Genomics Communications

PERSPECTIVE

Open Access

https://doi.org/10.48130/gcomm-0025-0006

Genomics Communications **2025**, 2: e005

Tips for improving genome annotation quality

Lan Lan[1,2,3*], Haifei Hu[4], Yong Jia[1,3], Xiaoni Zhang[2], Minlong Jia[5], Chengdao Li[1,3] and Zhiqiang Wu[2]

[1] *State Agricultural Biotechnology Centre (SABC), College of Science, Health, Engineering and Education, Murdoch University, WA 6150, Australia*
[2] *Shenzhen Branch, Guangdong Laboratory of Lingnan Modern Agriculture, Key Laboratory of Synthetic Biology, Ministry of Agriculture and Rural Affairs, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen 518120, China*
[3] *Western Crop Genetic Alliance, Murdoch University, WA 6150, Australia*
[4] *Rice Research Institute, Guangdong Academy of Agricultural Sciences & Guangdong Key Laboratory of New Technology in Rice Breeding, Guangzhou 510640, China*
[5] *College of Horticulture, Shanxi Agricultural University, Taiyuan 030031, China*
* Corresponding author, E-mail: 34648864@student.murdoch.edu.au

Abstract

With the advancement of sequencing technology and genome assembly algorithms, we can easily obtain high-quality genome assembly results, however, the remaining challenge is accurate genome annotation. As the cornerstone for downstream analysis, there are many details during the genome annotation process that are worth discussing or mentioning. To reduce the annotation errors, we briefly list six tips during the annotation process, and the noteworthy details in different widely used annotation tools/pipelines such as MAKER and EvidenceModeler, to improve annotation quality.

Citation: Lan L, Hu H, Jia Y, Zhang X, Jia M, et al. 2025. Tips for improving genome annotation quality. *Genomics Communications* 2: e005 https://doi.org/10.48130/gcomm-0025-0006

With the advances in long-read sequencing technology, we can easily generate a large amount of high-quality sequencing data using efficient sequencing platforms such as PacBio Revio (www.pacb.com/revio) and Nanopore duplex[1]. With these high-quality data and state-of-the-art assembly algorithms (such as hifiasm[2] and verkko[3]), we can conveniently generate a high-quality, error-free, and more continuous genome assembly than before. In general, a high-quality genome assembly is generated by following a combined strategy of scaffolding, polishing, and gap-filling processes, then the genome annotation could be considered the most important process for further in-depth analysis. Nowadays, genome annotation is facing challenges of the partial conservation of sequence patterns, variable intron length, the different distances between the genes, alternative splicing, TE insertions, and pseudogenes[4]. Accurate genome annotation is crucial and provides basic information to focus on species evolution, population genetic analysis, functional genomics, and so on.

There are a large number of well-established and widely used genome annotation pipelines, such as MAKER[5], MAKER-P[6], BRAKER[7], GETA (https://github.com/chenlianfu/geta), EvidenceModeler[8], EGAP (Eukaryotic Genome Annotation Pipeline) (https://github.com/ncbi/egapx), and so on. Although these pipelines are different in detail (for example, the different pipelines combine different software/tools), the core idea is almost the same: combining the different types of evidence and synthesizing them into non-redundant genome annotations. These different types of evidence include *ab initio* (*de novo*) evidence, expressed sequence tags (ESTs), or RNA-seq evidence, and homology-based evidence. Each type of evidence can be generated by corresponding input data and approaches. For *ab initio* evidence, we could generate the results by using the RNA-seq data and high-quality homologous protein sequences to train the gene models for the researchers' specific case, through AUGUSTUS[9], SNAP (https://github.com/KorfLab/SNAP), Helixer[10], and so on. For ESTs or RNA-seq evidence, we could obtain the gene models by using the RNA-seq data through the Program to Assemble Spliced Alignments (PASA)[8], StringTie[11], TransDecoder (https://github.com/TransDecoder/TransDecoder), exonerate (https://github.com/nathanweeks/exonerate), and so on, to generate the transcriptome information to support genome annotation. For the homology-based evidence, we use the homologous protein sequences to generate the results through miniprot[12], exonerate, GenomeThreader[13], and so on. Finally, these different types of evidence are combined to obtain final non-redundant high-quality genome annotation, usually, this combining process could be varied if researchers use different pipelines. Here we have summarized the commonly used genome annotation pipelines and the widely used tools in each procedure (Fig. 1).

