Assessing annotation consistency in the Gene Ontology

Dolan ME, Ni L, Camon E, Blake JA. A procedure for assessing GO annotation consistency. *Bioinformatics* 2005 Jun 1;21 Suppl 1:i136-i143. PMID: 15961450

SILS Biomedical Informatics Journal Club
http://ils.unc.edu/bioinfo/
2005-10-04

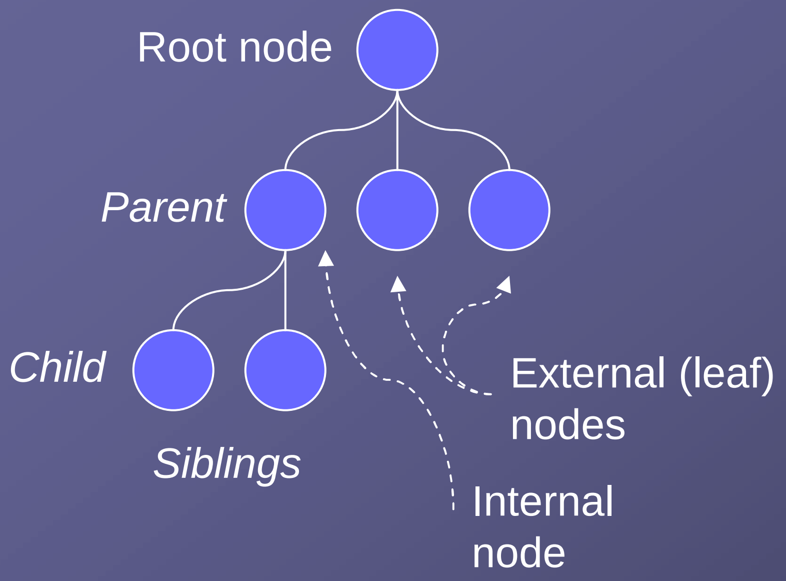
Gene Ontology (GO)

- A structure for classifying and linking genes and gene products from multiple organisms into three perspectives:
- molecular function – what activities is the entity involved in? (ex: binding)
- biological process – what process(es) is the entity involved in? (ex: cell growth)
- cellular component – where is the entity located? (ex: nucleus)

- organized in directed acyclic graphs (DAGs) - a ‘child’ entry can have many ‘parents’

Graph types: Trees vs DAGs

Tree



Root node

Parent

Child

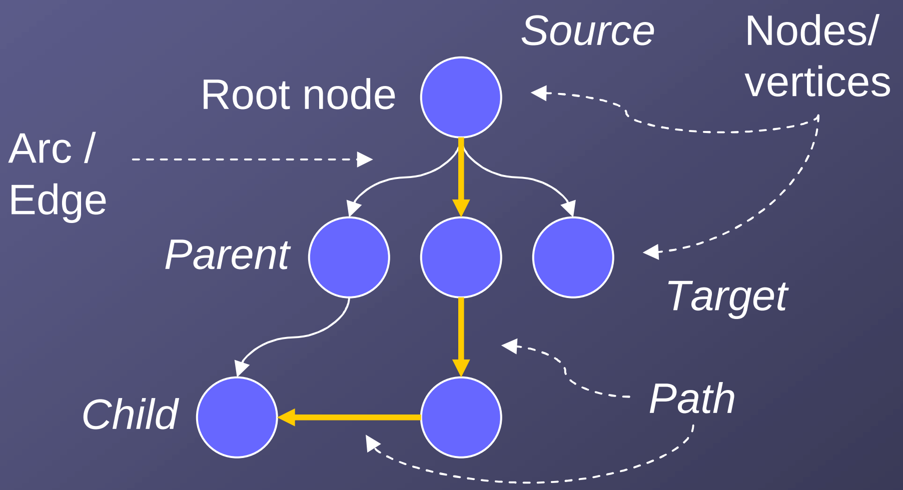
Siblings

External (leaf) nodes

Internal node

Depth = 2
(root = 0)

