bio-101 www.bio-protocol.org/e1010640 DOI:10.21769/BioProtoc.1010640

基于高通量测序的全长 DNA 条形码获取方法

Methods for Obtaining Full-length DNA Barcodes Using High-throughput Sequencing

杨琛涛[1]，周程冉[1]，刘山林[2]，周欣[2,*]

[1]深圳华大生命科学研究院，深圳 518083；[2]昆虫学系，中国农业大学，北京 100193

*通讯作者邮箱：xinzhou@cau.edu.cn

引用格式：杨琛涛，周程冉，刘山林，周欣.（2021）. 基于高通量测序的全长 DNA 条形码获取方法. Bio-101 e1010640. Doi：10.21769/BioProtoc.1010640.

How to cite: Yang, C. T., Zhou, C. R., Liu, S. L. and Zhou, X. (2021). Methods for Obtaining Full-length DNA Barcodes Using High-throughput Sequencing. Bio-101 e1010640. Doi: 10.21769/BioProtoc.1010640. (in Chinese)

摘要：DNA 条形码在分类学、分子生态学等领域中具有显著的应用价值。近年来，利用高通量测序技术批量获取标准参考条形码的方法经历了快速的发展，展现了巨大的发展前景。针对各领域对于动物线粒体 COI 条形码参考序列的大量常规需求，我们提出并搭建了一系列结合高通量测序（High throughput sequencing，HTS）技术和生物信息分析流程的方法，实现了标准 COI 条形码序列的经济、快速、高效获取。本系列方法简称为 HIFI-Barcode 方法，包含三个主要部分：(1) 测序前实验：需要对每个样品单独提取 DNA，并用 96 对带有特定长度和特定标签序列的条形码扩增引物进行聚合酶链反应扩增（Polymerase chain reaction，PCR），后收集并混合扩增产物。(2) 文库构建及测序：首先根据项目需求选择测序平台及技术，并根据待测序技术进行文库的构建与测序；包括 Illumina 平台和 MGISEQ 平台的 150 bp 配对末端读段测序技术（PE150 sequencing），以及 Pacific Biosciences (Pacbio) 平台的长读段单分子实时（SMRT）测序和 MGI2000 平台的单端 400 bp 测序技术（SE400）等。(3) 条形码数据获取：完成测序后，利用本团队研发的软件包进行分析，可以最终一次性获得 96 个高质量的全长 COI 条形码。HIFI-Barcode 方法可兼容多种测序平台及三种分析流程，为 DNA 条形码研究提供了可

选择的技术流程，拥有高准确性、低成本、多选择等优点，极大的提高了 DNA 条形码获取研究的效率。

关键词：COI 条形码，高通量测序，系统发育学

研究背景

在过去十年里，随着高通量测序技术的发展，我们见证了生物多样性研究中方法学和应用的巨大转变，例如采用标准脱氧核醣核酸 (Deoxyribonucleic acid，DNA) 序列用于快速准确地物种鉴定；利用高通量测序技术来分析复杂的环境样品 (如混合样品、环境 DNA (eDNA)、无脊椎动物来源的 DNA (iDNA) 等)。保存有规范记录信息的 DNA 条形码参考数据库也在全球科研人员的共同努力下逐渐建立起来。生命条形码数据库 BOLD (The Barcode of Life Data systems, http://v4.boldsystems.org/) 已经拥有了大约八百万条条形码序列，覆盖了包含动物、植物、真菌等生物在内的约 30 万种物种 (2020 年 2 月截止)。丰富的参考条形码数据为物种鉴定、系统进化关系的构建、种间交互作用和群落结构的研究，以及加深对生物多样性的理解提供了坚实的基础。

全球条形码参考数据库已经为多个生态系统的研究提供了重要帮助。早期的生态及生物多样性研究经常利用 Sanger 测序方法来进行条形码测序，随着高通量测序技术 (High-Throughput Sequencing Technologies，HTS) 的推广，DNA 宏条形码和线粒体宏基因组方法也被越来越多的研究者接受并使用。但是当研究某些新的环境样品时，条形码数据库中参考数据的缺乏仍会使得基于 HTS 的宏条形码组学研究遭遇瓶颈：得到的序列不能被有效分配到具体的物种，从而无法反应真实的生物多样性组成，使研究者难以深入了解环境的生态关系。

以昆虫 CO1 条形码为例，标准条形码获取成本较十多年前显著降低。在不包括样品收集及处理的成本情况下，传统的 Sanger 测序的平均生产成本约 10 美元，如果要构建 1 亿个样品的条形码数据集，则需要 10 亿美元。人类基因组在最初构建时，大约花费了 30 亿美元，但随着高通量技术的发展，目前一个基因组的测序成本已经降低为不到 600 美元。

先前基于高通量测序技术获取 DNA 条形码的方法也存在着各自的优缺点 (图 1)，比如：基于罗氏 454 测序平台构建单样品条形码的研究可以通过拼接获得 CO1 全长条形码，但成本较高；测序读长较短的部分 Illumina 测序平台也逐渐用于条形码数据的获

取，有研究者使用长度为 313 碱基对 (base pair，bp) 的 CO1 条形码进行分析，该方法虽然可以一次性测序获得全长序列，但由于条形码长度较短，因此特异性会显著降低；也有研究者利用两次聚合酶链式反应 (polymerase chain reaction，PCR) 增加标签序列和测序引物，继而通过测序来直接获取全长条形码，但相关操作比较繁琐。在基因组领域，短数据可以利用位置关系被拼接为准确的长片段；根据类似的原理，所在项目组已经专门开发过针对条形码的组装算法，可以被用于组装混合样品中的全长条形码。根据当时的研究背景，我们开发了基于 Illumina Hiseq 平台和 PacBio 平台的 HIFI-Barcode 方法，可以通过编码引物的方法，得到带有特定标签序列的 PCR 产物，从而实现一次性获得 96 个样品的 DNA 条形码的目标。该方法发表于 Liu *et al.*, 2017。HIFI-Barcode 方法有很高的准确性和效率，但是涉及到较为复杂的分析过程，所以我们认为基于高通量的条形码技术仍有提升的空间。随着国产测序仪 MGISEQ 2000 推出单端 400 bp 测序的模块，我们进一步在此平台上开发出了 HIFI-SE 的方法，可以不需要进行打断，在其后的数据分析阶段，只需要进行扩增子两端序列的简单拼接，就可得到完整的 COI 条形码序列。该方法发表于 Yang *et al.*, 2020。

