RESEARCH

Gene Ontology Meta Annotator for Plants (GOMAP)

Kokulapalan Wimalanathan[1,2†^] and Carolyn J. Lawrence-Dill[1,2,3*†]

Abstract

Annotating gene structures and functions to genome assemblies is necessary to make assembly resources useful for biological inference. Gene Ontology (GO) term assignment is the most used functional annotation system, and new methods for GO assignment have improved the quality of GO-based function predictions. The Gene Ontology Meta Annotator for Plants (GOMAP) is an optimized, high-throughput, and reproducible pipeline for genome-scale GO annotation of plants. We containerized GOMAP to increase portability and reproducibility and also optimized its performance for HPC environments. Here we report on the pipeline's availability and performance for annotating large, repetitive plant genomes and describe how GOMAP was used to annotate multiple maize genomes as a test case. Assessment shows that GOMAP expands and improves the number of genes annotated and annotations assigned per gene as well as the quality (based on F_{max}) of GO assignments in maize. GOMAP has been deployed to annotate other species including wheat, rice, barley, cotton, and soy. Instructions and access to the GOMAP Singularity container are freely available online at https://bioinformapping.com/gomap/. A list of annotated genomes and links to data is maintained at https://dill-picl.org/projects/gomap/.

Keywords: functional annotation; gene ontology; CAFA; plant genomes

Background

Plant genomes are notably repetitive and hard to assemble. As such, long-read sequencing technologies have been quickly and widely adopted [1, 2] to enable high-quality *de novo* assembly of plant genomes. The number of plant long-read, whole-genome sequencing datasets rapidly increasing (See figure 1) and would lead to increased number of high-quality plant genome assemblies in near future. In order to make the best use of high-quality assemblies for functional genomics applications, improved computational tools for gene structure and function prediction must also be developed and adopted.

In 1998, the Gene Ontology (GO) consortium released the first common vocabulary describing gene function across species, thus enabling a genome-wide and comparative approach to functional genomics [3]. GO is divided into three categories or sub-ontologies, namely cellular component (CC), molecular function

*Correspondence: triffid@iastate.edu

[1]Bioinformatics and Computational Biology Program, , Iowa State University, Ames, 50010 Iowa, USA

Full list of author information is available at the end of the article

†Co-corresponding authors kokul@bioinformapping.com

^Present Address: Greenlight Biosciences Inc., Medford, 02155, Massachusetts, USA

(MF) and biological process (BP). Various tools and approaches were developed to assign GO terms to genes, and a raft of statistical methods to interpret high-throughput experimental results for GO-based gene function implications were developed and released [4–8]. More recently, the Critical Assessment of protein Function Annotation (CAFA) competition has enticed research groups to develop tools that improve the accuracy and coverage of gene function prediction [5, 7, 9]. Unfortunately, methodologies developed through CAFA have not been widely adopted for annotating plant genomes, and existing plant-specific GO annotation pipelines mainly focus on subsets of GO terms rather than the full set of terms available [10, 11].

We sought to assess the performance of some of the best-performing methods produced through CAFA1 for assigning gene function to plant genomes and to produce an improved functional annotation dataset for maize. These efforts were successful, with improvements to prediction outcomes measured in terms of precision, recall, and coverage [12]. Obvious next steps were to generalize the developed pipeline and to apply it to other maize lines and additional plant species then to evaluate its performance for annotating gene function to those genomes. Here we present GOMAP (Gene Ontology Meta Annotator for Plants) pipeline that generalizes the methods used to produce the maize-GAMER datasets, with improvements to computational performance, reproducibility, and portability. We also present the utility of GOMAP by annotating genomes assemblies of four maize inbred lines, namely B73 RefGen_v4 (VB73v4), W22, PH207 and Mo17 [1, 2, 13, 14]. The GOMAP annotations are compared to the community annotations for B73v4 and PH207. Gramene annotated B73v4 RefGen_v4 using Ensembl Compara method and produced a high-confidence and high-coverage GO annotations [15]. The PH207 dataset was obtained from the supplemental tables of the PH207 genome sequencing paper [13]. GO terms for PH207 genes were annotated using InterProScan v5.0 that uses domain presence to assign GO terms to input protein sequences[13]. We compared GOMAP to community methods used to annotate B73v4 and the PH207 to illustrate the differences among datasets produced by three approaches for GO annotation.

Materials and Methods

Overview of the Annotation of input sequences

We selected the longest translated protein sequence for each gene model to represent each gene in the assembly. From there, the GOMAP system was used to apply sequence-similarity, domain-presence and mixed-method pipelines to annotate GO terms to the predicted protein set to produce a single unique and non-redundant GOMAP aggregate dataset as the result (See Figure 2). Sequence similarity searches are performed against two plant datasets: Arabidopsis and UniProt. The Arabidopsis dataset contains protein sequences downloaded from TAIR and curated GO annotations [16]. The UniProt dataset contains protein sequences from the top plants species that were ranked by number of curated GO annotations available in UniProt [17]. The first set of annotations is generated using BLAST to obtain reciprocal-best-hits between input and Arabidopsis sequences, and inheriting curated GO terms from Arabidopsis to the input sequence [18]. A second set of annotations is obtained using a similar approach, but instead of Arabidopsis the search is performed against the top ten annotated plant species in the UniProt database. Presence of valid domains in the input sequences is identified using the InterProScan5 pipeline. InterProScan uses fourteen types of protein signatures to detect putative domains in the input sequences, and assign GO terms [19]. As per documentation, InterProScan

only reports valid domains and GO annotations for the valid domains, so the GO annotations are not filtered based on scores for this step. Three mixed-method pipelines from the first iteration of the CAFA competition (CAFA1 tools) are used to annotate GO terms to the input sequences, namely Argot2.5, FANN-GO and PANNZER [7, 20–22]. Two CAFA1 tools require preprocessing of input sequences before they can be used to annotate GO terms. Argot2 requires the BLAST hits of the input sequences to the UniProt database and Pfam hits identified by HMMER search against the Pfam domain database [23–25]. PANNZER only requires the BLAST hits to the UniProt database for the annotation process. The 6 annotation datasets generated from previous steps are aggregated. Next any redundancy or duplication introduced by aggregation is removed to produce a final aggregate dataset. See Defoin-Platel et al. for the definitions of redundancy and duplication, and maize-GAMER for more details about the annotation methods used in GOMAP [12, 26]. For the analyses described here, non-plant-specific annotations were not removed. See the accompanying GitHub repository for a R script that can be used to filter for plant specific terms. The removal of non plant-specific GO terms did slignly reduce the number of annotations per GO category (See Supplentary table ST2). The GOMAP and community annotations retained about 99% the original annotations. This enables researchers to use such terms to formulate novel hypotheses about potential plant gene functions that could be inspired by data obtained in non-plant systems (e.g., genes involved in the initiation of neurons could be involved in initiation of root hairs, information on flagellar function in lung cells could inform ideas on flagellated sperm function in gymnosperms, etc.).

Implementation and Containerization of GOMAP

The GOMAP pipeline was developed by containerizing the refactoring maize-GAMER code into one singularity container [12, 27]. The GOMAP Pipeline is implemented in Python and R. Python code is used to run open-source tools for annotating GO terms, and R code is used to aggregate and clean annotation results. GOMAP was containerized to improve usability, portability and reproducibility. Containerization eliminates the need to install and configure dependencies. Singularity containerization was chosen because it works seamlessly in high performance computing (HPC) environments [27], and it has been widely adopted by HPC clusters. Several bottlenecks were encountered when containerizing GOMAP: the large size of the pipeline, long runtime on a single machine, and the use of MySQL and MATLAB by mixed-method pipelines.

