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

NOV 20, 2011, 38(11): 1705–1714

专论与综述

基于第二代测序技术的细菌基因组与转录组研究策略简介

刘万飞[1,2Δ] 王西亮[1,2Δ] 赵宇慧[1,2] 曾瀞瑶[1,2] 耿佳宁[1*] 胡松年[1*]

(1. 中国科学院北京基因组研究所基因组科学及信息重点实验室 北京 100029)
(2. 中国科学院研究生院 北京 100049)

摘 要: 随着基于第二代测序技术的细菌基因组与转录组研究越来越广泛, 选择合适的研究策略变得越来越重要。就基于第二代测序技术的细菌基因组和转录组研究策略进行综述, 并简要介绍细菌基因组和转录组研究中的机遇和挑战。综述细菌基因组与转录组研究的常规方法及步骤, 并简要地介绍存在的问题。细菌基因组和转录组研究策略为大多数细菌的研究提供了一个相对完整的研究路线, 同时也会促进其它领域的研究, 如生命形成、生物进化、基础代谢、疾病、药物等。

关键词: 细菌, 基因组, 转录组, 第二代测序技术

The brief introduction of research strategies for bacterial genome and transcriptome based on the next-generation sequencing technologies

LIU Wan-Fei[1,2Δ] WANG Xi-Liang[1,2Δ] ZHAO Yu-Hui[1,2] ZENG Jing-Yao[1,2]
GENG Jia-Ning[1*] HU Song-Nian[1*]

(1. *Key Laboratory of Genome Science and Information*, *Beijing Institute of Genomics*, *Chinese Academy of Sciences*, *Beijing* 100029, *China*)
(2. *Graduate University of Chinese Academy of Sciences*, *Beijing* 100049, *China*)

Abstract: Recently, many bacterial genome and transcriptome studies have been performed based on the next-generation sequencing technologies. Therefore, how to select the appropriate research strategies is becoming increasingly important. In this paper, we discussed the research strategies of bacterial genome and transcriptome based on the next-generation sequencing technologies, and stated the oppor-

基金项目: 中国科学院知识创新工程项目(No. KSCX2-EW-R-01-04)
*通讯作者: 耿佳宁: Tel: 86-10-82995362; ✉: gengjianing@big.ac.cn
胡松年: Tel: 86-10-82995362; ✉: husn@big.ac.cn
Δ共同第一作者
收稿日期: 2011-05-24; 接受日期: 2011-09-13

tunities and challenges in these fields briefly. We reviewed the conventional methods and procedures and presented the existing problems briefly for bacterial genome and transcriptome studies. The research strategies of bacterial genome and transcriptome provide a relatively complete pipeline for the majority of bacteria. Moreover, it will promote the research of other fields, such as the course of life, biological evolution, basal metabolism, disease and drugs.

Keywords: Bacteria, Genome, Transcriptome, The next-generation sequencing technologies

伴随着我国 1%人类基因组计划的完成, 基因组学及转录组学在我国取得了快速发展。以 454、Solexa 和 SOLiD 为代表的第二代测序技术的出现, 更使大规模应用基因组及转录组方法解决科学问题成为可能。与此同时, 遗传学的研究对象由少量基因及其功能转变为生物体的全基因组结构、基因功能、表观修饰、细胞调控等, 遗传学研究进入了基因组和后基因组时代。其中, 通过细菌基因组和转录组研究来揭示生命基本过程, 如生命形成、生物进化、基础代谢、疾病发生、药物靶点等, 成为生物学研究的重要手段。目前, 以高通量低成本为特点的第二代测序技术使得单个实验室或者研究组独立进行细菌基因组及转录组研究成为可能。然而, 如何将海量测序结果组装成一个完整的细菌基因组, 以及如何将 RNA 测序结果还原成细菌特定生理状态下的转录情况是科学家面临的主要问题。本文对现阶段基于第二代测序技术的细菌基因组及转录组研究策略进行综述, 旨在为细菌基因组及转录组研究提供帮助。

1 细菌基因组学研究

1.1 细菌基因组学简介

细菌基因组学是研究细菌全基因组 DNA 序列及其结构与功能的学科。1995 年, 科学家获得了流感嗜血杆菌(*Haemophilus influenzae* Rd)的全基因组序列[1], 这是第一个完整的基因组序列, 也是第一个完成的细菌基因组序列。紧接着古细菌詹氏甲烷球菌(*Methanococcus jannaschii*)基因组[2]、大肠杆菌(*Escherichia coli* K-12)基因组[3]等也相继完成。细菌基因组研究不仅有利于研究细菌的基本生命过程, 同时也对高等真核生物的基因组学及后基因组学研究提供了参考和平台。到目前为止, NCBI 上记录了 1 534 个细菌基因组, 包括了 103 个古细菌和 1 431 个真细菌(2011-4-24)[4], 其中中国科学家完成了 44 个细菌基因组的测序工作。此外关于基因组学的研究报告也在逐年增加, 如表 1[5]所示。

表 1 PubMed 检索到的有关基因组学和转录组学的文献[5]
Table 1 The literature number retrieved in PubMed database using "Genomics" and "Transcriptomics" as keywords[5]

年份 Year	基因组学文章 Articles about Genomics	基因组学综述 Reviews about Genomics	转录组学文章 Articles about Transcriptomics	转录组学综述 Reviews about Transcriptomics
2000	1 543	383	2	0
2001	2 072	559	4	2
2002	2 689	826	18	7
2003	3 570	984	23	13
2004	4 810	1 238	56	23
2005	6 109	1 417	88	52
2006	6 905	1 559	105	49
2007	7 295	1 613	133	55
2008	8 119	1 603	162	59
2009	8 551	1 579	234	75
2010	8 986	1 487	296	70

