HRannot: An accurate and user-friendly gene annotation pipeline for vertebrates

Siwen Wu[1,2*], Sisi Yuan[2*], Wei-ming Li[3*], Zhengchang Su[2#]

[1]Institute of Medical Biology, Chinese Academy of Medical Sciences, Peking Union Medical College, Kunming, Yunnan, China.

[2]Department of Bioinformatics and Genomics, the University of North Carolina at Charlotte, Charlotte, NC 28223, USA.

[3]The Second Affiliated Hospital of Kunming Medical University, Kunming, Yunnan, China.

*These authors contributed equally to the work.

[#]The author supervised this work and should be addressed to zcsu@uncc.edu (ZS).

Abstract

With the development of sequencing technology, genome assembly becomes more and more easy for individual labs. Many high-quality assemblies for various species have been released in NCBI these years, however, most of them lack annotations of protein-coding genes and pseudogenes due to the sophistication of the process. Although there are many gene annotation tools, they cannot either achieve accurate annotation or are not available for individual labs due to the complexity of their use and required computational resources. In this paper, we developed an accurate gene annotation pipeline HRannot, which could annotate both protein-coding genes and pseudogenes comprehensively for vertebrates, and it is also easy to use by individual labs. Based on both **h**omology method and **R**NA-seq data, HRannot outperforms all the existing *de novo*, homology and RNA-seq based gene annotation tools, and is also comparable or even performs better in some cases than NCBI's gene annotation pipeline. By using HRannot, we also helped to annotate six representative chicken genome assemblies released in NCBI which lack annotations, providing available gene annotation resources for more chicken breeds.

Keywords

Gene annotation, homology-based, RNA-seq data-based, protein-coding genes, pseudogenes.

Introduction

Rapid development of sequencing technologies has made genome assembly more and more accurate and achievable in individual labs. Since 1977, Sanger sequencing [1] had been the most commonly used technology to sequence genomes before the advent of next-generation sequencing (NGS) technology [2] developed in 2005. Since 2014, the single molecule real time (SMRT) sequencing platform [3] and Oxford Nanopore Technology (ONT) [4, 5] were able to produce reads up to 100 kbp in length, making genome assembly further gapless. Since 2018, the Earth BioGenome Project (EBP) aims to sequence and characterize the genomes of all eukaryotes on Earth over a period of 10 years [6], providing unprecedented opportunities to study the biological process of eukaryotes on Earth. In 2021, the Vertebrate Genomes Project (VGP) plans to assembly the genomes for all the roughly 70,000 extant vertebrate species, enabling a new era of discovery across the life sciences [7]. All these situations suggest that there are more and more available genome assemblies released by various consortium. However, due to the complexity of annotating protein-coding genes and pseudogenes, most of the assemblies lack gene annotation results. Until 2023, only 23% of 28,754 species released in NCBI were listed with the annotation of at least one annotated Coding Sequence (CDS) [8].

As we all known, subsequent annotation of all genes encoded or pseudogenized in the genome of an organism is important for the understanding of the biology and evolutionary history of the organism. To solve the problems related to annotation, many gene annotation tools have been developed these years, including *de novo* -based methods, RNA-seq-based methods and homology-based methods. However, none of these tools could sufficiently and accurately annotate both protein-coding genes and pseudogenes. For example, *de novo* gene

annotation tools AUGUSTUS [9], GENSCAN [10] and GeneMark-ES [11], based on a hidden Markov model, probabilistic model and self-training model, respectively, tend to cause high false-positive and false-negative rates; RNA-seq gene annotation tools Braker3 [12], CAARs [13] and FINDER [14], based on the combination of AUGUSTUS [9] and GeneMark-ETP [15], gene family alignments and raw sequence reads, respectively, often bias to high-expression genes, i.e. only genes expressed in the sampled tissues or experimental conditions will be detected; Homology gene annotation tools Genomix [16], GeMoMa [17] and Scipio [18], based on evolutionary conservation of sequence, intron position conservation and multigenic families, respectively, are likely to miss new genes that are uniquely exist in the target species. Although NCBI [19] and ENSEMBL [20] have developed annotation pipelines which performs well for predicting both protein-coding and pseudogenes, they are not available to individual research groups and are difficult to follow. In addition, their pipelines often need many manually correction for the annotations.

To fill these gaps, we developed an accurate and user-friendly gene annotation pipeline named HRannot. By using a combination of homology-based and RNA-seq data-based approach, our pipeline could annotate protein-coding genes and pseudogenes accurately for vertebrates. By comparing with the existing gene annotation tools, we found that HRannot outperforms these tools and is also comparable or even performs better in some cases than NCBI's annotation pipeline. Our previous studies in chicken [21] also demonstrate the reliability and usefulness of HRannot, thus in this paper, we also helped to annotate six representative genomes of different chicken breeds released in NCBI by using HRannot, providing researchers with good gene annotation resources to study the biological process for chicken.

Methods and Materials

Overall flowchart of HRannot

By using a combination of homology-based and RNA-seq-based method, HRannot could annotate both protein-coding genes and pseudogenes. The overall flowchart of HRannot is shown in Figure 1. Specifically, there are two steps of both homology-based and RNA-seq-based methods, including data preparations and gene finding procedures.