以上三种方法相互独立，测序及分析策略不同，但是在前期样品准备，DNA 提取和 PCR 操作等方面基本一致，所以在此操作手册中，我们将三种方法的实施步骤及相关比较纳入其中，方便研究者选择合适的建库测序平台和后续的分析方法。

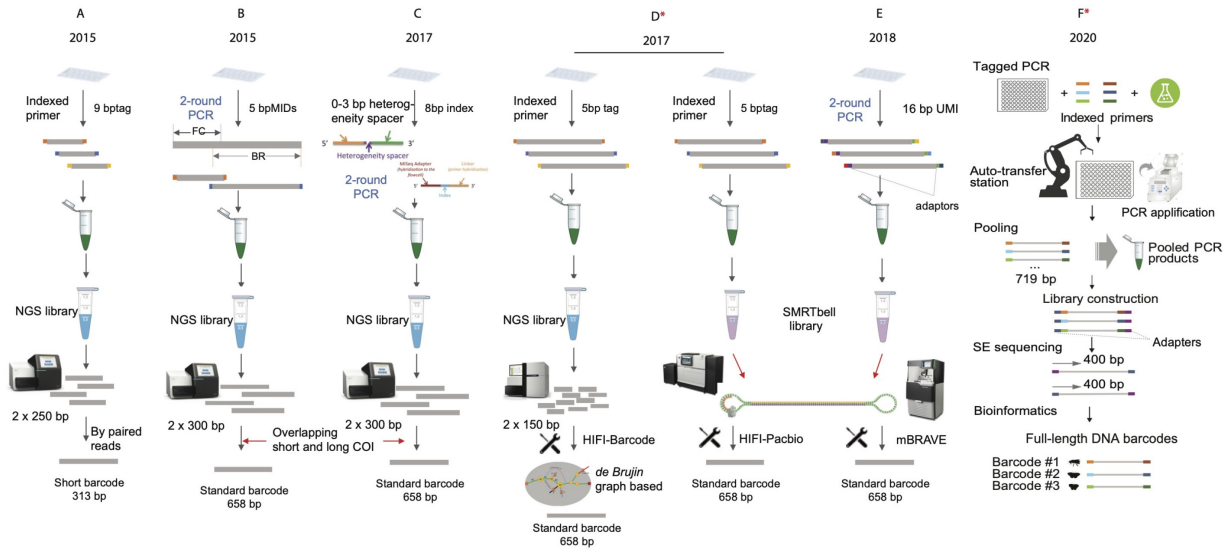

图 1. 高通量获取标准 DNA 条形码的策略比较，不同实验方案和测序策略 *代表本研究研制的方法策略。图片修改自 Yang *et al.*, 2020。A-E 分别代表的研究为 Meier *et al.*, 2016; Shokralla *et al.*, 2015; Cruaud *et al.*, 2017; Liu *et al.*, 2017; Hebert *et al.*, 2018; Yang *et al.*, 2020。

材料与试剂

1. 96 孔 PCR 板 (Axygen，PCR-96M2-HS-C)，1.5 ml 离心管 (Axygen)，2.5 μl，100 μl，200 μl，1 ml 枪头 (Axygen Filter Tips)
2. 96 方孔深孔板 (2 ml，PP MASTERBLOCK®, 96 Well)
3. 封口膜 (AxysealTM sealing film)
4. 琼脂糖 (NET WEIGHT, REGULAR AGAROSE G-10)
5. exTaq DNA 聚合酶 (TaKaRa Ex Taq™)
6. dNTP mix (TaKaRa Ex Taq™)
7. 10x primer buffer (TaKaRa Ex Taq™)
8. ddH_2O (广州誉维生物科技仪器有限公司 Unique 超纯水机制)
9. 合成引物 (上海生工)
10. 昆虫裂解液 Insect Lysis Buffer (见溶液配方)
11. 吸附缓冲液 Binding Buffer (见溶液配方)
12. 洗脱液 Wash Buffer (见溶液配方)
13. 吸附混合液 Binding Mix (见溶液配方)

注：对于 DNA 提取可根据自身实验室选择合适的方法和试剂主要试剂，本方法主要介绍玻璃纤维板 DNA 提取法 (Glass Fiber Plate method) 用昆虫腿提取 DNA 的方法。所用主要试剂见 10-13。

仪器设备

1. 0.2-2 μl, 2-20 μl, 20-200 μl, 100-1000 μl 移液枪 (Eppendrof)
2. 台式高速冷冻离心机 (Beckman，Allegra^{TM} 25R Centrifuge)
3. 水浴锅 (DK-8D 型，上海精宏实验设备有限公司) 或者恒温震荡仪 (MS-100 THERMO-SHAKER)，用于组织裂解

4. 96 孔 PCR 仪 (Thermo Fisher)
5. 电泳仪、电泳槽 (DYY-6C 型，北京市六一仪器厂)
6. 凝胶成像系统 (BIO-RAD)

软件

1. HIFI-barcode-hiseq，适用于 Illumina，MGISEQ 等二代测序平台 https://github.com/comery/HIFI-barcode-hiseq。
2. HIFI-barcode-pacbio，适用于 Pacbio 测序平台 https://github.com/comery/HIFI-barcode-pacbio。
3. HIFI-barcode-SE400，适用于 MGISEQ 2000 SE400 测序平台 https://github.com/comery/HIFI-barcode-SE400。

实验步骤

本实验流程包含了三种高通量条形码获取的流程，三种方法的测序前实验环节方法一致(对应步骤 1 至 3)，测序文库构建、测序及分析环节有差异。在完成 PCR 产物的混合后，可选择一种平台进行后续的建库测序。流程主要如图 1 所示，方法 1 HIFI-Barcode-Hiseq/MGISEQ 可选择平台及测序技术为 Hiseq 或 MGISEQ PE150；方法 2 HIFI-Barcode-Pacibio 可选择平台 Pacbio；方法 3 HIFI-Barcode-SE400 可选择 MGISEQ SE400。后续根据标签数目的增加，单条形码的测序成本还可以进一步降低。

表 1. 三种方法的多维度比较

比较项目	HIFI-Barcode-Hiseq/MGISEQ	HIFI-Barcode-Pacibio	HIFI-Barcode-SE400
测序成本	★ ★	★ ★ ★ ★ ★	★ ★ ★
测序错误率	★	★ ★ ★ ★	★ ★
建库复杂度	★ ★ ★	★ ★	★
数据分析复杂度	★ ★ ★ ★	★ ★	★ ★

