bio-101 www.bio-protocol.org/exxxx DOI:10.21769/BioProtoc.xxxx

内生镰刀菌基因组染色体级别组装和注释

Chromosome-Scale Genome Assembly and Annotation Method of Endophyte *Fusarium*

单晓亮 [1, 2]，袁志林 [1, 2,*]

[1] 中国林业科学研究院林木遗传育种国家重点实验室，北京；[2] 中国林业科学研究院亚热带林业研究所，杭州

*通讯作者邮箱: yuanzl@caf.ac.cn

摘要:镰刀菌（*Fusarium*）是一种丝状真菌，其包含许多农业上重要的植物病原体，也是霉菌毒素的产生者和机会性感染人类的病原体，但是我们在前期实验中发现了两种可以促进植物生长的内生镰刀菌：黄色镰刀菌（*F. culmorum*）和假禾谷镰刀菌（*F. pseudograminearum*），为了进一步解释这种现象的原因，我们对其进行了全基因组测序（WGS）。我们主要利用 PacBio 三代测序和 Illumina 二代测序技术相结合的方法，得到染色体级别的基因组。进一步结合 *de novo* 注释和同源的预测结果得到基因的结构注释，结合 NR 等数据库对基因集得到了功能注释，最终得到染色体级别的内生镰刀菌基因组组装结果和高质量的基因组注释结果。为后续研究人员开展内生镰刀菌比较基因组、进化选择分析、功能研究和共生互作提供高质量的参考基因组信息。

关键词: PacBio 测序，Illumina 测序，内生镰刀菌

材料与试剂

1. 内生镰刀菌 *Fusarium culmorum* Class2-1B、*Fusarium pseudograminearum* Class2-1C，分离自沿海滩涂植物滨麦 *Leymus mollis*，与植物共生可以促进植物生长和提高植物耐盐性 (Rodriguez 等，2008; Redman 等，2011; Pan 等，2018)

仪器设备

1. 三代测序仪 (Pacific Biosciences PacBio RS II)
2. 二代测序仪 (Illumina HiSeq 2500)

软件和数据库

1. MECAT2 (https://github.com/xiaochuanle/MECAT2)
2. BUSCO v2.0 (https://busco.ezlab.org/)
3. tRNAscan-SE (http://lowelab.ucsc.edu/tRNAscan-SE)
4. RepeatModeler: http://www.repeatmasker.org/RepeatModeler
5. RepeatMasker: http://repeatmasker.org
6. NR (https://www.ncbi.nlm.nih.gov/refseq/about/nonredundantproteins)
7. Swiss-Prot (https://www.uniprot.org/statistics/Swiss-Prot)
8. KEGG databases (https://www.genome.jp/kegg/kegg1.html)
9. Repbase database: https://www.girinst.org/server/RepBase
10. Fungi odb10 dataset: https://busco.ezlab.org/frames/fungi.htm
11. TRF (Tandem repeats finder) http://tandem.bu.edu/trf/trf.unix.help.html
12. LTR_FINDER http://tlife.fudan.edu.cn/tlife/ltr_finder
13. Augustus http://bioinf.uni-greifswald.de/augustus/
14. GlimmerHMM http://ccb.jhu.edu/software/glimmerhmm/
15. Piler http://www.drive5.com/piler
16. RepeatScout https://github.com/mmcco/RepeatScout
17. TrEMBL https://www.uniprot.org/statistics/TrEMBL
18. Interpro https://www.ebi.ac.uk/interpro/
19. *Fusarium culmorum* strain PV, whole genome shotgun sequencing project https://www.ncbi.nlm.nih.gov/nuccore/PVEM00000000
20. *Fusarium* pseudograminearum CS3096, whole genome shotgun sequencing project https://www.ncbi.nlm.nih.gov/nuccore/AFNW00000000

实验步骤

一、测序

1. 使用太平洋生物科学公司开发的单分子实时 (SMRT) 测序和 Illumina HiSeq 2500 测序技术来组装完整的基因组。测序在北京诺禾致源生物信息技术有限公司进行。
2. 取单孢分离后培养 15 天的内生镰刀菌 *Fusarium culmorum*、*Fusarium pseudograminearum* PDA 平板，使用 Omega 真菌 DNA 提取试剂盒提取