The uncompressed data required for GOMAP pipeline uses about 110 GB of local disk space. This large size is due to the inclusion of external tools and data, which results in a large container that creates issues during the development and distribution of GOMAP via free public resources. Some tools such as PANNZER were dependent on a back-end MySQL database, and FANN-GO included MATLAB specific code for the annotation. These two components complicated the complete containerization and subsequent efforts to run GOMAP on HPC systems. The original PANNZER code was updated to use a SQLite database, and the SQLite database file works seamlessly in HPC systems eliminating the need for MySQL. The FANN-GO code was updated to use open source GNU Octave instead of MATLAB. The ability to include Octave in the container enabled GOMAP to be run on any HPC system and completely enclose all the data and software required for GOMAP in the container. The pre-built GOMAP containers are currently shared via

CyVerse[1] and GitHub[2] [28]. Run time for GOMAP on a single machine on a single node in the Iowa State University HPC Condo Cluster[3] for 40,000 protein sequences takes between 12-14 days. To improve runtime, GOMAP was separated into different steps that run concurrently. Moreover, the 2 steps with longest runtime, InterProScan (1-2 days) and BLAST search against the UniProt sequence database (8-10 days), were parallelized. Most HPC environments have shorter time limits (e.g. 2-5 days), so parallelizing is necessary to complete these steps within such limitations.

Annotation of maize genomes as a test case

Two versions of the Maize B73 reference genome releases were annotated for the maize-GAMER project. At the completion of containerizing GOMAP, genomes of three more maize inbred lines had been released: W22, Mo17, and PH207. The GOMAP container was used to annotate the three newly released genomes and replicate the annotations for B73 RefGen_v4. The input files were downloaded from MaizeGDB as shown in Table **ST1**. The protein sequences downloaded for each genome were filtered to retain only the longest translated transcript for each gene. Each input fasta file from the 4 different maize genomes was annotated by GOMAP on the Pittsburgh Supercomputing Center (PSC) Bridges HPC Cluster [4]. GOMAP steps were run on Regular Shared Memory nodes. Each regular shared memory node is configured with two Intel Haswell (E5-2695 v3) CPUs (28 Total CPU cores) and 128 GB memory. In addition, two community annotation datasets for B73 RefGen_v4 and PH207 were obtained for comparison. The community annotation for B73 RefGen_v4 was downloaded from Gramene using GrameneMart tool [15, 29]. The community annotations for PH207 (PH207-community) were obtained from the supplemental methods of the original publication by Hirsch et al. [13]. The GOMAP datasets for the four inbred lines and the two community datasets were used for downstream comparison and evaluation.

Assessment and comparison of analysis and evaluation metrics for maize annotation datasets

Maize annotation datasets were assessed using two different metrics: analysis metrics and evaluation metrics. Analysis metrics were used to assess and compare the quantity of the annotations among the datasets, whereas the evaluation metrics assess the quality of the annotations by comparing against a gold-standard dataset produced by manual curation. The data and R code used to evaluate the datasets are available via GitHub [5].

Three analysis metrics, coverage, number of annotations normalized by gene count (i.e., number of annotations), and specificity, were used for the assessment and comparison of maize annotation datasets. Coverage represents the proportion of the total genes that have at least one GO annotation in the predicted dataset. The number of annotations represent the total number of annotations divided by the number of genes with at least one GO annotation. The specificity for a single annotation is calculated by counting the number of ancestral terms, and the mean specificity for all annotations represents the specificity of a dataset. See Defoin-Platel et al. for detailed definitions of the analysis metrics [26]. A general comparison of analysis metrics were performed for the four GOMAP and two community maize annotation datasets. As the next step, annotations were separated into each GO category and analysis metrics were calculated and compared for each GO category. The same GO-category-

[1] http://datacommons.cyverse.org/browse/iplant/home/shared/dillpicl/gomap/GOMAP
[2] https://github.com/Dill-PICL/GOMAP-singularity
[3] https://www.hpc.iastate.edu/guides/condo-2017
[4] https://www.psc.edu/resources/bridges/
[5] https://github.com/wkpalan/GOMAP-maize-analysis

specific approach was used for the generation and comparison of evaluation metrics.

A set of gold-standard annotations are required to calculate evaluation metrics. The gold-standard dataset used in maize-GAMER obtained from MaizeGDB was curated for the B73 RefGen_v3 gene models and not for B73 RefGen_v4 nor other inbred lines [30]. However, MaizeGDB has assigned gene models from B73 RefGen_v3 to other inbred lines' gene models and created a cross reference file[6]. This cross reference file was used to inherit curated GO terms from B73 RefGen_v3 to other inbred lines and create gold-standard datasets for all four inbred lines used in this project. The R script that was used to assign the GO terms is available as part of the GitHub repository. The gold-standard GO terms inherited from B73 RefGen_v3 to B73 RefGen_v4, Mo17, PH207, and Mo17 were used to calculate the protein-centric evaluation metrics defined by Clark and Radivojac [31] and used for the CAFA [5, 7, 9, 32]. The three protein-centric evaluation metrics calculated were Precision (Pr), Recall (Rc) and F_{max}.

Comparison of the GOMAP and the Community, and gold-standard annotations

The comparison of maize annotations produced by GOMAP to the community annotations was restricted to the two inbred lines that had community annotation datasets, namely B73v4 and PH207 [1, 13]. Analysis and evaluation metrics were generated for both datasets and compared to GOMAP-derived datasets. In addition, gold-standard annotations were overlapped with predicted annotations from the community and GOMAP datasets and directly compared. The gold-standard terms that contained only leaf terms were expanded to include all the ancestral terms to the root node, and the same expansion was performed for the predicted annotations. The intersection of gold-standard and predicted terms was performed in three types of objects: gold-standard genes, gold-standard GO terms, and gold-standard annotations. This analysis was used to identify the gold-standard genes, GO terms and GO annotations that were found in both predicted datasets (GOMAP and Community), only in one predicted dataset (GOMAP or Community), or not found in either dataset (only gold-standard). This comparison was performed separately for each GO category for both B73 and PH207.

[6] https://maizegdb.org/search/gene/download_gene_xrefs.php?relative=v4

Results

Annotation of Maize Genomes using GOMAP

The GOMAP container was tested by annotating GO terms to the protein coding genes of four maize inbred lines (B73, Mo17, W22, and PH207). The size the the number protein coding genes were similar among the maize lines as expected (See Table 1). The total predicted protein coding length varied slightly among the inbred lines. W22 has the highest total length, and B73v4 has the longest. The shortest genes that were annotated are less than five amino acids long in all inbreds except B73v3. These are potential annotation errors in the database, but are reported as valid gene models. The median and mean length of the genes in the annotations are similar but vary within a narrow range, and PH207 has the lowest median and mean gene length. Three inbred-lines have longest genes that are over 5000 amino acids long. The genes that are smaller than 50 amino acids present a challenge to predicting GO terms. Mo17 had the highest proportion of genes smaller than 50 amino acids in length (>1300), which incidentally has lowest annotated gene count. All the other inbred lines have less than 1% of genes shorter than 50 amino acids.

Run times of GOMAP steps for different maize genomes

Run times for GOMAP were determined using the PSC Bridges HPC cluster. The manual annotation process of maize-GAMER is complex with over 40 interdependent steps required for end-to-end annotation of a plant genome. To make the annotation process intuitive and convenient, the GOMAP annotation process combined the maize-GAMER steps into just seven discrete steps (see Table 2). The first four steps, seqsim, fanngo, domain, and mixmeth-blast, are setup to be run concurrently as independent processes. The last three steps, mixmeth-preproc, mixmeth, and aggregate, depend on the output of the first four steps. The total time taken to complete the annotation of the maize genomes were between thirty-three and thirty-six hours. The total predicted protein length and gene number had negligible impact on the total runtime of GOMAP for maize genomes, though runtimes of steps were impacted by the load of the cluster. Two parallelized steps, domain and mixmeth-blast, ran longer than other steps, but the runtime has been considerably shortened compared to the un-parallelized versions. The domain step runs for over five days without parallelization and mixmeth-blast runs for over ten days without parallelization. Notably, running steps 1-4 concurrently allows GOMAP to complete the annotation of maize genomes within twenty-four hours for each genome tested.

