Supporting Information

Table of Content

Supporting Texts

Supporting Tables

Supporting Figures

Reference

Supporting Texts

Text S1. Calculation of *p*-values for rate ratio tests

In this section, we explain how to calculate a p-value for a rate ratio in the context of Gene Ontology (GO) term annotation, adapting a more general derivation that was introduced by another earlier study (Fay, 2010). Consider a GO term q annotated to two taxa, t and t', with respective annotation rates:

$$\begin{cases} P_t(q) = n_t(q)/N_t \\ P_{t'}(q) = n_{t'}(q)/N_{t'} \end{cases} \qquad (S1)$$

Here, N_t and $N_{t'}$ are the total number of annotated proteins in taxon t and t', respectively, while $n_t(q)$ and $n_{t'}(q)$ are the subset of proteins annotated with q. Without loss of generality, we suppose $P_t(q) \leq P_{t'}(q)$. Note that $P_t(q)$ and $P_{t'}(q)$ are observed annotation rate, while the (unknown) real annotation rates of q annotated to t and t' are $\lambda_t(q)$ and $\lambda_{t'}(q)$, respective. The null and alternative of rate ratio test for q are:

$$\begin{cases} H_0\text{:} \quad RAR(q) = \dfrac{\lambda_t(q)}{\lambda_{t'}(q)} = RR \\ H_1\text{:} \quad RAR(q) = \dfrac{\lambda_t(q)}{\lambda_{t'}(q)} < RR \end{cases} \qquad (S2)$$

RR=0.1 is chosen as a biologically meaningful threshold for the difference in annotation rate; the chosen rate ratio threshold is ultimately arbitrary and constitutes a tunable parameter. For ease of discussion, we denote:

$$N(q) = n_t(q) + n_{t'}(q) \qquad (S3)$$

$$p(q) = \frac{N_t \cdot \lambda_t(q)}{N_t \cdot \lambda_t(q) + N_{t'} \cdot \lambda_{t'}(q)} \qquad (S4)$$

Since any protein among the pool of all $N(q)$ proteins associated with q can only belong to either taxon t or t' with probability $p(q)$ and $1 - p(q)$, respectively, the number of q-annotated proteins belonging to taxon t among all q-annotated proteins should follow a binomial distribution:

$$n_t(q) \sim binom(N(q), p(q)) \qquad (S5)$$

The p-value can thus be calculated as:

$$\begin{aligned} p - value(RAR(q) < RR) &= P\left(n_t(q) \geq binom\big(N(q), p(q)\big)\right) \\ &= 1 - P\left(n_t(q) < binom\big(N(q), p(q)\big)\right) \end{aligned} \qquad (S6)$$

Since $n_t(q)$ is a discrete variable (i.e. an integer), equation (S6) can be rewritten into:

$$\begin{aligned} p - value(RAR(q) < RR) &= 1 - P\left(n_t(q) - 1 \leq binom\big(N(q), p(q)\big)\right) \\ &= 1 - pbinom\big(n_t(q) - 1, N(q), p(q)\big) \end{aligned} \qquad (S7)$$

The term $pbinom\big(n_t(q) - 1, N(q), p(q)\big)$ is the value of Cumulative Distribution Function (CDF) for binomial distribution $binom\big(N(q), p(q)\big)$ at quantile $n_t(q) - 1$. Under any null hypothesis regarding the rate ratio parameter, equation (S4) can be re-written as:

$$p(q) = \frac{N_t \cdot \lambda_t(q)/\lambda_{t'}(q)}{N_t \cdot \lambda_t(q)/\lambda_{t'}(q) + N_{t'}} = \frac{N_t \cdot RR}{N_t \cdot RR + N_{t'}} \qquad (S8)$$

Therefore, equation (S7) can be rewritten into:

$$p - value(RAR(q) < RR) = 1 - pbinom\left(n_t(q) - 1, n_t(q) + n_{t'}(q), \frac{N_t \cdot RR}{N_t \cdot RR + N_{t'}}\right) \quad (S9)$$

Our program for the p-value calculation in our study is available at https://zhanglab.ccmb.med.umich.edu/RAR. Due to numerical limit for addition of double precision floating point numbers, all p-value <2.22E-16 will be output as 2.22E-16, because the mantissa of an IEEE 64-bit double precision floating point number has 52 bits, and 2^{-52}=2.22E-16. In other words, the minimal p-value that can be reported by this test is 2.22E-16 using double precision calculation.