1.2 细菌基因组研究策略

细菌基因组的研究策略主要分为 DNA 的提取及测序、基因组组装、基因组完成(Genome finishing)、基因预测、基因注释和基因组比较分析六大部分, 如图 1 所示。

首先是 DNA 的提取及测序。DNA 提取时要保证 DNA 纯度, 同时要避免 DNA 污染。目前主要应用的 DNA 测序技术是以 Roche 公司的 454 (http://www.454.com/)、Illumina 公司的 Solexa (http://www.illumina.com/)和 ABI 公司的 SOLiD (http://www.appliedbiosystems.com.cn/)为代表的第二代测序技术[6−8]。与第一代测序技术相比, 第二代测序技术在测序成本和测序速度方面有了极大的提高。454 采用焦磷酸测序法, 平均读长 400 bp, 每个循环能产生 400−600 Mb 序列, 耗时 7.5 h 左右。经过最近的升级, 454 测序长度已能达到 1 kb, 是进行未知基因组测序(De novo genome sequencing)的理想平台。然而, 454 在处理单碱基重复序列(Homopolymer)时会因为荧光信号强度判断错误而引入核苷酸缺失或插入。此外, 454 因测序通量较小, 使测序费用相对较高。Solexa 采用边合成边测序的方法, Solexa GAIIx 序列长度能达到 100 bp, 每个反应能产生 100 Gb 左右的碱基序列, 耗时约 9 d。升级版的 HiSeq 2000 在测序速度和测序通量上都有很大的提高, 序列长度可达到 120 bp, 每个反应能产生约 600 Gb 的碱基序列。Solexa 的优势在于测序通量大, 成本低, 但因为序列相对较短, 会增加后继序列拼接组装等分析的难度和计算量。SOLiD 采用双碱基编码原理, 利用 DNA 连接酶在连接过程中进行测序, 读长 50 bp, SOLiD 4 测序仪每个反应能产生 100 Gb 左右的数据, 耗时 6−7 d。SOLiD 与 Solexa 一样, 由于读长短, 后继分析相对复杂; 但是因为测序通量也较大, 测序费用相对较低。SOLiD 产生的序列是由 0、1、2、3 组成的 Colorspace, 因此数据处理时需要特殊的处理。相较基因组测序而言, SOLiD 在转录组测序上具有很大的优势, 因为它可以获得 RNA 的正负链信息, 这对于转录组研究, 尤其是 Antisense RNA 分析, 具有重要意义。最近, 一款新的 DNA 测序方法——粒子流(Ion-torrent)半导体基因组测序方法问世 (http://www.iontorrent.com/)[9]。此方法不依赖于光信号, 通过直接测知基于模板的 DNA 聚合酶合成反应产生的离子来获得序列信息, 并且成本和通量的性价比较好。因此, 测序时应该根据各个测序平台的优缺点来选择合适的测序平台。比如细菌 De novo 基因组测序, 可以利用 454 或 454 与 Solexa 相结合的测序方法, 即能降低测序成本, 又便于进行序列拼接组装。另外, 为了方便基因组组装, 我们可以适当的构建 Pair-end 或 Mate-pair 文库, 通过利用配对 Reads 之间的关系来确定 Contigs 之间的关系。

第二步, 基因组组装。常用的软件有 Newbler[10]、AMOScmp[11]、Phred/Phrap/Consed[12−13]和 Velvet[14]等, 可以根据自己的数据选择合适的组装软件, 也可以结合多种方法获得较好的组装结果。

第三步, 基因组完成, 即确定组装获得的 Contigs 之间的连接顺序并修补 Gaps。可以按照以下几个步骤进行: 首先, 计算 Contigs 和基因组的平均 Reads 覆盖度, 通过 Contigs 与基因组平均 Reads 覆盖度的比较, 获得 Unique contigs 和 Repeat contigs 以及 Repeat contigs 的重复次数。这一阶段, 可以过滤掉那些覆盖度明显低于基因组平均 Reads 覆盖度的 Contigs (可能来自污染 DNA 序列)以及一些较短的 Contigs (如长度小于 500 bp 的 Contigs, 这些 Contigs 会在修补 Gaps 时被填补回去)。其次, 根据 Contigs 之间的 Reads 连接数来确定 Contigs 之间可能的连接顺序。这一步还可以通过一些间接的方法来定位 Contigs 之间的连接关系: (1) 通过 Contigs 与近缘物种基因组的比较(如 MAUVE[15]和 MUMmer[16]), 获得 Contigs 之间的连接顺序; (2) 利用基因在基因组上排列顺序的保守性, 根据远缘物种基因组上基因的排列顺序来确定 Contigs 之间的连接顺序(把这些基因定位到 Contigs 上); (3) 将 Contigs 与 NCBI 的 nr/nt 库进行序列比对, 根据已知序列定位 Contigs 之间的连接顺序; (4) 进行随机 PCR 扩增来确定 Contigs 之间的连接顺序; (5) 利用

Pair-end 或者 Mate-pair 文库中的配对 Reads 获得 Contigs 之间的连接关系。最后, 根据 Contigs 之间可能的连接顺序设计引物, 并进行 PCR 扩增与测序来验证 Contigs 之间的连接顺序和修补 Gaps, 从而获得完整的基因组序列。在这一阶段, 可以先确定 Unique contigs 之间的关系, 然后再把 Repeat contigs 放回对应的位置。

