**HANNO**: efficient **H**igh-throughput **ANNO**tation of protein coding genes in eukaryote genomes

Heiner Kuhl1

1 Leibniz Institute of Freshwater Ecology and Inland Fisheries (IGB), Berlin, Germany

correspondence: [heiner.kuhl@igb-berlin.de](mailto:heiner.kuhl@igb-berlin.de)

**ABSTRACT**

**INTRODUCTION**

Gene annotation software plays a pivotal role in genomics, serving as a critical tool for the interpretation of raw genetic sequences. These programs automate the identification and functional characterization of genes within a genome, facilitating the understanding of biological functions and processes. By analyzing sequence data, gene annotation software predicts the locations of genes, determines their structure (untranslated regions (UTRs), coding sequence (CDS) and introns), and assigns potential functions based on homology to known genes. As genomic data continues to expand rapidly, gene annotation software remains indispensable for converting vast amounts of genetic information into meaningful biological insights. There are several prominent gene annotation approaches and software tools used widely in the field of genomics, each offering unique features and capabilities.

*Ab initio* methods typically only use intrinsic features of genomic sequence to predict gene models (open reading frames, splice-junctions, codon usage, GC-content etc.). Widely known tools for *ab initio* prediction are Genscan, GeneMark and Augustus (if used without hints). Those early *ab initio* methods had relatively low sensitivity and specificity compared to gene prediction methods that use evidence from protein homology and transcribed sequence. Recently, Helixer a tool that applies deep learning methods for *ab initio* gene prediction has shown improved accuracy.

Homology-based prediction methods generally provide more accurate gene models than *ab initio* methods, but require a known set of input proteins or mRNAs from a reasonably closely related, if not the same species for comparison with the genome of interest. Protein sequences are more conserved than nucleotide sequences due to the degeneracy of the genetic code and can therefore be used to predict genes between more distantly related organisms. Protein homology prediction is based on tools for aligning proteins with genomic sequences that take splice sites into account. The first generation of these tools, such as Genewise or Exonerate, were computationally intensive and a bottleneck for annotation pipelines. According to a recent study, only a few of the tools developed so far can be applied directly and at higher speed to complete vertebrate genomes, including Spaln2. The latest tool for protein splicing mapping, Miniprot, overcomes most of the limitations of its predecessors.

Another approach, namely to exploit existing annotations of reference genomes and transfer them to new genome assemblies, is called comparative annotation. Here, whole genome alignment of multiple assemblies allows the transformation of gene feature coordinates to the coordinates of the yet unannotated genomes. Important tools in this space are Cesar2 and Comparative Annotation Toolkit. One advantage of this approach is that ortholog genes between different genomes are well defined, on the other hand complex or repetitive regions of the genomes may be excluded from whole genome alignment and thus be missing in the annotation.

