GASCO: GENOME ANNOTATION BY SIMILARITY TO CONSENSUS OF ORTHOLOGS

A Thesis

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

M.S. Plant Breeding

by

Evan Rogers Rees

May 2021

ABSTRACT

We have developed a method for genome annotation based on evolutionary conservation independent of transcriptomic evidence. Newly assembled genomes may be lacking comprehensive mRNA transcription data, and even when such data is available, the transcriptome may be a poor proxy for the proteome. The approach described here leverages orthology information from annotations of multiple related species to construct a pseudo-ancestral consensus protein for a given orthogroup. This consensus is then used as a guide to inform gene prediction using several existing programs. Resulting gene predictions are translated and evaluated based on their similarity to the consensus sequence. Here we discuss development of this pipeline, named GASCO (Genome Annotation by Similarity to Consensus of Orthologs), including strengths and drawbacks of several annotation methods as applied to maize, and directions for further development.

BIOGRAPHICAL SKETCH

Evan Rees grew up outside of Boston, Massachusetts with his sister Jamie, raised by loving parents Janice and Tim. A two-time high school dropout, he went on to earned his GED in 2010 and worked in hospitality for six years between Boston, San Francisco, and London. In 2013, he opted for a career change and enrolled in Middlesex Community College intent on exploring the world of plants and agriculture. After earning his Associate's degree in 2015, he transferred to the University of Massachusetts, Amherst where he completed a Bachelor of Science in Plant and Soil Science in 2017. He started his graduate studies later that year and quickly became engrossed by the command line and bioinformatics programming, which continue to dominate his interests. He spends his free time making music, cooking, caring for plants, breaking and sometimes fixing things, and expanding his programming knowledge.

This thesis is dedicated to my mother, Janice, sister, Jamie, and late father, Tim.

Your love and support lifts me, always.

ACKNOWLEDGEMENTS

Cornell University is located on the traditional homelands of the Gayogo̲hó:nǫ' (the Cayuga Nation). The Gayogo̲hó:nǫ' are members of the Haudenosaunee Confederacy, an alliance of six sovereign Nations with a historic and contemporary presence on this land. The Confederacy precedes the establishment of Cornell University, New York State, and the United States of America. We acknowledge the painful history of Gayogo̲hó:nǫ' dispossession, and honor the ongoing connection of Gayogo̲hó:nǫ' people, past and present, to these lands and waters.

The GASCO source code was initially developed in cooperation with Mohamed El-Walid with additional contributions by Lynn Johnson. Improvements to the consensus algorithm were made on suggestion of Baoxing Song. The target region algorithm and parameters feature improvements suggested by Yaoyao Wu, which were implemented with the help of Zack Miller. Arun Seetharam is credited for the use of Mikado to evaluate gene predictions. Proofreading of this thesis, and moral and logistical support were provided by Sara Miller. Financial support was provided by the SIPS Denison Graduate Fellowship and the USDA-ARS.

Further, many mentors too numerous to name in full here have contributed to the author's personal and professional development. Chris Fiori, Laurie Ranger, David Kalivas, Lauren Maniatis, and Robert Kaulfuss at Middlesex Community College all helped stoke the author's interest in plants and agriculture. Susan Han, Sam Hazen, and Michelle DaCosta at UMass Amherst facilitated immersive experiences in plant science research and offered sound academic and career advice. Finally, Ed Buckler and his research group have provided a stimulating, challenging, and understanding environment and support system in which the author completed this thesis after much tribulation.

TABLE OF CONTENTS

INTRODUCTION

The cost of high-quality NGS data has continued to plummet in recent years. This has contributed to an explosion in the number of publicly available genome assemblies. The ambitious scope of projects such as the Earth BioGenome Project, which over the next decade aims to sequence and assemble the genomes of 1.5 million eukaryotic species, indicates that this trend will only continue to accelerate in coming years[1].

A key complement to any genome assembly is a high-quality annotation that describes salient functional regions in the genome and their hierarchy. A genome annotation should provide accurate models of all the genes and their functional products, particularly proteins. Consequently, numerous methods for annotating genome assemblies have been developed and continue to be improved[2, 3, 4, 5, 6]. The two main modes of gene prediction are *ab initio*, employing a probabilistic model of gene structure, and based on extrinsic evidence such as expressed sequence tags or, more recently, RNA-seq data. Each mode has drawbacks that are mitigated by a combined approach, as is typical of current gold-standard pipelines such as BRAKER[5].

Ab initio gene prediction relies upon a prior probabilistic notion of gene structure including splice sites, start and stop codons, and sequence motifs[7, 2, 6]. While *ab initio* gene prediction is powerful in that a previously unknown genome can be annotated with solely computational resources, it ultimately depends on the relevance and quality of the training set used to produce the model, and the quality of the genome assembly[8]. In practice, purely *ab initio* prediction in eukaryote genomes still requires extensive manual curation, and is almost always combined with extrinsic evidence for validation[5, 6].

In contrast, extrinsically-derived annotation relies upon direct evidence of a gene product to inform annotation. The foremost shortcoming of this approach is that in higher eukaryotes and specifically plants, the transcriptome is an imperfect proxy for the proteome. Eukaryotic transcriptome diversity is extensively driven by alternative splicing (AS), in which a single gene template can produce various transcript isoforms. In plants, the dominant mode of AS is intron retention, while animals primarily display exon-skipping[9]. Transcriptome diversity is marked in plants, and is thought to contribute to environmental adaptation by encoding alternative protein isoforms in response to biotic and abiotic stressors[10]. It is clear, however, that the presence of a transcript isoform is not itself indicative of a corresponding protein product. For example, in maize only around 36% of transcribed genes and 46% of high-abundance transcripts had a detectable protein protein product[11]. While some of this disparity results from lower sensitivity of proteomics, it may also be explained by bioenergetics: in Arabidopsis, the cost of transcription measured as a fraction of total energy budget is 1000-fold lower than that of translation[12]. The consequence is that plant transcriptome datasets are noisy, and may contribute to superfluous annotation of non-functional genes.