It is widely known that different evidence and pipelines could cause significantly varied annotation results[14], and the different parameters used in certain tools/pipelines may also generate slightly different annotation results. For example, for different cut-off e-values used in read/sequence alignment (e.g. the default e-value for blastx in MAKER is 1e-06, researchers can manually change it as 1e-5 or any other value), the different alignment results may slightly change the evidence, and finally change the annotation results. This means that even if we have chosen good benchmarked tools/pipelines and high confidence and sufficient evidence, the annotation results will still have prediction errors, including loss of genes[15], incorrect exon and gene boundaries[16], retention of non-coding sequences in coding exons[17], and fragmentation or fusion of gene models[18]. These errors can seriously affect downstream analysis, leading to bias in large-scale genomic research[19–21]. Here, we briefly discuss reasonable procedures to reduce annotation errors (Fig. 2).

First, the identification and masking of repeat regions is critical, for the repeat regions could lead to false evidence for gene annotation[22]. For example, the transposable elements (TEs) inserted into the genic region could be erroneously identified as extra exons. In maize, researchers have found that the TEs near or within genes would fragment gene space or even miss gene annotations[23]. By masking the TEs with different lengths (≥ 1 kb, ≥ 500 bp, and ≥ 80 bp), researchers found that unmasking short TEs (< 1 kb) would effectively prevent the fragmentation of genic regions and improve