第四步, 基因预测。常用的蛋白质编码基因预测软件有 Glimmer[17]、GeneMarkS[18]和 Prodigal[19], 通常可以任选其中一款软件进行预测, 也可以结合多个软件以获得较好的预测结果。此外, ZCURVE 是基于 DNA 序列 Z curve 理论的蛋白质编码基因识别软件, 具有较高的基因起始位点预测准确性[20]; GS-Finder 是不依赖于 rRNA 序列的细菌基因组翻译起始位点识别软件, 能大大提高翻译起始位点预测的准确性[21]; OperonDB 是比较常用的操纵子预测软件, 可以用来预测共同转录的基因簇[22]。另外, 非蛋白质编码基因的预测也有较成熟的软件, 通常用 RNAmmer[23]预测 rRNA、tRNAscan-SE[24]预测 tRNA 以及 Rfam[25]预测 Small RNA 等(可以利用 Splitter[26]将基因组分成较小序列, 然后用 Rfam 寻找 Small RNA; 或者下载并安装 Rfam database, 在本地寻找 Small RNA)。

第五步, 基因注释。这一步通常要整合多个数据库, 如 NCBI 的 nr 库、InterPro[27]、COG[28]和 KEGG[29]等, 通过序列比对进行预测基因的注释。此外, 还可以利用一些特定功能的软件或者数据库进行相应的分析, 如用 SignalP[30]预测信号肽、TMHMM[31]预测跨膜结构、ISfinder[32]预测插入序列、VFDB 预测毒力因子[33]、Islander 数据库查询基因组岛[34]、MobilomeFINDER[35]和 IslandViewer[36]鉴定基因组岛、PAIDB 预测潜在的致病岛[37]、Repeat-match 预测基因组重复序列、Tandem repeat Finder[38]寻找串联重复序列、CRISPR finder[39]预测 CRISPR 序列、Phage-finder[40]寻找噬菌体序列、TCDB[41]注释膜转运蛋白、Ori-Finder[42]寻找复制起始位点、ARDB[43]鉴定和注释抗菌素抗性基因、ACLAME[44]注释可变遗传因子(Mobile genetic elements)和 TADB[45]数据库搜索 Type2 toxin-antitoxin 位点等。另外, 有些基因是生物体生存不可或缺的基因, 即必需基因, 它们是生命的基础。DEG[46]数据库收集了一些物种的必需基因, 也可以用于注释必需基因, 这些必需基因是很好的抗菌药物靶基因。注释结束后, 对基因注释结果进行检查, 比如基因之间是否有 Overlap、是否存在假基因等, 可以利用 Mciobial Genome Submission Check[47]程序进行检查。

最后, 基因组比较分析。获得完整基因组及其注释后, 通常会进行相近物种之间或同一物种不同株之间的基因组比较分析。常用的细菌基因组比较分析软件和数据库有 ACT[48]、Mauve[15]、MUMmer[16]、MicrobesOnline[49]、mGenomeSubtractor[50]和 xBASE[51]等。ACT (Artemis Comparison Tool), 是一款进行基因组及其注释之间比较的可视化软件, 支持多种输入格式(EMBL, GenBank, FASTA 和 GFF 格式), 可以用来鉴定相似序列、插入、缺失、重排等。另外, 对于未注释序列, 可以寻找 CDS 序列并进行相应的比较。目前, 有两款基于 ACT 的网站, WebACT[52]和 DoubleACT (http://www.hpa-bioinfotools. org.uk/pise/double_act.html)。WebACT 提供了已知物种的基因组比较结果, 同时支持上传序列进行在线比较分析, 而 DoubleACT 提供生成 ACT 可读的基因组比较文件。Mauve 可以非常有效地构建多基因组比对结果, 而且可以容忍基因组重排和倒置, 同时 Mauve 会根据比对结果绘制不同基因组之间的进化树。MUMmer 可以用来比较完整的或者不完整的基因组序列(如 Contigs), 既可以在 DNA 水平比较, 也可以在蛋白质水平比较, 而且还可用于高等真核生物基因组之间的比较。另外, ACT 可以利用 MUMmer 的比较结果进行可视化操作。MicrobesOnline 是一个提供原核生物比较和功能基因组分析的数据库, 包含了 1 000 多个基因组和许多物种的芯片表达数据。MicrobesOnline 包含的模块有比较基因组浏览器、基因调控预测、系统发生搜索、代谢途径比较、操纵子预测、序列分析以及和其他微生物基因组资源的整合分析等。mGenomeSubtractor 是通过相近物种

基因组的在线比较分析得到保守的和特异性的基因组片段, 从而获得一个特定细菌的表型、环境适应和疾病相关等信息。xBASE 是一个细菌基因组比较注释软件, 只要提供完整或不完整的Fasta格式基因组序列, 就能获得基因组注释结果。xBASE 特别适合那些有参考基因组的基因组序列。另外, xBASE 开发了针对第二代测序数据的 xBASE-NG 模块, 可以进行 SNP 分析、新序列挖掘和基因组注释等。

通过细菌基因组比较分析有助于阐明微生物的许多特性(如耐高温、耐盐、降解塑料和抗药性等), 许多研究成果可应用于工业生产(如发酵)、环境治理(如分解石油)以及医药(如抗生素)等方面。当前对致病菌的研究主要集中于毒力因子、致病岛、耐药基因、耐药机制以及与寄主的关系等, 通过比较可以发现或预测致病菌的致病相关基因, 为疾病的诊断、药靶的寻找和疫苗或抗生素的研制提供理论支持。通过基因组比较还可以研究进化史、构建生物进化树。CVTree[53]是一个不基于序列比对的系统发生树分析工具。它是利用组成向量法(Composition vector)构建全基因组的系统发生树, 这样即避免了因为基因选择而带来的系统发生树的可变性, 又避免了不同长度和内容基因之间的序列比对。

