入这一领域，如：万种微生物基因组计划(Ten thousand kinds of microbial genome project)、人体肠道元基因组研究计划(MetaHIT)等。由于新一代测序技术(Next-generation sequencing techniques，NGS)测序读长较短、基因组结构的复杂性以及测序过程中的偏向性等原因，使得已完成测序的一些物种的基因组中含有数目不等的空缺区域。据统计，自 2008 年以来 GenBank 释放的 5 276 个微生物基因组序列中仅有 32% (1 692)是完整序列[6-7]。基因组空缺区域中可能存在重要的生物学信息，如果不能补齐所有的 Gap，不仅无法获得完整的基因组图谱，还会给后续的基因组信息解读(操纵子结构、基因调控、SNP 分析以及比较基因组等)造成困难[8-10]。因此，完整微生物基因组序列的获得需要在完成测序之后对空缺区域进行填充，即：将测序拼装后生成的叠联群(Contig)之间的 Gap 进行填充，然后按照一定的次序和方向拼装生成一条完整的基因组序列(完成图)，这个过程称之为基因组的 Gap closure (或补洞)。

Gap closure 的关键在于准确定位不同 Contig 之间的相对位置关系(Linkage 关系)，一旦位置关系确定，即可通过 PCR 扩增 Gap 区域序列或是文库克隆步移测序的方式关闭 Gap 区域。然而，由于一些微生物基因组 GC 含量高、重复序列数目多且长度大(插入序列、rDNA 操纵子、大片段重复等)以及 NGS 测序读长较短等原因造成测序偏向性高、拼接后生成过多的 Contig[6,11-12]，从而增加了 Gap closure 的难度。此外，缺少基因组参考序列或是与参考序列比对同源性低等因素，使得生成的 Contig 无法有效定位，也会导致 Gap closure 难度增加，因此 Contig 定位被认为是微生物基因组 Gap closure 过程中最困难和最耗时的阶段。一些生物信息学软件被开发用于微生物基因组的 Gap closure，并取得了一定的效果[7,13-14]；但对于基因组中的高度重复区域和低覆盖率区域的干扰仍无法有效解决，必需借助实验手段获得额外的序列信息才能最终完成基因组的 Gap closure，因而应用的范围和准确性受到一定限制。本文主要对微生物基因组 Gap closure 中常用的 6 种实验策略进行介绍，这些策略包括：参考序列比对(Reference alignment)；多引物 PCR (Multiplex PCR)；基因组步移(Genome walking)；基因组文库克隆末端测序(Clone-end Sequence)；末端配对(Paired-End)；基因组光学图谱(Whole genome mapping)。

1 参考序列比对

对于具有参考基因组序列(同种不同亚种、变种、血清型或近缘同属)的微生物来讲，这是一种非常省时并行之有效地方式。即：将测序产生的 Contig 序列与参考基因组序列进行比对(BLASTn，1e-100)，根据比对结果确定出不同 Contig 之间的相对位置关系，然后在相邻近的 Contig 末端设计向 Gap 区域延伸的引物，通过 PCR 扩增出 Gap 区域片段，最后将 PCR 产物测序后采用 Phred/Phrap/Consed 软件包填充相邻的 Contig 之间 Gap 区域[15-18]。该策略适用于待拼接的基因组序列与参考序列相似度较高的微生物基因组 Gap closure 工作，最好为同种的不同亚种、变种、血清群或是亲缘关系非常接近的同属物种。通过借助于参考序列提供的位置信息，可以降低拼接工作的难度，加速 Gap closure 工作的完成。如在问号勾端螺旋体菌株 IPAV (*Leptospira interrogans* IPAV)的拼接过程中，通过与序列相似度很高的不同血清型的菌株 *Leptospira interrogans* 56601 全基因组序列进行比对，从而将绝大多数 Contig 迅速定位，使得在较短时间内完成了拼接工作[19-20]。然而，许多微生物完成测序后由于缺少基因组参考序列或是与参考序列比对后同源性低，无法提供准确的定位信息，从而造成基因组 Gap closure 延滞，对于这些情况需要采用以下几种实验策略来完成这些微生物基因组的 Gap closure。

2 多引物 PCR

多引物 PCR (多重 PCR)是用多对引物同时对模板 DNA 上的多个区域进行扩增的技术，具有简洁高效、成本低等优点，已广泛应用于基因组拼接、病原菌检测、SNP 位点基因分型、突变分析以及模板定量分析等方面[21-31]。在 Gap closure 中，经常会遇到一些微生物基因组测序后由于 Gap 太多、缺少基因组参考序列造成拼接困难的情况，如果随机组合每一对引物进行 PCR，则工作量非常大。此时，可以考虑采用多引物 PCR 的策略(一般 Contig 数目小于 100 个)：将所有 Contig 末端朝向 Gap 区域扩增的引物进行随机组合(Pool)，每个 Pool 包含 4−20 条引物为宜；然后将不同编号的 Pool 两两组合后进行多引物 PCR 扩增。每一轮 PCR 产物切胶纯化后用相对应 Pool 组合中的引物进行测序，根据测序结果即可判定相对应的两个 Contig 的关系。接着去除关系明确的引物对，将剩余的引物再次进行组合，继续进行下一轮多引物 PCR 扩增，如此进行多轮次(2−10 轮)后，最终关闭所有的 Gap。多引物 PCR 极大地提高了关闭 Gap 的效率，应用该策略，已经有很多微生物的基因组完成拼接[32-33]，如作者在进行好客嗜酸两面菌(*Acidianus hospitalis* W1)的 Gap closure 时，仅采用一轮多引物 PCR，在一周之内即完成了该菌基因组的拼接工作[33]。

