A benchmark study of ab initio gene prediction methods in diverse eukaryotic organisms

Nicolas Scalzitti
Laboratoire ICube
Anne Jeannin-Girardon
Laboratoire ICube https://orcid.org/0000-0003-4691-904X
Pierre Collet
Laboratoire ICube
Olivier Poch
Laboratoire ICube
Julie Dawn Thompson (thompson@unistra.fr)
Laboratoire ICube https://orcid.org/0000-0003-4893-3478

Research article

Keywords: genome annotation, gene prediction, protein prediction, benchmark study

Posted Date: December 20th, 2019

DOI: https://doi.org/10.21203/rs.2.19444/v1

Version of Record: A version of this preprint was published at BMC Genomics on April 9th, 2020. See the published version at https://doi.org/10.1186/s12864-020-6707-9.

A benchmark study of *ab initio* gene prediction methods in diverse eukaryotic organisms

Nicolas Scalzitti[1], Anne Jeannin-Girardon[1], Pierre Collet[1], Olivier Poch[1], Julie D. Thompson[1]*

[1] *Department of Computer Science, ICube, CNRS, University of Strasbourg, Strasbourg, France*

***Corresponding author:**

Email: thompson@unistra.fr

Abstract

Background: The draft genome assemblies produced by new sequencing technologies present important challenges for automatic gene prediction pipelines, leading to less accurate gene models. New benchmark methods are needed to evaluate the accuracy of gene prediction methods in the face of incomplete genome assemblies, low genome coverage and quality, complex gene structures, or a lack of suitable sequences for evidence-based annotations.

Results: We describe the construction of a new benchmark, called G3PO (benchmark for Gene and Protein Prediction PrOgrams), designed to represent many of the typical challenges faced by current genome annotation projects. The benchmark is based on a carefully validated and curated set of real eukaryotic genes from 147 phylogenetically disperse organisms, and a number of test sets are defined to evaluate the effects of different features, including genome sequence quality, gene structure complexity, protein length, etc. We used the benchmark to perform an independent comparative analysis of the most widely used *ab initio* gene prediction programs and identified the main strengths and weaknesses of the programs. More

importantly, we highlight a number of features that could be exploited in order to improve the accuracy of current prediction tools.

Conclusions: The experiments showed that *ab initio* gene structure prediction is a very challenging task, which should be further investigated. We believe that the baseline results associated with the complex gene test sets in G3PO provide useful guidelines for future studies.

Keywords: genome annotation, gene prediction, protein prediction, benchmark study.

Background

The plunging costs of DNA sequencing [1] have made *de novo* genome sequencing widely accessible for an increasingly broad range of study systems with important applications in agriculture, ecology, and biotechnologies amongst others [2]. The major bottleneck is now the high-throughput analysis and exploitation of the resulting sequence data [3]. The first essential step in the analysis process is to identify the functional elements, and in particular the protein-coding genes. However, identifying genes in a newly assembled genome is challenging, especially in eukaryotes where the aim is to establish accurate gene models with precise exon-intron structures of all genes [3-5].

Experimental data from high-throughput expression profiling experiments, such as RNA-seq or direct RNA sequencing technologies, have been applied to complement the genome sequencing and provide direct evidence of expressed genes [6,7]. In addition, information from closely related genomes can be exploited, in order to transfer known gene models to the target genome. Numerous automated gene prediction methods have been developed that incorporate similarity information, either from transcriptome data or known gene models, including GenomeScan [8], GeneWise [9], FGENESH [10], Augustus [11], Splign [12], CodingQuarry [13], and LoReAN [14].

The main limitation of similarity-based approaches is in cases where transcriptome sequences or closely related genomes are not available. Furthermore, such approaches encourage the propagation of erroneous annotations across genomes and cannot be used to discover novelty [5]. Therefore, similarity-based approaches are generally combined with *ab initio* methods that predict protein coding potential based on the target genome alone. *Ab initio* methods typically use statistical models, such as Support Vector Machines (SVMs) or hidden Markov models (HMMs), to combine two types of sensors: signal and content sensors. Signal sensors exploit specific sites and patterns such as splicing sites, promotor and terminator sequences, polyadenylation signals or branch points. Content sensors exploit the coding versus non-coding sequence features, such as exon or intron lengths or nucleotide composition [15]. *Ab initio* gene predictors, such as Genscan [16], GlimmerHMM [17], GeneID [18], FGENESH [10], Snap [19], Augustus [20], and GeneMark-ES [21], can thus be used to identify previously unknown genes or genes that have evolved beyond the limits of similarity-based approaches.

Unfortunately, automatic *ab initio* gene prediction algorithms often make substantial errors and can jeopardize subsequent analyses, including functional annotations, identification of genes involved in important biological process, evolutionary studies, etc. [22-25]. This is especially true in the case of large “draft” genomes, where the researcher is generally faced with an incomplete genome assembly, low coverage, low quality, and high complexity of the gene structures. Typical errors in the resulting gene models include missing exons, non-coding sequence retention in exons, fragmenting genes and merging neighboring genes. Furthermore, the annotation errors are often propagated between species and the more “draft” genomes we produce, the more errors we create and propagate [3-5]. Other important challenges that have attracted interest recently include the prediction of small proteins/peptides coded by short open reading frames (sORFs) [26,27] or the identification of

events such as stop codon recoding [28]. These atypical proteins are often overlooked by the standard gene prediction pipelines, and their annotation requires dedicated methods or manual curation.

The increased complexity of today's genome annotation process means that it is timely to perform an extensive benchmark study of the main computational methods employed, in order to obtain a more detailed knowledge of their advantages and disadvantages in different situations. Some previous studies have been performed to evaluate the performance of the most widely used *ab initio* gene predictors. One of the first studies [29] compared 9 programs on a set of 570 vertebrate sequences encoding a single functional protein, and concluded that most of the methods were overly dependent on the original set of sequences used to train the gene models. More recent studies have focused on gene prediction in specific genomes, usually from model or closely-related organisms, such as mammals [30], human [31,32] or eukaryotic pathogen genomes [33], since they have been widely studied and many gene structures are available that have been validated experimentally. To the best of our knowledge, no recent benchmark study has been performed on complex gene sequences from a wide range of organisms.

Here, we describe the construction of a new benchmark, called G3PO – benchmark for Gene and Protein Prediction PrOgrams, containing a large set of complex eukaryote genes from very diverse organisms (from human to protists). The benchmark consists of 1793 reference genes and their corresponding protein sequences from 147 species and covers a range of gene structures from single exon genes to genes with over 20 exons. A crucial factor in the design of any benchmark is the quality of the data included. Therefore, in order to ensure the quality of the benchmark proteins, we constructed high quality multiple sequence alignments (MSA) and identified the proteins with inconsistent sequence segments that might indicate potential sequence annotation errors. The benchmark thus contains both Confirmed

and Unconfirmed proteins and represents many of the typical prediction errors presented above. We believe the benchmark allows a more realistic evaluation of the currently available gene prediction tools on challenging data sets.

We used the G3PO benchmark to compare the accuracy and efficiency of five widely used *ab initio* gene prediction programs, namely Genscan, GlimmerHMM, GeneID, Snap and Augustus. Our initial comparison highlighted the difficult nature of the test cases in the G3PO benchmark, since 69% of the exons and 71% of the Confirmed protein sequences were badly predicted by all five gene prediction programs. Different benchmark tests were then designed in order to identify the main strengths and weaknesses of the different programs, but also to investigate the impact of the genomic environment, the complexity of the gene structure, or the nature of the final protein product on the prediction accuracy.

Results

Benchmark data sets

The G3PO benchmark contains 1793 proteins from a diverse set of organisms (Additional file 1: Table S1), which can be used for the evaluation of gene prediction programs. The proteins were extracted from the Uniprot [34] database, and are divided into 20 orthologous families (called BBS1-21, excluding BBS14) that are representative of complex proteins, with multiple functional domains, repeats and low complexity regions (Additional file 1: Table S2). The benchmark test sets cover many typical gene prediction tasks, with different gene lengths, protein lengths and levels of complexity in terms of number of exons (Additional file 1: Fig. S1). For each of the 1793 proteins, we identified the corresponding genomic sequence and the exon map in the Ensembl [35] database. We also extracted the same genomic sequences with additional DNA regions ranging from 2,000 to 10,000 nucleotides upstream and downstream of the gene, in order to represent more realistic genome annotation tasks.

Additional file 1: Fig. S2 shows the distribution of various features of the 1793 benchmark test cases, at the genome level (gene length, GC content), gene structure level (number and length of exons, intron length), and protein level (length of main protein product).

Phylogenetic distribution of benchmark sequences

The protein sequences used in the construction of the G3PO benchmark were identified in 147 phylogenetically diverse eukaryotic organisms, ranging from human to protists (Fig. 1A and Additional file 1: Table S3). The majority (72%) of the proteins are from the Opisthokonta clade, which includes 1236 (96.4%) Metazoa, 25 (1.9%) Fungi and 22 (1.7%) Choanoflagellida sequences (Fig. 1B). The next largest groups represented in the database are the Stramenopila (172), Euglenozoa (149) and Alveolata (99) sequences. More divergent species are included in the 'Others' group, containing 57 sequences from 6 different clades, namely Apusozoa, Cryptophyta, Diplomonadida, Haptophyceae, Heterolobosea and Parabasalia.

Exon map complexity

The benchmark was designed to cover a wide range of test cases with different exon map complexities, as encountered in a realistic complete genome annotation project. The test cases in the benchmark range from single exon genes to genes with 40 exons (Additional file 1: Fig. S2). In particular, the different species included in the benchmark present different challenges for gene prediction programs. To illustrate this point, we compared the number of exons in the human genes to the number of exons in the orthologous genes from each species (Fig. 2). Three main groups can be distinguished: i) Chordata, ii) other Opisthokonta (Mollusca, Platyhelminthes, Panarthropoda, Nematoda, Cnidaria, Fungi and Choanoflagellida) and iii) other Eukaryota (Amoebozoa, Euglenozoa, Heterolobosza, Parabasalia, Rhodophyta,

Viridiplantae, Stramenopila, Alveolata, Rhizaria, Cryptophyta, Haptophyceae). As might be expected, the sequences in the Chordata group generally have a similar number of exons compared to the Human sequences. The sequences in the 'other Opisthokonta' group have greater heterogeneity, as expected due to their phylogenetic divergence, although some classes, such as the insects are more homogeneous. The genes in this group have three times less exons on average, compared to the Chordata group. The 'other Eukaryota' group includes diverse clades ranging from Viridiplantae and Protists, although the exon map complexity is relatively homogeneous within each clade. For example, in the Euglenozoa clades, all sequences have less than 20% of the number of exons compared to human.

Quality of protein sequences

The protein sequences included in the benchmark were extracted from the public databases, and it has been shown previously that these resources contain many sequence errors [22-25]. Therefore, we evaluated the quality of the protein sequences in G3PO using a homology-based approach (see Methods), similar to that used in the GeneValidator program [23]. We thus identified protein sequences containing potential errors, such as inconsistent insertions/deletions or mismatched sequence segments (Additional file 1: Fig. S3 and Methods). Of the 1793 proteins, 889 (49.58%) protein sequences had no identified errors and were classified as 'Confirmed', while 904 (50.42%) protein sequences had from 1 to 8 potential errors (Fig. 3A) and were classified as 'Unconfirmed'. The 904 Unconfirmed sequences contain a total of 1641 errors, *i.e.* each sequence has an average of 1.8 errors. We further characterized the Unconfirmed sequences by the categories of error they contain (Fig. 3B) and by orthologous protein family (Additional file 1: Fig. S4A and B). All the protein families contain Unconfirmed sequences, regardless of the number or length of the sequences, although the ratio of Confirmed to Unconfirmed sequences is not the same in all families. For

example, the BBS6, 11, 12, 18 families, that are present mainly in vertebrate species, have more Confirmed sequences (68.5%, 80.0%, 52.3%, 61.1% respectively). Inversely, the majority of sequences in the BBS8 and 9 families, that contain many phylogenetically disperse organisms, are Unconfirmed (68.8%, 73.3% respectively). The majority of the 1641 errors (58.4%) are internal (*i.e*. do not affect the N- or C-termini) and 31% are internal mismatched segments, while N-terminal errors (378=23.0%) are more frequent than C-terminal errors (302=18.4%). At the N- and C-termini, deletions are more frequent than insertions (280 and 145, respectively), in contrast to the internal errors, where insertions are more frequent (304 compared to 143).

The distributions of various features are compared for the sets of 889 Confirmed and 904 Unconfirmed sequences in Additional file 1: Fig. S2. There are no significant differences in gene length (p-value=0.735), GC content (p-value=0.790), number of exons (p-value=0.073), and exon/intron lengths (p-value=0.690 / p-value=0.949) between the Confirmed and Unconfirmed sequences. The biggest difference is observed at the protein level, where the Confirmed protein sequences are 13% shorter than the Unconfirmed proteins (p-value=8.75×10^{-9}). We also compared the phylogenetic distributions observed in the Confirmed and Unconfirmed sequence sets (Fig. 1C and D). Two clades had a higher proportion of Confirmed sequences, namely Opisthokonta (691/1283=54%) and Stramenopila (88/172=51%). In contrast, Alveolata (24/99=24%), Rhizaria (5/21=24%) and Choanoflagellida (5/22=22%) had fewer Confirmed than Unconfirmed sequences.

Quality of genome sequences

The genomic sequences corresponding to the reference proteins in G3PO were extracted from the Ensembl database. In all cases, the soft mask option was used (see Methods) to localize repeated or low complexity regions. However, some sequences still contained

undetermined nucleotides, represented by 'n' characters, probably due to genome sequencing or assembly errors. Undetermined (UDT) regions were found in 283 (15.8%) genomic sequences from 58 (39.5%) organisms, of which 281 sequences (56 organisms) were from the metazoan clade (Additional file 1: Fig. S5). Of these 283 sequences, 133 were classified as Confirmed and 150 were classified as Unconfirmed.

We observed important differences between the characteristics of the sequences with UDT regions and the other G3PO sequences, for both Confirmed and Unconfirmed proteins (Additional file 1: Table S4). The average length of the 283 gene sequences with UDT regions (95584 nucleotides) is 6 times longer than the average length of the 1510 genes without UDT (15934 nucleotides), although the protein sequences have similar average lengths (551 amino acids for UDT sequences compared to 514 amino acids for non UDT sequences). Sequences with UDT regions have twice as many exons, three times shorter exons and five times longer introns than sequences without UDT.

Evaluation metrics

The benchmark includes a number of different performance metrics that are designed to measure the quality of the gene prediction programs at different levels. At the nucleotide level, we study the ability of the programs to correctly classify individual nucleotides found within exons or introns. At the exon level, we applied a strict definition of correctly predicted exons: the boundaries of the predicted exons should exactly match the boundaries of the benchmark exons. At the protein level, we compare the predicted protein to the benchmark sequence and calculate the percent sequence identity. It should be noted that, due to their strict definition, scores at the exon level are generally lower. For example, in some cases, the predicted exon boundary may be shifted by a few nucleotides, resulting in a low exon score but high nucleotide and protein level scores.

Evaluation of gene prediction programs

We selected five widely used gene prediction programs: Augustus, Genscan, GeneID, GlimmerHMM and Snap. These programs all use Hidden Markov Models (HMMs) trained on different sets of known protein sequences and take into account different context sensors, as summarized in Table 1. The genomic sequences for the 1793 test cases in the G3PO benchmark were used as input to the selected gene prediction programs and a series of tests were performed (outlined in Fig. 4), in order to identify the strong and weak points of the different algorithms, as well as to highlight specific factors affecting prediction accuracy.

Computational runtime

First, we compared the CPU time required for each program to process the benchmark sequences (Fig. 5A and Additional file 1: Table S5). Using the gene sequences only with 0Kb flanking regions (representing a total length of 51,110,612 nucleotides), Augustus required the largest CPU time (1826 seconds), taking >3.5 times as long as the second slowest program, namely Genscan (484 seconds). GeneID was the fastest program and completed the gene prediction for the 1793 genomic regions, including 10Kb upstream/downstream flanking nucleotides (total length of 86,970,612 nucleotides), in 260 seconds.

Gene prediction accuracy

In order to estimate the overall accuracy of the five gene prediction programs, the genes predicted by the programs were compared to the benchmark sequences in G3PO. At this stage, we included only the 889 Confirmed proteins, and used the genomic sequences corresponding to the gene region only (0Kb flanking regions) (Fig. 4 – Initial tests) as input.

Fig. 5(B-D) and Additional file 1: Table S6 show the mean quality scores at different levels: nucleotide, exon structure and final protein sequence (defined in Methods).

At the nucleotide level (Fig. 5B), most of the programs have higher specificities than sensitivities (with the exception of GlimmerHMM), meaning that they tend to underpredict. F1 scores range from 0.377 for GeneID to 0.528 for Augustus, meaning that it has the best accuracy.

At the exon level (Fig. 5C), Augustus and Genscan achieve higher sensitivities (0.27, 0.26 respectively) and specificities (0.30, 0.31 respectively) than the other programs. Nevertheless, the number of mis-predicted exons remains high with 73% and 74% Missing Exons and 69% and 68% Wrong Exons respectively for Augustus and Genscan. At this level, GlimmerHMM has the lowest sensitivity and specificity, indicating that the predicted splice boundaries are not accurate. To further investigate the complementarity of the different programs, we plotted the number of Correct Exons (*i.e*. both 5' and 3' exon boundaries correctly predicted) identified by at least one of the programs (Fig. 6A). A total of 130 exons were found by all five programs, suggesting that they are relatively simple to identify. More importantly, 705 exons were correctly predicted by only one program, while 5507 (69.3%) exons were not predicted correctly by any of the programs.

As might be expected, the nucleotide and exon scores are reflected at the protein level (Fig. 5D), with Augustus again achieving the best score, obtaining 78% sequence identity overall and predicting 188 of the 889 (21.2%) Confirmed proteins with 100% accuracy. GlimmerHMM and GeneID have the lowest scores in terms of perfect protein predictions (28, 32 respectively). Again, we investigated the complementarity of the programs, by plotting the number of proteins that were perfectly predicted (100% identity) by at least one of the programs (Fig. 6B). Only 7 proteins are perfectly predicted by all five programs, while 117 proteins were predicted with 100% accuracy by a single program. These were mostly

predicted by Augustus (66), followed by Genscan (34). 634 (71%) of the 889 benchmark proteins were not predicted perfectly by any of the programs included in this study.

Analysis of factors affecting gene prediction quality

Based on the results of our initial comparison of gene prediction accuracy, and particularly the complementarity of the programs highlighted in Fig. 6, we decided to investigate further the different factors that may influence the performance of the prediction programs. Fig. 4 provides an overview of the different tests performed, including: i) factors associated with the input genomic sequence, ii) factors associated with the gene structure, and iii) factors associated with the protein product.

Factors associated with the input genomic sequence

We first evaluated the genome context and the effect of adding flanking sequences upstream and downstream of the benchmark gene sequence used as input to the prediction programs, using the 889 Confirmed benchmark tests. We added different flanking sequence lengths ranging from 2Kb to 10Kb, and calculated the same quality scores as above, at the nucleotide, exon and protein levels (Fig. 7 and Additional file 1: Table S7).