One approach to distilling genes with functional products is to take into account protein evidence from related species. Such genes should be under greater selection pressure and their protein products relatively conserved. This conservation can be represented in the form of protein orthology relationships, for which numerous comprehensive datasets with varying granularity have been published[13, 14, 15]. A consensus sequence for a given orthogroup, representing a pseudo-ancestral protein intermediate to each ortholog, may provide a reliable hint for predicting gene structure in unannotated genomes. We have implemented such an approach, titled Genome Annotation by Similarity to Consensus of Orthologs (GASCO), and

applied it to annotate version 5 of the maize B73 genome.

In the context of GASCO, we explored three existing gene-prediction tools: AUGUSTUS[6], exonerate[16], and GenomeThreader[17]. AUGUSTUS employs a Hidden Markov Model trained on known gene structures, and has also been extended to incorporate block profiles derived from multiple sequence alignments (MSAs) which model highly conserved regions in order to improve prediction accuracy[6, 18]. Exonerate is an alignment program used in the BRAKER annotation program to detect splice sites, and can efficiently provide optimal spliced alignments for protein sequences to a target nucleotide database[16]. GenomeThreader is a similarly efficient splice-aware aligner, but incorporates one of several pretrained splice site models[17]. Due to the sensitivity of each approach and the large search space constituted by many plant genomes, it is computationally prohibitive to do a direct exhaustive search with any of these methods. We therefore borrowed the approach used in BUSCO[8] and BRAKER[5] to reduce the search space by identifying candidate regions using BLAST[19], and then applied the gene prediction software to each region.

Genome annotations even for highly-studied model organisms are being constantly revised as annotation methods are developed, and new assemblies, molecular data, and other sources of evidence become available. For example, the annotation associated with version 3 of the maize B73 reference genome was revised to include or exclude genes on four occasions between 2013 and 2014[20]. Here we explore the idea of leveraging protein orthology to inform gene prediction, circumventing noise in the transcriptome, and provide an additional prong of evidence to take into account in genome annotation and modeling of likely functional proteins.

METHODS

The starting point for GASCO was a set of protein MSAs obtained from the eggNOG database[13]. The eggNOG database is constructed from the annotated genomes of 5090 core organisms including 477 representative eukaryotes sourced from Ensembl. The component annotations are generated by numerous methods, generally relying on transcriptomics from model species, and orthogroups are defined at each taxonomic level based on Smith-Waterman reciprocal best hits[13]. All MSAs for orthogroups defined within in the Poales order were downloaded from eggNOG and processed using a custom Python script to generate protein consensus sequences and block profiles. The consensus algorithm is detailed in **Figure 2.1**. The consensus residue for each position was determined by the maximum BLOSUM score of all possible residues at that position given the alignment. Positions with gap frequency $\geq$0.5 were then removed to produce a consensus protein sequence written to FASTA format for use with GASCO. AUGUSTUS-PPX[18] mode also required a protein block profile as input, representing probabilities of each residue at each position. Block profiles were generated from the same residue counts matrix as the consensus algorithm, again excluding columns with gap frequency $\geq$0.5. A pseudocount value of 0.1 was added to all counts, and the matrix was normalized over column sums to produce a probability matrix representing the block profile.

The GASCO pipeline is outlined in **Figure 2.2**. GASCO consists of three broad steps: 1) target region identification; 2) gene structure prediction; 3) filtering of gene predictions. Identification of target regions was carried out using TBLASTN [19] with parameters shown in **Table 2.2**, using the orthogroup consensus sequences as queries and the genome as target. The user-settable parameters

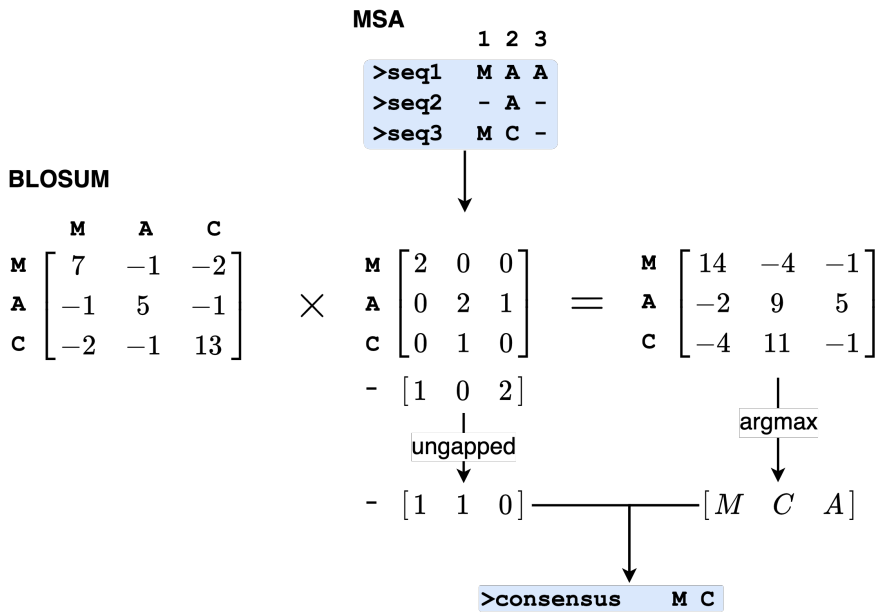


Figure 2.1: Detail of MSA consensus algorithm. Gapped alignments are converted to a matrix of residue counts by position. The dot product of a modified BLOSUM matrix and the residue counts matrix yields a weighted score of residue by position. The alignment consensus is the highest scoring residue for each column. Gapped columns are then masked to produce a consensus sequence.

mergeWithin, *discardBelow*, and *expandBy* were applied to process TBLASTN hits into target regions. Hits for each orthogroup were merged if the distance between the end coordinate of the upstream hit and start coordinate of the downstream hit was less than *mergeWithin*, with a default value of 20,000bp. Merged hits were then filtered to exclude those with query identity below *discardBelow* with a default value of 0.5, and up to 25 regions with the highest query identity were retained. Finally, regions were expanded up and downstream by up to *expandBy* or the end of the target sequence, with a default value of 5,000bp.