Assessment and comparison of the analysis metrics for maize annotations

Coverage, number of annotations, and specificity (See Table 3) were calculated for the GOMAP and community datasets. High coverage of around 100% is observed for all GOMAP datasets. In comparison, the community datasets for B73v4 and PH207 have about 77% and 45% overall coverage, respectively. The gold-standard datasets only cover around 3-4% of genes and provide only a smaller number of genes to calculate the CAFA evaluation metrics (Table 3). The annotations were separated by each category to get a more clear picture of the coverage (See Figure 4). The coverage changes substantially across the categories for all datasets. The GOMAP datasets have the highest coverage in the biological process category for all inbred lines (i.e., ∼100%), and have lower coverage other categories (CC:86-92%; MF:82-95%). However, both community datasets have highest coverage in the molecular function category. The PH207 community dataset had the lowest coverage among annotation datasets in all three GO categories, and the PH207-community dataset covered only about ∼10% genes in the cellular component category. The Gramene dataset had higher coverage than the PH207-community, but had lower coverage than GOMAP in all GO categories. This indicates that GOMAP produces higher-coverage datasets than both Gramene or PH207-community methods.

The number of annotations were normalized by dividing the total number of annotations by the number of genes. This normalization allows for comparison among different datasets for the same genome and different genomes. The number of annotations vary among the inbred lines and datasets. B73v3 has the highest number of annotations among the GOMAP datasets, even though W22 had the highest number of protein-coding genes. GOMAP datasets had the highest number of annotations across all inbred lines, followed by the community datasets. The gold-standard datasets had the lowest number of annotations by a significant margin (See Table 3). In some inbreds, such as B73v4 and Mo17, GOMAP had nine times the number of annotations than the corresponding gold-standard dataset. The community datasets also have higher number of annotations than the gold-standard datasets, but the magnitude of difference was lower (∼1-3x). The number of annotations were separated by GO category and compared among each

other. This allowed for the number of annotations to be compared among different inbred lines, annotation sources, and GO categories. The highest number of annotations was seen in GOMAP datasets in the BP category (∼7 annotations per gene), which is significantly higher than the community datasets in BP (B73v4:∼3x; PH207∼6x) and GOMAP datasets in other GO categories. GOMAP datasets have a higher number of annotations compared to the community datasets in all GO categories, but the magnitude of difference is not as high in CC and MF categories. The PH207 community dataset shows the lowest number of annotations across all three GO categories, and this number is especially low in the CC category. In comparison, GOMAP shows lowest number of annotations in the MF category. Gramene datasets for B73v4 has the highest number of annotations in MF and has the lowest in CC.

Specificity indicates the number of ancestral terms for a given annotations given the GO hierarchy, and the mean of all annotations for a particular dataset. Specificity represents a measure of information provided by a specific term. This metric is higher in the community datasets and gold-standard datasets in all three categories (see Table 3), compared to coverage and number of annotations. The Gramene dataset for B73v4 has higher specificity than even the gold-standard dataset. The GOMAP datasets also had lower specificity than the gold-standard datasets. A more detailed analysis separated by each GO category allowed similar comparisons for coverage and number of annotations (see Figure 4). All datasets had higher specificity in BP and MF categories than CC. The Gramene B73v4 dataset has highest specificity across all GO categories, but achieved significantly higher specificity in the BP category. The PH207 community dataset has higher specificity than GOMAP only in BP category, but GOMAP has slightly higher coverage in both CC and MF categories.

Assessment and comparison of the evaluation metrics for maize annotations

The evaluation metrics were calculated by comparing the predicted annotations to the gold-standard datasets. Three protein-centric evaluation metrics from CAFA were used to assess the annotations: Precision (Pr), Recall(Rc), and F_{max}. Precision measures the proportion of predicted annotations that overlap gold-standard. Recall measures the proportion of gold-standard annotations that are correctly predicted. F_{max} is the harmonic mean of Pr and Rc and provides a single number for comparison among different methods. The evaluation metrics were calculated separately for each GO category (see Figure 5). An important factor to notice is the total number of gold-standard annotations are imbalanced and are skewed toward the CC category (See Table 4). This skewed distribution of gold-standard data directly affects the calculation of the evaluation metrics, and this is indicated by the wider standard error bars seen in MF and BP categories in Figure 5. Evaluation metrics compare the performance of the methods used for annotation, thus the following conventions are used to describe the annotation methods for maize datasets. The community method used to annotate B73v4 is called "Gramene" and the community method used to annotate PH207 is called "PH207-community" in the following section.

All methods had higher Precision in the CC category compared to other categories, while BP category had the lowest Precision overall. PH207-community method achieved the highest Precision among all datasets in all three GO categories. Furthermore, PH207-community has substantially higher Precision than GOMAP in CC and BP categories. Gramene also obtained higher precision for B73v4 than GOMAP in

all three categories, although the magnitude of difference was lower. The method employed by the PH207-community is more precise in comparison to other methods. Recall values did not show a clear performance trend as seen with Precision. The recall performance varied among the methods, and no single methods performed better than other methods across all GO categories. GOMAP achieved better Recall in both CC and BP categories, but Gramene showed slightly better recall than GOMAP in the MF category. It was clear both GOMAP and Gramene outperformed PH207-community method in all categories, and the recall was more than 5-10x higher for GOMAP in both CC and BP categories. GOMAP is the only method that achieved higher or comparable performance to other methods in all three categories.

F_{max} gives a single number for the comparison of the performance of the three methods. Similar to Recall, no one method showed higher performance in all three GO categories. Gramene showed higher performance in MF and BP categories, but GOMAP had higher F_{max} in CC category. The higher precision achieved by Gramene edged Gramene ahead of GOMAP in both categories, and higher recall edged GOMAP ahead in the CC category. PH207-community method had lower F_{max} in all three categories, and especially lower by a significant margin in the CC category. PH07-community method showed comparable although slightly lower performance than GOMAP only in MF category. The performance of the PH207-community method was affected by the lower recall observed in all categories.

Comparison of to GOMAP the Community and Curated Annotations

A comparison of genes, GO terms, and annotations between the GOMAP dataset and community dataset was performed for B73v4 and PH207 in each GO category. This comparison was restricted to the gold-standard terms to provide biological validity to the data that was being compared. The recall values of less than one observed in all datasets across all GO categories indicate that no method managed to predict all the annotations in the gold-standard dataset (Figure 5). The comparison allowed for the identification of unique genes and GO terms that were only annotated by a particular method. The comparative proportions of the comparisons are presented in Figure 6 and absolute numbers are presented in Table 4.

