PAINT Curation Update

Huaiyu Mi
On behalf of all the PAINTers

August 31st, 2015
GO Consortium Meeting, Washington DC.

PAINT Overall Progress

	8/25/2015	10/2014
Total # families	1914	429
Total # sequences	306,293	67,427
Total # sequences with IBA	264,322	59,742

PAINT Curation Update

name	# families	# total seqs	#IBA seqs	# nodes painted	# IBD BP annot	# IBD MF annot	# IBD CC annot
MAF	983	139451	116886	2108	2528	1578	1844
PG	466	78243	60485	1167	1356	747	726
HM	148	33727	27337	417	514	321	286
KRC	101	10396	7413	192	154	106	368
DHL	92	8788	6109	144	150	118	107
CJM	30	8408	5805	105	217	70	92
RAMA	17	5124	3895	50	44	26	32
PDTHOMAS	14	4886	3476	78	48	84	35
CURATOR	10	3991	1459	30	30	27	13
SUZI	7	3792	2943	61	47	45	18
JD	13	3316	2782	96	71	90	38
LN	12	3204	2565	47	54	25	18
MONICA	9	2467	1974	23	40	14	10
RANJANA	1	343	454	1	1	1	1
JBLAKE	11	157	85	8	7	1	6

New GO annotations from PAINT curation

	PAINT annotations		Literature curation	
	genes	annotations	genes	annotations
Human	5118	20720	4227	29472
Mouse	5870	25248	3350	25727
Fly	2992	11762	2000	11624
Worm	3470	14842	2042	9680
Yeast	1286	3825	1685	9420
MOD (12)	44630	190153	21550	125863
Non-MOD (92)	219692	951995	2326	6576

Only a fraction of literature annotations are used

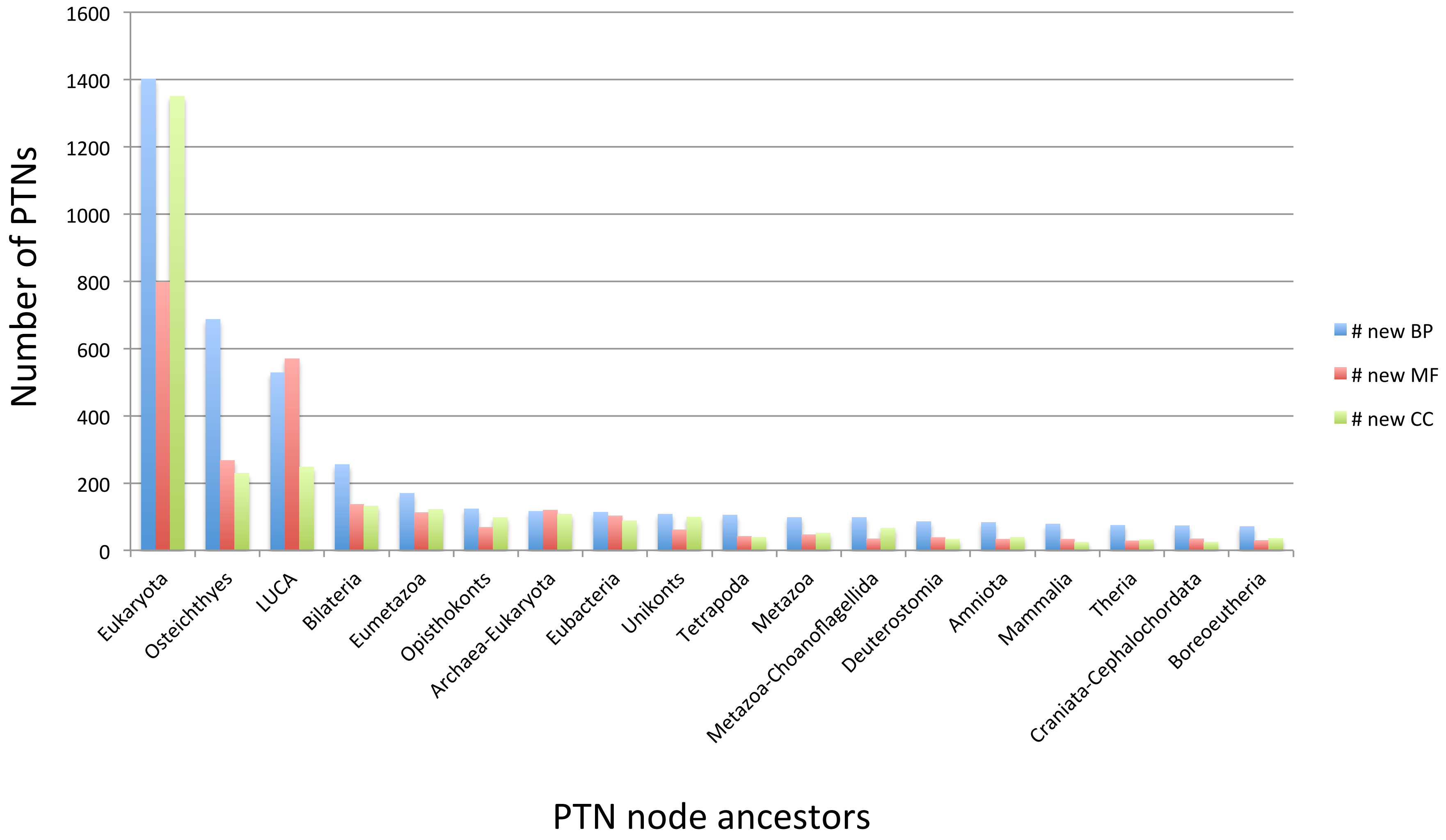
Ontology	Average # GO terms by literature curation/family	Average # GO terms propagated in PAINT/ family
Biological process	26.38	3.9
Cellular component	9.43	2.03
Molecular function	6.34	2.39

Most manually-curated GO terms are not used in PAINT annotations.

