Ontology Engineering

Supplementary Information

Gil Alterovitz[1,2,3], Michael Xiang[1,2], David P. Hill[4], Jane Lomax[5], Jonathan Liu[6], Michael Cherkassky[2], Jonathan Dreyfuss[1,2], Chris Mungall[7], Midori A. Harris[5], Mary E. Dolan[4], Judith A. Blake[4], and Marco F. Ramoni[1,2]

Supplementary Notes 1

We proposed twelve terms to the GO Consortium for validation of our approach and to see if it would be implemented in the ontology. As a result of our analysis, eleven of the twelve GO terms (92%) that were evaluated were moved by GO curators to produce a more optimal location in the ontology. All of the modifications were found to retain the logical/ontological correctness. The structural changes repositioned the GO term in a location where it was no longer found to be of inappropriate specificity (Figure S5). All of the modifications were found to retain the logical/ontological correctness. Only in the case of one GO term, "pigmentation" (GO:0043473), was the placement in the graph (level 1) not altered. In this case, the "pigmentation" GO term appears out of place for its level due to under-annotation of the genes involved in the process. The Mouse Genome Informatics (MGI) resource had 44 genes annotated to the GO term "pigmentation." However, a search at MGI for genes in mice having a "pigmentation" phenotype or defect returned 417 genes. This example illustrates that many more potential genes can be annotated to the GO term "pigmentation" than have currently been annotated. In addition, it highlights an annotation bias in which classic mutations resulting in coat-color pigmentation defects have not been a high priority for MGI GO curators. This case illustrates an important point about the methods that GO curators used to evaluate the appropriateness of suggested changes. First, curators examined the terms and their placement in the graph without respect to annotations, but rather with the idea of whether the biology could be better represented by modifying the placement of a term. In the case of "pigmentation" they decided that the term was in the most appropriate place. The annotation status of the term was then examined to

Nature Biotechnology, vol. 28, no. 2 Ramoni, M et al.

explain why the information content suggested a term re-placement. Identification of annotation biases such as this can be used to target areas of the ontology for future curation. In addition, it should be possible to take known annotation biases into account to improve GO analyses. This analysis shows that the methods used in this study can be used to both improve the ontology itself as well as to identify areas of the ontology that should be addressed by gene annotators.

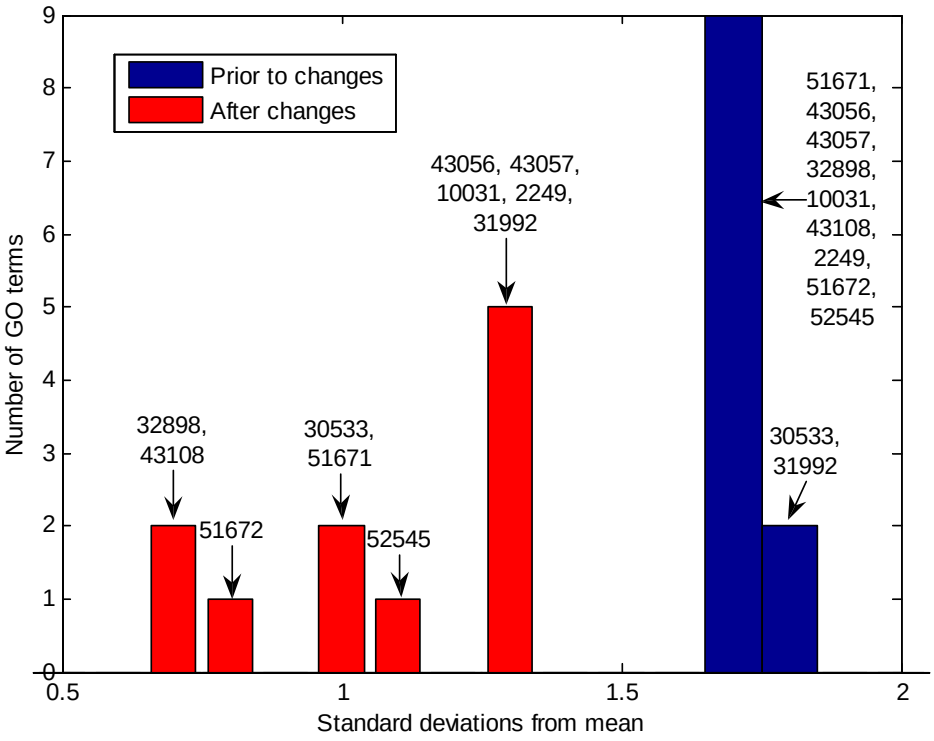
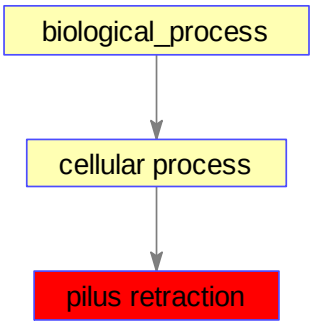


