

Article Title

Dennis Psaroudakis¹, Ha Vu¹, Colleen Yanarella¹, Steven Cannon¹, Darwin Campbell¹, Parnal Joshi¹, Iddo Friedberg^{1,4}, Kokulapalan Wimalanathan^{1,2}, Carolyn J. Lawrence-Dill^{1,2,3*}

¹ Bioinformatics and Computational Biology, Iowa State University, Ames, IA, USA

² Department of Genetics, Development, and Cell Biology, Iowa State University, Ames, IA, USA

³ Department of Agronomy, Iowa State University, Ames, IA, USA

⁴ Department of Veterinary Microbiology, Iowa State University, Ames, IA, USA

Correspondence*:

Carolyn J. Lawrence-Dill

triffid@iastate.edu

2 ABSTRACT

Abstract length and content varies depending on article type. Refer to <http://www.frontiersin.org/about/AuthorGuidelines> for abstract requirement and length according to article type.

Keywords: Text Text Text Text Text Text Text Text

1 INTRODUCTION

Hello, how are we doing?

2 METHODS

8 2.1 Generating Predictions

Used GOMAP on condo lalala. Input files are (usually) published along results.

10 2.2 Clean up

All functional annotation sets were cleaned up the following way (using definitions from the Gene Ontology version 2019-07-01):

1. Any annotations where the GO accession was marked as obsolete were removed.
2. Some terms in the GO have 'alternative ids'. When naively removing duplicates, two entries will not be recognized as duplicates if they have different accessions pointing to the same GO term. Therefore, all GO accessions were changed to their respective 'main id' and the dataset was again scanned for duplicates.
3. Any annotations with modifiers (NOT, contributes.to...) were removed since no tool used in the further analysis can handle them.

Table 1 provides information on the number of annotations that were removed this way from each dataset. All further analyses were performed on the cleaned datasets since we assume the user will only be interested in still valid and non-redundant functional annotations.

2.3 Quantitative Evaluation

lalala lololo table xyz

2.4 Quality Evaluation

Quality evaluation of gene function predictions is not trivial and usually done by comparing the set of predicted functions of a gene against a *gold standard* consisting of annotations that are assumed to be correct. We used annotations that were created or in some way curated with human participation for gold standards. There are a plethora of different metrics to perform the comparison of predictions against this gold standard. When we first published GOMAP (Wimalanathan et al., 2018), we used a modified version of the hierarchical evaluation metrics originally introduced in (Verspoor et al., 2006) because they were simple, clear, and part of an earlier attempt at unifying and standardizing GO annotation comparisons (Defoin-Platel et al., 2011). In the meantime, Plyusnin et al. (2019) have published an approach for evaluating different metrics showing substantial differences within the robustness of different approaches. TODO DESCRIBE THEIR APPROACH We have applied their method on the Gold Standards available to us to determine which evaluation metric is the most appropriate in our case. The results of this analysis can be seen in TODO.

We then evaluated our predictions and the other annotation sets using the best performing metrics as well as the one we previously used (Table TODO).

2.5 Phylogenetic Tree Construction

To demonstrate that a more top-level and holistic use of whole-genome functional predictions can still be useful we devised some simple ways of applying phylogenetic methods to our predictions. ### Distance Based ### Character Based

2.6 Ensuring Reproducibility

containerization, github...

3 RESULTS

... a quantitative comparison of the datasets in Table.

3.1 Quality Evaluation

TODO If it turns out that our predictions are good with hF but bad with more appropriate metrics, explanation would be that score thresholds for the prediction tools used in the GOMAP pipeline have been chosen to maximize this hF value. It now seems reasonable to re-adjust these thresholds to maximize a different metric which will likely result in a drop in hF score but increase in other metrics. Again emphasizes the importance of choosing the right evaluation metric. Also shows how comparison between different pipelines/predictions can be difficult if chose different metric or optimized for different metric. Also: if an annotation is not present in the gold standard, there is no way of knowing whether that gene truly doesn't have that function or whether it has just never been characterized/examined. So we cannot distinguish between a biologically true negative and an actually false negative in the gold standard. This

Table 1. Number of removed annotations during cleanup.

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

[Download this table \(CSV\)](#)

57 poses a problem when annotations are predicted that are not found in the gold standard: Is this truly a wrong
58 prediction or is the gold standard incomplete? Especially in our case where the predictions not only contain
59 more annotations than the gold standard, but are also more diverse. In effect this means that a quality

Table 2. Quantitative metrics of the cleaned functional annotation sets. CC, MF, BP, and A refer to the aspects of the GO: Cellular Component, Molecular Function, Biological Process, and Any/All.