At the nucleotide level, the sensitivity and specificity of Augustus, Genscan, GeneID and Snap is not significantly affected by the addition of the flanking sequences. For GlimmerHMM (p-value=1.60×10^{-30}), a significant increase in sensitivity is observed when 2Kb flanking sequences are added, compared to the gene sequences only (0Kb). This is probably due to the addition of specific signals in the genomic environment of the gene, such as the promoter, enhancers/silencers, etc. that are taken into account in the program prediction models. In terms of specificity, the addition of 2Kb flanking sequences generally increases the quality of the programs, except for Snap. At the exon level, the effect of the flanking

298 sequences is not the same for the different programs. For example, the sensitivity of Augustus
299 and Genscan is highest when the input sequence has no flanking regions, while for GeneID
300 (p-value=4.115×10^{-5}), GlimmerHMM (p-value=4.408×10^{-17}) and Snap (p-value=0.023),
301 sensitivity is significantly improved by adding at least 2Kb flanking nucleotides. Similar
302 results are observed in terms of specificity. At the protein level, for Augustus (p-
303 value=0.0058), Genscan (p-value=8.9×10^{-8}) and Snap (p-value=0.037), the sequence identity
304 compared to the benchmark protein sequence decreases as the length of the flanking
305 sequences increases. For GlimmerHMM and GeneID, the maximum sequence identity is
306 achieved for 2Kb flanking nucleotides. For Augustus and Genscan, the addition of the
307 flanking sequences also reduces the number of proteins perfectly predicted (100% identity).
308 This is especially true for Genscan, where we observe a loss of more than 30% of perfectly
309 predicted proteins between 0Kb and 2Kb. On the other hand, for GeneID, GlimmerHMM and
310 Snap, the number of perfectly predicted proteins increases, especially when 2Kb flanking
311 DNA is provided.

312 Since the greatest effect of adding upstream/downstream flanking sequences was generally
313 observed for a length of 2Kb, the remaining analyses described in this work are all based on
314 the gene sequences with 2Kb upstream/downstream flanking regions.

315 Next, we studied the relative robustness of the programs to the presence of UDT regions in
316 the genomic sequences, generally due to genome sequencing or assembly errors. This test was
317 limited to the Confirmed sequences from the metazoan clade, since the sequences with UDT
318 regions were almost exclusively found in this clade. Of the 675 metazoan sequences, 133
319 were found to have UDT regions. We therefore compared the 542 Confirmed sequences
320 without UDT (-UDT) regions with the 133 Confirmed sequences with UDT regions (+UDT).
321 Fig. 8 and Additional file 1: Table S8 show the average scores obtained for these two
322 sequence sets, at the nucleotide, exon and protein levels. As might be expected, a reduction in

sensitivity and specificity was observed at the nucleotide and exon levels for almost all programs (except specificity of Augustus at the exon level) for the +UDT sequences, and at the protein level, very few +UDT proteins are predicted with 100% accuracy. Overall, Augustus and Genscan perform better, although GlimmerHMM predicts the highest number of proteins with 100% accuracy for the +UDT sequences.

Since the UDT regions affected the programs to different extents, the analyses described in the following sections are all based on the set of 756 Confirmed sequences that have no UDT regions.

Factors associated with the gene structure

We first evaluated the effect of the Exon Map Complexity (EMC), represented by the number of exons in the Confirmed benchmark tests (Additional file 1: Fig. S6). Fig. 9 shows the quality scores at the exon and protein levels, for sequences with the number of exons ranging from 1 to 20. Overall, we observed a tendency for the five programs to achieve better sensitivity and specificity for the genes with more exons. This may be because most of these more complex sequences are from well-studied vertebrate genomes. For very complex exon maps (≥20 exons), all the programs seem to perform less well, although this may be an artifact due to the small number of these sequences in the benchmark (Additional file 1: Fig. S6A). For single exon genes, all the programs tend to perform worse, although the 3' boundary of the cDNA is predicted better than the 5' boundary. Similarly, the 3' exon boundaries are generally predicted better than the 5' boundaries by all the programs, for genes with a small number of exons. At the protein level, Augustus and GlimmerHMM achieve higher sequence identity for genes with ≤8 exons, while Augustus and Genscan are more accurate for genes with more exons. Most of the perfectly predicted proteins (with 100% sequence identity) have less than 3 exons.

We then assessed the effect of exon lengths on the prediction quality of the five programs, using the 756 Confirmed sequences without UDT regions. Fig. 10A and Additional file 1: Table S9A show the proportion of Correct exons (both 5' and 3' exon boundaries correctly predicted) depending on the exon length. The short exons (<50 nucleotides) are generally the least accurate, with the best program, Augustus, achieving only 17% Correct short exons. Medium length exons (50-200 nucleotides) are predicted better than longer exons (>200 nucleotides) for Augustus and Genscan.

To further investigate the exon prediction, each exon predicted by a gene prediction program was classified as 'Correct' if both exon boundaries were correctly predicted, 'Wrong (5')' or 'Wrong (3')' if the 5' or 3' exon boundary was badly predicted respectively, and 'Wrong' if both boundaries were badly predicted. In some cases, the predicted exon has good 5' and 3' exon boundaries, however they correspond to 2 different benchmark exons, so these exons are classed as 'Wrong (Fusion)'. Fig. 10B and Additional file 1: Table S9B show the number of Correct, Wrong, Wrong (5'), Wrong (3') and Wrong (Fusion) exons, according to the exon lengths. Overall, there are more 'Wrong' exons than 'Correct' exons for all exon lengths and for all the programs. Interestingly, the number of predicted exons with only one boundary correctly predicted, *i.e*. Wrong (5') or Wrong (3'), is small for all exon lengths, except for exons with >200 nucleotides.

Factors associated with the protein product

In this section, prediction accuracy is measured at the protein level and is estimated by the percent sequence identity of the predicted protein compared to the benchmark protein.

First, we investigated the effect of protein length on protein prediction quality. We divided the 756 Confirmed sequences without UDT regions into five groups, with different protein lengths ranging from 50 to 1000 amino acids (Additional file 1: Fig. S7). Note that the very

large proteins (>1000 amino acids) in the benchmark are all classified as Unconfirmed and are therefore not included in this study. Fig. 11 and Additional file 1: Table S10 show the mean accuracies obtained by the five programs for the different length proteins. The prediction accuracy generally decreases for shorter proteins and for protein lengths >650 amino acids. For proteins with <100 amino acids, GlimmerHMM achieves the best results with 68% sequence identity and five (26%) perfectly predicted proteins (100% identity), while Augustus obtains only 43% sequence identity and no perfectly predicted proteins.

We then studied the phylogenetic origin of the proteins and the availability of suitable species models in the different programs. Fig. 12 and Additional file 1: Table S11 show the performance of the five gene prediction programs for the sequences in the different clades in G3PO. The accuracy of each program is highly variable between the different clades, probably due to the availability of suitable prediction models for some species. For the sequences in the Craniata clade, Augustus achieves the highest accuracy (76%), while Snap has the lowest accuracy (39%). In contrast, Augustus obtains lower accuracy (38%) for Fungi proteins, compared to the highest accuracy obtained by GlimmerHMM (58%). The proteins in the Euglenoezoa clade are predicted with the highest accuracy by all the programs, although this might be explained by their low EMC. Choanoflagellida and Cnidaria proteins are the least well predicted, but these clades contain only a few sequences (5 and 6 sequences respectively) and this result remains to be confirmed.

Effect of protein sequence errors

Finally, we investigated the performance of the prediction programs for the 904 Unconfirmed sequences, where potential sequence errors were observed in the benchmark sequences. As mentioned above, the G3PO benchmark sequences were extracted from the Uniprot database, which means that many of the proteins are not supported by experimental

evidence. In this test, we wanted to estimate the prediction accuracy of the five gene prediction programs for the Unconfirmed benchmark sequences. Since the Unconfirmed sequences could not be used as a ground truth, here we measured prediction accuracy based on a closely related Confirmed sequence (see Methods). Table 2 shows the prediction accuracies achieved by each program for the sets of Confirmed and Unconfirmed sequences. As might be expected, the Unconfirmed sequences are predicted with lower accuracy than the Confirmed sequences by all five programs. Augustus and Genscan achieved the highest accuracy (63%, 61% respectively) for the Unconfirmed sequences. For comparison purposes, we also calculated the accuracy scores for the Unconfirmed benchmark proteins. The benchmark proteins had higher accuracy (76%) than any of the methods tested here, implying that the more complex pipelines used to curate proteins in Uniprot can effectively improve the results of *ab initio* methods.

Discussion

Thanks to cheap genome sequencing, consortia such as the Genome 10K [36], Bird 10K [37], the Cephseq consortium for cephalopods [38], or the Earth Biogenome Project [39], can now produce eukaryotic genome sequences on a very large scale. Recently, the new sequencing technologies have also been used to improve genome annotation by providing an overview of the genome regions that are actively transcribed. Nevertheless, when transcriptome data is not available or coverage of the transcriptome is shallow, computational annotation strategies play an important role in genome annotation.

Several recent reviews [3,22-23] have highlighted the fact that automated annotation strategies still have difficulty correctly identifying protein-coding genes. This failure might be explained by the quality of the draft genome assemblies, the complexity of eukaryotic exon maps, high levels of genetic sequence divergence or deviations from canonical genetic

423 characteristics [40]. As a result, although we have access to a large amount of genome data,
424 this comes at the expense of annotation quality.

425 In order to improve gene prediction quality, it is essential to benchmark the existing
426 different gene prediction strategies to assess their reliability, to identify the most promising
427 approaches, but also to limit the spread of errors in protein databases [41]. An ideal
428 benchmark for gene prediction programs should include proteins encoded by real genomic
429 sequences. Unfortunately, most of the protein sequences in the public databases have not been
430 verified by experimental means, with the exception of the manually annotated Swiss-Prot
431 sequences (representing only 0.3% of UniProt), and contain many sequence annotation errors.
432 It is therefore dangerous to use them to estimate the accuracy of the prediction programs.

433 We therefore constructed a new high-quality benchmark, called G3PO, containing 1793
434 orthologous sequences from 20 different protein families. The benchmark is designed to be as
435 representative as possible of the living world. To achieve this, we included sequences from
436 phylogenetically diverse organisms in the main clades, *e.g*. Vertebrates, Fungi, Plants and
437 Protists. We also included sequences with a wide range of different genomic and protein
438 characteristics, from simple single exon genes to very long and complex genes with over 20
439 exons.

440 The protein sequences in the benchmark were extracted from the Uniprot database, where
441 a 'canonical' protein isoform is defined based on cross-species conservation and the
442 conservation of protein structure and function. Consequently, programs that predicted more
443 minor isoforms were penalized in our evaluations. Unfortunately, there is currently no ideal
444 solution to this. In the future, gene prediction programs will need to evolve to predict all
445 isoforms for a gene.

446 The *ab initio* gene prediction programs included in the benchmark study are based on
447 statistical models that are trained using known proteins and typically perform well at

448 predicting conserved or well-studied genes [33,42]. However, *ab initio* prediction accuracy
449 has been previously shown to decrease in some special cases, such as small proteins [43],
450 organism-specific genes or other unusual genes [44-46]. Our goal was therefore to identify the
451 strengths and weaknesses of the programs, but also to highlight genomic and protein
452 characteristics that could be incorporated to improve the prediction models.

453 First, we ensured the quality of the 1793 protein sequences in the benchmark using a
454 homology-based method, and 1641 errors were identified in 904 of the sequences. In other
455 words, more than 50% of the reference proteins from Uniprot had sequence errors, including
456 deletions, insertions and mismatched segments. Sequence errors were found in all protein
457 families and in all main clades. Since accurate and reliable data sets are crucial for
458 benchmarking studies, the main results of this work were established using the set of
459 Confirmed protein sequences only. Second, we characterized the test sets in the benchmark
460 using different features at the genome, gene structure and protein levels. This in-depth
461 characterization allowed us to investigate the impact of these features on gene prediction
462 accuracy.

463 Our initial experiment to measure the overall quality of the gene prediction programs
464 generally confirmed previous findings in terms of program ranking, with Augustus and
465 Genscan achieving the best overall accuracy scores. However, it should be noted that
466 Augustus is also the most computationally expensive method, taking over 1 hour to process
467 the 87Mb corresponding to the 1793 benchmark sequences, compared to the fastest program,
468 GeneID, which required only 4 minutes.

469 We then performed a more in-depth study of the different factors affecting prediction
470 accuracy. At the genome level, we first analyzed the impact of the genomic context, by
471 extending the benchmark gene sequences, *i.e.* the cDNA, by different lengths (2Kb-10Kb) at
472 the 5' and 3' ends. An increase in accuracy was generally observed when at least 2Kb flanking

regions were added, reflecting the fact that all the programs try to model *in vivo* gene translation systems to some extent by taking into account the different regulatory signals found within and outside the gene [47]. We also investigated the effect of undetermined regions in the gene sequences, due to genome sequencing or assembly errors. We observed a negative effect of these regions on the accuracy of all the prediction programs, even when they occur outside the coding exons of the genes.

At the gene structure level, we studied the effects of the number and length of exons on prediction accuracy. We found that the number of exons affects the accuracy of all the programs and that gene prediction is generally more difficult for complex exon maps, as might be expected. We also observed that the precise exon boundaries are difficult to predict for all the programs, although they tend to predict the 3' boundaries better than the 5' boundaries for genes with a small number of exons. Concerning the effect of exon length, the programs appear to be optimized for intermediate length exons (50-200 nucleotides), since none of the programs was able to reliably predict exons that were shorter (<50 nucleotides) or longer (>200 nucleotides).

At the protein level, protein length had a similar effect to that observed for exon length, since the programs seem to be optimized for intermediate length proteins (300-650 amino acids). This result confirms previous findings that smaller proteins (less than 100 amino acids) are often missed in genome annotations [43], although we also demonstrated that long proteins are also more likely to be badly predicted. The phylogenetic origin of the benchmark sequences had a large effect on prediction accuracy, with different programs producing the best results depending on the specific species. The two best scoring programs, Augustus and Genscan use different strategies, since Augustus includes >100 different species models (for Animals, Alveolata, Plants and Algae, Fungi, Bacteria, Archaea), while Genscan has only

three models (human/Vertebrates, *Arabidopsis*, Maize). The species models are trained on species-specific training sets, and probably contribute to the success of the Augustus program.

Each of the analyses performed here highlights different strengths or weaknesses of the prediction programs, as summarized in the heat map shown in Fig. 13. The in-depth characterization of the benchmark sequences and the detailed information extracted from the analyses provide essential elements that could be used to improve model training and therefore gene prediction. It may be interesting to further analyze the weaknesses identified, including small proteins, very long proteins, proteins coded by a large number of exons, proteins from non-model organisms, etc.

Finally, the Unconfirmed sequences identified in this study represent a goldmine for the identification of atypical gene features, for example atypical regulatory signals or splice sites, that are not fully taken into account in the current prediction models. More than 50% of the original reference protein sequences extracted from public databases were found to contain at least one error. They therefore represent very challenging test cases that were not resolved by the combined *ab initio* and similarity-based curation processes used to annotate these proteins. We accurately located the errors within these badly predicted sequences and classified them into 9 groups. Here, we performed a preliminary analysis using the erroneous sequences that confirmed our idea that all the prediction programs are less accurate for these proteins. A more comprehensive analysis of these proteins will be published elsewhere.

Conclusions

The complexity of the genome annotation process and the recent activity in the field mean that it is timely to perform an extensive benchmark study of the main computational methods employed, in order to obtain a more detailed knowledge of their advantages and disadvantages in different situations. Currently, most of the programs used for gene prediction

are based on statistical approaches and perform relatively well in intermediate cases. However, they have difficulty identifying more extreme cases, such as very short or very long proteins, complex exon maps, or genes from less well studied species. Recently, artificial intelligence approaches have been applied to some specific tasks, for example DeepSplice [48] or SpliceAI [49] for the prediction of splice sites. The further development of these approaches should contribute to production of high quality gene predictions that can be leveraged downstream to improve functional annotations, evolutionary studies, prediction of disease genes, etc.

Methods

Benchmark test sets

To construct a benchmark set of eukaryotic genes, we selected the 20 human Bardet-Biedl Syndrome (BBS) proteins (Additional file 1: Table S2). Based on this initial gene set, we extended the test sets using the pipeline shown in Fig. 14 and described in detail below.

(i) For each of the 20 human proteins, orthologous proteins were identified in 147 eukaryotic organisms (Additional file 1: Table S1) using OrthoInspector version 3.0 [50], which was built using proteins from the Uniprot Reference Proteomes database [34] (Release 2016_11). This resulted in a total of 1793 protein sequences, of which 65 (3.6%) were found in the curated Swissprot database. The number of proteins in each BBS family is provided in Additional file 1: Table S2. BBS 6,10,11,12,15, 16 and 18 are specific to Metazoa (with some exceptions), and therefore contain fewer sequences than the other families.

(ii) Since the reference protein sequences extracted from the Uniprot database may contain errors, we identified potentially unreliable sequences based on multiple sequence alignments (MSA). MSAs were constructed for each protein family using

the Pipealign2 tool (http://www.lbgi.fr/pipealign) and manually refined to identify and correct misaligned regions. The SIBIS (version 1.0) program [51] using a Bayesian framework combined with Dirichlet mixture models and visual inspection, was used to identify inconsistent sequence segments. These segments might indicate that different isoforms are defined as the canonical sequence for different organisms, or they might indicate a badly predicted protein (Additional file 1: Fig. S3). SIBIS classifies the potential sequence errors into 9 categories: N-terminal deletion, N-terminal extension, N-terminal mismatched segment, C-terminal deletion, C-terminal extension, C-terminal mismatched segment, internal deletion, internal insertion and internal mismatched segment. Of the 1793 protein sequences identified in step (i), 889 proteins had no errors (called "Confirmed") and 904 proteins had at least one potential error (called "Unconfirmed"). At this stage, the BBS14 protein was excluded from the benchmark because the MSA contained too many misalignments.

(iii) For each orthologous protein, the genomic sequence was extracted from the Ensembl database [35]. Genomic sequences were extracted with the 'soft mask' option, *i.e.* repeated or low complexity regions are replaced by lower case nucleotides. These are generally ignored by gene prediction programs. We also found regions with 'n' characters, which are used to indicate undetermined nucleotides probably caused by genome misassembly or sequencing errors. Additional file 1: Table S4 summarizes the general statistics of these 283 sequences with undetermined (UDT) regions. Finally, we identified the Ensembl transcript corresponding to the Uniprot protein sequence, (generally the 'canonical transcript' from APPRIS [52]) in order to construct the exon map by extracting the positions of all exons/introns.

(iv) To make the benchmark set more challenging, we also extracted genomic sequences corresponding to 2Kb, 4Kb, 6Kb, 8Kb, 10Kb upstream and downstream of the gene sequence.