Annotations rare propagated

- Biological processes that are **indirectly** controlled by the gene product
  - System level processes, e.g., reproduction, locomotion, development, behavior, etc.
  - Processes controlled by proteins involved in transcription
  - Phenotypes
- Cellular component annotations from high-throughput experiments
- Molecular function – most binding and protein binding

Evolution of new function

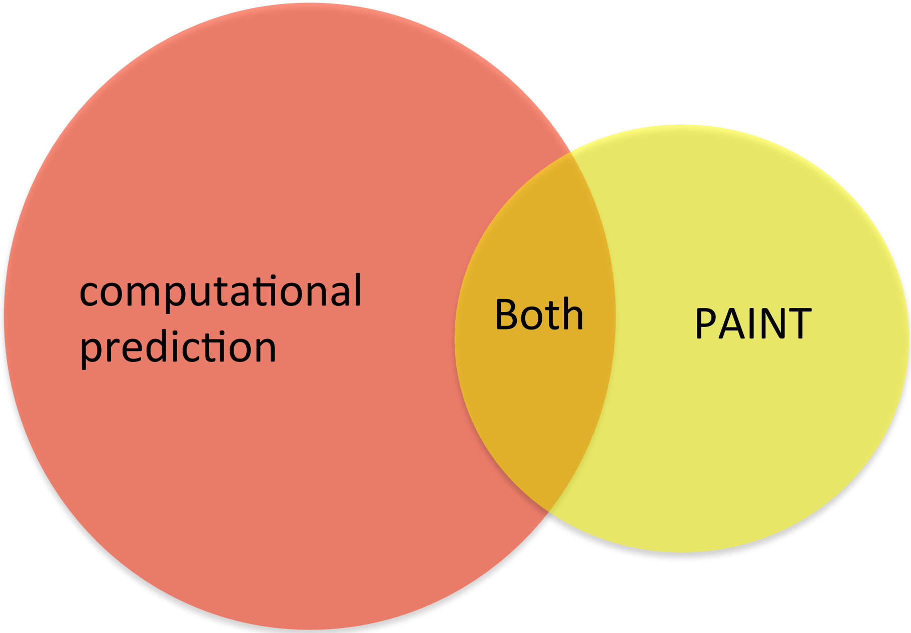


Comparison of phylogenetic annotations vs. computational predictions

	PAINT annotation	Computational prediction	overlap
HUMAN	20,720	69,471	7,567 (36%, 11%)
MOUSE	25,248	64,598	5,531 (22%, 8.5%)
FLY	11,762	9,718	2,623 (22%, 27%)
WORM	14,842	16,855	2,636 (18%, 16%)
YEAST	3,825	12,604	794 (21%, 6.3%)
ECOLI	1,595	6,803	419 (22%, 6.2%)
MODs	190,153	346,165	46,146 (24%, 13%)

If PAINT annotates to a more general term, it is considered as “overlap”.
If PAINT annotates to a more specific term, it is not considered as overlap, because new information is generated by PAINT.

Comparison of phylogenetic annotations vs. computational predictions



- Annotations only predicted in computational annotations
  - Annotations can't be inferred phylogenetically
  - Annotated GO terms that are not usually used in PAINT, such as system processes, behavior terms and protein binding.
  - Annotations are to more general GO terms.
- Annotations only predicted in PAINT
  - Annotations are to more specific GO terms
  - Annotations are to GO terms that are not predicted by computational annotations

Some annotations predicted by computational annotation only are not likely to be predicted in PAINT

	Electronic annotation only	No phylogenetic inference	To a more general GO term	To a term not propagated in PAINT	No inference + unpainted GO terms
HUMAN	61,904	10,899	9,431	13,586	21,476
MOUSE	59,067	11,818	7,448	10,606	20,219
FLY	7,095	1,850	854	1,612	2,864
WORM	14,219	3,321	3,840	2,730	4,795
YEAST	11,810	2,768	3,085	1,623	3,679
ECOLI	6,384	1,542	1,777	1,239	2,310
MOD	300,019	55,524	41,721	63,222	103,743

PAINT only annotations are to more specific GO terms or new GO terms

	PAINT only annotations	More specific than computational annotations	Additional new PAINT annotations
HUMAN	13,153	4,145	9,008
MOUSE	19,717	3,670	16,047
FLY	9,139	1,192	7,947
WORM	12,206	2,830	9,376
YEAST	3,031	879	2,152
ECOLI	5,460	1,377	4,083
MOD	144,007	26,276	117,731

Current Issues

- PANTHER 10.0 release
  - A small number of painted nodes can’t be tracked forwardly
  - A few painted nodes are moved to a different family (family merge or split)
  - A small number of families need to be reviewed
  - Some proteins are not in the UniProt reference proteome build, so they are missing from Panther10
  - Some proteins that were in families in Panther 9 are orphans in Panther 10 (less than 20)
  - These will be looked at and fixed; also, used as learning experience for next Panther release

On its way

- PAINT touchup script

  To automatically fix some annotation errors in PAINT:

  - Taxon constraints not respected
  - Lack of primary literature (if annotations are deleted after the PAINT annotation)
  - Flag when new literature annotations are added
  - Update of obsolete GO terms when a ‘replace by’ exists
  - Currently working to make touchup do what we expect