2 细菌转录组学研究

2.1 细菌转录组学简介

细菌转录组学是从 RNA 水平研究基因表达的情况。目前细菌转录组学主要应用于基因注释校正、基因表达、操纵子鉴定、转录起始位点(TSS)鉴定、新基因鉴定、Small RNA 分析等内容。此外, 通过细菌转录组比较分析, 研究细菌在不同环境、宿主等情况下的基因表达变化, 从而阐述宿主、环境对细菌的影响、细菌致病机制等。目前转录组学的研

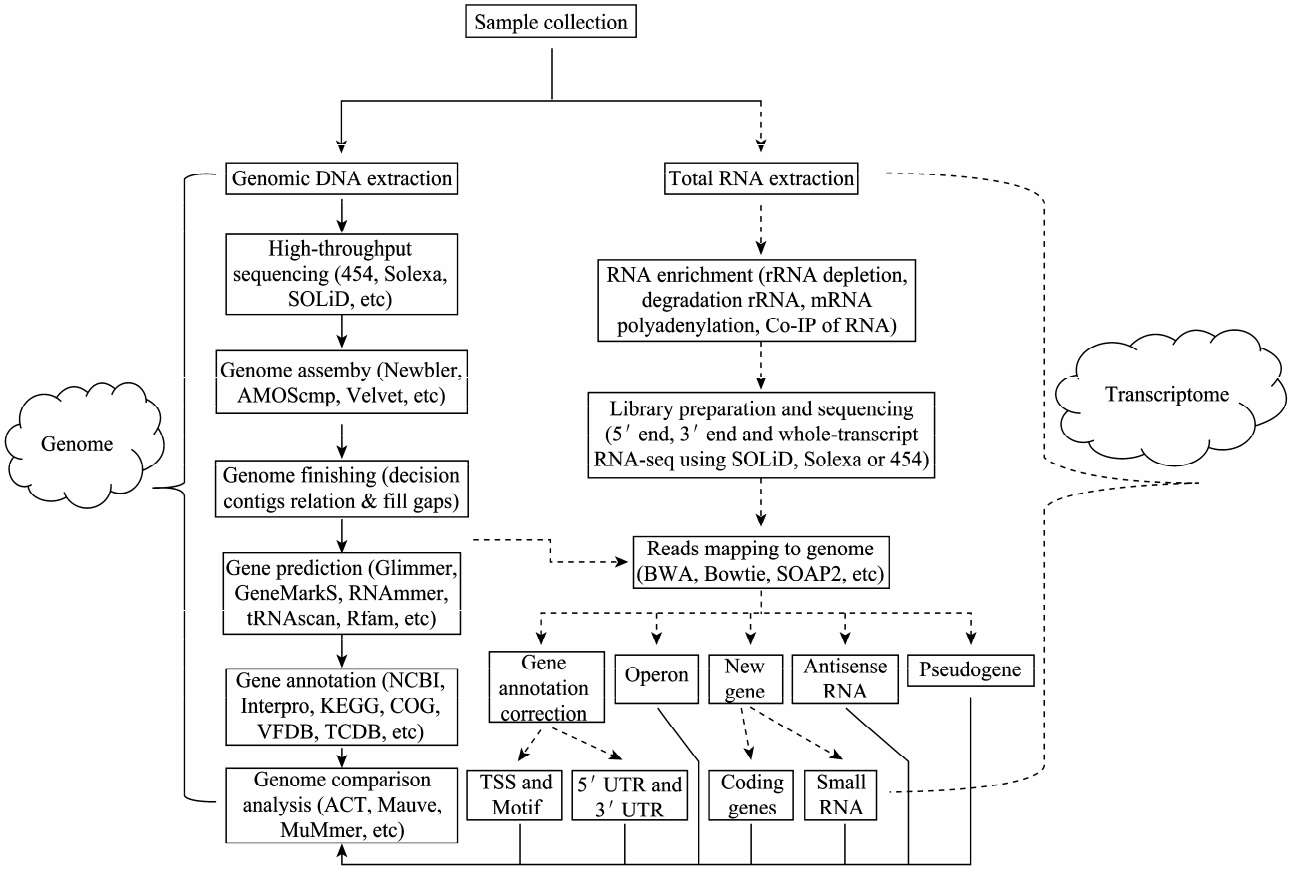

图 1 基于第二代测序技术的细菌基因组与转录组研究策略图

Fig. 1 The research strategies of bacteria genome and transcriptome based on the next-generation sequencing technologies