Procedure of Homology-based method

For the data preparation step, we firstly collected as many as possible well-annotated homologous genomes and then mapped their corresponding CDSs to the target assembly using Splign (2.0.0) with default settings [22]. For each template gene whose CDSs can be mapped to the same contig/scaffold/chromosome, we concatenate all the mapped genomic regions to form putative full-length CDSs, and predict it to be an intact gene by including its introns (intervals of CDSs), if and only if its length is an integer time of three and the last three nucleotides form a stop codon but no stop codon appeared in the middle of the sequence (a true open reading frame (ORF)). If the putative full-length CDSs contains a premature stop codon in its middle (nonsense mutation), or its length is not an integer time of three (ORF shift mutation), we verify the mutations by mapping the high-quality DNA sequencing reads (could be Illumina DNA short reads or HiFi long reads) from the same individual to the genomic locus using bowtie (2.4.1) [23] with no gaps and mismatches permitted. If the locus can be completely covered by at least 10 short reads at each nucleotide position (the number can be determined by users, we used the average genome coverage/4 in this study of different

species), we consider the pseudogenization (nonsense mutation or ORF shift mutation) is fully supported by the high-quality DNA sequencing reads, and predict the sequence to be a pseudogene. Otherwise, we consider the pseudogenization is not supported by the high-quality sequencing reads, and call the sequence a partially supported gene, because the pseudogenization might be artificially caused by errors of the target assembly. Finally, all intact genes and partially supported genes are considered as homology-based protein-coding genes (Figure 1).

Since our HRannot supports the mapping results of Splign [22] from CDSs of multiple homologous genomes as the input, when determining putative genes for each locus, we firstly sort the homologous genomes according to their average mapping identity to the target genome in the descending order, and then ensure putative gene/pseudogene set according to the mapping results of the first genome according to the methods mentioned above. For each putative gene/pseudogene in the current gene set, we update its coordinates when there is a set of mapping in the the next homologous genomes whose average mapping identity is higher than the average mapping identity of the current gene/pseudogene and the loci from mapping start to mapping end overlaps more than 0.95 (the number could be set by users) with the current gene/pseudogene. If the mapping loci of the next homologous genomes have no overlap with any of the genes/pseudogenes in the current gene set, and its average mapping identity is greater than 0.99 (the number could be set by users), we added the mapping loci in the current gene sets. The process is cyclical until the mapping results of all homologous genomes has been checked.

Procedure of RNA-seq-based method

For the data preparations of RNA-seq-based method, we firstly collected as many as possible RNA-seq data from various tissues of the target species or their homologous species, and then mapped all the RNA-seq reads to the rRNA database [25] using bowtie (2.4.1) [23] to remove potential rRNAs. At the same time, we used infernal (1.1.2) [26] to examine the target assembly against Rfam database [27] to annotate non-coding RNAs. We then aligned the cleaning reads to the target assembly using STAR (2.7.0c) [28] with default settings and assembled transcripts using Trinity (2.8.5) [29] with its genome-guided option according to the mapping results. We next mapped the well-assembled transcripts to the target assembly using Splign (2.0.0) [22] with default settings, and removed those that at least partially overlapped non-coding RNAs, protein-coding genes or pseudogenes predicted by the homology-based method. For each of the remaining mappings, we used the same approach mentioned in the homology-based method to get the putative full-length CDSs. If it contained at least one ORF, the longest one was at least 300 bp (the number could be determined by users), and the average identity score of the mapping is at least 0.985 (the number could be determine by users), we predicted the locus corresponding to the longest CDSs as a protein-coding gene (Figure 1).

Annotations of the genomes of human, chicken, mouse and turtle using different tools

For the annotations of these different species, the target assemblies, homologous genomes, corresponding RNA-seq data and high-quality DNA sequencing data used in this study are summarized in Table S1. During annotations, for our HRannot, we used the parameters mentioned above in each step to annotate the genomes.

For AUGUSTUS [9], we used its *de novo* function and default settings. For Braker3 [12], we used its ETP mode (annotate using both RNA-seq data and protein data) with default

settings. For GeMoMa [17], we used both RNA-seq data and homologous annotations with its default settings. Specifically, for the homologous genome used in both HRannot and GeMoMa [17], for human and mouse, we used all the available annotations of mammalian (totally 242 genomes) released in NCBI RefSeq (Table S2); For chicken, we used all the available annotations of avian (totally 145 genomes) released in NCBI RefSeq (Table S3); For turtle, we used all the available annotations of reptile (totally 47 genomes) released in NCBI RefSeq (Table S4). For the protein data used in Braker3 [12], we used all the proteins corresponding to the homologous annotations used in HRannot and GeMoMa [17].

Annotation of the genome of a Tibetan ground-tit individual

For the annotation of the Tibetan ground-tit individual, the target assembly, homologous assembly, corresponding RNA-seq data and high-quality DNA sequencing data used in this study are summarized in Table S1. Similarly, we used all the available annotations of avian (totally 145 genomes) released in NCBI RefSeq (Table S3) as the homologous genomes. For HRannot, we used the parameters mentioned above in each step to annotate the genome.