进行多引物 PCR 的反应体系不应过小(至少 50 μL)，模板 DNA 的量大于 50 ng；在引物设计时需注意减少引物间的交联，避免造成扩增效率降低，每条引物的终浓度不低于 0.3 μmol/L。此外，多引物 PCR 产物纯化过程中，应使用长板胶进行电泳，使目的条带尽量分开，避免长度相近的不同目的片段混合，混淆测序结果。如果进行多轮多引物 PCR 扩增后，还有少量 Contig 没有找到关系，这可能由于以下 3 种情况造成：(1) 已经确定的 Contig 位置关系中有错误；(2) Gap 区域较大；(3) Gap 区域局部 GC 含量过高导致常规 PCR 条件无法得到扩增产物。对于第一种情况，一般是由于重复序列造成的，此时可以再从 Contig 内部距离原先设计引物的位点 1−2 kb 的位置重新设计引物以避开重复序列(特殊情况下需要再远一些，根据重复区域的长度而定)，然后再次进行 PCR 验证。如果无法扩增出目的条带，则该位点可能拼错，需要将该位点的一对引物和剩余的引物重新组合以确定正确的位置关系，直至将所有 Gap 填充。第三种情况可以使用 LA (Long accurate) *Taq* 酶与 GC buffer (Ⅰ或Ⅱ，TaKaRa，专门用于高 GC 片段扩增)的组合对 Gap 区域进行扩增，也可尝试在扩增体系中加入 DMSO 以消除模板 DNA 的局部二级结构，增加引物与模板的特异性结合，降低解链温度。

3 基因组步移

对于上述第二种情况，可以先考虑采用 Long PCR 进行扩增[34]，如果依然没有效果，可以采用基因组步移的方式从关系未知 Contig 的末端向相邻的未知 Gap 区域进行延伸，直至与其它的 Contig 建立联系。通过这种方式可以确定位置关系待定的 Contig 的位置信息，同时也可以鉴别先前是否存在拼接错误。从已知序列向邻近未知序列步移的方法有多种，如反向 PCR[35-38]、连接介导 PCR[39-42]、TAIL-PCR (Thermal asymmetric interlaced PCR)[43-45]、随机引发 PCR 等[46-48]。这些方法各具特点，反向 PCR 和连接介导 PCR 由于需要酶切和连接等步骤，操作较繁琐，成本高且耗时较长，在基因组步移中应用相对较少。TAIL-PCR 操作简单，应用广泛，但由于扩增产物长度较小(小于 1 kb)且易于扩增出重复片段，在 Gap closure 中也较少应用。Gap closure 中常用以下 2 种步移方式。

3.1 寻靶 PCR (Site-finding PCR)

寻靶 PCR 技术由北京大学生命科学学院陈章良实验室首先提出，主要用于由已知序列向未知序列区域进行基因组步移，该技术首先采用寻靶引物(Site-finder primer，SFP)在低温下进行一轮随机单

引物延伸,然后在反应体系中加入已知序列最内侧的特异引物(Gene-specific primer，GSP1)进行 PCR 扩增，将扩增后的 PCR 产物稀释 500−1 000 倍(根据实际情况)后作为模板，用已知序列外侧的特异引物(GSP2，GSP3)和靶向引物(SFP1，SFP2)分别进行巢式 PCR 扩增；将扩增得到的 PCR 产物切胶回收后用 *Not* Ⅰ酶切后与载体(pBlueScript)连接，然后使用特异的外侧引物和载体的引物进行测序，从而得到未知区域序列[49]。

寻靶 PCR 技术的关键在于寻靶引物的设计，引物的最末端选取在基因组中出现频率最高的 4 个碱基组合,以提高在低温下随机错配发生的概率(寻找靶位点)；紧挨着 4 个碱基的前面为 6 个随机碱基，使得寻靶引物可以与靶点位置的 10 个碱基产生较紧密结合，便于接下来巢式 PCR 的进行。由于扩增反应采用的是高保真的 *Pfu* 酶,产物为平末端，为了与载体定向连接，故在随机碱基的后面设计 *Not* Ⅰ酶切位点，用于产生粘性末端。寻靶引物的左侧设计有靶向引物序列，靶向引物 SFP1 (20 bp)与 SFP2 (24 bp)为嵌套引物，用于进行巢式 PCR，之间有 9 个碱基的重合。巢式 PCR 的目的在于保证扩增出序列的保真性,当没有目标序列扩增时，靶向引物扩增后产生颈环结构，酶切后将无法连入载体。进行 Site-finding PCR 模板的量需要大于 40 ng (20 μL 体系)，所用的聚合酶必需是 LA*Taq* 酶或是高保真的 *Pfu* 酶。另外，特异性引物的退火温度最好和酶的最适延伸温度一致(68 ℃，PCR 延伸和退火为一步)。与其它几种方法相比，寻靶 PCR 扩增效率高、特异性好、扩增片段长(1−3 kb)，但扩增后需要酶切和连接等步骤，仍然存在着操作繁琐、成本较高的问题，因此，近年来该方法得到了进一步的改进[50-51]，但由于该方法扩增产物片段长、特异性好，在基因组拼接中有着特殊的应用价值。