究也在逐年增加, 如表 1 所示, 但是相对于基因组学来说, 细菌转录组学尚处于发展阶段。

2.2 细菌转录组研究策略

细菌转录组的研究策略主要包括总 RNA 的提取和目的 RNA 的富集、RNA 建库及测序、Reads mapping 和后继分析等(图 1)。

首先进行总 RNA 的提取和目的 RNA 的富集。细菌 mRNA 的半衰期很短, 又极易降解, 因此防止其降解是十分重要的。仪器用具都要经过严格的处理, 最好是 RNA 实验专用; 材料一定要新鲜, 切忌使用反复冻融的材料; 整个 RNA 提取过程中动作应迅速, 并需要加入 RNA 酶抑制剂。细菌 mRNA 一般只占总 RNA 的 1%−5%, 因此, 总 RNA 提取完成后一般要进行 mRNA 的富集, 主要有以下 4 种方法[54]: (1) rRNA 捕获法。根据 16S rRNA 和 23S rRNA 保守区域序列设计探针并将探针固定在磁珠上, 利用探针和 rRNA 杂交去除 rRNA。在一般情况下该方法可以去掉大部分 rRNA, 但效果因不同的基因组而异。(2) 降解 rRNA 和 tRNA 法。在原核生物中 rRNA 和 tRNA 等加工过的 RNA 含有 5′单磷酸(5′P)而 mRNA 含有 5′三磷酸(5′PPP), 利用特异性降解 5′单磷酸 RNA 分子的核酸外切酶(5′→3′)降解 rRNA 和 tRNA。该方法一般仅可以去掉细菌 10%−20%的 rRNA。(3) 多腺苷酸 mRNA 选择法。利用大肠杆菌 polyA 聚合酶能够在 mRNA 末尾添加 polyA 而不能在 rRNA 末尾添加 polyA 的特性, 在 mRNA 末尾加上 polyA, 然后利用 Oligo (dT)探针捕获处理后的 mRNA。该方法简单快速, 但 mRNA 中会有初级转录本的加工产物和未翻译的 mRNA 降解产物。(4) 抗体法。通过抗体捕获与特定蛋白质相互作用的 RNA, 比如 Small RNA 可以和 Hfq 蛋白质相互作用, 可以利用这一特性采取免疫共沉淀的方法获得 Small RNA。该方法特异性好, 效率高, 可用于 Small RNA 的分析。4 种 mRNA 富集方法各有特点, rRNA 捕获法需要预先知道 rRNA 序列; 多腺苷酸 mRNA 选择法依赖 polyA 聚合酶; 抗体法比较适合 Small RNA 等的分析; 而降解 rRNA 和 tRNA 法也依赖特异性核酸外切酶。

第二步, 选择合适的建库方法和测序平台进行测序。测序文库包括 5′末端测序文库、3′末端测序文库、完整转录本测序文库等, 5′末端测序文库和 3′末端测序文库适合 UTR 区和操纵子的研究, 而完整转录本测序文库适合进行不同样本或者不同时期转录本之间表达量的比较分析以及新基因鉴定等。测序平台主要有 SOLiD、Solexa 和 454 三种。SOLiD 具有链特异性, 适合研究 Small RNA、发现新基因等, Solexa 测序成本较低, 而 454 适合进行基因组序列未知物种的转录组分析。大家要根据研究对象和研究目的选择合适的建库方法和测序平台, 比如鉴定转录起始位点应该构建 5′末端测序文库和采用链特异性测序方法(SOLiD), 而研究操纵子要构建完整转录本测序文库。需要说明的是链特异性测序方法越来越受欢迎。此方法可以明确转录本的方向、解决正负链基因重叠问题和提高基因(包括操纵子)注释的准确性, 特别适合鉴定转录起点、反义 RNA 和发现新基因。目前链特异性测序方法主要分为两类, 一类是在 mRNA 的两端分别连接不同的接头, 使 5′和 3′易于区分便于寻找转录模板; 另一类是对 cDNA 的一条链进行化学修饰产生标记, 如用重亚硫酸盐处理 RNA 或者利用 dUTP 来合成 cDNA 的第二条链[55]。

第三步, Reads mapping 和后继分析。首先将 Reads 进行 Mapping, 从而获得 Reads 在基因组上的位置, 常用的 Mapping 软件有 BWA[56]、Bowtie[57]、SOAP2[58]等。其次, 根据 Reads mapping 结果, 进行后继分析。TSS 鉴定: 根据基因组上的 Reads coverage 来鉴定转录起始位点; 5′UTR 和 3′UTR 鉴定: 鉴定编码基因中转录但不翻译的区域, 即 5′UTR 和 3′UTR; Operon 鉴定: 根据 TSS、Reads coverage 等鉴定 Operon (也就是细菌中的转录单元); 新基因鉴定: 找出之前没有注释但是表达的区域, 并根据是否存在 ORF 区分为蛋白质编码基因(Coding gene)和非蛋白质编码基因(Small RNA gene); Antisense RNA 鉴定: 根据 Small RNA 与其他基因的位置关系确定是否是 Antisense RNA; Pseudogenes 分析: 比如研究假基因的表达情况; 保守结构域的鉴定: 利用 MEME[59]等软件鉴定保守结构域; ncRNA 预测和鉴定: 可以利用 sRNAFinder[60]、nocoRNAc[61] 等软件

进行预测, 也可以通过与 Rfam database 比较来预测 Small RNA[25]。另外, 对那些比较重要的或者感兴趣的基因, 可以通过 Real-time PCR 来验证 RNA-seq 的结果。

此外, 为了方便科学家直观地了解细菌基因组及转录组图谱, 生物信息学家开发了许多基因组可视化软件, 比如 Artemis[62]和 Integrated Genome Browser (IGB)[63]等。

3 细菌基因组学与转录组学研究的机遇与挑战

3.1 细菌基因组学研究机遇与挑战

伴随着国际人类基因组计划的进行, 细菌基因组学也获得了快速的发展。随着第二代测序技术的出现, 细菌基因组学研究迎来了第二次高峰。目前, 细菌基因组测序多采用 454 或者 454 加 Solexa 的方式, 不但加快了数据产出, 而且有利于基因组拼接。获得细菌完整基因组后, 就要进行细菌基因组的分析和注释。常用的细菌基因组分析和注释工具如上文所述, 我们也可以参考 Pau Stothard 等的综述[64]。细菌的研究现在多集中于模式细菌(如大肠杆菌)和致病菌, 主要研究细菌的毒素、运动、粘附和生物膜形成、分泌系统、细胞表面蛋白、代谢及应激反应等[65]。此外, 通过相似物种基因组比较分析来揭示病原菌相关遗传线索[66], 也是细菌基因组学研究的一个重要方向。