Target regions were then searched with AUGUSTUS, exonerate, or GenomeThreader, with parameters from **Table 2.2**, again using the orthogroup consensus (exonerate and GenomeThreader) or orthogroup profile (AUGUSTUS) to query the target region. AUGUSTUS and GenomeThreader each require a 'species' argument, which

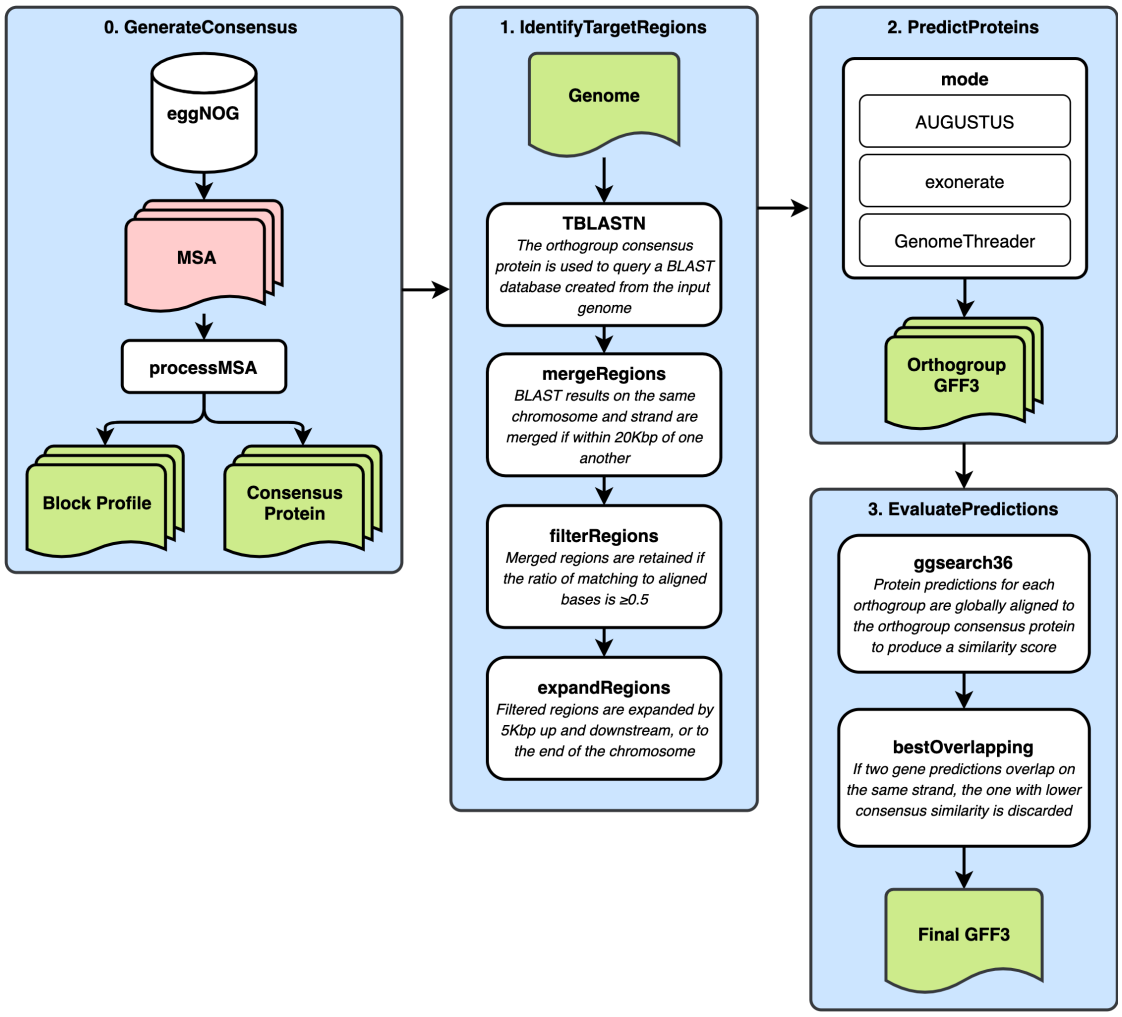


Figure 2.2: Overview of major steps in GASCO pipeline. The MSA consensus algorithm in step 0.GenerateConsensus is detailed in Figure 2.1.

uses a pre-trained model to aid in detection of splice junctions and exon structure. To avoid biasing predictions to the maize reference genome, rice was used in both cases as a common outgroup. When using AUGUSTUS it was necessary to search each target region individually due to program constraints, while exonerate and GenomeThreader allowed all target regions for a given orthogroup to be searched simultaneously.

Filtering of gene predictions was carried out in two parts. First, predictions for each orthogroup were translated and aligned back to the orthogroup consensus

sequence using ggsearch36 from the FASTA36[21] package with parameters shown in **Table 2.2**. This produced a global alignment and consensus similarity score for each predicted protein. BLASTP was also used to produce local alignments of predicted proteins against consensus sequences, but these were not considered during filtering. Next, predictions for all orthogroups were grouped by chromosome and strand, and checked for overlaps. If two predictions on the same strand overlapped, the one with lower consensus similarity was discarded.

Version 5 of the maize B73 reference genome [22] was selected for benchmarking GASCO due to the high quality of the annotation and assembly itself. GASCO was run in each mode (AUGUSTUS, exonerate, GenomeThreader) against the genome target with a query of 31,003 orthogroups from the eggNOG Poales database. GenomeThreader was also run independently of GASCO in standalone mode to serve as a benchmark. The results from each run were written to a GFF3 file used for further analysis. This GFF was then compared to the B73 v5 reference annotation using Mikado[23], with parameters shown in **Table 2.2**. Categories assigned to each prediction by Mikado are summarized in **Table 2.1**.