DAG



Source

Nodes/ vertices

Root node

Arc / Edge

Parent

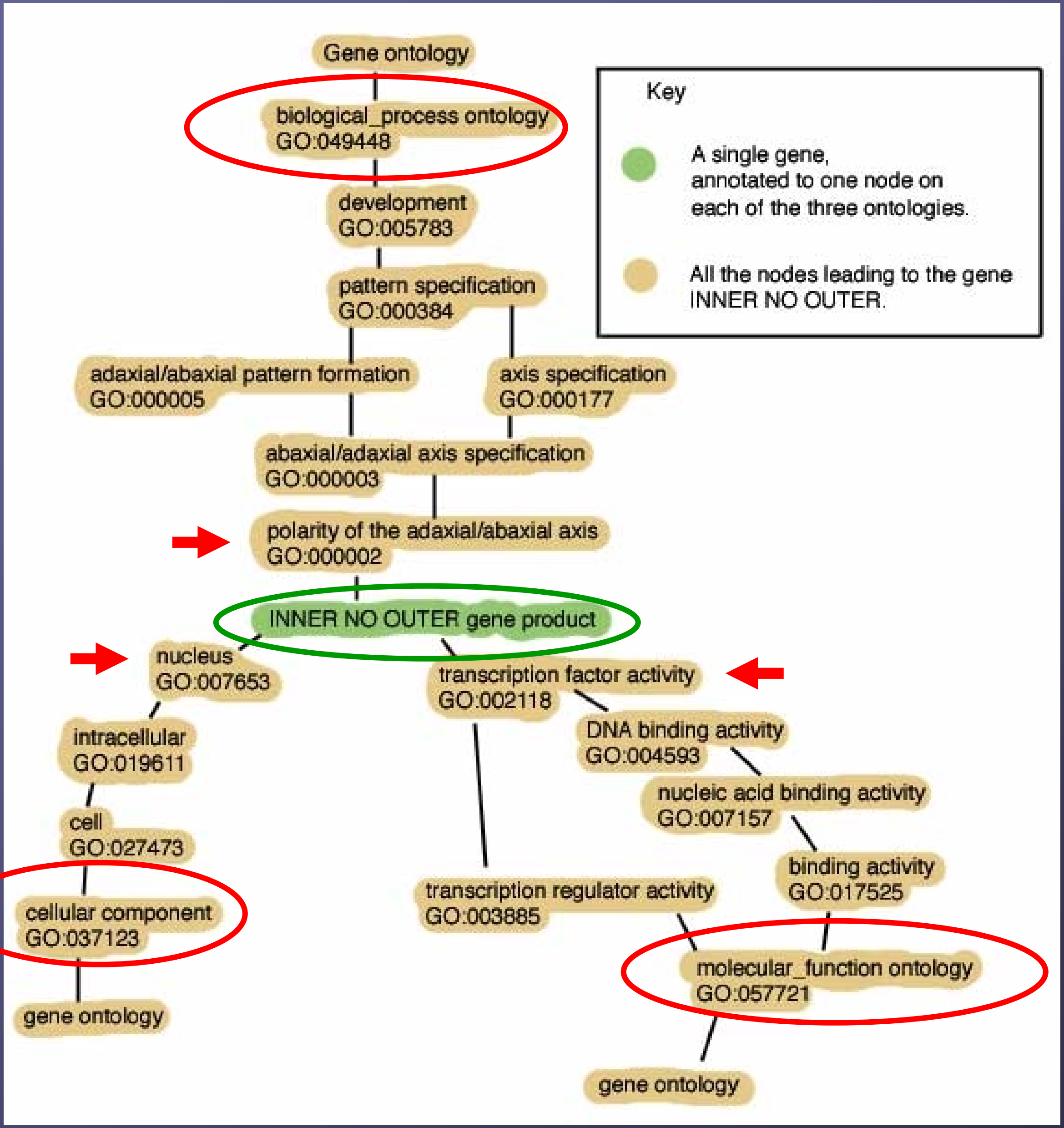
Target

Path