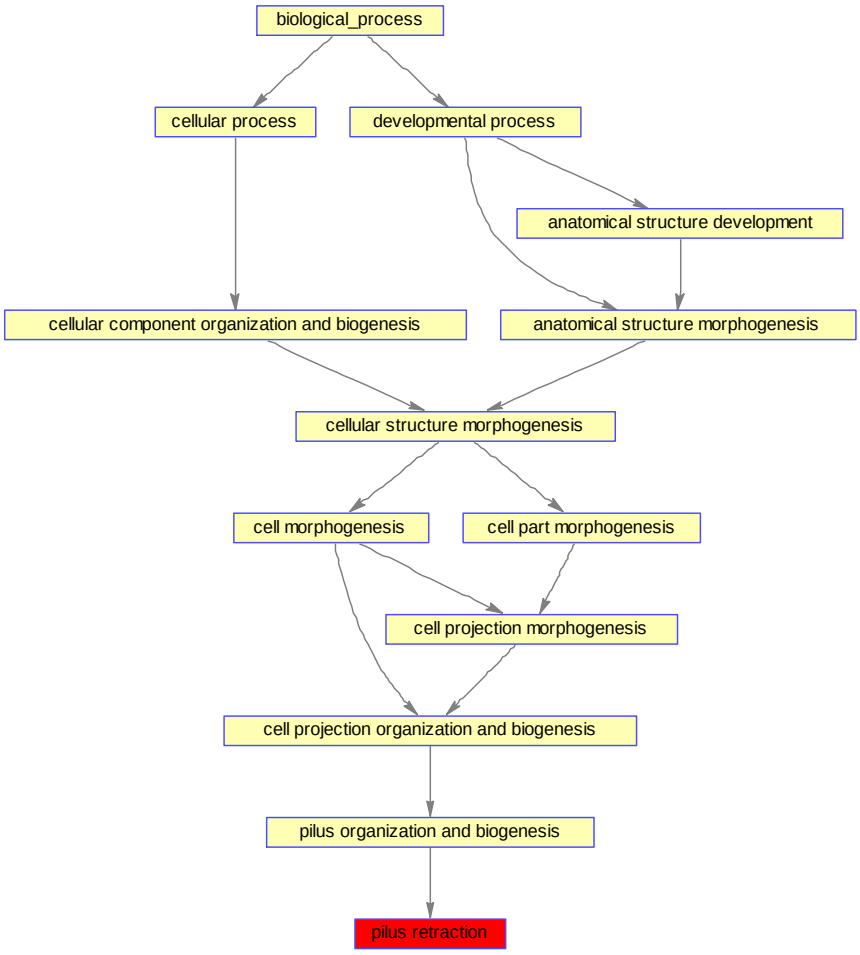
Figure S1: Extent of deviation from level mean for GO terms moved by the GO Consortium, before and after the changes. GO terms: GO:0051671, “induction of autolysin activity in another organism”; GO:0043056, “forward locomotion”; GO:0043057, “backward locomotion”; GO:0032898, “neurotrophin production”; GO:0010031, “circumnutation”; GO:0043108, “pilus retraction”; GO:0002249, “lymphocyte anergy”; GO:0051672, “cell wall peptidoglycan catabolic process in another organism”; GO:0052545, “callose localization”; GO:0030533, “triplet codon-amino acid adaptor activity”; GO:0031992, “energy transducer activity.”

