Standardized genome-wide function prediction enables comparative functional genomics in plants: a new application area for Gene Ontologies

Leila Fattel[1,*], Dennis Psaroudakis[2,*,†], Colleen F. Yanarella[1,‡], Kevin Chiteri[1,‡], Haley A. Dostalik[1,‡], Parnal Joshi[4,‡], Dollye C. Starr[1,‡], Ha Vu[5,‡], Kokulapalan Wimalanathan[5,§], and Carolyn J. Lawrence-Dill[1,5,¶]

[1] *Department of Agronomy, Iowa State University*
[2] *Department of Plant Pathology and Microbiology, Iowa State University*
[3] *Department of Ecology, Evolution and Organismal Biology, Iowa State University*
[4] *Department of Veterinary Microbiology and Preventative Medicine, Iowa State University*
[5] *Department of Genetics, Development and Cell Biology, Iowa State University*

(Dated: April 26, 2021)

Background Genome-wide gene function annotations are useful for hypothesis generation and for prioritizing candidate genes responsible for phenotypes of interest. We functionally annotated the genes of 18 crop plant genomes across 14 species using the GOMAP pipeline.

Results By comparison to existing GO annotation datasets available for a subset of these genomes, GOMAP-generated datasets cover more genes, assign more GO terms, and produce datasets similar in quality (based on precision and recall metrics using existing gold standards as the basis for comparison). From there, we sought to determine whether the datasets could be used in tandem to carry out comparative functional genomics analyses. As a test of the idea and a proof of concept, we created parsimony and distance-based dendrograms of relatedness based on functions for all 18 genomes. These dendrograms were compared to well-established species-level phylogenies to determine whether trees derived through the analysis of gene function agree with known evolutionary histories, which they largely do. Where discrepancies were observed, we determined branch support based on jack-knifing then removed individual annotation sets by genome to identify the annotation sets causing errant relationships.

Conclusions Based on the results of these analyses, it is clear that for genome assembly and annotation products of similar quality, GOMAP-derived functional annotations used together across species do retain sufficient biological signal to recover known phylogenetic relationships, indicating that comparative functional genomics across species based on GO data hold promise as a tool for generating novel hypotheses about gene function and traits.

Keywords: Gene function; ontology; plants; comparative genomics; functional genomics

I. BACKGROUND

The Gene Ontology (GO) is a vocabulary organized as a directed acyclic graph, which makes it computationally useful as an organized classification system of gene functions [2, 48]. GO-based gene function annotation is the association of particular GO terms to specific genes. Functions may be assigned to genes based on different types of evidence for the association. For example, functional predictions can be inferred from experiments (EXP), inferred from expression pattern (IEP), and more [4]. Computational pipelines often are used to generate functional predictions for newly sequenced genomes, where the genome is first sequenced and assembled, then genes are predicted, then functions are associated with those gene predictions. Genome-wide gene function prediction datasets are frequently used to analyze gene expression studies, to prioritize candidate genes linked to a phenotype of interest, to determine experiments aimed at characterizing functions of genes, and more [49, 7, 43]. Clearly, how well a gene function prediction set models reality is determined by the correctness of the genome assembly coupled with how well the software used to predict functions performs.

GOMAP (the Gene Ontology Meta Annotator for Plants) is a gene function prediction system for plants that generates high-coverage and reproducible functional annotations [51]. The system employs multiple functional prediction approaches, including sequence similarity, protein domain presence, and mixed-method pipelines developed to compete in the Critical Assessment of Function Annotation (CAFA) Challenge [54], a community challenge that has advanced the performance of gene function prediction pipelines over the course of five organized competitions [1].

We previously annotated gene functions for the maize B73 genome and demonstrated that GOMAP's predicted functions were closer to curated gene-term associations from the literature than those of other community functional annotation datasets ([52]). Using the newly containerized GOMAP system ([51]), we report the functional annotation of 18 plant genomes across the 14 crop plant species shown in Table I.

We were curious to find out whether the datasets could be somehow used together as a set to reveal biologically relevant and interesting perspectives. As a first step in that direction, we describe here a method by which we

*Contributed equally.
†Current address: Department of Molecular Biology, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), 06466 Seeland, OT Gatersleben, Germany
‡Listed alphabetically.
§kokul@bioinformapping.com
¶triffid@iastate.edu

used gene function annotations to generate dendrograms of genome-level similarity in function (see Figure 1 for process overview). This idea is similar to that of Zhu et al., who determined the evolutionary relationships among microorganisms based on whole-genome functional similarity [55]. Here we expand on that approach, analyzing genome-wide GO assignments to generate parsimony and distance-based dendrograms. We compared these with well-established species phylogenies (Figure 2) to determine whether trees derived from gene function agree with evolutionary histories and to better consider whether these datasets could be used together for comparative functional genomics analyses.

II. RESULTS OF ANALYSES

A. Overview

Figure 1 gives an overview of the general workflow of analyses. In brief, gene function annotation sets were created and compiled for each genome. For those with existing annotation sets available, the datasets were compared. From there, matrices that included genomes as rows and terms as columns were generated. These were used directly to build parsimony trees or to create distance matrices for neighbor-joining tree construction [12, 13, 44]. In subsequent analyses, jackknifing was used to remove terms (columns) or to remove genomes (rows) to map the source of signal for treebuilding results [53].

B. Functional Annotation Sets Produced

Table II shows quantitative attributes of each of the annotation sets. In summary, GOMAP covers all annotated genomes with at least one annotation per gene (at least one of which is in the Biological Process aspect), and provides between 3.8 and 12.1 times as many annotations as Gramene or Phytozome [47, 15].

Quality evaluation of gene function predictions is not trivial. Most often datasets are assessed by comparing the set of predicted functions for a given gene to a *Gold Standard* consisting of annotations that are assumed to be correct. This assumption of correctness can be based on any number of criteria. Here we used as our Gold Standard all annotations present in Gramene that had a non-IEA (non-Inferred by Electronic Annotation) evidence code. This enabled us to assess ten of the eighteen annotation sets, as shown in Table II. (Note that the IEA and non-IEA annotation sets from Gramene frequently contain overlaps, indicating that some of the predicted annotations were manually confirmed afterwards by a curator and that in such cases, a new annotation was asserted with the new evidence code rather than simply upgrading the evidence code from IEA to some other code, thus preserving the IEA annotations in Gramene that are produced by the Ensembl analysis pipeline [39]).

There are many different metrics that have been used to evaluate the quality of predicted functional annotations. For the maize B73 GOMAP [52] annotation assessment, we used a modified version of the hierarchical evaluation metrics originally introduced in [50] because they were simple, clear, and part of an earlier attempt at unifying and standardizing GO annotation comparisons [8]. In the meantime, [38] published an approach for evaluating different metrics showing differences among the robustness of different approaches to quality assessment. Here we used the SimGIC2 and Term-centric Area Under Precision-Recall Curve (TC-AUCPCR) metrics recommended by [38]. We also evaluated with the F_{max} metric, because it is very widely-used (e.g., by [54]), but according to [38] it is actually a flawed metric. Results of the quality assessments for the 10 genomes where a Gold Standard was available are shown in Table III and Figure 9. While evaluation values differ between metrics and the scores are not directly comparable, a few consistent patterns emerge: GOMAP annotations are almost always better than Gramene and Phyotzome annotations in the Cellular Component and Molecular Function aspect, with the only three exceptions being the Molecular Function aspect for *Triticum aestivum* using the TC-AUCPCR and the F_{max} metric and the Cellular Component aspect for *Medicago truncatula* A17 using the F_{max} metric. Conversely, GOMAP predictions achieve consistently lower quality scores in the Biological Process aspect with the exception of *Brachypodium dystachion*, *Oryza sativa*, and *Sorghum bicolor* with the TC-AUCPR metric. Generally, annotations that are better in one aspect are also better in the other two aspects (there are few cross-overs in Figure 9), but the ranking of annotations does not necessarily hold across metrics. The Phytozome annotation for *Oryza sativa* is especially bad, potentially because it is based on a modified structural annotations than the Gold Standard and the other annotations.

C. Phylogenetic Tree Analyses

After visualizing the neighbor-joining tree and the parsimony tree of the 18 plant genomes using iTOL [36], the two tree topologies, rooted at *P. lambertiana*, were compared to one another and to the topology of the expected tree (Figure 2). For both the neighbor-joining (Figure 3a) and parsimony trees (Figure 3b), one common difference is noted: *S. bicolor* is not at the base of the *Z. mays* clade as expected, and is clustered with *B. distachyon* instead. Notable differences between the neighbor-joining and parsimony tree are the following: *C. sativa* appears at the base of the dicots instead of *G. raimondii* in the neighbor-joining tree, while *G. raimondii* is grouped with *A. hypogaea* and *C. sativa* is grouped with *G. max* in the parsimony tree. Second, *O. sativa* was expected to be at the base of the BOP clade, but appears at the base of *Z. mays* in the neighbor-joining tree, but at the base of all angiosperms in the parsimony tree. Differences among

Table I: Functional annotation sets generated by GOMAP. More information about each dataset including the source of the input to GOMAP can be found at the respective DOI.

Species	Germplasm/Line	Assembly/Annotation	Dataset DOI
Arachis hypogaea	Tifrunner	Arachis hypogaea assembly 1.0	[28]
Brachypodium distachyon	Bd21	Bd21.v3.1.r1	[21]
Cannabis sativa	Hemp	NCBI Cannabis sativa GCA_900626175.1	[33]
Glycine max	Williams 82	Joint Genome Institute (JGI) Wm82.a4.v1	[31]
Gossypium raimondii	Cotton D	Gossypium raimondii JGI v2.1	[24]
Hordeum vulgare	–	IBSC_PGSB_r1	[18]
Medicago truncatula	R108_HM340	R108: v1.0	[20]
Medicago truncatula	A17_HM341	Mt4.0v2	[19]
Oryza sativa	japonica	IRGSP 1.0	[29]
Phaseolus vulgaris	G19833	DOE-JGI and USDA-NIFA annotation 2.0	[22]
Pinus lambertiana	Sugar Pine	TreeGenesDB sugar pine assembly v1.5	[32]
Sorghum bicolor	BTx623	BTx623.v3.0.1.r1	[30]
Triticum aestivum	Chinese Spring	IWGSC RefSeq 1.1	[6]
Vigna unguiculata	IT97K-499-35	JGI annotation v1.1	[23]
Zea mays[*]	Mo17	Zm-Mo17-REFERENCE-CAU-1.0	[25]
Zea mays[*]	PH207	Zm-PH207-REFERENCE_NS-UIUC_UMN-1.0	[26]
Zea mays[*]	W22	Zm-W22-REFERENCE-NRGENE-2.0 Zm00004b.1	[27]
Zea mays[*]	B73	RefGen_V4 Zm00001d.2	[34]

Latest overview at https://dill-picl.org/projects/gomap/gomap-datasets/

[*] Previously published in [51].

relationships within the *Z. mays* clade constaining B73, PH207, W22, and Mo17 were disregarded given the high degree of similarity across annotation sets and the fact that these relationships are not clear given the complex nature of within-species relationships

Due to differences between the function-based dendrograms and the expected tree, jackknifing analysis was carried out by removing terms (columns in underlying datasets) to determine the degree to which the underlying datasets support specific groupings based on functional term assignments. This analysis was carried out for both neighbor-joining and parsimony trees. First, trees were generated by omitting 5% then 10%, then 15% on up to 95% of the dataset to determine the threshhold at which the tree topologies deviated from those generated using the full dataset. That threshold was at 45%; therefore we used trees generated with 40% of the data removed for reporting branch support for the topology (Figure 3). Comparing the two trees, the topologies were similar at jackknife values up to 40% but the support values for parsimony were comparatively lower. Based on this robustness for neighbor-joining treebuilding in general, we carried out all subsequent analyses using neighbor-joining treebuilding methods.

To map the source of discrepancies to specific gene annotation sets, we generated various neighbor-joining trees excluding one genome each time, an additional tree with both *Medicago* genomes excluded, and another with all *Z. mays* genomes excluded. To exemplify this, see the monocot clade in Figure 2 and the lower (monocot) clade in Figure 3a. When the neighbor-joining tree was generated, two species are misplaced: *S. bicolor* and *O. sativa*. As shown in Figure 4a, removal of *O. sativa* corrects one error (itself) but does not correct the errant grouping of *S. bicolor* with *B. distachyon*. In 4b, it is shown that the removal of *S. bicolor* corrects the errant grouping of itself and *B. distachyon*, but *O. sativa* placement remains incorrect. However, as shown in 4d, the removal of *B. distachyon* generates a tree where all relationships are in agreement with known species-level relationships. (Note well: all individual annotation sets were progressively removed, not just these three shown in the example.)