虽然第二代测序技术给细菌基因组学研究带来了新的机遇, 但是也带来了一些新的问题, 比如基于焦磷酸测序的 454 测序方法常在单碱基重复序列区域出现插入/缺失, 会导致注释基因的移码突变; Solexa 测序法获得的 Reads 长度较短而影响拼接结果; 细菌基因组组装及分析流程较繁琐, 亟待新的方法或高度整合型的处理流程来加快分析过程等。

3.2 细菌转录组学研究机遇与挑战

转录组学是研究生物体基因表达和 RNA 调控的有力工具。伴随着第二代测序技术的出现, 在真核生物中进行了大量基于高通量测序技术的转录组学研究[67]。然而, 由于科学家普遍认为细菌转录组比较简单、细菌 mRNA 不易富集等原因, 细菌转录组学被大大忽视。随着测序能力的大幅度提高和一些特定细菌 mRNA 富集方法的出现[68−71], 细菌转录组越来越受到关注。

虽然细菌转录组学领域的研究目前还处于发展阶段, 但是它提供了在基因组水平上研究细菌 RNA 调控机制的重要手段。通过深度测序的转录组研究, 人们发现许多调控元件(如 Small RNA、Riboswitches 和 Cis-antisense 调控因子等)[72−76]参与了原核生物的生理和病理过程。此外, 有效的转录组分析可以改善基因组的注释, 转录组分析将来可能会成为基因组注释的一个标准构件。

目前原核转录组研究表明原核生物的调控复杂性和冗余性远远超过了最初的预料[54]。进一步的原核转录组研究可能揭示 Small RNA 调控网络、顺式作用元件(Cis-acting element)、环境依赖性功能开关(Riboswitch)和长反义转录本等的重要作用。转录组学方法的局限性在于它们需要成千上万的细胞作为材料, 这就无法确定正义转录本和反义转录本是互相排斥的还是可以同时转录的。因此, 单细胞转录组研究是转录组研究的新方向, 同时单细胞转录组将促进非培养细菌的研究和更精确地研究不同时间、不同环境下转录组的变化。此外, 随着转录组学研究的深入, 便捷的转录组学分析流程及软件会大大促进细菌转录组学的发展。

4 总结

细菌基因组学和转录组学的发展, 不但可以用来研究生命形成、生物进化、基础代谢、疾病发生、药物靶点等, 同时也可以相互促进各自学科的发展。比如基因组序列为转录组数据注释提供了参考, 同时转录组数据可以用来校正基因组注释信息、发现新基因和促进功能基因组学的发展。细菌基因组学和转录组学研究策略为大多数细菌研究提供了一个相对完整的研究路线, 同时也会促进单个实验室或者研究组进行细菌基因组及转录组的研究。此外, 细菌转录组学还需要科学家们进一步的深入研究, 并把细菌转录组分析变成为细菌的常规分析。

参 考 文 献

[1] Fleischmann RD, Adams MD, White O, et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd[J]. Science, 1995, 269(5223): 496−512.

[2] Bult CJ, White O, Olsen GJ, et al. Complete genome sequence of the methanogenic archaeon, Methanococcus jannaschii[J]. Science, 1996, 273(5278): 1058−1073.

[3] Blattner FR, Plunkett G 3rd, Bloch CA, et al. The complete genome sequence of *Escherichia coli* K-12[J]. Science, 1997, 277(5331): 1453−1462.

[4] NCBI: http: //www. ncbi. nlm. nih. gov/genomes/lproks. Cgi.

[5] Sayers EW, Barrett T, Benson DA, et al. Database resources of the National Center for Biotechnology Information[J]. Nucleic Acids Res, 2009, 37(Database issue): D5−D15.

[6] Margulies M, Egholm M, Altman WE, et al. Genome sequencing in microfabricated high-density picolitre reactors[J]. Nature, 2005, 437(7057): 376−380.

[7] Turcatti G, Romieu A, Fedurco M, et al. A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis[J]. Nucleic Acids Res, 2008, 36(4): e25.

[8] Shendure J, Porreca GJ, Reppas NB, et al. Accurate multiplex polony sequencing of an evolved bacterial genome[J]. Science, 2005, 309(5741): 1728−1732.

[9] Rothberg JM, Hinz W, Rearick TM, et al. An integrated semiconductor device enabling non-optical genome sequencing[J]. Nature, 2011, 475(7356): 348−352.

[10] 454 sequencing: http://454. com.

[11] Pop M, Phillippy A, Delcher AL, et al. Comparative genome assembly[J]. Briefings in Bioinformatics, 2004, 5(3): 237−248.

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

[13] Gordon D. Viewing and editing assembled sequences using Consed[J]. Curr Protoc Bioinformatics, 2003, Chapter 11.

[14] Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs[J]. Genome Research, 2008, 18(5): 821−829.

[15] Darling ACE, Mau B, Blattner FR, et al. Mauve: multiple alignment of conserved genomic sequence with rearrangements[J]. Genome Research, 2004, 14(7): 1394−1403.

[16] Kurtz S, Phillippy A, Delcher AL, et al. Versatile and open software for comparing large genomes[J]. Genome Biology, 2004, 5(2): R12.

[17] Delcher AL, Bratke KA, Powers EC, et al. Identifying bacterial genes and endosymbiont DNA with Glimmer[J]. Bioinformatics, 2007, 23(6): 673−679.

[18] Besemer J, Lomsadze A, Borodovsky M. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions[J]. Nucleic Acids Research, 2001, 29(12): 2607−2618.