Annotations of the genome of the six chicken breeds

For the annotations of the six chicken breeds, the target assembly, homologous assembly, corresponding RNA-seq data and high-quality DNA sequencing data used in this study are summarized in Table S1. Similarly, we used all the available annotations of avian (totally 145 genomes) released in NCBI RefSeq (Table S3) as the homologous genomes. For HRannot, we used the parameters mentioned above in each step to annotate the genome.

Assessment of genes overlapping with NCBI's annotation

For the genes/pseudogenes annotated by different tools, we refer these overlap more than 50% with genes/pseudogenes from NCBI's annotation and identity of the mapping parts were 100% compared with genes/pseudogenes of NCBI's annotation as the genes overlapping with NCBI's annotation.

BUSCO assessment

We used the CDSs and exons of the protein-coding genes and pseudogenes annotated in each species against BUSCO (5.7.1) [30] to assess the quality of the annotation results from different tools.

For the database used in the BUSCO (5.7.1) [30] assessment, for human and mouse, we used mammalia_odb10; For chicken and the other birds, we used aves_odb10; For turtle, we used tetrapoda_odb10.

Precision and Recall assessment

For the Precision and Recall assessment for each tool, we used the following formulas:

Precision = # Overlapped genes / # Predicted genes

Recall = # Overlapped genes / # Genes predicted by NCBI

Results

HRannot outperforms the existing gene annotation tools

We used AUGUSTUS [9], Braker3 [12], GeMoMa [17], which represents the annotation methods of *de novo* based, RNA-seq + protein based, Homology + RNA-seq based, respecitvely, and our HRannot to annotate the protein-coding genes and pseudogenes for four different vertebrate species, including human, chicken, mouse and turtle, and compared the annotation results with

those from NCBI. As shown in Table 1, for the annotations of all the four species, HRannot could achieve the nearest number of protein-coding genes (20,717, 18,272, 20,624 and 19,670, for human, chicken, mouse and turtle, respectively) compared with NCBI's annotation (20,090, 17,485, 20,482 and 20,259, for human, chicken, mouse, turtle, respectively). By contrast, GeMoMa [17] predicted the largest number of protein-coding genes for each species (59,002, 29,054, 39,398 and 38,766 for human, chicken, mouse, turtle, respectively), which is 1.7-2.9 times higher than those of NCBI's annotation (Table 1). For AUGUSTUS [9] and Braker3 [12], both of them had different levels of discrepancy compared with NCBI's annotations (Table 1).

For the protein-coding genes overlapped with NCBI's annotation, AUGUSTUS [9] achieved the highest number in the annotations of human and turtle, which is 15,424 and 15,955, respectively, with HRannot achieved the second highest number, which is 15,340 and 14,862, respectively (Table 1). In the annotations of chicken, GeMoMa [17] had the highest number (13,300) and HRannot had the second highest number (13,071), with the fact that GeMoMa [17] predicted 10,782 more genes than HRannot (Table 1). In mouse, HRannot got the highest number (15,667), and GeMoMa [17] got the second highest number (15,385) (Table 1). By comparing the precision and recall of these annotations, we found that AUGUSTUS [9] tends to have high precision but relatively low recall, Braker3 [12] and GeMoMa [17] tends to have high recall but relatively low precision, and HRannot always had the most balanced performance on precision and recall compared with the other tools (Table 1, Figure 2).

For the BUSCO [30] completeness assessment, HRannot and GeMoMa [17] could achieve more than 90% completeness for all the annotations of the four species (Table 1), suggesting that the annotations of these two tools are very complete. For Braker3 [12], its

BUSCO [30] completeness values are 81.4%-92.1% for the four species, relatively lower than those of HRannot and GeMoMa [17] (Table 1). For AUGUSTUS [9], its BUSCO [30] completeness values was 35.5%-71.9% (Table 1), suggesting that it cannot annotate genes completely for these species, although its precision and recall are relatively high compared to the other tools, which is because there are many fragmental annotations of the tool (Figure 3, Table S5). Specifically, HRannot achieved the highest BUSCO [30] completeness value in mouse and turtle (Table 1). In human, HRannot was only 0.1% lower than that of GeMoMa [17], with the fact that GeMoMa [17] predicted 38,285 more protein-coding genes than HRannot (Table 1). In chicken, the BUSCO [30] completeness value of GeMoMa [17] is 1% higher than that of HRannot, with the number of its protein-coding genes is 10,782 more than that annotated by HRannot (Table 1). By checking the details of the BUSCO [30] assessment results, we found that except for chicken, HRannot had the highest value of Complete and single-copy BUSCOs (Figure 3, Table S5), suggesting that there is little over-prediction in the annotation. The Complete and duplicated BUSCOs value is lower than 1% for HRannot in all the four annotations (Figure 3, Table S5), suggesting again that HRannot could achieve a complete but non-redundant annotations.