Here, we focus on the change to “pilus retraction” as a specific case study. Initially, this GO term was in level 2 as a child of “cellular process” (Figure S2.a), where it was found

to be too specific for its graphical position. As a consequence of our analysis, review of this term by the GO Consortium lead to movement of the term to level 9, where it is of appropriate specificity (Figure S2.b). The movement also resulted in creation of a new term, “pilus organization and biogenesis” (GO:0043711), which serves as the new parent of “pilus retraction.” The change places “pilus retraction” in a new context in the GO graph that is more sensible biologically, which is confirmed quantitatively by the information theoretic analysis.



a)



b)

Figure S2: The graphical placement of “pilus retraction” before (a) and after (b) information-theoretic repositioning.

Supplementary Notes 2

Our approach necessitates quantification of term specificity, *i.e.* information content. Intuitively, descriptive specificity correlates inversely with annotation frequency. For example, the GO term "metabolism," which annotates approximately 40% of human genes, reveals little about actual gene function; on the other hand, the GO term "carbohydrate metabolism," which annotates fewer than 2% of human genes, imparts more information. Mathematically, the information content (in bits) of an ontology term A_n is the self-information (also called "surprisal") of the term, denoted by $I(A_n)$, which is related to the definition of Shannon information[1]: $I(A_n) = -\log_2 p(A_n)$, where

$$p(A_n) = \frac{|k(A_n)|}{\left|\bigcup_{m=1}^{j} k(A_m)\right|}$$

Here, $k(A_n)$ is the number of genes annotated by term A_n, j is the total number of ontology terms, and $p(A_n)$ is the probability of observing a gene, chosen randomly, and finding that it is annotated by term A_n; *i.e.*, annotation frequency.

Supplementary Notes 3

Figure S3 plots the average information content (in bits) of each GO level. Error bars indicate one standard deviation in information content. As would be expected, the descriptive specificity of GO terms generally increases with GO level. The mean information content occasionally decreases from one level to the next. Such an occurrence is an "information bottleneck" (Figure S3): most of the gene annotation information of the previous level is transmitted to the next through only a few terms. The larger the decrease in information content, the more severe the "information bottleneck." We quantify the extent of this structural variation by defining the inter-level variability metric as the area between the curve of mean information content by level and its monotonically increasing convex hull.

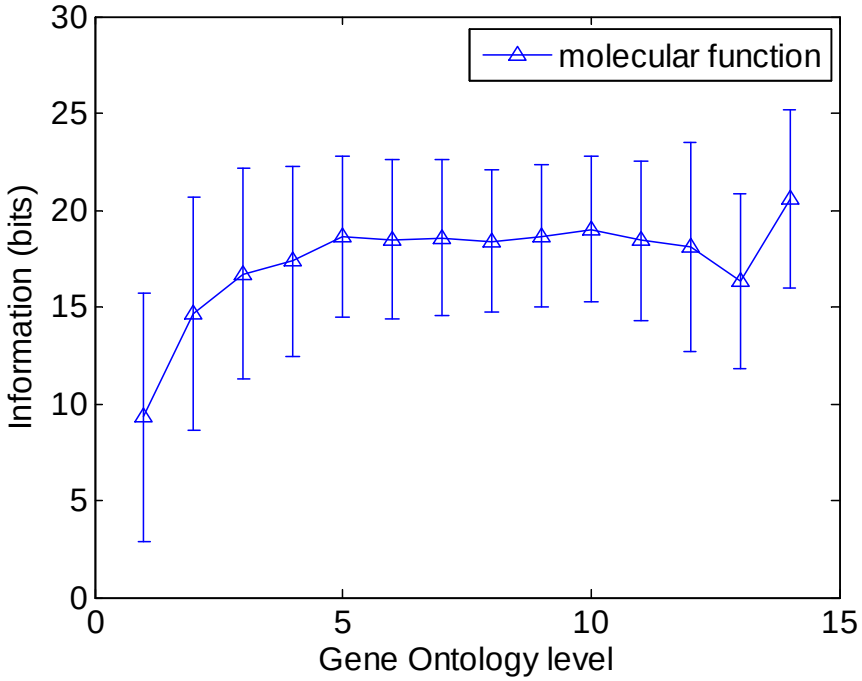


Figure S3: An information bottleneck (molecular function) between levels 12 and 13.