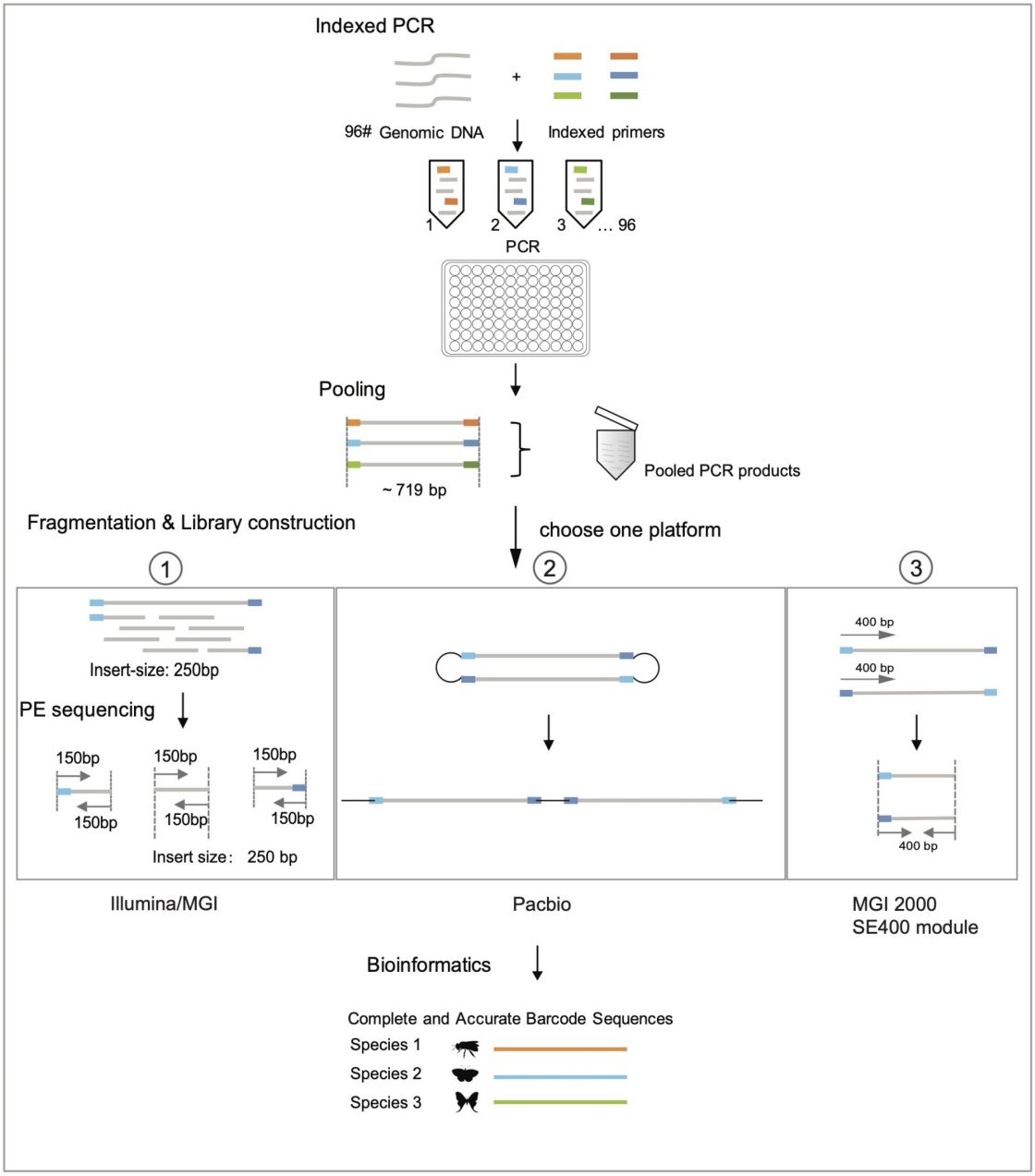

图 2. 实验操作及分析流程图

1. 动物全基因组 DNA 提取

1.1 简介

本流程采用利用玻璃纤维板 DNA 提取法 (Glass Fiber Plate method) 进行昆虫全基因组 DNA 的提取。该方法是加拿大 DNA 条形码中心提供的批量提取 DNA 的方法，研究者可以根据自身项目需求采取不同的基因组 DNA 提取方法。

1.2 主要流程

1) 获取昆虫组织样品：

a) 利用酒精清洁实验台，并准备好相关材料和试剂。

b) 在 96 孔板的每个孔中加入 30 μl 的无水乙醇防止昆虫组织因为静电蹦跳。

c) 利用干净的镊子夹取足量昆虫组织按照顺序放入 96 孔板中。

d) 取样后加上封盖密封，进行短暂离心；第五：离心后打开盖子，并 37 °C 加热蒸发酒精，获得用于提取 DNA 的昆虫组织样品。

2) 昆虫组织裂解：在带裙边的 96 孔板的每个孔中加入 50 μl 昆虫裂解液，并将 1) 中获取的组织样品依次添加到添加了裂解液的孔中，在 56 °C 温浴 6 h 使样品充分裂解。

3) 裂解后，首先将板子进行离心 (转速为 1500 × *g*, 15 s) 以移除盖子上的冷凝水。然后，加 100 μl 的 binding mix 溶液到每个孔中，1000 × *g* 震荡离心 20 s。

4) 换板：移除封盖，将板中溶液加入到玻璃纤维板 (简称 GF 板) 中，GF 板放在方形盒上，用密封膜封盖 GF 板。

5) 吸附：7000 × *g* 离心 5 min，基因组 DNA 将会被吸附在 GF 板的玻璃纤维膜 (简称 GF 膜) 上。

6) 两次洗脱：第一次：每孔加入 180 μl 的 PWB 洗脱液，用新密封膜密封，5000 × *g* 离心 2 min。第二次：每孔加入 750 μl 的 WB 洗脱液，重新用密封膜密封，5000 × *g* 离心 5 min。

7) 去除残留酒精：揭开封盖后，将 GF 板放置在 56 °C 中温浴 30 min。

8) 孵育：在 GF 板下放置收集板；56 °C 预热 ddH_2O 后添加 45 μl 到每个孔的 GF 膜上，室温孵育 1 min，并密封。

9) 收集 DNA：封装好的板子放到干净的方孔板并 5000 × *g* 离心 5 min；移走并丢弃 GF 板；DNA 已经提取并收集好。

2. 条形码序列扩增与检测

2.1 引物标签设计

1) 标签 (index) 设计考虑的因素：由于需要使用高通量测序技术对昆虫条形码 COI 序列进行测序和分析，因此在设计引物标签时需要综合考虑以下因

素：a) 标签可识别性：标签差异碱基数目、标签数量等；b) 成本相关因素：测序读长、序列长度、标签长度；c) 扩增效率：是否会影响扩增效率，是否容易造成非特异性扩增等。

