doi:10.3969/j.issn.1672-5565.2015.02.07

ABI PGM 测序平台用于细菌基因组 de novo 测序的评价

黄方亮
(浙江大学生命科学学院大型仪器平台,杭州 310058)

摘 要:为了探索加快细菌基因组研究的方法,利用 ABI PGM 测序平台测定了 1 株单细胞硫还原地杆菌的基因组序列。测序共获得 1.4 Gbp 数据,平均读长为 177 bp。通过多个拼接软件并采用合适的组装策略,得到一个完整细菌基因组 3.55 Mbp 和一条完整质粒序列 110 kbp。测定基因组序列与参考基因组 kn400 序列的相似性达到 94%,参考基因组 91%的基因能在测定基因组中找到相似基因。通过本研究表明采用 ABI PGM 测序平台结合灵活的拼接策略可快速构建细菌基因组精细图谱,为进一步的功能注释及深入的信息分析提供准确的数据,大大加快研究进程。

关键词:PGM 测序平台;细菌基因组测序
中图分类号:Q75 **文献标志码**:A **文章编号**:1672-5565(2015)-02-116-04

Evaluation of PGM sequencing platform using in bacterial genome de novo sequencing

HUANG Fangliang
(*Equipment and Technology Service Platform of College of Life Sciences Zhejiang university*, *Hangzhou* 310058, *China*)

Abstract: In order to speed up bacterial genome exploration, we performed the genome sequencing of Geobacter sulfurreducens using PGM. Totally, 1.4 Gbp raw data were obtained with an average read length of 177 bp. 2 contigs were assembled by multiple software calculations using appropriate assembly strategies. The size of whole obtained genome and plasmid was measured to be 3.55 Mbp and 110 kbp, respectively. The sequenced genome identified 94% of reference genome strain KN400 and 91% genes of KN400 were tested to be orthologous in the sequenced genome. This study proved that the use of ABI PGM sequencing platform with splicing flexible strategy can rapidly build bacteria genome map. By providing accurate data for the functional annotation and in-depth information analysis, it will greatly accelerate research progress.

Keywords: ABI PGM Sequencing Platform; Bacterial Genome de novo Sequencing

随着测序技术的迅速发展和测序成本的急速降低,细菌全基因组精细测序成为科学家研究目的细菌的基本要求[1]。2005 年罗氏 454 测序仪出现后,一次开机产生上百万条数据的高通量测序技术大大加快了基因组研究的进程[2],2012 年 454 测序仪发明人 Jonathan Rothberg 博士在焦磷酸测序[3]的基础上,发明了新一代测序仪 ABI PGM,它的测序通量更有弹性,能够使用 314、316、318 三种芯片,分别出 10 M,100 M,1 G 测序数据。用半导体检测技术替代了冷光 CCD 拍照成像技术检测 DNA 信号,测序成本更低,原始数据占用的计算机资源更少[4]。一张芯片上机测序只要 3 小时。利用 ABI PGM 318 芯片配合本来用于 5500 测序仪上的 mate pair 试剂盒,使 ABI PGM 测序平台成为细菌基因组精细测序的强大工具。

本研究中,我们希望快速得到目的菌株完整基因组序列。为此,构建了 200 bp 短片段文库和 3 KB mate pair 文库,接上不同的接头,使用 PGM 测序。得到的数据用 CLC Bio Genomics work bench 6.0 (CLC Bio, Aarhus, Denmark)软件拼接,采用合适的拼接策略后,两周左右就得到完整的目的细菌基因组精细图谱。

收稿日期:2015-01-13;**修回日期**:2015-03-12.
基金项目:浙江省教育厅科研项目(Y201328526)。
作者简介:黄方亮,男,博士,研究方向:细菌基因组测序;E-mail:huangfl@ zju.edu.cn.