Gene prediction methods

The programs tested are listed in Table 1 with the main features, including the HMM model used to differentiate intron/exon regions, and the specific signal sensors used to detect the presence of functional sites. Transcriptional signal sensors include the initiator or cap signal located at the transcriptional start site and the upstream TATA box promoter signal, as well as the polyadenylation signal (a consensus AATAAA hexamer) located downstream of the coding region and the 3' UTR. Translational signals include the "Kozak sequence" located immediately upstream of the start codon [53]. For higher eukaryotes, splice site signals are also incorporated, including donor and acceptor sites (GT-AG on the intron sequence) and the branch point [yUnAy] [54] (underlined A is the branch point at position zero and y represents pyrimidines, n represents any nucleotide) located 20–50 bp upstream of the AG acceptor.

All programs were run on an Intel(R) Xeon(R) CPU E5-2695 v2 @ 2.40Ghz, 12 cores, with 256 Go RAM. Each prediction program was run with the default settings, except for the species model to be used. As the benchmark contains sequences from a wide range of species, we selected the most pertinent training model for each target species, based on the taxonomic proximity between the target and model species.

Evaluation metrics

The performance of the gene prediction programs is based on the measures used in [29], calculated at three different levels: nucleotides, exons and complete proteins. The significance of pairwise comparisons of the evaluation metrics was evaluated using the standard t-test.

At the nucleotide level, we measure the accuracy of a gene prediction on a benchmark sequence by comparing the predicted state (exon or intron) with the true state for each nucleotide along the benchmark sequence. Nucleotides correctly predicted to be in either an exon or an intron are considered to be True Positives (TP) or True Negatives (TN) respectively. Conversely, nucleotides incorrectly predicted to be in exons or introns are considered to be False Positives (FP) or False Negatives (FN) respectively. We then calculated different performance statistics, defined below.

Sensitivity measures the proportion of benchmark nucleotides that are correctly predicted:

$$Sn = \frac{TP}{TP + FN}$$

The specificity measure that is most widely used in the context of gene prediction is the proportion of nucleotides predicted in exons that are actually in exons:

$$Sp = \frac{TP}{TP + FP}$$

The F1 score represents the harmonic mean of the sensitivity and specificity values:

$$F1 = 2 * \frac{Sp * Sn}{Sp + Sn}$$

At the exon structure level, we measure the accuracy of the predictions by comparing predicted and true exons along the benchmark gene sequence. An exon is considered correctly predicted (TP), when it is an exact match to the benchmark exon, *i.e.* when the 5' and 3' exon boundaries are identical. All other predicted exons are then considered FP. Sensitivity and specificity are then defined as before.

Since the definition of TP and TN exons above is strict, we also calculated two additional measures similar to those defined in [29] (Additional file 1: Fig. S8). First, true exons with or without overlap to predicted exons are considered to be Missing Exons (ME) and the MEScore is defined as:

$$MEScore = \frac{ME}{Total\ number\ of\ true\ exons}$$

Second, predicted exons with or without overlap to true exons are considered Wrong Exons (WE). The WEScore is defined as:

$$WEScore = \frac{WE}{Total\ number\ of\ predicted\ exons}$$

We also determined the proportion of correctly predicted 5' and 3' exon boundaries, as follows:

$$5' = \frac{\text{number of true } 5' \text{ exon boundaries correctly predicted} * 100}{\text{number of correct predicted exons} + \text{number of wrong exons}}$$

$$3' = \frac{\text{number of true } 3' \text{ exon boundaries correctly predicted} * 100}{\text{number of correct predicted exons} + \text{number of wrong exons}}$$

At the protein level, we measure the accuracy of the protein products predicted by a program. Since a program may predict more than one transcript for a given gene sequence in the benchmark, we calculate the percent identity between the benchmark protein and all predicted proteins and the predicted protein with the highest percent identity score is selected. To calculate the percent identity score between the benchmark protein and the predicted

642 protein, we construct a pairwise alignment using the MAFFT software (version 7.307) [55]
643 and the percent identity is then defined as:
644
645 $$\%\, Identity = \frac{Number\ of\ identical\ amino\ acids\ * 100}{Length\ of\ benchmark\ protein}$$
646
647 ## Evaluation metric for Unconfirmed benchmark proteins
648 Since the Unconfirmed proteins in the benchmark are badly predicted and have at least
649 one identified sequence error, the %Identity score defined above for the Confirmed sequences
650 cannot be used. Instead, we compare the protein sequences predicted by the programs with the
651 most closely related Confirmed sequence found in the corresponding MSA. Thus, for a given
652 Unconfirmed sequence, *E*, we calculated the sequence identity between *E* (excluding the
653 sequence segments with predicted errors) and all the orthologous sequences in the
654 corresponding MSA. If a Confirmed orthologous sequence, *V*, was found that shared ≥50%
655 identity with *E*, then the sequence *V* was used as the reference protein to evaluate the program
656 prediction accuracy.
657 As before, a pairwise alignment between the prediction protein and sequence *V* was
658 constructed using MAFFT and the %Identity score was calculated. Finally, the accuracy score
659 was normalized by the sequence identity shared between the *E* and *V* benchmark sequences.
660 $$Accuracy = \frac{\%Identity(P,V)\ * 100}{\%Identity(E,V)}$$
661 ## Abbreviations
662 AA: Amino acid
663 BBS: Bardet-Biedl syndrome
664 Bp: Base pair
665 DNA: Deoxyribonucleic acid

EMC: Exon map complexity

F1: F1 score

FN: False negative

FP: False positive

HMM: Hidden Markov Model

Kb: Kilobase

ME: Missing exon

MSA: Multiple Sequence Alignment

RNA: Ribonucleic acid

Sn: Sensitivity

Sp: Specificity

TN: True negative

TP: True positive

UDT: Undetermined region

UTR: Untranslated region

WE: Wrong exon

Declarations

Ethics approval and consent to participate

Not applicable. All data presented in this article was extracted from publicly available sources.

Consent for publication

Not applicable

Availability of data and materials

The DNA and protein sequences used in the G3PO benchmark are available at http://git.lbgi.fr/scalzitti/Benchmark_study.

Competing interests

The authors declare that they have no competing interests.

Funding

NS was supported by funds from the Swiss foundation BIONIRIA. This work was also supported by the ANR projects Elixir-Excelerate: GA-676559 and RAinRARE: ANR-18-RAR3-0006-02, and Institute funds from the French Centre National de la Recherche Scientifique, the University of Strasbourg.

Authors' contributions

NS developed the benchmark, performed the program benchmarking, and produced all graphical presentations. AJG and PC advised on the feature content of the test sets and supervised the comparative analyses. OP and JDT supervised the production and exploitation of the benchmark. All authors participated in the definition of the original study concept. All authors read and approved the final manuscript.

Acknowledgements

The authors would like to thank the BISTRO and BICS Bioinformatics Platforms for their assistance.

Additional information

Additional file 1 Additional Tables S1-11, Additional Figures 1-8.

References

1. DNA Sequencing Costs: Data | NHGRI. https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data. Accessed 30 Oct 2019.

2. Matz MV. Fantastic Beasts and How To Sequence Them: Ecological Genomics for Obscure Model Organisms. Trends in Genetics. 2018;34:121–32.

3. Salzberg SL. Next-generation genome annotation: we still struggle to get it right. Genome Biol. 2019;20:92, s13059-019-1715–2.

4. Mudge JM, Harrow J. The state of play in higher eukaryote gene annotation. Nat Rev Genet. 2016;17:758–72.

5. Danchin A, Ouzounis C, Tokuyasu T, Zucker J-D. No wisdom in the crowd: genome annotation in the era of big data - current status and future prospects. Microb Biotechnol. 2018;11:588–605.

6. Ozsolak F, Platt AR, Jones DR, Reifenberger JG, Sass LE, McInerney P, et al. Direct RNA sequencing. Nature. 2009;461:814–8.

731 7. Workman RE, Tang AD, Tang PS, Jain M, Tyson JR, Zuzarte PC, et al. Nanopore native
732 RNA sequencing of a human poly(A) transcriptome. Nat Methods; 2019 (in press).

733 8. Yeh R-F, Lim LP, Burge CB. Computational Inference of Homologous Gene Structures in
734 the Human Genome. Genome Research. 2001;11:803–16.

735 9. Birney E. GeneWise and Genomewise. Genome Research. 2004;14:988–95.

736 10. Solovyev V, Kosarev P, Seledsov I, Vorobyev D. Automatic annotation of eukaryotic
737 genes, pseudogenes and promoters. Genome Biology. 2006;:12.

738 11. Stanke M, Schöffmann O, Morgenstern B, Waack S. Gene prediction in eukaryotes with a
739 generalized hidden Markov model that uses hints from external sources. BMC
740 Bioinformatics. 2006;7:62.

741 12. Kapustin Y, Souvorov A, Tatusova T, Lipman D. Splign: algorithms for computing
742 spliced alignments with identification of paralogs. Biol Direct. 2008;3:20.

743 13. Testa AC, Hane JK, Ellwood SR, Oliver RP. CodingQuarry: highly accurate hidden
744 Markov model gene prediction in fungal genomes using RNA-seq transcripts. BMC
745 Genomics. 2015;16:170.

746 14. Cook DE, Valle-Inclan JE, Pajoro A, Rovenich H, Thomma BPHJ, Faino L. Long-Read
747 Annotation: Automated Eukaryotic Genome Annotation Based on Long-Read cDNA
748 Sequencing. Plant Physiol. 2019;179:38–54.

749 15. Huang Y, Chen S-Y, Deng F. Well-characterized sequence features of eukaryote genomes
750 and implications for ab initio gene prediction. Computational and Structural Biotechnology
751 Journal. 2016;14:298–303.

752 16. Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA.
753 Journal of Molecular Biology. 1997;268:78–94.

754 17. Salzberg SL, Pertea M, Delcher AL, Gardner MJ, Tettelin H. Interpolated Markov Models
755 for Eukaryotic Gene Finding. Genomics. 1999;59:24–31.

756 18. Guigó R, Knudsen S, Drake N, Smith T. Prediction of gene structure. Journal of
757 Molecular Biology. 1992;226:141–57.

758 19. Korf I. Gene finding in novel genomes. BMC Bioinformatics. 2004;5:59.

759 20. Stanke M, Waack S. Gene prediction with a hidden Markov model and a new intron
760 submodel. Bioinformatics. 2003;19 Suppl 2:ii215–25.

761 21. Lomsadze A. Gene identification in novel eukaryotic genomes by self-training algorithm.
762 Nucleic Acids Research. 2005;33:6494–506.

763 22. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO:
764 assessing genome assembly and annotation completeness with single-copy orthologs.
765 Bioinformatics. 2015;31:3210–2.

766 23. Drăgan M-A, Moghul I, Priyam A, Bustos C, Wurm Y. GeneValidator: identify problems
767 with protein-coding gene predictions. Bioinformatics. 2016;32:1559–61.

768 24. Nishimura O, Hara Y, Kuraku S. Evaluating Genome Assemblies and Gene Models Using
769 gVolante. In: Kollmar M, editor. Gene Prediction. New York, NY: Springer New York; 2019.
770 p. 247–56.

771 25. Kemena C, Dohmen E, Bornberg-Bauer E. DOGMA: a web server for proteome and
772 transcriptome quality assessment. Nucleic Acids Research. 2019;47:W507–10.

773 26. Delcourt V, Staskevicius A, Salzet M, Fournier I, Roucou X. Small Proteins Encoded by
774 Unannotated ORFs are Rising Stars of the Proteome, Confirming Shortcomings in Genome
775 Annotations and Current Vision of an mRNA. Proteomics. 2018;18:1700058.

776 27. Mat-Sharani S, Firdaus-Raih M. Computational discovery and annotation of conserved
777 small open reading frames in fungal genomes. BMC Bioinformatics. 2019;19:551.

778 28. Rajput B, Pruitt KD, Murphy TD. RefSeq curation and annotation of stop codon recoding
779 in vertebrates. Nucleic Acids Research. 2019;47:594–606.

780 29. Burset M, Guigó R. Evaluation of Gene Structure Prediction Programs. Genomics.
781 1996;34:353–67.

782 30. Rogic S, Mackworth AK, Ouellette FBF. Evaluation of Gene-Finding Programs on
783 Mammalian Sequences. Genome Research. 2001;11:817–32.

784 31. Guigo R. An Assessment of Gene Prediction Accuracy in Large DNA Sequences.
785 Genome Research. 2000;10:1631–42.

786 32. Guigó R, Flicek P, Abril JF, Reymond A, Lagarde J, Denoeud F, et al. EGASP: the
787 human ENCODE Genome Annotation Assessment Project. Genome Biology. 2006;:31.

788 33. Goodswen SJ, Kennedy PJ, Ellis JT. Evaluating High-Throughput Ab Initio Gene Finders
789 to Discover Proteins Encoded in Eukaryotic Pathogen Genomes Missed by Laboratory
790 Techniques. PLoS ONE. 2012;7:e50609.

791 34. The UniProt Consortium. UniProt: the universal protein knowledgebase. Nucleic Acids
792 Res. 2017;45:D158–69.

793 35. Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, et al. The Ensembl
794 genome database project. Nucleic Acids Res. 2002;30:38–41.

795 36. Koepfli K-P, Paten B, the Genome 10K Community of Scientists, O'Brien SJ. The
796 Genome 10K Project: A Way Forward. Annu Rev Anim Biosci. 2015;3:57–111.

797 37. Zhang G. Bird sequencing project takes off. Nature. 2015;522:34–34.

798 38. Albertin CB, Bonnaud L, Brown CT, Crookes-Goodson WJ, da Fonseca RR, Di Cristo C,
799 et al. Cephalopod genomics: A plan of strategies and organization. Stand Genomic Sci.
800 2012;7:175–88.

801 39. Lewin HA, Robinson GE, Kress WJ, Baker WJ, Coddington J, Crandall KA, et al. Earth
802 BioGenome Project: Sequencing life for the future of life. Proc Natl Acad Sci USA.
803 2018;115:4325–33.

804 40. Wilbrandt J, Misof B, Panfilio KA, Niehuis O. Repertoire-wide gene structure analyses: a
805 case study comparing automatically predicted and manually annotated gene models. BMC
806 Genomics. 2019;20:753.

807 41. Schnoes AM, Brown SD, Dodevski I, Babbitt PC. Annotation Error in Public Databases:
808 Misannotation of Molecular Function in Enzyme Superfamilies. PLoS Comput Biol. 2009;5.

809 42. Yandell M, Ence D. A beginner's guide to eukaryotic genome annotation. Nature Reviews
810 Genetics. 2012;13:329–42.

811 43. Sberro H, Fremin BJ, Zlitni S, Edfors F, Greenfield N, Snyder MP, et al. Large-Scale
812 Analyses of Human Microbiomes Reveal Thousands of Small, Novel Genes. Cell.
813 2019;178:1245-1259.e14.

814 44. Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M. Gene prediction in
815 novel fungal genomes using an ab initio algorithm with unsupervised training. Genome Res.
816 2008;18:1979–90.

817 45. Reid I, O'Toole N, Zabaneh O, Nourzadeh R, Dahdouli M, Abdellateef M, et al.
818 SnowyOwl: accurate prediction of fungal genes by using RNA-Seq and homology
819 information to select among ab initio models. BMC Bioinformatics. 2014;15:229.

820 46. Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. BRAKER1: Unsupervised
821 RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS: Table 1.
822 Bioinformatics. 2016;32:767–9.

823 47. Matera AG, Wang Z. A day in the life of the spliceosome. Nat Rev Mol Cell Biol.
824 2014;15:108–21.

825 48. Zhang Y, Liu X, MacLeod J, Liu J. Discerning novel splice junctions derived from RNA-
826 seq alignment: a deep learning approach. BMC Genomics. 2018;19. doi:10.1186/s12864-018-
827 5350-1.

828 49. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D,
829 Li YI, et al. Predicting Splicing from Primary Sequence with Deep Learning. Cell.
830 2019;176:535-548.e24.

831 50. Nevers Y, Kress A, Defosset A, Ripp R, Linard B, Thompson JD, et al. OrthoInspector
832 3.0: open portal for comparative genomics. Nucleic Acids Res. 2019;47 Database
833 issue:D411–8.

834 51. Khenoussi W, Vanhoutrève R, Poch O, Thompson JD. SIBIS: a Bayesian model for
835 inconsistent protein sequence estimation. Bioinformatics. 2014;30:2432–9.

836 52. Rodriguez JM, Maietta P, Ezkurdia I, Pietrelli A, Wesselink J-J, Lopez G, et al. APPRIS:
837 annotation of principal and alternative splice isoforms. Nucleic Acids Res. 2013;41 Database
838 issue:D110–7.

839 53. Kozak M. Possible role of flanking nucleotides in recognition of the AUG initiator codon
840 by eukaryotic ribosomes. Nucleic Acids Res. 1981;9:5233–52.

841 54. Gao K, Masuda A, Matsuura T, Ohno K. Human branch point consensus sequence is
842 yUnAy. Nucleic Acids Res. 2008;36:2257–67.

843 55. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7:
844 Improvements in Performance and Usability. Mol Biol Evol. 2013;30:772–80.

845

Tables and Figures

Gene predictor	Signal sensors	Content sensors	Algorithm model	Organism-specific models
Genscan (version 1.0)	Promoter (15 bp), cap site (8 bp), TATA to cap site distance of 30 to 36 bp, donor (−3 to +6 bp)/acceptor (−20 to +3) splice sites, polyadenylation, translation start/stop sites	Intergenic, 5'-/3'-UTR, exon/introns in 3 phases, forward/reverse strands	3-periodic fifth-order Markov model (GHMM)	3 models
GlimmerHMM (version 3.02)	Donor (16 bp)/ acceptor (29 bp) splice sites, start/stop codons	Exon/intron in one frame,intron length 50-1500 bp, total coding length >200 bp	Hidden Markov model (GHMM)	5 models
GeneID (version 1.4)	Donor/acceptor splice sites (−3 to +6 bp), start/stop codons	First/initial/last exon, single-exon gene, intron, intron length >40 bp, intergenic distance >300 bp	Fifth-order Markov model (HMM)	66 models
SNAP (version 2006-07-28)	Donor (-3 to +6 bp) /acceptor (-24 to +3) splice sites, translation start (-6 to +6 bp) /stop (-6 to +3 bp) sites	intergenic, single-exon gene, first/initial/last exon, introns in 3 phases	Fourth-order Markov model (GHMM)	11 models
Augustus (version 3.3.2)	Donor (-3 to +6 bp) /acceptor (-5 to +1 bp) splice sites, branch point (32 bp), translation start (-20 to +3)/stop (3 bp) sites	intergenic, single exon gene, first/initial/last exon, short/long introns in 3 phases and forward/reverse strands	Fourth-order Markov model (GHMM)	109 models

Table 1. Main characteristics of the gene prediction programs evaluated in this study. GHMM: Generalized hidden Markov model; UTR: Untranslated regions.

	Confirmed proteins (%Identity)	Unconfirmed proteins (%Identity)
Augustus	77.07	63.31
Genscan	66.57	61.14
GeneID	53.05	50.11
GlimmerHMM	63.93	53.17
Snap	56.57	47.19

Table 2. Effect of protein sequence quality measured at the protein level. %Identity indicates the average sequence identity observed between the predicted and benchmark protein sequences for the test sets of Confirmed and Unconfirmed proteins.