2) 标签设计：首先，选定标签长度：利用 Barcrawl 程序 (v.30May2012) 生成 5-10 nt 长度的标签若干，在保证碱基差异大于或等于 2 情况下且标签数目大于或等于 96 条的情况下，对标签数目和长度进行比较，后选定为 5 nt 长的标签序列集合。对应命令行为"barcrawl -l 5 -m 2 -g 70 -a 5 -b 5 -p 5 -c 30 –nd"，命令行对应的主要条件为：两标签序列之间差异大于或等于 2，GC 含量在 30%-70%之间。共生成了 160 条标签，从生成的标签集合中随机选取 96 条作为后续研究使用的引物标签 (附录 A 表 6)。

2.2 标签引物合成

按照标签表 (附录 A 表 6) 设计出含标签序列的条形码引物：在标准引物的上 (下) 游的 5'端添加正向 (反向) 标签序列，形成新的带有标签序列的 96 对 COI 基因标准引物。加标签后的引物由上海生工有限公司合成。

标准引物序列为：

上游引物 LCO1490：5'-TAAACTTCAGGGTGACCAAAAAATCA-3'。

下游引物 HCO2198：5'-GGTCAACAAATCATAAAGATATTGG-3'。

例如：编号 001 的引物对序列为：

上游引物 LCO1490：5'-AAAGCTAAACTTCAGGGTGACCAAAAAATCA-3'。

下游引物 HCO2198：5'-AAAGCGGTCAACAAATCATAAAGATATTGG-3'。

2.3 PCR

1) 实验组：在 96 孔 PCR 板中依次加入带有标签的引物、昆虫 DNA 和扩增需要的试剂；每个孔包含 1 个 PCR 扩增体系 (表 2)。

表 2. PCR 体系

组分	每个 PCR 体系加入量
ddH_2O	16.2 μl
10 μM 正向标签引物	1 μl
10 μM 反向标签引物	1 μl

10x primer buffer	3 μl
dNTP mix	2.5 μl
exTaq	0.3 μl
DNA 样本或无菌水	1 μl (1 μl DNA 样本中的 DNA 含量约 200 ng)

2) 空白对照：95 个昆虫 DNA 分别对应 96 孔板的 95 个孔，剩余的 1 个作为空白对照，空白对照孔中需要将反应体系中的 1 μl DNA 替换为 1 μl ddH_2O，其余成分不变。

3) 反应条件：反应条件与常规 COI 条形码 PCR 反应一致，具体为：94 °C 预变性 1 min，然后进入 5 个循环：94 °C 变性 30 s、45 °C 退火 40 s、72 °C 延伸 1 min，5 个循环反应结束后进入 35 个循环：94 °C 变性 30 s、51 °C 退火 40 s、72 °C 延伸 1 min，循环结束后，72 °C 延伸 10 min，12 °C 保持。

2.4 电泳，分装与质量检测

1) 电泳：PCR 扩增完成后，电泳检测 PCR 扩增产物。

2) 分装：从 PCR 板的每个孔 (对应一个 PCR 反应产物 25 μl) 中吸取 5 μl 进行混合，获得一份 480 μl 的混合结果后，再将混合物均分为 5 份储存，每份包含 96 μl 的混合产物。

3) 分别选取一份扩增混合产物送去进行 DNA 质量检测，包括 Qubit 检测和电泳检测，并确认是否能够进行建库测序。

3. 高通量测序技术选择与测序

得到 PCR 产物之后，可以根据选择的方法对测序平台进行选择，然后进行相应的标准文库的构建和测序。

3.1 HIFI-Barcode-Hiseq/MGISEQ

PE150 测序。PCR 产物质量检测合格后，选取一份混合产物送至测序公司进行建库与测序。建库插入片段为 250 bp 小片段。建库后，使用 Illumina Hiseq 4000 (或 BGISEQ 系列平台) 进行双端测序，测序读长为 150 bp，数据量约 2-5 Gb。插入片段长度和测序读长为固定参数，不能随意修改。

3.2 HIFI-Barcode-PacBio

PacBio 测序。选取一份扩增产物混合液进行 PacBio RSII 平台质量检测、建库与测序，数据量约 2-5 Gb。

注：由于三代测序所需的 DNA 质量和总量都比较高，所以三代测序可能需要 3-4 管 (96 μl/管)。

3.3 HIFI-Barcode-SE400

SE400 测序。选取一份扩增产物混合液进行 MGI 2000 平台的 SE400 平台建库与测序，数据量约 2-5 Gb。

测序数据分析

1. 简介

针对相同的 PCR 产物，在不同的平台上进行建库测序，最终根据数据特点的不同，我们设计了对应的三种分析程序来还原 DNA 条形码序列。

2. 基于二代测序数据的条形码获取- HIFI-Barcode-Hiseq/MGISEQ

获取每个 PCR 板的二代测序数据后，分别对两份数据进行分析。首先进行原始数据的过滤，再进行条形码数据的组装与分析；条形码组装的详细流程见图 3。为了方便使用，已将流程打包为软件包 HIFI-barcode (https://github.com/comery/HIFI-barcode-hiseq)。