With this observation in hand, we sought to determine the minimum number of genomes that could be removed to create a tree that matched the expected tree topology. The removal of the three genomes was required to generate function-based trees consistent with known phylogenetic relationships. They are *C. sativa*, *G. max*, and *B. distachyon* (Figure 5). Jackknifing analysis was also carried out for this dataset with support shown. Branch support is generally higher than that for the full dataset (i.e., branch support is higher in Figure 5 than in Figure 3a).

D. Potential Causes of Errant Groupings

As a first step in investigating whether and how the comparative quality of assemblies and annotation sets underlying the predicted gene function datasets could mislead treebuilding, we have begun work to assess the quality of each genome assembly and structural anno-

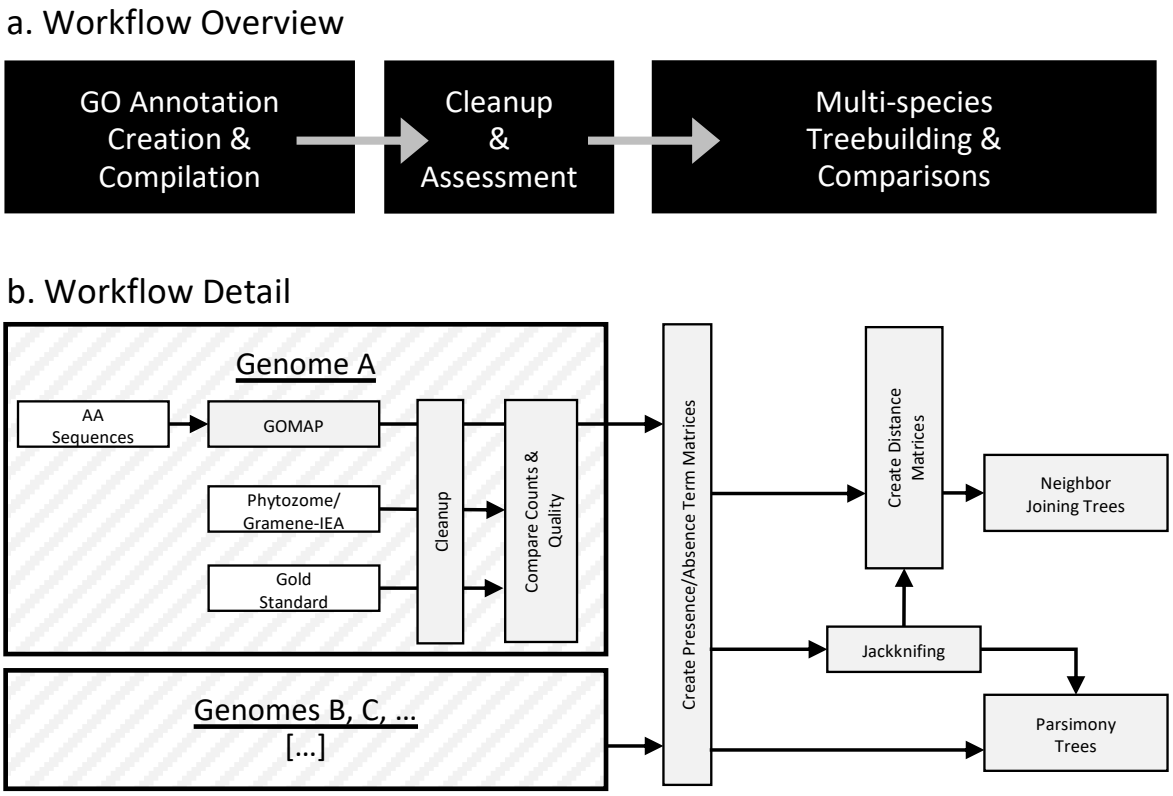


Figure 1: Data workflow schema. The workflow overview is shown in panel 'a' with steps represented as black boxes and the flow of information and processes indicated by arrows. Details are shown in panel 'b' where the upper large hatched box shows process detail for a single genome and the lower hatched box represents additional genomes for which the details of processing are identical. White boxes represent input datasets. Arrows indicate the flow of information and processes.

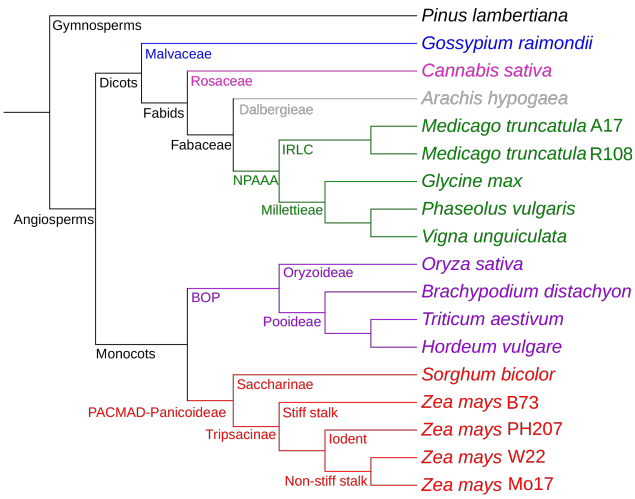


Figure 2: Phylogenetic relationships among species. Cladogram is rooted by the gymnosperm *Pinus lambertiana* (black). Among angiosperms, dicots clades include Malvaceae (blue), Rosaceae (magenta), Dalbergieae (grey), and NPAAA (green). Monocots include members of the BOP (purple) and PACMAD-Panicoideae (red) clades.

tation set using GenomeQC [37]. Although the following analysis and interpretation are preliminary, assembly quality (Figure 6), structural annotation measures of quality (Figure 7), and proportion of single-copy BUSCOS (short for benchmarking universal single-copy genes; [46]) can be compared across assembly and annotation products. As shown in Figure 6, among the genomes available via GenomeQC, *G. max* and *B. distachyon* have comparative low values for percent of the genome that is useful (as calculated based on scaffold number and scaffold size relative to published genome sizes). In addition, as shown in Figure 7, the number of gene models <200 base pairs in length is comparatively high for *G. max*. Compounding these issues, for *G. max*, structural annotations for BUSCOS are duplicated for many *G. max* genes.

In the case of *C. sativa*, the fact that the line sequenced is not inbred may account for misplacement in the function-based dendrograms. This means that to generate an assembly, there are likely regions where alleles are not aligned, which would inflate the length of the assembly. Indeed, the comparatively low-quality assembly for *Cannabis* genome has been noted by others [14], and our preliminary investigations indicate that the assembly length is in fact longer than expected based on

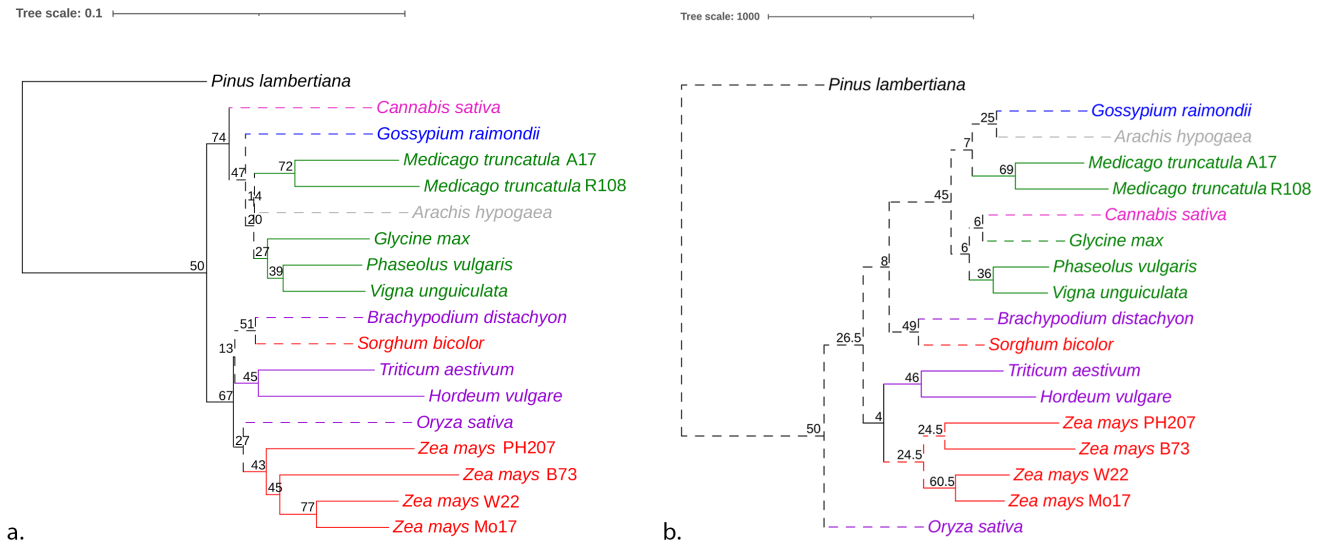


Figure 3: Neighbor-joining and Parsimony Trees. Phylograms are colored and rooted as described in Figure 1. For both neighbor-joining (a) and parsimony (b), node values represent the jackknifing support values derived by removing 40% of GO terms in the dataset. Dashed lines mark deviations from known phylogenetic relationships. Tree scales are shown above each, with NJ showing distances and parsimony showing changes in character state.

C-values for genome sizes reported previously [45]. In addition, we found 184 GO terms missing from *C. sativa* that were present in the rest of the dicots, and 93 additional terms that were only in *C. sativa* but none of the other dicots in our study. This finding could confirm the poor quality of the *C. sativa* genome used in our dataset. More efforts to determine in detail the causes of these discrepancies are currently underway.

III. DISCUSSION

In this study, we used the GOMAP pipeline to produce whole-genome GO annotations for eighteen genome assembly and annotation sets from fourteen plant species [51]. Assessments of the number of terms predicted as well as the quality of predictions (based on F-score) indicate that GOMAP functional prediction datasets cover more genes, produce more predictions per gene, and are of similar quality to smaller prediction datasets produced by other systems, thus supporting the notion that these high-coverage datasets are a useful addition for researchers who are interested in genome-level analyses, including efforts aimed at prioritizing candidate genes for downstream analyses. Given that we can now produce high-quality, whole genome functional annotations for plants in a straightforward way, we intend to produce more of these over time (indeed we recently annotated *Vitis vinifera* Pinot Noir grape [9], *Brassica rapa* (doi in process), *Musa acuminata*[10], *Theobroma cacao* [11], *Solanum lycopersicum* [40], and *Solanum pennellii* [41]).

With eighteen genome functional annotations in hand, we sought to determine whether and how researchers could use multispecies GO annotation datasets to perform comparative functional genomics analyses. As a first step in that direction and a proof of concept, we adapted phylogenetic tree-building methods to use the gene function terms assigned to genes represented in genomes to build dendrograms of functional relatedness and hypothesized that if the functions were comparable across species, the resulting trees would closely match evolutionary relationships. To our delight and surprise, the neighbor-joining and parsimony trees (Figure 3) did resemble known phylogenies, but were not exact matches to broadly accepted phylogenetic relationships.

After removing the minimum number of genomes that resulted in restoration of the expected evolutionary relationships, we found that the individual species that may be responsible for the discrepancies observed in Figure 3 were *C. sativa*, *G. max*, and *B. distachyon*. We hypothesize in a general way that the following reasons could account for these errant relationships:

1) Quality of sequencing and coverage assembly: genomes of similarly high sequence coverage that have excellent gene calling would be anticipated to create the best source for functional annotation. Genomes of comparatively lower, or different, character would be anticipated to mislead treebuilding and other comparative genomics approaches.

2) Reference guided assembly: if one genome is used to guide the assembly and annotation of another genome, some similarities may result naturally due to the inheritance of information.

3) Shared selected or natural traits: species that have been selected for, e.g., oilseeds may share genes involved in synthesis of various oils. Other shared traits would be

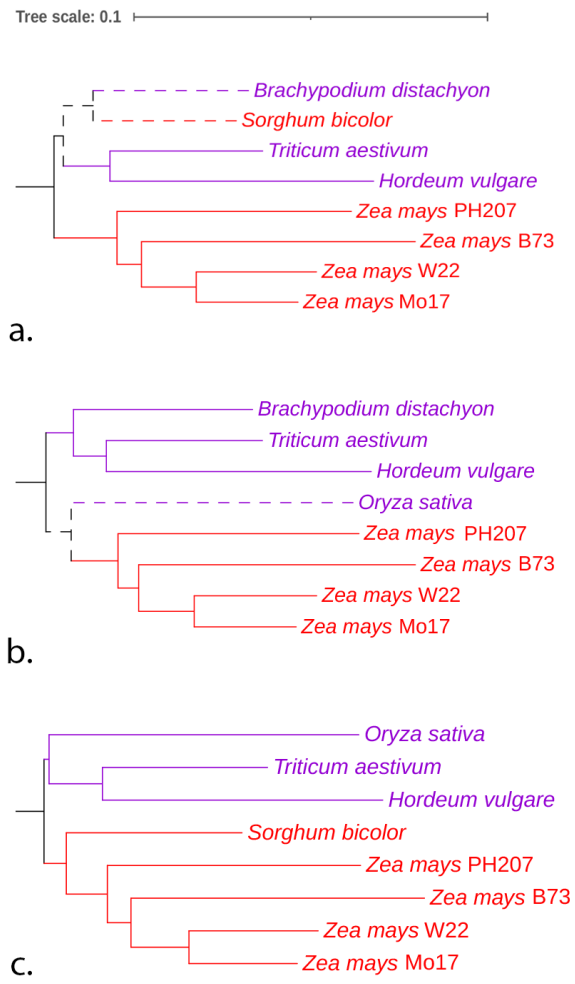


Figure 4: Restoring monocot relationships. Phylograms are colored and rooted as described in Figure 1. Dashed lines mark deviations from known phylogenetic relationships. Monocot topology changes with removal of a single species: a) *O. sativa*, b) *S. bicolor*, and c) *B. distachyon*. Tree scale is shown above.