Supplementary Notes 4

The third measure of structural variation, quantified by the topological variability metric, arises from inefficiencies in the branch structure of the ontology. Application of the entropy rate measure (*H*) to a random walk on a weighted graph (*G*) quantifies the uniformity in branch structure[2]:

$$H(G) = -\sum_i \sum_j \frac{W_{ij}}{2W} \log \frac{W_{ij}}{W_i},$$

Here, W_{ij} is the weight from term *i* to term *j*, W_i is the sum of all weights emanating from term *i*, and *W* is the sum of all edge weights in *G*. *Wij* is the proportion of one child's accessions relative to the total number of children accessions. A higher entropy rate is more desirable, as it indicates that the branch structure is more uniform and no particular terms or annotation outcomes are favored over others. Since a lower value for the other two variability metrics is better, the topological metric is defined as the

reciprocal of the entropy rate, so its interpretation is consistent with the other structural variability metrics.

Supplementary Notes 5

Figures S4, S5, and S6 show the three metrics used to quantify structural elements within GO as well as the best-fit line.

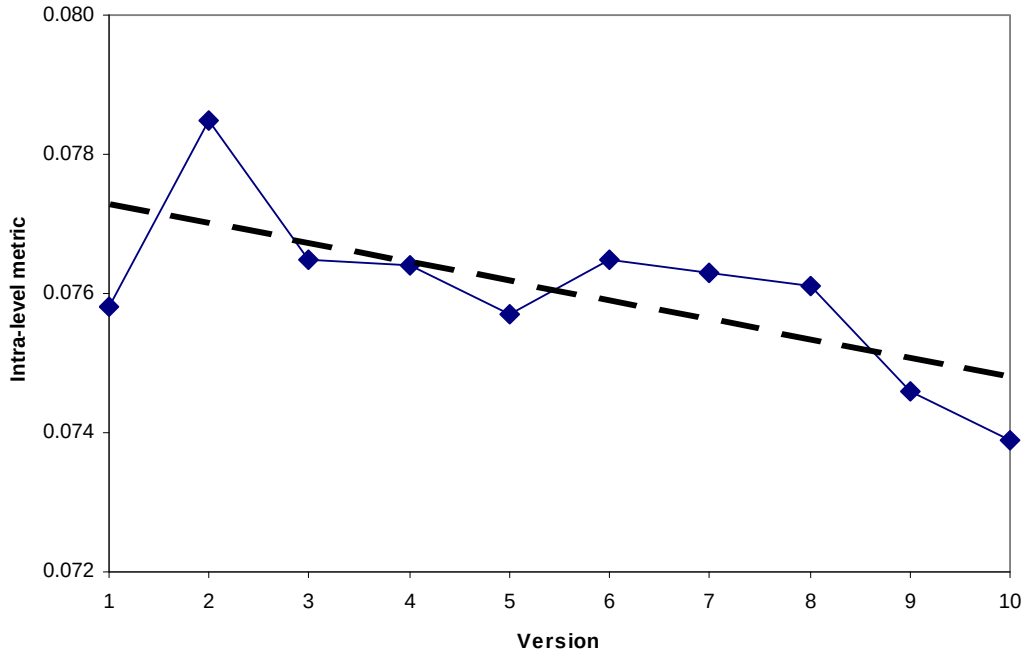


Figure S4: Intra-level metric across GO versions.