DNA，DNA 浓度大于 100 ng/μl，DNA 纯度（$OD_{260/280}$ 在 1.8-2.0 之间；$OD_{260/230}$ 在 2.0-2.2 之间），使用 50 mg DNA 构建 PacBio 和 Illumina 测序文库。

3. 对 PacBio 文库，构建每个菌株的 20 kb 插入片段大小的标准 SMRTbell 文库，用 PacBio Sequel II 系统对 PacBio 长读序列进行测序。
4. 为了完善基于 PacBio long-read 的基因组组装，在 Illumina HiSeq 2500 上对插入大小为 500 bp 的双端 Illumina DNA 文库进行了测序。
5. 基于 Illumina Short Reads 的数据，分析了两个基因组的 K-mer 分布，并估计了两个基因组的大小。

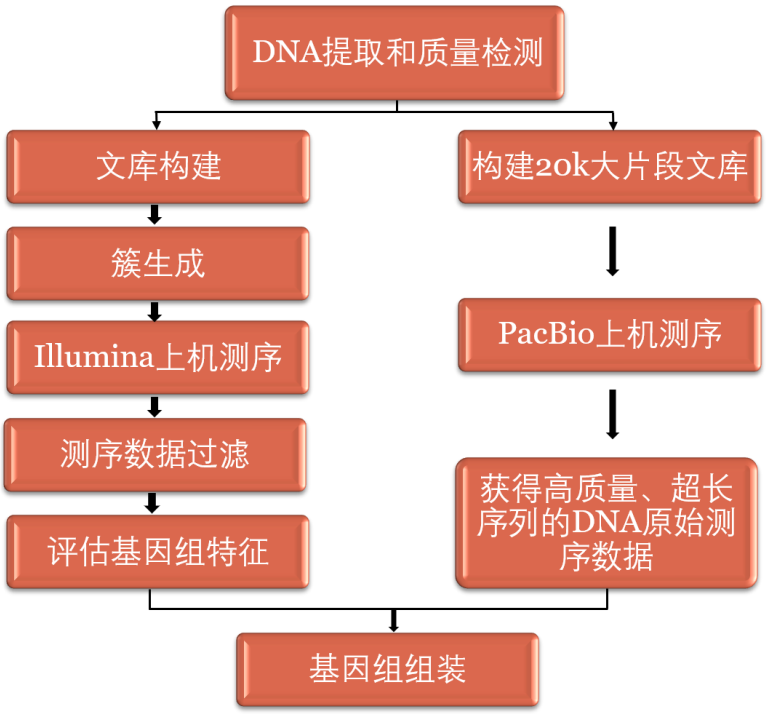

图 1 Illumina 和 PacBio 测序流程图

图 1 展示了第二代测序 Illumina 和第三代测序 PacBio 技术的测序流程，结合二代和三代测序数据进行了高质量的基因组组装。

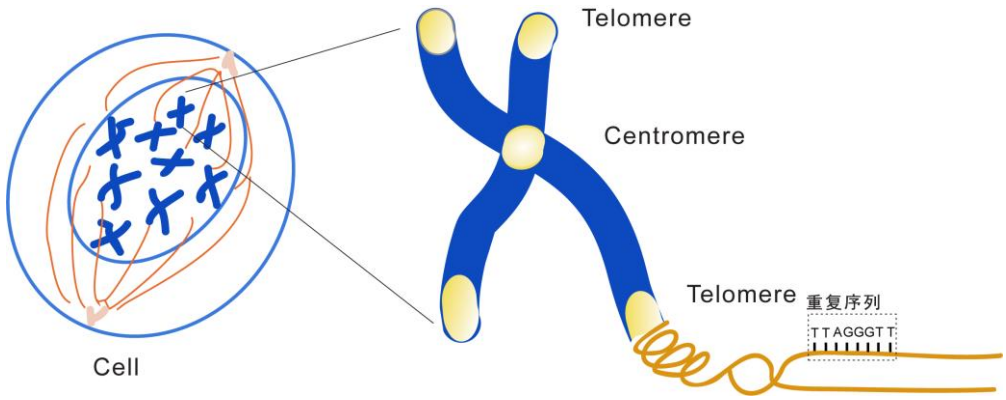

图 2 真菌染色体结构模式图

图 1 模式图展示了真菌的染色体两端具有端粒结构，在基因组组装中，染色体端粒到端粒的组装代表染色体的完整性，也是高质量基因组组装结果的标志。

二、基因组组装和注释

1. 获得了 16.7 GB 的 long-read 数据 *F. culmorum* Class2-1B，其中 Scaffold N50 的长度为 9.63 M；而 *F. pseudograminearum* Class2-1C，获得了 19.7 GB 的 long-read 数据，Scaffold N50 的长度为 9.15 M。
2. 利用 MECAT2 进行了基因组组装和纠错。然后使用 Pilon (v1.22) 用 Illumina 短读测序对三代组装结果进行纠错和修正。
3. Class2-1B 和 Class2-1C 都分别得到了 6 和 7 个 Scaffold，参照两株镰刀菌的参考基因组：*F. culmorum* PV，*F. Pseudograminearum* CS3096 (Schmidt, Ruth, 等，2018; Gardiner DM 等，2012) 。端粒是真核生物染色体末端的 DNA 重复序列，作用是保持染色体的完整性和控制细胞分裂周期。将 Class2-1B 组装成的四条染色体，其中两条是端粒对端粒，而另外两条只在一端有一个已识别的端粒。将 Class2-1C 组装成的四条染色体，其中三条两端都有端粒结构，而另外一条只在一端有一个已识别的端粒。在 Scaffold 末端发现了 TTAGGG 的串联重复序列 (或互补 DNA 链序列，AATCCC) 。Class2-1B 和 Class2-1C 的 Scaffold 至少有一端含有端粒结构，每条 Scaffold 都接近完整染色体的长度 (Aksenova 和 Mirkin，2019) 。如上所示，两个内生镰刀菌基因组的染色体都含有图 2 中的端粒结构，说明我们得到了两个高质量组装的基因组。通过 BLAST 搜索鉴定了 Class2-1B 中的 2 个短 Scaffold 为线粒体基因