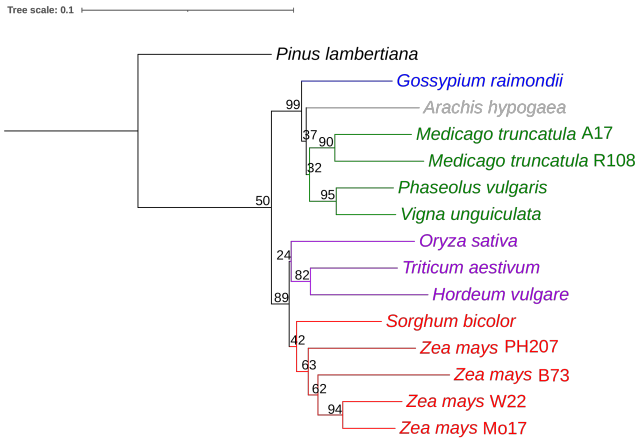


Figure 5: Restoring known phylogentic relationships to the NJ tree via removal of a minimal number of species. Phylograms are colored and rooted as described in Figure 1. Node values represent the jackknifing support values derived by removing 40% of GO terms in the dataset. 3 genomes have been removed: *C. sativa*, *G. max*, and *B. distachyon*. Tree scale is shown above.

anticipated to cause similarities for species with those. shared traits.

For the species noted in this paper as misleading tree-building based on whole-genome gene function, we find that the most likely cause of discrepancies between phylogenetic trees and those inferred from gene functions is the simplest: differences in quality of input sequence assemblies and gene structure annotations. The particular differences are our current focus for investigation and are anticipated to result in updates and additions to this manuscript.

In conclusion, we have demonstrated that the GOMAP system produces datasets can be used together for comparative functional genomics analyses if the datasets are derived from comparably high-quality assemblies and gene annotations. We look forward not only to developing systems to support comparative functional genomics tools, but also to seeing the tools other research groups will develop to approach these datasets for use to formulate useful comparative functional genomics hypotheses.

IV. METHODS

A. Acquiring Input Datasets

For each of the 18 genomes listed in Table I, information on how to access input annotation products are listed in the DOI shown. For each, one representative translated peptide sequence per protein coding gene was selected and used as the input for GOMAP. Unless the authors of the genome provided a set of representative sequences designated as canonical, we chose the longest translated peptide sequence as the representative for each gene model. Non-IUPAC characters and trailing asterisks (*) were removed from the sequences, and headers were simplified to contain only non-special characters. The corresponding script for each dataset can be found at the respective DOI that required these changes. Based on this input, GOMAP yielded a functional annotation set spanning all protein-coding genes in the genome. Using the Gene Ontology version releases/2020-10-09, this functional annotation set was cleaned up by removing duplicates, annotations with modifiers, and obsolete GO terms. Any terms containing alternative identifiers were merged to their respective main identifier, which uncovered a few additional duplicates, which were also removed. Table IV shows the number of annotations removed from each dataset produced.

To compare the quality of GOMAP predictions to currently available functional predictions

	A	B	C	D	E	F	G	H	I	J	K
1		MaizeMo17_	MaizeB73_v	MaizeW22_	Sorghum_bic	Rice_japonic	Arabidopsis_	Brassica_rap	Brachypodiu	Setaria_itali	Glycine_max_
2	Number of scaffolds	2560	596	306	4773	15	7	1103	35	337	17185
3	Total size of scaffolds	2182615441	2134339606	2133868603	675363888	382778125	119668634	353140194	270739461	405868399	955377461
4	Total scaffold length as percent	99.2097928	97.0154366	96.9940274	92.5156011	89.0181686	88.6434326	79.7156194	76.2646369	82.8302855	86.8524965
5	useful amount of scaffold seque	2166421525	2134248774	2132523330	658598248	382775990	119668634	349517073	270688160	403931411	903991371
6	% of estimated genome that is	98.4737057	97.0113079	96.9328786	90.2189381	89.0176721	88.6434326	78.8977591	76.2501859	82.4349818	82.1810337
7	Longest scaffold	32176138	39317442	83688764	14037432	45064769	30427671	45156810	37944854	58970307	1208265
8	Shortest scaffold	1007	5568	711	76	2135	154478	2324	1361	1005	53
9	Number of scaffolds > 1K nt	2560	596	305	4443	15	7	1103	35	337	15209
10	Number of scaffolds > 10K nt	2216	591	291	2349	14	7	1089	32	86	8373
11	Number of scaffolds > 100K nt	475	366	130	958	14	7	99	28	15	3233
12	Number of scaffolds > 1M nt	304	296	97	96	12	5	10	25	9	3
13	Number of scaffolds > 10M nt	69	69	62	4	12	5	10	8	9	0
14	N50	10204498	10679169	35520101	1310030	30828668	23459830	28928902	21988513	47252588	182839
15	L50	69	62	19	78	6	3	6	5	4	1548
16	NG50	9989738	10214929	33636442	952806	30828668	23459830	28493056	14353953	42145699	155995
17	LG50	70	66	20	102	6	3	7	7	5	1976
18	%A	26.1620236	26.1751525	26.1136293	28.0624282	27.4791688	31.9409245	31.3947797	26.7967177	26.6161279	32.6149569
19	%C	23.0431074	23.0885848	22.9215831	21.9374973	21.2048113	18.0091535	18.3175258	23.2102519	22.7891655	17.3815369
20	%G	23.0346257	23.1036049	22.9355355	21.9440298	21.2106374	17.9907293	18.2998537	23.191945	22.8044083	17.3804671
21	%T	26.1511802	26.1942841	26.125969	28.0560447	27.4769294	31.9035905	31.3996908	26.8010853	26.6017409	32.6230391
22	Total Number of Ns	35119661	30699779	40613559	0	10060957	185738	2076994	0	4823979	0
23	%N	1.60906316	1.43837367	1.90328303	0	2.62840438	0.15521026	0.58814999	0	1.18855743	0

Figure 6: Assembly quality table directly from GenomeQC. Note well that the species and genome versions here are not guaranteed to agree with previous datasets in this analysis.

	A	B	C	D	E	F	G	H	I	J	K	L
1		MaizeMo17_C	MaizeB73_v	MaizeW22_	Sorghum_bicolor_N	Rice_japonica_	Arabidopsis_	Brassica_rapa_	Brachypodium_	Setaria_italica_	Glycine_max_v2.1_con	
2	Number of gene models (bp)	38620	39498	40691	31705	30540	33467	46250	30002	30125	54881	
3	Minimum gene length (bp)	21	111	87	22	21	3	150	58	54	22	
4	Maximum gene length (bp)	146217	128402	154495	137963	93056	27265	26363	59987	111270	161416	
5	Average gene length (bp)	4076	4173	4330	3942.2	3232.7	2043.2	1859.4	3644.1	3554.6	4001.8	
6	Number of exons	272323	1209198	313830	296871	138233	324691	0	297600	254939	568844	
7	Average number of exons per gene model	7	30	7	9.4	4.5	9.7	0	9.9	8.5	10.4	
8	Average exon length (bp)	297	284	292	341	367.2	307.3	0	346.7	332.4	295.1	
9	Number of transcripts	46530	133786	51717	42297	29548	53886	46250	41064	37633	75844	
10	Average number of transcripts per gene model	1	3	1	1.3	1	1.6	1	1.4	1.2	1.4	
11	Number of gene models less than 200bp length	1225	37	64	595	862	1847	605	1302	594	3251	

Figure 7: Structural annotation table directly from GenomeQC. Note well that the species and genome versions here are not guaranteed to agree with previous datasets in this analysis.

from Gramene and Phytozome, we downloaded IEA annotations from Gramene (version 63, [47], https://www.gramene.org/) and Phytozome (version 12, [15], https://phytozome.jgi.doe.gov/) for each species with functional annotations of the same genome version. These datasets were cleansed of duplications and redundancies. Similarly cleaned non-IEA annotations from Gramene served as the Gold Standard wherever they were available. More detailed information on how we accessed these datasets can be found at https://github.com/Dill-PICL/GOMAP-Paper-2019.1/blob/master/data/go_annotation_sets/README.md.

B. Quantitative and Qualitative Evaluation

The number of annotations in each cleansed dataset was determined and related to the number of protein coding genes (based on transcripts in the input FASTA file). This was done for separately for each GO aspect as well as in total (see Table II).

The ADS software version published in [38] is available from https://bitbucket.org/plyusnin/ads/. We used version b6309cb (also included in our code as a submodule) to calculate SimGIC2, TC-AUCPCR, and F_{max} quality scores. To provide the information content required for the SimGIC2 metric, the Arabidopsis GOA from https://www.ebi.ac.uk/GOA/arabidopsis_release was used in version 2021-02-16.

C. Cladogram Construction

For clustering, we first collected all GO terms annotated to any gene in each genome into a list and removed the duplicates, yielding a one-dimensional set of GO terms for each genome (T). Next, we added all ancestor terms for each term in this set T and once again removed the duplicates, yielding a set S. These sets with added ancestors served as a starting point of our

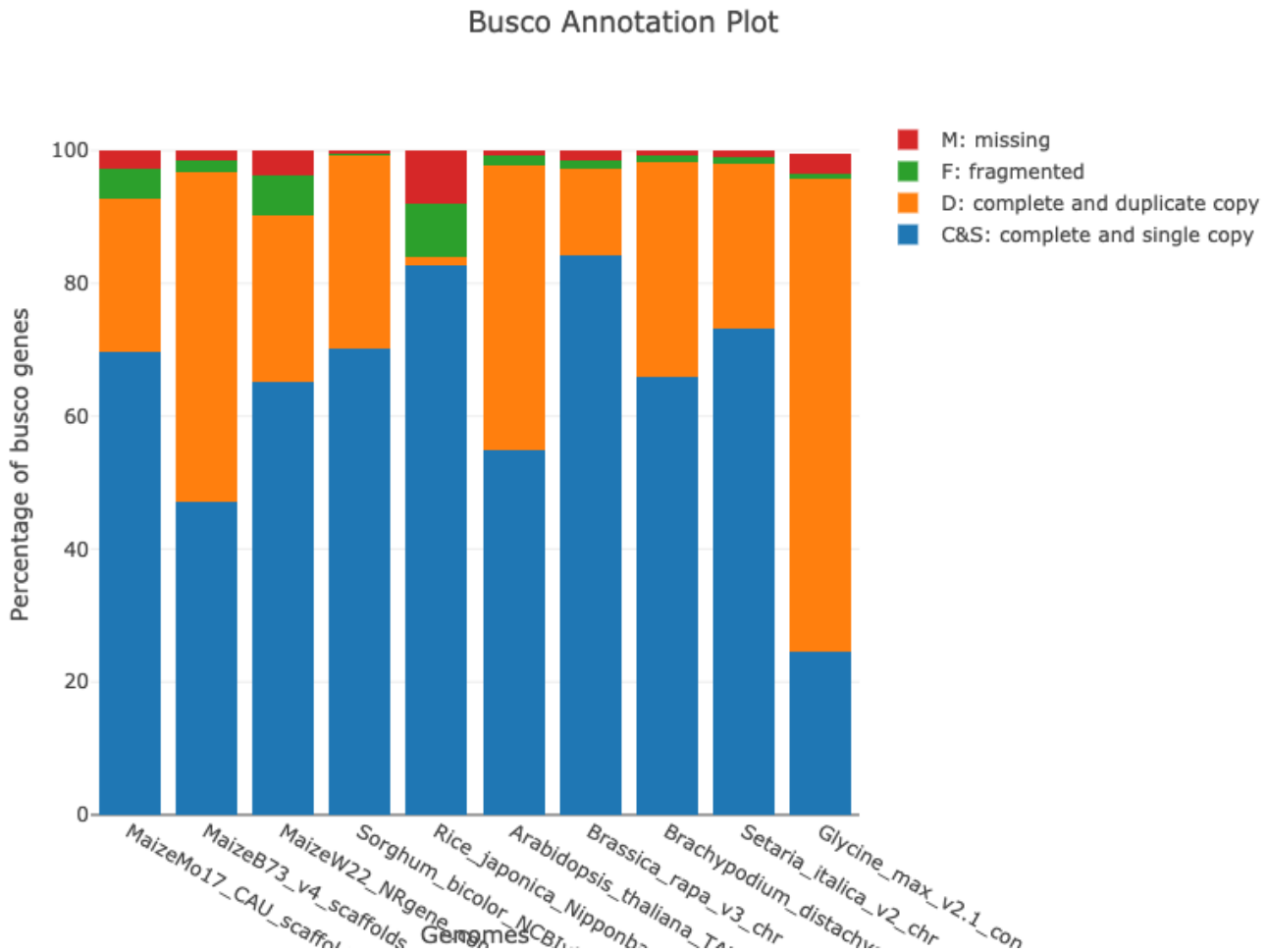


Figure 8: BUSCO plot directly from GenomeQC. Note well that the species and genome versions here are not guaranteed to agree with previous datasets in this analysis. Complete and single-copy genes are shown in blue, complete and duplicte copies in orange, fragmented copies in green, and missing copies in red.