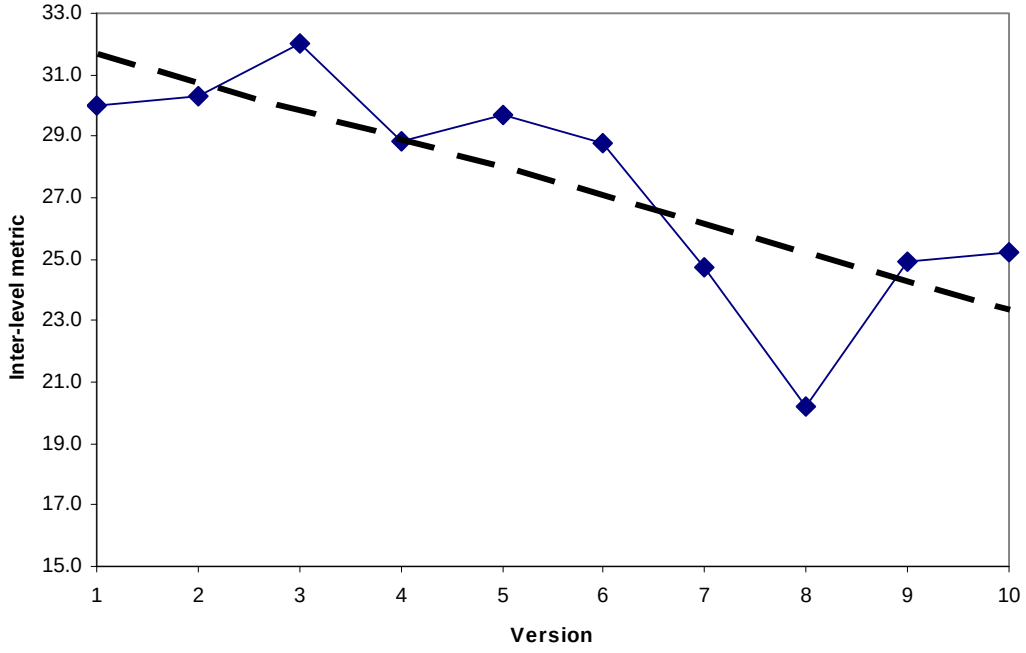


Figure S5: Inter-level metric across GO versions

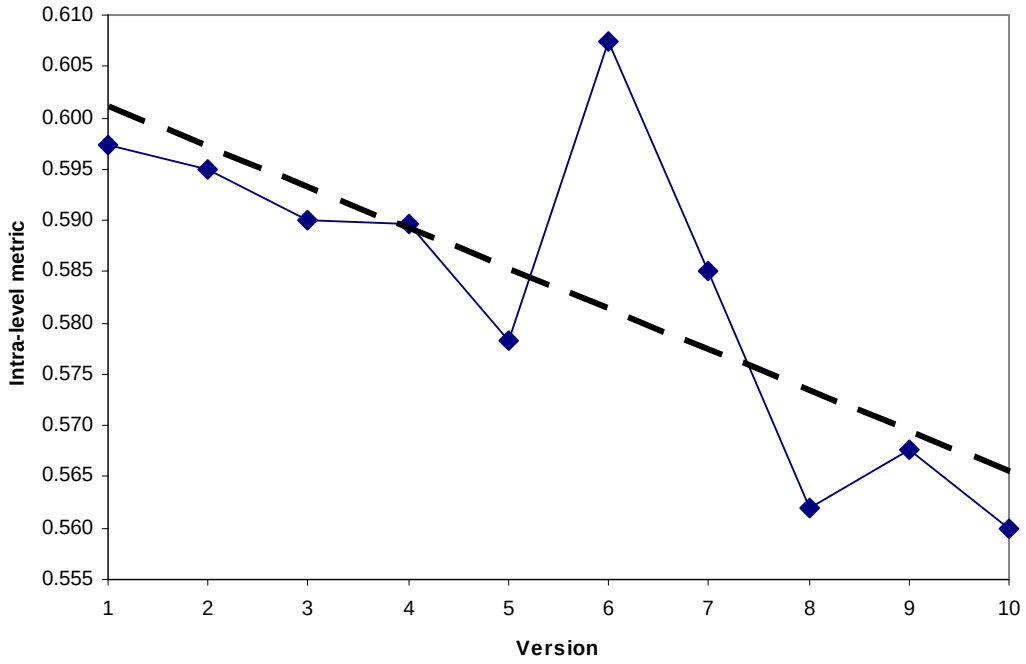


Figure S6: Topological metric across GO versions

Supplementary Notes 6

We altered the structure of the April 1, 2007 release of GO according to our intra-level variability metric. For a robust assessment of the effect of these changes on the overall distribution of intra-level variability scores of all nodes, we applied a paired Wilcoxon test to the intra-level variability of all nodes before and after structural modification. This test gave a one-tailed p-value < 10^{-3}, indicating significant improvement.