Supporting Tables

Table S1. Species selected for this study (Taxon identifiers are given in parentheses).

Taxon	Species (Taxon identifier)
Animals (3 species, vertebrates)	*Homo sapiens* (9606), *Mus musculus* (10090), *Gallus gallus* (9031).
Animals (2 species, invertebrates)	*Drosophila melanogaster* (7227), *Caenorhabditis elegans* (6239).
Bacteria (7 species)	*Escherichia coli* (83333), *Clostridium difficile* (272563), *Bacillus subtilis* (224308), *Helicobacter pylori* (85962), *Mycoplasma genitalium* (243273), *Pseudomonas aeruginosa* (208963), *Salmonella typhimurium* (99287).
Archaea (3 species)	*Methanococcus maripaludis* (267377), *Sulfolobus acidocaldarius* (330779), *Thermoplasma acidophilum* (273075)
Fungi (3 species)	*Saccharomyces cerevisiae* (559292), *Schizosaccharomyces pombe* (284812), *Candida albicans* (237561).
Plants (2 species)	*Arabidopsis thaliana* (3702), *Selaginella moellendorffii* (88036).

Table S2. Overall statistics of annotations for GO terms with rate ratio <0.1 by automated rate ratio analysis. The data reported here includes both GO terms confirmed and rejected as misannotations by our manual inspection. For the subset of potential misannotations confirmed by our manual inspection, the overall statistics are reported in Table 1 of the main text.

UniProt-GOA release	Analysis type	Number of annotations for GO terms with rate ratio <0.1		
		GO terms	Proteins[a]	Annotations[b]
2019-06-03	Kingdom	2369	7553	9126
	Phylum	618	2107	2706
	Both	2987	9660	11832
2018-11-06	Kingdom	2448	7592	9243
	Phylum	651	2133	2739
	Both	3099	9725	11982

[a] For each GO term with rate ratio <0.1, we retrieve the list of proteins in the taxon with the minimal (but non-zero) number of proteins annotated with this GO term. "Proteins" are the number of unique proteins over all GO terms with rate ratio <0.1.

[b] "Annotations" refers to the number of protein-GO term associations. For example, if GO:0005739 "mitochondrion" and GO:0005634 "nucleus" are both annotated to two proteins P39615 and P12295, this table will count 2 GO terms, 2 proteins, and 4 annotations.

Table S3. List of potential misannotated GO terms detected by kingdom-level analysis for UniProt-GOA release 2019-06-03, ranked in ascending order of annotation rate ratio (fifth column).

[a] "Aspect" refers to three aspects of GO terms: Molecular Function (MF), Biological Process (BP), and Cellular Component (CC). The full definition for each GO term is available at http://purl.obolibrary.org/obo/go.obo.

[b] The five integers are the number of proteins annotated with the specific GO term in the four kingdoms (Animals, Archaea, Bacteria, Fungi, and Plants), where the shaded number corresponds to the kingdom to which the GO term is potentially misannotated.

[c] "Frequency" is the number of potential misannotations associated with the most frequent evidence code, divided by all potential misannotations.

[d] "P-value" is the *p*-value of rate ratio test (Text S1). In the whole kingdom-level analysis, we perform the rate ratio test 2155 times for the 2155 GO terms with rate ratio $\theta < 0.1$, this column also reports "Q-value", which is the *p*-value after adjusting for multiple testing by controlling false discovery rate (Benjamini and Hochberg, 1995).