3.2 单引物 PCR

单引物 PCR 利用引物在较低退火温度下随机引发扩增实现[52]，可分为两个反应步骤：单引物随机扩增和巢式 PCR。该方法需要在位置关系待定的 Contig 末端依次设计一条内侧引物(特异性引物)和一条外侧引物(用于非特异性扩增)。首先，在扩增体系中只加入外侧引物，利用外侧引物在较低退火温度下(35 ℃)进行随机非特异性扩增，然后将扩增产物稀释(500−1 000 倍)后作为模板，在扩增体系中加入内侧特异性引物和外侧引物进行巢式 PCR 扩增(特异性扩增)。将扩增出的产物片段切胶回收后用内侧和外侧引物分别测序，测序得到的序列即为 Gap 区域的序列 如此不断向 Gap 区域延伸，直至与其它位置关系待定的 Contig 建立联系，关闭之间的 Gap。利用该策略，郑华军等成功关闭了一个长达 10 kb 的 Gap，最终获得了德氏乳杆菌保加利亚种 2038 菌株的基因组完成图[53]。需要注意的是设计外侧引物时长度不能太长，一般 11−17 个碱基即可，以提高随机引发扩增的概率；另外,加入模板的浓度需大于 10 ng (20 μL 体系)。

总体而言，基因组步移的效率比较低，扩增出的片段长度一般在 1−3 kb 左右。该策略主要在还剩少数几个 Gap、不能通过常规 PCR 方式填充的情况下使用，如果 Gap 区域内含有重复序列，则不能采用该策略。

4 基因组文库克隆末端测序

如果一个待 Gap closure 的微生物基因组缺少参考序列,并且 Contig 数目比较多时(大于 200 个)可以考虑构建基因组文库、利用对克隆进行末端测序(End sequence)的方式进行 Contig 定位。该方法通过对克隆中插入片段的两个末端采用载体上的通用引物进行常规测序(Sanger 法，长度 600−1 000 bp)，然后将测序得到的末端序列与所有 Contig 序列进行比对(BLASTn，1e-100)；如果一个插入片段的两个末端分别比对到两个不同的 Contig 上，那么就可以确定这两个 Contig 相邻接(或存在 Linkage 关系)，它们之间的 Gap 大小即为该克隆中插入片段的长度，通过 PCR 扩增后测序或直接以该克隆进行引物步移(Primer walking)测序的方式即可得到

Gap 区域序列。

基因组 Gap closure 工作通常采用插入片段为 6−8 kb 的质粒文库和 30−40 kb 的 Fosmid 文库。质粒文库所用的载体一般为 pUC18 和 pSMART (Lucigen 公司)，两者均为平末端载体，便于目的片段的随机插入。pUC18 常用于革兰氏阴性菌的克隆，而 pSMART 则适用于革兰氏阳性菌。pSMART 是一种低拷贝的载体，在载体的克隆位点前后均含有转录终止子，可以抑制插入片段的转录，从而减少克隆阳性菌基因组片段时由于部分基因过量表达对宿主大肠杆菌产生的毒害致死作用。Fosmid 文库一般用于构建微生物基因组的整体构架，帮助从全局上确定基因组的拼接情况。通过一定覆盖率(5−10 倍)的克隆末端测序，可以从整体上大致确定整个基因组的框架、验证先前拼接结果的正确性，以及确定一些由于 Gap 较大(大于 15 kb)的原因造成无法定位的关系。Fosmid 文库插入片段长度较大(30−40 kb)，一般用于微生物基因组 Gap closure 的后期，与 6−8 kb 质粒文库相互配合使用效果最好。如果经过文库克隆末端测序后仍然有关系未知的 Contig 存在，则可以在这些 Contig 的两侧设计探针引物对 Fosmid 文库进行筛选，将筛选到的阳性 Fosmid 克隆首先进行末端测序以确定相应 Contig 之间的位置关系，然后将对应的克隆 DNA 抽提出来(要求大于 5 μg)打断构建 1.6−4.0 kb 的 pSMART 亚克隆文库，按照 Fosmid 插入片段长度的 10 倍克隆覆盖率随机挑选亚克隆进行常规 Sanger 法测序。最后，将测得的序列采用 Phred/Phrap/Consed 软件包拼装后即可得到 Gap 区域的未知序列，如作者在喜温硫杆菌 SM-1 基因组 Gap closure 时，就采用该方法将最后 3 个大于 15 kb 的 Gap 关闭[11]。