Supplementary Notes 7

Here we examine the changes in the enrichments of the 10,117 DNA microarray experimental gene signatures from human and model organisms mapped to homologous human genes. The set of differentially expressed genes in each experiment were extracted from the GEO[3] via Exalt[4]. Gene symbols were mapped to UniProt IDs[5] and a Fisher exact test for enrichment was performed for each GO category using Matlab and R, where all human genes were used as the background. This was done for the unmodified GO structure from April 1, 2007 and for that structure modified by moving terms upwards according to the intra-level variability metric. For each experiment with at least three proteins in its signature, the enrichment p-values between the modified and unmodified GO structures were compared using the Kolmogorov-Smirnov (KS) test, where 95% of the KS p-values were found to be below 5%. These KS p-values were aggregated over all experimental gene signatures using Fisher's combined probability test[6] which yielded a combined p-value < 10^{-3}.

Nearly all (97.5%) of the comparisons had a different set of significantly enriched GO categories when using the modified versus unmodified GO structures, where significance was defined as a False Discovery Rate[7] below 10%. Further, these sets differed by 14.6%, on average, where the proportion different was calculated as the number of GO categories that were significant according to only one GO structure (i.e., significant when using the modified but not the unmodified version, or vice versa) divided by the number significant according to either GO structure.

Supplementary Notes 8

We examined genes and gene ontology terms linked by the new approach after enrichment analysis of the microarray experimental gene signatures. If a gene signature is found to be enriched for a particular GO term, then the genes in that signature are annotated with that GO term significantly more than would be expected by chance alone. We examined genes that were inside a gene signature associated with a new GO term, but not annotated for that gene ontology term. When such term appears several times for the same gene, it suggests that the gene could potentially be associated with that new term. Out of fourteen that we proposed to the Gene Ontology Consortium for corroboration, 12 of these were accepted and actually integrated into the Gene Ontology Annotation Database. One (O95271, "TRF1-Interacting Ankyrin-Related") was not accepted and one (P19404, NADH dehydrogenase (Ubiquinone) flavoprotein 2, 24kDa) is being considered for annotation to a less specific terms than the one proposed (see Table S1). While the links may be valid, GOC could not find experimental evidence found for both of the highly specific proposed GO terms and, in the case of O95271, the annotation could not be done based on similarity with ankyrin.

As an example of our results, gene analysis enrichment using our modified ontology structure found the term "Translation factor activity" (GO:008135) to be associated with the gene "Transcription factor SOX-11" (P35716). The enrichment was found in the modified ontology structure, but was not found when doing the same analysis in the original ontology. Certainly, this is suggested in the name of the term and protein in this particular case. But, we wanted to have a more formal process for corroboration of this and the other proposed annotations. The GOC provided this and evaluated them using the same strict standards used during its normal annotation process (see Table 1 below).

Nature Biotechnology, vol. 28, no. 2 Ramoni, M et al.

Table S1: List of Gene Ontology Annotations Proposed

Uniprot ID	Uniprot Name	Organism	GO Term	GO Term Name	Notes
Q5T8M9	Actin, Alpha 1 Skeletal Muscle	Human	30239	Myofibril Assembly	Accepted by GOC and integrated into GOA. Modified GO (April 1, 2007) with GOA (April 10, 2007) versus modified GO (April 1, 2007) with GOA (April 10, 2007).
Q5T8M9	Actin, Alpha 1 Skeletal Muscle	Human	15629	Actin Cytoskeleton	Accepted by GOC and integrated into GOA. Modified GO (April 1, 2007) with GOA (April 10, 2007) versus modified GO (April 1, 2007) with GOA (April 10, 2007).
Q96D31	Calcium Release-Activated Calcium Chanel Protein 1	Human	51924	Regulation of Calcium Ion Transport	Accepted by GOC and integrated into GOA. Modified GO (April 1, 2007) with GOA (April 10, 2007) versus modified GO (April 1, 2007) with GOA (April 10, 2007).
O14924	Regulator of G-protein signaling 12	Human	30695	GTPase regulator activity	Accepted by GOC and integrated into GOA. Enrichment of gene signatures at GO level 2. Modified GO (April 1, 2007) with GOA (April 10, 2007) versus modified GO (April 1, 2007) with GOA without enforcing annotation inheritance (April 10, 2007).
Q00534	Cyclin-dependent Kinase 6	Human	45786	Negative Regulation of Progression through Cell Cycle	Accepted by GOC and integrated into GOA. Modified GO (April 1, 2007) with GOA (April 10, 2007) versus modified GO (April 1, 2007) with GOA without enforcing annotation inheritance (April 10, 2007).
P35716	Transcription Factor SOX-11	Human	08135	Translation Factor Activity	Accepted by GOC and integrated into GOA. Modified GO (April 1, 2007) with GOA (April 10, 2007) versus modified GO (April 1, 2007) with GOA without enforcing annotation inheritance (April 10, 2007).
Q14746	Component of Oligomeric Golgi Complex 2	Human	05795	Golgi Stack	Accepted by GOC and integrated into GOA. Enrichment of gene signatures at GO level 2. Modified GO (April 1, 2007) with GOA (April 10, 2007) versus modified GO (April