Category	**Description**
Match	Concordance between two transcript models
Extension	One model extends the intron chain of the other
Alternative	Exon chains of both models overlap, but differ in significant ways
Intronic	One model is completely contained in the intron chains of the other
Overlap	The exon sequences of the two models overlap
Fragment	The prediction is a fragment of the reference, usually on the opposite strand
Fusion	The prediction is a fusion between two or more reference models
Unmatched	The prediction has no close match in the reference

Table 2.1: Description of Mikado class codes. Reproduced from Mikado documentation[23].

GASCO was implemented primarily in Kotlin, with helper scripts in Bash and Python, and deployed as a Docker image for cross-platform compatibility. Analyses described here were run on CentOS 7.6, and GASCO was additionally

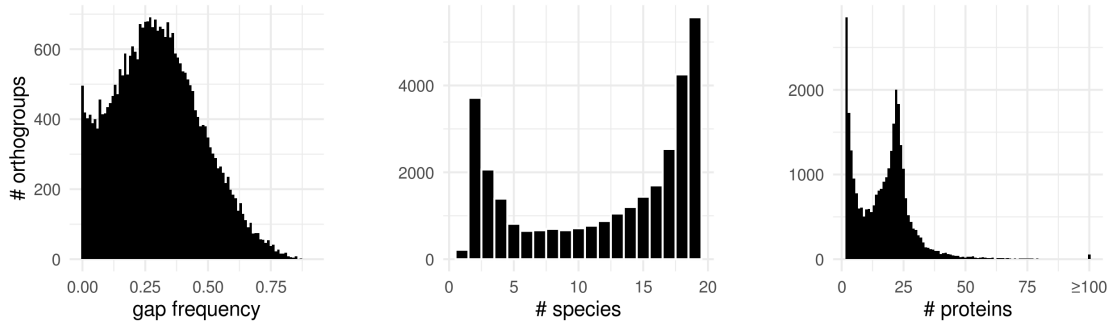
command	**parameters**
AUGUSTUS	–genemodel=complete –gff3=on –UTR=off –introns=on –species=rice
exonerate	–model protein2genome –refine full –showalignment yes –showvulgar yes –showtargetgff yes
GenomeThreader	-species rice -gff3out -skipalignmentout
tblastn	-outfmt 6 qseqid sseqid pident length mismatch gapopen qstart qend sstart send evalue bitscore qlen slen nident positive
ggsearch36	-m8CBs -p -d 0
mikado	compare –protein-coding

Table 2.2: Command line parameters for programs used by GASCO.

tested on macOS 11. The source code has been made publicly available at https://bitbucket.org/bucklerlab/gasco.

RESULTS

Protein consensus sequences and block profiles were derived from eggNOG MSAs and used to annotate the maize B73 v5 reference genome. At the Poales level, the eggNOG database contains 31,003 orthogroups constructed from genome annotations of 19 species. Species representation within each orthogroup varied, with 3,932 orthogroups containing proteins from just one or two species (**Figure 3.1b**). The smallest orthogroups contained as few as 2 and as many as 100 proteins, with the distribution centered at roughly 24 proteins (**Figure 3.1c**). Orthogroups with fewer proteins and/or representing fewer species contained less evidence on which to base the consensus protein, which negatively impacts the reliability of GASCO predictions. The alignment gap frequency of eggNOG MSAs was also visualized as an indicator of potentially noisy alignments that could negatively impact consensus quality (**Figure 3.1a**).



(a) Gap frequency of eggNOG MSAs.

(b) Distribution of species representation by eggNOG orthogroup.

(c) Distribution of eggNOG orthogroup size.

Figure 3.1: Summary of eggNOG orthogroups and multiple sequence alignments.

The number of proteins and orthogroups predicted using each method was tabulated and is shown in **Table 3.1**. AUGUSTUS and exonerate each produced a large number of raw annotations, but after filtering overlapping genes this was reduced to 28,789 and 34,065 proteins representing 16,683 and 18,420 orthogroups,

respectively. GenomeThreader produced far fewer raw annotations, resulting in only 22,654 proteins and 13,570 orthogroups after filtering. In standalone mode, GenomeThreader produced a similar number of genes and orthogroups as when run within GASCO after filtering. By comparison, the eggNOG MSAs contained 25,207 maize proteins from 18,695 orthogroups.

mode	total genes	final genes	orthogroup count
AUGUSTUS	272,769	28,739	16,683
exonerate	133,131	34,065	18,420
GenomeThreader	32,144	22,654	13,570
GenomeThreader*	26,980	23,703	14,459

Table 3.1: GASCO results for each method. An asterisk (*) denotes the program was run standalone, outside of GASCO.

Figure 3.2 displays the consensus similarity of GASCO proteins versus that of the maize proteins in the eggNOG database. Proteins were matched based on ordered consensus similarity. Points along the diagonal indicated agreement of GASCO predictions with the existing reference, while points above or below the diagonal indicated out-performance of one model over the other. While predictions from all three methods clustered along the diagonal to the upper right, the predictions from AUGUSTUS and GenomeThreader showed distinct groupings of low-similarity proteins below the diagonal. This was also reflected in the results from GenomeThreader when run standalone, although to a lesser extent than when run within GASCO. The most similar matches from each mode, shown on the y-axis of the top row of **Figure 3.2**, were also visualized in **Figure 3.3**. The clusters of off-diagonal proteins in Figure 3.2 were much less pronounced in this view, particularly for standalone GenomeThreader.