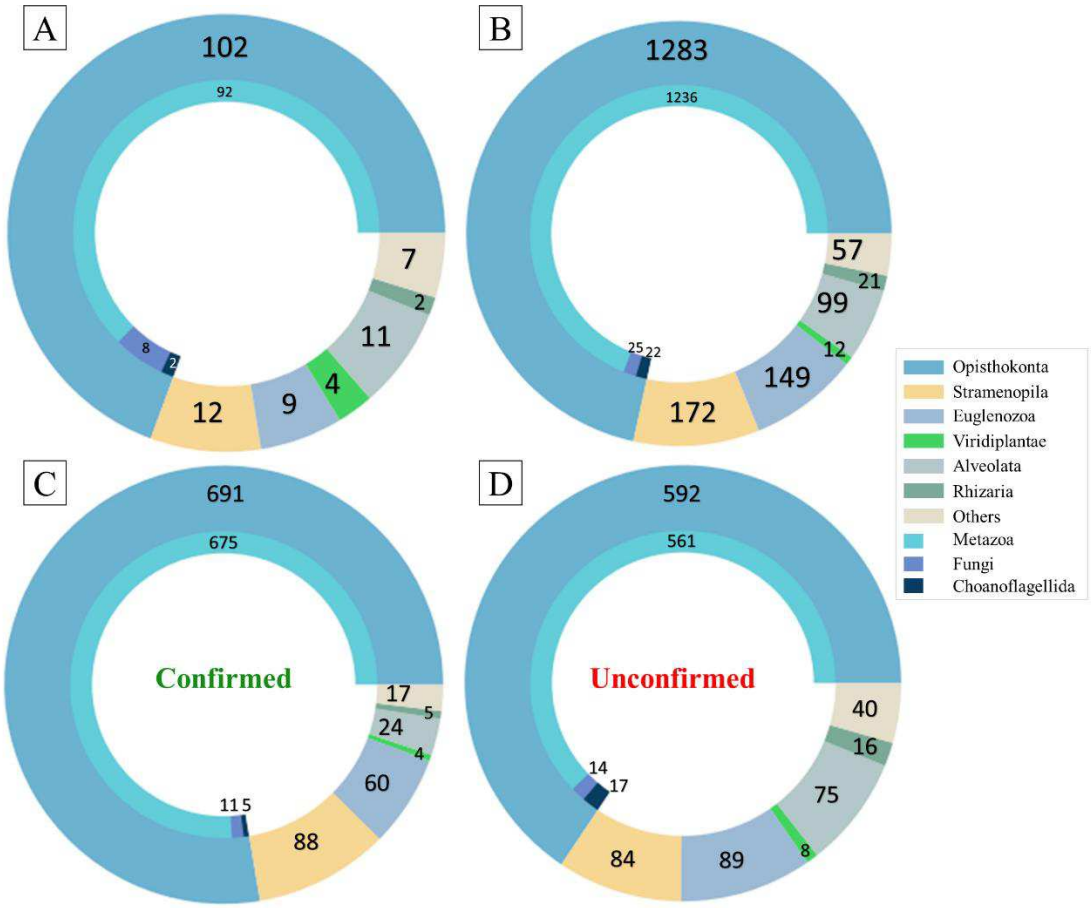

Fig. 1. Phylogenetic distribution of the 1793 test cases in the G3PO benchmark. A) Number of species in each clade. B) Number of sequences in each clade. C) Number of sequences in each clade in the Confirmed test set. D) Number of sequences in each clade in the Unconfirmed test set. The 'Others' group corresponds to: Apusozoa, Cryptophyta, Diplomonadida, Haptophyceae, Heterolobosea, Parabasalia.

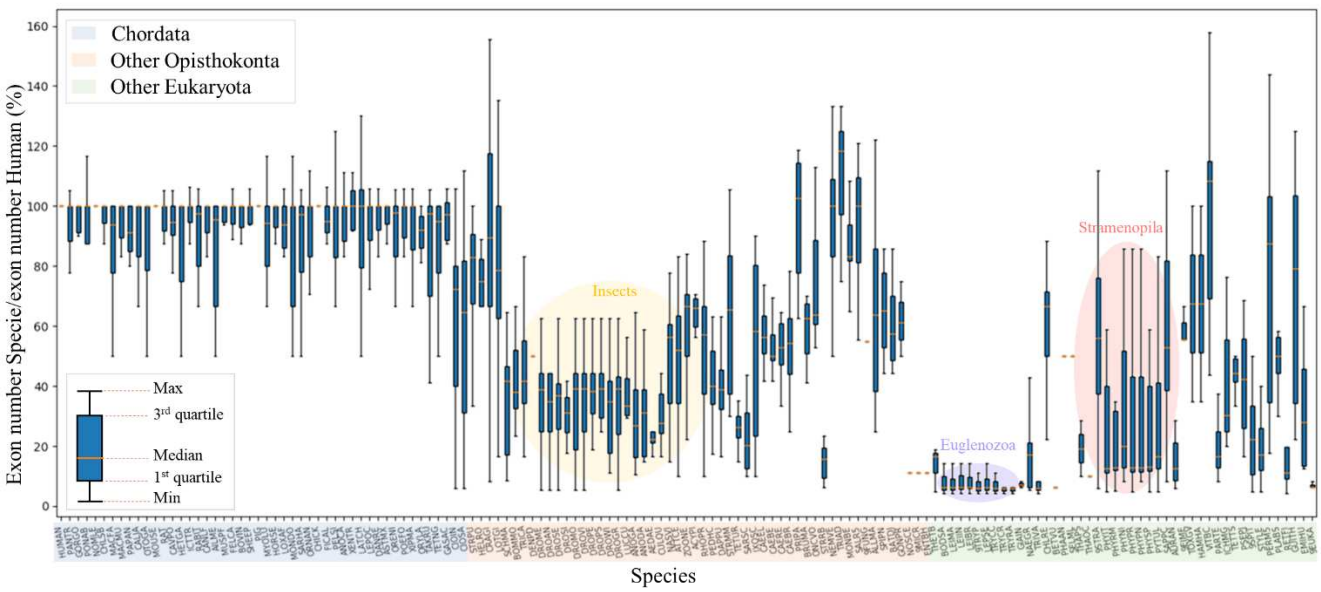

Fig. 2. Exon map complexity for each species. Each box plot represents the distribution of the ratio of the number of exons in the gene of a given species (Exon Number Species), to the number of exons in the orthologous human gene (Exon number Human), for all genes in the benchmark. Notable clades include Insects (BOMMO to PEDHC), Euglenozoa (BODSA to TRYRA) or Stramenopila (THAPS to AURAN).

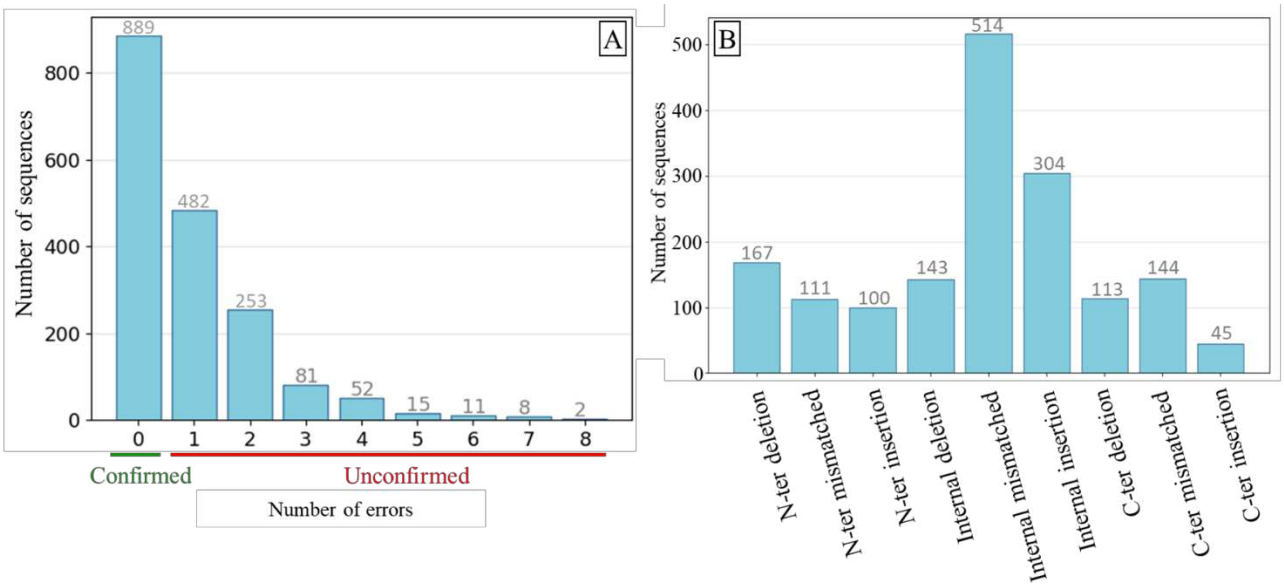

Fig. 3. A) Number of identified sequence errors in the 1793 benchmark proteins. B) Number of 'Unconfirmed' protein sequences for each error category.

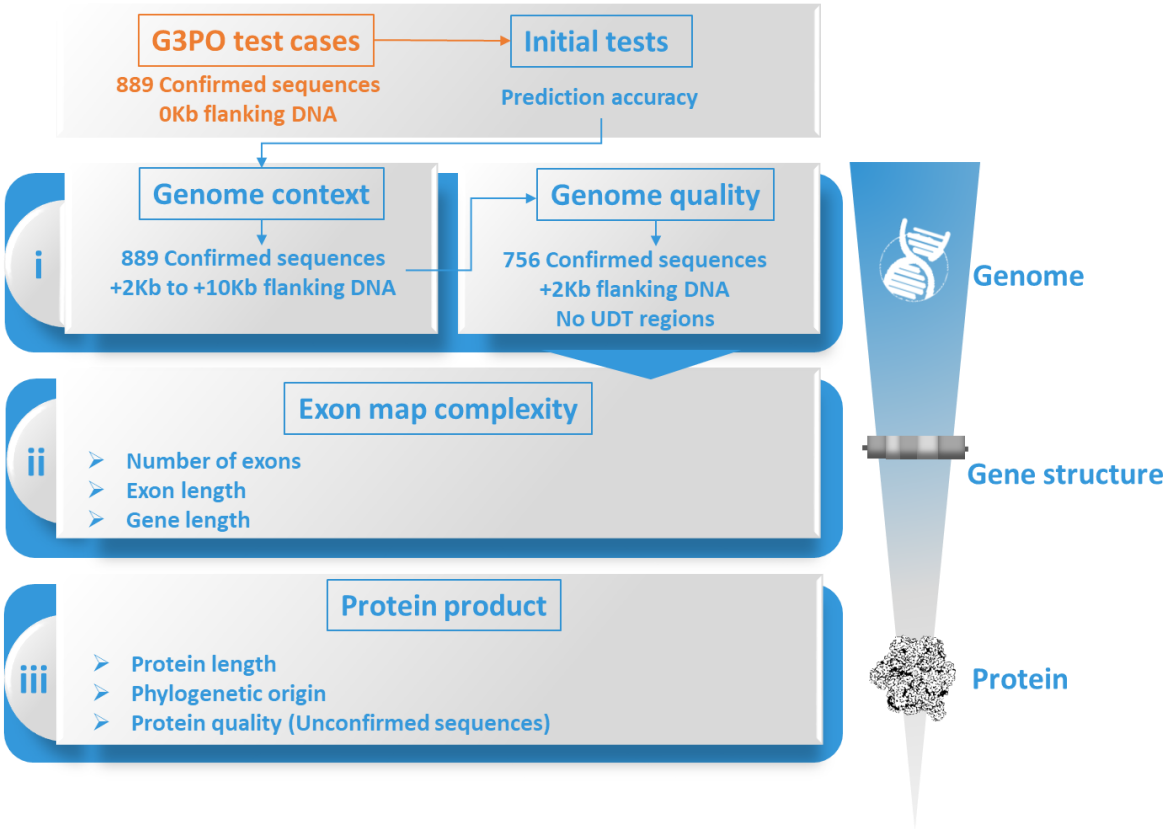

Fig. 4. Workflow of different tests performed to evaluate gene prediction accuracy. The initial tests are based on the 889 confirmed proteins and their genomic sequences corresponding to the gene region only (0Kb flanking sequence). At the genome level, effect of genome context and genome quality are tested, and 756 confirmed sequences with +2Kb flanking sequences and no undetermined (UDT) regions are selected. These are used at the gene structure and protein levels, to investigate effects of factors linked to exon map complexity and the final protein product.

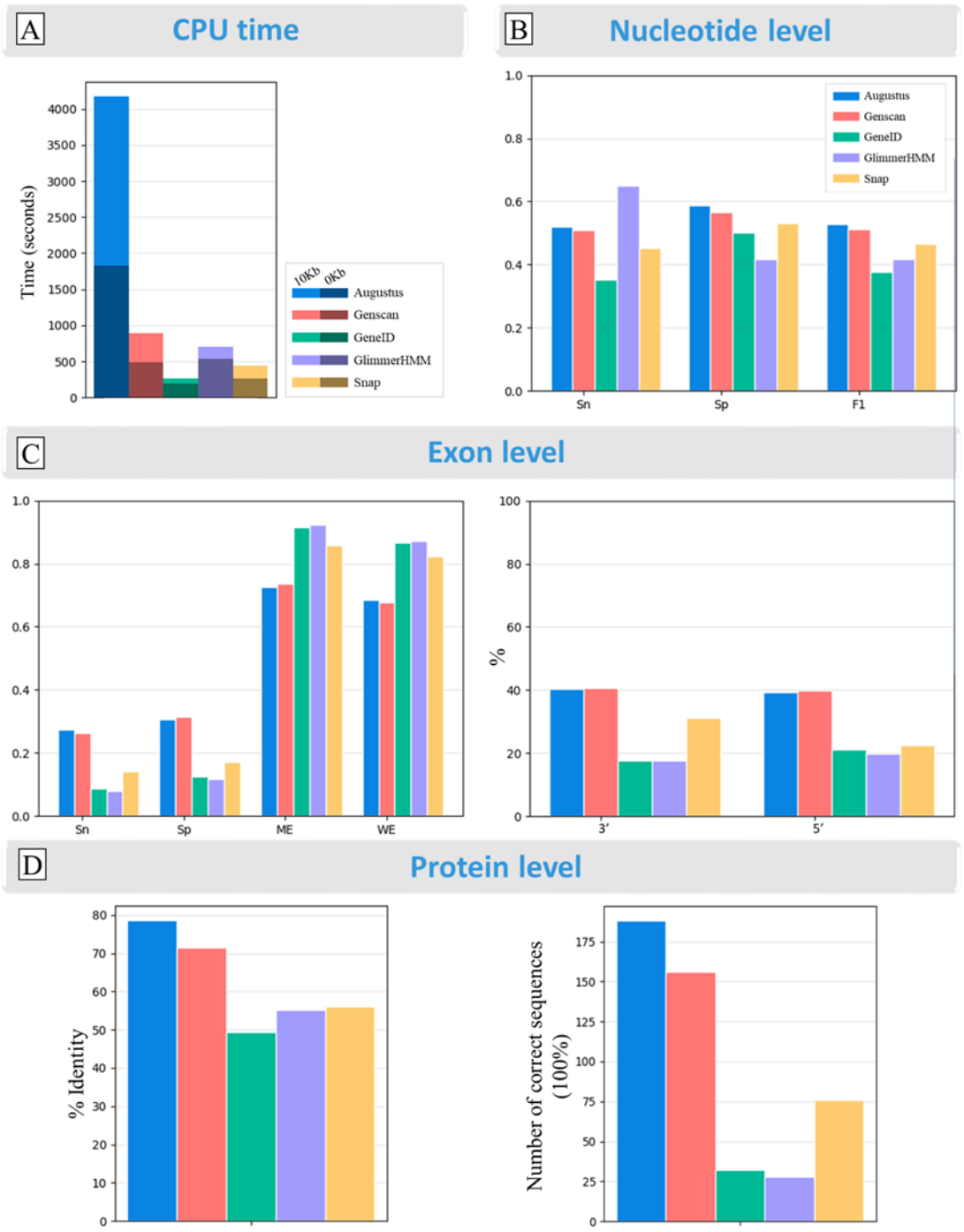

Fig. 5. A) Total time required to process the 1793 genomic sequences covering the gene region only (dark colors) and with 10Kb upstream/downstream flanking regions (light colors). B-D) Overall performance of the five gene prediction programs, using the 889 Confirmed sequences, at the nucleotide, exon and protein levels. Sn=sensitivity; Sp=specificity; F1=F1 score; ME=Missing Exon; WE=Wrong Exon; 3' and 5' are the percentage of correctly predicted 3' and 5' exon boundaries. %Identity indicates the average sequence identity observed between the predicted proteins and the Confirmed benchmark sequences.

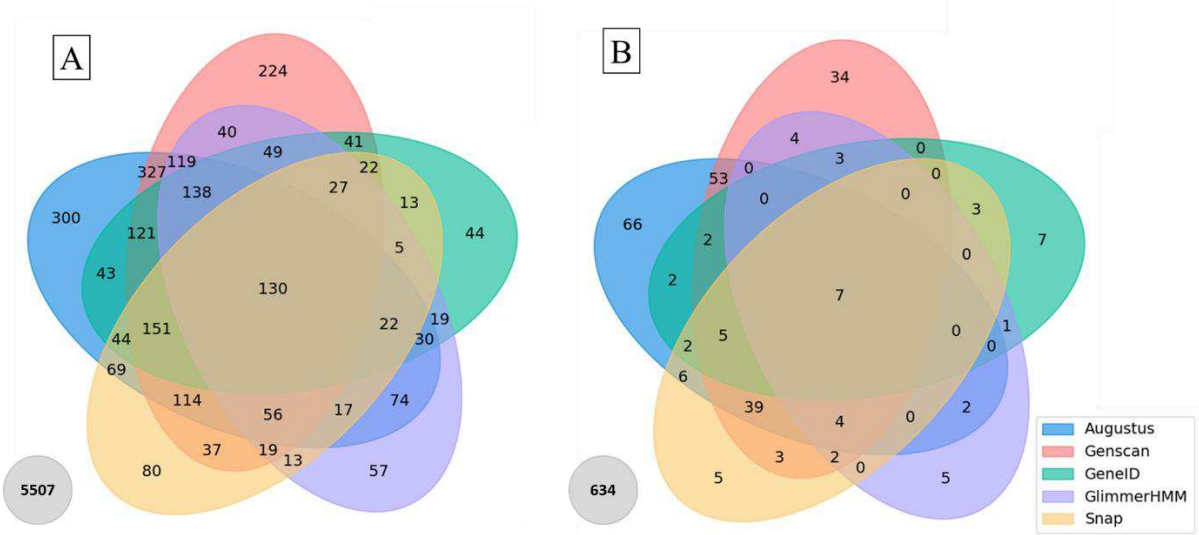

Fig. 6. Venn diagrams representing A) the number of correct exons predicted by each program, and B) the number of perfectly predicted proteins by each program. The grey circles indicate the number of exons/proteins badly predicted by all programs.