tree-building analyses: pairwise distances between the genomes were calculated using the Jaccard distance as a metric of the dissimilarity between any two sets a and b.

$$d_{ab} = 1 - \frac{|S_a \cap S_b|}{|S_a \cup S_b|} \tag{1}$$

Then a neighbor-joining tree was constructed on the pairwise distance matrix using PHYLIP [12]. Additionally, term sets S of all genomes were combined into a binary matrix (with rows corresponding to genomes and columns corresponding to GO terms, values of 0 or 1 indicating whether a term is present or absent in the given set). PHYLIP pars was used to construct a parsimony tree from this binary matrix.

P. lambertiana, a gymnosperm, was included in the dataset as an outgroup to the angiosperms to separate between the monocot and dicot plants. iTOL [35] was used to visualize the trees using their Newick format, and root them at *P. lambertiana*. Moreover, a cladogram representing the known phylogeny of the included taxa was created by hand based on known evolutionary relationships [17, 5, 16, 3, 42]. This was used to compare the generated phylogenetic relationship based on functional similarity with the evolutionary relationships of the plant genomes.

Jackknifing analysis was carried out for both parsimony and neighbor-joining trees to assess the support for each clade based on the proportion of jackknife trees showing the same clade. To this end, 40% of the terms in T were randomly removed, ancestors of the remaining terms were added and trees constructed as above. The majority rule consensus tree of 100 individual trees was calculated with the jackknife values represented on each branch. The tree was then visualized using iTOL using its Newick format, and rooted again at *P. lambertiana*.

V. AVAILABILITY OF SOURCE CODE AND SUPPORTING DATA

All used data and source code are freely available at https://github.com/Dill-PICL/GOMAP-Paper-2019.1 under the terms of the MIT license. All software requirements and dependencies are packaged into a Singularity container so no other setup is required. We will provide a DOI through Zenodo for the final version of the manuscript after reviews and corrections are incorporated.

An up-to-date list of all available annotation sets can be found at https://dill-picl.org/projects/gomap/gomap-datasets/.

VI. DECLARATIONS

A. Competing Interests

The author(s) declare that they have no competing interests.

B. Funding

This work has been supported by the Iowa State University Plant Sciences Institute Faculty Scholars Program to CJLD, the Predictive Plant Phenomics NSF Research Traineeship (#DGE-1545453) to CJLD (CFY is a trainee), and IOW04714 Hatch funding to Iowa State University.

C. Author's Contributions

LF, DP, DFY, KC, HAD, PJ, DCS, HV, and KW generated annotations for plants as described in this paper. DP and CJLD co-conceived the idea for phylogenetic analysis. LF worked with DP to create dendrograms and compare those to phylogenetic trees. LF carried out assembly and annotation metric comparisons. LF, DP, and CJLD wrote the manuscript. All authors read, offered suggestions to improve, and approved the final copy of the manuscript.

VII. ACKNOWLEDGEMENTS

Thanks to Steven Cannon for help to understand phylogenetic relationships among dicots and helpful discussions and to Toby Kellogg for discussions and ideas on how to consider the data. Thanks to Nancy Manchanda for reviewing documentation and for checking genome versions used as input for GenomeQC. Thanks to Darwin Campbell who guided the deposition of datasets with CyVerse for release and DOI assignment.

VIII. AUTHORS' INFORMATION

KW created the GOMAP system during his time as a graduate student at Iowa State University. LF, DP, CFY, KC, and PJ are currently graduate students. HAD and DCS are undergraduate students who annotated grape and canola, respectively. Each graduate and undergraduate student annotated at least one genome over the course of a research rotation lasting no more than one semester.

References

[1] URL: https://www.biofunctionprediction.org.

[2] Michael Ashburner et al. "Gene Ontology: tool for the unification of biology". In: *Nature genetics* 25.1 (May 2000), pp. 25–29. ISSN: 1061-4036. DOI: 10.1038/75556. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3037419/ (visited on 03/12/2021).

[3] Nasim Azani et al. "A new subfamily classification of the Leguminosae based on a taxonomically comprehensive phylogeny: The Legume Phylogeny Working Group (LPWG)". en. In: *TAXON* 66.1 (2017). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.12705/661.3, pp. 44–77. ISSN: 1996-8175. DOI: https://doi.org/10.12705/661.3. URL: https://onlinelibrary.wiley.com/doi/abs/10.12705/661.3 (visited on 03/12/2021).

[4] D. Binns et al. "QuickGO: a web-based tool for Gene Ontology searching". In: *Bioinformatics* 25.22 (Sept. 2009), pp. 3045–3046. DOI: 10.1093/bioinformatics/btp536. URL: https://doi.org/10.1093/bioinformatics/btp536.

[5] Steven B. Cannon and Randy C. Shoemaker. "Evolutionary and comparative analyses of the soybean genome". eng. In: *Breeding Science* 61.5 (Jan. 2012), pp. 437–444. ISSN: 1344-7610. DOI: 10.1270/jsbbs.61.437.

[6] Carolyn Lawrence Dill. *GOMAP Wheat Reference Sequences 1.1*. 2019. DOI: 10.25739/65KF-JZ20. URL: http://datacommons.cyverse.org/browse/iplant/home/shared/commons_repo/curated/Carolyn_Lawrence_Dill_GOMAP_Wheat_RefSeq1.1_HC_December_2018.r1.

[7] Ana Conesa and Stefan Götz. "Blast2GO: A Comprehensive Suite for Functional Analysis in Plant Genomics". en. In: *International Journal of Plant Genomics* 2008 (Apr. 2008), pp. 1–12. ISSN: 1687-5370, 1687-5389. DOI: 10.1155/2008/619832. URL: https://www.hindawi.com/journals/ijpg/2008/619832/ (visited on 03/12/2021).

[8] Michael Defoin-Platel et al. "AIGO: Towards a unified framework for the Analysis and the Inter-comparison of GO functional annotations". In: *BMC Bioinformatics* (2011). ISSN: 14712105. DOI: 10.1186/1471-2105-12-431. URL: https://doi.org/10.1186/1471-2105-12-431.

[9] Haley Dostalik and Carolyn Lawrence-Dill. *Carolyn_Lawrence_Dill_GOMAP_Grape_Genoscope_12x_January_202* 2021. DOI: 10.25739/JTFK-Q888. URL: http://datacommons.cyverse.org/browse/iplant/home/shared/commons_repo/curated/Carolyn_Lawrence_Dill_GOMAP_Grape_Genoscope_12x_January_2021.r1.

[10] Leila Fattel and Carolyn Lawrence-Dill. *Carolyn_Lawrence_Dill_GOMAP_Banana_NCBI_ASM31385v2_February_2021.v1*. 2021. DOI: 10.25739/YT7W-GS55. URL: http://datacommons.cyverse.org/browse/iplant/home/shared/commons_repo/curated/Carolyn_Lawrence_Dill_GOMAP_Banana_NCBI_ASM31385v2_February_2021.r1.

[11] Leila Fattel and Carolyn Lawrence-Dill. *Carolyn_Lawrence_Dill_GOMAP_Cacao_NCBI_CriolloV2_March_2021.v1*. 2021. DOI: 10.25739/9QC0-N310. URL: http://datacommons.cyverse.org/browse/iplant/home/shared/commons_repo/curated/Carolyn_Lawrence_Dill_GOMAP_Cacao_NCBI_CriolloV2_March_2021.r1.

[12] Joseph Felsenstein. *PHYLIP (phylogeny inference package), version 3.5 c*. Joseph Felsenstein., 1993.

[13] Walter M. Fitch. "Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology". In: *Systematic Zoology* 20.4 (Dec. 1971), p. 406. DOI: 10.2307/2412116. URL: https://doi.org/10.2307/2412116.

[14] Shan Gao et al. "A high-quality reference genome of wild Cannabis sativa". en. In: *Horticulture Research* 7.1 (May 2020). Number: 1 Publisher: Nature Publishing Group, pp. 1–11. ISSN: 2052-7276. DOI: 10.1038/s41438-020-0295-3. URL: https://www.nature.com/articles/s41438-020-0295-3 (visited on 03/12/2021).

[15] David M. Goodstein et al. "Phytozome: a comparative platform for green plant genomics". In: *Nucleic Acids Research* 40.D1 (Nov. 2011), pp. D1178–D1186. DOI: 10.1093/nar/gkr944. URL: https://doi.org/10.1093/nar/gkr944.

[16] Candice N. Hansey et al. "Maize (Zea mays L.) genome diversity as revealed by RNA-sequencing". eng. In: *PloS One* 7.3 (2012), e33071. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0033071.

[17] Elizabeth A. Kellogg. "Evolutionary History of the Grasses". en. In: *Plant Physiology* 125.3 (Mar. 2001), pp. 1198–1205. ISSN: 0032-0889, 1532-2548. DOI: 10.1104/pp.125.3.1198. URL: https://academic.oup.com/plphys/article/125/3/1198-1205/6109905 (visited on 03/12/2021).

[18] Carolyn Lawrence-Dill. *GOMAP Barley Reference Sequences $IBSC_PGSB_r1$*. 2019. DOI: 10.25739/ZVGV-8E37. URL: http://datacommons.cyverse.org/browse/iplant/home/shared/commons_repo/curated/Carolyn_Lawrence_Dill_GOMAP_Barley_IBSC_PGSB-1.0_May_2019.r1.

[19] Carolyn Lawrence-Dill. *GOMAP Barrel Clover $A17_HM341Mt4.0v2$*. 2019. DOI: 10.25739/PY38-YB08. URL: http://datacommons.cyverse.org/browse/iplant/home/shared/commons_repo/curated/Carolyn_Lawrence_Dill_GOMAP_BarrelClover_A17_HM341_Mt4.0v2_August_2019.r1.

[20] Carolyn Lawrence-Dill. *GOMAP Barrel Clover $R108_HM340v1.0$*. 2019. DOI: 10.25739/2SQC-J140. URL: http://datacommons.cyverse.org/browse/iplant/home/shared/commons_repo/curated/Carolyn_Lawrence_Dill_GOMAP_BarrelClover_R108_HM340_v1.0_August_2019.r1.

[21] Carolyn Lawrence-Dill. *GOMAP Bdistachyon.Bd21.v3.1.r1*. 2019. DOI: 10.25739/DW2T-3G82. URL: http://datacommons.cyverse.org/browse/iplant/home/shared/commons_repo/curated/Carolyn_Lawrence_Dill_GOMAP_Bdistachyon.Bd21.v3.1_November_2019.r1.

[22] Carolyn Lawrence-Dill. *GOMAP Common Bean DOE-JGI and USDA-NIFA v2.0*. 2019. DOI: 10.25739/1YWE-EW96. URL: http://datacommons.cyverse.org/browse/iplant/home/shared/commons_repo/curated/Carolyn_Lawrence_Dill_GOMAP_CommonBean_DOE-JGI-USDA-NIFA.2.0_August_2019.r1.

[23] Carolyn Lawrence-Dill. *GOMAP Cowpea IT97K-499-35 JGI annotation v1.1*. 2019. DOI: 10.25739/CDX9-WR97. URL: http://datacommons.cyverse.org/browse/iplant/home/shared/commons_repo/curated/Carolyn_Lawrence_Dill_GOMAP_Cowpea_JGI.v1.1_August_2019.r1.