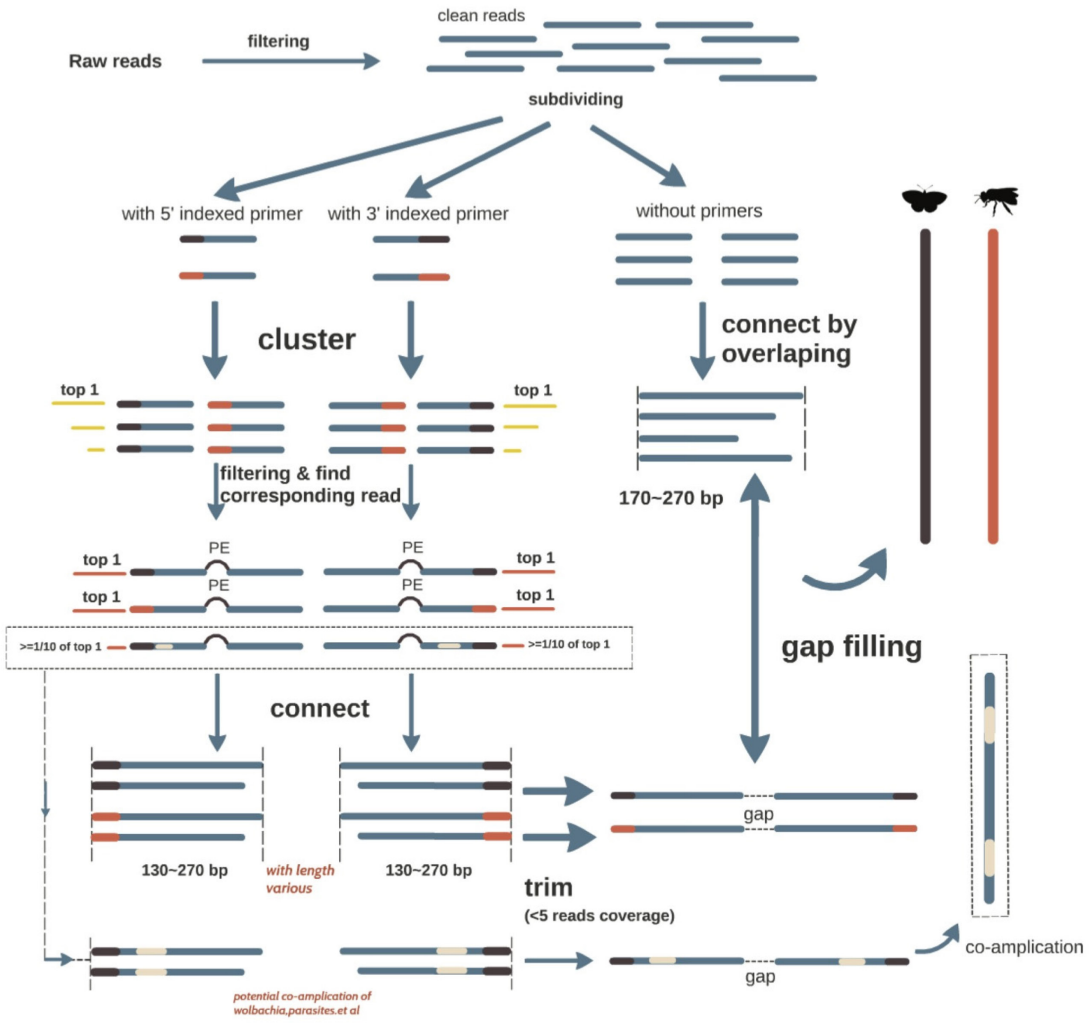

图 3. HIFI-Barcode 方法条形码组装流程 HIFI-Barcode 程序包含数据过滤和拆分、排序和聚类、双端拼接、间隙填充等步骤。

2.1 数据过滤

按照以下条件对原始测序数据进行质量控制和数据过滤：

1) 去除有接头污染的序列：至少能比对到接头 15 碱基且最多有 3 个错配。

2) 去除 N 数目大于 10 的序列。

3) 去除低质量 reads：30%的碱基质量值低于 33 的序列。

2.2 序列拆分和过滤

1) 拆分：由于 96 个样品分别对应一对已知的标签序列，因此可以结合测序序列双末端所包含的标签序列和引物序列的信息，将所有 reads 序列进行拆

分。可以分为三类：带有 96 种前端 (条形码 5'末端) 标签引物的 reads，带有 96 种尾端 (条形码 3'末端) 标签引物的 reads 和没有引物序列的 reads。

2) 去冗余：将拆分得到的含有标签引物的单端序列按照 98%的相似度聚类，减少冗余数据。

3) 丰度排序与过滤：根据 2) 的聚类结果将序列按照丰度高低进行排序，并记录相关丰度信息；根据丰度信息保留三类序列进行下游分析：a) 保留聚类结果中最高丰度的序列作为昆虫目标条形码序列进行后续的连接和组装；b) 与最高丰度序列差异大于 2%的序列也被保留，作为候选结果；c) 选择聚类结果中丰度不小于相同标签的最高丰度的 1/10 的序列，作为疑似共生物 (如潜在的寄生虫、沃尔巴克氏体感染或肠道微生物等) 条形码序列进行后续分析。

2.3 双端拼接

1) 拼接：首先，根据 reads 的双端信息把属于同一条片段的两端序列挑选出来；然后，将每一对 PE reads 按照以下条件拼接起来：a) 序列之间具有高于 95%的相似性重叠区；b) 重叠区域达到一定的重叠长度：前两类有引物信息的 reads 拼接长度需要在 130-270 bp 间，第三类无引物信息的 reads 拼接后的序列长度限制在 170-270 bp 之间。

2) 深度过滤：统计每种标签对应的拼接序列的各个碱基的深度，将覆盖深度不足 5 的碱基删除；从而获得可以代表标签对应的昆虫条形码前后两端的一对序列。

2.4 间隙 (gap) 填充

1) 输入数据：带标签的双端拼接结果作为两端信息输入，中间无标签的拼接结果统一作为中间信息输入。

2) 组装算法：采用 SOAPBarcode 算法 (图 4) 对条形码序列进行填充，从而获得高精度的完整的昆虫 COI 基因条形码。a) 每一对两端数据的前端为起始点，末端为终点，并用 kmer 构建 *de bruijin* 图，从起点到终点查找潜在的连接路径；b) 采用以下策略保证连接正确性：第一，删除在分叉处 kmer 丰度小于平均 kmer 丰度 10%的路径 (蓝色)；第二，如果经过第一步过滤后还有多条出度 (out degree，OD) 存在，则针对不同 ODs 和位于最后分

叉前的 kmer 间进行 reads 数目统计，后删除丰度小于平均 reads 丰度 10% 的 ODs (黑色)；第三，删除超出了预先设定的长度的路径 (红色)。

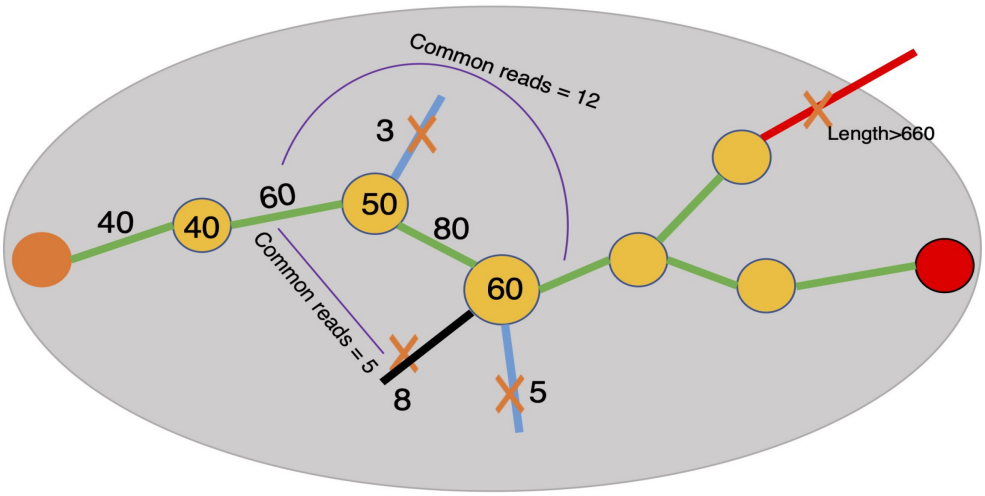

图 4. SOAPBarcode 组装算法 (引自 Liu *et al.*, 2013)