#	GO term (Aspect)[a]	Number of proteins for Archaea, Animals, Bacteria, Fungi, Plants[b]	Most frequent evidence code (frequency)[c]	P-value[d] (Q-value)	Rate ratio	GO term name	Examples of potential misannotations
1	GO:0030435 (BP)	0 1 276 166 0	IEA (1.00)	2.22E-16 (4.42E-13)	4.61E-4	sporulation resulting in formation of a cellular spore	Q22236
2	GO:0005634 (CC)	0 28809 6 5925 7060	IBA (1.00)	2.22E-16 (4.42E-13)	8.68E-4	nucleus	P39615 P94593 P56397 P47264 P47343 P67073
3	GO:0000329 (CC)	0 5 0 404 0	IBA (1.00)	2.22E-16 (4.42E-13)	1.35E-3	fungal-type vacuole membrane	P91354 A0A1D5PS38 Q8SWW2 Q8N1S5 Q8BWY7
4	GO:0009507 (CC)	0 0 2 2 2694	IBA (1.00)	2.22E-16 (4.42E-13)	2.00E-3	chloroplast	P56141 P00929 Q5AEN1 P00431
5	GO:0009252 (BP)	0 0 182 0 1	IBA (1.00)	2.22E-16 (4.42E-13)	2.04E-3	peptidoglycan biosynthetic process	A0A1I9LPE3
6	GO:0005730 (CC)	0 3076 2 751 552	IBA (1.00)	2.22E-16 (4.42E-13)	2.29E-3	nucleolus	O31774 O25455
7	GO:0042597 (CC)	0 0 352 17 3	IEA (1.00)	2.22E-16 (4.42E-13)	3.16E-3	periplasmic space	Q66GJ0 D8T634 D8T7W8
8	GO:0000917 (BP)	5 2 61 42 0	IEA (1.00)	3.34E-13 (6.65E-10)	4.17E-3	division septum assembly	Q9NQY0 Q9JI08

9	GO:0001216 (MF)	0 3 62 0 3	IEA (0.67)	1.88E-12 (3.74E-9)	6.15E-3	bacterial-type RNA polymerase transcriptional activator activity, sequence-specific DNA binding	F1NFT7 Q9BYN7 Q6PGC9 Q700C7 Q8LPR5 Q9FMX2
10	GO:0005739 (CC)	0 6571 18 2277 1716	IBA (1.00)	2.22E-16 (4.42E-13)	6.78E-3	mitochondrion	O32176 O34893 P32732 P36430 P39615 P45867 P52035 P96608 O25598 P56397 P47343 P47352 P41795 P67073 Q7CPZ2 Q8ZJV1 Q8ZKS2 Q8ZLT2
11	GO:0033309 (CC)	0 1 0 13 0	IBA (1.00)	3.06E-3 (1.00)	8.39E-3	SBF transcription complex	Q9N4L7
12	GO:0030907 (CC)	0 1 0 12 0	IBA (1.00)	5.48E-3 (1.00)	9.09E-3	MBF transcription complex	Q9N4L7
13	GO:0005199 (MF)	0 4 0 47 59	IEA (0.75)	1.50E-9 (2.99E-6)	9.29E-3	structural constituent of cell wall	Q3Y407 Q5FC67 A0A3Q2U6M6 P14923
14	GO:0030428 (CC)	0 5 15 56 0	IBA (1.00)	4.93E-11 (9.81E-8)	9.74E-3	cell septum	G5ECD6 B7Z099 Q8IPN4 Q9I7P4 Q9VNW7
15	GO:0005759 (CC)	0 854 4 320 134	IBA (1.00)	5.70E-11 (1.14E-7)	1.07E-2	mitochondrial matrix	P50866 O25926 Q56063 Q8ZRC0
16	GO:0005741 (CC)	0 550 2 157 64	IBA (1.00)	9.28E-6 (1.85E-2)	1.09E-2	mitochondrial outer membrane	O25090 Q93GS9
17	GO:0006696 (BP)	1 9 0 81 10	IEA (0.56)	2.71E-14 (5.39E-11)	1.21E-2	ergosterol biosynthetic process	A0A1D5PJH0 F1NFJ9 Q9VDI6 A0A1W2PQ47 E9PNM1 P37268 Q9UKR5 P53798 Q9ERY9
18	GO:0000032 (BP)	0 5 1 45 3	IBA (1.00)	2.31E-8 (4.59E-5)	1.21E-2	cell wall mannoprotein biosynthetic process	O16315 P34650 A0A1D5Q008 P34949 Q924M7
19	GO:0043231 (CC)	1 3159 5 79 220	IBA (0.80)	8.70E-12 (1.73E-8)	1.25E-2	intracellular membrane-bounded organelle	O05496 O31853 O34539 A0A0H2ZF23 P29768 Q9HLZ0