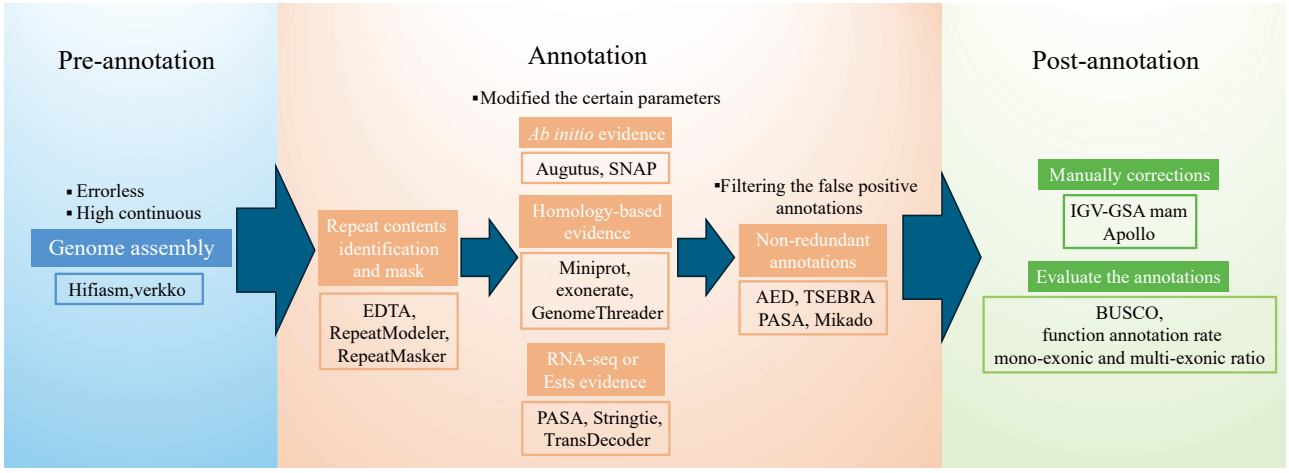
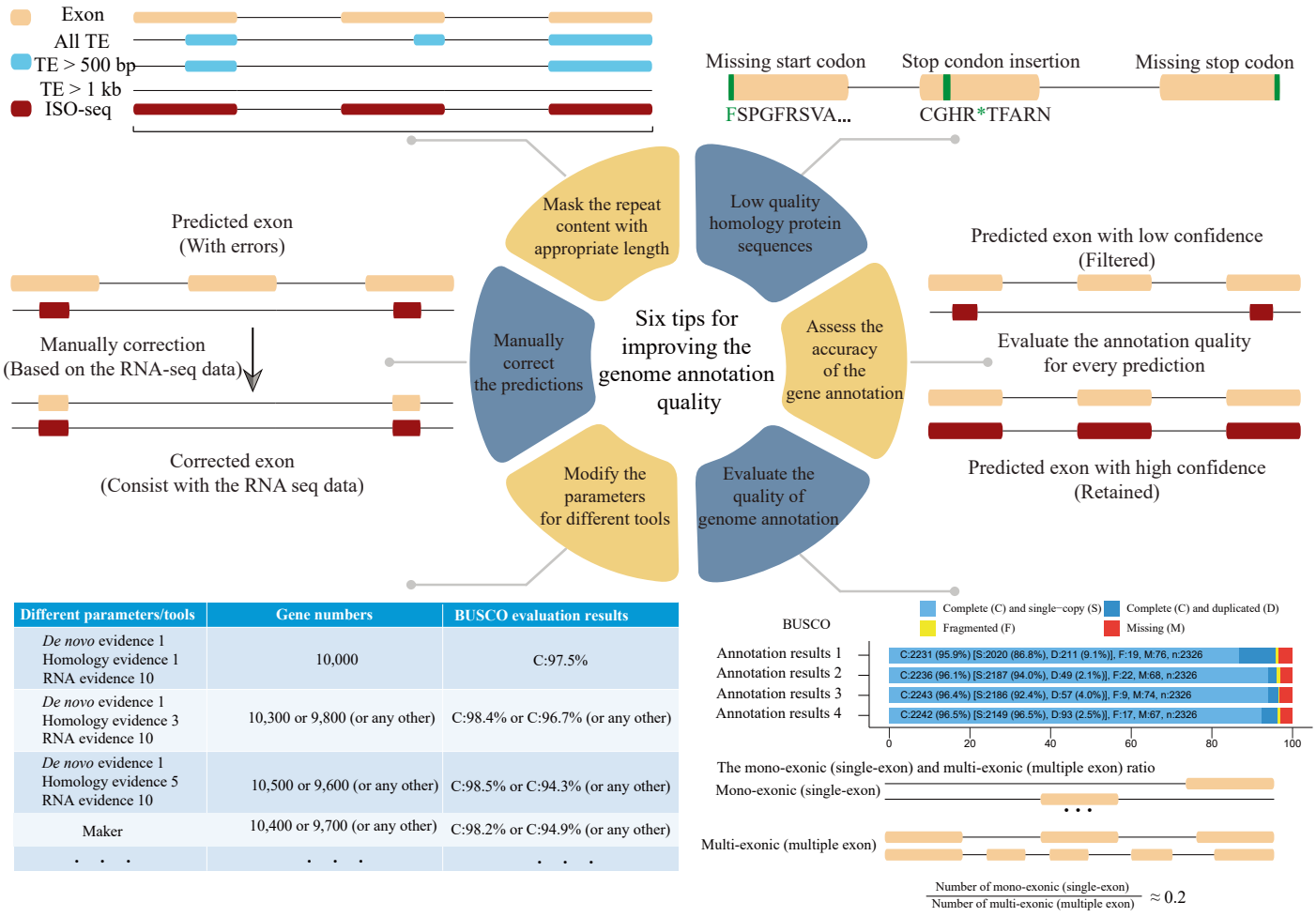


Fig. 1 Recommended flowchart for genome annotation. The tools/pipelines listed in the figure only represent part of those commonly used in genome annotations. Researchers need to carefully consider and use the appropriate tools/pipelines for their specific scenario.



