**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)

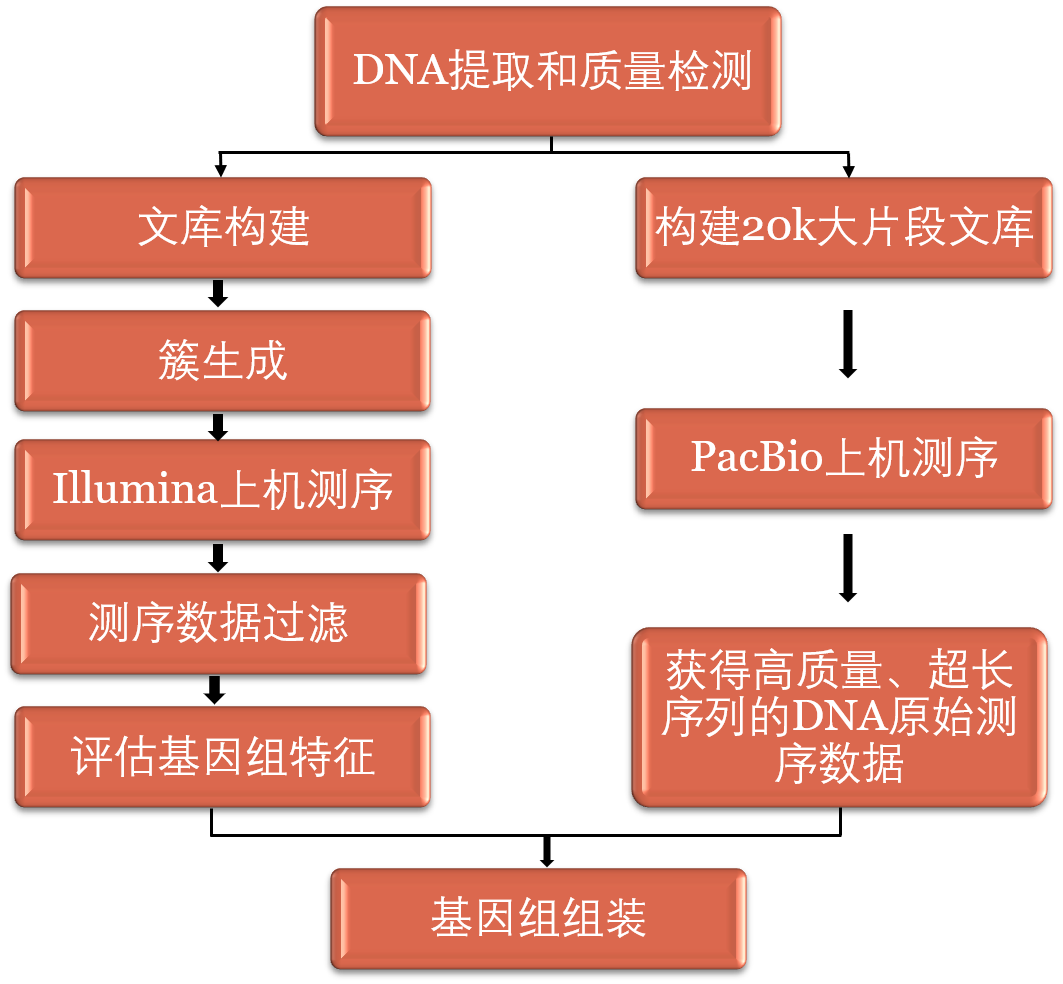
**仪器设备**

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**

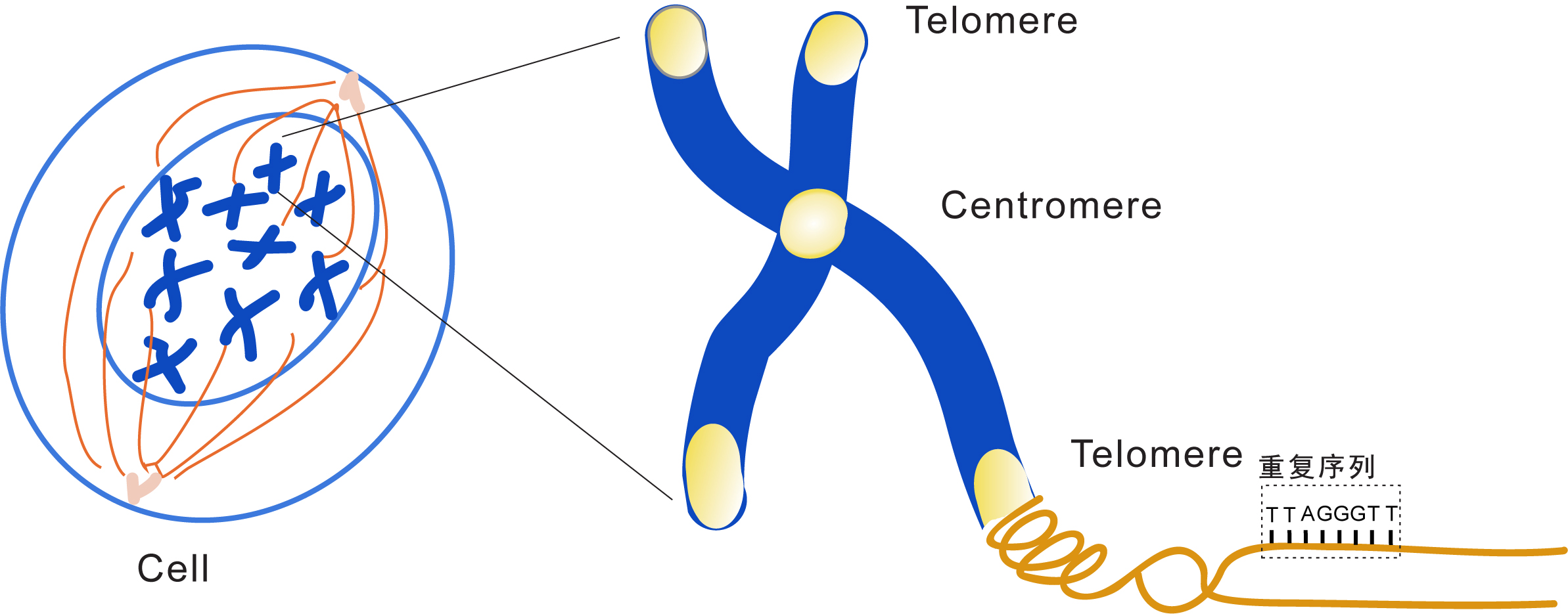
1. MECAT2 (https://github.com/xiaochuanle/MECAT2)
2. BUSCO v2.0 (https://busco.ezlab.org/)
3. tRNAscan-SE (http://lowelab.ucsc.edu/tRNAscan-SE)
4. RepeatModeler: http://www.repeatmasker.org/RepeatModeler
5. RepeatMasker: http://repeatmasker.org
6. NR (<https://www.ncbi.nlm.nih.gov/refseq/about/nonredundantproteins>)
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**