[24] Carolyn Lawrence-Dill. *GOMAP Gossypium raimondii JGI v2.1*. 2020. DOI: 10.25739/A13T-ZH47. URL: http://datacommons.cyverse.org/browse/iplant/home/shared/commons_repo/curated/Carolyn_Lawrence_Dill_GOMAP_Gossypium_raimondii_JGI_v2.1_January_2020.r1.

[25] Carolyn Lawrence-Dill. *GOMAP Maize Zm-Mo17-REFERENCE-CAU-1.0 Zm00014a.1*. 2019. DOI: 10.25739/M634-CN58. URL: http://datacommons.cyverse.org/browse/iplant/home/shared/commons_repo/curated/Carolyn_Lawrence_Dill_GOMAP_maize.Mo17.AGPv1_April_2019.r1.

[26] Carolyn Lawrence-Dill. *GOMAP Maize Zm-PH207-$REFERENCE_NS$ − $UIUC_UMN$ − 1.0$Zm00008a.1$*. 2019. DOI: 10.25739/DM9S-AA15. URL: http://datacommons.cyverse.org/browse/iplant/home/shared/commons_repo/curated/Carolyn_Lawrence_Dill_GOMAP_maize.PH207.UIUC_UMN-1.0_April_2019.r1.

[27] Carolyn Lawrence-Dill. *GOMAP Maize Zm-W22-REFERENCE-NRGENE-2.0 Zm00004b.1*. 2019. DOI: 10.25739/E4VA-9F09. URL: http://datacommons.cyverse.org/browse/iplant/home/shared/commons_repo/curated/Carolyn_Lawrence_Dill_GOMAP_maize.W22.AGPv2_April_2019.r1.

[28] Carolyn Lawrence-Dill. *GOMAP Peanut IPGI 1.0*. 2019. DOI: 10.25739/CHAB-0E35. URL: http://datacommons.cyverse.org/browse/iplant/home/shared/commons_repo/curated/Carolyn_Lawrence_Dill_GOMAP_Peanut_Tifrunner.IPGI.1.0_August_2019.r1.

[29] Carolyn Lawrence-Dill. *GOMAP Rice Reference Sequences 2.0*. 2019. DOI: 10.25739/53G0-J859. URL: http://datacommons.cyverse.org/browse/iplant/home/shared/commons_repo/curated/Carolyn_Lawrence_Dill_GOMAP_Rice_IRGSP-1.0_April_2019.r2.

[30] Carolyn Lawrence-Dill. *GOMAP Sbicolor.BTx623.v3.0.1.r1*. 2019. DOI: 10.25739/4TY0-YE98. URL: http://datacommons.cyverse.org/browse/iplant/home/shared/commons_repo/curated/Carolyn_Lawrence_Dill_GOMAP_Sbicolor.BTx623.v3.0.1_November_2019.r1.

[31] Carolyn Lawrence-Dill. *GOMAP Soybean JGI-Wm82.a4.v1*. 2019. DOI: 10.25739/59EC-1719. URL: http://datacommons.cyverse.org/browse/iplant/home/shared/commons_repo/curated/Carolyn_Lawrence_Dill_GOMAP_Soybean_JGI-Wm82.a4.v1_April_2019.r1.

[32] Carolyn Lawrence-Dill. *GOMAP TreeGenesDB sugar pine assembly v1.5*. 2020. DOI: 10.25739/JVS4-XR88. URL: http://datacommons.cyverse.org/browse/iplant/home/shared/commons_repo/curated/Carolyn_Lawrence_Dill_GOMAP_SugarPine_TreeGenesDB-1.5_January_2020.r1.

[33] Carolyn Lawrence-Dill. $GOMAP_{C}annabis_{s}ativa_{N}CBI - cs10_{J}anuary_{2}020$. 2020. DOI: 10.25739/AB9Z-2Z86. URL: http://datacommons.cyverse.org/browse/iplant/home/shared/commons_repo/curated/Carolyn_Lawrence_Dill_GOMAP_Cannabis_NCBI-cs10_January_2020.r1.

[34] Carolyn Lawrence-Dill. *maize-GAMER Annotaions for maize.B73.AGPv4.r1*. 2017. DOI: 10.7946/P2M925. URL: http://datacommons.cyverse.org/browse/iplant/home/shared/commons_repo/curated/Carolyn_Lawrence-Dill_maize-GAMER_maize.B73_RefGen_v4_Zm00001d.2_Oct_2017.r1.

[35] Ivica Letunic and Peer Bork. "Interactive Tree Of Life (iTOL) v4: recent updates and new developments". en. In: *Nucleic Acids Research* 47.W1 (July 2019), W256–W259. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkz239. URL: https://academic.oup.com/nar/article/47/W1/W256/5424068 (visited on 03/12/2021).

[36] Ivica Letunic and Peer Bork. "Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation". In: *Nucleic Acids Research* (Apr. 2021). DOI: 10.1093/nar/gkab301. URL: https://doi.org/10.1093/nar/gkab301.

[37] Nancy Manchanda et al. "GenomeQC: a quality assessment tool for genome assemblies and gene structure annotations". In: *BMC Genomics* 21.1 (Mar. 2020), p. 193. ISSN: 1471-2164. DOI: 10.1186/s12864-020-6568-2. URL: https://doi.org/10.1186/s12864-020-6568-2 (visited on 03/12/2021).

[38] Ilya Plyusnin, Liisa Holm, and Petri Törönen. "Novel comparison of evaluation metrics for gene ontology classifiers reveals drastic performance differences". In: *PLOS Computational Biology* 15.11 (Nov. 2019). Ed. by Predrag Radivojac, e1007419. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1007419. URL: http://dx.plos.org/10.1371/journal.pcbi.1007419.

[39] S. C. Potter. "The Ensembl Analysis Pipeline". In: *Genome Research* 14.5 (May 2004), pp. 934–941. DOI: 10.1101/gr.1859804. URL: https://doi.org/10.1101/gr.1859804.

[40] Dennis Psaroudakis and Carolyn Lawrence-Dill. *Carolyn_Lawrence_Dill_GOMAP_Solanum_lycopersicum_ITAG4.1.v1_April_2021.r1*. 2021. DOI: 10.25739/ZH2V-4P15. URL: http://datacommons.cyverse.org/browse/iplant/home/shared/commons_repo/curated/Carolyn_Lawrence_Dill_GOMAP_Solanum_lycopersicum_ITAG4.1.v1_April_2021.r1.

[41] Dennis Psaroudakis and Carolyn Lawrence-Dill. *Carolyn_Lawrence_Dill_GOMAP_Solanum_pennellii_Bolger2014.v1_April_2021.r1*. 2021. DOI: 10.25739/FHR4-CX67. URL: http://datacommons.cyverse.org/browse/iplant/home/shared/commons_repo/curated/Carolyn_Lawrence_Dill_GOMAP_Solanum_pennellii_Bolger2014.v1_April_2021.r1.

[42] Mark N. Puttick et al. "The Interrelationships of Land Plants and the Nature of the Ancestral Embryophyte". English. In: *Current Biology* 28.5 (Mar. 2018). Publisher: Elsevier, 733–745.e2. ISSN: 0960-9822. DOI: 10.1016/j.cub.2018.01.063. URL: https://www.cell.com/current-biology/abstract/S0960-9822(18)30096-4 (visited on 03/12/2021).

[43] Seung Yon Rhee and Marek Mutwil. "Towards revealing the functions of all genes in plants". en. In: *Trends in Plant Science* 19.4 (Apr. 2014), pp. 212–221. ISSN: 1360-1385. DOI: 10.1016/j.tplants.2013.10.006. URL: https://www.sciencedirect.com/science/article/pii/S1360138513002343 (visited on 03/12/2021).

[44] N Saitou and M Nei. "The neighbor-joining method: a new method for reconstructing phylogenetic trees." In: *Molecular Biology and Evolution* (July 1987). DOI: 10.1093/oxfordjournals.molbev.a040454. URL: https://doi.org/10.1093/oxfordjournals.molbev.a040454.

[45] Koichi Sakamoto et al. "Characterization Genome Sizes and Morphology of Sex Chromosomes in Hemp (Cannabis sativa L.)." In: *CYTOLOGIA* 63.4 (1998), pp. 459–464. DOI: 10.1508/cytologia.63.459. URL: https://doi.org/10.1508/cytologia.63.459.

[46] Felipe A. Simão et al. "BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs". en. In: *Bioinformatics* 31.19 (Oct. 2015), pp. 3210–3212. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btv351. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btv351 (visited on 03/12/2021).

[47] Marcela K Tello-Ruiz et al. "Gramene 2021: harnessing the power of comparative genomics and pathways for plant research". In: *Nucleic Acids Research* 49.D1 (Nov. 2020), pp. D1452–D1463. DOI: 10.1093/nar/gkaa979. URL: https://doi.org/10.1093/nar/gkaa979.

[48] The Gene Ontology Consortium et al. "The Gene Ontology resource: enriching a GOld mine". en. In: *Nucleic Acids Research* 49.D1 (Jan. 2021), pp. D325–D334. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkaa1113. URL: https://academic.oup.com/nar/article/49/D1/D325/6027811 (visited on 03/12/2021).

[49] Iris Tzafrir et al. "Identification of Genes Required for Embryo Development in Arabidopsis". en. In: *Plant Physiology* 135.3 (July 2004). Publisher: American Society of Plant Biologists Section: GENOME ANALYSIS, pp. 1206–1220. ISSN: 0032-0889, 1532-2548. DOI: 10.1104/pp.104.045179. URL: http://www.plantphysiol.org/content/135/3/1206 (visited on 03/12/2021).

[50] Karin Verspoor et al. "A categorization approach to automated ontological function annotation". In: *Protein Science* 15.6 (2006). ISSN: 09618368. DOI: 10.1110/ps.062184006. URL: https://dx.doi.org/10.1110%7B%5C%%7D2Fps.062184006.

[51] Kokulapalan Wimalanathan and Carolyn J. Lawrence-Dill. "Gene Ontology Meta Annotator for Plants (GOMAP)". In: *bioRxiv* (2021). DOI: 10.1101/809988. eprint: https://www.biorxiv.org/content/early/2021/02/25/809988.full.pdf. URL: https://www.biorxiv.org/content/early/2021/02/25/809988.

[52] Kokulapalan Wimalanathan et al. "Maize GO Annotation-Methods, Evaluation, and Review (maize-GAMER)". In: *Plant Direct* 2.4 (Apr. 2018), e00052. ISSN: 24754455. DOI: 10.1002/pld3.52. URL: http://doi.wiley.com/10.1002/pld3.52.

[53] C. F. J. Wu. "Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis". In: *The Annals of Statistics* 14.4 (Dec. 1986). DOI: 10.1214/aos/

1176350142. URL: https://doi.org/10.1214/aos/1176350142.

[54] Naihui Zhou et al. "The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens". In: *Genome Biology* 20.1 (Nov. 2019). DOI: 10.1186/s13059-019-1835-8. URL: https://doi.org/10.1186/s13059-019-1835-8.

[55] Chengsheng Zhu et al. "Functional Basis of Microorganism Classification". en. In: *PLOS Computational Biology* 11.8 (Aug. 2015). Publisher: Public Library of Science, e1004472. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1004472. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004472 (visited on 03/12/2021).

Table II: Quantitative metrics of the cleaned functional annotation sets. CC, MF, BP, and A refer to the aspects of the Gene Ontology: Cellular Component, Molecular Function, Biological Process, and Any/All. GOMAP covers all genomes with at least one annotation per gene and provides substantially more annotations than Gramene63 or Phytozome, especially in the BP aspect.