Different parameters/tools	Gene numbers	BUSCO evaluation results
De novo evidence 1 Homology evidence 1 RNA evidence 10	10,000	C:97.5%
De novo evidence 1 Homology evidence 3 RNA evidence 10	10,300 or 9,800 (or any other)	C:98.4% or C:96.7% (or any other)
De novo evidence 1 Homology evidence 5 RNA evidence 10	10,500 or 9,600 (or any other)	C:98.5% or C:94.3% (or any other)
Maker	10,400 or 9,700 (or any other)	C:98.2% or C:94.9% (or any other)
. . .	. . .	. . .

Fig. 2 Six tips for improving genome annotation quality.

the annotation quality according to the evaluation results of Benchmarking Universal Single-Copy Ortholog (BUSCO)[24] in maize[23]. Therefore, we recommend that the researchers mask the TEs of limited length rather than all TEs when performing genome annotation. Moreover, one of the most commonly used repeat content annotation pipelines, Extensive de novo TE Annotator (EDTA), its default output of repeat-masked genome sequences named '[genome_prefix].fa.mod.MAKER.masked', masked the TE > 1 kb which means you could mask the genome by filtering different TE lengths. By masking the TE with different lengths, it is possible to explore reasonable and suitable annotation results for your specific genomes.

Second, erroneous gene predictions could be propagated while we generate homology-based evidence. The low-quality proteins

from the public databases containing the sequence or prediction errors could lead to inaccurate annotation results[25,26], for example, some genes may contain multiple stop codons. Other annotation processes may further use these errors to generate the new annotation results and eventually obtain a wide dissemination. To address the challenges caused by low-quality homology evidence, we recommend utilizing the high-confidence protein from uniprot (www.uniprot.org/help/downloads) and the already known well-annotated results from related species or perform quality control by yourself before the annotation process.

Third, genome annotation results using different pipelines may vary and cause bias. There are already benchmarking works showing that MAKER has a lower performance than the combination pipeline of TSEBRA[27] and BRAKER[4,7]. Even using the same pipeline, e.g. EvidenceModeler, users must manually select the weight value for different evidence to generate the appropriate results, while modifying the weight value will undoubtedly slightly change the annotation results. In practice, we recommend that researchers set the weight value from the EST or RNA-seq evidence to 10 and the other set to 1. Also, the researchers could modify the weight value for their specific case, which would lead to slightly different annotation results, and then choose the best one as the final annotation result. Moreover, we need to be cautious of default parameters used in certain tools. Some species may have a specific genome architecture, such as the Chinese pine containing many genes with an extremely long intron (> 10 kb)[28], when we use Trinity[29] to assemble the RNA-seq data on the genome-guide mode, the '--genome_guided_max_intron' should use the larger value instead of the default value of 10,000. In practice, we may need to try several iterations of different parameters, or even different tools/pipelines, and compare the different annotation results and choosing the outperformed one as the final result.

Fourth, we need to use appropriate evaluation methods to assess the accuracy of the gene annotation. If the researchers use MAKER, they could calculate the annotation edit distance (AED), which can be used to measure the congruence between the annotation and its supporting evidence[30]. The AED value ranges from 0 to 1, the annotations with an AED of 0 referring to the perfect agreement with the evidence provided, while 1 refers to the complete lack of evidence to support the annotation. In practice, the annotations with AED values lower than 0.5 (≤ 0.5) could be considered good annotations, and lower than 0.3 could be considered high-quality annotations[5]; so we can filter out the annotations with large AED values to reduce the false positive annotation results. For those who used the BRAKER pipeline, we could use TSEBRA (https://github.com/Gaius-Augustus/TSEBRA) to filter out the low-confidence annotations according to the provided evidence and increase the accuracy. Furthermore, PASA can be used to update annotation results, and incorporate PASA alignment evidence to correct exon boundaries for any annotation pipeline, as well as Mikado[31] dose.