In addition, except for HRannot, all the other three tools could not annotate pseudogenes for the target assembly. As shown in Table 1, HRannot predicted 784, 541, 56 and 1,952 pseudogenes for human, chicken, mouse and turtle, respectively, with 134, 9, 24 and 170 overlapped with NCBI's annotation for pseudogenes of the four species, respectively. Notably, all the pseudogenes predicted by HRannot were supported by several high-quality sequencing reads of the corresponding individual of the species (Methods and Materials), suggesting that

the pseudogenes predicted by HRannot are reliable. However, since HRannot mainly predicted unprocessed pseudogenes, and NCBI mainly predicted process pseudogenes, there are a few overlaps between the pseudogenes predicted by the two tools.

HRannot are comparable or even performs better in some cases than NCBI's annotation pipeline

To further demonstrate the prediction accuracy of HRannot, we annotated the genome of a Tibetan ground-tit individual (Table S1) by using HRannot and compared the results with the annotations from NCBI. By using the CDS isoforms of 145 well-annotated avian species from NCBI RefSeq (Table S3) as the homologous template, the homology-based method of HRannot predicted 15,352 intact genes, 448 partially supported genes and 81 pseudogenes (Table 2). By using the RNA-seq reads from multiple tissues of the individual (Table S1), the RNA-seq-based method annotated 102 new genes for the individual (Table 2). In total, HRannot predicted 15,902 protein-coding genes and 81 pseudogenes for the Tibetan ground-tit individual (Table 2), while NCBI annotated 15,814 protein-coding genes and 83 pseudogenes for it (Table 2). There are 14,030 overlapped protein-coding genes predicted by the two pipelines, suggesting that HRannot could annotate most of the protein-coding genes (88.7%) predicted by NCBI's annotation pipeline (Table 2). Besides, there is no overlapped peudogenes predicted by the two pipelines (Table 2). For the BUSCO [30] completeness value, HRannot is similar to that of NCBI's annotation, which are 99.0% and 99.5%, respectively (Table 2), both of which are of really high complete.

For the inconsistent protein-coding genes and pseudogenes predicted by the two pipelines, we found that NCBI's pipeline could mistakenly predict protein-coding genes as

pseudogenes, or mistakenly predict pseudogenes as protein-coding genes. For examples, genes *PTCD*2 and *DIO*3 are annotated as protein-coding genes by NCBI's pipeline, while they are annotated as pseudogenes by HRannot since there are stop codons in the middle of their ORFs (Figures 4a and 4b); gene *LOC*102109485 (named *LOC*107212260 in our annotation) is annotated as pseudogenes by NCBI's pipeline, while it is annotated as protein-coding genes by HRannot since there are neither stop codons nor ORF shift mutations in its ORFs (Figure 4c). Thus, these results suggest that our HRannot is at least comparable to NCBI's pipeline in accuracy, and even more sensitive in some cases than NCBI's pipeline.

Annotation of six representative chicken genomes in NCBI

In NCBI, there are several genome assemblies for different chicken breeds which lack gene annotations. To provide available annotations for these assemblies, we annotated six representative chicken genomes which were assembled at chromosome-level released in NCBI by using HRannot. As shown in Table 3, we annotated 16,827-17,544 homology-based genes and 604-797 RNA-seq-based genes in each of the assemblies. In total, there are 17,434-18,218 protein-coding genes and 923-1,128 pseudogenes in the chicken breeds of Huxu, Cobb, Ross, Hailanhe, Silkie and Lueyang (Table 3). To further verify the accuracy and completeness of the annotations, we used BUSCO completeness value [30] to evaluate the annotation results. As shown in Table 3 and Figure 3, all of annotations achieved more than 97% BUSCO completeness, with smaller than 0.5% Complete and duplicated BUSCOs, suggesting that HRannot has achieved complete and non-redundant annotations for the six chicken breeds.

Discussion

We developed a homology- and RNA-seq-based gene annotation pipeline HRannot that demonstrates superior performance compared to several established tools, including AUGUSTUS [9], Braker3 [12] and GeMoMa [17], three popular and widely used gene annotation tools. Through comprehensive benchmarking, our pipeline consistently achieved higher annotation accuracy, as evidenced by a BUSCO completeness score [30] exceeding 90% across diverse test genomes (Table 1). Importantly, our method maintains a well-balanced trade-off between precision and recall, addressing a common limitation observed in other annotation tools that often sacrifice one for the other (Figure 2). While Braker3 [12] and GeMoMa [17] rely heavily on homology and transcript evidence, respectively, our pipeline integrates both sources more effectively, resulting in improved gene model prediction. Compared to AUGUSTUS [9], which primarily depends on *ab initio* prediction, our pipeline shows substantial gains in both structural accuracy and biological relevance. These results underscore the robustness and reliability of our approach, making it particularly valuable for annotating newly sequenced or non-model genomes.

Besides, HRannot is also comparable or even more sensitive in some cases compared with NCBI's gene annotation pipeline, i.e. correcting some pseudogenes or protein-coding genes that are mistakely annotated as the other type by NCBI's pipeline (Figure 4). In addition, a particularly advantage of our pipeline lies in its ability to identify unprocessed pseudogenes, which retain intron-exon structures and closely resemble their functional counterparts. In contrast, the NCBI pipeline primarily annotates processed pseudogenes, which lack introns and often arise from retrotransposition. The detection of unprocessed pseudogenes is especially challenging due to their structural similarity to active genes and their typically low expression

levels. Our pipeline's success in capturing these elements highlights its enhanced sensitivity and specificity in distinguishing subtle genomic features. Although the BUSCO [30] completeness value of NCBI's annotations are always higher than those of HRannot, there are always many manually correction of NCBI's annotation, while HRannot is completely automatic, causing the relatively lower BUSCO [30] completeness value.