GOMAP has annotated more gold-standard genes in both B73v4 and PH207 across all three GO categories than Gramene and PH207-community methods. The majority of genes have annotation from both GOMAP and Gramene for B73v3, but GOMAP and PH207-community methods have annotated a majority of the genes only in MF and BP categories. Due to higher coverage observed in Gramene and GOMAP, the portion gold-standard genes both is higher in all GO categories compared to the proportion of genes annotated by GOMAP and the PH207-community method. Only a few gold-standard genes from CC and BP categories in B73v4 have been annotated by only Gramene, but a larger number of gold-standard genes in CC and BP have been annotated only by GOMAP. No genes were annotated by the PH207-community method that were not annotated by GOMAP, and a substantially higher proportion of PH207 genes have only been annotated by GOMAP. The same trend is also observed in GO terms annotated by different methods. The majority of GO terms were annotated by both methods for B73v4, but GOMAP annotates more terms to gold-standard genes than Gramene. Gramene has annotated only a few terms in BP and one term in MF that were not annotated by GOMAP to any gold-standard genes. GO terms annotated by the PH207-community method are a subset of GOMAP GO terms, and GOMAP has annotated more than twice the number of GO

terms annotated by the PH207-community method to gold-standard genes in both the CC and BP categories. Unfortunately, a proportion of GO terms in the gold-standard data has not been annotated by any method for both inbreds. This number varies among GO categories, but is higher in CC and BP than MF. Next, the comparison was performed using the gold-standard annotations (i.e., Curated Gene-GO term pairs). GOMAP outperforms Gramene in the proportion of gold-standard annotations that are correctly predicted in CC and BP, but Gramene outperforms GOMAP for MF. Although the number of gold-standard annotations in MF that are only predicted by GOMAP (31) are similar to Gramene (35), the F_{max} difference is significant. The PH207-community annotations are a subset of GOMAP annotations and a substantial number of annotations are only found in GOMAP. This is expected based on the recall values seen in Figure 5. Smaller proportions of gold-standard annotations are not predicted by either method in CC ($\sim$11%) and MF ($\sim$8%) categories, but this number increases to ($\sim$40%) in the BP category.

Discussion

Over the course of the maize-GAMER project our main goal was to improve the maize GO annotation landscape, and develop a reproducible method for annotating plant genomes. During the GOMAP project, we focused on developing a reproducible and high-throughput pipeline that can produce high-coverage and high-quality plant GO annotations. Furthermore, we also wanted the pipeline to be portable across different systems, and be usable by researchers of different backgrounds with minimal effort. We acheived high-quality and high-coverage annotation by streamlining and generalizing GAMER code. We containerized GOMAP for portability and reproducibility, which decreases the effort needed to run the GOMAP pipeline. Moreover, we parallelized the time-consuming steps and decreased the overall runtime from a few weeks to a few days. Since we released GOMAP, six graduate students have annotated gene functions to thirteen plant genomes over the course of an eight-week rotation, and one undergraduate student annotated the grape genome over the course of a single semester for research course credit [33–47].

The comparison of GOMAP to the community annotations illustrates that the GOMAP datasets are higher coverage than other methods as evidenced by coverage and number of annotations. GOMAP shines in recall due to higher coverage, but this comes at the cost of precision. Moreover, a careful comparison of the gold-standard annotations also confirmed that GOMAP does indeed have a significantly lower number of FNs than other methods. We can accept the sacrifice in precision as long as the potential FPs fall within an acceptable margin for a high-coverage annotation pipeline such as GOMAP. We have used the current set of maize gold-standard annotations to optimize the balance between precision and recall for GOMAP. The gold-standard data available for maize are incomplete and sparse and inflate the number of FPs. The inflation of FPs leads to underestimation of precision, and under-optimized annotation parameters. At present, the accurate identification of FPs with incomplete gold-standard data has been difficult even for larger-scale efforts such as CAFA. Moreover, the sparse gold-standard data also leads to an inflation of FPs for methods that have higher number of annotations, and in this case the Gramene and GOMAP are affected more than PH207-community approach. PH207-method has higher precision in CC and BP categories, and PH207-community in these two categories have lower number of annotations and coverage. In comparison, both GOMAP and Gramene have lower precision in those categories, indicating that some of the correct predictions have been classified as FPs. We

expect the gradual accumulation of gold-standard annotations will not only improve the optimization of annotation methods, but also precision metric calculation.

Comparison of the methods also indicates that GOMAP annotation quality is comparable to the Gramene method. We designed GOMAP not as a replacement for Gramene but as a supplemental source of annotations. Gramene has been a important resource that provides the plant community with high-quality annotations and invaluable community outreach, and is a federally funded organization. The latest updates from Gramene indicate large-scale curation efforts to improve functional annotations [48]. The curation efforts will improve the annotations of the plant genomes currently available in Gramene, but will not be easily transferable to newly assembled genomes. Unfortunately, Gramene doesn't include all newly released plant genomes. For example, Three out of the four inbred lines that were annotated in this paper are not currently available in Gramene. We expect GOMAP to allows researchers to annotate their own plant genomes or translated transcriptomes in a high-throughput manner and produce annotations of comparable quality to sophisticated methods employed by Gramene. This reduces the time for functional annotation of newly assembled genomes and leads to better understanding of the sequenced genomes.

We started developing GOMAP after the first round of the CAFA competition (CAFA1) results had been announced. GAMER and subsequently GOMAP were developed based on three of the top performing CAFA1 methods. Overall performance of CAFA1 methods were better than naive methods such as BLAST or Pfam. We are grateful for the effort to organize the CAFA competitions and the function prediction community for developing these methods for GOMAP. Moreover, the CAFA competition standardized evaluation methods and provided an unbiased and effective method to compare across different annotation methods. While methods that participated in CAFA improved the quality of the predictions, they were not assessed in the context of annotating non-model plants nor for genome-wide performance. GOMAP bridges the gap between the top performing functional annotation methods and adapts them to a plant-specific context. Over the course of the GOMAP project, we also assessed and optimized the quantity of annotations produced for entire genomes. We have seen continuous improvement in the function prediction methods over CAFA2 and CAFA3. The top performing methods of CAFA2 and CAFA3 have improved the quality of the annotations further as evidenced by F_{max}. We expect GOMAP can be further improved by adding top performing methods from CAFA2 and CAFA3 to the system. Assessing newer tools could also allow us to decouple GOMAP from external methods such as Argot2 and create a self-contained pipeline without sacrificing the quality of annotations produced.

Conclusions

Competing interests

The authors declare that they have no competing interests.

Author's contributions

K.W. designed the pipeline and the computational implementation and analyzed the data. K.W. and C.J.L-D. wrote the manuscript.

Acknowledgments

Thanks to: R. Walls and D. Campbell for generating data DOIs and hosting GOMAP data on CyVerse; N. Weeks for helping adapt FANN-GO to use GNU Octave instead of MATLAB; I. Braun, S. Cannon, A. Jain, G. Kandoi, and N. Weeks for testing the GOMAP pipeline and for valuable suggestions. Thanks to K. Chiteri, H. Dostalik, L. Fattel, P. Joshi, H. Vu, D. Psaroudakis, and C. Yanarella for using GOMAP to annotate plant genomes. Finally, thanks to all members of the Dill Plant Informatics and Computational Lab (dill-picl.org) for in-depth discussions about the project and for offering helpful suggestions.

This work has been supported by the XSEDE startup allocation awarded to K.W. and C.J.L-D; funding from the Iowa State University Plant Sciences

Institute Faculty Scholars Program to C.J.L-D.; and funding from the National Science Foundation [IOS #1027527] to C.J.L-D.

Author details

[1]Bioinformatics and Computational Biology Program, , Iowa State University, Ames, 50010 Iowa, USA. [2]Department of Genetics Development and Cell Biology, Iowa State University, Ames, 50010 Iowa, USA. [3]Department of Agronomy, Iowa State University, Ames, 50010 Iowa, USA.