Fifth, the genome annotation results could be manually corrected by Apollo[32], IGV - GSAman (Genome Sequence Annotation Manipulator, an extended version of Integrative Genomics Viewer) (https://gitee.com/CJchen/IGV-sRNA), and so on. Based on the datasets from RNA-seq and homologous proteins, we could jointly use them to manually correct the annotations. By visually checking each annotation, a significant improvement could be observed compared to the original annotations[33,34].

Sixth, evaluate the quality of genome annotation results globally. The BUSCO is the most commonly used tool to assess the quality. Empirically, the BUSCO complete value should be at least 90% to be considered as a good annotation result. The BUSCO evaluation process works by evaluating the presence of predefined highly conserved orthologous genes, and there are already 462 available datasets that could be used for evaluation. The most important thing is that we need to choose the appropriate datasets to evaluate the results, for example, the same annotation results could get different BUSCO completeness scores through the different BUSCO datasets[35]. We could also calculate the mono-exonic (single-exon), and multi-exonic (multiple-exon) ratios to evaluate the annotation quality, which can be done by gFACs[36]. By assessing a large set of model plant genomes' annotation results, the ideal ratio was near 0.2[4,37]. This means that, theoretically, the more the ratio deviates from 0.2, the higher the likelihood of erroneous annotations. Moreover, we could also compare the annotated genes with the public databases and calculate the rates. Higher rates of over 80% by the different databases indicate better prediction results[4].

From the above six tips, you would notice that the most important dataset is the RNA-seq data. Whatever the tools/pipelines we use for the genome annotation or correction, high-confidence evidence is always needed, and in most cases the sufficient depth, multiple tissues, and different periods of RNA-seq data is always treated as the most high-confidence evidence, for some genes may be expressed only in specific tissues or developmental period.

It is also worth mentioning that how the long-read transcriptome data was used in the genome annotation. In general, short-read data could produce predicting errors because of error mapping, as the single read cannot cover the full length of the gene. The use of long-read sequencing can increase the accuracy of automated genome annotation by improving genome mapping of sequencing data, correctly identifying intron-exon boundaries, directly identifying alternatively spliced transcripts, identifying transcription start and end sites, and providing accurate strand orientation for single exon genes[38,39]; and several studies have already shown that long reads paired with short reads can improve annotation quality[40,41], also benchmarking work has proven that using high-quality long reads could significantly improve the annotation results compared to the short-read-based annotations[42]. In addition, deep learning and machine learning approaches have recently been adopted for genomic annotation (reviewed by Chen et al.[43]). We hope that this discussion could inspire researchers in the annotation process and help them to choose the optimal pipeline/parameters to obtain annotation results with fewer errors for further downstream analysis.

Author contributions

The authors confirm contribution to the paper as follows: study conception and design: Lan L, Wu Z, Li C; draft manuscript preparation: Lan L, Hu H, Jia Y, Zhang X, Jia M. All authors reviewed the results and approved the final version of the manuscript.

Data availability

Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Acknowledgments

This work was funded by the Chinese Academy of Agricultural Sciences Elite Youth Program (110243160001007), The Agricultural Science and Technology Innovation Program, and the Funding of Major Scientific Research Tasks, Kunpeng Institute of Modern Agriculture at Foshan (KIMA-ZXFR2024004).

Conflict of interest

The authors declare that they have no conflict of interest. Dr. Zhiqiang Wu is the Editorial Board member of *Genomics Communications* who was blinded from reviewing or making decisions on the manuscript. The article was subject to the journal's standard procedures, with peer-review handled independently of this Editorial Board member and the research groups.