Given its demonstrated effectiveness, our pipeline HRannot offers a compelling alternative to current standards and holds promise for broad applications in genomics research. The accurate annotations for the six representative chicken breeds could also provide foundational resources for the study related to chicken breeds.

Declarations

Consent for publication
Not applicable.

Availability of data and materials
The code and pipeline description are available at https://github.com/zhengchangsulab/HRannot.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
ZS supervised and conceived the project; SW, SY and ZS designed the pipeline; SW and ZS performed data analysis; and SW and ZS wrote the manuscript.

References

1. Sanger, F., S. Nicklen, and A.R. Coulson, *DNA sequencing with chain-terminating inhibitors.* Proc Natl Acad Sci U S A, 1977. **74**(12): p. 5463-7.
2. Mardis, E.R., *Next-generation DNA sequencing methods.* Annu Rev Genomics Hum Genet, 2008. **9**: p. 387-402.

3. van Dijk, E.L., et al., *Ten years of next-generation sequencing technology.* Trends Genet, 2014. **30**(9): p. 418-26.
4. Clarke, J., et al., *Continuous base identification for single-molecule nanopore DNA sequencing.* Nat Nanotechnol, 2009. **4**(4): p. 265-70.
5. Eisenstein, M., *Oxford Nanopore announcement sets sequencing sector abuzz.* Nat Biotechnol, 2012. **30**(4): p. 295-6.
6. Lewin, H.A., et al., *Earth BioGenome Project: Sequencing life for the future of life.* Proc Natl Acad Sci U S A, 2018. **115**(17): p. 4325-4333.
7. Rhie, A., et al., *Towards complete and error-free genome assemblies of all vertebrate species.* Nature, 2021. **592**(7856): p. 737-746.
8. Brůna, T., et al., *Galba: genome annotation with miniprot and AUGUSTUS.* BMC Bioinformatics, 2023. **24**(1): p. 327.
9. Stanke, M. and S. Waack, *Gene prediction with a hidden Markov model and a new intron submodel.* Bioinformatics, 2003. **19 Suppl 2**: p. ii215-25.
10. Burge, C. and S. Karlin, *Prediction of complete gene structures in human genomic DNA.* J Mol Biol, 1997. **268**(1): p. 78-94.
11. Lomsadze, A., et al., *Gene identification in novel eukaryotic genomes by self-training algorithm.* Nucleic Acids Res, 2005. **33**(20): p. 6494-506.
12. Gabriel, L., et al., *BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA.* Genome Res, 2024. **34**(5): p. 769-777.
13. Rey, C., et al., *CAARS: comparative assembly and annotation of RNA-Seq data.* Bioinformatics, 2019. **35**(13): p. 2199-2207.
14. Banerjee, S., et al., *FINDER: an automated software package to annotate eukaryotic genes from RNA-Seq data and associated protein sequences.* BMC Bioinformatics, 2021. **22**(1): p. 205.
15. Brůna, T., A. Lomsadze, and M. Borodovsky, *GeneMark-ETP significantly improves the accuracy of automatic annotation of large eukaryotic genomes.* Genome Res, 2024. **34**(5): p. 757-768.
16. Coghlan, A. and R. Durbin, *Genomix: a method for combining gene-finders' predictions, which uses evolutionary conservation of sequence and intron-exon structure.* Bioinformatics, 2007. **23**(12): p. 1468-75.
17. Keilwagen, J., F. Hartung, and J. Grau, *GeMoMa: Homology-Based Gene Prediction Utilizing Intron Position Conservation and RNA-seq Data.* Methods Mol Biol, 2019. **1962**: p. 161-177.
18. Kollmar, M., *Predicting Genes in Closely Related Species with Scipio and WebScipio.* Methods Mol Biol, 2019. **1962**: p. 193-206.
19. NCBI, https://www.ncbi.nlm.nih.gov/refseq/annotation_euk/process/#references.
20. ENSEMBL, https://useast.ensembl.org/info/genome/genebuild/index.html.
21. Wu, S., et al., *Annotations of four high-quality indigenous chicken genomes identify more than one thousand missing genes in subtelomeric regions and micro-chromosomes with high G/C contents.* BMC Genomics, 2024. **25**(1): p. 430.
22. Kapustin, Y., et al., *Splign: algorithms for computing spliced alignments with identification of paralogs.* Biol Direct, 2008. **3**: p. 20.

23. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nat Methods, 2012. **9**(4): p. 357-9.
24. Chen, M.M., et al., *The GFF3toolkit: QC and Merge Pipeline for Genome Annotation.* Methods Mol Biol, 2019. **1858**: p. 75-87.
25. Quast, C., et al., *The SILVA ribosomal RNA gene database project: improved data processing and web-based tools.* Nucleic Acids Res, 2013. **41**(Database issue): p. D590-6.
26. Nawrocki, E.P. and S.R. Eddy, *Infernal 1.1: 100-fold faster RNA homology searches.* Bioinformatics, 2013. **29**(22): p. 2933-5.
27. Kalvari, I., et al., *Rfam 14: expanded coverage of metagenomic, viral and microRNA families.* Nucleic Acids Res, 2021. **49**(D1): p. D192-d200.
28. Dobin, A., et al., *STAR: ultrafast universal RNA-seq aligner.* Bioinformatics, 2013. **29**(1): p. 15-21.
29. Grabherr, M.G., et al., *Full-length transcriptome assembly from RNA-Seq data without a reference genome.* Nat Biotechnol, 2011. **29**(7): p. 644-52.
30. Simão, F.A., et al., *BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs.* Bioinformatics, 2015. **31**(19): p. 3210-2.

Table 1: Comparison of the annotations of different tools from several aspects

Species	Category	HRannot		AUGUSTUS		Braker3		GeMoMa		NCBI	
		Protein-coding gene	Pseudogene	Protein-coding gene	Pseudogene	Protein-coding gene	Pseudogene	Protein-coding gene	Pseudogene	Protein-coding gene	Pseudogene
Human	# Annotation by the tool	20717	784	23489	-	24869	-	59002	-	20090	15534
	# Overlap with NCBI's annotation	15340	134	15424	-	11425	-	13456	-	-	-
	BUSCO completeness	91.1%	-	61.3%	-	81.4%	-	91.2%	-	99.3%	-
Chicken	# Annotation by the tool	18272	541	15898	-	18984	-	29054	-	17485	240
	# Overlap with NCBI's annotation	13071	9	12370	-	12080	-	13300	-	-	-
	BUSCO completeness	96.7%	-	71.9%	-	92.1%	-	97.7%	-	97.7%	-
Mouse	# Annotation by the tool	20624	56	10897	-	18848	-	39398	-	20482	5091
	# Overlap with NCBI's annotation	15667	24	9116	-	13361	-	15385	-	-	-
	BUSCO completeness	93.2%	-	35.5%	-	86.0%	-	92.5%	-	98.4%	-
Turtle	# Annotation by the tool	19670	1952	19641	-	20848	-	38766	-	20259	705
	# Overlap with NCBI's annotation	14862	170	15955	-	12956	-	13309	-	-	-
	BUSCO completeness	94.7%	-	53.7%	-	89.7%	-	90.9%	-	97.8%	-

Table 2: Comparison of gene annotation results of the Tibetan ground-tit genomes by HRannot and NCBI's pipeline

Methods	Homology-based			# RNA-based genes	# Total protein-coding genes	# Total pseudogenes	BUSCO completeness
	# Intact genes	# Partially supported genes	# Pseudogenes				
HRannot	15,352	448	81	102	15,902	81	99.0%
NCBI	-	-	-	-	15,814	83	99.5%
Common	-	-	-	-	14,030	0	-

Table 3: Annotations of six representative chicken genomes released in NCBI

Assembly	Breeds	# Protein-coding genes		# Pseudogenes	BUSCO completeness
		# Homology-based	# RNA-seq-based		
GGswu	Huxu	17544	674	923	97.2%
bGalGal5.pri	Cobb	17200	604	945	97.3%
bGalGal4.pri	Ross	16827	607	1126	97.3%
ASM3450984v1	Hailanhe	17281	681	961	97.2%
ASM3450986v1	Silkie	16886	755	1128	97.2%
ASM3450988v1	Lueyang	17223	797	972	97.2%