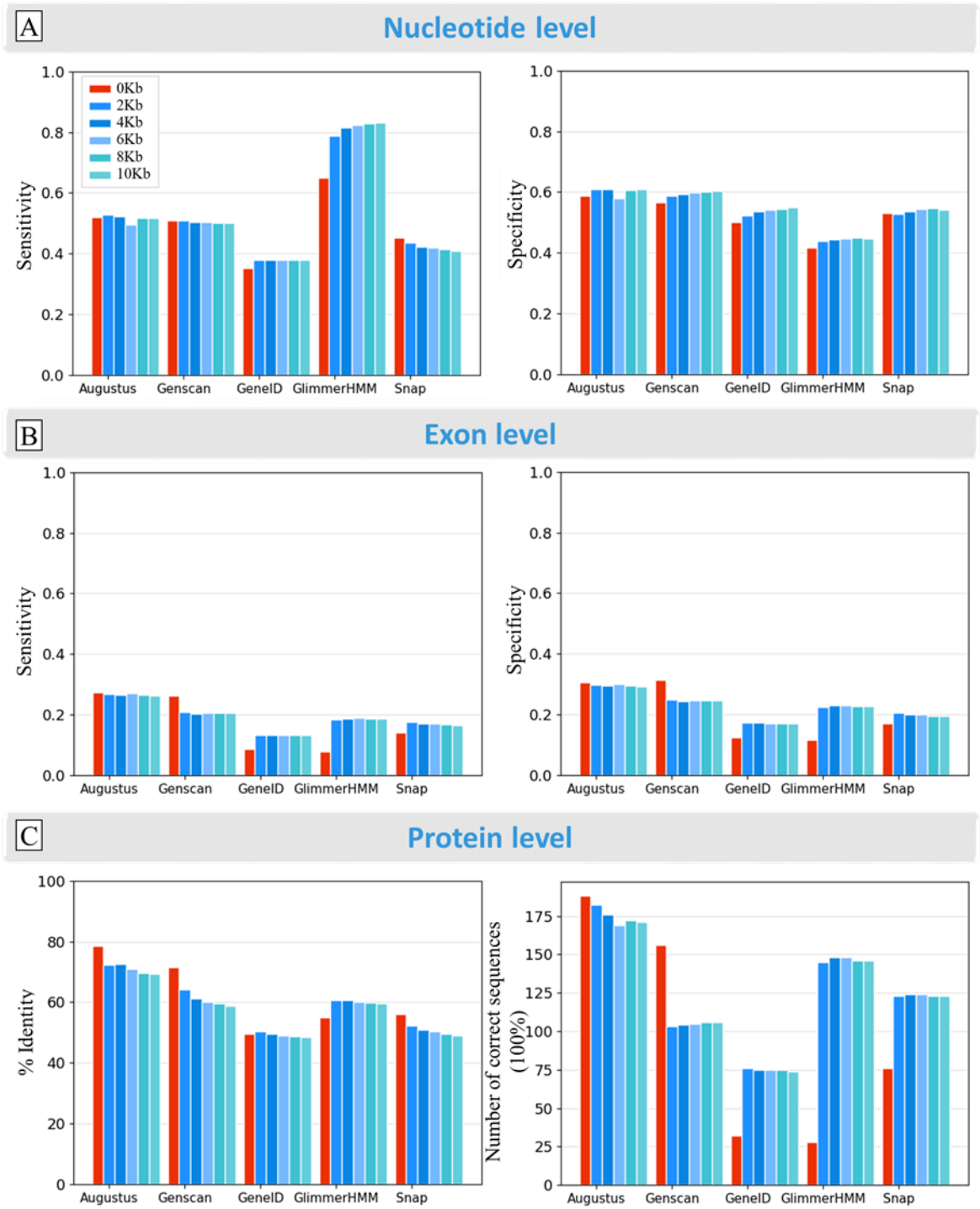

Fig. 7. Effect of the genomic context based on the different lengths of upstream/downstream flanking genomic sequences on the performance of the five gene prediction programs. A) sensitivity and specificity of prediction of coding nucleotides. B) sensitivity and specificity of exon prediction. C) accuracy of protein sequence prediction (% identity) and number of proteins correctly predicted with 100% identity.

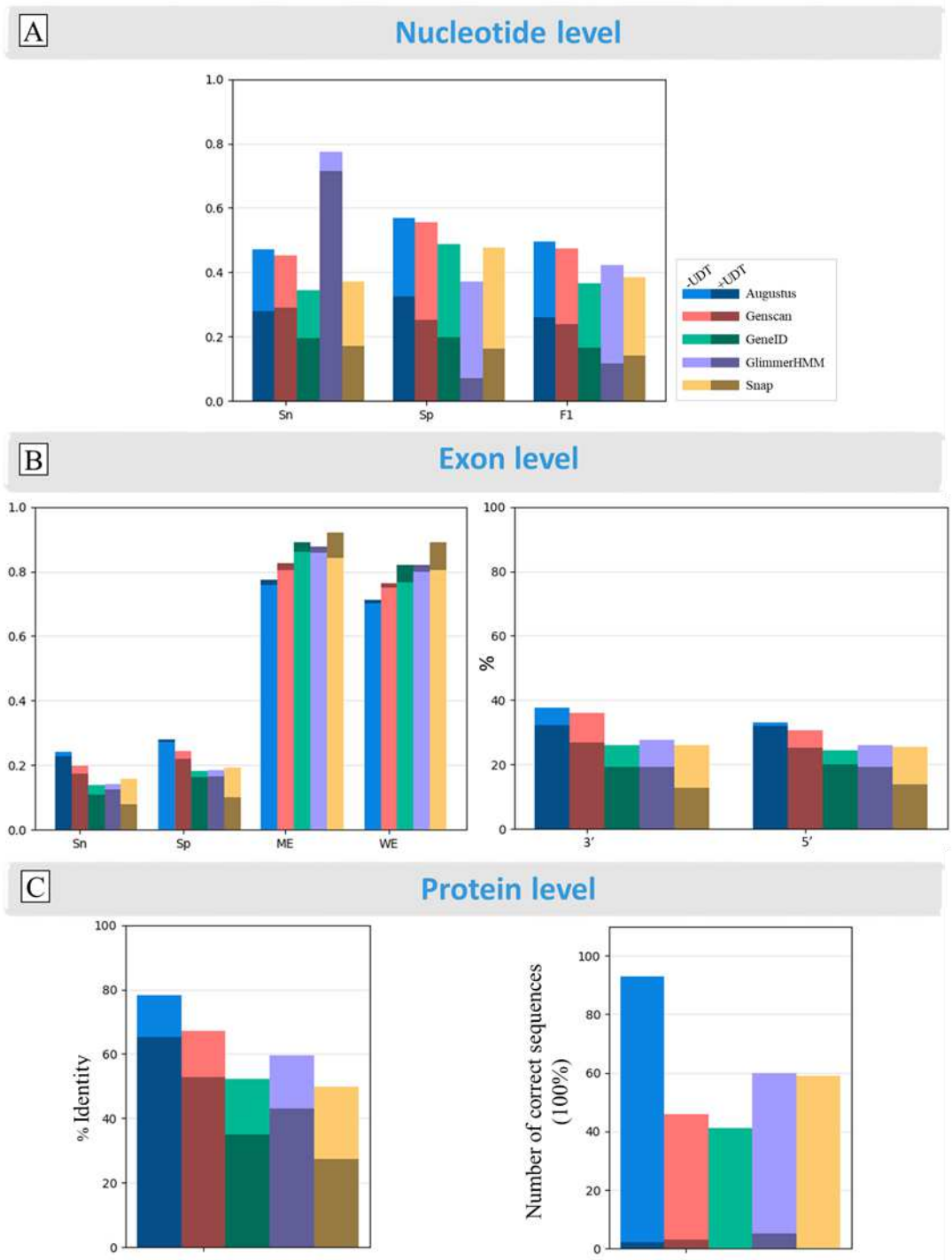

Fig. 8. Effect of undetermined sequence regions (UDT) on prediction performance of the five gene prediction programs, using Confirmed benchmark sequences from Metazoa, where 542 sequences have no undetermined regions (-UDT: light colors) and 133 sequences have undetermined regions (+UDT: dark colors). A) sensitivity and specificity of nucleotide prediction. B) sensitivity and specificity of exon prediction. C) accuracy of protein sequence prediction (% identity) and number of proteins correctly predicted with 100% identity. Sn=sensitivity; Sp=specificity; F1=F1 score; ME=Missing Exons; WE=Wrong Exons; 3' and 5' are the percentage of correctly predicted 3' and 5' exon boundaries. %Identity indicates the sequence identity observed between the predicted proteins and the Confirmed benchmark sequences.

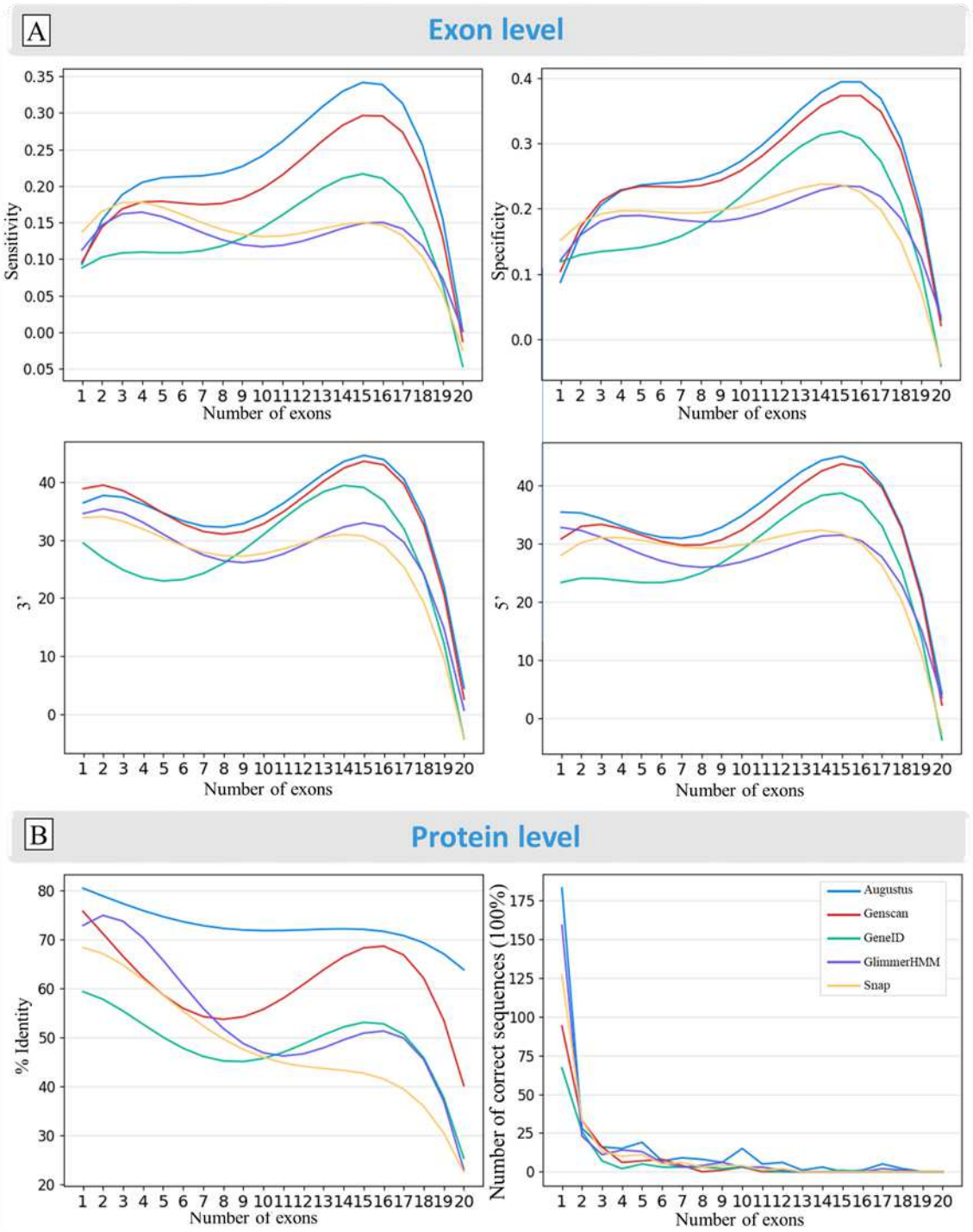

Fig. 9. Effect of exon map complexity on prediction quality at the A) exon and B) protein levels. A 4th degree polynomial curve fitting was used to represent the results more clearly. Sequences with 21-24 exons were not included, due to the low number of sequences in the benchmark with these exon counts. 3' and 5' are the proportion of correctly predicted 3' and 5' exon boundaries respectively. %Identity indicates the sequence identity observed between the predicted proteins and the Confirmed benchmark sequences.

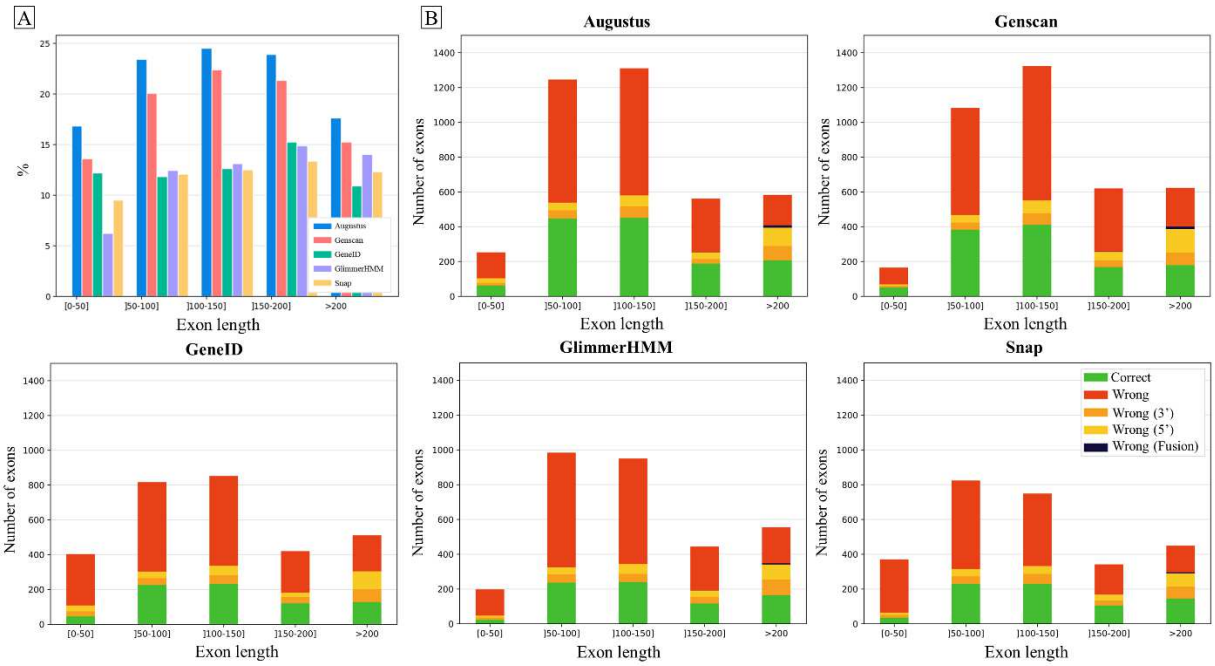

Fig. 10. Effect of exon length on exon prediction quality. A) Proportion of benchmark exons correctly predicted depending on the exon length. B) Number of exons predicted correctly, with one of the 5' or 3' exon boundaries correct, or with both boundaries wrongly predicted, for each of the five programs.

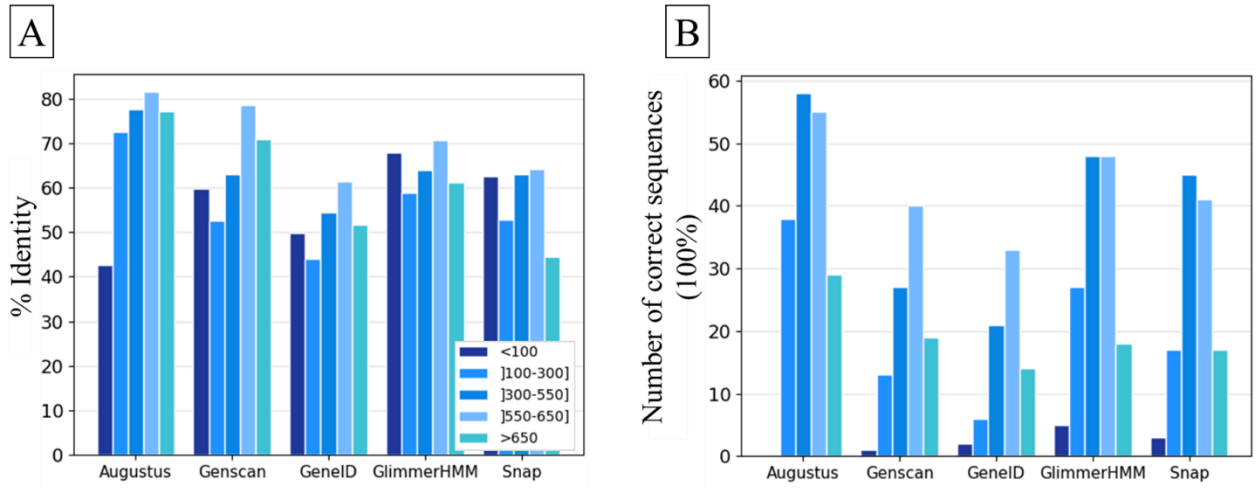

Fig. 11. Effect of protein length on prediction accuracy: A) average percent identity between the predicted and the benchmark protein sequences, B) number of proteins perfectly predicted with 100% sequence identity.

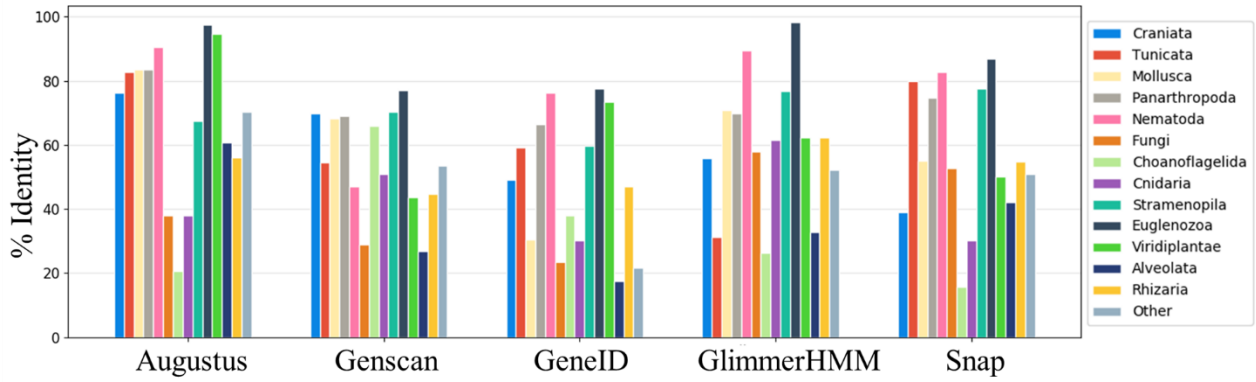

Fig. 12. Prediction performance for sequences from different clades. The 'Other' group contains the Apusozoa, Cryptophyta, Diplomonadida, Haptophyceae, Heterolobosea, Parabasalia clades, as well as Placozoa, Annelida and urchin. % Identity indicates the average percent identity between the predicted and the benchmark protein sequences.