Dates

Received 15 December 2024; Revised 28 January 2025; Accepted 6 March 2025; Published online 26 March 2025

References

1. Koren S, Bao Z, Guarracino A, Ou S, Goodwin S, et al. 2024. Gapless assembly of complete human and plant chromosomes using only nanopore sequencing. *Genome Research* 34:1919–30
2. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. 2021. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature methods* 18:170–75
3. Rautiainen M, Nurk S, Walenz BP, Logsdon GA, Porubsky D, et al. 2023. Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nature Biotechnology* 41:1474–82
4. Vuruputoor VS, Monyak D, Fetter KC, Webster C, Bhattarai A, et al. 2023. Welcome to the big leaves: Best practices for improving genome annotation in non-model plant genomes. *Applications in Plant Sciences* 11:e11533
5. Holt C, Yandell M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* 12:491
6. Campbell MS, Holt C, Moore B, Yandell M. 2014. Genome annotation and curation using MAKER and MAKER-P. *Current Protocols In Bioinformatics* 48:4.11.1–4.11.39
7. Hoff KJ, Lomsadze A, Borodovsky M, Stanke M. 2019. Whole-genome annotation with BRAKER. In *Gene prediction. Methods in Molecular Biology*, ed. Kollmar M. New York: Humana. pp. 65–95. doi: 10.1007/978-1-4939-9173-0_5
8. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, et al. 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology* 9:R7
9. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, et al. 2006. AUGUSTUS: *ab initio* prediction of alternative transcripts. *Nucleic Acids Research* 34:W435–W439
10. Stiehler F, Steinborn M, Scholz S, Dey D, Weber APM, et al. 2020. Helixer: cross-species gene annotation of large eukaryotic genomes using deep learning. *Bioinformatics* 36:5291–98
11. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, et al. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology* 33:290–95
12. Li H. 2023. Protein-to-genome alignment with miniprot. *Bioinformatics* 39:btad014
13. Gremme G, Brendel V, Sparks ME, Kurtz S. 2005. Engineering a software tool for gene structure prediction in higher organisms. *Information and Software Technology* 47:965–78
14. Shen JS, Lan L, Kan SL, Cheng HF, Peng D, et al. 2024. A haplotype-resolved genome for Rhododendron× pulchrum and the expression analysis of heat shock genes. *Journal of Systematics and Evolution* 62:489–504
15. Denton JF, Lugo-Martinez J, Tucker AE, Schrider DR, Warren WC, Hahn MW. 2014. Extensive error in the number of genes inferred from draft genome assemblies. *PLoS computational biology* 10:e1003998
16. Guigó R, Agarwal P, Abril JF, Burset M, Fickett JW. 2000. An assessment of gene prediction accuracy in large DNA sequences. *Genome research* 10:1631–42
17. Drăgan MA, Moghul I, Priyam A, Bustos C, Wurm Y. 2016. GeneValidator: identify problems with protein-coding gene predictions. *Bioinformatics* 32:1559–61
18. Guigó R, Flicek P, Abril JF, Reymond A, Lagarde J, et al. 2006. EGASP: the human ENCODE genome annotation assessment project. *Genome biology* 7:S2
19. Prosdocimi F, Linard B, Pontarotti P, Poch O, Thompson JD. 2012. Controversies in modern evolutionary biology: the imperative for error detection and quality control. *BMC Genomics* 13:5
20. Weisman CM, Murray AW, Eddy SR. 2022. Mixing genome annotation methods in a comparative analysis inflates the apparent number of lineage-specific genes. *Current Biology* 32:2632–2639.e2
21. Söllner JF, Leparc G, Zwick M, Schönberger T, Hildebrandt T, et al. 2019. Exploiting orthology and de novo transcriptome assembly to refine target sequence information. *BMC Medical Genomics* 12:69
22. Yandell M, Ence D. 2012. A beginner's guide to eukaryotic genome annotation. *Nature Reviews Genetics* 13:329–42
23. Benson CW, Heringer P, Ou S. 2024. Four Strategies for Whole-Genome Annotation of Transposable Elements and Repeats in Maize. *Cold Spring Harbor Protocols*
24. Seppey M, Manni M, Zdobnov EM. 2019. BUSCO: assessing genome assembly and annotation completeness. In *Gene prediction. Methods in Molecular Biology*, ed. Kollmar M. New York: Humana. pp. 227-45. doi: 10.1007/978-1-4939-9173-0_14
25. Salzberg SL. 2019. Next-generation genome annotation: we still struggle to get it right. *Genome Biology* 20:92
26. Tørresen OK, Star B, Mier P, Andrade-Navarro MA, Bateman A, et al. 2019. Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. *Nucleic Acids Research* 47:10994–1006
27. Gabriel L, Hoff KJ, Brůna T, Borodovsky M, Stanke M. 2021. TSEBRA: transcript selector for BRAKER. *BMC Bioinformatics* 22:566
28. Niu S, Li J, Bo W, Yang W, Zuccolo A, et al. 2022. The Chinese pine genome and methylome unveil key features of conifer evolution. *Cell* 185:204–217.e14
29. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, et al. 2013. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols* 8:1494–512
30. Eilbeck K, Moore B, Holt C, Yandell M. 2009. Quantitative measures for the management and comparison of annotated genomes. *BMC Bioinformatics* 10:67
31. Venturini L, Caim S, Kaithakottil GG, Mapleson DL, Swarbreck D. 2018. Leveraging multiple transcriptome assembly methods for improved gene structure annotation. *GigaScience* 7:giy093
32. Dunn NA, Unni DR, Diesh C, Munoz-Torres M, Harris NL, et al. 2019. Apollo: democratizing genome annotation. *PLoS Computational Biology* 15:e1006790
33. Feng J, Zhang W, Chen C, Liang Y, Li T, et al. 2024. The pineapple reference genome: Telomere-to-telomere assembly, manually curated annotation, and comparative analysis. *Journal of Integrative Plant Biology* 66:2208–25
34. Liao B, Shen X, Xiang L, Guo S, Chen S, et al. 2022. Allele-aware chromosome-level genome assembly of Artemisia annua reveals the correlation between ADS expansion and artemisinin yield. *Molecular Plant* 15:1310–28
35. Lan L, Leng L, Liu W, Ren Y, Reeve W, et al. 2023. The haplotype-resolved telomere-to-telomere carnation (*Dianthus caryophyllus*) genome reveals the correlation between genome architecture and gene expression. *Horticulture Research* 11:uhad244
36. Caballero M, Wegrzyn J. 2019. gFACs: gene filtering, analysis, and conversion to unify genome annotations across alignment and gene prediction frameworks. *Genomics, Proteomics & Bioinformatics* 17:305–10
37. Jain M, Khurana P, Tyagi AK, Khurana JP. 2008. Genome-wide analysis of intronless genes in rice and Arabidopsis. *Functional & integrative genomics* 8:69–78