5 末端配对

2007 年，罗氏 454 公司首先推出了商业应用的末端配对(Paired-End)技术，用于帮助确定 Contig 之间的相对位置关系[54]。Paired-End 文库构建的原理如下(以 3 kb 长度插入片段为例)：将基因组 DNA 采用 HydroShear 随机片段化(Genomic Solutions 公司，英国)，截取 3 kb 左右的组分回收。将回收后的片段末端进行修饰，然后利用甲基化酶甲基化，以避免被限制性内切酶 *Eco*R Ⅰ切断，接着加上末端含有发夹结构的接头。接头上含有生物素标签以及 *Eco*R Ⅰ识别的未甲基化的位点，未连接上发夹状接头的片段通过核酸外切酶去除。将两端含有接头的插入片段进行 *Eco*R Ⅰ酶切，去除发夹结构露出粘性末端，然后连接成环。接下来使用氮气将环化的片段打断，回收 200−300 bp 的片段，然后用带有链亲和素的磁珠回收带有生物素标记的片段。将回收得到的片段末端加上 Paired-End 接头，该接头上含有用于 emPCR (Emulsion-based clonal amplification)扩增的引物结合序列和用于 454 测序的引物结合序列。进行 emPCR 扩增，破油后回收扩增产物，然后上样，进行 454 测序。如果测序得到的 Reads 两端比对到两个不同的 Contig，就认为这两个 Contig 存在着 Link 关系，它们之间的 Gap 大约为 3 kb 左右。采用 Newbler 拼接软件(454 Life Sciences，Roche)将测序得到的 Reads 和基因组测序的 Reads 进行组装，软件可以估计出 Gap 的大小并用字母 N 代替 Gap 区域的碱基序列，最终生成包含依次相连接的多个 Contig 位置信息的框架图(Scaffold)。在进行的基因组 Gap closure 工作时，首先将 Scaffold 内部的 Contig 之间的 Sequence gap 通过 PCR 测序的方式填充，然后可以用多引物 PCR 的策略将不同 Scaffold 之间的 Physical gap 关闭。

Paired-End 技术具有通量高、插入片段随机性好、高效省时等优点，较传统 Gap closure 方式具有明显的优势。该方法测序所得的 Reads 可以和基因组 454 测序的 Reads 直接进行拼装，生成包含 Link 关系的 Scaffold。而没有被用于构建 Scaffold 的 Reads 可以当做 Shot-gun reads 拼装到 Contig 序列中去，补强相应位置的低值碱基，提高 Contig 序列的质量。Paired-End 文库插入片段的长度可以达到 3−8 kb，甚至 20 kb，能够跨过较大的 Gap。由于具有高效、省时、随机性好等优点，该技术已

经得到广泛应用[54-60]。除 Paired-End 之外，ABI、Illumina 等主流测序公司也已推出各自的双端测序(Mate-pair)技术用于搭建 Scaffold[61-65]。

6 基因组光学图谱技术

基因组光学图谱技术是一种利用单分子 DNA 的限制性内切酶图谱生成高分辨率、有序的全基因组限制性内切酶酶切图谱的方法[66-67]。2010 年，OpGen 公司(盖瑟斯堡，美国)面向市场推出了基于光学图谱专利技术的自动化的基因组拼接和数据分析系统——Argus™工作站，使研究人员能够在一个工作日内获得待测微生物基因组的高质量全基因组限制性内切酶酶切图谱。该系统主要由 Mapcard 加工工作站、光学扫描系统和数据处理工作站等组成，通过限制性内切酶对固定于 MapCard DNA surface 区域中的单分子 DNA 进行原位切割，使切割后的 DNA 片段顺序保持不变。DNA 片段经荧光染料染色后置于荧光显微镜下，采集每个限制性内切酶片段的大小和顺序的信息，信息经转换处理后生成单个 DNA 分子的限制性内切酶酶切位点图谱，最后根据全部 DNA 分子限制性内切酶酶切位点图谱的相互重叠部分拼接得到全基因组限制性内切酶酶切位点图谱。该系统最初被应用于微生物基因组序列组装、比较基因组学以及菌株分类[9,68-72]，目前已经开始在人类和动植物全基因组组装中应用[73-75]。

作为对基因组 454-Sanger 杂交测序法的一种新的替代技术[76]，光学图谱技术提供了一种全新的微生物基因组测序和拼接策略，虽然该技术不能直接获得基因组 DNA 的碱基序列，但它可以提供微生物基因组整体有序的物理框架结构。通过与新一代测序技术的结合，光学图谱技术可以提供基因组绝大多数 Contig 的位置和顺序信息，从而辅助基因组的 Scaffold 构建，在复杂微生物基因组(多重复序列)的 Gap closure 中已经被证明行之有效。如在嗜线虫致病杆菌(*Xenorhabdus nematophila*)基因组的 Gap closure 过程中，由于含有多达几百个转座子序列以及 7 套核糖体 RNA 序列(16-23-5S)，导致经过 4 个月的拼接后还剩余 36 个 Contig 无法找到 Linkage 关系完成拼接，借助于光学图谱技术研究人员通过将嗜线虫致病杆菌高质量的基因组酶切物理图谱(基因组限制性内切酶酶切图谱)与基因组序列模拟酶切图谱(根据测序序列信息计算机模拟所生成基因组 *In silico* 酶切图谱)进行比较，最终将 36 个 Contig 定位，并且通过比较图谱之间的差异还发现几处由于新一代测序技术自身缺陷(读长短、碱基插入和缺失)所造成的 Contig 本身拼接错误，经过纠正，在一个月内完成了嗜线虫致病杆菌基因组的 Gap closure[70]。

除了基因组序列模拟的酶切图谱，Argus™系统还可以整合其它一些序列的酶切物理图谱信息(文库克隆序列、Gap closure PCR 产物以及 Paired-End 序列等)用于辅助微生物基因组的 Gap closure 工作[77]。