3) 组装结果输出：每个标签选取分值最好且长度与标准条形码相似的序列作为对应昆虫的条形码输出；根据前期丰度过滤后作为备选序列的双端序列所填充出来的条形码则作为非目标条形码输出。

命令行：

$ python3 HIFIBarcode.py all -outpre hifi -index 5 -q1 test1.fq.gz -q2 test2.fq.gz -primer indexed_primer.txt

注：“indexed_primer.txt”文件详见附录 B

综上，可获得结合了二代测序技术和生物信息分析技术得到的目标和非目标条形码。

3. 基于三代测序数据的条形码获取- HIFI-Barcode-PacBio

三代测序技术具有较长的读长，可以直接获得全长条形码序列，但由于三代测序技术的错误率较二代测序更高，因此需要对数据进行比较过滤后再获取全长条形码数据。分析方法如下：

1) 提取一致性循环序列 (consensus circular sequence，CCS)：从测序公司获得过滤后的 h5 格式文件后，直接使用 Pacbio 官方提供的流程 PacbioSmartAnalysis pipeline 将下机文件转换成 FASTA 格式。

2) 丰度过滤：三代测序结果中条形码全长可以被完全测通，且会产生多个循环，因此，本环节将循环数小于 15 次的序列直接过滤。

3) 引物序列定位：利用动态规划算法与通用引物的序列，定位到引物序列对应的位置，以此定位到条形码的双端；由于三代测序错误率较二代测序更高，我们在引物定位环节允许 2 个碱基的错配和 1 个碱基的插入或缺失。

4) 标签序列定位与样品匹配：将通用引物两端各自向外延伸 5 bp，获得标签序列。由于标签序列之间至少有 2 个差异，因此若标签序列中有连续 4 bp 或以上的序列能够匹配到某一样本对应的标签序列，则判定该序列可能为对应样品条形码。

5) 确定样品条形码：a) 由于存在 PCR 扩增错误和 PacBio 测序错误，因此需要选取循环数交高的序列；b) 循环数最高且能够成功翻译为蛋白的序列作为目标条形码序列输出；c) 其它输出的条形码作为非目标条形码输出。

命令行：

```
# step1, 从 H5 文件中提取 CCS 序列
$ source /path/PicBio/smrtanalysis/current/etc/setup.sh
$ fofnToSmrtpipeInput.py my_inputs.fofn > my_inputs.xml
$ smrtpipe.py --params=settings.xml xml:input.xml
# step2, 从 H5 文件中提取序列循环数信息
$ python bin/ccs_passes.py  data/*.ccs.h5 >ccs_passes.lst
# step3, 通过循环数信息过滤低质量序列
$ awk '$2>=15{print $1}' ccs_passes.lst >ccs_passes_15.lst
$ perl ./bin/fish_ccs.pl ccs_passes_15.lst
data/reads_of_insert.fasta >ccs_passes_15.fa
# step4, 根据引物序列拆分序列到对应样品
$ perl ./bin/1.primer_like_extract.pl -p experiment_data/primer.fa -index
experiment_data/index.xls -fa ccs_passes_15.fa -cm 2 -cg 1
# 通过聚类得到丰度最高的目标序列
$ cd 02.assignment/
$ perl ../bin/2.cluster_count_passes_length.pl -ccs ccs.successfully_assigned.fa -pattern
check.ccs_passes_15.fa.log -passes ../ccs_passes.lst
$ perl ../bin/change_name-location.pl cluster.top1.fas >hifi-barcode-pacbio.cluster.top1.fa
*具体输入文件格式见 github 页面：https://github.com/comery/HIFI-barcode-pacbio
```