38. Minoche AE, Dohm JC, Schneider J, Holtgräwe D, Viehöver P, et al. 2015. Exploiting single-molecule transcript sequencing for eukaryotic gene prediction. *Genome Biology* 16:184
39. Abdel-Ghany SE, Hamilton M, Jacobi JL, Ngam P, Devitt N, et al. 2016. A survey of the sorghum transcriptome using single-molecule long reads. *Nature Communications* 7:11706
40. Wei C, Yang H, Wang S, Zhao J, Liu C, et al. 2018. Draft genome sequence of *Camellia sinensis* var. *sinensis* provides insights into the evolution of the tea genome and tea quality. *Proceedings of the National Academy of Sciences of the United States of America* 115:E4151−E4158
41. Xie M, Chung CYL, Li MW, Wong FL, Wang X, et al. 2019. A reference-grade wild soybean genome. *Nature Communications* 10:1216
42. Paniagua A, Agustín-García C, Pardo-Palacios FJ, Brown T, De Maria M, et al. 2024. Evaluation of strategies for evidence-driven genome annotation using long-read RNA-seq. *Genome Research* 35:1−12
43. Chen Z, Ain NU, Zhao Q, Zhang X. 2024. From tradition to innovation: conventional and deep learning frameworks in genome annotation. *Briefings in Bioinformatics* 25:bbae138