Genome	Genes	Dataset	Genes Annotated[%] ^a				Annotations ^b				Median Ann. per G. ^c			
			CC	MF	BP	A	CC	MF	BP	A	CC	MF	BP	A
<i>Arachis hypogaea</i>	67,124	GOMAP	85.85	84.68	100.00	100.00	150,525	132,144	493,145	775,814	2	2	6	10
<i>Brachypodium distachyon</i>	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
<i>Cannabis sativa</i>	33,677	GOMAP	94.22	95.48	100.00	100.00	85,755	73,614	262,741	422,110	2	2	6	11
<i>Glycine max</i>	52,872	GOMAP	86.95	88.92	100.00	100.00	126,470	113,068	416,989	656,527	2	2	6	11
<i>Gossypium raimondii</i>	37,505	GOMAP	93.00	92.37	100.00	100.00	95,419	84,910	307,470	487,799	2	2	6	11
<i>Hordeum vulgare</i>	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
<i>Medicago truncatula</i> 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
<i>Medicago truncatula</i> R108	55,706	GOMAP	72.10	90.14	100.00	100.00	108,388	107,499	381,831	597,718	1	2	5	9
<i>Oryza sativa</i>	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
<i>Phaseolus vulgaris</i>	27,433	GOMAP	94.48	93.06	100.00	100.00	70,987	64,022	229,230	364,239	2	2	6	11
<i>Pinus lambertiana</i>	31,007	GOMAP	92.67	95.91	100.00	100.00	71,247	68,315	212,248	351,810	2	2	5	10
<i>Sorghum bicolor</i>	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
<i>Triticum aestivum</i>	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
<i>Vigna unguiculata</i>	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
<i>Zea mays</i> 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
<i>Zea mays</i> 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
<i>Zea mays</i> 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
<i>Zea mays</i> 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

score as calculated above may not only describe the quality of the prediction, but to some extent also the completeness of the gold standard itself. At least we can see here that gold standards with a median of 3 annotations per gene resulted in higher quality scores than gold standards with less annotations per gene, even though predictions were generated the same way in all cases. TODO maybe put a figure with regression quality score/median annotations per gene or something In conclusion this means that truly making a statement about the quality of a prediction set would require the ideal and complete gold standard. The scores we can generate so far are by far not as meaningful.

REFERENCES

- Defoin-Platel, M., Hindle, M. M., Lysenko, A., Powers, S. J., Habash, D. Z., Rawlings, C. J., et al. (2011). AIGO: Towards a unified framework for the Analysis and the Inter-comparison of GO functional

Table 3. Quality evaluation of the used GO annotation sets.

Genome	Dataset	SimGIC2			TC-AUCPCR			Fmax		
		CC	MF	BP	CC	MF	BP	CC	MF	BP
<i>Brachypodium distachyon</i>	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
<i>Hordeum vulgare</i>	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
<i>Medicago truncatula</i> 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
<i>Oryza sativa</i>	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
<i>Sorghum bicolor</i>	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
<i>Triticum aestivum</i>	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
<i>Zea mays</i> 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
<i>Zea mays</i> Mo17	GOMAP	0.399521	0.464265	0.225632	0.236209	0.239598	0.125599	0.744360	0.743026	0.537489
<i>Zea mays</i> PH207	GOMAP	0.394481	0.436266	0.224226	0.221709	0.221266	0.117086	0.743111	0.718933	0.533092
<i>Zea mays</i> W22	GOMAP	0.397602	0.463499	0.223511	0.210198	0.217609	0.113262	0.743783	0.742341	0.535572

Download this table (CSV)

- 69 annotations. *BMC Bioinformatics* doi:10.1186/1471-2105-12-431
- 70 Plyusnin, I., Holm, L., and Törönen, P. (2019). Novel comparison of evaluation metrics for gene
71 ontology classifiers reveals drastic performance differences. *PLOS Computational Biology* 15, e1007419.
72 doi:10.1371/journal.pcbi.1007419
- 73 Verspoor, K., Cohn, J., Mniszewski, S., and Joslyn, C. (2006). A categorization approach to automated
74 ontological function annotation. *Protein Science* doi:10.1110/ps.062184006
- 75 Wimalanathan, K., Friedberg, I., Andorf, C. M., and Lawrence-Dill, C. J. (2018). Maize GO Annotation-
76 Methods, Evaluation, and Review (maize-GAMER). *Plant Direct* 2, e00052. doi:10.1002/pld3.
77 52