[19] Hyatt D, Chen GL, Locascio PF, et al. Prodigal: prokaryotic gene recognition and translation initiation site identification[J]. BMC Bioinformatics, 2010, 11(1): 119.

[20] Guo FB, Ou HY, Zhang CT. ZCURVE: a new system for recognizing protein-coding genes in bacterial and archaeal genomes[J]. Nucleic Acids Res, 2003, 31(6): 1780−1789.

[21] Ou HY, Guo FB, Zhang CT. GS-Finder: a program to find bacterial gene start sites with a self-training method[J]. Int J Biochem Cell Biol, 2004, 36(3): 535−544.

[22] Ermolaeva MD, White O, Salzberg SL. Prediction of operons in microbial genomes[J]. Nucleic Acids Res, 2001, 29(5): 1216−1221.

[23] Lagesen K, Hallin P, Rødland EA, et al. RNAmmer: consistent and rapid annotation of ribosomal RNA genes[J]. Nucleic Acids Research, 2007, 35(9): 3100−3108.

[24] Schattner P, Brooks AN, Lowe TM. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs[J]. Nucleic Acids Research, 2005, 33(Issue suppl 2):W686−W689.

[25] Gardner PP, Daub J, Tate JG, et al. Rfam: updates to the RNA families database[J]. Nucleic Acids Research, 2009, 37(Database issue): D136−D140.

[26] splitter: http: //emboss. bioinformatics. nl/cgi−bin/emboss/ splitter.

[27] Hunter S, Apweiler R, Attwood TK, et al. InterPro: the integrative protein signature database[J]. Nucleic Acids Research, 2009, 37(Database issue): D211−D215.

[28] Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families[J]. Science, 1997, 278(5338): 631−637.

[29] KEGG: http: //www. genome. jp/kegg/.

[30] Bendtsen JD, Nielsen H, von Heijne G, et al. Improved prediction of signal peptides: SignalP 3. 0[J]. Journal of Molecular Biology, 2004, 340(4): 783−795.

[31] Krogh A, Larsson B, von Heijne G, et al. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes[J]. Journal of Molecular Biology, 2001, 305(3): 567−580.

[32] Siguier P, Perochon J, Lestrade L, et al. ISfinder: the reference centre for bacterial insertion sequences[J]. Nucleic Acids Research, 2006, 34(Database issue): D32−D36.

[33] Yang J, Chen LH, Sun LL, et al. VFDB 2008 release: an enhanced web-based resource for comparative pathogenomics[J]. Nucleic Acids Research, 2008, 36(Database issue): D539−D542.

[34] Mantri Y, Williams KP. Islander: a database of integrative islands in prokaryotic genomes, the associated integrases and their DNA site specificities[J]. Nucleic Acids Res, 2004, 32(Database issue): D55−D58.

[35] Ou HY, He XY, Harrison EM, et al. MobilomeFINDER: web-based tools for *in silico* and experimental discovery of bacterial genomic islands[J]. Nucleic Acids Res, 2007, 35(Web Server issue): W97−W104.

[36] Langille MGI, Brinkman FSL. IslandViewer: an integrated interface for computational identification and visualization of genomic islands[J]. Bioinformatics, 2009, 25(5): 664−665.

[37] Yoon SH, Park YK, Lee S, et al. Towards pathogenomics: a web-based resource for pathogenicity islands[J]. Nucleic Acids Res, 2007, 35(Database issue): D395−D400.

[38] Benson G. Tandem repeats finder: a program to analyze DNA sequences[J]. Nucleic Acids Research, 1999, 27(2): 573−580.

[39] Grissa I, Vergnaud G, Pourcel C. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats[J]. Nucleic Acids Research, 2007, 35(Web Server issue): W52−W57.

[40] Fouts DE. *Phage_Finder*: Automated identification and classification of prophage regions in complete bacterial genome sequences[J]. Nucleic Acids Research, 2006, 34(20): 5839−5851.

[41] Saier MH Jr, Yen MR, Noto K, et al. The Transporter Classification Database: recent advances[J]. Nucleic Acids Research, 2009, 37(Database issue): D274−D278.

[42] Gao F, Zhang CT. Ori-Finder: a web-based system for finding *oriCs* in unannotated bacterial genomes[J]. BMC Bioinformatics, 2008, 9: 79.

[43] Liu B, Pop M. ARDB-Antibiotic Resistance Genes Database[J]. Nucleic Acids Res, 2009, 37(Database issue): D443−447.

[44] Leplae R, Lima-Mendez G, Toussaint A. ACLAME: a CLAssification of Mobile genetic Elements, update 2010[J]. Nucleic Acids Res, 2010, 38(Database issue): D57−D61.

[45] Shao YC, Harrison EM, Bi DX, et al. TADB: a web-based resource for Type 2 toxin-antitoxin loci in bacteria and archaea[J]. Nucleic Acids Res, 2011, 39(Database issue): D606−D611.

[46] Zhang R, Lin Y. DEG 5. 0, a database of essential genes in both prokaryotes and eukaryotes[J]. Nucleic Acids Res, 2009, 37(Database issue): D455−D458.

[47] Microbial Genome Submission Check: http://www. ncbi. nlm.nih.gov/genomes/frameshifts/frameshifts.cgi.

[48] Carver TJ, Rutherford KM, Berriman M, et al. ACT: the Artemis Comparison Tool[J]. Bioinformatics, 2005, 21(16): 3422−3423.

[49] Dehal PS, Joachimiak MP, Price MN, et al. MicrobesOnline: an integrated portal for comparative and functional genomics[J]. Nucleic Acids Res, 2010, 38(Database issue): D396−D400.