Child

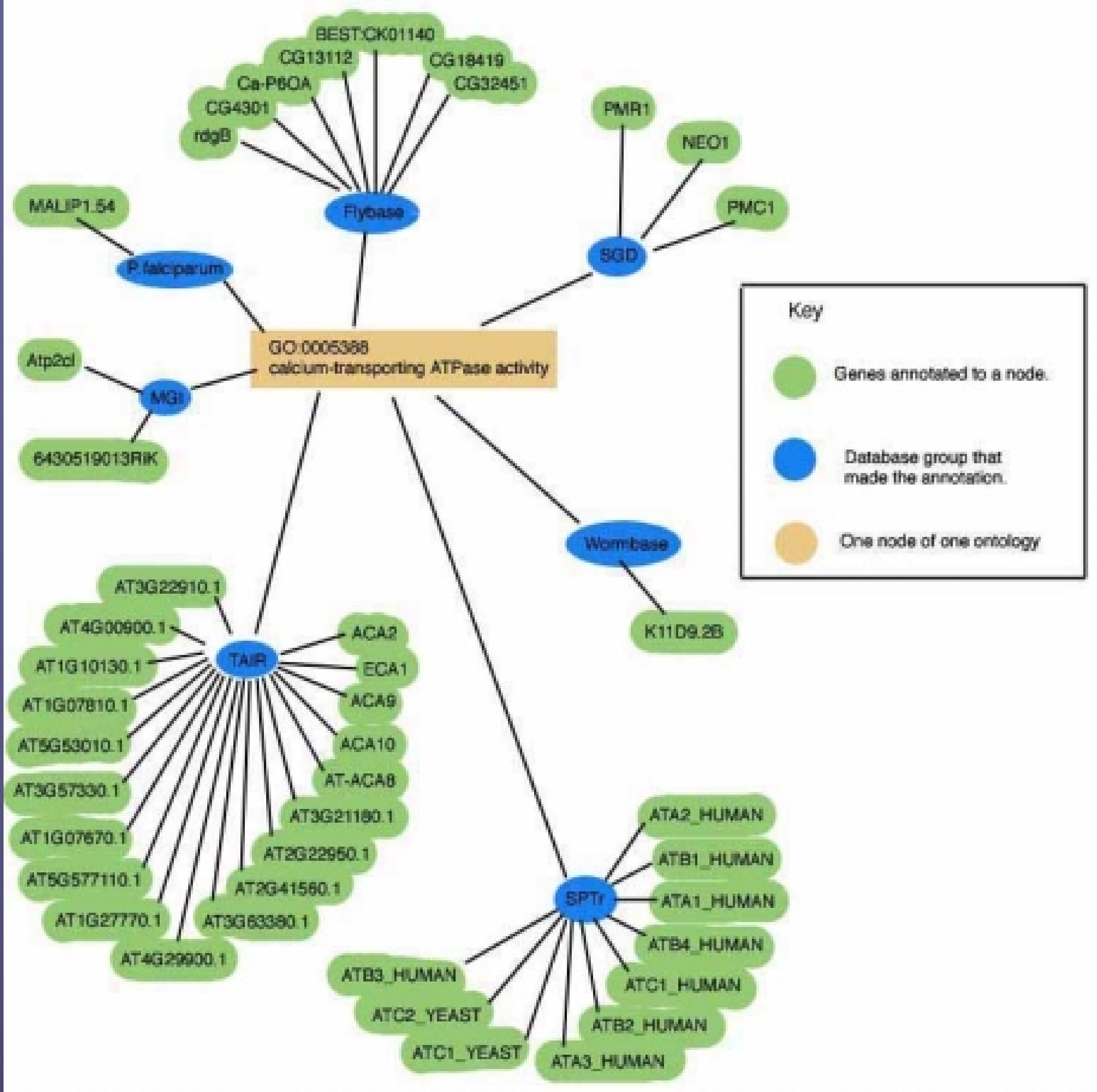
“Nodes & edges”
“Vertices & arcs”
Enables distance calculations