			Genes Annotated[%][a]				Annotations[b]				Median Ann. per G.[c]			
Genome	Genes	Dataset	CC	MF	BP	**A**	CC	MF	BP	**A**	CC	MF	BP	**A**
Arachis hypogaea	67,124	GOMAP	85.85	84.68	100.00	**100.00**	150,525	132,144	493,145	**775,814**	2	2	6	**10**
Brachypodium distachyon	34,310	GOMAP	81.33	85.35	100.00	**100.00**	74,172	69,213	255,397	**398,782**	2	2	6	**10**
		GoldStandard	21.54	19.53	18.20	**26.66**	10,985	10,436	11,120	**32,673**	1	1	1	**3**
		Gramene63-IEA	33.12	49.29	38.29	**63.60**	21,658	36,372	23,899	**82,026**	1	1	1	**3**
		Phytozome12	10.25	37.21	26.86	**43.11**	4,186	18,597	11,070	**34,060**	0	1	1	**2**
Cannabis sativa	33,677	GOMAP	94.22	95.48	100.00	**100.00**	85,755	73,614	262,741	**422,110**	2	2	6	**11**
Glycine max	52,872	GOMAP	86.95	88.92	100.00	**100.00**	126,470	113,068	416,989	**656,527**	2	2	6	**11**
Gossypium raimondii	37,505	GOMAP	93.00	92.37	100.00	**100.00**	95,419	84,910	307,470	**487,799**	2	2	6	**11**
Hordeum vulgare	39,734	GOMAP	88.57	91.76	100.00	**100.00**	86,489	79,727	272,420	**438,636**	2	2	5	**10**
		GoldStandard	28.23	26.30	23.43	**35.64**	15,734	15,391	15,267	**46,414**	1	1	1	**3**
		Gramene63-IEA	36.19	50.90	41.71	**65.03**	29,826	44,789	29,425	**104,178**	1	1	1	**3**
Medicago truncatula A17	50,444	GOMAP	83.79	86.69	100.00	**100.00**	104,902	99,155	363,608	**567,665**	2	2	6	**10**
		GoldStandard	25.45	23.26	21.51	**32.12**	17,938	18,416	18,461	**54,827**	1	1	1	**3**
		Gramene63-IEA	34.25	50.84	40.26	**66.14**	32,753	63,470	40,441	**137,001**	1	1	1	**3**
		Phytozome12	8.87	36.05	25.83	**41.07**	5,315	25,950	15,576	**47,098**	0	1	1	**2**
Medicago truncatula R108	55,706	GOMAP	72.10	90.14	100.00	**100.00**	108,388	107,499	381,831	**597,718**	1	2	5	**9**
Oryza sativa	35,825	GOMAP	79.78	83.31	100.00	**100.00**	71,306	64,150	248,304	**383,760**	2	2	6	**9**
		GoldStandard	29.95	27.29	25.33	**37.57**	15,492	15,176	16,536	**47,339**	1	1	1	**3**
		Gramene63-IEA	32.21	45.83	36.75	**60.13**	21,935	37,425	24,255	**83,645**	1	1	1	**3**
		Phytozome12	10.31	40.10	29.18	**46.09**	4,361	20,842	12,451	**37,884**	0	1	1	**2**
Phaseolus vulgaris	27,433	GOMAP	94.48	93.06	100.00	**100.00**	70,987	64,022	229,230	**364,239**	2	2	6	**11**
Pinus lambertiana	31,007	GOMAP	92.67	95.91	100.00	**100.00**	71,247	68,315	212,248	**351,810**	2	2	5	**10**
Sorghum bicolor	34,129	GOMAP	82.44	85.98	100.00	**100.00**	75,145	69,659	259,004	**403,808**	2	2	6	**10**
		GoldStandard	34.48	32.91	30.90	**42.84**	16,837	17,614	17,850	**52,593**	1	1	1	**3**
		Gramene63-IEA	35.91	52.11	42.36	**67.41**	23,608	39,418	27,074	**90,313**	1	1	1	**3**
		Phytozome12	10.54	39.19	27.90	**45.10**	4,246	19,724	11,432	**35,599**	0	1	1	**2**
Triticum aestivum	107,891	GOMAP	88.53	90.98	100.00	**100.00**	259,318	217,467	785,051	**1,261,836**	2	2	6	**10**
		GoldStandard	2.98	2.78	2.56	**3.82**	4,727	4,512	4,793	**14,035**	1	1	1	**3**
		Gramene63-IEA	29.12	58.62	38.72	**70.41**	47,595	111,889	62,977	**222,721**	0	1	1	**2**
Vigna unguiculata	29,773	GOMAP	91.21	91.08	100.00	**100.00**	74,791	67,734	242,847	**385,372**	2	2	6	**11**
		Phytozome12	13.91	45.68	34.14	**53.06**	5,107	19,962	12,209	**37,534**	0	1	1	**2**
Zea mays B73.v4	39,324	GOMAP	93.16	94.92	100.00	**100.00**	87,648	81,665	278,305	**447,618**	2	2	6	**10**
		GoldStandard	37.92	34.78	32.67	**46.85**	22,531	21,292	23,153	**67,285**	1	1	1	**3**
		Gramene63-IEA	39.16	58.16	48.21	**73.87**	30,189	53,748	35,276	**119,273**	1	1	1	**3**
Zea mays Mo17	38,620	GOMAP	86.98	90.87	100.00	**100.00**	86,074	78,650	277,395	**442,119**	2	2	6	**10**
		GoldStandard	27.56	25.20	23.73	**33.98**	16,128	15,384	16,489	**48,220**	1	1	1	**3**
Zea mays PH207	40,557	GOMAP	86.55	90.61	100.00	**100.00**	88,962	84,910	288,208	**462,080**	2	2	6	**10**
		GoldStandard	28.18	25.82	24.26	**34.66**	17,370	16,580	17,791	**51,984**	1	1	1	**3**
Zea mays W22	40,690	GOMAP	90.77	92.58	100.00	**100.00**	93,622	84,450	289,364	**467,436**	2	2	6	**10**
		GoldStandard	25.40	23.15	21.80	**31.29**	15,518	14,818	15,850	**46,402**	1	1	1	**3**

Download this table (CSV)

[a] How many genes in the genome have at least one GO term from the CC, MF, BP aspect annotated to them? A = How many at least one from any aspect? (A = CC ∪ MF ∪ BP)

[b] How many annotations in the CC, MF, and BP aspect does this dataset contain? A = How many in total? A = CC + MF + BP

[c] Take a typical gene that is present in the annotation set. How many annotations does it have in each aspect? A = How many in total? Please note that A ≠ CC + MF + BP

Table III: Qualitative metrics of functional annotation sets predicted by GOMAP, Gramene, and Phytozome.

		SimGIC2			TC-AUCPCR			Fmax		
Genome	Dataset	CC	MF	BP	CC	MF	BP	CC	MF	BP
Brachypodium distachyon	GOMAP	0.404149	0.464127	0.223830	0.233442	0.230701	0.118526	0.741361	0.740897	0.526881
	Gramene63-IEA	0.317801	0.420859	0.349406	0.129163	0.192507	0.111361	0.691016	0.738542	0.650325
	Phytozome12	0.370264	0.370521	0.352206	0.112582	0.136832	0.085628	0.717759	0.697076	0.660603
Hordeum vulgare	GOMAP	0.400087	0.470012	0.238177	0.237231	0.261399	0.130784	0.745272	0.750213	0.560096
	Gramene63-IEA	0.306119	0.426601	0.381010	0.157352	0.228797	0.136002	0.680996	0.742638	0.665696
Medicago truncatula A17	GOMAP	0.371795	0.451258	0.213407	0.272809	0.282650	0.139032	0.730838	0.726991	0.531406
	Gramene63-IEA	0.329600	0.437274	0.343561	0.176497	0.265887	0.133503	0.701093	0.749900	0.654297
	Phytozome12	0.358311	0.367257	0.363013	0.144247	0.170863	0.110386	0.717307	0.698429	0.661233
Oryza sativa	GOMAP	0.408945	0.482650	0.248207	0.298502	0.303384	0.159724	0.751121	0.757181	0.559221
	Gramene63-IEA	0.328761	0.423191	0.341193	0.167619	0.265410	0.135451	0.711309	0.738732	0.643827
	Phytozome12	0.049975	0.041007	0.044279	0.000003	0.000003	0.000002	0.470134	0.266628	0.239256
Sorghum bicolor	GOMAP	0.404852	0.466708	0.224011	0.316873	0.337380	0.169883	0.746540	0.742001	0.534258
	Gramene63-IEA	0.323037	0.400241	0.353135	0.177038	0.260198	0.154157	0.711107	0.712170	0.653591
	Phytozome12	0.356091	0.348264	0.340124	0.151947	0.177579	0.110483	0.715714	0.675147	0.641535
Triticum aestivum	GOMAP	0.410582	0.489881	0.229271	0.050762	0.030610	0.019360	0.736476	0.762420	0.533897
	Gramene63-IEA	0.362452	0.476685	0.395112	0.040992	0.043701	0.027872	0.737769	0.762059	0.670953
Zea mays B73.v4	GOMAP	0.417455	0.467339	0.245373	0.302761	0.290371	0.153011	0.759504	0.746870	0.564707
	Gramene63-IEA	0.303231	0.416301	0.346308	0.175735	0.250075	0.138275	0.662987	0.732860	0.647725
Zea mays Mo17	GOMAP	0.399521	0.464265	0.225632	0.236209	0.239598	0.125599	0.744360	0.743026	0.537489
Zea mays PH207	GOMAP	0.394481	0.436266	0.224226	0.221709	0.221266	0.117086	0.743111	0.718933	0.533092
Zea mays W22	GOMAP	0.397602	0.463499	0.223511	0.210198	0.217609	0.113262	0.743783	0.742341	0.535572

Download this table (CSV)

IX. SUPPLEMENTARY

Table IV: Number of removed annotations during cleanup.

Genome	Dataset	Obsolete Annotations	Duplicates	Annotations with Modifiers
Arachis hypogaea	GOMAP	3437	13	912
Brachypodium distachyon	GOMAP	2512	49	789
	GoldStandard	21	204	0
	Gramene63-IEA	166	114	0
	Phytozome12	99	18	0
Cannabis sativa	GOMAP	1714	6	757
Glycine max	GOMAP	3333	10	930
Gossypium raimondii	GOMAP	1781	7	822
Hordeum vulgare	GOMAP	1877	8	815
	GoldStandard	1	9	0
	Gramene63-IEA	282	147	0
Medicago truncatula A17	GOMAP	2673	10	798
	GoldStandard	2	23	0
	Gramene62-IEA	429	251	0
	Gramene63-IEA	309	243	0
	Phytozome12	132	17	0
Medicago truncatula R108	GOMAP	4168	7	803
Oryza sativa	GOMAP	1642	7	869
	GoldStandard	37	833	0
	Gramene61-IEA	242	28	0
	Gramene63-IEA	238	64	0
	Phytozome12	119	19	0
Phaseolus vulgaris	GOMAP	1190	6	783
Pinus lambertiana	GOMAP	1839	4	587
Sorghum bicolor	GOMAP	2384	66	783
	GoldStandard	178	219	0
	Gramene63-IEA	278	198	0
	Phytozome12	131	12	0
Triticum aestivum	GOMAP	9624	17	1132
	GoldStandard	1	5	0
	Gramene61-IEA	706	88	0
	Gramene63-IEA	584	319	0
Vigna unguiculata	GOMAP	1269	6	811
	Phytozome12	122	27	0
Zea mays B73.v4	GOMAP	2077	89	848
	GoldStandard	50	633	0
	Gramene63-IEA	306	140	0
Zea mays Mo17	GOMAP	2346	83	823
	GoldStandard	36	1489	0
Zea mays PH207	GOMAP	2676	82	830
	GoldStandard	37	2702	0
Zea mays W22	GOMAP	2681	88	840
	GoldStandard	30	499	0

Download this table (CSV)