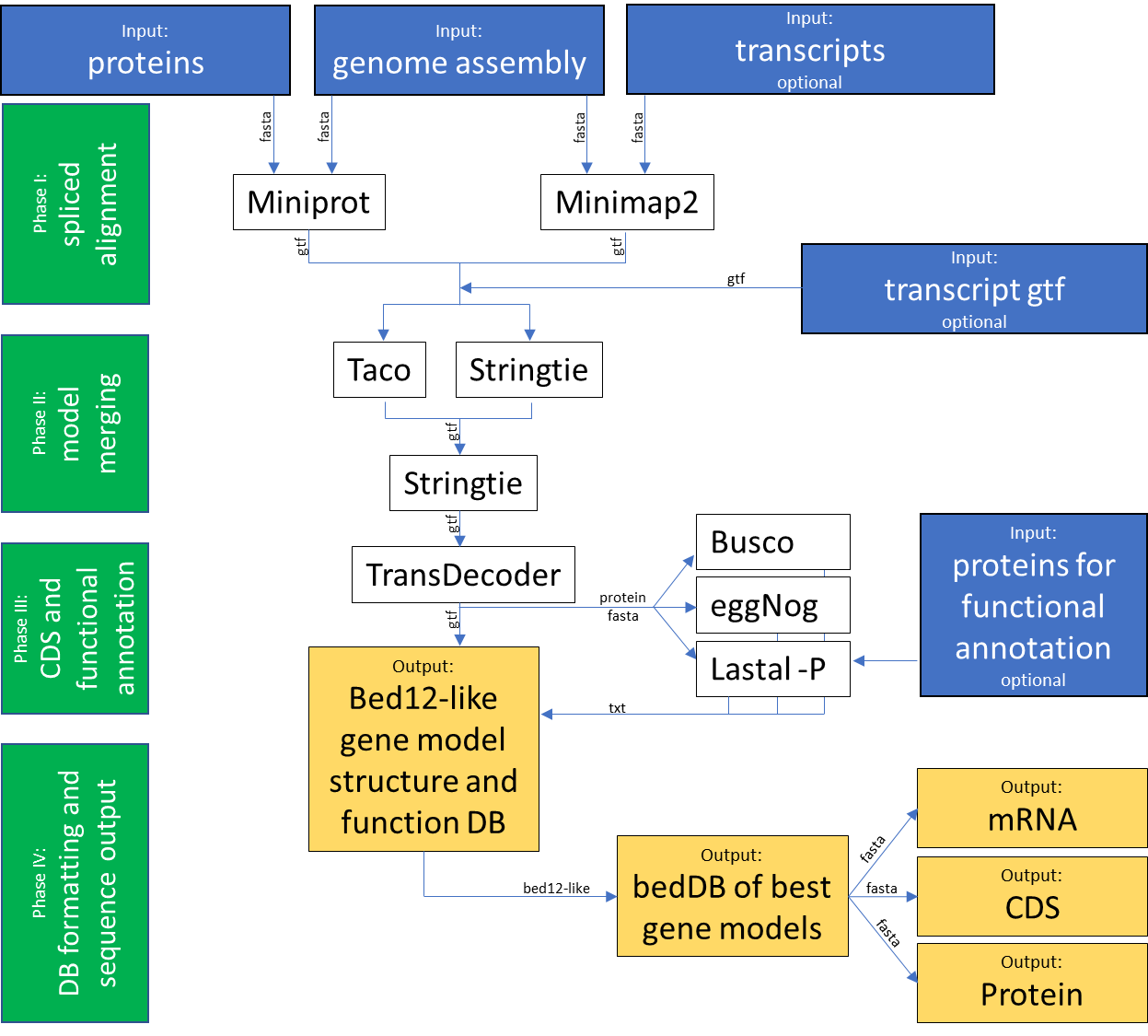
Several bioinformatic pipelines have implemented and combined the different approaches mentioned above, among them Braker3, Maker2, Pasa, the Ensemble annotation pipeline and NCBI GNOMON. These tools, among others that have recently been benchmarked, form the backbone of modern genomics research, enabling scientists to decode the functional elements of genomes with high precision and efficiency.

HANNO is a new annotation pipeline that utilizes state-of-the-art methods for spliced protein and transcript alignment (Miniprot and Minimap2) to enable high-throughput genome annotation of large eukaryote genomes within few hours. By integrating eggNog, it also provides extensive functional annotations. In the absence of transcriptome data of the target species for annotation, HANNO can be used to easily transfer NCBI REFSEQ annotations even from distantly related species. Therefore, HANNO provides high-speed annotation that is urgently needed in genome projects that are dealing with several thousand species.

**METHODS**

*Design of the HANNO pipeline*

Protein sequences are splice-aligned with the genome assembly using Miniprot with gtf output. Subsequently, transcript sequences (if provided) are splice-aligned with the genome assembly using Minimap2 (-x splice), supported by a splice junction file, generated from the prior protein alignments. Minimap2 sam output is converted to gtf format. The strand of the mRNA alignments in the gtf file is corrected using the information in the sam “ts:” fields, if necessary. Optionally, a Stringtie gtf file of reference guided transcript assemblies can be added by the user. The resulting gtf files of genomic exon coordinates from protein and transcript alignments are combined using StringTie and Taco and the genomic coordinates of CDS exons are calculated by TransDecoder (this step was parallelized to remove a significant bottleneck of the pipeline). All resulting gene-models are functionally annotated by eggNog, best protein matches (LAST aligner) and Busco. A single best gene model is chosen from a cluster of gene models according to scoring of its functional annotation or its CDS length (if no functional annotation assigned). The gene models are stored in a bed12-like database format that comprises all structural and functional annotation of one gene model in each line. This format is readable by IGV (use file suffix ”.bed”) and functional annotations of interest may also be displayed in IGV by concatenating them to the bed name field. HANNO consists of several PERL and AWK scripts. The main script “HANNO.v0.4.pl” is governing the whole pipeline shown in Figure 1. The current versions of the HANNO pipeline scripts, installation and usage instructions can be downloaded from <https://github.com/HMPNK/HANNO>.