GO annotation


Gene ontology
biological_process ontology
GO:049448
development
GO:005783
pattern specification
GO:000384
adaxial/abaxial pattern formation
GO:000005
axis specification
GO:000177
abaxial/adaxial axis specification
GO:000003
polarity of the adaxial/abaxial axis
GO:000002
INNER NO OUTER gene product
nucleus
GO:007653
transcription factor activity
GO:002118
intracellular
GO:019611
DNA binding activity
GO:004593
nucleic acid binding activity
GO:007157
cell
GO:027473
binding activity
GO:017525
transcription regulator activity
GO:003885
cellular component
GO:037123
molecular_function ontology
GO:057721
gene ontology
gene ontology
Key
A single gene, annotated to one node on each of the three ontologies.
All the nodes leading to the gene INNER NO OUTER.

http://geneontology.org/GO.nodes.shtml

GO multi–organism annotation



http://geneontology.org/GO.annotation.example.shtml

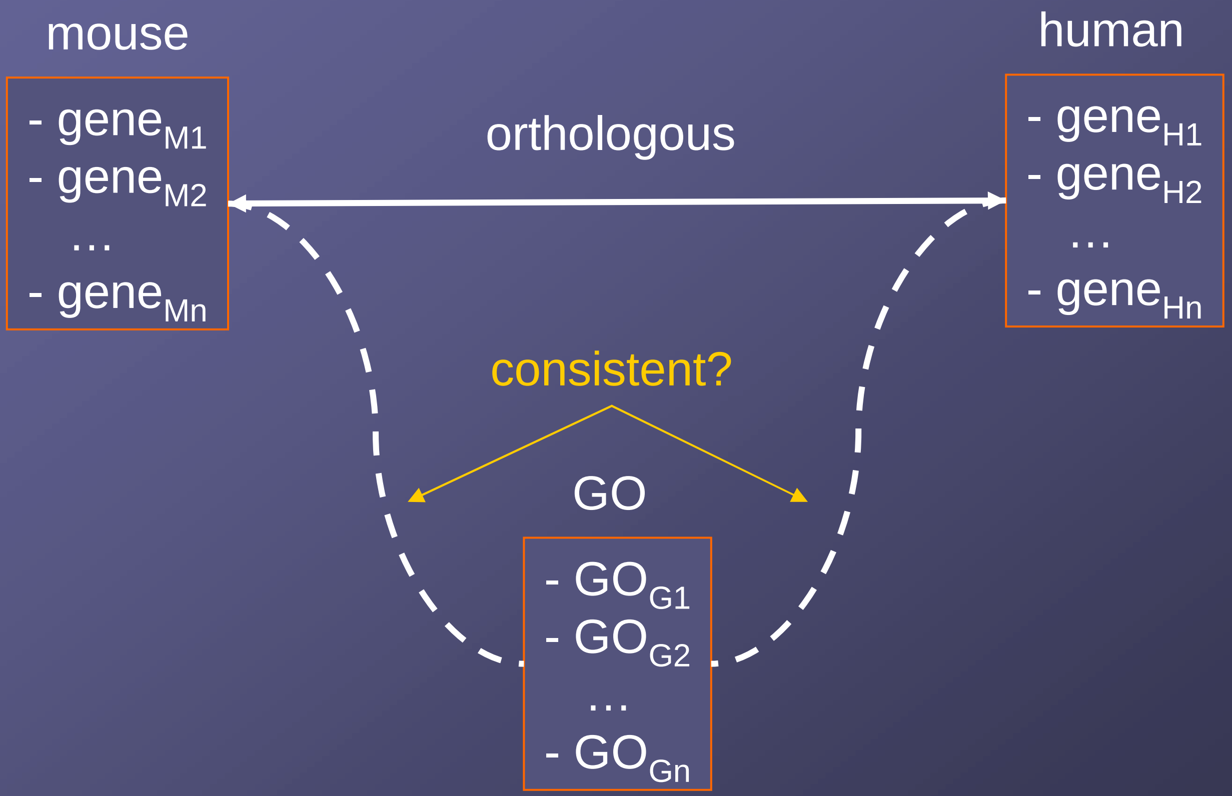
Objectives (Dolan, et al.)