1　材料与方法

1.1　菌株培养和核酸提取

单细胞硫还原地杆菌菌株由浙大热能所提供,挑取单克隆菌落,在 37 ℃下用改进过的 LB 液体培养基密闭振荡培养过夜。取 200 mL 菌液最高速离心 1 min,弃上清,将沉淀转入研钵,加液氮研磨,研磨充分后加入 1 mL Plant DNAzol ,2 μL 2-ME(β-巯基乙醇)继续研磨,转移裂解产物至 1.5 mL 离心管中。将离心管置 65 ℃水浴 30 min。加 750 μL 氯仿,混合均匀。12 000 rpm,离心 5 min。小心取上清(避免吸取中间蛋白层),转入一新的 1.5 mL 管(体积大约有600 μL)。加 0.7 体积的异丙醇(约 420 μL),12 000 rpm,离心 10 min。弃上清,加入 1 mL 75%乙醇至离心管中,颠倒数次以重悬 DNA,直立离心管 1 min 至 DNA 团块沉至管底,倾去或吸除洗涤液。细小的 DNA 沉淀团块容易在倾倒洗涤液时丢失,可室温3 000 rpm,离心 3 ~ 5 min,然后倾去或吸除洗涤液。重复清洗 1 次。最后简短离心,用枪头小心吸弃残留液体。室温静置数分钟(约 10 min)使残余乙醇挥发,注意不要完全晾干 DNA。加入适量(100 ~ 200 μL)灭菌双蒸水或 TE 缓冲液,使 DNA 沉淀溶解。向 DNA 溶液中加入终浓度为40 μg·mL^{-1} 的 RNase A,37 ℃孵育 30 min,-20 ℃保存。

1.2　基因组测序文库构建及 PGM 测序

取 200 ng 目的细菌基因组 DNA,用 millipore 水稀释到 50 μL 体积,放入 Biorupt,参数:Power Level:L ,Time ON:0.5 min,Time OFF:0.5 min,Number of 15-min Cycles:3。超声破碎到 250 bp 左右,用 Ion Xpress™ Plus Fragment Library Kit 构建 200 bp 左右测序文库。取 3 μg 基因组 DNA,用 millipore 水稀释到 150 μL 体积,利用 hydroshear 核酸片断化仪打断到 3 KB,参数:Standard Shearing Assembly,SC 13,20cycles。1%凝胶电泳回收纯化,使用 5500 SOLID MATE-PAIRED LIBRARY KIT 构建 3 KB mate-pair 文库。两个文库接不同的接头,上 PGM 测序,PGM 测序参照 ABI PGM 操作手册。

1.3　测序数据 de novo 拼接

将两个文库数据导入 CLC Bio Genomics work bench 6.0,用 trimed 功能去除低质量数据后,以 de novo 模块拼接。参数使用如下:word size values 范围是 25 ~ 40 核苷酸,bubble sizes 选择 50 bp, 60 bp, 70 bp三种,Map reads back to contig(slow):mismatch cost:2,insertion cost:2,deletion cost:3,length fraction:0.5,similarity fraction:0.8。将得到的最理想拼接结果做为参照序列,比对得到的两个文库数据,从而填补 scaffold 序列中的 gap,并根据落在两个不同 scaffold 上的成对 mate-pair 数据,确认 scaffold 间的关系。不同参数条件拼接出来的 contigs 重新 mapping 回拼好的 scaffold 上,消除 gap。拼接策略见图 1。