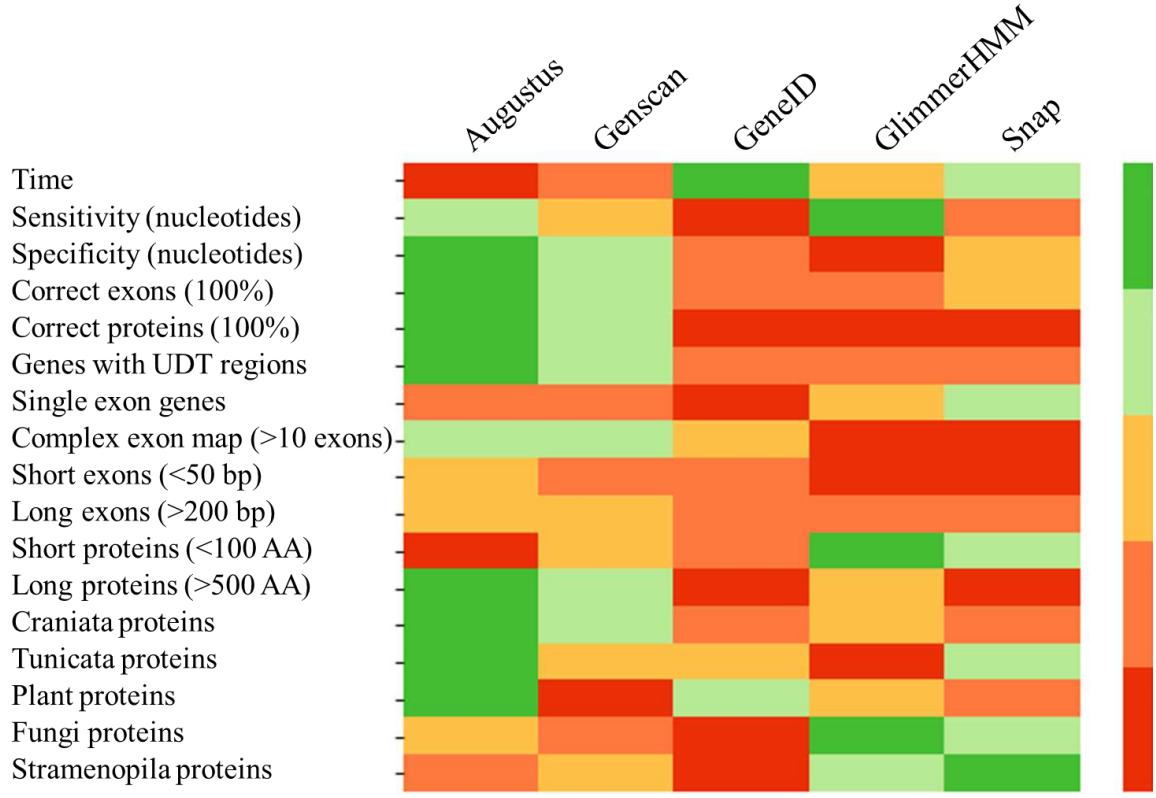

Fig. 13. Strengths and weaknesses of the gene prediction programs evaluated in this study. Heatmap colors are: dark green = best program, light green = 2nd best program, yellow = 3rd best program, orange = 4th best program, red = 5th best program.

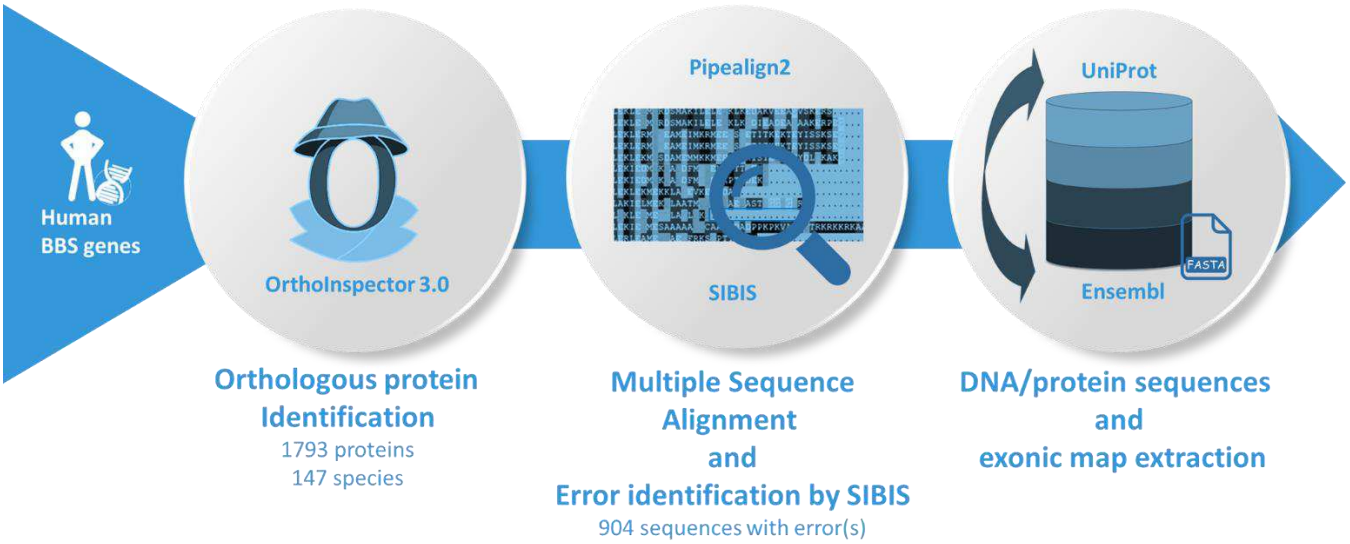

Fig. 14. Schematic view of the pipeline used to construct the benchmark.

Figures

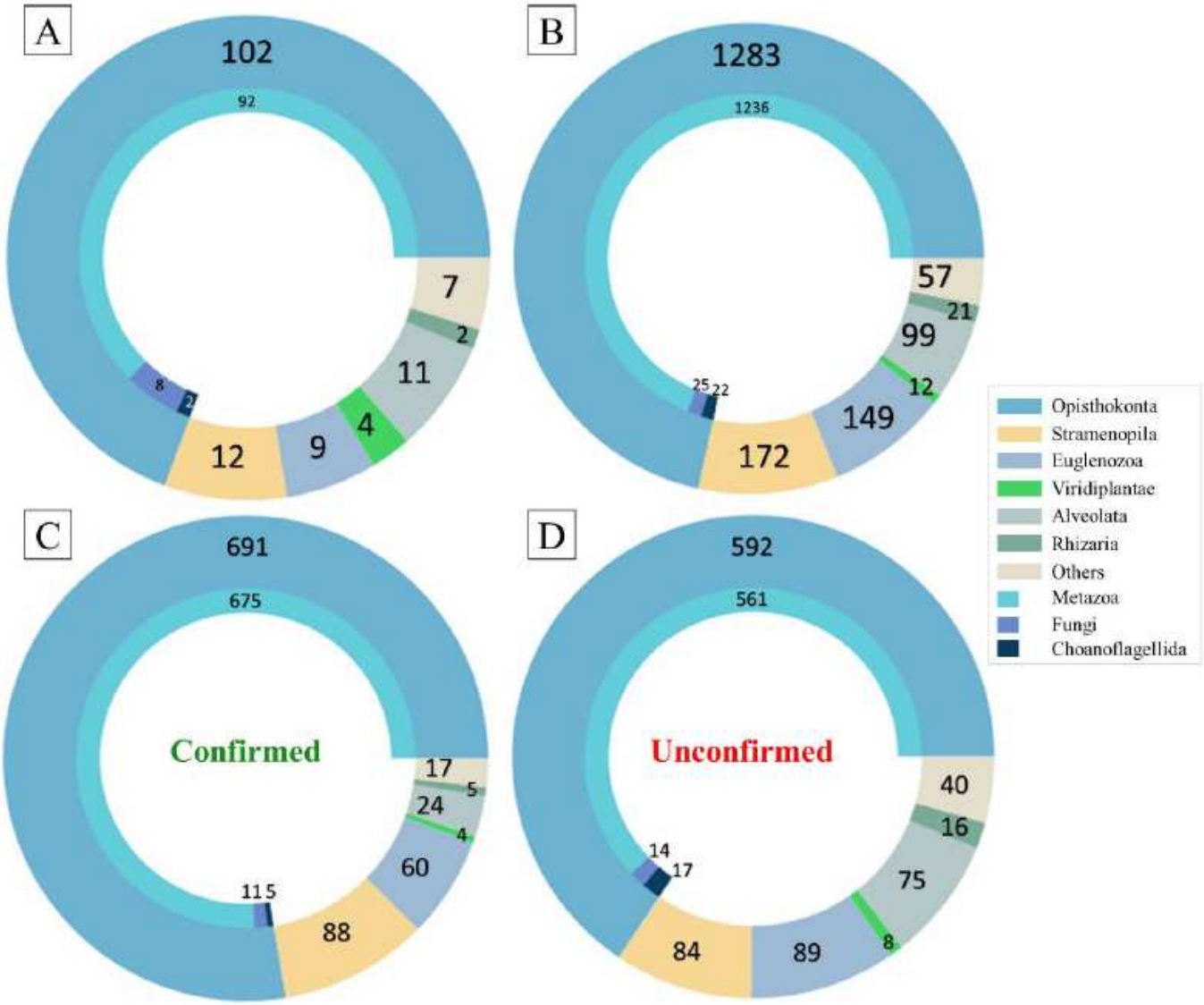

Figure 1

Phylogenetic distribution of the 1793 test cases in the G3PO benchmark. A) Number of species in each clade. B) Number of sequences in each clade. C) Number of sequences in each clade in the Confirmed test set. D) Number of sequences in each clade in the Unconfirmed test set. The ‘Others’ group corresponds to: Apusozoa, Cryptophyta, Diplomonadida, Haptophyceae, Heterolobosea, Parabasalia.

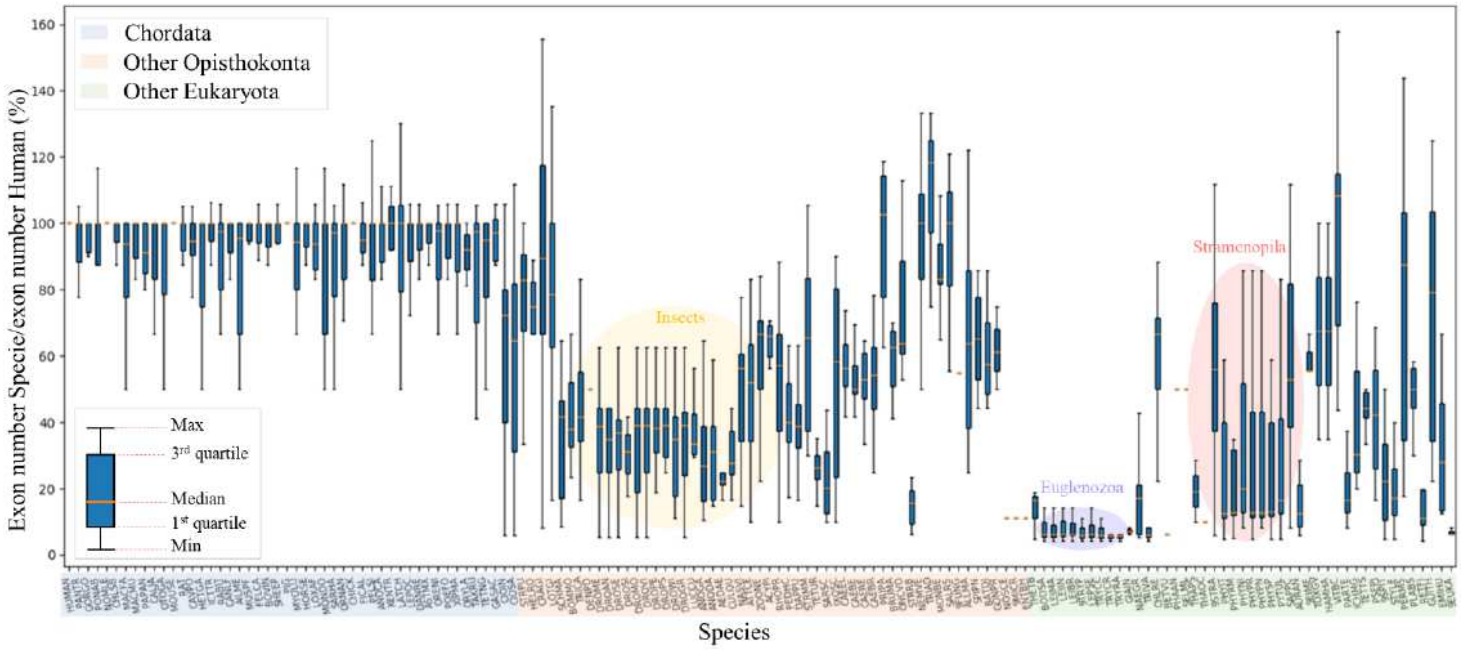

Figure 2

Exon map complexity for each species. Each box plot represents the distribution of the ratio of the number of exons in the gene of a given species (Exon Number Species), to the number of exons in the orthologous human gene (Exon number Human), for all genes in the benchmark. Notable clades include Insects (BOMMO to PEDHC), Euglenozoa (BODSA to TRYRA) or Stramenopila (THAPS to AURAN).

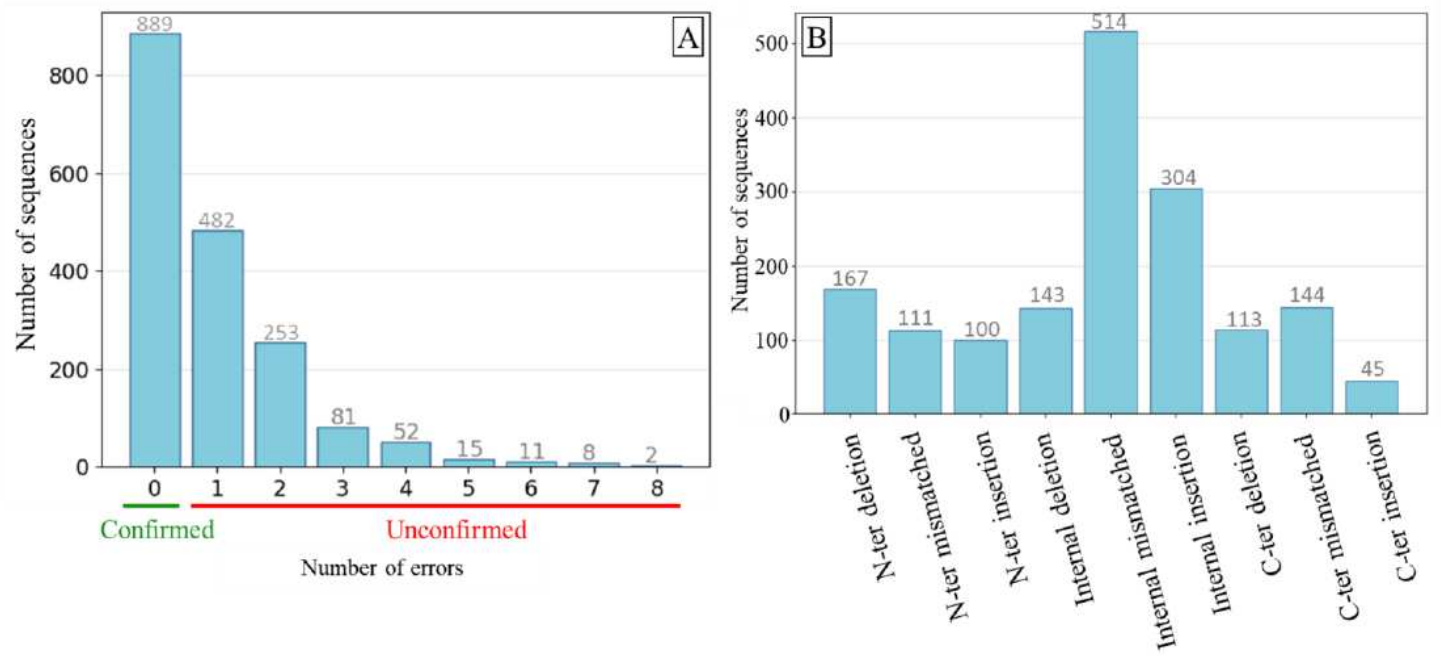

Figure 3

A) Number of identified sequence errors in the 1793 benchmark proteins. B) Number of 'Unconfirmed' protein sequences for each error category.