- Multiple groups of individuals independently create GO annotations via differing methods and contexts
- Goal: create methods to assess consistency of GO annotation across databases for orthologous genes

Methods

- Check for consistency by “compar[ing] annotations between genes that share close evolutionary relationships [orthologous genes], and are likely (although not necessarily) to function in similar ways” [i136]
- Uses pre-existing curated orthology sets
- Uses pre-existing simplified form of GO (GO_Slims)
- Focused on Molecular Function ontology

Mouse/Human annotation consistency



mouse

- gene$_{M1}$
- gene$_{M2}$

…

- gene$_{Mn}$

orthologous

human

- gene$_{H1}$
- gene$_{H2}$

…

- gene$_{Hn}$

consistent?

GO

- GO$_{G1}$
- GO$_{G2}$

…

- GO$_{Gn}$

Data

- 14,908 mouse-human orthology pairs in MGI dataset (2004-11-12) [current stats]
- 11,860 curated mouse-human ortholog pairs
- RQ: How many ortholog pairs have annotations in both databases?

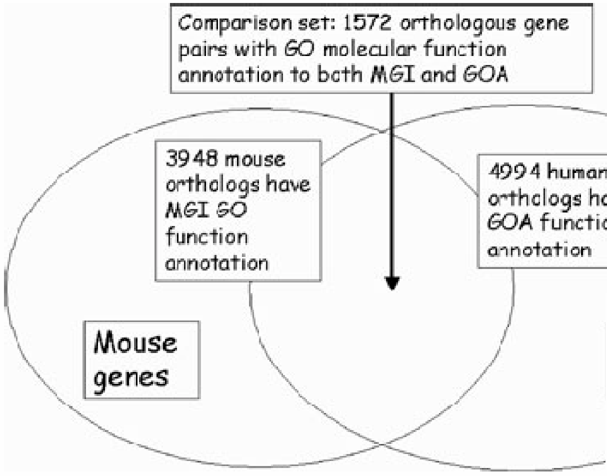
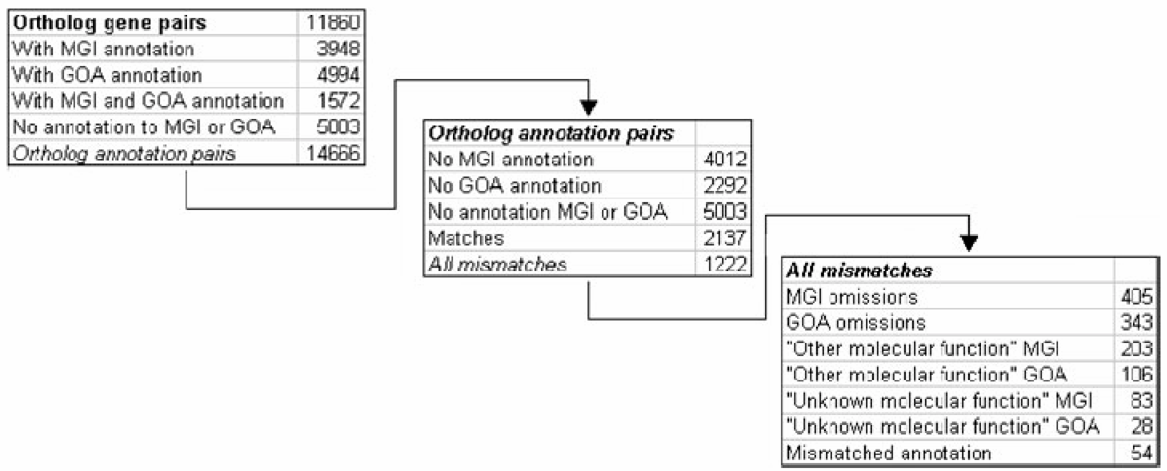