[50] Shao YC, He XY, Harrison EM, et al. mGenomeSubtractor: a web-based tool for parallel *in silico* subtractive hybridization analysis of multiple bacterial genomes[J]. Nucleic Acids Res, 2010, 38(Web Server issue): W194−W200.

[51] Chaudhuri RR, Loman NJ, Snyder LAS, et al. xBASE2: a comprehensive resource for comparative bacterial genomics[J]. Nucleic Acids Res, 2008, 36(Database issue): D543−D546.

[52] Abbott JC, Aanensen DM, Rutherford K, et al. WebACT-an online companion for the Artemis Comparison Tool[J]. Bioinformatics, 2005, 21(18): 3665−3666.

[53] Xu Z, Hao BL. CVTree update: a newly designed phylogenetic study platform using composition vectors and whole genomes[J]. Nucleic Acids Res, 2009, 37(Web Server issue): W174−W178.

[54] Sorek R, Cossart P. Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity[J]. Nature Reviews Genetics, 2010, 11(1): 9−16.

[55] Levin JZ, Yassour M, Adiconis X, et al. Comprehensive comparative analysis of strand-specific RNA sequencing methods[J]. Nat Methods, 2010, 7(9): 709−715.

[56] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform[J]. Bioinformatics, 2009, 25(14): 1754−1760.

[57] Langmead B, Trapnell C, Pop M, et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome[J]. Genome Biology, 2009, 10(3): R25.

[58] Li RQ, Yu C, Li YR, et al. SOAP2: an improved ultrafast tool for short read alignment[J]. Bioinformatics, 2009, 25(15): 1966−1967.

[59] Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers // Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology. 1994, 2: 28−36.

[60] Tjaden B. Prediction of small, noncoding RNAs in bacteria using heterogeneous data[J]. J Math Biol, 2008, 56(1/2): 183−200.

[61] Herbig A, Nieselt K. nocoRNAc: characterization of non-coding RNAs in prokaryotes[J]. BMC Bioinformatics,

2011, 12: 40.
[62] Rutherford K, Parkhill J, Crook J, et al. Artemis: sequence visualization and annotation[J]. Bioinformatics, 2000, 16(10): 944−945.
[63] Nicol JW, Helt GA, Blanchard SG Jr, et al. The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets[J]. Bioinformatics, 2009, 25(20): 2730−2731.
[64] Stothard P, Wishart DS. Automated bacterial genome analysis and annotation[J]. Current Opinion in Microbiology, 2006, 9(5): 505−510.
[65] Duchaud E, Boussaha M, Loux V, et al. Complete genome sequence of the fish pathogen *Flavobacterium psychrophilum*[J]. Nature Biotechnology, 2007, 25(7): 763−769.
[66] Guzmán E, Romeu A, Garcia-Vallve S. Completely sequenced genomes of pathogenic bacteria: a review[J]. Enferm Infecc Microbiol Clin, 2008, 26(2): 88−98.
[67] Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics[J]. Nature Reviews Genetics, 2009, 10(1): 57−63.
[68] Yoder-Himes DR, Chain PSG, Zhu Y, et al. Mapping the Burkholderia cenocepacia niche response via high-throughput sequencing[J]. Proceedings of the National Academy of Sciences of the United States of America, 2009, 106(10): 3976−3981.
[69] Sharma CM, Hoffmann S, Darfeuille F, et al. The primary transcriptome of the major human pathogen *Helicobacter pylori*[J]. Nature, 2010, 464(7286): 250−255.
[70] Frias-Lopez J, Shi YM, Tyson GW, et al. Microbial community gene expression in ocean surface waters[J]. Proceedings of the National Academy of Sciences of the United States of America, 2008, 105(10): 3805−3810.
[71] Sittka A, Lucchini S, Papenfort K, et al. Deep sequencing analysis of small noncoding RNA and mRNA targets of the global post-transcriptional regulator, Hfq[J]. PLoS Genetics, 2008, 4(8): e1000163.
[72] Perkins TT, Kingsley RA, Fookes MC, et al. A strand-specific RNA-Seq analysis of the transcriptome of the typhoid bacillus *Salmonella typhi*[J]. PLoS Genet, 2009, 5(7): e1000569.
[73] Wurtzel O, Sapra R, Chen F, et al. A single-base resolution map of an archaeal transcriptome[J]. Genome Res, 20(1): 133−141.
[74] Güell M, van Noort V, Yus E, et al. Transcriptome complexity in a genome-reduced bacterium[J]. Science, 2009, 326(5957): 1268−1271.
[75] Toledo-Arana A, Dussurget O, Nikitas G, et al. The Listeria transcriptional landscape from saprophytism to virulence[J]. Nature, 2009, 459(7249): 950−956.
[76] Passalacqua KD, Varadarajan A, Ondov BD, et al. Structure and complexity of a bacterial transcriptome[J]. J Bacteriol, 2009, 191(10): 3203−3211.

稿件书写规范

论文中有关正、斜体的约定

物种的学名: 菌株的属名、种名(包括亚种、变种)用拉丁文斜体。属的首字母大写, 其余小写, 属以上用拉丁文正体。病毒一律用正体, 首字母大写。

限制性内切酶: 前 3 个字母用斜体, 后面的字母和编码正体平排, 例如: *Bam*H I、*Msp* I、*Sau*3A I 等。

氨基酸和碱基的缩写: 氨基酸缩写用 3 个字母表示时, 仅第一个字母大写, 其余小写, 正体。碱基缩写为大写正体。

基因符号用小写斜体, 蛋白质符号首字母大写, 用正体。