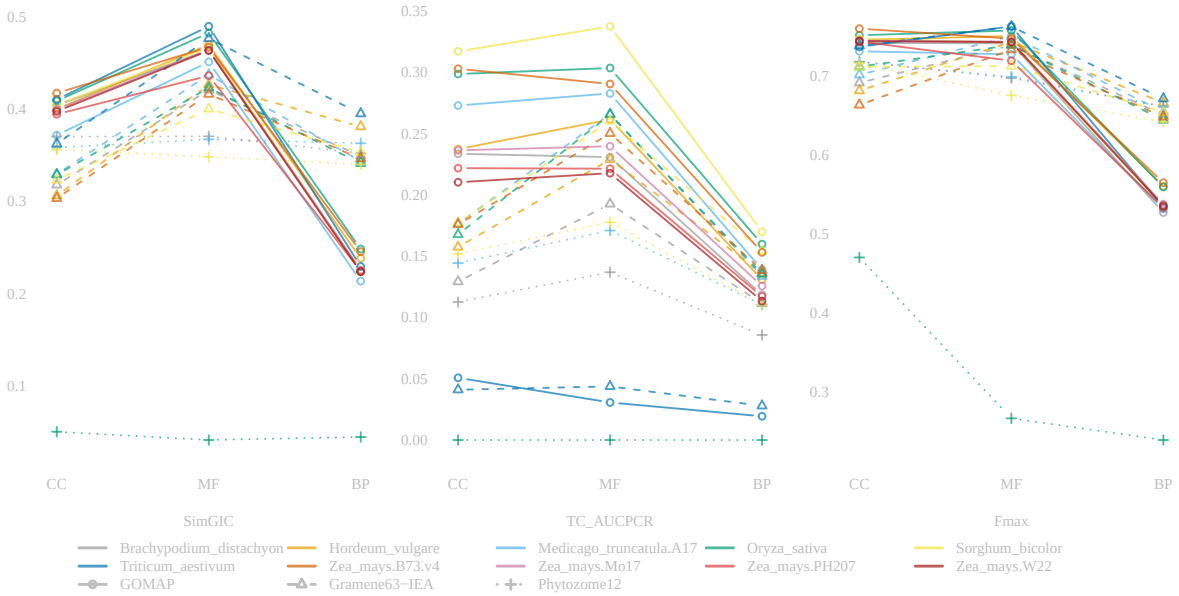
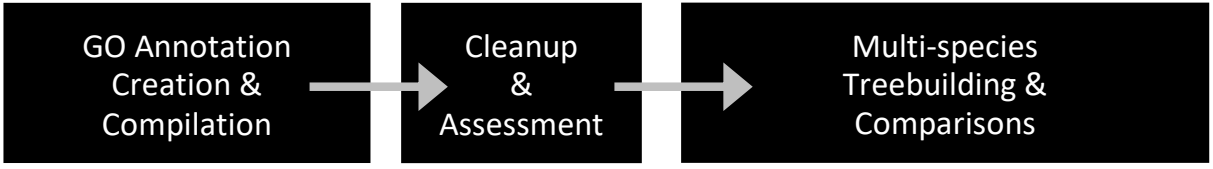
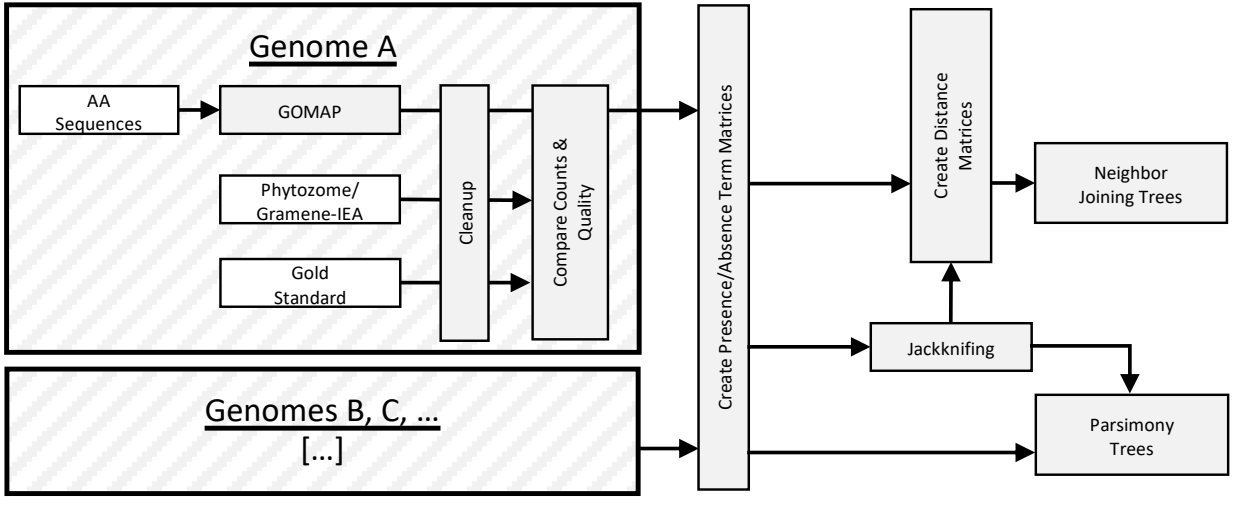


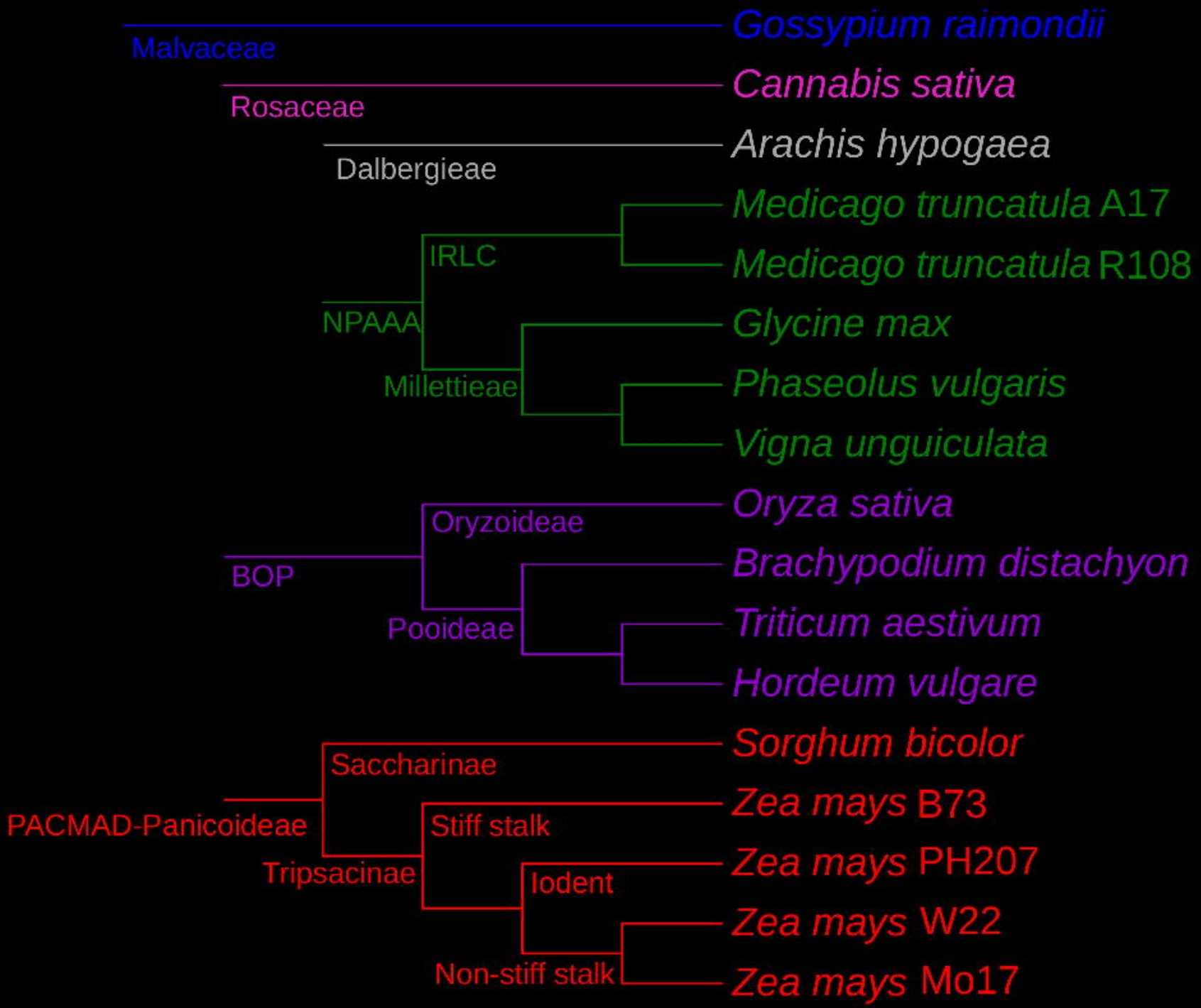
Figure 9: Quality scores of the predicted annotation visualized. Line and symbol color designate the genome while line and symbol type designate prediction method/dataset. Please note that scales are different for each metric (values are not directly comparable).

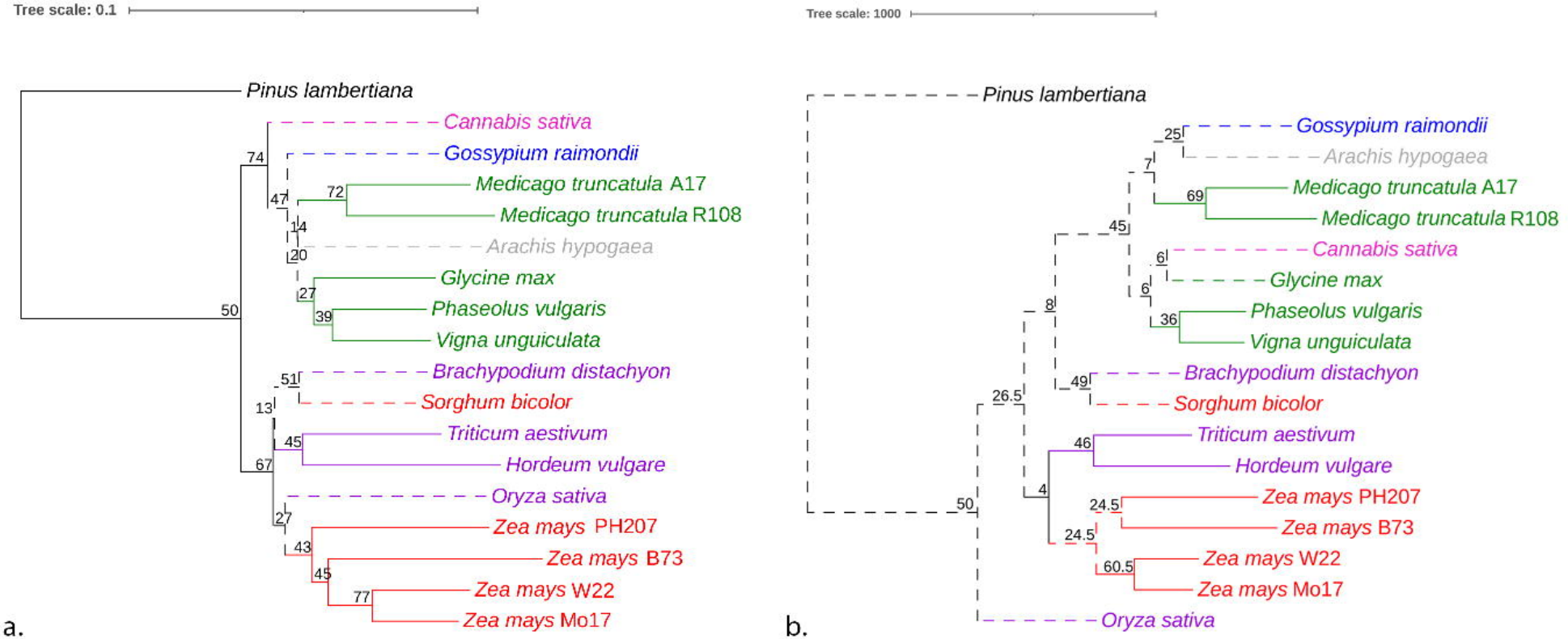
a. Workflow Overview

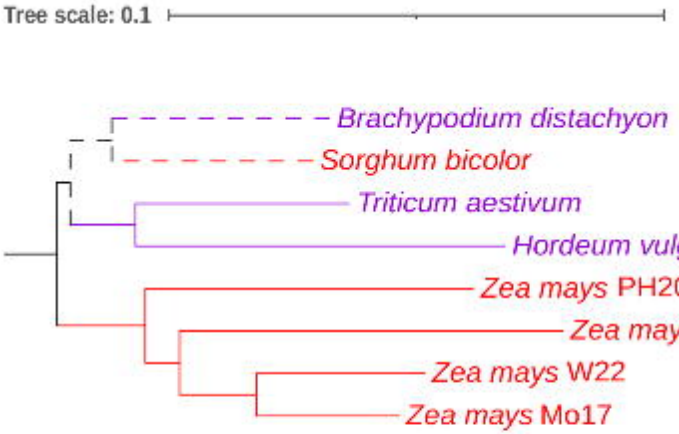


b. Workflow Detail

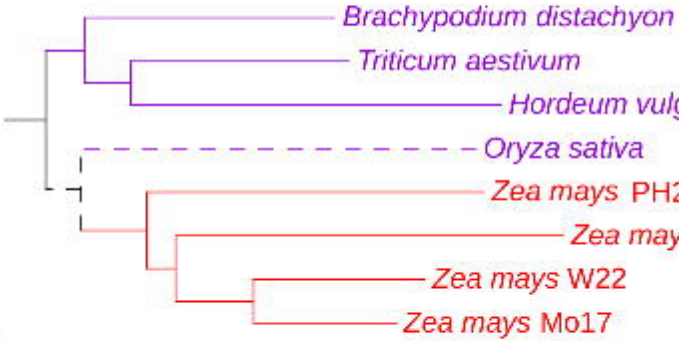



Malvaceae
Gossypium raimondii
Rosaceae
Cannabis sativa
Dalbergieae
Arachis hypogaea
IRLC
Medicago truncatula A17
Medicago truncatula R108
NPAAA
Glycine max
Millettieae
Phaseolus vulgaris
Vigna unguiculata
Oryzoideae
Oryza sativa
BOP
Brachypodium distachyon
Pooideae
Triticum aestivum
Hordeum vulgare
Saccharinae
Sorghum bicolor
PACMAD-Panicoideae
Zea mays B73
Stiff stalk
Tripsacinae
Zea mays PH207
Iodent
Zea mays W22
Non-stiff stalk
Zea mays Mo17

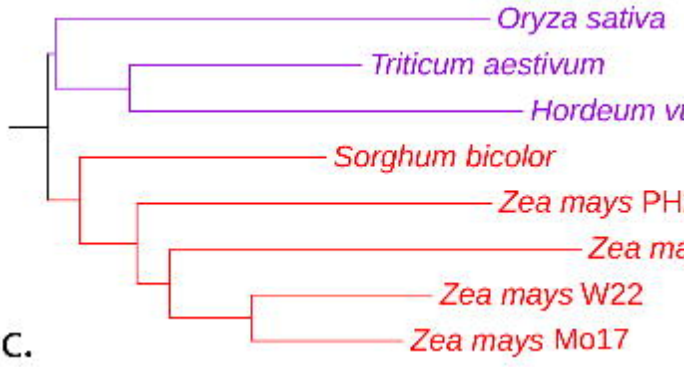

Tree scale: 0.1
Pinus lambertiana
Cannabis sativa
Gossypium raimondii
Medicago truncatula A17
Medicago truncatula R108
Arachis hypogaea
Glycine max
Phaseolus vulgaris
Vigna unguiculata
Brachypodium distachyon
Sorghum bicolor
Triticum aestivum
Hordeum vulgare
Oryza sativa
Zea mays PH207
Zea mays B73
Zea mays W22
Zea mays Mo17
74
47
72
14
20
27
39
50
51
13
45
67
27
43
45
77
a.
Tree scale: 1000
Pinus lambertiana
Gossypium raimondii
Arachis hypogaea
Medicago truncatula A17
Medicago truncatula R108
Cannabis sativa
Glycine max
Phaseolus vulgaris
Vigna unguiculata
Brachypodium distachyon
Sorghum bicolor
Triticum aestivum
Hordeum vulgare
Zea mays PH207
Zea mays B73
Zea mays W22
Zea mays Mo17
Oryza sativa
25
7
69
45
6
6
36
8
49
26.5
46
4
24.5
24.5
60.5
50
b.