组，总 GC 含量为 31.2%。同样，通过 BLAST 搜索鉴定了 Class2-1C 中的 3 个短 Scaffold 为线粒体基因组，总 GC 含量为 34.6%，进一步分别比较它们的同种镰刀菌线粒体基因组时，发现这两个线粒体基因组都显示出大于 98%的序列同源性 (Kulik 等，2020) .

4. 通过结合 de novo 注释和基于同源的预测结果进行蛋白质编码基因的结构注释 (Rigden, 2017) 。使用 Maker (v.2.31.9) 分别在 Class2-1B 和 Class2-1C 中预测了 11450 和 11221 个完整的蛋白编码基因模型。发现 Class2-1B 和 Class2-1C 中分别有 97.06%和 96.93%的基因可以在 InterProScan、Gene Ontology、KEGG 以及 NR 数据库被注释。
5. 使用 BUSCO (Benchmark Universal Single-Copy Orologs) Fungi odb10 数据库 (v.4.0.6) 对基因注释和基因组组装质量进行评估，结果显示 Class2-1B 和 Class2-1C 的基因注释和基因组组装质量分别为 98.8%和 99.1% (总共搜索了 758 个保守核心蛋白) ，这表明俩个基因组的组装质量是非常高的 (Simão 等，2015)。
6. 对于转座子 (TEs) 注释，RepeatMasker (v.4.07) 用于 Repbase 数据库 (v.23.06) (Bao 等，2015) 来识别已知的 TEs。同时，还使用 RepeatModeler (v1.0.11) 和 LTR finder (Jurka 等，2005) 进行从头检测。在 Class2-1B 和 Class2-1C 中分别鉴定出约 1.55Mb 和 2.04Mb 的 TEs (占总基因组的 4.13%和 5.37%) 。

结果分析

表 1. 黄色镰刀菌和假禾谷镰刀菌的基因组特点和预测特征

Characteristics	*F. culmorum*	*F. pseudograminearum*
Total genome size (Mb)	40.05	42.90
Nuclear genome size (Mb)	39.91	42.76
Mitogenome size (bp)	136,406	136,045
N50 Scaffold length (Mb)	9.63	9.15
Chromosome numbers	4	4
Scaffolds numbers	6	7
Genome coverage	443	519
G+C (%)	47.4	47.0
N50 Scaffold average (Mb)	1.19	1.81
Total transposable elements (Mb)	1.55	2.04
The total number of gene	11450	11221
Average gene length (bp)	1653	1633
Genome BUSCO (%)	98.8	99.1

致谢

本 protocol 的研究工作得到课题“内生镰刀菌促进树木生长和耐盐性的分子调控机制研究”资助经费，课题编号为 76B2018001。

参考文献：

1.Rodriguez, R. J., Henson, J., Van Volkenburgh, E., Hoy, M., Wright, L., Beckwith, F., Kim, Y. O. and Redman, R. S. (2008). Stress tolerance in plants via habitat-adapted symbiosis. *ISME J.* 2: 404–416.