References

1. Jiao Y, Peluso P, Shi J, Liang T, Stitzer MC, Wang B, et al. Improved maize reference genome with single-molecule technologies. Nature. 2017 Jun;546(7659):524–527. Available from: http://dx.doi.org/10.1038/nature22971.
2. Sun S, Zhou Y, Chen J, Shi J, Zhao H, Zhao H, et al. Extensive intraspecific gene order and gene structural variations between Mo17 and other maize genomes. Nature Genetics. 2018 jul;50(9):1289–1295. Available from: http://dx.doi.org/10.1038/s41588-018-0182-0.
3. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000 May;25(1):25–29. Available from: http://dx.doi.org/10.1038/75556.
4. Grossmann S, Bauer S, Robinson PN, Vingron M. Improved detection of overrepresentation of Gene-Ontology annotations with parent child analysis. Bioinformatics. 2007 Nov;23(22):3024–3031. Available from: http://dx.doi.org/10.1093/bioinformatics/btm440.
5. Jiang Y, Oron TR, Clark WT, Bankapur AR, D'Andrea D, Lepore R, et al. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. Genome Biol. 2016 Sep;17(1):184. Available from: http://dx.doi.org/10.1186/s13059-016-1037-6.
6. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, et al. GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol. 2010 May;28(5):495–501. Available from: http://dx.doi.org/10.1038/nbt.1630.
7. Radivojac P, Clark WT, Oron TR, Schnoes AM, Wittkop T, Sokolov A, et al. A large-scale evaluation of computational protein function prediction. Nat Methods. 2013 Mar;10(3):221–227. Available from: http://dx.doi.org/10.1038/nmeth.2340.
8. Young MD, Wakefield MJ, Smyth GK, Oshlack A. Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol. 2010 Feb;11(2):R14. Available from: http://dx.doi.org/10.1186/gb-2010-11-2-r14.
9. Zhou N, Jiang Y, Bergquist TR, Lee AJ, Kacsoh BZ, Crocker AW, et al. The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. Genome Biology. 2019 nov;20(1):244. Available from: http://dx.doi.org/10.1186/s13059-019-1835-8.
10. Amar D, Frades I, Danek A, Goldberg T, Sharma SK, Hedley PE, et al. Evaluation and integration of functional annotation pipelines for newly sequenced organisms: the potato genome as a test case. BMC Plant Biol. 2014 Dec;14:329. Available from: http://dx.doi.org/10.1186/s12870-014-0329-9.
11. Zwaenepoel A, Diels T, Amar D, Van Parys T, Shamir R, Van de Peer Y, et al. Morphdb: prioritizing genes for specialized metabolism pathways and gene ontology categories in plants. Front Plant Sci. 2018 Mar;9:352. Available from: http://dx.doi.org/10.3389/fpls.2018.00352.
12. Wimalanathan K, Friedberg I, Andorf CM, Lawrence-Dill CJ. Maize GO Annotation-Methods, Evaluation, and Review (maize-GAMER). Plant Direct. 2018 Apr;2(4):e00052. Available from: http://doi.wiley.com/10.1002/pld3.52.
13. Hirsch CN, Hirsch CD, Brohammer AB, Bowman MJ, Soifer I, Barad O, et al. Draft Assembly of Elite Inbred Line PH207 Provides Insights into Genomic and Transcriptome Diversity in Maize. Plant Cell. 2016 nov;28(11):2700–2714. Available from: http://dx.doi.org/10.1105/tpc.16.00353.
14. Springer NM, Anderson SN, Andorf CM, Ahern KR, Bai F, Barad O, et al. The maize W22 genome provides a foundation for functional genomics and transposon biology. Nature Genetics. 2018 jul;50(9):1282–1288. Available from: http://dx.doi.org/10.1038/s41588-018-0158-0.
15. Tello-Ruiz MK, Naithani S, Stein JC, Gupta P, Campbell M, Olson A, et al. Gramene 2018: unifying comparative genomics and pathway resources for plant research. Nucleic Acids Res. 2018 jan;46(D1):D1181–D1189. Available from: http://dx.doi.org/10.1093/nar/gkx1111.
16. Berardini TZ, Reiser L, Li D, Mezheritsky Y, Muller R, Strait E, et al. The Arabidopsis information resource: Making and mining the "gold standard" annotated reference plant genome. Genesis. 2015 Aug;53(8):474–485. Available from: http://dx.doi.org/10.1002/dvg.22877.
17. Consortium U. UniProt: a hub for protein information. Nucleic Acids Res. 2015 Jan;43(Database issue):D204–12. Available from: http://dx.doi.org/10.1093/nar/gku989.
18. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990 Oct;215(3):403–410. Available from: http://dx.doi.org/10.1016/S0022-2836(05)80360-2.
19. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014 May;30(9):1236–1240. Available from: http://dx.doi.org/10.1093/bioinformatics/btu031.
20. Falda M, Toppo S, Pescarolo A, Lavezzo E, Di Camillo B, Facchinetti A, et al. Argot2: a large scale function prediction tool relying on semantic similarity of weighted Gene Ontology terms. BMC Bioinformatics. 2012 Mar;13 Suppl 4:S14. Available from: http://dx.doi.org/10.1186/1471-2105-13-S4-S14.
21. Clark WT, Radivojac P. Analysis of protein function and its prediction from amino acid sequence. Proteins. 2011 Jul;79(7):2086–2096. Available from: http://dx.doi.org/10.1002/prot.23029.
22. Koskinen P, Törönen P, Nokso-Koivisto J, Holm L. PANNZER: high-throughput functional annotation of uncharacterized proteins in an error-prone environment. Bioinformatics. 2015 May;31(10):1544–1552. Available from:

http://dx.doi.org/10.1093/bioinformatics/btu851.

23. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic Acids Res. 2014 Jan;42(Database issue):D222–30. Available from: http://dx.doi.org/10.1093/nar/gkt1223.
24. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W29–37. Available from: http://dx.doi.org/10.1093/nar/gkr367.
25. Consortium U. UniProt: a hub for protein information. Nucleic Acids Res. 2015 Jan;43(Database issue):D204–12. Available from: http://dx.doi.org/10.1093/nar/gku989.
26. Defoin-Platel M, Hindle MM, Lysenko A, Powers SJ, Habash DZ, Rawlings CJ, et al. AIGO: towards a unified framework for the analysis and the inter-comparison of GO functional annotations. BMC Bioinformatics. 2011 Nov;12:431. Available from: http://dx.doi.org/10.1186/1471-2105-12-431.
27. Kurtzer GM, Sochat V, Bauer MW. Singularity: Scientific containers for mobility of compute. PLoS ONE. 2017 May;12(5):e0177459. Available from: http://dx.doi.org/10.1371/journal.pone.0177459.
28. Merchant N, Lyons E, Goff S, Vaughn M, Ware D, Micklos D, et al. The iPlant Collaborative: Cyberinfrastructure for Enabling Data to Discovery for the Life Sciences. PLoS Biol. 2016 Jan;14(1):e1002342. Available from: http://dx.doi.org/10.1371/journal.pbio.1002342.
29. Spooner W, Youens-Clark K, Staines D, Ware D. GrameneMart: the BioMart data portal for the Gramene project. Database: the Journal of Biological Databases and Curation. 2012 feb;2012:bar056. Available from: http://dx.doi.org/10.1093/database/bar056.
30. Andorf CM, Cannon EK, Portwood JL, Gardiner JM, Harper LC, Schaeffer ML, et al. MaizeGDB update: new tools, data and interface for the maize model organism database. Nucleic Acids Research. 2016 jan;44(D1):D1195–201. Available from: http://dx.doi.org/10.1093/nar/gkv1007.
31. Clark WT, Radivojac P. Information-theoretic evaluation of predicted ontological annotations. Bioinformatics. 2013 jul;29(13):i53–61. Available from: http://dx.doi.org/10.1093/bioinformatics/btt228.
32. Zhou G, Wang J, Zhang X, Guo M, Yu G. Predicting functions of maize proteins using graph convolutional network. BMC Bioinformatics. 2020 dec;21(Suppl 16):420. Available from: http://dx.doi.org/10.1186/s12859-020-03745-6.
33. Psaroudakis D, Larence-Dill C. GOMAP Soybean JGI-Wm82.a4.v1. CyVerse Data Commons; 2019. Available from: http://dx.doi.org/10.25739/59ec-1719.
34. Vu H, Larence-Dill C. GOMAP Rice Reference Sequences 2.0. CyVerse Data Commons; 2019. Available from: http://dx.doi.org/10.25739/53g0-j859.
35. Psaroudakis D, Larence-Dill C. GOMAP Wheat Reference Sequences 1.1. CyVerse Data Commons; 2019. Available from: http://dx.doi.org/10.25739/65kf-jz20.
36. Yanarella C, Larence-Dill C. GOMAP Barley Reference Sequences IBSC_PGSB_r1. CyVerse Data Commons; 2019. Available from: http://dx.doi.org/10.25739/zvgv-8e37.
37. Psaroudakis D, Larence-Dill C. GOMAP Bdistachyon.Bd21.v3.1.r1. CyVerse Data Commons; 2019. Available from: http://dx.doi.org/10.25739/dw2t-3g82.
38. Wimalanathan K, Larence-Dill C. GOMAP Sbicolor.BTx623.v3.0.1.r1. CyVerse Data Commons; 2019. Available from: http://dx.doi.org/10.25739/4ty0-ye98.
39. Psaroudakis D, Larence-Dill C. GOMAP Peanut IPGI 1.0. CyVerse Data Commons; 2019. Available from: http://dx.doi.org/10.25739/chab-0e35.
40. Psaroudakis D, Larence-Dill C. GOMAP Barrel Clover R108_HM340 v1.0. CyVerse Data Commons; 2019. Available from: http://dx.doi.org/10.25739/2sqc-j140.
41. Psaroudakis D, Larence-Dill C. GOMAP Barrel Clover A17_HM341 Mt4.0v2. CyVerse Data Commons; 2019. Available from: http://dx.doi.org/10.25739/py38-yb08.
42. Psaroudakis D, Larence-Dill C. GOMAP Common Bean DOE-JGI and USDA-NIFA v2.0. CyVerse Data Commons; 2019. Available from: http://dx.doi.org/10.25739/1ywe-ew96.
43. Psaroudakis D, Larence-Dill C. GOMAP Cowpea IT97K-499-35 JGI annotation v1.1. CyVerse Data Commons; 2019. Available from: http://dx.doi.org/10.25739/cdx9-wr97.
44. Yanarella C, Larence-Dill C. GOMAP TreeGenesDB sugar pine assembly v1.5. CyVerse Data Commons; 2020. Available from: http://dx.doi.org/10.25739/jvs4-xr88.
45. Chiteri K, Larence-Dill C. GOMAP_Cannabis_sativa_NCBI-cs10_January_2020. CyVerse Data Commons; 2020. Available from: http://dx.doi.org/10.25739/ab9z-2z86.
46. Joshi P, Larence-Dill C. CyVerse Data Commons;.
47. Dostalik H, Larence-Dill C. Carolyn_Lawrence_Dill_GOMAP_Grape_Genoscope_12x_January_2021.r1. CyVerse Data Commons; 2021. Available from: http://dx.doi.org/10.25739/jtfk-q888.
48. Tello-Ruiz MK, Naithani S, Gupta P, Olson A, Wei S, Preece J, et al. Gramene 2021: harnessing the power of comparative genomics and pathways for plant research. Nucleic Acids Research. 2020 nov;Available from: https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkaa979/5973447.