20	GO:0001525 (BP)	0 697 0 1 3	IEA (1.00)	2.97E-7 (5.90E-4)	1.26E-2	angiogenesis	Q5ANE0 Q8H0V2 D8R2U2 D8RXL5
21	GO:0005778 (CC)	0 192 1 65 47	IBA (1.00)	1.20E-2 (1.00)	1.32E-2	peroxisomal membrane	O34703
22	GO:0016575 (BP)	0 259 1 64 50	IEA (1.00)	1.32E-2 (1.00)	1.34E-2	histone deacetylation	O07595
23	GO:0032543 (BP)	0 247 4 230 28	IBA (1.00)	3.28E-07 (6.53E-4)	1.49E-2	mitochondrial translation	O30509 P36430 O25372 P47346
24	GO:0042128 (BP)	2 6 27 4 28	IEA (1.00)	5.65E-07 (1.12E-3)	1.62E-2	nitrate assimilation	Q9XTQ8 A0A1D5PZ75 P07850 Q9VWP4 P51687 Q8R086
25	GO:0010973 (BP)	0 1 0 5 0	IMP (1.00)	2.62E-1 (1.00)	2.18E-2	positive regulation of division septum assembly	Q9N4A7
26	GO:0016558 (BP)	0 50 1 36 11	IBA (1.00)	1.80E-1 (1.00)	2.38E-2	protein import into peroxisome matrix	O34703
27	GO:0005750 (CC)	0 58 1 30 23	IBA (1.00)	3.03E-1 (1.00)	2.86E-2	mitochondrial respiratory chain complex III	O26064
28	GO:0016573 (BP)	1 366 0 117 73	IEA (1.00)	5.21E-1 (1.00)	3.77E-2	histone acetylation	Q6LWX7
29	GO:0006122 (BP)	0 68 2 33 23	IEA (0.50)	5.55E-1 (1.00)	5.21E-2	mitochondrial electron transport, ubiquinol to cytochrome c	O26064 A0A0H2ZGY7
30	GO:0004402 (BP)	1 311 0 69 68	IEA (1.00)	1.00 (1.00)	6.39E-2	histone acetyltransferase activity	Q6LWX7
31	GO:0006314 (BP)	0 0 1 9 4	IEA (1.00)	1.00 (1.00)	9.52E-2	intron homing	O34479

Table S4. List of potential misannotated GO terms detected by phylum-level analysis for UniProt-GOA release 2019-06-03, ranked in ascending order of annotation rate ratio (fifth column).

[a] "Aspect" refers to three aspects of GO terms: Molecular Function (MF), Biological Process (BP), and Cellular Component (CC). The full definition for each GO term is available at http://purl.obolibrary.org/obo/go.obo.

[b] The two integers are the number of proteins annotated with the specific GO term in vertebrates and invertebrates, respective, where the shaded number corresponds to the group to which the GO term is potentially misannotated. GO:0007565 and GO:0021987 are potentially misannotated to both invertebrates and one species (*Gallus gallus*) of vertebrate.

[c] "Frequency" is the number of potential misannotations associated with the most frequent evidence code, divided by all potential misannotations.

[d] "P-value" is the *p*-value of rate ratio test (Text S1). In the whole phylum-level analysis, we perform the rate ratio test 617 times for the 617 GO terms with rate ratio $\theta < 0.1$, this column also reports "Q-value", which is the *p*-value after adjusting for multiple testing by controlling false discovery rate (Benjamini and Hochberg, 1995).