Comparison set: 1572 orthologous gene pairs with GO molecular function annotation to both MGI and GOA
3948 mouse orthologs have MGI GO function annotation
4994 human orthologs have GOA function annotation
Mouse genes

fig 3



Ortholog gene pairs	11860
With MGI annotation	3948
With GOA annotation	4994
With MGI and GOA annotation	1572
No annotation to MGI or GOA	5003
Ortholog annotation pairs	14666

Ortholog annotation pairs	
No MGI annotation	4012
No GOA annotation	2292
No annotation MGI or GOA	5003
Matches	2137
All mismatches	1222

All mismatches	
MGI omissions	405
GOA omissions	343
"Other molecular function" MGI	203
"Other molecular function" GOA	106
"Unknown molecular function" MGI	83
"Unknown molecular function" GOA	28
Mismatched annotation	54

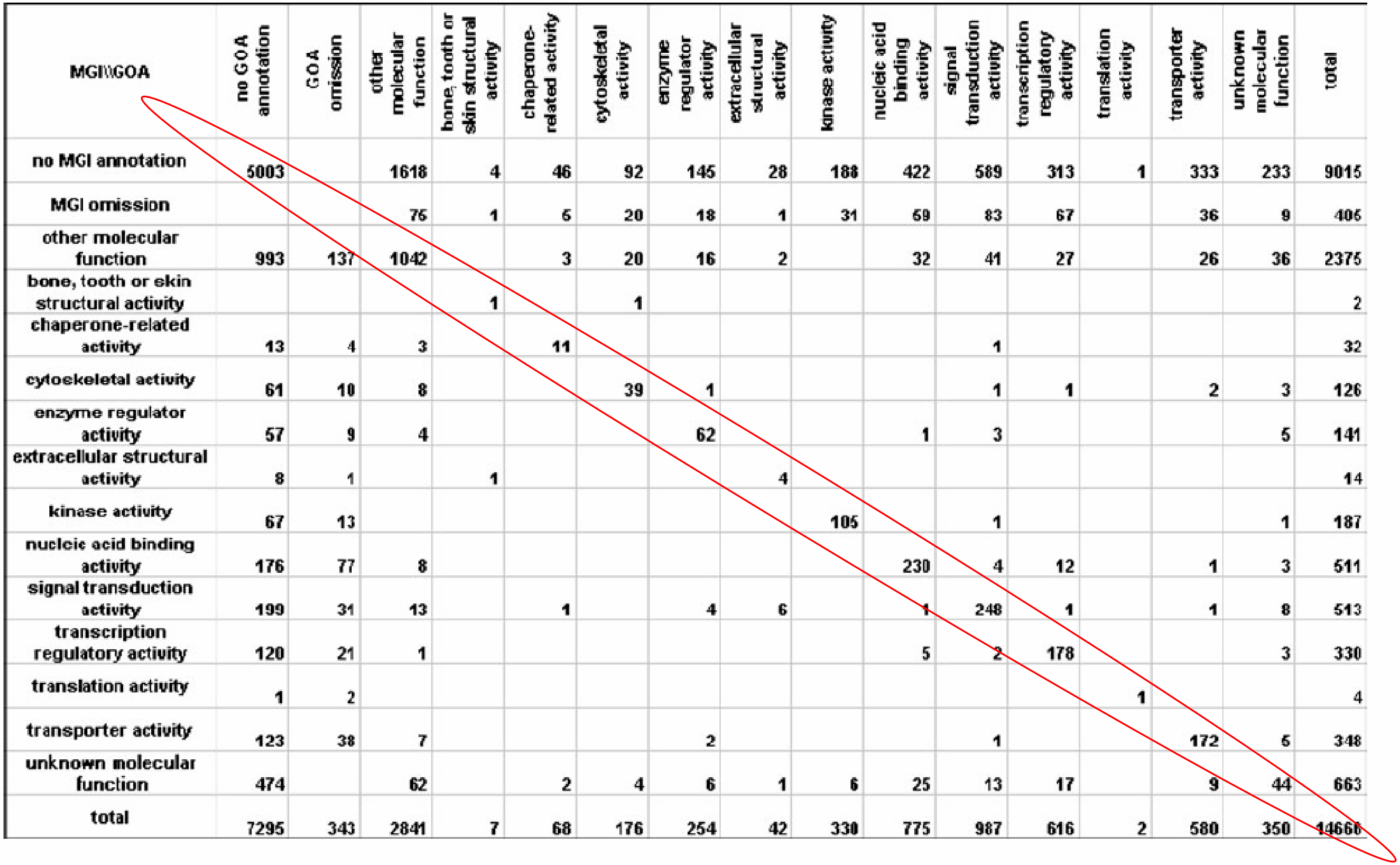
fig 4

Results

- 2,137 matches from 1,572 jointly-annotated pairs (some pairs had multiple annotations)
- 1,222 mismatches in seven case types:

1. mismatches that correctly reflect the difference in the experimental evidence for the mouse and human genes;
2. incomplete annotation;
3. Annotation based on static out-of-date automated cross-reference tables;
4. annotation errors;
5. mismatches with ‘unknown molecular function’ for one gene and a known molecular function for its ortholog;
6. annotation mismatch due to the GO structure;
7. annotation mismatch due to our GO_Slim definition.

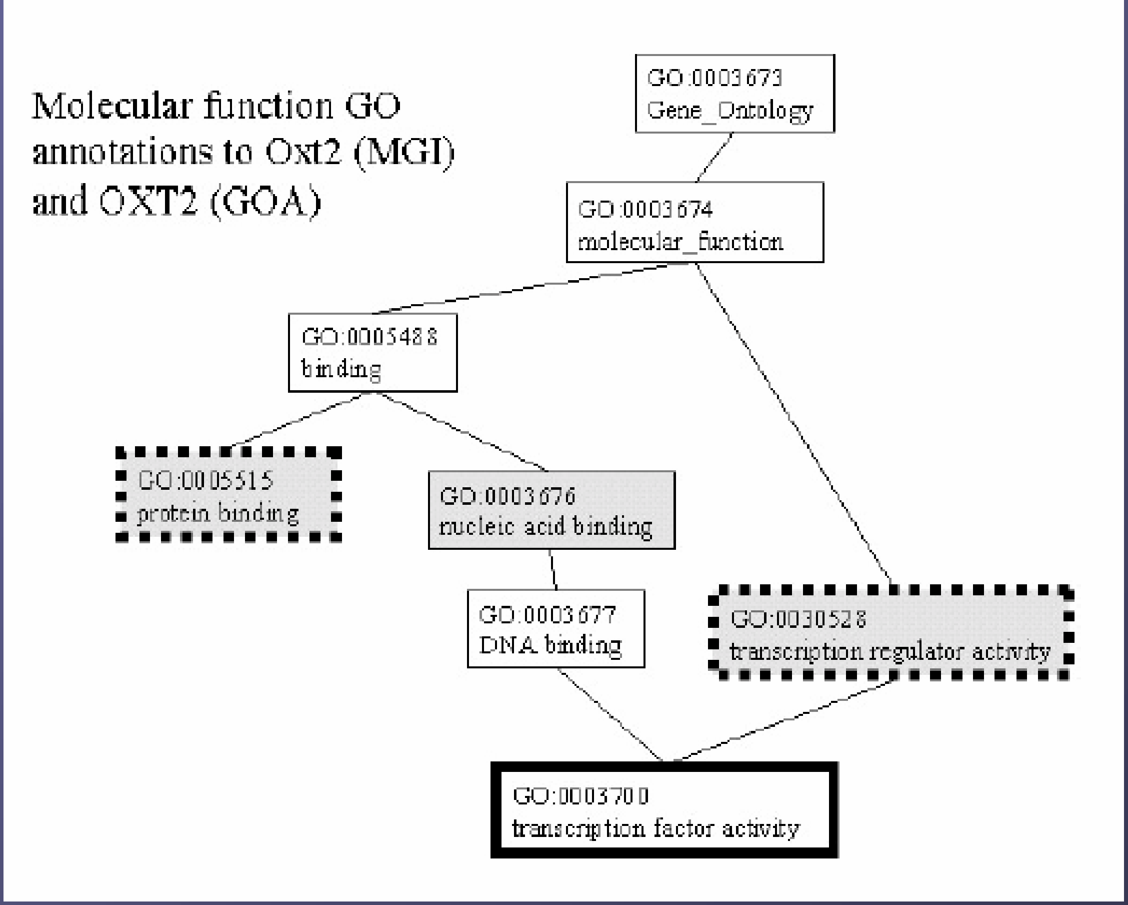
Results (table 2)