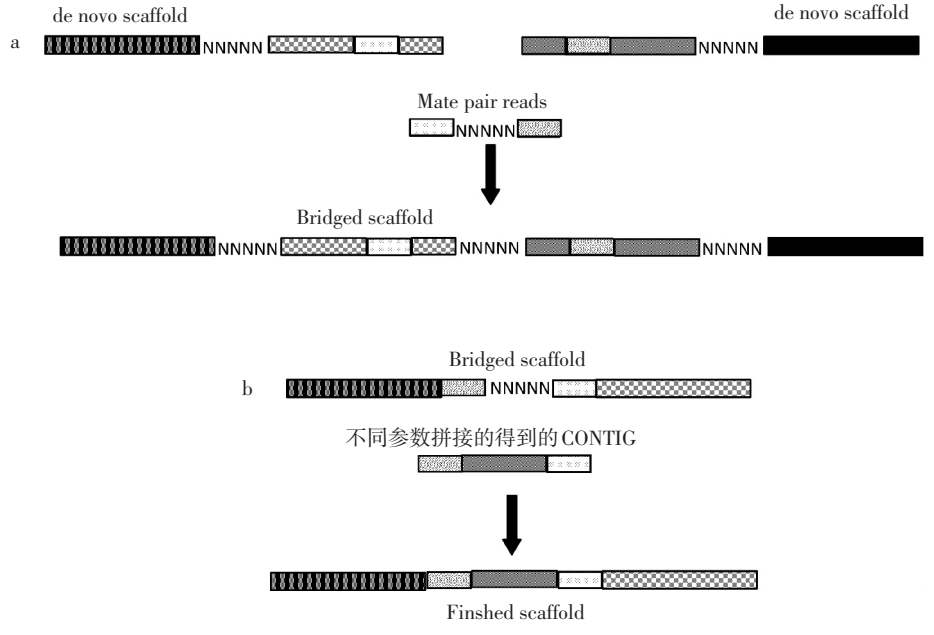

图 1　拼接策略示意图

Fig.1　Schematic diagram of assembly method

注:(a) 利用 3 KB mate pair 数据确定 scaffold 间关系;(b) 利用不同参数条件下得到的 contig 序列,填补 scaffold 中的 gap,得到完整序列。

Notes:(a) Scaffold ordering phase:using 3 KB mate pair data to determine the relationship between scaffolds; (b) Genome finish phase:fill gap by contig mapping.

1.4 基因组 FINISH

经 1.3 拼接后,得到成环的基因组序列,根据缺少的 gap,设计基于 gap 的引物。经 PCR 扩增后,利用一代测序仪 3130 的数据,补全序列,从而构建完整环状基因组。

1.5 基因预测注释分析

将基因组数据提交到 RAST(Rapid Annotation using Subsystem Technology)[5] 网站,得到 3 822 个预测基因。结合另外几个原核生物基因预测软件 Glimmer[6],Genemarker[7],FgeneSB[8] 校正预测结果。利用 RAST 网站 Compare 模块中的 function based 功能与其它基因组做功能比较。KEGG 模块看基因组中基因所在 pathway 信息。并与 InterPro[9],COG[10] 数据库比对确认预测基因生化代谢功能。对于非蛋白质编码基因 rRNA 和 tRNA 的预测,分别用 RNAmmer[11] 和 tRNAscanSE[12] 确认。

1.6 基因组比较分析

选取单细胞硫还原地杆菌生物型菌株 kn400[13] 做为参考序列,运用 NCBI 网站的 Blast2SEQ 软件比较两个基因组相似性。根据预测的基因,用 RAST 网站的 compare 基于 sequence based 查找参考基因组中的同源基因。

2 结 果

2.1 测序数据量和基因组拼接

两个文库共获得 8.1 M 条序列,1.4 Gbp 碱基,数据详情见表 1。将数据导入 CLC 分析软件,经过 trimed 后,还有 7.8 M 条序列可用,序列统计见图 2。经过多次 de novo 拼接,调整各种参数,最后 word size values 选 35, bubble sizes 选择 60 bp,组装成 16 个 scaffolds,总长 3.66 M,N50 为 492 k,最大长度 889 k。将 16 个 scaffolds 序列做为参照序列,把两个文库的数据 mapping 上去,找到 16 个 scaffolds 间的前后关系,并补上 scaffold 中的 gap。经过多轮的 mapping 最终将基因组拼接成一个环状染色体序列 3.55 M,并发现一个完整的质粒序列 110 KB,基因组 G+C 含量 61%。环状染色体序列中还有 4 个不能通过序列拼接确定的 gap,用 PRIMER3 在线软件设计引物[14],PCR 扩增测序后,拼回原来的位置得到一个完整的环状染色体序列,扩增产物电泳图见图 3。

表 1 200 bp 及 3 KB mate pair 文库数据统计情况

Table 1 200 bp and 3 KB mate pair sequencing library data

	碱基数(bp)	大于 Q20(bp)	序列数	平均读长(bp)	Trimed 后序列数	Trimed 后平均读长(bp)
200 bp 文库	574 924 448	477 804 821	3 242 186	177	3 031 287	180
3 KB mate pair 文库	469 488 963	402 844 367	4 967 970	71.4	4 799 204	67