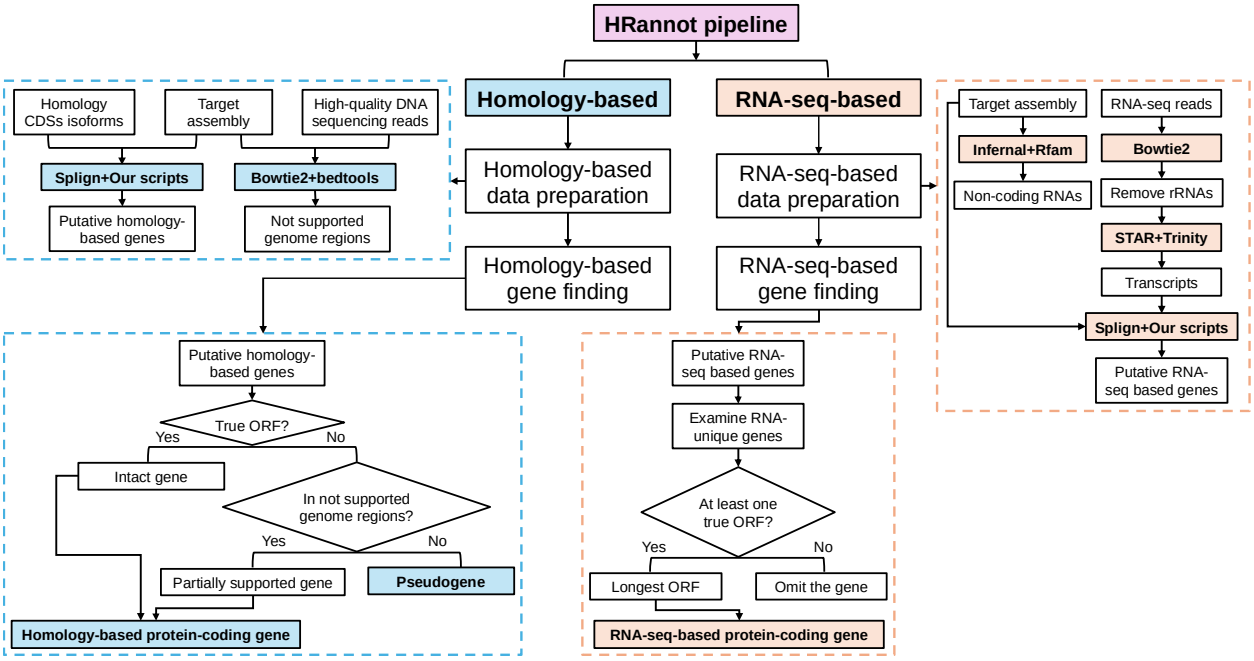

Figure 1. Flowchart of the HRannot pipeline

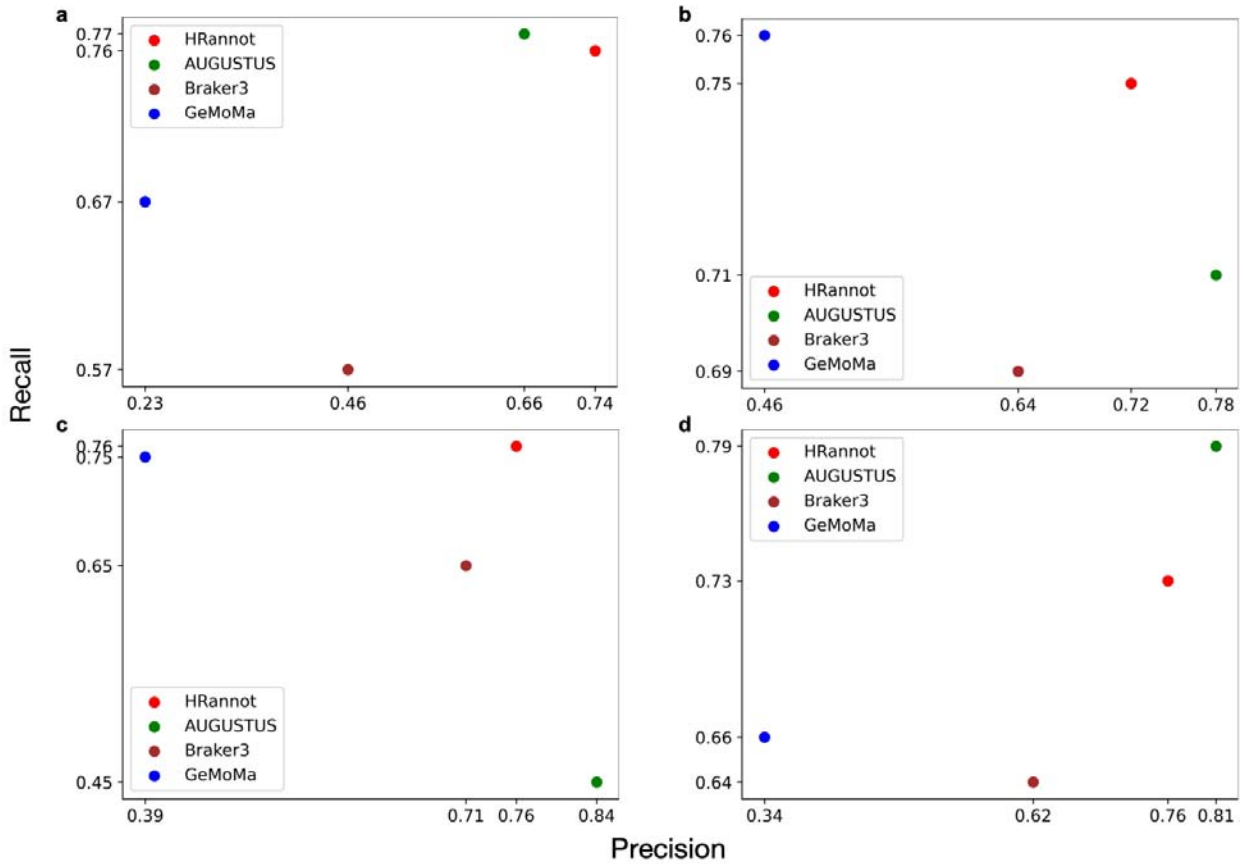

Figure 2. Precision and Recall assessment of the annotations from different tools for human, chicken, mouse and turtle, respectively (a-d).