1. **测序**
2. 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。
3. 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 (OD260/280 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。
4. 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。
5. 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。
6. 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。

1. **基因组组装和注释**
2. 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。
3. 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。
4. 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) .
5. 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。
6. 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)。
7. 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.](https://pubmed.ncbi.nlm.nih.gov/18256707/) *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](https://pubmed.ncbi.nlm.nih.gov/21750695/). *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.](http://en.cnki.com.cn/Article_en/CJFDTotal-LYKX201805010.htm) *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](https://pubmed.ncbi.nlm.nih.gov/29718180/)." *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](https://pubmed.ncbi.nlm.nih.gov/23028337/). *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](https://pubmed.ncbi.nlm.nih.gov/30764567/). *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](https://pubmed.ncbi.nlm.nih.gov/32528440/). *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](https://pubmed.ncbi.nlm.nih.gov/26059717/). *Bioinformatics* 31: 3210–3212.
9. Bao, W., Kojima, K. K. and Kohany, O. (2015). [Repbase Update, a database of repetitive elements in eukaryotic genomes](https://pubmed.ncbi.nlm.nih.gov/26045719/). *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](https://pubmed.ncbi.nlm.nih.gov/16093699/). *Cytogentic and Genome Research* 110: 462-467.
12. Benson, G. (1999). [Tandem repeats finder: a program to analyze DNA sequences.](https://pubmed.ncbi.nlm.nih.gov/9862982/) *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](https://pubmed.ncbi.nlm.nih.gov/15961478/). *Bioinformatics* 21: i351-i358.
14. Edgar, R. C. and Myers, E. W. Piler: *(*2005). [Identification and Classification of genomic repeats](https://pubmed.ncbi.nlm.nih.gov/15961452/). *Bioinformatics* 21: i152-158.
15. Xu, Z. and Wang, H. Ltr\_ (2007). [Finder: an efficient tool for the prediction of full-length ltr retrotransposons](https://pubmed.ncbi.nlm.nih.gov/17485477/). *Nucl. Acids Res.* 35: W265-268.
16. Kent, W. J. (2002). [BLAT-the BLAST-like alignment tool](https://pubmed.ncbi.nlm.nih.gov/11932250/). *Genome Res.* 12: 656–664.
17. Guy, S. and Ewan, B. (2005). [Automated generation of heuristics for biological sequence comparison](https://pubmed.ncbi.nlm.nih.gov/15713233/). *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](https://pubmed.ncbi.nlm.nih.gov/16845043/)" *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](https://pubmed.ncbi.nlm.nih.gov/20436464/). *Nat. Biotechnol.* 28: 511-515.
20. Majoros, W. H., Pertea, M. and Salzberg, S. L. [TigrScan and GlimmerHMM: two open](https://pubmed.ncbi.nlm.nih.gov/15145805/)
21. Carson, H. and Mark, Y. (2011). [MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects](https://pubmed.ncbi.nlm.nih.gov/22192575/). *BMC Bioinformatics* 12: 491.
22. Bairoch, A. and Apweiler, R. (2000). [The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000.](https://pubmed.ncbi.nlm.nih.gov/10592178/) *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](https://pubmed.ncbi.nlm.nih.gov/11590104/). *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](https://pubmed.ncbi.nlm.nih.gov/10802651/). *Nat Genet* 25: 25-29.
25. Kanehisa, M. and Goto, S. (2000). [KEGG: kyoto encyclopedia of genes and genomes](https://pubmed.ncbi.nlm.nih.gov/10592173/). *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](https://pubmed.ncbi.nlm.nih.gov/15608160/). *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](https://pubmed.ncbi.nlm.nih.gov/9023104/). *Nucleic Acids Res.*