Figures

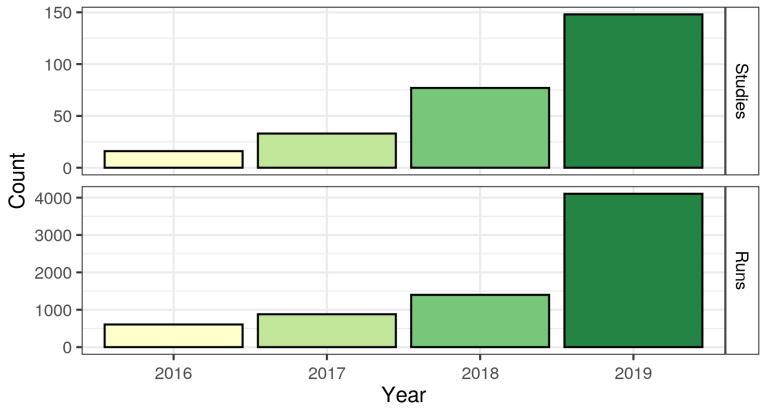

Figure 1: **Long-read sequencing datasets published for plants in the NCBI Short Read Archive.**

Top histogram: number of studies or experiments 2012-2019. Bottom: number of runs deposited for the same time period (each study is made up of one or more runs).

NCBI SRA was queried on Jan 30, 2021 with the following parameters: ((”pacbio smrt”[Platform]) OR ”oxford nanopore”[Platform]) AND Embryophyta[Organism] AND wgs[strategy])

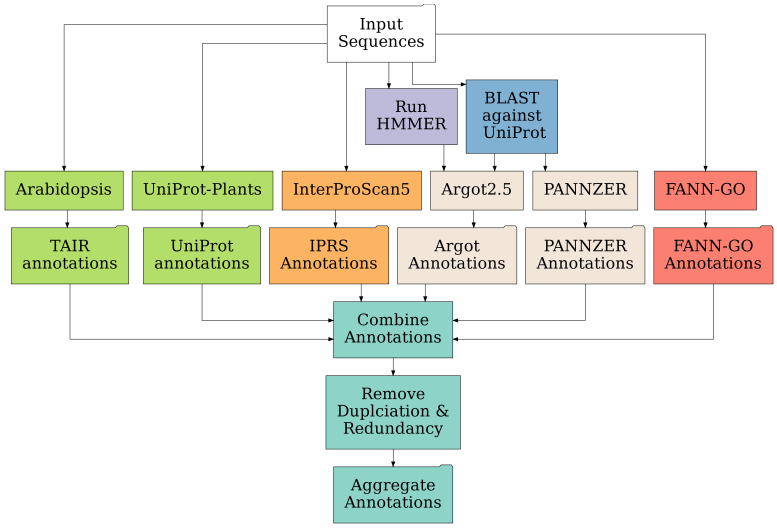

Figure 2: **Overview of processes used to create GOMAP Annotations**

Top: data inputs are shown as a white box. Sequence similarity components are shown in lime green. InterProScan domain-based annotations are shown in orange. For the CAFA processes (Argot2.5 and PANNZER; shown in tan boxes) pre-processing steps are shown in purple and blue, respectively. The CAFA process FANN-GO does not require preprocessing (red). Once each annotation type is produced, these are combined, duplicates and redundancies are removed, and the aggregate dataset is assembled (turquoise).

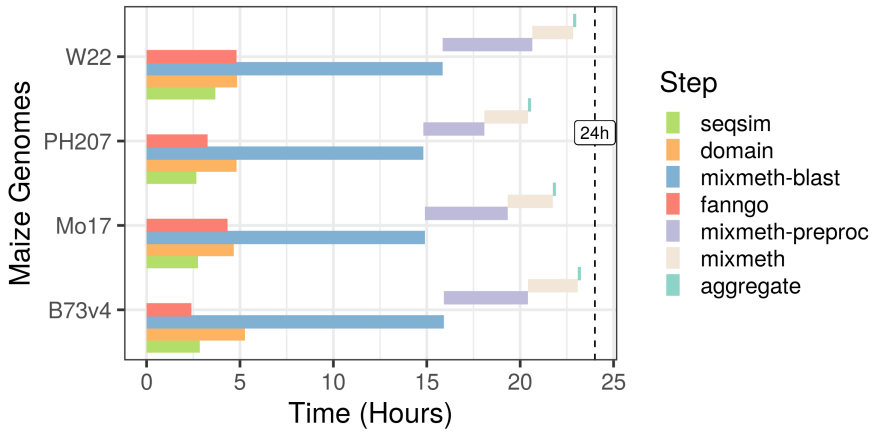

Figure 3: **Comparison of runtime for GOMAP steps across four maize genomes.**

Steps are color-coded as shown in the figure key at right. Start time occurs at zero. The four steps shown simultaneously within a single maize genome (i.e., seqsim, domain, mixmeth-blast, and fanngo) run in parallel. For all maize genomes, the full annotation time took less than 24 hours on the PSC Bridges cluster.