****

**Figure 1:** Flow chart of the HANNO Pipeline.

*Benchmarking of HANNO on Fish, Amphibia, Bird and Mammal genomes*

To benchmark HANNO, high-quality Refseq annotations for a representative fish, amphibian, bird and mammal genome were chosen. Protein sequences and mRNA sequences for *Perca flavescens*, *Bufo bufo*, *Taeniopygia guttata* and *Homo sapiens* were downloaded (see Supplementary Data 1 for accessions and filenames). We benchmarked the completeness of the reference proteins by BUSCO using (actinopterygii\_odb9, vertebrata\_odb9, aves\_odb9 and mammalia\_odb10 ortholog databases) to corroborate their high-quality (overall >=98.0% complete and <=1.2% missing BUSCOs). In a first step HANNO was run on the very same genome assembly from which the annotations were derived, to analyze the performance under perfect conditions. Then HANNO was run to annotate genome assemblies of diverged species. These assemblies were chosen to cover a divergence time range that goes back up to 108 - 224MY for each clade according to [www.timetree.org](http://www.timetree.org) and had to fulfill highest standards regarding contig N50 / scaffold N50. The command lines for running HANNO on those genome assemblies are shown in Supplementary Data 1.

**RESULTS**

*HANNO runtime on vertebrate genomes*

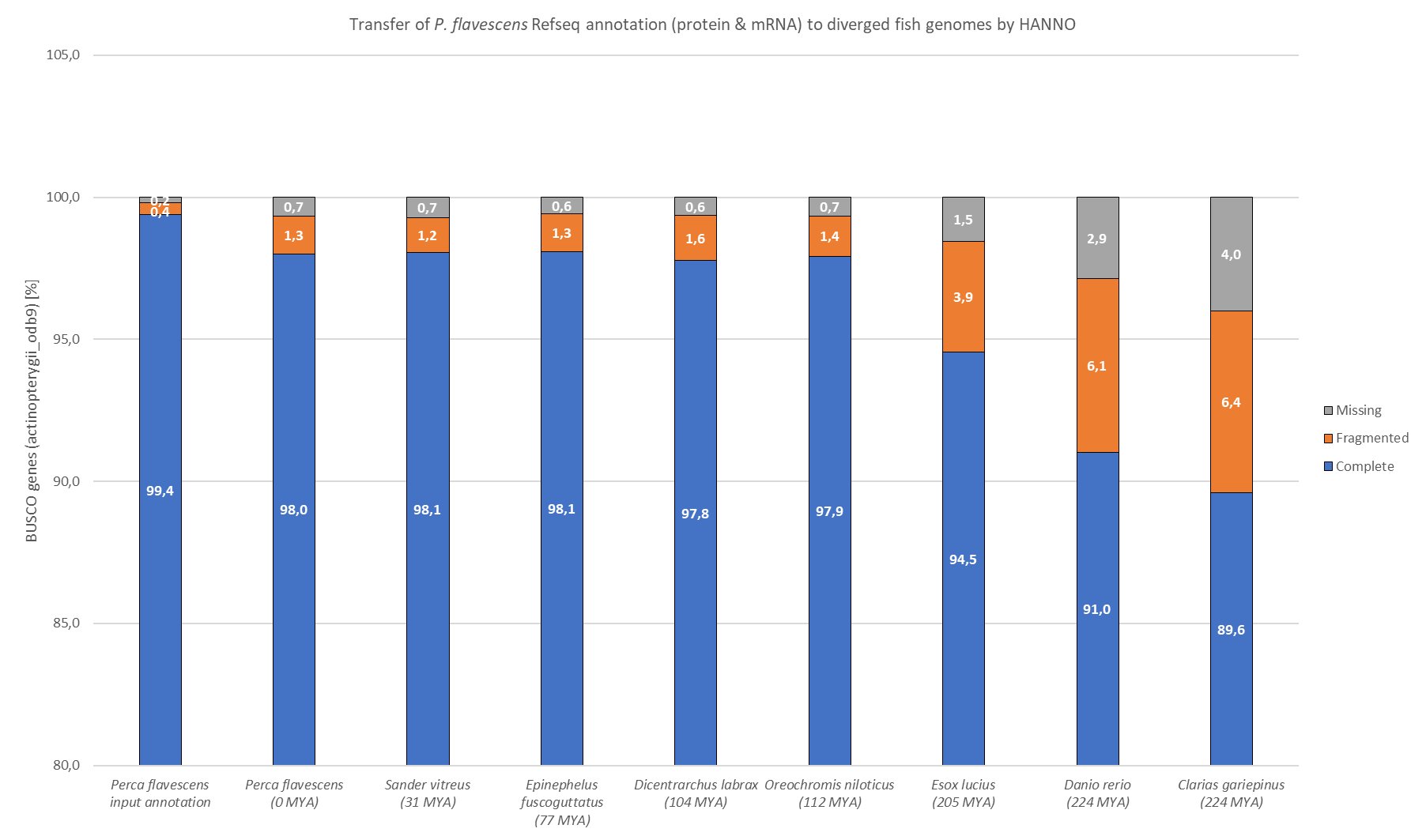
Due to the highly efficient splice alignment (miniprot, minimap2) and model merging (Stringtie, Taco) methods HANNO could finish test runs within a few hours on 6 - 12 CPU threads for different vertebrate clades (Table 1). In this regard our compute server (96 CPU threads, 1TB RAM) processed up to 11 genomes in parallel, which allowed us to annotate at a rate of about 3 genomes per hour for Fish, Amphibia and Birds. Mammals were outliers (1.5/h) due to using the large mammalia\_odb10 database with BUSCO. These numbers can be improved to even higher speed, if time consuming functional annotation steps like BUSCO and eggNog are optionally switched off.

**Table1:** HANNO benchmark runtimes on different vertebrate clades (HPC-server: Intel(R) Xeon(R) CPU E7-8890 v4 @ 2.20GHz; 96-threads; 1TB RAM). Note that we ran 6 (Birds, Amphibians), 8 (Fish) and 11 (Mammals) genome annotations in parallel.