Comparison with the B73 v5 reference annotation indicated that the low-similarity clusters were largely from proteins with no direct correspondence to

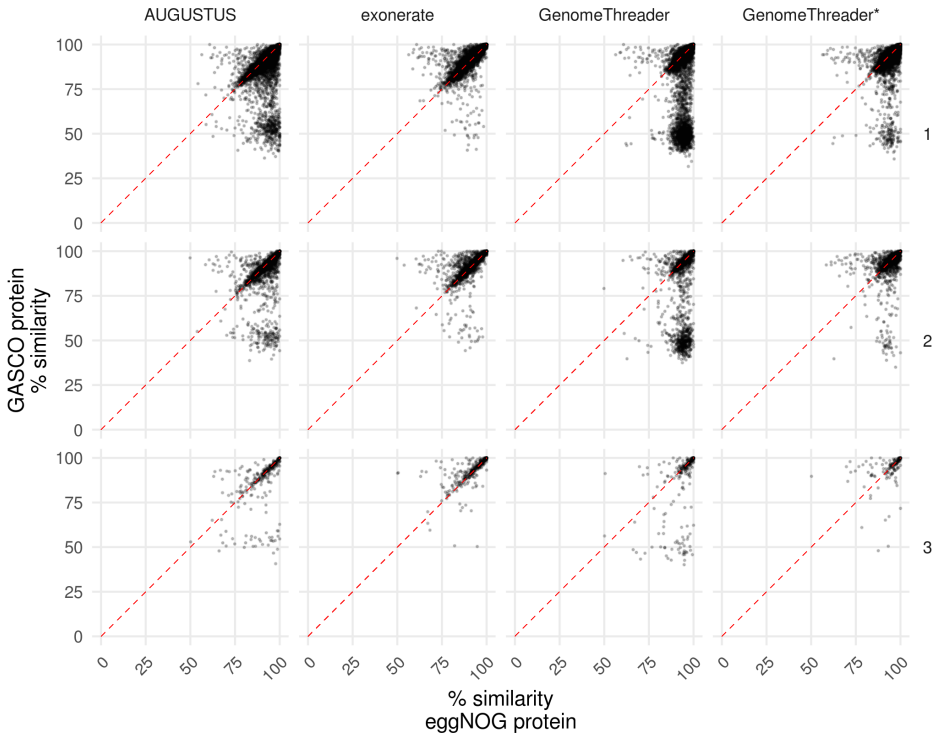


Figure 3.2: Consensus similarity of GASCO proteins vs eggNOG proteins, given as a percentage. Comparisons in each column are made according to ranked consensus similarity of each protein. An asterisk (*) denotes the program was run standalone, outside of GASCO.

the reference annotation or orthogroups with less evidence (**Figures 3.5a-3.5b**). The predominant prediction classes were direct matches, followed by alternative isoforms, intron chain extensions, and finally unmatched genes. The breakdown of each class was similar between exonerate, GenomeThreader, and GenomeThreader standalone, while AUGUSTUS yielded a greater proportion of alternative isoforms. The comparisons also suggested a substantial disagreement in exon boundaries between GASCO results and the reference annotation, warranting further exploration of the disjoint intervals.

The full GASCO pipeline took between 10 and 15 hours to complete depending on the mode (**Figure 3.6**). Most of the wall time (60-95%) was due to the initial TBLASTN search, with the actual gene prediction taking up most of the remainder.

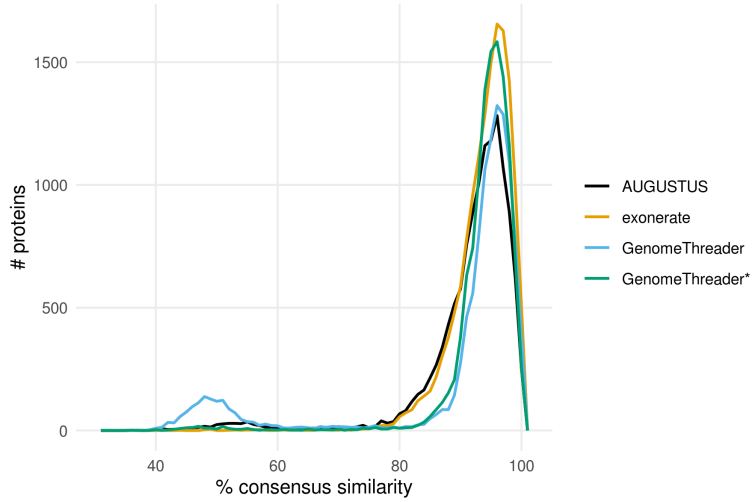


Figure 3.3: Consensus similarity distribution for each GASCO mode. For each orthogroup, the protein with highest consensus similarity is shown. This corresponds to the y-axis of the first row in Figure 3.2. An asterisk (*) denotes the program was run standalone, outside of GASCO.

In the GenomeThreader mode, wall time attributed to the prediction step was negligible, although this was not true when it was run standalone. RAM usage did not exceed 8GB for AUGUSTUS, and 4GB for exonerate and GenomeThreader.