Tree scale: 0.1
Brachypodium distachyon
Sorghum bicolor
Triticum aestivum
Hordeum vulgare
Zea mays PH207
Zea mays B73
Zea mays W22
Zea mays Mo17

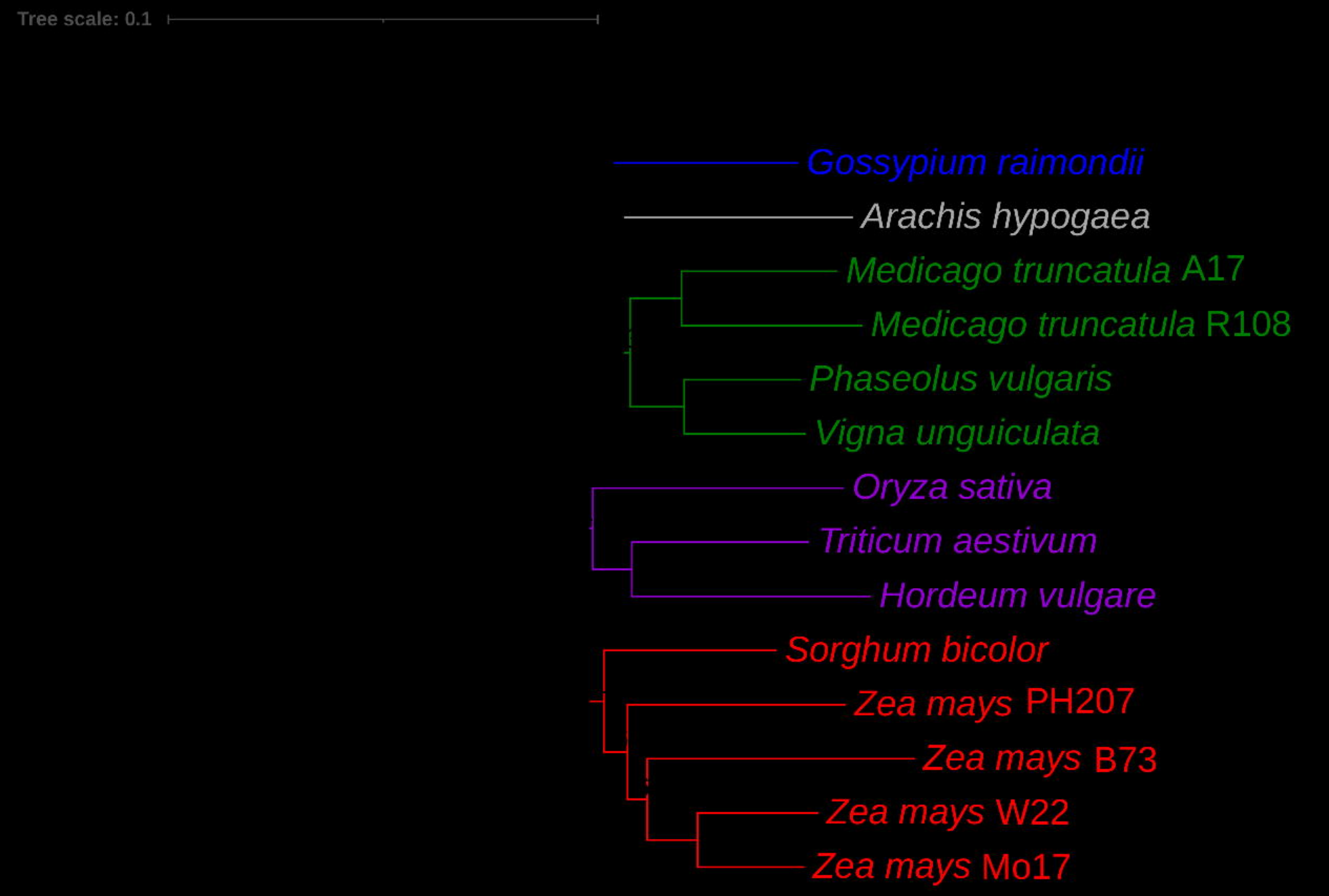
a.


Brachypodium distachyon
Triticum aestivum
Hordeum vulgare
Oryza sativa
Zea mays PH207
Zea mays B73
Zea mays W22
Zea mays Mo17

b.

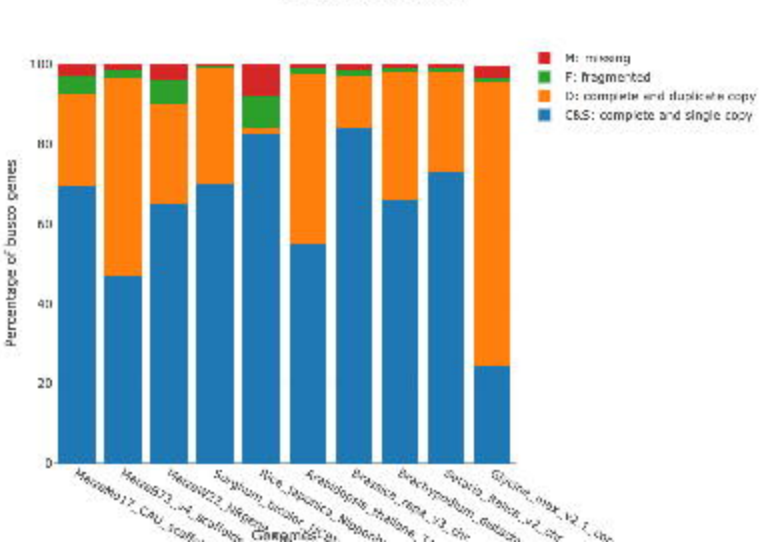

Oryza sativa
Triticum aestivum
Hordeum vulgare
Sorghum bicolor
Zea mays PH207
Zea mays B73
Zea mays W22
Zea mays Mo17

c.


Tree scale: 0.1
Gossypium raimondii
Arachis hypogaea
Medicago truncatula A17
Medicago truncatula R108
Phaseolus vulgaris
Vigna unguiculata
Oryza sativa
Triticum aestivum
Hordeum vulgare
Sorghum bicolor
Zea mays PH207
Zea mays B73
Zea mays W22
Zea mays Mo17

	A	B	C	D	E	F	G	H	I	J	K
1		MaizeMo17_	MaizeB73_v	MaizeW22_	Sorghum_bic	Rice_japonic	Arabidopsis_	Brassica_rap	Brachypodiu	Setaria_itali	Glycine_max_
2	Number of scaffolds	2560	596	306	4773	15	7	1103	35	337	17185
3	Total size of scaffolds	2182615441	2134339606	2133868603	675363888	382778125	119668634	353140194	270739461	405868399	955377461
4	Total scaffold length as percent	99.2097928	97.0154366	96.9940274	92.5156011	89.0181686	88.6434326	79.7156194	76.2646369	82.8302855	86.8524965
5	useful amount of scaffold seque	2166421525	2134248774	2132523330	658598248	382775990	119668634	349517073	270688160	403931411	903991371
6	% of estimated genome that is	98.4737057	97.0113079	96.9328786	90.2189381	89.0176721	88.6434326	78.8977591	76.2501859	82.4349818	82.1810337
7	Longest scaffold	32176138	39317442	83688764	14037432	45064769	30427671	45156810	37944854	58970307	1208265
8	Shortest scaffold	1007	5568	711	76	2135	154478	2324	1361	1005	53
9	Number of scaffolds > 1K nt	2560	596	305	4443	15	7	1103	35	337	15209
10	Number of scaffolds > 10K nt	2216	591	291	2349	14	7	1089	32	86	8373
11	Number of scaffolds > 100K nt	475	366	130	958	14	7	99	28	15	3233
12	Number of scaffolds > 1M nt	304	296	97	96	12	5	10	25	9	3
13	Number of scaffolds > 10M nt	69	69	62	4	12	5	10	8	9	0
14	N50	10204498	10679169	35520101	1310030	30828668	23459830	28928902	21988513	47252588	182839
15	L50	69	62	19	78	6	3	6	5	4	1548
16	NG50	9989738	10214929	33636442	952806	30828668	23459830	28493056	14353953	42145699	155995
17	LG50	70	66	20	102	6	3	7	7	5	1976
18	%A	26.1620236	26.1751525	26.1136293	28.0624282	27.4791688	31.9409245	31.3947797	26.7967177	26.6161279	32.6149569
19	%C	23.0431074	23.0885848	22.9215831	21.9374973	21.2048113	18.0091535	18.3175258	23.2102519	22.7891655	17.3815369
20	%G	23.0346257	23.1036049	22.9355355	21.9440298	21.2106374	17.9907293	18.2998537	23.191945	22.8044083	17.3804671
21	%T	26.1511802	26.1942841	26.125969	28.0560447	27.4769294	31.9035905	31.3996908	26.8010853	26.6017409	32.6230391
22	Total Number of Ns	35119661	30699779	40613559	0	10060957	185738	2076994	0	4823979	0
23	%N	1.60906316	1.43837367	1.90328303	0	2.62840438	0.15521026	0.58814999	0	1.18855743	0

	A	B	C	D	E	F	G	H	I	J	K	L
1		MaizeMo17_C	MaizeB73_v	MaizeW22_I	Sorghum_bicolor_N	Rice_japonica_I	Arabidopsis_	Brassica_rapa_	Brachypodium_	Setaria_italica_	Glycine_max_v2.1_con	
2	Number of gene models (bp)	38620	39498	40691	31705	30540	33467	46250	30002	30125	54881	
3	Minimum gene length (bp)	21	111	87	22	21	3	150	58	54	22	
4	Maximum gene length (bp)	146217	128402	154495	137963	93056	27265	26363	59987	111270	161416	
5	Average gene length (bp)	4076	4173	4330	3942.2	3232.7	2043.2	1859.4	3644.1	3554.6	4001.8	
6	Number of exons	272323	1209198	313830	296871	138233	324691	0	297600	254939	568844	
7	Average number of exons per gene model	7	30	7	9.4	4.5	9.7	0	9.9	8.5	10.4	
8	Average exon length (bp)	297	284	292	341	367.2	307.3	0	346.7	332.4	295.1	
9	Number of transcripts	46530	133786	51717	42297	29548	53886	46250	41064	37633	75844	
10	Average number of transcripts per gene model	1	3	1	1.3	1	1.6	1	1.4	1.2	1.4	
11	Number of gene models less than 200bp length	1225	37	64	595	862	1847	605	1302	594	3251	
12												


Busco Annotation Plot
M: missing
F: fragmented
D: complete and duplicated copy
CBS: complete and single copy
Percentage of busco genes
100
80
60
40
20
0
Maize_Mo17_CAU_1.0
Maize_B73_v4_scaffolds
Maize_PH207_1.0_scaffolds
Sorghum_bicolor_NCBIv3
Rice_Japonica_Nipponbare_IRGSP1.0
Arabidopsis_thaliana_TAIR10
Brassica_rapa_v3.0
Brachypodium_distachyon_v3.0
Setaria_italica_v2.0
Glycine_max_v2.1