#	GO term (Aspect)[a]	Number of proteins for vertebrates, invertebrates[b]	Most frequent evidence code (frequency)[c]	P-value[d] (Q-value)	Rate ratio	GO term name	Examples of potential misannotations
1	GO:0048749 (BP)	5 214	IEA (0.40)	2.22E-16 (4.42E-13)	7.62E-3	compound eye development	F1ND76 Q9BU40 Q920C1
2	GO:0001750 (CC)	323 1	IEA (1.00)	6.98E-4 (1.00)	9.48E-3	photoreceptor outer segment	P06002
3	GO:0007565 (BP)	256 1	IBA (1.00)	4.87E-3 (1.00)	1.20E-2	female pregnancy	Q9VA76 E1C1R3 E1C688
4	GO:0006954 (BP)	1380 9	IBA (1.00)	2.24E-10 (4.47E-07)	2.00E-2	inflammatory response	M9NDW9 P08953 P15330 P98149 Q9VIA4 Q9VJX9 Q9VLE6 Q9VPH1 Q9VVJ1
5	GO:0016028 (CC)	4 60	IEA (0.50)	5.25E-4 (1.00)	2.18E-2	rhabdomere	A0A1D5PMV1 Q12866 Q60805
6	GO:0001525 (BP)	969 7	IEA (0.71)	3.96E-07 (7.88E-4)	2.21E-2	angiogenesis	Q4V5H1 Q7JRE4 Q9V3C1 Q9VPX8

7	GO:0021987 (BP)	243 2	IBA (1.00)	3.12E-2 (1.00)	2.52E-2	cerebral cortex development	G5ECG0 A0A0B4JD97 A0A1D5NW33 A0A1D5NW61 A0A1D5NZL7 A0A1D5P4G9 A0A1D5P8D7 A0A1D5P8Z8 A0A1D5PHE0 A0A1D5PI40 A0A1D5PT51 A0A1D5PTW6 A0A1D5PWG7 A0A1D5PXK9 A0A1D5PZD9 A0A1L1RTU2 E1BTY2 E1BW44 E1C6L0 E1C9F6 F1NCL7 F1NE63 F1NGP6 F1NI11 F1NJK2 F1NLP0 F1NTE7 F1NY57 F1P241 F1P2P9 O42108 P10288 P28673 Q04929 Q5ZMC9 Q5ZMT0 Q6R6I2 Q91987 Q9PTR5 Q9W601
8	GO:0035003 (CC)	5 53	IEA (0.40)	6.39E-3 (1.00)	3.08E-2	subapical complex	F1NJ32 Q86UT5 A0A0R4J0D4 Q99MJ6
9	GO:0005118 (MF)	1 8	IEA (1.00)	6.78E-1 (1.00)	4.08E-2	sevenless binding	Q8K3J9
10	GO:0044548 (MF)	74 1	IEA (1.00)	6.20E-1 (1.00)	4.13E-2	S100 protein binding	Q9W3Y3
11	GO:0042622 (CC)	46 2	IBA (1.00)	9.38E-1 (1.00)	1.28E-1	photoreceptor outer segment membrane	Q22875 Q9VEK6
12	GO:0001755 (BP)	179 10	IBA (1.00)	1.88E-1 (1.00)	1.64E-1	neural crest cell migration	Q17330 Q19764 Q95XP4 Q9TYS4 A0A0B4KG38 Q24322 Q24323 Q7KK54 Q7YU67 Q9VTT0

Table S5. List of potential misannotated GO terms detected by kingdom-level analysis for UniProt-GOA release 2018-11-06, ranked in ascending order of annotation rate ratio (fifth column).

[a] “Aspect” refers to three aspects of GO terms: Molecular Function (MF), Biological Process (BP), and Cellular Component (CC). The full definition for each GO term is available at http://purl.obolibrary.org/obo/go.obo.

[b] The five integers are the number of proteins annotated with the specific GO term in the five kingdoms (Archaea, Animals, Bacteria, Fungi, and Plants), where the shaded number corresponds to the kingdom to which the GO term is potentially misannotated.

[c] “Frequency” is the number of potential misannotations associated with the most frequent evidence code, divided by all potential misannotations.

[d] “P-value” is the p-value of rate ratio test (Text S1). In the whole kingdom-level analysis, we perform the rate ratio test 2206 times for the 2206 GO terms with rate ratio $\theta < 0.1$, this column also reports “Q-value”, which is the p-value after adjusting for multiple testing by controlling false discovery rate (Benjamini and Hochberg, 1995).