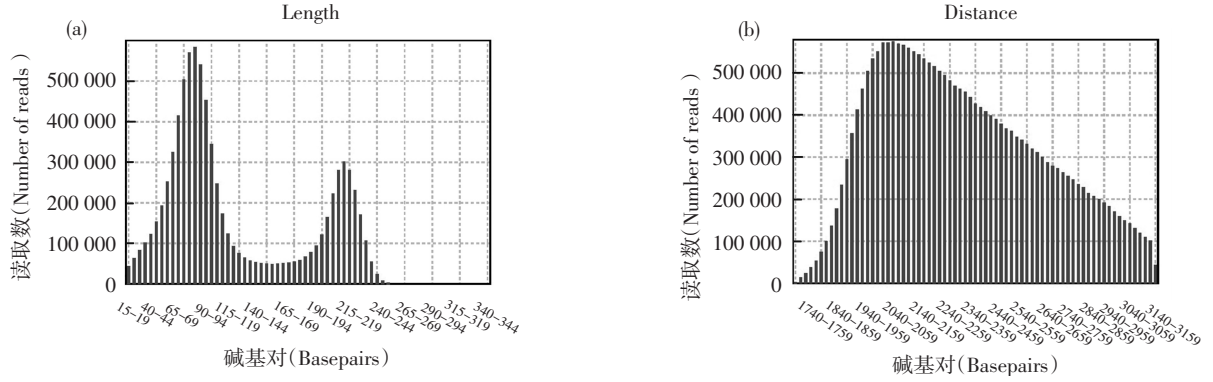

图 2 序列示意图

Fig.2 Schematic diagram of sequence

注:(a) 200 bp 文库和 3 KB mate pair 文库序列读长分布;(b) 3 KB mate pair 数据在基因组上的实际定位统计,峰值出现在 2.1 KB,范围在 1.7 KB~3.1 KB 间。

Notes:(a) Read length distribution of the 200 bp library and 3 KB mate pair library; (b) Distance of 3 KB mate pair library data locate in genome, peak appeared in the 2.1 KB, ranging from between 1.7 KB~3.1 KB.

2.2 与参考序列比较结果

选择基因组大小为 3.7 M 的 kn400 做为参考,进行基因组比对,结果显示两个基因组序列相似度 94%。参考基因组中 91%的基因能在测定基因组的预测基因中找到,相似度≥95%的基因占 52%,95%>相似度≥30%的基因占 39%。

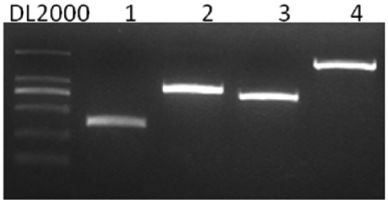

图 3　电泳图

Fig.3　Electrophoresis

注:1,2,3,4 分别是四个 gap PCR 产物电泳条带。

Notes:1~4 is PCR amplification products of 4 gaps.

3　讨　论

目前,得到细菌全基因组序列完整图谱已经是高质量细菌文章发表的必备条件。而很多时候科学家在高通量测序完成后,得到的是几十个独立的 scaffolds,要找到它们之间的关系,拼接成环状完整的基因组,还需要订购很多的引物,几个月的时间做 PCR 扩增,费时费力。采用 200 bp 文库加 3 KB mate pair 文库,用 PGM 318 芯片测序后,得到 1.4 G原始数据,经过高质量筛选后,余下 881 M 数据,覆盖基因组 266 倍左右,软件初步拼接得到 16 个 scaffolds。将 16 个 scaffolds 做为参考序列,把所有测序数据 mapping 上去,通过定位在两个不同 scaffolds 上的多个成对的 mate-pair 序列来确定 scaffolds 间的前后关系,也可以结合软件 SSPACE 来辅助确认 scaffolds 间的关系。确认关系排好顺序的 scaffolds 被拼接到一起,做为参考序列,再做 mapping,通过 mapping 结果可以进一步确认是否正确拼接 scaffolds。如此反复,直到拼接成环状序列。过程中可以结合 gap 修复软件 Gapfiller[15],SOAPdenovo GapCloser v1.12r6 来关闭 gaps[16]。可能是因为重复序列的关系,环状基因组中还是会有 4 个 gap 无法修复,最终通过设计引物 PCR 扩增,3130 测序,拼接出完整的基因组数据。拼接完成后还检测到一个完整的质粒序列。