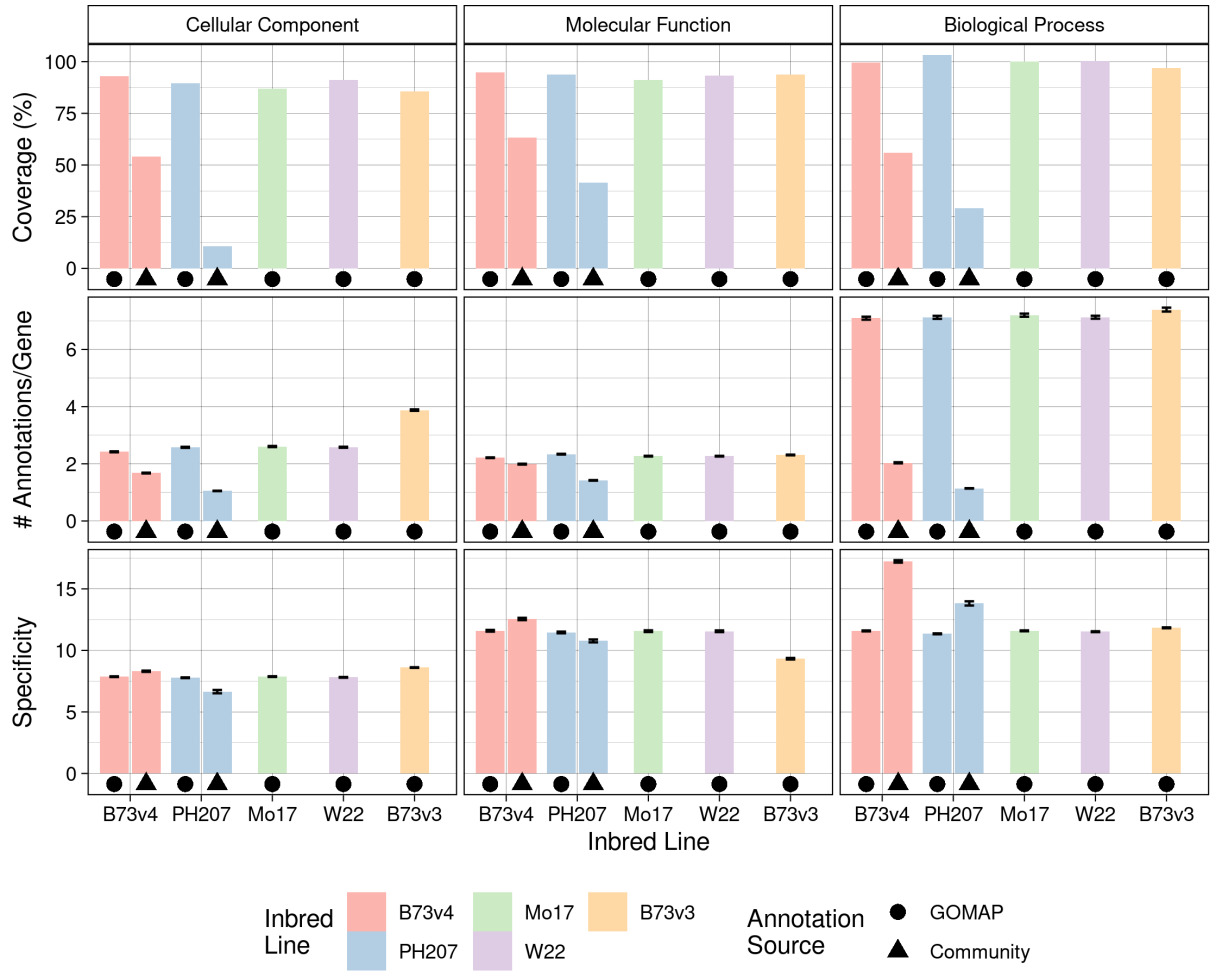

Figure 4: **Analysis metrics calculated for the maize annotations from community and GOMAP annotations.**

Left column: Cellular Component. Middle column: Molecular Function. Right column: Biological Process. Top row: Percentage of genes with an annotation. Middle row: number of annotations per gene. Bottom row: specificity of the annotations. Inbred lines are color-coded as shown in the key at bottom. GOMAP annotations are denoted by a filled circle. Community annotations are denoted by a filled triangle. Coverage is shown as an overall percentage, but both number of annotation per gene and specificity are represented as mean values across all annotations in the dataset. Error bars indicate standard error. The confidence interval is very small so the high and low error bars overlap each other for most datasets.

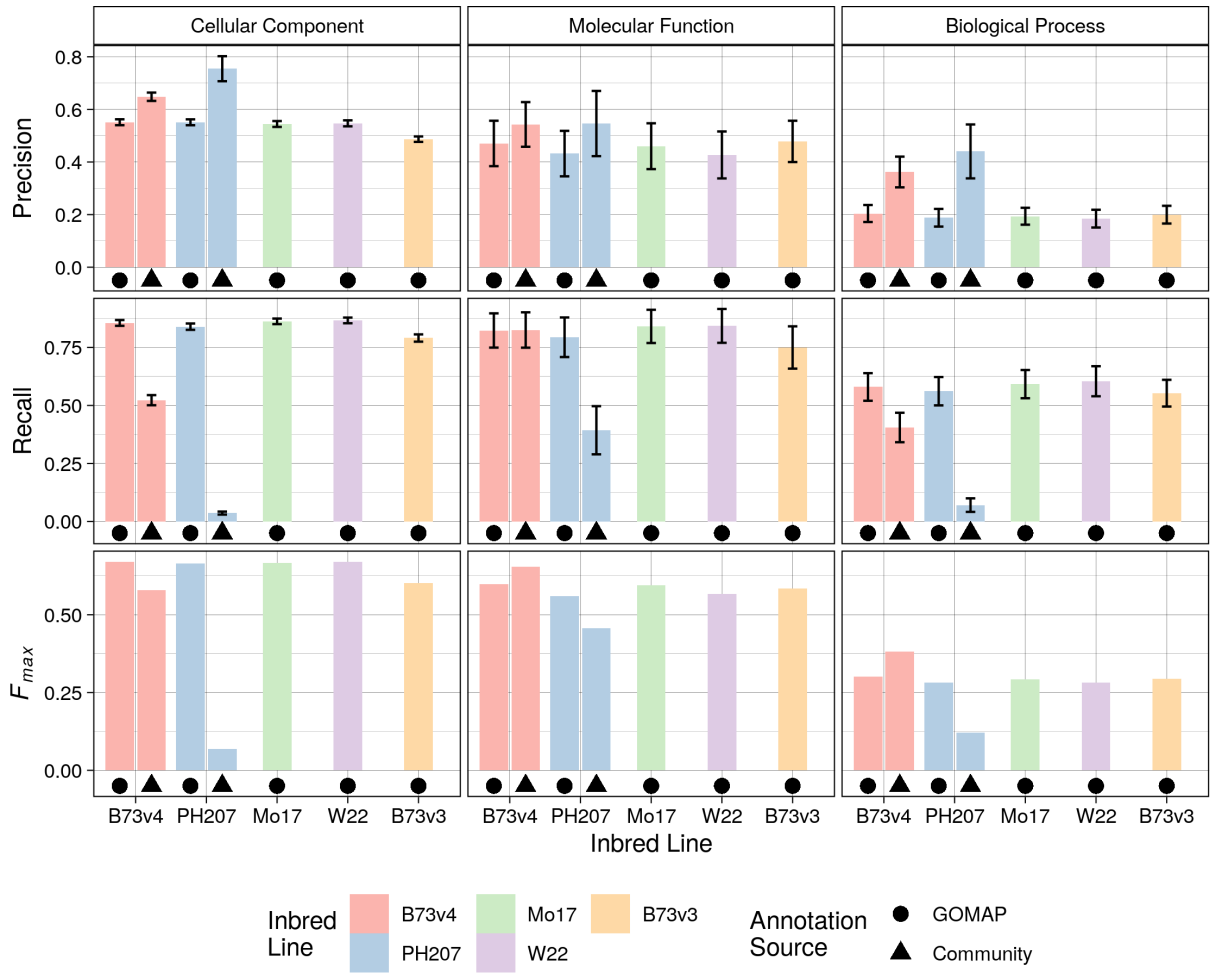

Figure 5: **Evaluation metrics calculated for the maize annotations from the community and GOMAP.**

Left column: Cellular Component. Middle column: Molecular Function. Right column: Biological Process. Top row: Precision. Middle row: Recall. Bottom row: F_max. Inbred lines are color-coded as shown in the key at bottom. GOMAP annotations are denoted by a filled circle. Community annotations are denoted by a filled triangle. Precision and recall are shown as the mean value of all annotations with error bars indicating standard error, but F_max is represented as an absolute value for a specific dataset.