2.Redman, R. S., Kim, Y. O., Woodward, C. J. D. A., Greer, C., Espino, L., Doty, S. L. and Rodriguez, R. J. (2011). Increased fitness of rice plants to abiotic stress via habitat adapted symbiosis: A strategy for mitigating impacts of climate change. *PLoS One* 6: 1-10.

3.Pan, X. Y., Sun, H. J. and Yuan, Z. L. (2018). Toxin accumulation of three Leymus mollis-associated endophytic Fusarium Isolates and their effects 200 on growth and salt tolerance of Liquidambar styraciflua seedlings. *For. Res*. 31: 64–73.

4.Schmidt, R., Durling, M. B., de Jager, V., Menezes, R. C., Nordkvist, E., Svatoš, A., Dubey, M., Lauterbach, L., Dickschat, J. S., Karlsson, M. et al. (2018).

"Deciphering the genome and secondary metabolome of the plant pathogen Fusarium culmorum." *FEMS microbiology ecology* 94.6: fiy078.

5.Gardiner, D. M., McDonald, M. C., Covarelli, L., Solomon, P. S., Rusu, A. G., Marshall, M., Kazan, K., Chakraborty, S., McDonald, B. A. and Manners, J. M. (2012). Comparative pathogenomics reveals horizontally acquired novel virulence genes in fungi infecting cereal hosts. *PLoS Pathog*, 8(9): e1002952.

6.Aksenova, A. Y. and Mirkin, S. M. (2019). At the beginning of the end and in the middle of the beginning: structure and maintenance of telomeric dna repeats and interstitial telomeric sequences. *Genes (Basel)* 10: 118.

7.Kulik, T., Brankovics, B., van Diepeningen, A. D., Bilska, K., Żelechowski, M., Myszczyński, K., Molcan, T., Stakheev, A., Stenglein, S, Beyer, M. et al. (2020).Diversity of mobile genetic elements in the mitogenomes of closely related Fusarium culmorum and F. graminearum sensu stricto strains and its implication for diagnostic purposes. *Front. Microbiol*. 11: 1–14.

8.Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31: 3210–3212.

9.Bao, W., Kojima, K. K. and Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA*. 6: 4–9.

10.Rigden, D. J. (2017). From protein structure to function with bioinformatics: second edition.

11.Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. *Cytogentic and Genome Research* 110: 462-467.

12.Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* 27: 573-580.

13.Price, A. L., Jones, N. C. and Pevzner, P. A. (2005). De novo identification of repeat families in large genomes. *Bioinformatics* 21: i351-i358.

14.Edgar, R. C. and Myers, E. W. Piler: *(*2005). Identification and Classification of genomic repeats. *Bioinformatics* 21: i152-158.

15.Xu, Z. and Wang, H. Ltr_ (2007). Finder: an efficient tool for the prediction of full-length ltr retrotransposons. *Nucl. Acids Res.* 35: W265-268.

16.Kent, W. J. (2002). BLAT-the BLAST-like alignment tool. *Genome Res.* 12:

656–664.

17.Guy, S. and Ewan, B. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* 6: 31

18.Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S. and Morgenstern, B. (2006). "AUGUSTUS: ab initio prediction of alternative transcripts" *Nucleic Acids Research* 34: W435-W439.

19.Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J. and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28: 511-515.

20.Majoros, W. H., Pertea, M. and Salzberg, S. L. TigrScan and GlimmerHMM: two open

21.Carson, H. and Mark, Y. (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* 12: 491.

22.Bairoch, A. and Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucl. Acids Res.* 28: 45-48.

23.Zdobnov, E. M. and Apweiler, R. (2001). InterProScan - an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* 17: 847-848.

24.Ashburner, M. Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T. et al. (2000). Gene Ontology: tool for the unification of biology. *Nat Genet* 25: 25-29.

25.Kanehisa, M. and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28: 27-30.

26.Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S. R. and Bateman, A. (2005). Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res* 33: D121-4.

27.Todd M. Lowe and Sean R. Eddy. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*