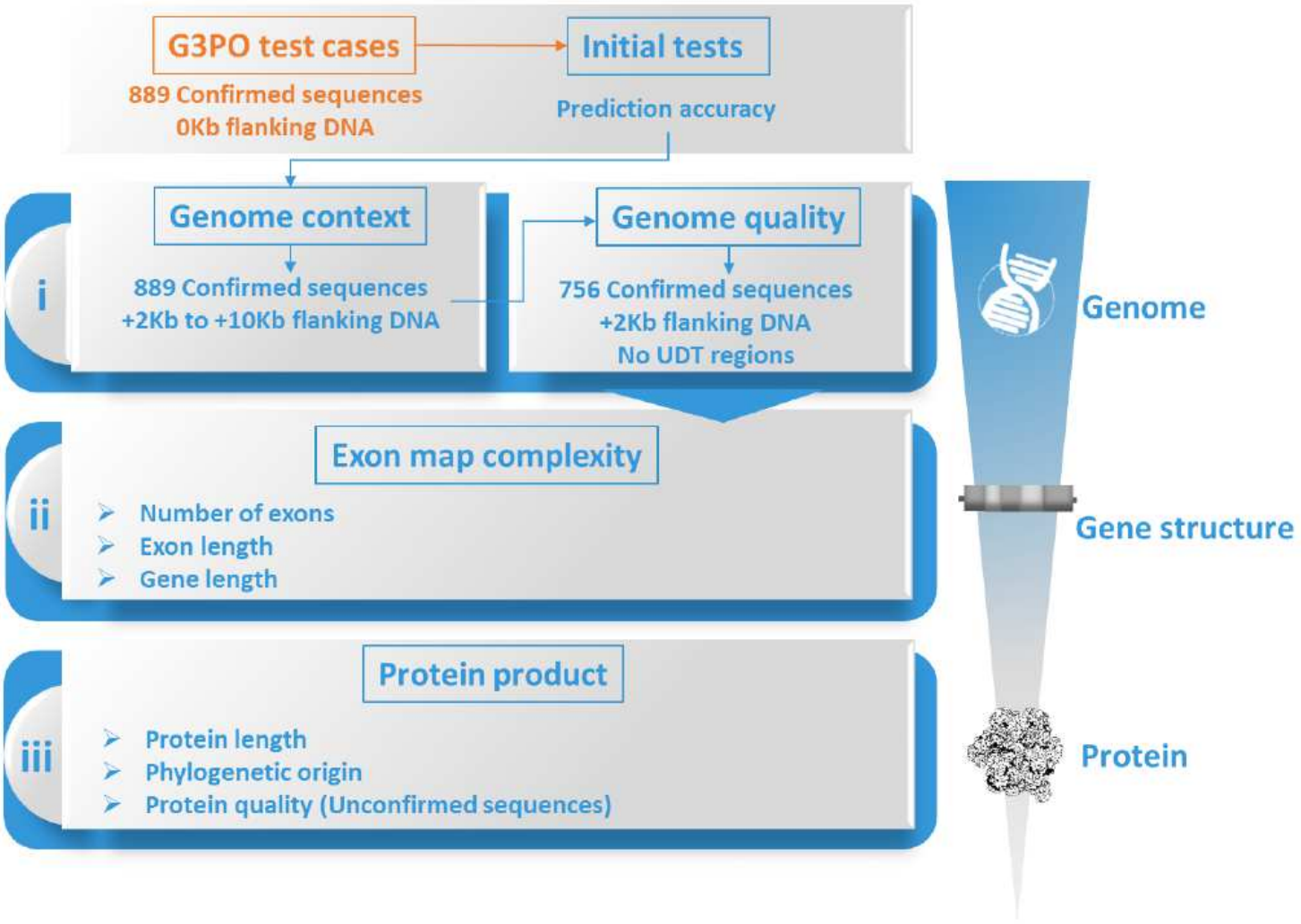

Figure 4

Workflow of different tests performed to evaluate gene prediction accuracy. The initial tests are based on the 889 confirmed proteins and their genomic sequences corresponding to the gene region only (0Kb flanking sequence). At the genome level, effect of genome context and genome quality are tested, and 756 confirmed sequences with +2Kb flanking sequences and no undetermined (UDT) regions are selected. These are used at the gene structure and protein levels, to investigate effects of factors linked to exon map complexity and the final protein product.

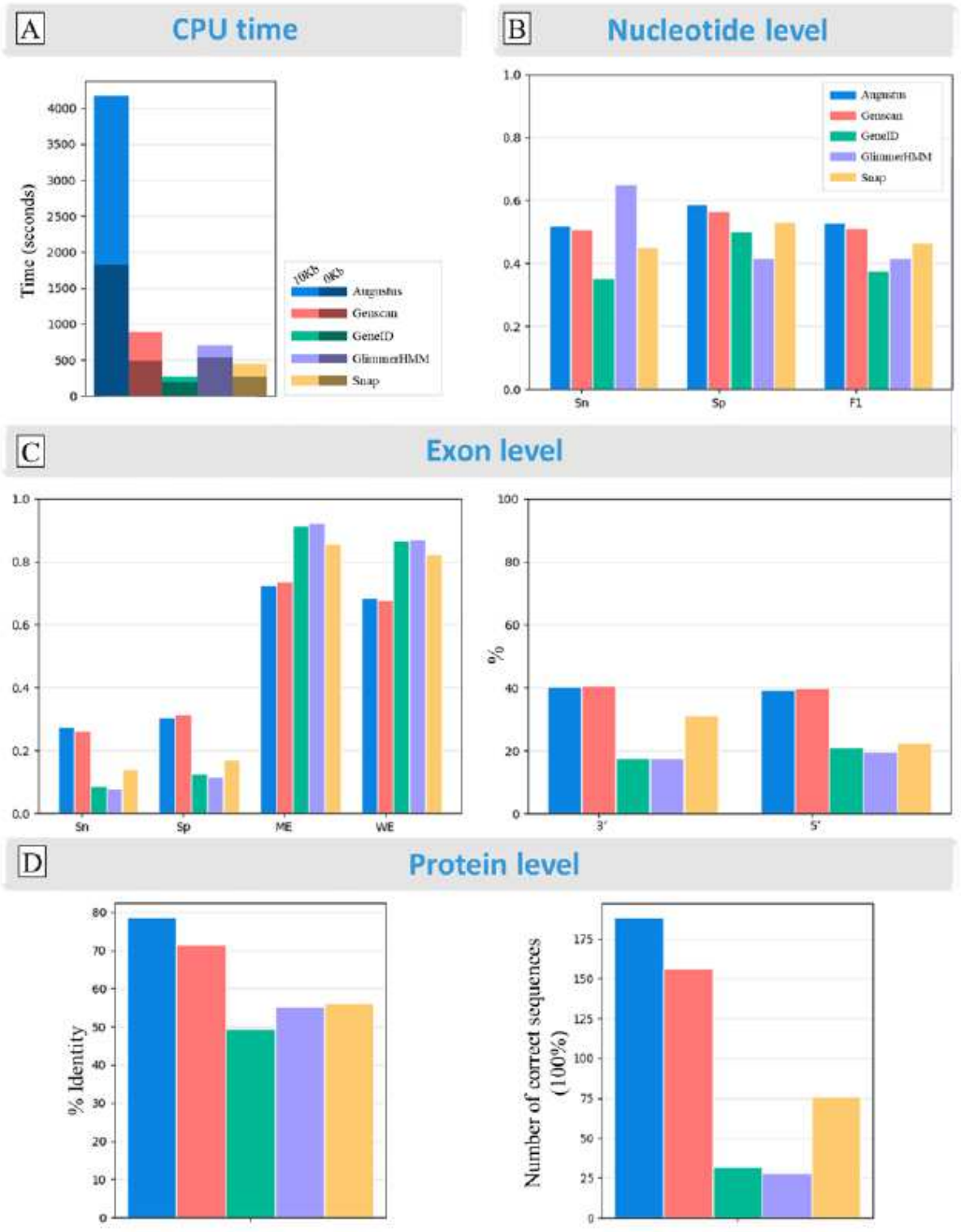

Figure 5

A) Total time required to process the 1793 genomic sequences covering the gene region only (dark colors) and with 10Kb upstream/downstream flanking regions (light colors). B-D) Overall performance of the five gene prediction programs, using the 889 Confirmed sequences, at the nucleotide, exon and protein levels. Sn=sensitivity; Sp=specificity; F1=F1 score; ME=Missing Exon; WE=Wrong Exon; 3' and 5' are the

percentage of correctly predicted 3’ and 5’ exon boundaries. %Identity indicates the average sequence identity observed between the predicted proteins and the Confirmed benchmark sequences.

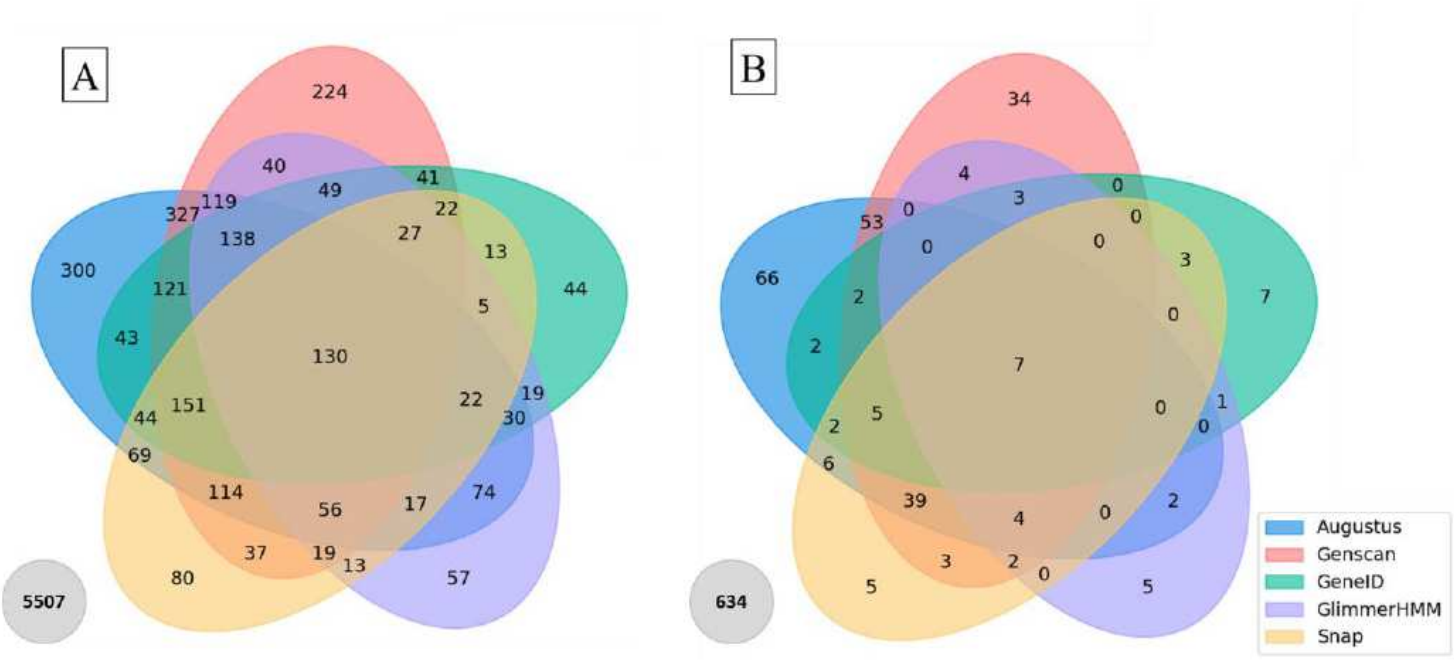

Figure 6

Venn diagrams representing A) the number of correct exons predicted by each program, and B) the number of perfectly predicted proteins by each program. The grey circles indicate the number of exons/proteins badly predicted by all programs.

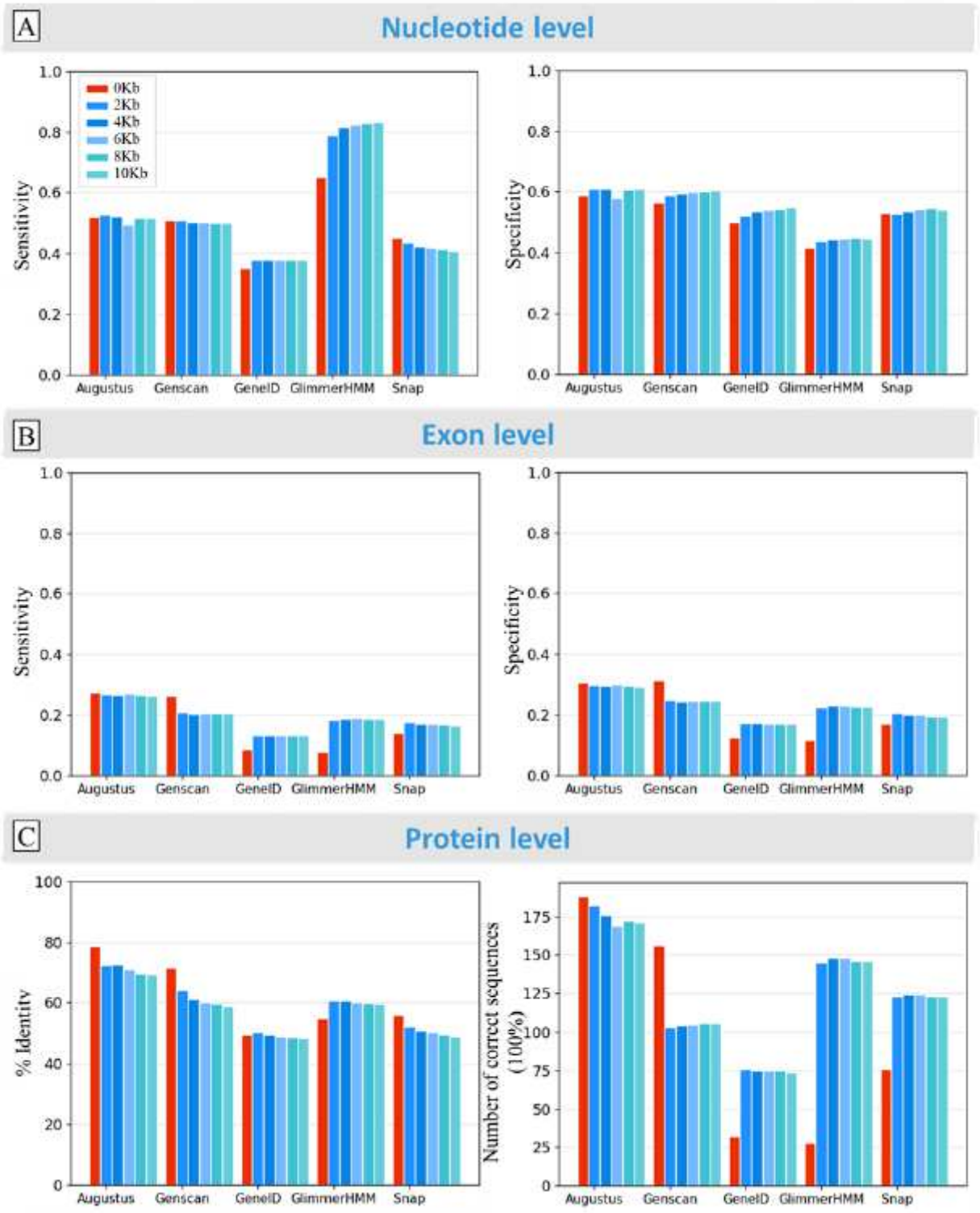

Figure 7

Effect of the genomic context based on the different lengths of upstream/downstream flanking genomic sequences on the performance of the five gene prediction programs. A) sensitivity and specificity of prediction of coding nucleotides. B) sensitivity and specificity of exon prediction. C) accuracy of protein sequence prediction (% identity) and number of proteins correctly predicted with 100% identity.

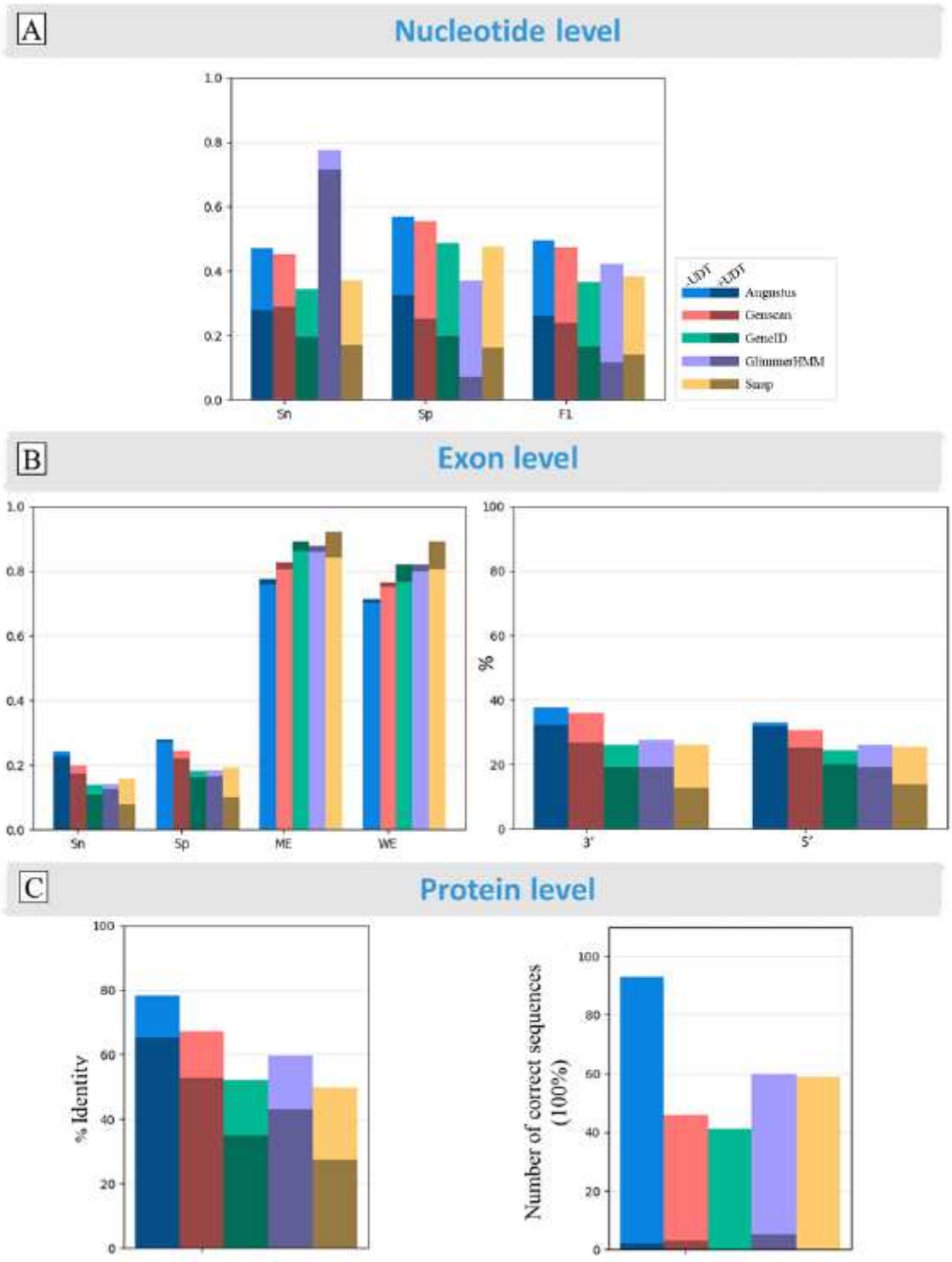

Figure 8

Effect of undetermined sequence regions (UDT) on prediction performance of the five gene prediction programs, using Confirmed benchmark sequences from Metazoa, where 542 sequences have no undetermined regions (-UDT: light colors) and 133 sequences have undetermined regions (+UDT: dark colors). A) sensitivity and specificity of nucleotide prediction. B) sensitivity and specificity of exon prediction. C) accuracy of protein sequence prediction (% identity) and number of proteins correctly

predicted with 100% identity. Sn=sensitivity; Sp=specificity; F1=F1 score; ME=Missing Exons; WE=Wrong Exons; 3’ and 5’ are the percentage of correctly predicted 3’ and 5’ exon boundaries. %Identity indicates the sequence identity observed between the predicted proteins and the Confirmed benchmark sequences.