PGM 测序平台还应用到了另外几个细菌基因组的研究中,都得到完整的细菌基因组图谱。但经过实验发现如果目的细菌中出现多个质粒,且质粒间的序列高度相似时,虽然可以得到完整的基因组数据,却很难保证得到完整的质粒序列。必须将质粒分离开单独测序才行。本研究实验结果证明 PGM 单次上机成本较低,一天就能完成两张 318 芯片测序,一张 318 芯片数据足够满足 4 M 左右细菌基因组的精细图拼接。因此采用 ABI PGM 测序平台结合合适的拼接软件,采用灵活的拼接策略可以快速构建细菌基因组精细图谱,为进一步的基因功能注释和深入的信息分析提供准确的数据,能够大大加快细菌基因组研究的进程。

参考文献(References)

[1] BARBOSA E G, ABURJAILE F F, RAMOS R T, et al. Value of a newly sequenced bacterial genome[J]. World J Biol Chem,2014, 5(2): 161-168.

[2] YANG Y, XIE B, YAN J. Application of next-generation sequencing technology in forensic science[J]. Genomics Proteomics Bioinformatics, 2014, 12(5): 190-197.

[3] RONAGHI M, UHLEN M, NYREN P. A sequencing method based on real-time pyrophosphate[J]. Science, 1998, 281(5375): 363-365.

[4] MERRIMAN B,ROTHBERG J M. Progress in ion torrent semiconductor chip based sequencing[J]. Electrophoresis, 2012, 33(23): 3397-3417.

[5] OVERBEEK R, OLSON R, PUSCH G D, et al. The SEED and the rapid annotation of microbial genomes using subsystems technology (RAST)[J]. Nucleic Acids Res, 2014, 42(Database issue): 206-214.

[6] DELCHER A L, BRATKE K A, POWERS E C, et al. Identifying bacterial genes and endosymbiont DNA with Glimmer[J]. Bioinformatics, 2007, 23(6): 673-679.

[7] HOLLAND M M, PARSON W. GeneMarker(R) HID: A reliable software tool for the analysis of forensic STR data [J]. J. Forensic Sci, 2011, 56(1): 29-35.

[8] VICTOR S, ASAF S. Automatic annotation of microbial genomes and metagenomic sequences in metagenomics and its applications in agriculture[J]. Biomedicine and Environmental Studies, 2011: 61-78.

[9] HUNTER S, JONES P, MITCHELL A, et al. InterPro in 2011: new developments in the family and domain prediction database[J]. Nucleic Acids Res, 2012, 40 (Database issue): D306-312.

[10] TATUSOV R L, KOONIN E V, LIPMAN D J. A genomic perspective on protein families[J]. Science, 1997, 278 (5338): 631-637.

[11] LAGESEN K, HALLIN P, RODLAND E A, et al. RNAmmer: consistent and rapid annotation of ribosomal RNA genes[J]. Nucleic Acids Res, 2007, 35(9): 3100-3108.

[12] SCHATTNER P, BROOKS A N, LOWE T M. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs[J]. Nucleic Acids Res, 2005, 33(Web Server issue): W686-689.

[13] BUTLER J E, YOUNG N D, AKLUJKAR M, et al. Comparative genomic analysis of Geobacter sulfurreducens KN400, a strain with enhanced capacity for extracellular electron transfer and electricity production [J]. BMC Genomics, 2012, 13: 471.

[14] UNTERGASSER A, CUTCUTACHE I, KORESSAAR T, et al. Primer3-new capabilities and interfaces[J]. Nucleic Acids Res, 2012, 40(15): e115.

[15] NADALIN F, VEZZI F, POLICRITI A. GapFiller: a de novo assembly approach to fill the gap within paired reads [J]. BMC Bioinformatics, 2012, 13 (Suppl 14): S8.

[16] LUO R, LIU B, XIE Y, et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler[J]. Gigascience, 2012, 1(1): 18.