7 结语

新一代测序技术的商业化应用，使得测序成本不断降低，测序通量不断提高，越来越多的微生物基因组得到测定，极大地促进了微生物基因组学的发展。然而，由于测序读长短、数据量大、基因组结构复杂等因素却并没有降低基因组数据拼接和 Gap closure 的难度。由于一些微生物基因组重复序列较多、GC 含量高、测序覆盖率低以及测序读长短等因素会使得序列拼接后产生很多 Gap，如果再缺少参考基因组序列或是与参考序列同源性低，Gap closure 难度将会很大。提高测序覆盖率在一定程度上可以有效减少基因组中的 Gap，但成本相对较高，并且对于一些复杂的微生物基因组效果有限，如 454 二代测序覆盖率为 10×时，喜温硫杆菌 SM-1 测序后生成的 Gap 数目为 400 个；测序覆盖率提高至 25×时，Gap 数目减少至 280 个；但当测序覆盖率继续提高至 38×时，Gap 数目进入平台期(276 个)，相比 25×测序覆盖率时仅减少了 4 个，已经不能再单纯通过提高覆

盖率来减少 Gap 数目。因此，对于复杂的微生物基因组，需要将基因组的 Gap closure 分为几个阶段，针对不同阶段采用相应的策略进行：如果 Gap 数目大于 200 个，可以通过构建基因组文库、Paired-End 测序或者采用基因组光学图谱技术的策略确定 Contig 之间的相对位置和顺序，然后再依次关闭 Contig 之间的 Gap 区域；当 Gap 数目小于 100 个时，可以采用多引物 PCR 的策略寻找 Linkage 信息，关闭所有能够关闭的 Gap；如果最后还剩余几个 Gap 无法关闭，则可以采用基因组步移或文库筛选的策略，如作者在对喜温硫杆菌 SM-1 基因组 Gap closure 时，通过结合构建基因组文库、Paired-End 测序、多引物 PCR 以及 Fosimid 文库筛选等多种策略最终完成了 SM-1 基因组的 Gap closure[11]。6 种常用的微生物基因组 Gap closure 策略比较见表 1。

研究表明，第三代测序技术对微生物基因组的 Gap closure 也具有促进作用，2011 年太平洋生物科学公司(Pacific biosciences，美国)面向市场推出了第一台商业化的单分子实时测序(Single molecule real time DNA sequencing，SMRT)系统 PacBio RS，由于其测序读长长(平均 2−3 kb)、测序无偏向性等优点使得在微生物基因组 Gap closure 中显示出良好的应用前景[12,78-79]，但由于测序通量较小、测序成本较高以及测序错误率较高(约 15%)等因素使其在实际应用中受到了一定限制。目前，太平洋生物科学公司正在对 PacBio RS 测序系统进行改进，以期进一步提高测序读长和测序通量、降低测序错误率。此外，随着生物信息学的不断发展，用于微生物基因组 Gap closure 的生物学软件也将不断完善。我们有理由相信，今后微生物基因组的 Gap closure 成本将会越来越低，难度将会越来越小，所花费的时间也会越来越短。本文主要结合作者以及借鉴上海人类基因组研究中心多年的微生物基因组 Gap closure 经验，对微生物基因组 Gap closure 中常用的 6 种实验策略进行了综述，以期为广大科研同行提供参考借鉴。

表 1 6 种常用的微生物基因组 Gap closure 策略比较
Table 1 Comparisons of the frequently-used strategies for microbial genome gap closure

	参考序列比对 Reference alignment	多引物 PCR Multiplex PCR	基因组步移 Genome walking	文库克隆末端测序 Clone-end sequence	末端配对 Paired-End	基因组光学图谱 Whole genome mapping
适用条件 Applicable conditions	有参考序列	Gaps≤100	Gaps≤5	Gaps≥200	Gaps≥200	Gaps≥200
成本 Costs	低	较低	低	高	高	高
耗时 Time needed	少	较多	少	多	少	少
操作难易 Ease of operation	易	易	较难	难	难	难
文库构建 Library construction	否	否	否	是	是	否

致谢: 本文撰写和试验过程中得到中国科学院微生物研究所刘双江研究员课题组和上海人类基因组研究中心王升跃研究员课题组的悉心指导和帮助，在此表示衷心感谢。

参 考 文 献

[1] Kawarabayasi Y, Hino Y, Horikawa H, et al. Complete genome sequence of an aerobic hyper-thermophilic crenarchaeon, *Aeropyrum pernix* K1[J]. DNA Research, 1999, 6(2): 83-101,145-152.

[2] Ng WV, Kennedy SP, Mahairas GG, et al. Genome sequence of *Halobacterium* species NRC-1[J]. Proceedings of the National Academy of Sciences, 2000, 97(22): 12176-12181.

[3] Bolotin A, Wincker P, Mauger S, et al. The complete genome sequence of the lactic acid bacterium *Lactococcus lactis* ssp. *lactis* IL1403[J]. Genome Research, 2001, 11(5): 731-753.

[4] Parkhill J, Wren BW, Thomson NR, et al. Genome sequence of *Yersinia pestis*, the causative agent of plague[J]. Nature, 2001, 413(6855): 523-527.

[5] Shimizu T, Ohtani K, Hirakawa H, et al. Complete genome sequence of *Clostridium perfringens*, an anaerobic flesh-eater[J]. Proceedings of the National Academy of Sciences, 2002, 99(2): 996-1001.

[6] English AC, Richards S, Han Y, et al. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology[J]. PLoS One, 2012, 7(11): e47768. DOI: 10.1371/journal.pone.0047768.