4. 基于 SE400 测序数据的条形码获取-HIFI-Barcode-SE400

SE400 平台是数据可以应用 Python 包 HIFI-SE 来完成分析。

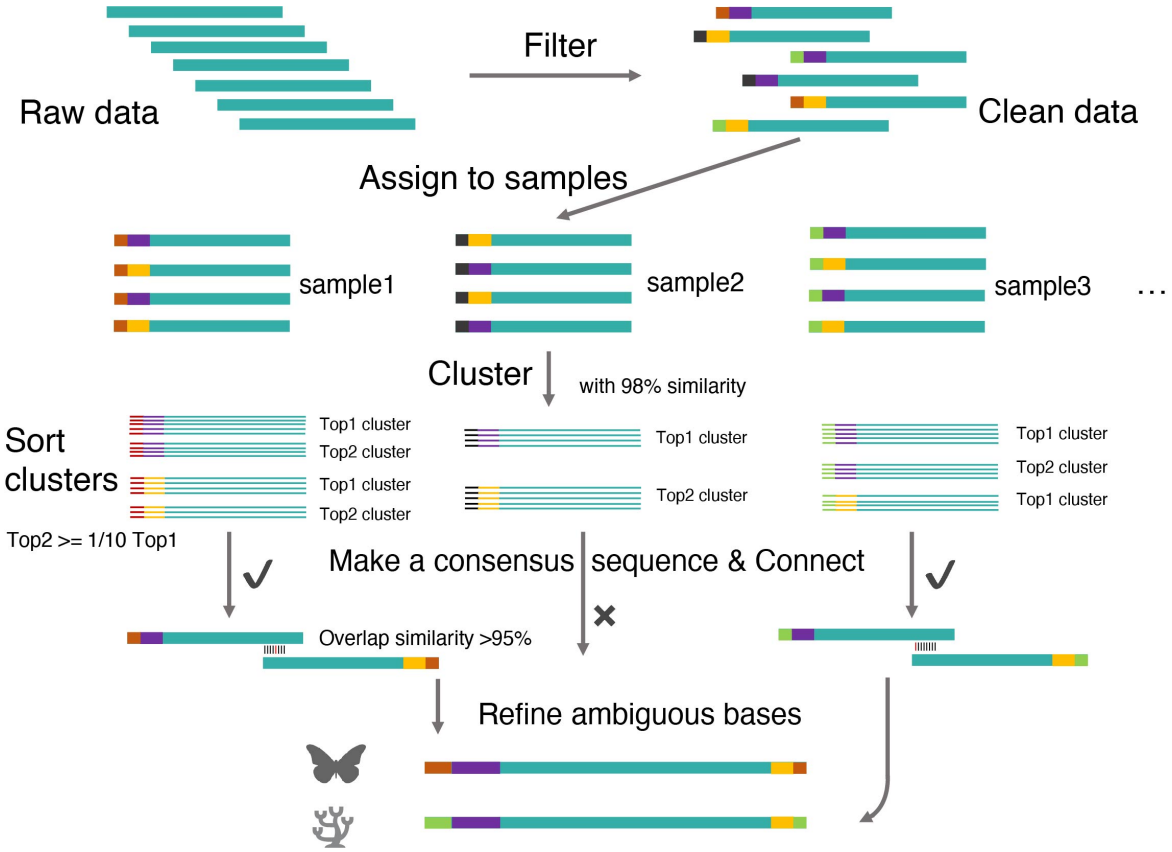

图 5. HIFI-SE 数据分析流程

1) 数据过滤，去掉包含 M 个 N 的序列，删除 expected error > 10 (expected error 定义见 Yang *et al.*, 2020)的序列，或者可以根据 quality 的参数设置阈值。
2) 序列拆分，由于 96 个样品分别对应一对已知的标签序列，因此可以结合测序序列双末端所包含的标签序列和引物序列的信息，将所有 reads 序列进行拆分。主要可以分为两类：带有 96 种前端 (条形码 5'末端) 标签引物的 reads，带有 96 种尾端 (条形码 3'末端) 标签引物的 reads。其他未分配的 reads 主要是因为测序错误引起的引物或标签序列的错配情况，或者是 PCR 产生的嵌合体。
3) 组装，可分为两种模式，cluster 模式和 consensus 模式。在 cluster 模式下，对应每个单独的样品，会对 5'和 3'的序列分别进行聚类，然后选取丰度最高的两个 cluster 序列，根据前后端的 overlap 信息将其拼接起来。而在第二种模式下，不对序列进行过滤，而是用所有序列直接生成 consensus 序列，用前后端分别得到的 consensus 序列直接拼接。一般来讲，如果一个样品中 target 物种

的丰度很高且没有污染，可以考虑用第二种模式进行组装，速度会比较快。但是如果孔增产物比较复杂时，建议使用第一种模式进行组装。

4) 物种鉴定，HIFI-SE 程序包含了一个 taxonomy 的程序可以允许在本地进行 COI 序列的物种鉴定，具体实现是将序列提交到 BOLD 服务器进行物种鉴定，然后在返回结果显示。可以极大的提高物种注释的效率。

命令行：

```
$ python3 HIFI-SE.py all -outpre hifi -trim -e 5 -raw test.raw.fastq -index 5 -primer index_primer.list -mode 1 -cid 0.98 -oid 0.95 -seqs_lim 50000 -threads 4 -tp 2
```

溶液配方

表 3. DNA 提取试剂、材料

试剂名/材料名	中文名	缩写
Disodium ethylenediamine tetraacetate • $2H_2O$	乙二胺四乙酸二水	EDTA
Ethyl alcohol (anhydrous)	无水乙醇	EtOH 96%
Guanidine thiocyanate	异硫氰酸胍	GuSCN
Molecular biology grade water	分子生物等级用水	ddH_2O
Polyethylene glycol sorbitan monolaurate	吐温-20	Tween-20
Proteinase K	蛋白酶 K	
Sodium chloride	氯化钠	NaCl
Sodium dodecyl sulfate	十二烷基硫酸钠	SDS
Sodium hydroxide	氢氧化钠	NaOH
t-Octylphenoxypolyethoxyethanol	叔辛基苯氧基聚乙烯乙氧基乙醇	Triton X-100
Tris(hydroxymethyl)aminometane	三 (羟甲基) 氨基甲苯	Trizma base
Tris(hydroxymethyl)aminometane hydrochloride	三 (羟甲基) 氨基甲烷盐酸 分	Trizma HCl
AcroPrepTM 96 1 ml filter plate with 3.0 µm Glass Fiber media over 0.2 µm Bio-Inert membrane, natural housing	PALL AcroPrep 96 孔滤板,1 ml，3.0 µm，玻璃纤维/0,2Bio-Inert 膜	PALL

试剂名/材料名	中文名	缩写
AxysealTM sealing film	封膜	self-adhering cover
Eppendorf® twin.tec 96-well microplates	96 孔板	microplate
PP MASTERBLOCK®, 96 Well, 2 ml	96 孔方孔盒	square-well block
SBS Receiver Plate Collar	PALL 离心适配圈	PALL collar
Others	其它常规仪器或材料：移液器和枪头、离心机、一次性手套等	

表 4. 预备液配置表

预备液名称	主要成分	含量	体积 (加 ddH_2O)
1 M Tris-HCl	Trizma® base Trizma® HCl	26.5 g 44.4 g	500 ml
1 M Tris-HCl	Trizma® base Trizma® HCl	9.7 g 66.1 g	500 ml
0.1 M Tris-HCl	Trizma® base	6.06 g	500 ml
1 M NaCl	NaCl	29.22 g	500 ml
0.5 M EDTA	EDTA NaOH	186.1 g 20.0 g	1000 ml
Proteinase K (20 mg/ml)	Proteinase K	100 mg	5 ml

表 5. DNA 提取试剂

混合液名称	成分	添加量	体积 (加 ddH_2O 后)
昆虫裂解液 Insect Lysis Buffer	GuSCN 0.5 M EDTA, pH 8.0 1 M Tris-HCl, pH 8.0 Triton X-100 Tween-20	16.5 g 12 ml 6 ml 1 ml 10 ml	200 ml
吸附缓冲液 Binding Buffer	GuSCN 0.5 M EDTA, pH 8.0	354.6 g 20 ml	500 ml

混合液名称	成分	添加量	体积 (加 ddH_2O 后)
	0.1 M Tris-HCl, pH 6.4	50 ml	
	Triton X-100	20 ml	
洗脱液 Wash Buffer	EtOH 96%	300 ml	475 ml
	1 M NaCl	23.75 ml	
	1 M Tris-HCl, pH 7.4	4.75 ml	
	0.5 M EDTA, pH 8.0	0.475 ml	
吸附混合液 Binding Mix	Binding Buffer	50 ml	100 ml
	EtOH 96%	50 ml	
	蛋白洗脱液	100 ml	
	Binding Buffer	26 ml	
	EtOH 96%	70 ml	

致谢

本研究获得科技部科技基础资源调查专项《中国东部传粉昆虫资源调查与评估》(2018FY100403) 以及深圳市科创委基金 (NO. JCYJ20170817150755701) 的资助。

竞争性利益声明

无经济或非经济性竞争性利益。

参考文献

1. Ivanova, N. V., Dewaard, J. R. and Hebert, P. D. N. (2006). An inexpensive, automation-friendly protocol for recovering high-quality DNA. *Mol Ecol Notes* 6(4): 998-1002.
2. Liu, S., Li, Y., Lu, J., Su, X., Tang, M., Zhang, R., Zhou, L., Zhou, C., Yang, Q., Ji, Y., Yu, D. W. and Zhou, X. (2013). SOAPBarcode: revealing arthropod biodiversity through assembly of Illumina shotgun sequences of PCR amplicons. *Methods Ecol Evol* 4(12): 1142-1150.
3. Liu, S., Yang, C., Zhou, C. and Zhou, X. (2017). Filling reference gaps via assembling DNA barcodes using high-throughput sequencing—moving toward barcoding the world. *GigaScience* 6(12): gix104.