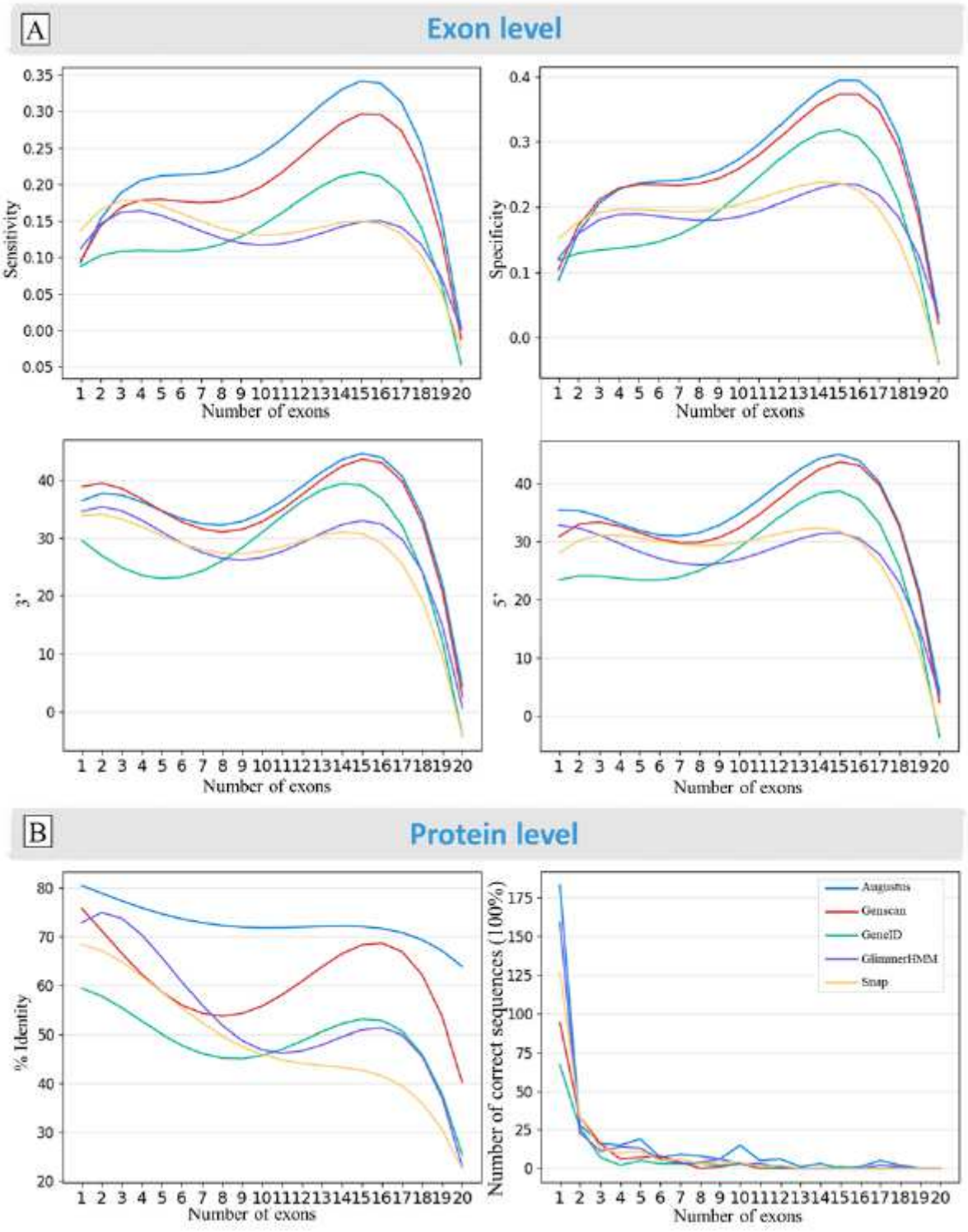

Figure 9

Effect of exon map complexity on prediction quality at the A) exon and B) protein levels. A 4th degree polynomial curve fitting was used to represent the results more clearly. Sequences with 21-24 exons were

not included, due to the low number of sequences in the benchmark with these exon counts. 3’ and 5’ are the proportion of correctly predicted 3’ and 5’ exon boundaries respectively. %Identity indicates the sequence identity observed between the predicted proteins and the Confirmed benchmark sequences.

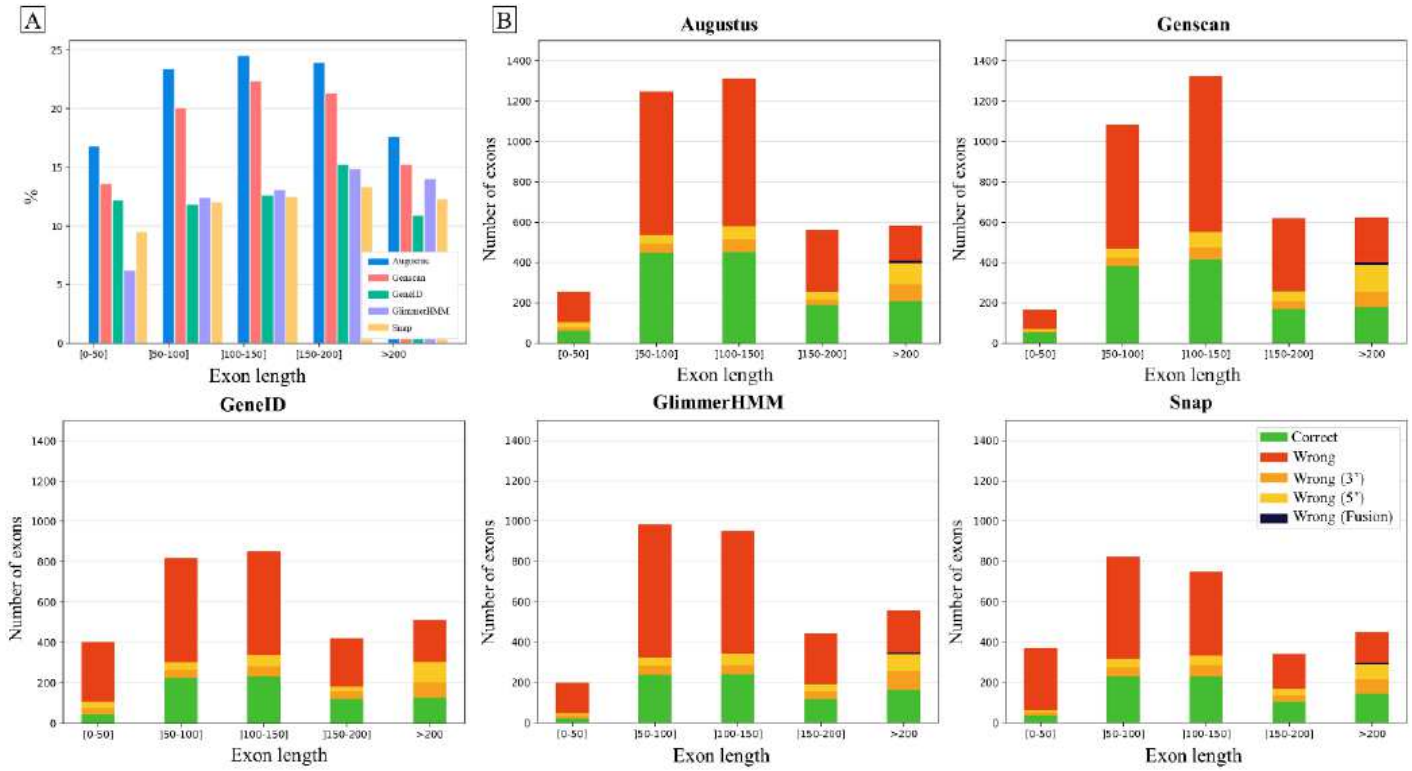

Figure 10

Effect of exon length on exon prediction quality. A) Proportion of benchmark exons correctly predicted depending on the exon length. B) Number of exons predicted correctly, with one of the 5’ or 3’ exon boundaries correct, or with both boundaries wrongly predicted, for each of the five programs.

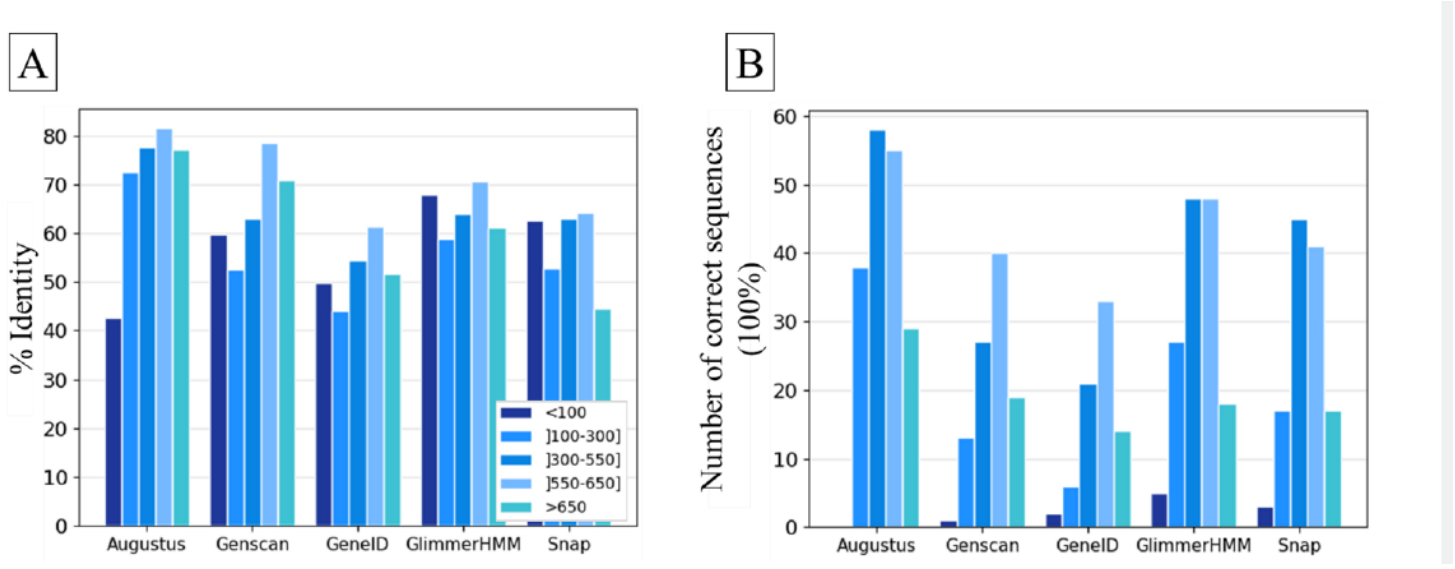

Figure 11

Effect of protein length on prediction accuracy: A) average percent identity between the predicted and the benchmark protein sequences, B) number of proteins perfectly predicted with 100% sequence identity.

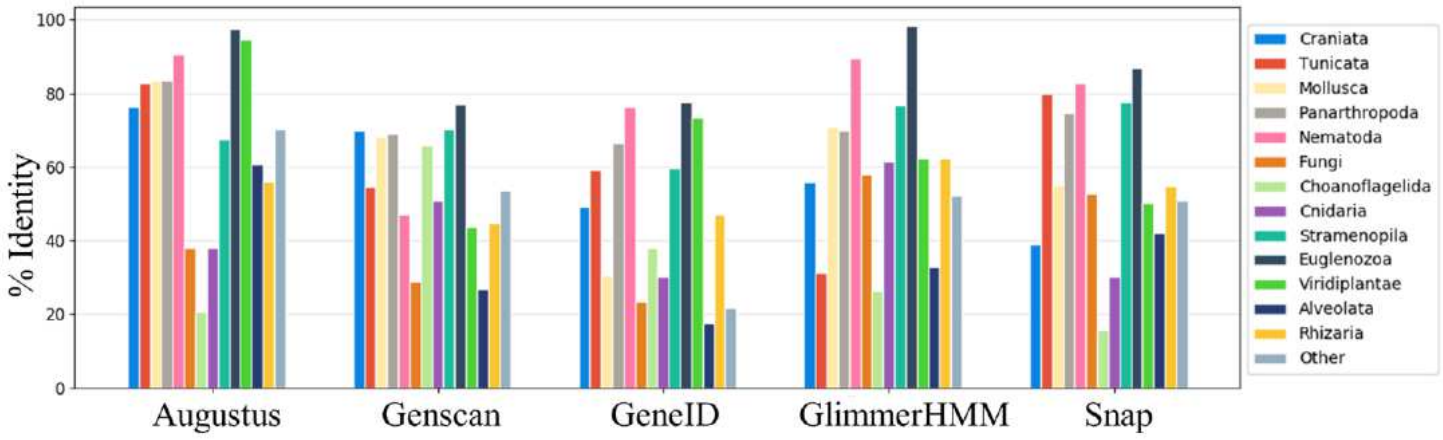

Figure 12

Prediction performance for sequences from different clades. The 'Other' group contains the Apusozoa, Cryptophyta, Diplomonadida, Haptophyceae, Heterolobosea, Parabasalia clades, as well as Placozoa, Annelida and urchin. % Identity indicates the average percent identity between the predicted and the benchmark protein sequences.

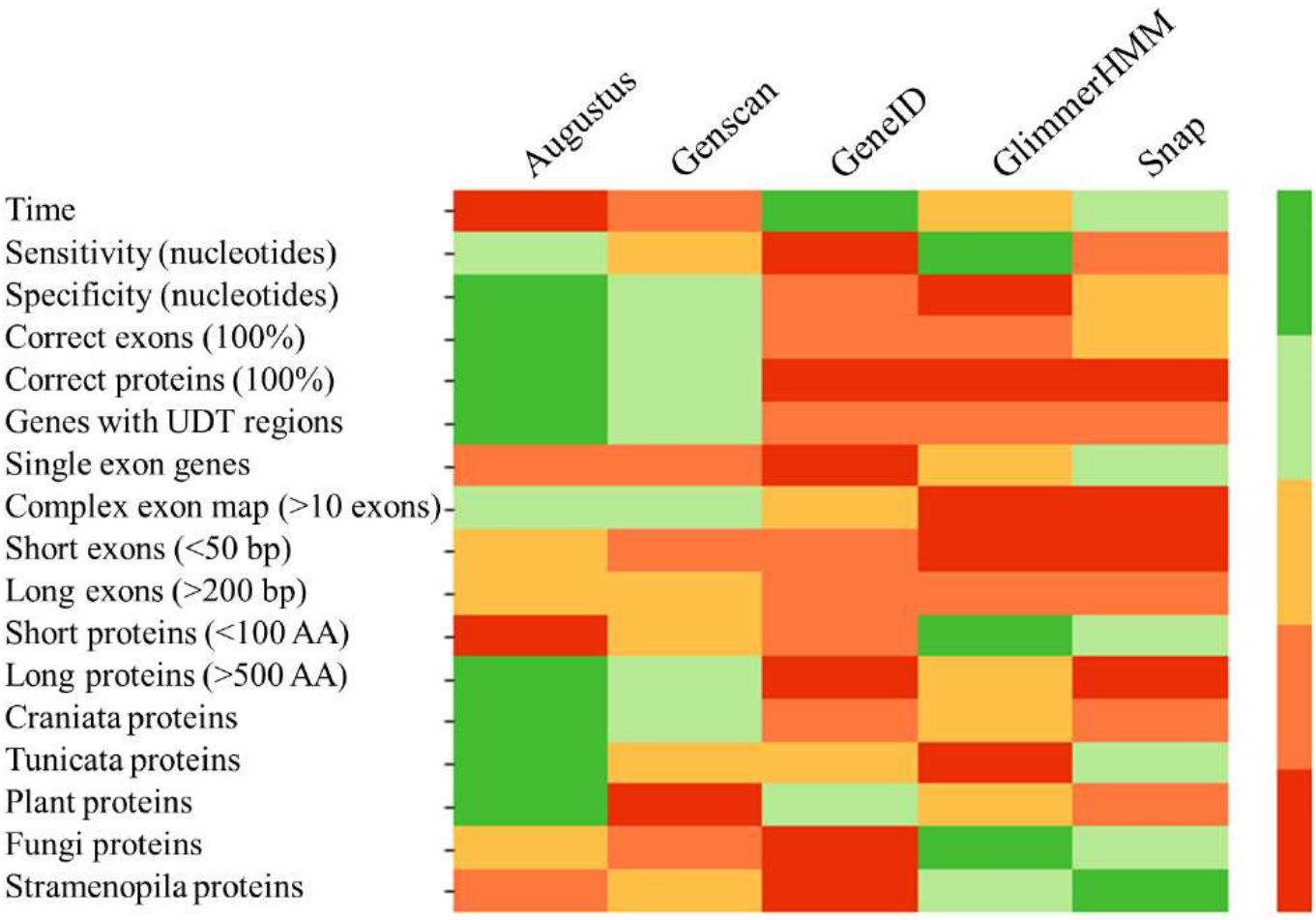

Figure 13

Strengths and weaknesses of the gene prediction programs evaluated in this study. Heatmap colors are: dark green = best program, light green = 2nd best program, yellow = 3rd best program, orange = 4th best program, red = 5th best program.

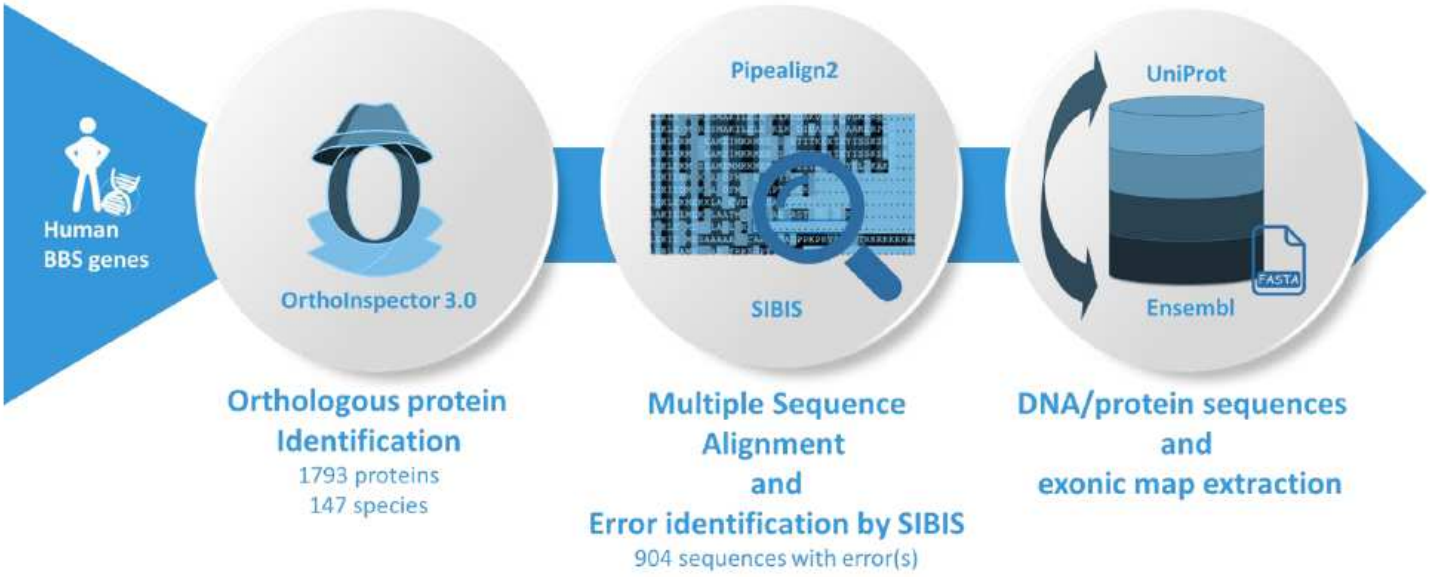

Figure 14

Schematic view of the pipeline used to construct the benchmark.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.pdf