[7] Tang B, Wang Q, Yang M, et al. ContigScape: a Cytoscape plugin facilitating microbial genome gap closing[J]. BMC Genomics, 2013, 14: 289. DOI: 10.1186/1471-2164-14-289.

[8] Fraser CM, Eisen JA, Nelson KE, et al. The value of complete microbial genome sequencing (you get what you pay for)[J]. Journal of Bacteriology, 2002, 184(23): 6403-6405.

[9] Nagarajan N, Cook C, Di Bonaventura M, et al. Finishing genomes with limited resources: lessons from an ensemble of microbial genomes[J]. BMC Genomics, 2010, 11: 242. DOI: 10.1186/1471-2164-11-242.

[10] Alkan C, Sajjadian S, Eichler EE. Limitations of next-generation genome sequence assembly[J]. Nature Methods, 2011, 8(1): 61-65.

[11] You XY, Guo X, Zheng HJ, et al. Unraveling the *Acidithiobacillus caldus* complete genome and its central metabolisms for carbon assimilation[J]. Journal of Genetics and Genomics, 2011, 38(6): 243-252.

[12] Sergey K, Gregory PH, Timothy PL, et al. Reducing assembly complexity of microbial genomes with single-molecule sequencing[J]. Genome Biology, 2013, 14(9): R101.

[13] Van Hijum SA, Zomer AL, Kuipers OP, et al. Projector 2: contig mapping for efficient gap-closure of prokaryotic genome sequence assemblies[J]. Nucleic Acids Research, 2005, 33(Suppl 2): W560-W566.

[14] Boetzer M, Pirovano W. Toward almost closed genomes with GapFiller[J]. Genome Biology, 2012, 13(6): R56. DOI: 10.1186/gb-2012-13-6-r56.

[15] Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities[J]. Genome Research, 1998, 8(3): 186-194.

[16] Ewing B, Hillier L, Wendl MC, et al. Base-calling of automated sequencer traces using phred. I. Accuracy assessment[J]. Genome Research, 1998, 8(3): 175-185.

[17] Gordon D, Abajian C, Green P. Consed: a graphical tool for sequence finishing[J]. Genome Research, 1998, 8(3): 195-202.

[18] Gordon D, Desmarais C, Green P. Automated finishing with autofinish[J]. Genome Research, 2001, 11(4): 614-625.

[19] Ren SX, Fu G, Jiang XG, et al. Unique physiological and pathogenic features of *Leptospira interrogans* revealed by whole-genome sequencing[J]. Nature, 2003, 422(6934): 888-893.

[20] Zhong Y, Chang X, Cao XJ, et al. Comparative proteogenomic analysis of the *Leptospira interrogans* virulence-attenuated strain IPAV against the pathogenic strain 56601[J]. Cell Research, 2011, 21(8): 1210-1229.

[21] Chehab FF, Kan YW. Detection of specific DNA sequences by fluorescence amplification: a color complementation assay[J]. Proceedings of the National Academy of Sciences, 1989, 86(23): 9178-9182.

[22] Burgart LJ, Robinson RA, Heller MJ, et al. Multiplex polymerase chain reaction[J]. Modern Pathology, 1992, 5(3): 320-323.

[23] Tettelin H, Radune D, Kasif S, et al. Optimized multiplex PCR: efficiently closing a whole-genome shotgun sequencing project[J]. Genomics, 1999, 62(3): 500-507.

[24] Altinok I, Capkin E, Kayis S. Development of multiplex PCR assay for simultaneous detection of five bacterial fish pathogens[J]. Veterinary Microbiology, 2008, 131(3/4): 332-338.

[25] Hu MX, Zhuo K, Liao JL. Multiplex PCR for the simultaneous identification and detection of *Meloidogyne incognita*, *M. enterolobii*, and *M. javanica* using DNA extracted directly from individual galls[J]. Phytopathology, 2011, 101(11): 1270-1277.

[26] Szalanski AL, Tripodi AD, Austin JW. Multiplex polymerase chain reaction diagnostics of bed bug (Hemiptera: Cimicidae)[J]. Journal of Medical Entomology, 2011, 48(4): 937-940.

[27] Beadling C, Neff TL, Heinrich MC, et al. Combining highly multiplexed PCR with semiconductor-based sequencing for rapid cancer genotyping[J]. Journal of Molecular Diagnostics, 2013, 15(2): 171-176.

[28] Esteves LM, Bulhoes SM, Brilhante MJ, et al. Three multiplex snapshot assays for SNP genotyping in candidate innate immune genes[J]. BMC Research Notes, 2013, 6: 54. DOI: 10.1186/1756-0500-6-54.

[29] Kaewmanee M, Phoksawat W, Romphruk A, et al. Development of a multiplex polymerase chain reaction-sequence-specific primer method for NKG2D and NKG2F single-nucleotide polymorphism typing using isothermal multiple displacement amplification products[J]. Tissue Antigens, 2013, 81(6): 419-427.