4. Yang, C., Zheng, Y., Tan, S., Meng, G., Rao, W., Yang, C., Bourne, D. G., O'Brien, P. A., Xu, J., Liao,, S., Chen, A., Chen, X., Jia, X., Zhang, A. and Liu, S. (2020). Efficient COI barcoding using high throughput single-end 400 bp sequencing. *BMC Genomics* 21(1): 1-10.

附录

附录 A，5 bp 标签序列

表 6. 标签序列详情

正向 index	序列 (5'→3')	反向 index	序列 (5'→3')	正向 index	序列 (5'→3')	反向 index	序列 (5'→3')
For001	AAAGC	Rev001	AAAGC	For049	GACTT	Rev049	GACTT
For002	AACAG	Rev002	AACAG	For050	GAGAT	Rev050	GAGAT
For003	AACCT	Rev003	AACCT	For051	GAGTA	Rev051	GAGTA
For004	AACTC	Rev004	AACTC	For052	GATAG	Rev052	GATAG
For005	AAGCA	Rev005	AAGCA	For053	GATCT	Rev053	GATCT
For006	AAGGT	Rev006	AAGGT	For054	GATGA	Rev054	GATGA
For007	AAGTG	Rev007	AAGTG	For055	GATTC	Rev055	GATTC
For008	AATGG	Rev008	AATGG	For056	GCAAA	Rev056	GCAAA
For009	ACACT	Rev009	ACACT	For057	GCTAT	Rev057	GCTAT
For010	ACAGA	Rev010	ACAGA	For058	GCTTA	Rev058	GCTTA
For011	ACCAT	Rev011	ACCAT	For059	GGAAT	Rev059	GGAAT
For012	ACCTA	Rev012	ACCTA	For060	GGATA	Rev060	GGATA
For013	ACGAA	Rev013	ACGAA	For061	GGTTT	Rev061	GGTTT
For014	ACGTT	Rev014	ACGTT	For062	GTAGA	Rev062	GTAGA
For015	ACTGT	Rev015	ACTGT	For063	GTCAT	Rev063	GTCAT
For016	AGAAC	Rev016	AGAAC	For064	GTGAA	Rev064	GTGAA
For017	AGACA	Rev017	AGACA	For065	GTGTT	Rev065	GTGTT
For018	AGAGT	Rev018	AGAGT	For066	GTTAC	Rev066	GTTAC
For019	AGATG	Rev019	AGATG	For067	GTTCA	Rev067	GTTCA
For020	AGCTT	Rev020	AGCTT	For068	TAAGG	Rev068	TAAGG
For021	AGGAT	Rev021	AGGAT	For069	TACTG	Rev069	TACTG

For022	AGTAG	Rev022	AGTAG	For070	TAGGA	Rev070	TAGGA
For023	AGTCT	Rev023	AGTCT	For071	TAGTC	Rev071	TAGTC
For024	AGTGA	Rev024	AGTGA	For072	TATCG	Rev072	TATCG
For025	AGTTC	Rev025	AGTTC	For073	TATGC	Rev073	TATGC
For026	ATACC	Rev026	ATACC	For074	TCACA	Rev074	TCACA
For027	ATCAC	Rev027	ATCAC	For075	TCAGT	Rev075	TCAGT
For028	CAAAG	Rev028	CAAAG	For076	TCATG	Rev076	TCATG
For029	CAACT	Rev029	CAACT	For077	TCCAA	Rev077	TCCAA
For030	CAATC	Rev030	CAATC	For078	TCCTT	Rev078	TCCTT
For031	CAGAA	Rev031	CAGAA	For079	TCGAT	Rev079	TCGAT
For032	CATAC	Rev032	CATAC	For080	TCGTA	Rev080	TCGTA
For033	CATCA	Rev033	CATCA	For081	TCTCT	Rev081	TCTCT
For034	CCAAT	Rev034	CCAAT	For082	TGAAG	Rev082	TGAAG
For035	CGATT	Rev035	CGATT	For083	TGACT	Rev083	TGACT
For036	CGTAT	Rev036	CGTAT	For084	TGAGA	Rev084	TGAGA
For037	CGTTA	Rev037	CGTTA	For085	TGCTA	Rev085	TGCTA
For038	CTAAC	Rev038	CTAAC	For086	TGGAA	Rev086	TGGAA
For039	CTACA	Rev039	CTACA	For087	TGTAC	Rev087	TGTAC
For040	CTATG	Rev040	CTATG	For088	TGTCA	Rev088	TGTCA
For041	CTCAA	Rev041	CTCAA	For089	TGTGT	Rev089	TGTGT
For042	CTGAT	Rev042	CTGAT	For090	TTACG	Rev090	TTACG
For043	CTGTA	Rev043	CTGTA	For091	TTAGC	Rev091	TTAGC
For044	CTTAG	Rev044	CTTAG	For092	TTCTC	Rev092	TTCTC
For045	CTTCT	Rev045	CTTCT	For093	TTGAC	Rev093	TTGAC
For046	GAAAC	Rev046	GAAAC	For094	TTGCA	Rev094	TTGCA
For047	GAACA	Rev047	GAACA	For095	TTGGT	Rev095	TTGGT
For048	GAATG	Rev048	GAATG	For096	TTTCC	Rev096	TTTCC

附录 B，indexed_primer.txt 文件的内容及格式要求：

Rev001AAAGCTAAACTTCAGGGTGACCAAAAAATCA

Rev002AACAGTAAACTTCAGGGTGACCAAAAAATCA

…

…

For095TTGGTGGTCAACAAATCATAAAGATATTGG

For096TTTCCGGTCAACAAATCATAAAGATATTGG