#	GO term (Aspect)[a]	Number of proteins for Archaea, Animals, Bacteria, Fungi, Plants[b]	Most frequent evidence code (frequency)[c]	P-value[d] (Q-value)	Rate ratio	GO term name	Examples of potential misannotations
1	GO:0030435 (BP)	0 1 276 166 0	IEA (1.00)	2.22E-16 (4.42E-13)	4.75E-3	sporulation resulting in formation of a cellular spore	Q22236
2	GO:0005634 (CC)	0 28653 5 5657 7077	IBA (1.00)	2.22E-16 (4.42E-13)	7.58E-3	nucleus	P39615 P12295 P56397 P47343 P67073
3	GO:0009507 (CC)	0 0 1 2 2676	IBA (1.00)	2.22E-16 (4.42E-13)	1.01E-3	chloroplast	P0DM85 Q5AEN1 P00431
4	GO:0009506 (CC)	0 5 0 0 1019	IBA (1.00)	2.22E-16 (4.42E-13)	1.73E-3	plasmodesma	Q9TZM3 F1NRK4 A0A0B4KHT3 Q7JXU8 Q80VQ1
5	GO:0000329 (CC)	0 5 0 323 0	IBA (1.00)	2.22E-16 (4.42E-13)	1.74E-3	fungal-type vacuole membrane	P91354 A0A1D5PS38 Q8SWW2 Q8N1S5 Q8BWY7
6	GO:0009252 (BP)	0 0 185 0 1	IBA (1.00)	2.22E-16 (4.42E-13)	2.00E-3	peptidoglycan biosynthetic process	A0A1I9LPE3
7	GO:0005730 (CC)	0 3142 2 742 552	IBA (1.00)	2.22E-16 (4.42E-13)	2.31E-3	nucleolus	O31774 O25455
8	GO:0042597 (CC)	0 0 348 10 3	IEA (1.00)	2.22E-16 (4.42E-13)	3.19E-3	periplasmic space	Q66GJ0 D8T7W8 D8T634
9	GO:0045944 (BP)	0 4504 2 381 261	IEA (1.00)	2.22E-16 (4.42E-13)	3.39E-3	positive regulation of transcription by RNA polymerase II	O34482 Q8ZLD3

10	GO:0000790 (CC)	0 845 2 421 38	IBA (1.00)	2.22E-16 (4.42E-13)	4.07E-3	nuclear chromatin	O34482 Q8ZLD3
11	GO:0000917 (BP)	5 2 61 50 0	IEA (1.00)	7.20E-13 (1.43E-9)	4.30E-3	division septum assembly	Q9NQY0 Q9JI08
12	GO:0030907 (CC)	0 1 2 24 0	IBA (1.00)	5.70E-06 (1.13E-2)	4.68E-3	MBF transcription complex	Q9N4L7 O34482 Q8ZLD3
13	GO:0005618 (CC)	5 11 46 165 735	IBA (1.00)	2.22E-16 (4.42E-13)	5.29E-3	cell wall	H8ESF9 Q7K705 H8ESF7 H8ESF8 Q9U1R8 H8ESF6 H8ESG0 F1NZI4 Q9VKD9 Q32M88 Q8BP56
14	GO:0005199 (MF)	0 3 0 47 58	IEA (0.67)	4.45E-10 (8.87E-7)	7.17E-3	structural constituent of cell wall	Q3Y407 Q5FC67 P14923
15	GO:0000978 (MF)	0 1734 2 221 38	IBA (1.00)	1.51E-8 (3.01E-5)	7.76E-3	RNA polymerase II proximal promoter sequence-specific DNA binding	O34482 Q8ZLD3
16	GO:0001228 (MF)	0 1656 2 80 52	IBA (1.00)	2.19E-7 (4.37E-4)	9.22E-3	DNA-binding transcription activator activity, RNA polymerase II-specific	O34482 Q8ZLD3