[30] Vogt PH, Bender U. Human Y chromosome microdeletion analysis by PCR multiplex protocols identifying only clinically relevant AZF microdeletions[J]. Methods in Molecular Biology, 2013, (927): 187-204.
[31] Zhu J, Chen L, Mao Y, et al. Multiplex allele-specific amplification from whole blood for detecting multiple polymorphisms simultaneously[J]. Genetic Testing and Molecular Biomarkers, 2013, 17(1): 10-15.
[32] Liu LJ, You XY, Zheng H, et al. Complete genome sequence of *Metallosphaera cuprina*, a metal sulfide-oxidizing Archaeon from a hot spring[J]. Journal of Bacteriology, 2011, 193(13): 3387-3388.
[33] You XY, Liu C, Wang SY, et al. Genomic analysis of *Acidianus hospitalis* W1 a host for studying crenarchaeal virus and plasmid life cycles[J]. Extremophiles, 2011, 15(4): 487-497.
[34] Cheng S, Chang SY, Gravitt P, et al. Long PCR[J]. Nature, 1994, 369(1994): 684-685.
[35] Benkel BF, Fong Y. Long range-inverse PCR (LR-IPCR): extending the useful range of inverse PCR[J]. Genetic Analysis: Biomolecular Engineering, 1996, 13(5): 123-127.
[36] Pang KM, Knecht DA. Partial inverse PCR: a technique for cloning flanking sequences[J]. Biotechniques, 1997, 22(6): 1046-1048.
[37] Erster O, Liscovitch M. A modified inverse PCR procedure for insertion, deletion, or replacement of a DNA fragment in a target sequence and its application in the ligand interaction scan method for generation of ligand-regulated proteins[J]. Methods in Molecular Biology, 2010(634): 157-174.
[38] Pavlopoulos A. Identification of DNA sequences that flank a known region by inverse PCR[J]. Methods in Molecular Biology, 2011(772): 267-275.
[39] Mueller PR, Wold B. *In vivo* footprinting of a muscle specific enhancer by ligation mediated PCR[J]. Science, 1989, 246(4931): 780-786.
[40] Riley J, Butler R, Ogilvie D, et al. A novel, rapid method for the isolation of terminal sequences from yeast artificial chromosome (YAC) clones[J]. Nucleic Acids Research, 1990, 18(10): 2887-2890.
[41] Jones DH, Winistorfer SC. Sequence specific generation of a DNA panhandle permits PCR amplification of unknown flanking DNA[J]. Nucleic Acids Research, 1992, 20(3): 595-600.
[42] Yuanxin Y, Chengcai A, Li L, et al. T-linker-specific ligation PCR (T-linker PCR): an advanced PCR technique for chromosome walking or for isolation of tagged DNA ends[J]. Nucleic Acids Research, 2003, 31(12): e68.
[43] Liu YG, Mitsukawa N, Oosumi T, et al. Efficient isolation and mapping of *Arabidopsis thaliana* T-DNA insert junctions by thermal asymmetric interlaced PCR[J]. Plant Journal, 1995, 8(3): 457-463.
[44] Liu YG, Whittier RF. Thermal asymmetric interlaced PCR: automatable amplification and sequencing of insert end fragments from P1 and YAC clones for chromosome walking[J]. Genomics, 1995, 25(3): 674-681.
[45] Zhou Z, Ma H, Qu L, et al. Establishment of an improved high-efficiency thermal asymmetric interlaced PCR for identification of genomic integration sites mediated by phiC31 integrase[J]. World Journal of Microbiology & Biotechnology, 2012, 28(3): 1295-1299.
[46] Ohara O, Dorit RL, Gilbert W. One-sided polymerase chain reaction: the amplification of cDNA[J]. Proceedings of the National Academy of Sciences, 1989, 86(15): 5673-5677.
[47] Wesley CS, Ben M, Kreitman M, et al. Cloning regions of the Drosophila genome by microdissection of polytene chromosome DNA and PCR with nonspecific primer[J]. Nucleic Acids Research, 1990, 18(3): 599-603.
[48] Wu C, Zhu S, Simpson S, et al. DOP-vector PCR: a method for rapid isolation and sequencing of insert termini from PAC clones[J]. Nucleic Acids Research, 1996, 24(13): 2614-2615.
[49] Tan G. SiteFinding-PCR: a simple and efficient PCR method for chromosome walking[J]. Nucleic Acids Research, 2005, 33(13): e122.
[50] Wang S, He J, Cui Z, et al. Self-formed adaptor PCR: a simple and efficient method for chromosome walking[J]. Applied and Environmental Microbiology, 2007, 73(15): 5048-5051.
[51] Gröning JA, Tischler D, Kaschabek SR, et al. Optimization of a genome-walking method to suit GC-rich template DNA from biotechnological relevant Actinobacteria[J]. Journal of Basic Microbiology, 2010, 50(5): 499-502.
[52] Telenius H, Carter NP, Bebb CE, et al. Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer[J]. Genomics, 1992, 13(3): 718-725.
[53] Hao P, Zheng H, Yu Y, et al. Complete sequencing and pan-genomic analysis of *Lactobacillus delbrueckii* subsp. *bulgaricus* reveal its genetic basis for industrial yogurt production[J]. PLoS One, 2011, 6(1): e15964. DOI: 10.1371/journal.pone.0015964.
[54] Korbel JO, Urban AE, Affourtit JP, et al. Paired-end mapping reveals extensive structural variation in the human genome[J]. Science, 2007, 318(5849): 420-426.
[55] Fullwood MJ, Wei CL, Liu ET, et al. Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses[J]. Genome Research, 2009, 19(4): 521-532.
[56] Sindi S, Helman E, Bashir A, et al. A geometric approach for classification and comparison of structural variants[J]. Bioinformatics, 2009, 25(12): i222-i230.
[57] Liang C, Liu X, Yiu SM, et al. De novo assembly and characterization of *Camelina sativa* transcriptome by paired-end sequencing[J]. BMC Genomics, 2013, 14(1): 146. DOI: 10.1186/1471-2164-14-146.
[58] Ng KP, Yew SM, Chan CL, et al. Draft genome sequence of the first isolate of extensively drug-resistant (XDR) *Mycobacterium tuberculosis* in Malaysia[J]. Genome Announcements, 1(1): pii: e00056-12.
[59] Robinson DR, Wu YM, Kalyana-Sundaram S, et al. Identification of recurrent NAB2-STAT6 gene fusions in solitary fibrous tumor by integrative sequencing[J]. Nature Genetics, 2013, 45(2): 180-185.
[60] Voet T, Kumar P, Van Loo P, et al. Single-cell paired-end genome sequencing reveals structural variation per cell cycle[J]. Nucleic Acids Research, 2013, 42(11):