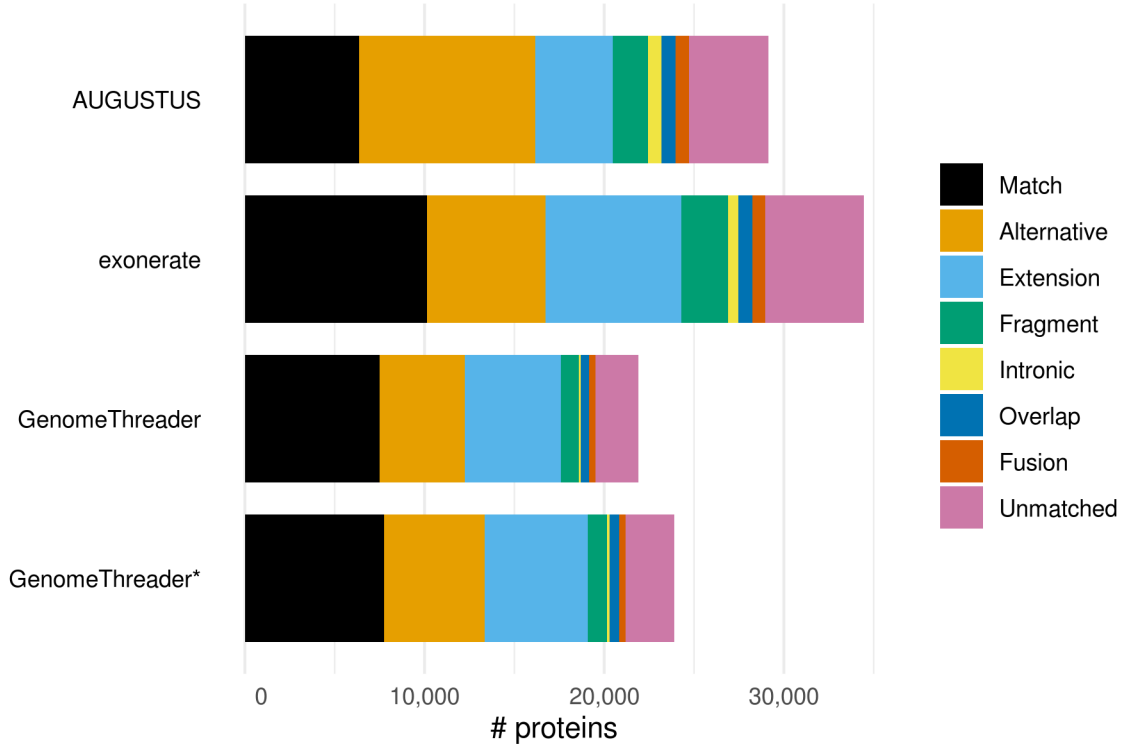
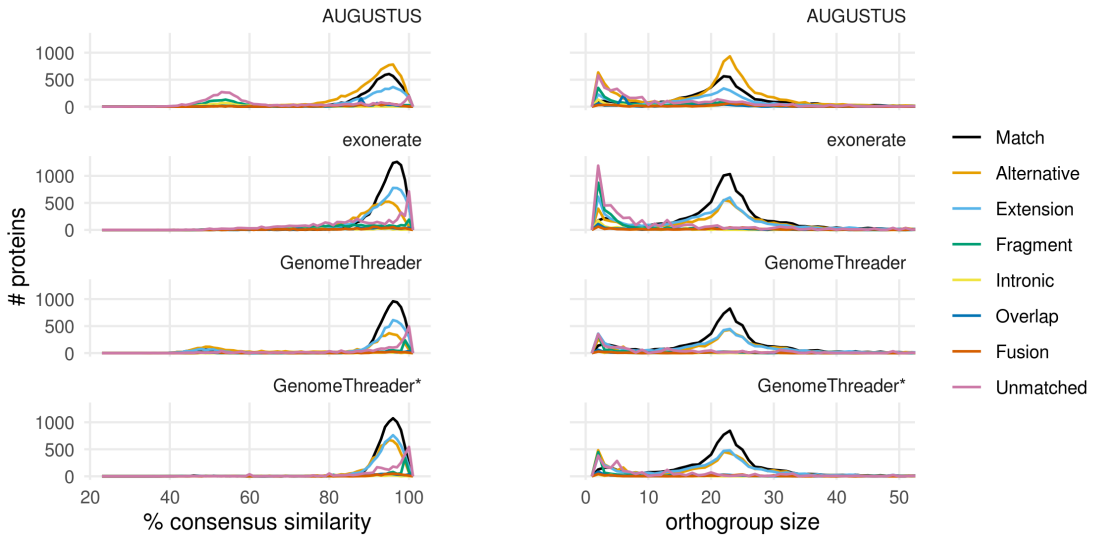


Figure 3.4: Breakdown of Mikado classification of GASCO predictions for each mode and GenomeThreader standalone. Categories are summarized in Table 2.1.



(a) Consensus similarity of GASCO results for each mode categorized by correspondence to B73 v5 annotation. Categories are summarized in Table 2.1.

(b) Orthogroup size of GASCO results for each mode categorized by correspondence to B73 v5 annotation. Categories are summarized in Table 2.1.

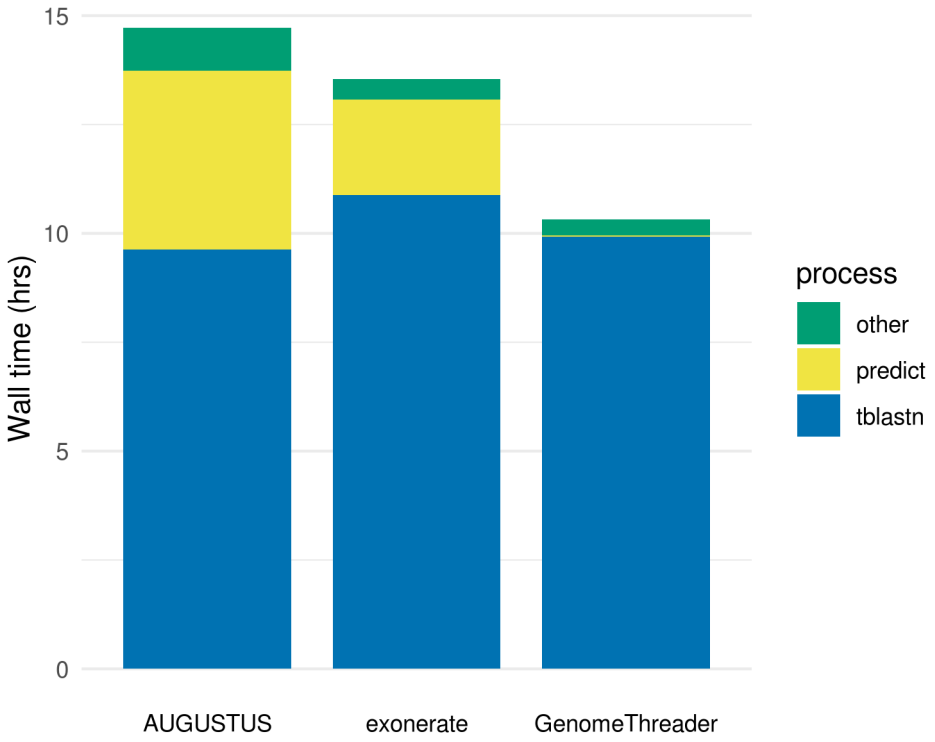


Figure 3.6: Wall time of GASCO in each mode broken down by process.