MGI\\GOA	no GOA annotation	GOA omission	other molecular function	bone, tooth or skin structural activity	chaperone-related activity	cytoskeletal activity	enzyme regulator activity	extracellular structural activity	kinase activity	nucleic acid binding activity	signal transduction activity	transcription regulatory activity	translation activity	transporter activity	unknown molecular function	total
no MGI annotation	5003		1618	4	46	92	145	28	188	422	589	313	1	333	233	9015
MGI omission			76	1	6	20	18	1	31	69	83	67		36	9	406
other molecular function	993	137	1042		3	20	16	2		32	41	27		26	36	2375
bone, tooth or skin structural activity				1		1										2
chaperone-related activity	13	4	3		11						1					32
cytoskeletal activity	61	10	8			39	1				1	1		2	3	126
enzyme regulator activity	57	9	4				62			1	3				5	141
extracellular structural activity	8	1		1				4								14
kinase activity	67	13							105		1				1	187
nucleic acid binding activity	176	77	8							230	4	12		1	3	511
signal transduction activity	199	31	13		1		4	6		1	248	1		1	8	513
transcription regulatory activity	120	21	1							5	2	178			3	330
translation activity	1	2											1			4
transporter activity	123	38	7				2				1			172	5	348
unknown molecular function	474		62		2	4	6	1	6	25	13	17		9	44	663
total	7295	343	2841	7	68	176	254	42	330	775	987	616	2	580	350	14666

Fig. 2. Consistency of mouse–human ortholog GO annotation as a confusion matrix. In this chart the MGI GO_Slim annotations define the rows and GOA annotations define columns. Diagonal elements of matrix represent consistent MGI and GOA annotations: e.g. 39 orthologs are annotated to 'cytoskeletal activity' Molecular Function by MGI for the mouse ortholog and by GOA for the human ortholog. Off-diagonal elements represent mismatches and potential inconsistencies: e.g. four orthologs are annotated to 'signal transduction activity' Molecular Function by MGI for the mouse ortholog and to 'enzyme regulator activity' by GOA for the human ortholog.

Results (fig 5)



Questions

- The method’s precision is uncertain because orthologous genes don’t necessarily have the same function
- How many of the other 13,336 orthologous pairs should be annotated with the same GO terms? (14,908 - 1,572)
- The use of GO_Slims obscures mis-matches at more granular levels.
- Is there a discovery component, or is this only useful for quality control?
- How do we represent 3-way consistency? Or *n*-way?