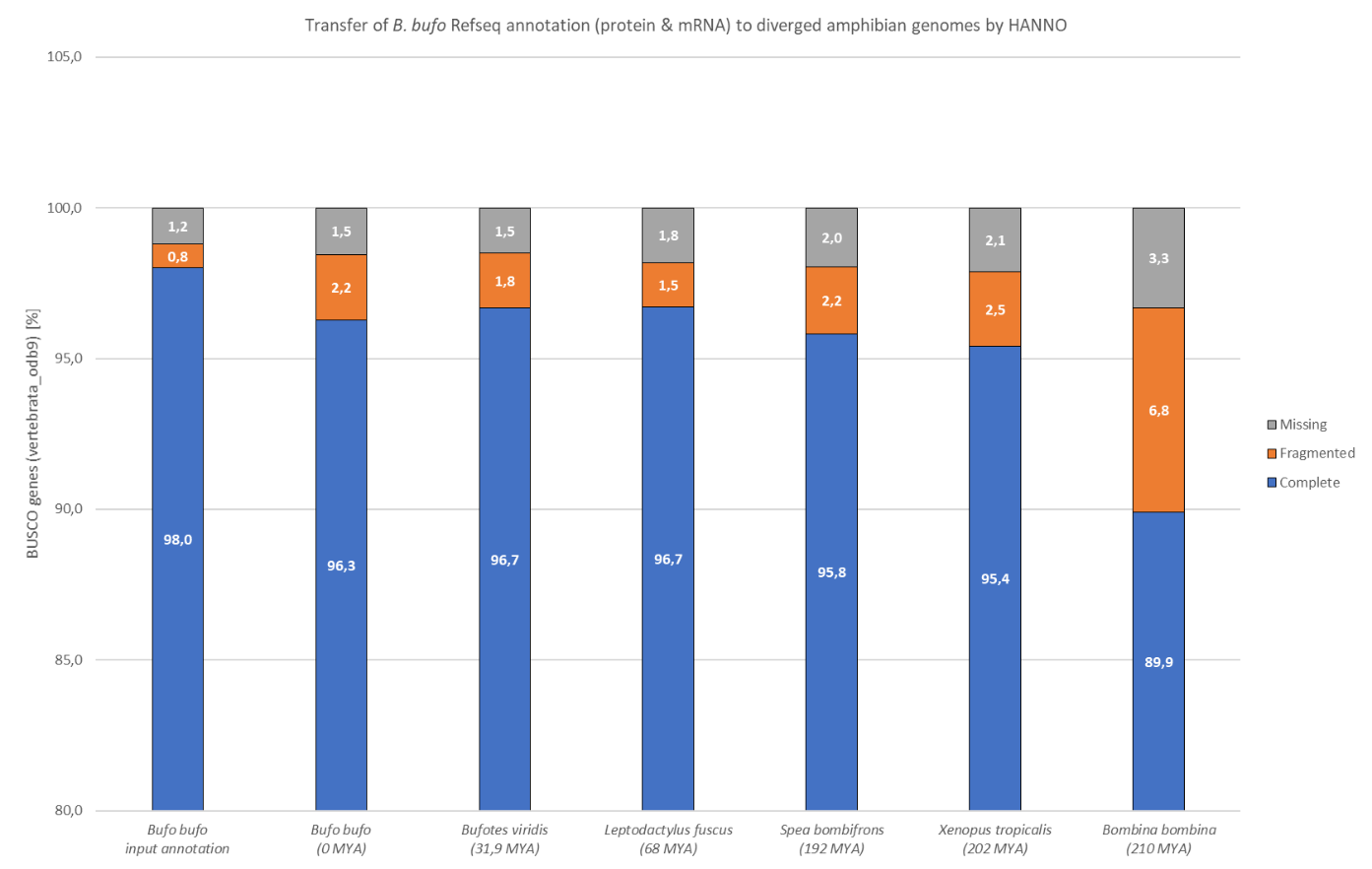
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **#threads used per genome** | **#parallel genome annotations** | **approx. wall-clock time** | **#genomes per hour** |
| **Fish** | 8 | 8 | 3h:00m | 2.67 |
| **Amphibians** | 12 | 6 | 2h:00m | 3.00 |
| **Birds** | 12 | 6 | 1h:45m | 3.43 |
| **Mammals** | 6 | 11 | 7h:20m | 1.50 |