					1, 2007) with GOA without enforcing annotation inheritance (April 10, 2007).
Q14563	Semaphorin 3A	Human, Mouse (O08665)	48841	Regulation of Axon Extension Involved in Axon Guidance	Accepted by GOC and integrated into GOA. Modified GO (April 1, 2007) with GOA (April 10, 2007) versus modified GO (April 1, 2007) with GOA (April 10, 2007).
P18827	Syndecan-1	Human, Mouse (P 18828)	48627	Myoblast Development	Accepted by GOC and integrated into GOA. Modified GO (April 1, 2007) with GOA (April 10, 2007) versus modified GO (April 1, 2007) with GOA (April 10, 2007).
P18827	Syndecan-1	Human	55002	Striated Muscle Cell Development	Accepted by GOC and integrated into GOA. Modified GO (April 1, 2007) with GOA (April 10, 2007) versus modified GO (April 1, 2007) with GOA (April 10, 2007).
Q86VI1	Exocyst complex component 3-like protein	Human, Mouse (Q8BI71)	06887	Exocytosis	Accepted by GOC and integrated into GOA. Modified GO (April 1, 2007) with GOA (April 10, 2007) versus modified GO (April 1, 2007) with GOA without enforcing annotation inheritance (April 10, 2007).
Q5T8M9	Actin, Alpha 1, Skeletal Muscle	Human	30017	Sarcomere	Accepted by GOC and integrated into GO. Modified GO (April 1, 2007) with GOA (April 10, 2007) versus modified GO (April 1, 2007) with GOA (April 10, 2007).
O95271	TRF1-Interacting Ankyrin-Related	Human	15662	ATPase activity, coupled to transmembrane movement of ions, phosphorylative mechanism	Not accepted by GOC. Enrichment of gene signatures at GO level 11. Modified GO (April 1, 2007) with GOA (April 10, 2007) versus modified GO (April 1, 2007) with GOA without enforcing annotation inheritance (April 10, 2007).
P19404	NADH dehydrogenase (Ubiquinone) flavoprotein 2, 24kDa	Human	16628	Oxidoreductase activity, acting on the CH-CH group of donors, NAD or NADP as acceptor	Being investigated for less specific terms. Enrichment of gene signatures at GO level 3. Modified GO (April 1, 2007) with GOA (April 10, 2007) versus modified GO (April 1, 2007) with GOA (April 10, 2007).

References

1. MacKay, D.J.C. *Information theory, inference, and learning algorithms*, xii, 628 p. (Cambridge University Press, Cambridge, U.K. ; New York, 2003).
2. Cover, T.M. & Thomas, J.A. *Elements of Information Theory*, (Wiley-Interscience New York, NY, 2006).
3. Edgar, R. & Barrett, T. NCBI GEO standards and services for microarray data. *Nat Biotechnol* 24, 1471-2 (2006).
4. Yi, Y., Li, C., Miller, C. & George, A.L., Jr. Strategy for encoding and comparison of gene expression signatures. *Genome Biol* 8, R133 (2007).
5. Alterovitz, G., Patek, D., Kohane, I.S. & Ramoni, M. Human Protein Meta-Interaction Database (HPMD) Potentiates Integration for Meta-Analysis. *Proceedings of IEEE Genomic Signal Processing and Statistics (GENSIPS)* (2005).
6. Fisher, R.A. Combining independent tests of significance. *American Statistician* 2(1948).
7. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society B* 57, 289-300 (1995).
8. Moretti, S., Procopio, A., Boemi, M. & Catalano, A. Neuronal semaphorins regulate a primary immune response. *Curr Neurovasc Res* 3, 295-305 (2006).