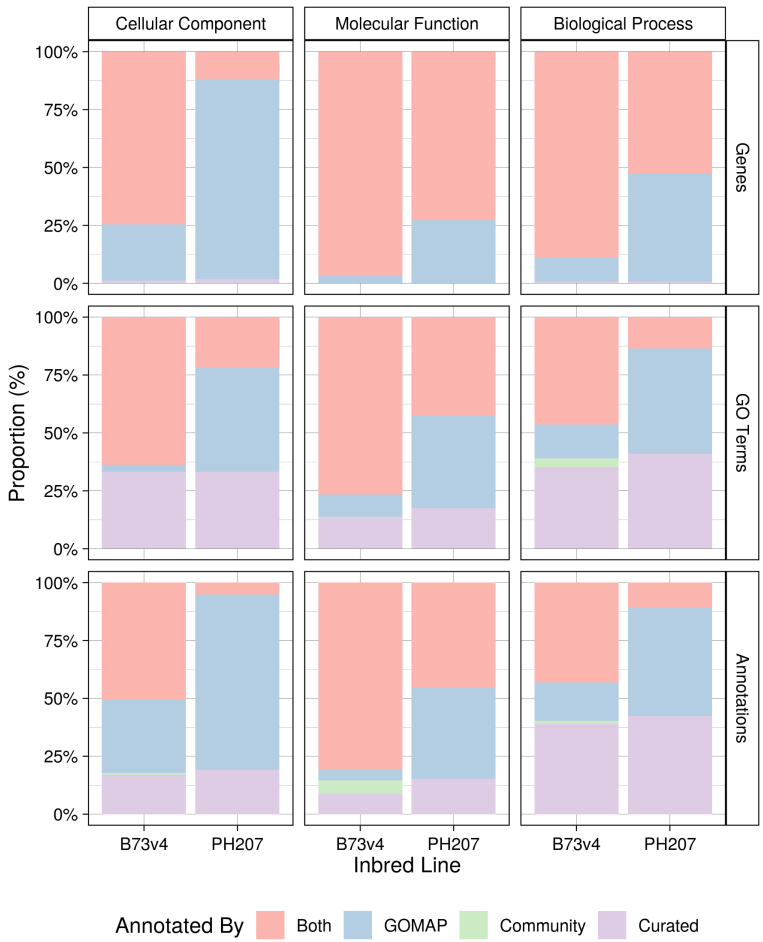

Figure 6: **Comparison of the GOMAP and community annotations based on whether gold standard terms were annotated.**

Left column: Cellular Component. Middle column: Molecular Function. Right column: Biological Process. Top row: Percentage of genes with at least one annotation. Middle row: proportion of unique GO terms recovered. Bottom row: proportion of expanded GO annotations recovered. Gold standard genes or annotations recovered by both the community and GOMAP methods are shown in pink. Those recovered by GOMAP but not the community method are shown in blue. Those recovered by the community annotation but not GOMAP are shown in green. Those annotated in the gold-standard that were not recovered by either method are shown in lavender.

Tables

Table 1: Comparison of maize input sequences

Inbred	Gene Count [1]	Total AA	Length				Small Proteins(%)[2]
			min	mean	median	max	
B73v3	39,475	14,382,005	25	364.33	306	4,743	0.56
B73v4	39,324	15,373,604	2	390.95	316	5,267	0.86
Mo17	38,620	14,640,283	5	379.09	306	5,426	3.44
PH207	40,557	14,311,872	2	352.88	280	4,947	0.33
W22	40,690	15,439,503	1	379.44	304	5,426	0.25

[1]This indicates the protein coding genes

[2]Any protein smaller than 50 amino acids was classified as small

Table 2: Comparison of the runtime of different GOMAP steps on PSC Bridges Cluster

Num	Step	Description	Depends On	Nodes	B73v4	Mo17	PH207	W22
1	seqsim	Runs sequence similarity steps	NA	1	2h51m	2h45m	2h40m	3h41m
2	domain[1]	Run InterProScan5	NA	10	5h16m	4h40m	4h49m	4h51m
3	mixmeth-blast[1]	Run the BLAST step against UniProt	NA	10	15h55m	14h54m	14h49m	15h51m
4	fanngo	Run FANN-GO	NA	1	2h24m	4h20m	4h40m	4h49m
5	mixmeth-preproc	Convert output from UniProt BLAST and Run HMMER	3	1	4h30m	4h26m	3h16m	5h57m
6	mixmeth	Submit jobs to Argot2.5 web-server and Run PANNZER	5	1	2h40m	2h25m	2h20m	2h11m
7	aggregate	Combine all GO annotations, clean, and generate aggregate dataset	1,2,4,6	1	0h10m	0h10m	0h10m	0h10m

[1]These steps have been parallelized to be run on MPI based HPC workload manager.

Table 3: Analysis metrics of GOMAP annotations for maize genomes.

Inbred	Source	Total Genes	Coverage (%)		Annotations/Gene		Specificity	
			Curated	Predicted	Curated	Predicted	Curated	Predicted
B73v3	GOMAP	39,475	4.10	99.98	1.23	13.05	11.98	10.56
B73v4	Community	39,324	4.12	76.94	1.23	4.30	11.98	13.03
B73v4	GOMAP	39,324	4.12	100.00	1.23	11.46	11.98	10.85
W22	GOMAP	40,690	3.08	100.00	1.24	11.58	12.10	10.78
Mo17	GOMAP	38,620	3.42	100.00	1.26	11.53	12.00	10.85
PH207	Community	40,557	3.34	45.44	1.26	2.21	12.06	11.31
PH207	GOMAP	40,557	3.34	100.00	1.26	11.48	12.06	10.67

Table 4: Number of gold-standard genes, GO terms, and annotations that were assigned by GOMAP, the community annotation, and the gold-standard.

Inbred Line	Type	Cellular Component			Molecular Function			Biological Process		
		Genes	Terms	Annotations	Genes	GO Terms	Annotations	Genes	Terms	Annotations
B73v4	Both	980	64	8,014	56	183	518	126	271	1,305
	GOMAP	317	4	5,013	2	23	31	15	85	504
	Community	2	0	94	0	1	36	0	22	35
	Curated	14	34	2,703	0	32	57	1	206	1,184
	Total	1313	102	15824	58	239	642	142	584	3028
PH207	Both	151	22	784	40	100	283	71	77	304
	GOMAP	1,104	46	11,712	15	94	245	63	260	1,319
	Curated	24	34	2,987	0	41	96	1	234	1,194
	Total	1279	102	15483	55	235	624	135	571	2817

[a]The gold-standard data overlaps with only GOMAP

[b]The gold-standard data overlaps with only Community

[c]The gold-standard data does not overlap with GOMAP or Community

Additional Files

Additional file 1 — GOMAP-Supplemental

The supplemental tables are provided as a PDF file that can be opened with the free software.