*HANNO BUSCO completeness over divergence time for vertebrate genomes*

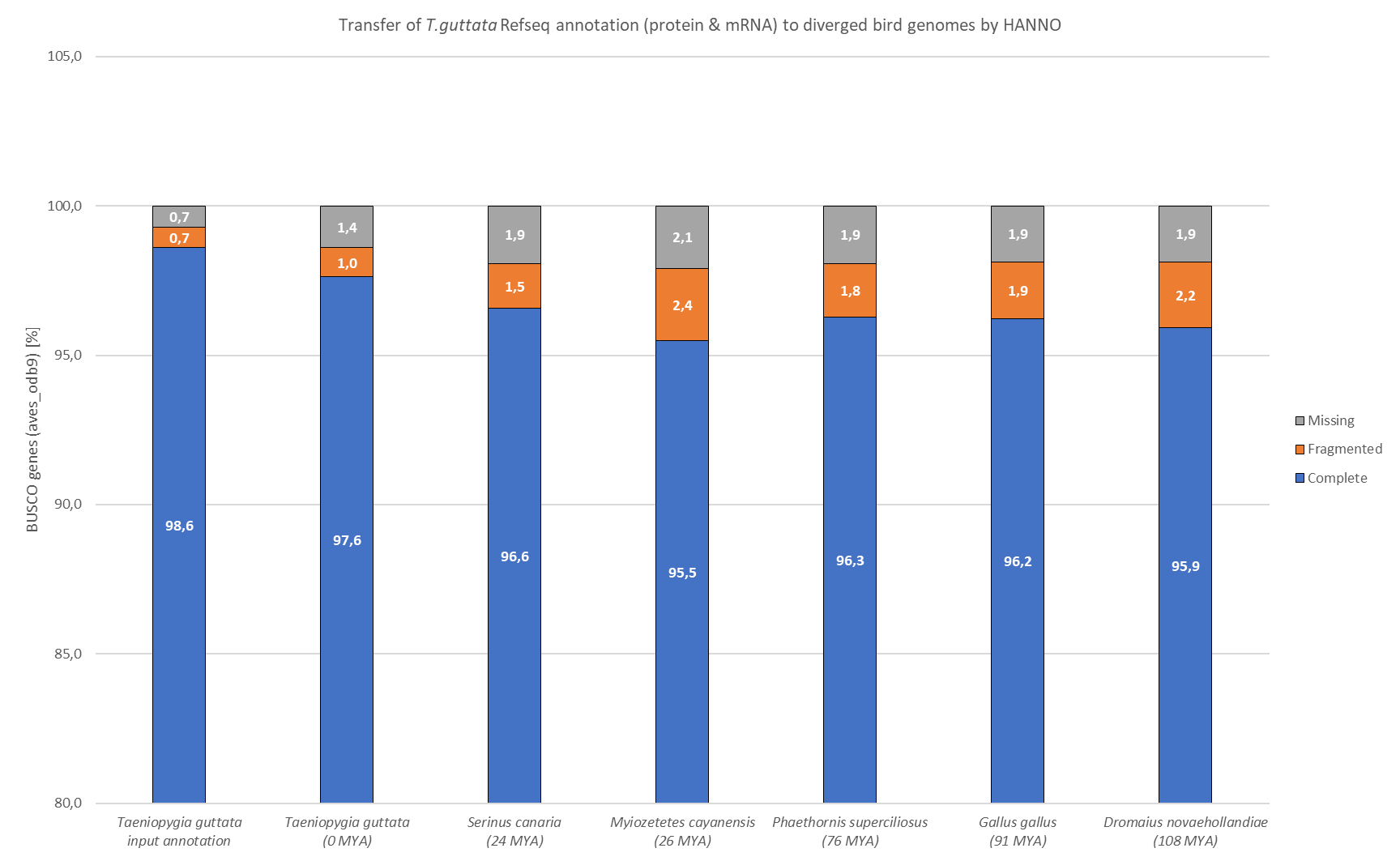
HANNO was benchmarked on 8 fish, 6 amphibian, 6 bird and 11 mammal genomes, by transferring Refseq annotations (from *P. flavescens, B. bufo, T. guttata* and *H. sapiens*) to closely related and diverged species’ genomes. Results in Figure 2 – 5 show the trends for annotated "Complete" and "Missing" BUSCOs for these genome annotations ordered by divergence time from reference species (from www.timetree.org). The number of complete BUSCOs slightly decreases with divergence time, vice versa the number of missing BUSCOs increases. Annotating species with divergence time of less than 100 MYA typically resulted in >95% Complete BUSCOs and <2% Missing BUSCOs for all tested clades.