DISCUSSION

We implemented GASCO, a pipeline for genome annotation based on orthogroup consensus proteins, and compared the results of several component gene prediction programs when run against the maize B73 reference genome. Exonerate generally outperformed AUGUSTUS and GenomeThreader in terms of both the number and proportion of matches to the reference annotation (**Figure 3.5a**, **Table 3.1**). Furthermore, the vast majority of predictions in each mode were above 80% consensus similarity, indicating high fidelity to the protein consensus sequence used to inform gene prediction.

Despite generally promising results, the cases where GASCO and the reference annotation differ require further explanation. One way to investigate the disagreements would be to examine evolutionary conservation at the genomic level in disjoint intervals using a metric such as genome evolutionary rate profiling (GERP)[24]. For a given disjoint interval, greater conservation relative to the common intervals would lend credibility to the containing model and vice-versa. GERP is also a convenient metric as it is derived solely from genomic alignments, and is therefore not biased towards any particular annotation method.

The use of orthogroup consensus sequences was effective at predicting many genes in a manner consistent with the reference annotation. In cases of disagreement, it was likely that the consensus was derived from a lower quality MSA (**Figure 3.5b**). This suggests that further refinement of the input MSAs would be beneficial in producing more accurate consensus sequences and improving the performance of GASCO. Orthogroup size and species representation are evidently key factors, as they correspond to the breadth of protein diversity necessary for an accurate consensus. To supplement data available from eggNOG, additional

orthology databases such as OMA[14] and OrthoDB[15] should be explored.

Another area for improvement is the method by which target regions are identified (**Figure 2.2**). The current algorithm is extremely permissive when merging TBLASTN results in order to capture divergent genes and accommodate large introns, and regions are generously expanded by up to 5,000bp up and downstream. This may cause problems in the case of tandem duplicates, which are likely to be merged into a single region which may interfere with the ability of the gene prediction software to discriminate between copies, particularly since they will share the same consensus sequence. An alternate algorithm has been proposed in which TBLASTN results are first filtered based on E-value, then merged taking into account the distance between intervals on both the target and the query sequence, and finally filtered based on query coverage. The associated parameters *maxEvalue*, *maxTargetGap*, *maxTargetOverlap*, *maxQueryGap*, *maxQueryOverlap*, and *minQueryCoverage* allow for fine-tuning based on genome characteristics. It is intended that this will improve prediction of tandem duplicates, as well as improve overall performance by reducing the quantity and size of regions to be searched, although this will have to be balanced with sensitivity. While this algorithm has been designed and a test implementation has been written, it has not yet been incorporated into the main GASCO pipeline pending further testing.

The overall resource usage of GASCO was moderate, and dominated by the initial TBLASTN search. RAM usage was insignificant at less than 8GB, while the CPU requirement was between 1,000 and 1,500 CPU hours using the parameters described previously. The wall time and CPU usage may be reduced by finding an alternate sequence search algorithm to TBLASTN. One promising alternative is DIAMOND, which is reportedly 2-4 orders of magnitude faster than equivalent

BLAST algorithms[25]. DIAMOND does not currently support searching a translated nucleotide database with a protein query, but an examination of the source code indicates that this functionality may be forthcoming.

Overall, GASCO was shown to be an effective annotation method. It can be further applied to systematically annotate genomes given a set of high quality reference proteins. Moreover, GASCO can expedite annotation of newly assembled plant genomes and re-annotation of existing ones while ensuring consistency in its results, minimizing or eliminating transcriptomics bias. As a complement to the existing arsenal of genomics research, GASCO allows researchers to leverage the torrent of genomic data to fuel life science discoveries.

BIBLIOGRAPHY

[1] Harris A. Lewin, Gene E. Robinson, W. John Kress, William J. Baker, Jonathan Coddington, Keith A. Crandall, Richard Durbin, Scott V. Edwards, Félix Forest, M. Thomas P. Gilbert, Melissa M. Goldstein, Igor V. Grigoriev, Kevin J. Hackett, David Haussler, Erich D. Jarvis, Warren E. Johnson, Aristides Patrinos, Stephen Richards, Juan Carlos Castilla-Rubio, Marie-Anne van Sluys, Pamela S. Soltis, Xun Xu, Huanming Yang, and Guojie Zhang. Earth BioGenome Project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences*, 115(17):4325–4333, April 2018.

[2] John Besemer and Mark Borodovsky. GeneMark: Web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Research*, 33(Web Server issue):W451–W454, July 2005.

[3] Tomáš Brůna, Katharina J. Hoff, Alexandre Lomsadze, Mario Stanke, and Mark Borodovsky. BRAKER2: Automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genomics and Bioinformatics*, 3(1), January 2021.

[4] Brandi L. Cantarel, Ian Korf, Sofia M.C. Robb, Genis Parra, Eric Ross, Barry Moore, Carson Holt, Alejandro Sánchez Alvarado, and Mark Yandell. MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research*, 18(1):188–196, January 2008.

[5] Katharina J. Hoff, Simone Lange, Alexandre Lomsadze, Mark Borodovsky, and Mario Stanke. BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics*, 32(5):767–769, March 2016.

[6] Mario Stanke, Oliver Keller, Irfan Gunduz, Alec Hayes, Stephan Waack, and Burkhard Morgenstern. AUGUSTUS: Ab initio prediction of alternative transcripts. *Nucleic Acids Research*, 34(suppl_2):W435–W439, July 2006.

[7] Steven L. Salzberg, Mihaela Pertea, Arthur L. Delcher, Malcolm J. Gardner, and Hervé Tettelin. Interpolated Markov Models for Eukaryotic Gene Finding. *Genomics*, 59(1):24–31, July 1999.

