TBD

The Pickles

Febcember 32nd 3023

Methods

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 respecitve 'main id' and the dataset was again scanned for duplicates.

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.

Results

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

Table 1: Number of removed annotations during cleanup.

Genome	Dataset	Obsolete Annotations	Duplicates
Triticum_aestivum	GOMAP	285	0
Hordeum_vulgarum	GOMAP	101	0
Glycine_max	GOMAP	203	0
Arachis_hypogaea	GOMAP	0	0
Zea_mays.PH207	GOMAP	798	76
Zea_mays.Mo17	GOMAP	726	77
Phaseolus_vulgaris	GOMAP	0	0
Vigna_unguiculata	GOMAP	0	0
Medicago_truncatula.R108	GOMAP	0	0
Zea_mays.B73.v3	GOMAP	1107	70
Zea_mays.W22	GOMAP	754	82
Medicago_truncatula.A17	GOMAP	0	0
Zea_mays.B73.v4	GOMAP	752	83
Oryza_sativa	GOMAP	111	2

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

Genome	Genes	Dataset	Annotations[a] C	F	P	A	Annotated Genes [%][b] C	F	P	A	Median Ann. per G.[c] C	F	P	A
Arachis_hypogaea		GOMAP	153433	132944	493799	**780176**	576.67	568.55	671.23	**671.24**	2	2	6	**10**
Glycine_max		GOMAP	129215	113827	417555	**660597**	460.20	470.34	528.71	**528.72**	2	2	6	**11**
Hordeum_vulgarum		GOMAP	88130	80282	272823	**441235**	352.37	364.70	397.33	**397.34**	2	2	5	**10**
Medicago_truncatula.A17		GOMAP	107362	99719	364065	**571146**	423.25	437.36	504.43	**504.44**	2	2	6	**10**
Medicago_truncatula.R108		GOMAP	112343	108031	382322	**602696**	403.32	502.20	557.06	**557.06**	1	2	5	**9**
Oryza_sativa		GOMAP	72780	64685	248700	**386165**	286.19	298.53	358.24	**358.25**	2	2	6	**9**
Phaseolus_vulgaris	100	GOMAP	72005	64583	229630	**366218**	259.34	255.39	274.32	**274.33**	2	2	6	**11**
Triticum_aestivum		GOMAP	267741	218623	785960	**1272324**	956.04	981.87	1078.90	**1078.91**	2	2	6	**10**
Vigna_unguiculata		GOMAP	75867	68313	243278	**387458**	271.73	271.24	297.72	**297.73**	2	2	6	**11**
Zea_mays.B73.v3		GOMAP	135211	87420	291251	**513882**	348.66	380.73	394.68	**394.69**	3	2	6	**11**
Zea_mays.B73.v4		GOMAP	88827	82251	278719	**449797**	367.17	373.37	393.23	**393.24**	2	2	6	**10**
Zea_mays.Mo17		GOMAP	87567	79214	277787	**444568**	336.18	351.05	386.19	**386.20**	2	2	6	**10**
Zea_mays.PH207		GOMAP	90617	85500	288677	**464794**	351.70	367.62	405.56	**405.57**	2	2	6	**10**
Zea_mays.W22		GOMAP	95390	85039	289780	**470209**	369.87	376.85	406.89	**406.90**	2	2	6	**10**

[a] How many annotations in the C, F, and P aspect does this dataset contain? A = How many in total? $A = C + F + P$

[b] How many genes in the genome have at least one GO term from the C, F, P aspect annotated to them? A = How many at least one from any aspect? ($A = C \cup F \cup P$)

[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? Ask your favorite statistician why $A \neq C + F + P$