****

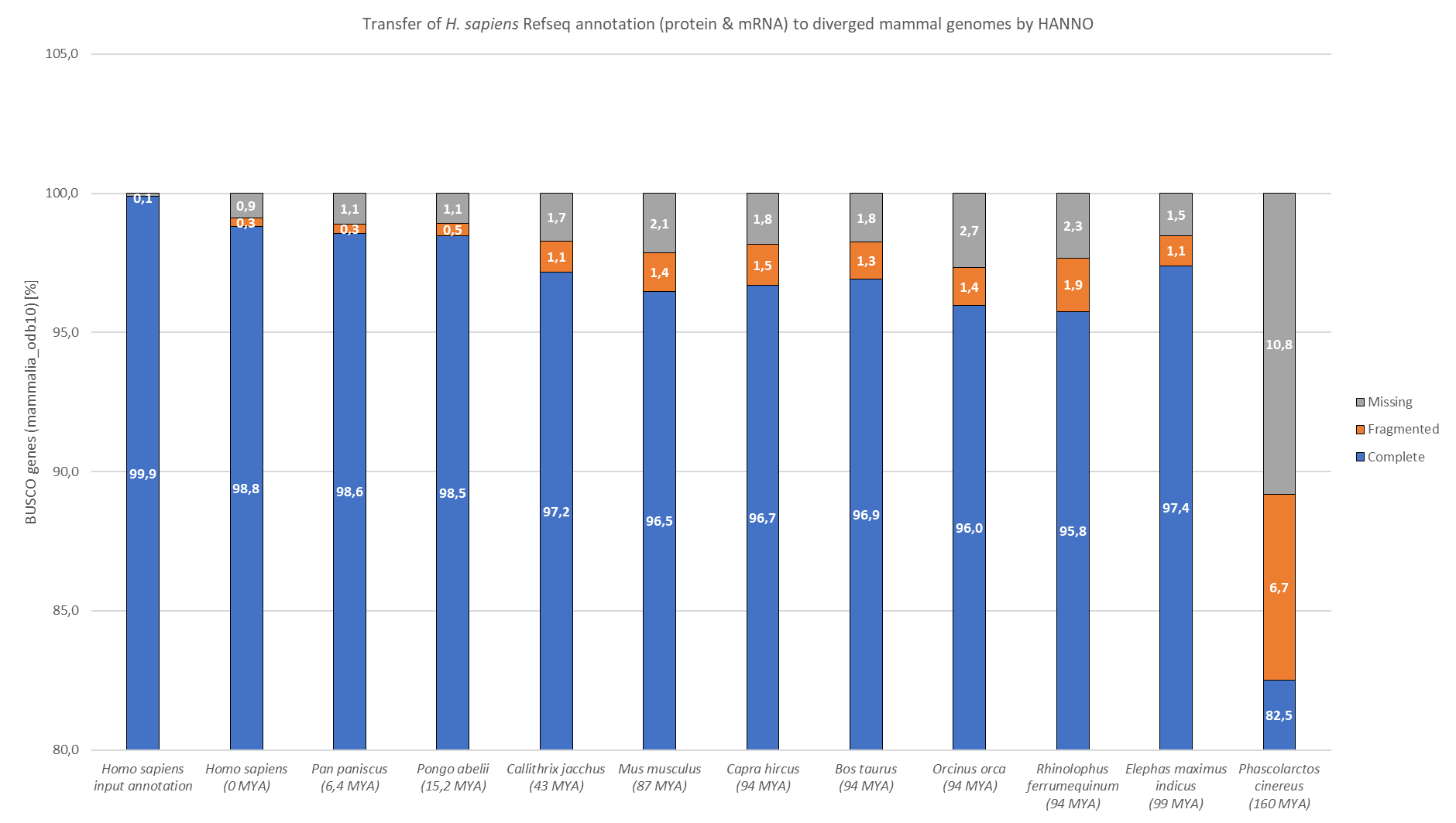
**Figure 2:** HANNO annotated complete, fragmented and missing BUSCO fractions [%] for diverged fish genomes. Note that the Y-axis is limited to the range 80 - 100% for better resolution.

****

**Figure 3:** HANNO annotated complete, fragmented and missing BUSCO fractions [%] for diverged amphibian genomes. Note that the Y-axis is limited to the range 80 - 100% for better resolution.

****

**Figure 4:** HANNO annotated complete, fragmented and missing BUSCO fractions [%] for diverged bird genomes. Note that the Y-axis is limited to the range 80 - 100% for better resolution.



**Figure 5:** HANNO annotated complete, fragmented and missing BUSCO fractions [%] for diverged mammal genomes. Note that the Y-axis is limited to the range 80 - 100% for better resolution.

*HANNO UTR completeness*

**DISCUSSION**

* Spaln
* Speed
* Improvement of Miniprot results by mRNA mappings
* Outlook on annotation in large projects.

**REFERENCES**