[8] Mathieu Seppey, Mosè Manni, and Evgeny M. Zdobnov. BUSCO: Assessing Genome Assembly and Annotation Completeness. In Martin Kollmar, editor, *Gene Prediction: Methods and Protocols*, Methods in Molecular Biology, pages 227–245. Springer, New York, NY, 2019.

[9] Anireddy S.N. Reddy, Yamile Marquez, Maria Kalyna, and Andrea Barta. Complexity of the Alternative Splicing Landscape in Plants. *The Plant Cell*, 25(10):3657–3683, October 2013.

[10] Qiuyue Chen, Yingjia Han, Haijun Liu, Xufeng Wang, Jiamin Sun, Binghao Zhao, Weiya Li, Jinge Tian, Yameng Liang, Jianbing Yan, Xiaohong Yang, and Feng Tian. Genome-Wide Association Analyses Reveal the Importance of Alternative Splicing in Diversifying Gene Function and Regulating Phenotypic Variation in Maize. *The Plant Cell*, 30(7):1404–1423, July 2018.

[11] Justin W. Walley, Ryan C. Sartor, Zhouxin Shen, Robert J. Schmitz, Kevin J. Wu, Mark A. Urich, Joseph R. Nery, Laurie G. Smith, James C. Schnable, Joseph R. Ecker, and Steven P. Briggs. Integration of omic networks in a developmental atlas of maize. *Science (New York, N.Y.)*, 353(6301):814–818, August 2016.

[12] Michael Lynch and Georgi K. Marinov. The bioenergetic costs of a gene. *Proceedings of the National Academy of Sciences*, 112(51):15690–15695, December 2015.

[13] Jaime Huerta-Cepas, Damian Szklarczyk, Davide Heller, Ana Hernández-Plaza, Sofia K Forslund, Helen Cook, Daniel R Mende, Ivica Letunic, Thomas Rattei, Lars J Jensen, Christian von Mering, and Peer Bork. eggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research*, 47(D1):D309–D314, January 2019.

[14] Adrian M Altenhoff, Clément-Marie Train, Kimberly J Gilbert, Ishita Mediratta, Tarcisio Mendes de Farias, David Moi, Yannis Nevers, Hale-Seda Radoykova, Victor Rossier, Alex Warwick Vesztrocy, Natasha M Glover, and Christophe Dessimoz. OMA orthology in 2021: Website overhaul, conserved isoforms, ancestral gene order and more. *Nucleic Acids Research*, 49(D1):D373–D379, January 2021.

[15] Evgenia V Kriventseva, Dmitry Kuznetsov, Fredrik Tegenfeldt, Mosè Manni, Renata Dias, Felipe A Simão, and Evgeny M Zdobnov. OrthoDB v10: Sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Research*, 47(D1):D807–D811, January 2019.

[16] Guy St C. Slater and Ewan Birney. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, 6(1):31, February 2005.

[17] Gordon Gremme, Volker Brendel, Michael E. Sparks, and Stefan Kurtz. Engineering a software tool for gene structure prediction in higher organisms. *Information and Software Technology*, 47(15):965–978, December 2005.

[18] Oliver Keller, Martin Kollmar, Mario Stanke, and Stephan Waack. A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics*, 27(6):757–763, March 2011.

[19] Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, October 1990.

[20] John L Portwood, II, Margaret R Woodhouse, Ethalinda K Cannon, Jack M Gardiner, Lisa C Harper, Mary L Schaeffer, Jesse R Walsh, Taner Z Sen, Kyoung Tak Cho, David A Schott, Bremen L Braun, Miranda Dietze, Brittney Dunfee, Christine G Elsik, Nancy Manchanda, Ed Coe, Marty Sachs, Philip Stinard, Josh Tolbert, Shane Zimmerman, and Carson M Andorf. MaizeGDB 2018: The maize multi-genome genetics and genomics database. *Nucleic Acids Research*, 47(D1):D1146–D1154, January 2019.

[21] William R. Pearson. Finding Protein and Nucleotide Similarities with FASTA. *Current protocols in bioinformatics*, 53:3.9.1–3.925, March 2016.

[22] Matthew B. Hufford, Arun S. Seetharam, Margaret R. Woodhouse, Kapeel M. Chougule, Shujun Ou, Jianing Liu, William A. Ricci, Tingting Guo, Andrew Olson, Yinjie Qiu, Rafael Della Coletta, Silas Tittes, Asher I. Hudson, Alexandre P. Marand, Sharon Wei, Zhenyuan Lu, Bo Wang, Marcela K. Tello-Ruiz, Rebecca D. Piri, Na Wang, Dong won Kim, Yibing Zeng, Christine H. O'Connor, Xianran Li, Amanda M. Gilbert, Erin Baggs, Ksenia V. Krasileva, John L. Portwood, Ethalinda K. S. Cannon, Carson M. Andorf, Nancy Manchanda, Samantha J. Snodgrass, David E. Hufnagel, Qiuhan Jiang, Sarah Pedersen, Michael L. Syring, David A. Kudrna, Victor Llaca, Kevin Fengler, Robert J. Schmitz, Jeffrey Ross-Ibarra, Jianming Yu, Jonathan I. Gent, Candice N. Hirsch, Doreen Ware, and R. Kelly Dawe. De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. *bioRxiv*, page 2021.01.14.426684, January 2021.

[23] Luca Venturini, Shabhonam Caim, Gemy George Kaithakottil, Daniel Lee Mapleson, and David Swarbreck. Leveraging multiple transcriptome assembly methods for improved gene structure annotation. *GigaScience*, 7(giy093), August 2018.

[24] Gregory M. Cooper, Eric A. Stone, George Asimenos, Eric D. Green, Ser-

afim Batzoglou, and Arend Sidow. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Research*, 15(7):901–913, July 2005.

[25] Benjamin Buchfink, Klaus Reuter, and Hajk-Georg Drost. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods*, 18(4):366–368, April 2021.