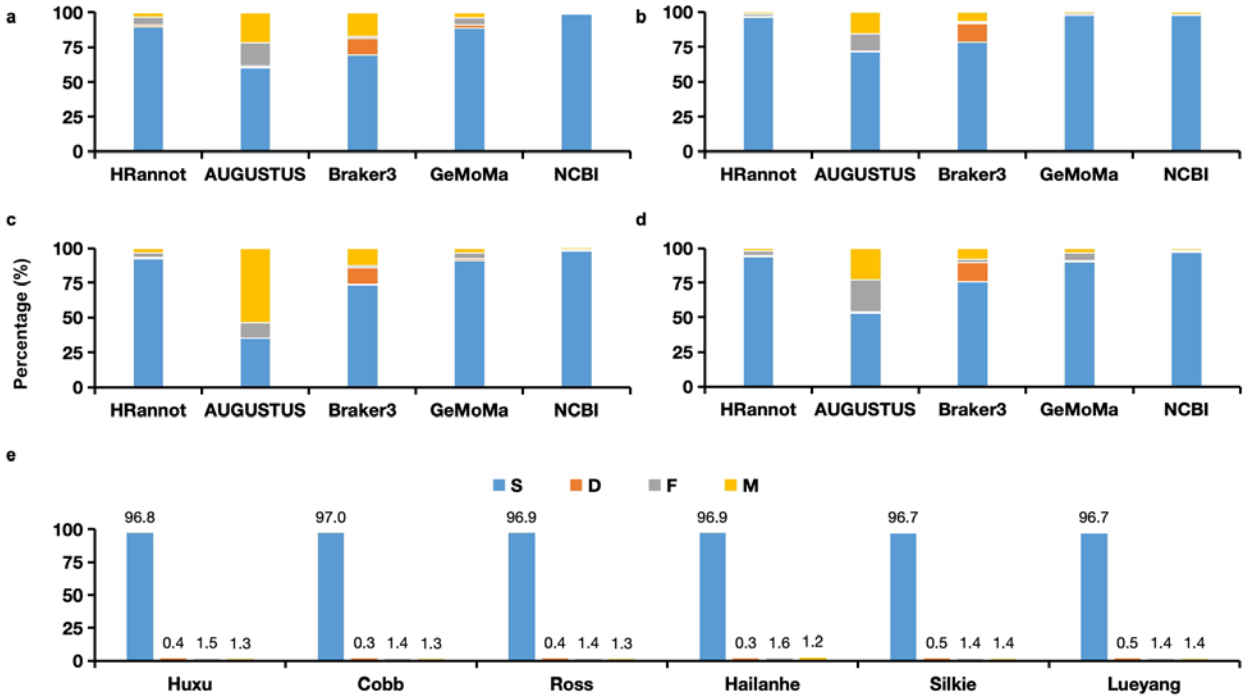

Figure 3. BUSCO completeness of the annotations for different species. **a-d.** BUSCO completeness values of the annotations from different annotation tools for human, chicken, mouse and turtle, respectively. **e.** BUSCO completeness values for the annotations of the six different chicken breeds by using HRannot. In each subfigures, S represents Complete and single-copy BUSCOs, D represents Complete and duplicated BUSCOs, F represents Fragmented BUSCOs, and M represents Missing BUSCOs. Specifically, BUSCOs completeness in Tables mentioned above consists of S and D.

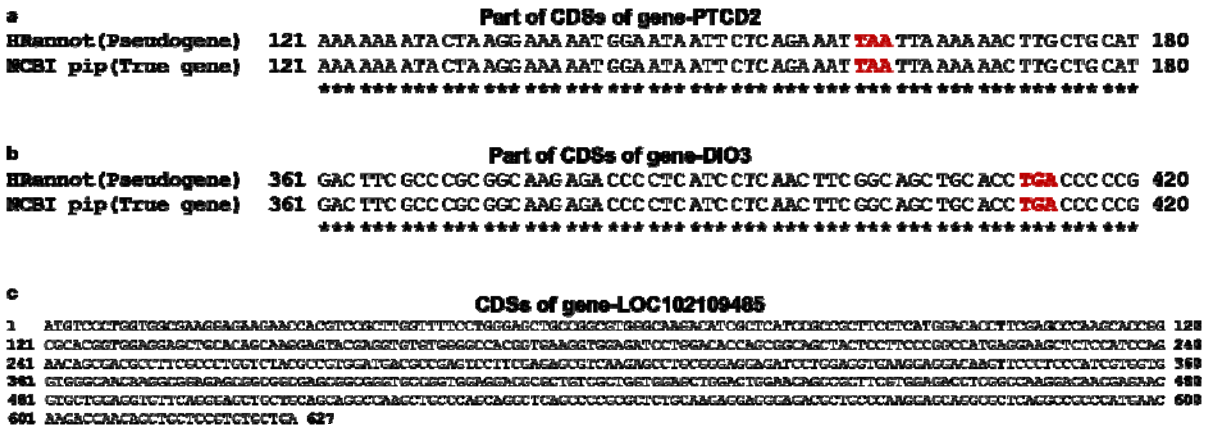

Figure 4. Examples of different annotations for the same gene/pseudogene from the two gene annotation pipelines. **a.** Part of ORFs of gene *PTCD*2 annotated by both the two pipelines contains a stop codon labeled in red. **b.** Part of ORFs of gene *DIO*3 annotated by both the two pipelines contains a stop codon labeled in red. **c.** The ORF of gene *LOC*102109485 annotated by the two pipelines. The sequence contains no pseudogenization mutations.