6119-6138.
[61] Medvedev P, Stanciu M, Brudno M. Computational methods for discovering structural variation with next-generation sequencing[J]. Nature Methods, 2009, 6(11): S13-S20.
[62] Belova T, Zhan B, Wright J, et al. Integration of mate pair sequences to improve shotgun assemblies of flow-sorted chromosome arms of hexaploid wheat[J]. BMC Genomics, 2013, 14: 222. DOI: 10.1186/1471-2164-14-222.
[63] Jiao X, Hooper SD, Djureinovic T, et al. Gene rearrangements in hormone receptor negative breast cancers revealed by mate pair sequencing[J]. BMC Genomics, 2013, 14: 165. DOI: 10.1186/1471-2164-14-165.
[64] Moulin L, Mornico D, Melkonian R, et al. Draft Genome Sequence of *Rhizobium mesoamericanum* STM3625, a Nitrogen-Fixing Symbiont of Mimosa pudica Isolated in French Guiana (South America)[J]. Genome announcements, 2013, 1(1): pii: e00066-12.
[65] Van Heesch S, Kloosterman WP, Lansu N, et al. Improving mammalian genome scaffolding using large insert mate-pair next-generation sequencing[J]. BMC Genomics, 2013, 14: 257. DOI: 10.1186/1471-2164-14-257.
[66] Schwartz DC, Li X, Hernandez LI, et al. Ordered restriction maps of *Saccharomyces cerevisiae* chromosomes constructed by optical mapping[J]. Science, 1993, 262(5130): 110-114.
[67] Levy-Sakin M, Ebenstein Y. Beyond sequencing: optical mapping of DNA in the age of nanotechnology and nanoscopy[J]. Current Opinion in Biotechnology, 2013, 24(4): 690-698. DOI: 10.1016/j.copbio.2013.01.009.
[68] Zhou S, Kile A, Bechner M, et al. Single-molecule approach to bacterial genomic comparisons via optical mapping[J]. Journal of Bacteriology, 2004, 186(22): 7773-7782.
[69] Zhou S, Kile A, Kvikstad E, et al. Shotgun optical mapping of the entire *Leishmania major* Friedlin genome[J]. Molecular and Biochemical Parasitology, 2004, 138(1): 97-106.
[70] Latreille P, Norton S, Goldman BS, et al. Optical mapping as a routine tool for bacterial genome sequence finishing[J]. BMC Genomics, 2007, 8: 321. DOI: 10.1186/1471-2164-8-321.
[71] Nagarajan N, Read TD, Pop M. Scaffolding and validation of bacterial genome assemblies using optical restriction maps[J]. Bioinformatics, 2008, 24(10): 1229-1235.
[72] Turner PC, Yomano LP, Jarboe LR, et al. Optical mapping and sequencing of the *Escherichia coli* KO11 genome reveal extensive chromosomal rearrangements, and multiple tandem copies of the *Zymomonas mobilis pdc* and *adhB* genes[J]. Journal of Industrial Microbiology & Biotechnology, 2011, 39(4): 629-639.
[73] Zhou S, Bechner MC, Place M, et al. Validation of rice genome sequence by optical mapping[J]. BMC Genomics, 2007, 8: 278. DOI: 10.1186/1471-2164-8-278.
[74] Zhang Q, Chen W, Sun L, et al. The genome of *Prunus mume*[J]. Nature Communications, 2012, 3: 1318. DOI: 10.1038/ncomms2290.
[75] Dong Y, Xie M, Jiang Y, et al. Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (*Capra hircus*)[J]. Nature Biotechnology, 2013, 31(2): 135-141.
[76] Goldberg SM, Johnson J, Busam D, et al. A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes[J]. Proceedings of the National Academy of Sciences, 2006, 103(30): 11240-11245.
[77] Skiadas J, Aston C, Samad A, et al. Optical PCR: genomic analysis by long-range PCR and optical mapping[J]. Mammalian Genome, 1999, 10(10): 1005-1009.
[78] Bashir A, Klammer AA, Robins WP, et al. A hybrid approach for the automated finishing of bacterial genomes[J]. Nature Biotechnology, 2012, 30(7): 701-707.
[79] Ribeiro FJ, Przybylski D, Yin S, et al. Finished bacterial genomes from shotgun sequence data[J]. Genome Research, 2012, 22(11): 2270-2277.