17	GO:0009245 (BP)	0 6 85 4 29	IBA (1.00)	1.11E-14 (2.21E-11)	9.25E-3	lipid A biosynthetic process	Q9U241 Q20122 E1C4C0 Q94519 O14561 Q9CR21 A0A1D8PDT0 Q5AHH7 Q10217 P32463 Q9FGJ4 P53665 O80800 Q8VZA5 F4IAT8 F4IF99 P0DKB8 F4JGP6 F4JIP6 Q9SU91 P0DKB9 Q8LEA0 P0DKB7 F4IAW1 A0A1I9LRA2 A0A1I9LRA1 A0A2H1ZEC5 A0A1P8B4F5 A0A1P8B4E8 A0A1I9LRA0 F4JEP7 D8R888 D8SHE4 D8QUJ6 D8T1B0 D8T5N1 D8QYK7 D8QQP3 D8T1A9
18	GO:0005739 (CC)	0 6557 22 1966 1792	IBA (1.00)	2.22E-16 (4.42E-13)	9.60E-3	mitochondrion	P96608 P39615 O34893 P45867 O32176 P32732 P36430 P0AGL5 P12295 P37686 P76553 P32099 P60340 P56397 P47343 P47352 P67073 P41795 Q8ZJV1 Q8ZLT2 Q7CPZ2 Q8ZKS2
19	GO:0007507 (BP)	0 915 0 1 0	IEA (1.00)	8.11E-4 (1.00)	9.73E-3	heart development	Q59TT8
20	GO:0030428 (CC)	0 5 15 56 0	IBA (1.00)	1.00E-10 (2.00E-7)	1.00E-2	cell septum	G5ECD6 Q9I7P4 Q9VNW7 B7Z099 Q8IPN4

21	GO:0001525 (BP)	0 737 0 1 3	IEA (1.00)	4.57E-8 (9.10E-5)	1.15E-2	angiogenesis	Q5ANE0 Q8H0V2 D8R2U2 D8RXL5
22	GO:0043231 (CC)	1 3162 5 66 204	IBA (0.80)	2.90E-12 (5.77E-9)	1.21E-2	intracellular membrane-bounded organelle	O05496 O31853 O34539 A0A0H2ZF23 P29768 Q9HLZ0
23	GO:0006696 (BP)	1 9 0 81 10	IEA (0.56)	7.37E-14 (1.47E-10)	1.25E-2	ergosterol biosynthetic process	F1NFJ9 A0A1D5PJH0 Q9VDI6 P37268 Q9UKR5 E9PNM1 A0A1W2PQ47 Q9ERY9 P53798
24	GO:0000032 (BP)	0 5 2 45 2	IBA (1.00)	4.03E-8 (8.02E-5)	1.25E-2	cell wall mannoprotein biosynthetic process	P34650 O16315 A0A1D5Q008 P34949 Q924M7
25	GO:0010973 (BP)	0 1 0 9 0	IMP (1.00)	3.40E-2 (1.00)	1.25E-2	positive regulation of division septum assembly	Q9N4A7
26	GO:0005778 (CC)	0 190 1 65 48	IBA (1.00)	1.20E-2 (1.00)	1.32E-2	peroxisomal membrane	O34703
27	GO:0016575 (BP)	0 238 1 64 56	IEA (1.00)	1.32E-2 (1.00)	1.34E-2	histone deacetylation	O07595
28	GO:0005741 (CC)	0 570 2 124 65	IBA (1.00)	2.26E-4 (4.51E-1)	1.38E-2	mitochondrial outer membrane	O25090 Q93GS9
29	GO:0032543 (BP)	0 245 4 233 28	IBA (1.00)	2.45E-7 (4.88E-4)	1.47E-2	mitochondrial translation	O30509 P36430 O25372 P47346
30	GO:0042128 (BP)	2 6 46 4 46	IEA (1.00)	1.49E-6 (2.97E-3)	1.71E-2	nitrate assimilation	Q9XTQ8 P07850 A0A1D5PZ75 Q9VWP4 P51687 Q8R086
31	GO:0005759 (CC)	0 857 6 283 136	IBA (1.00)	6.71E-8 (1.34E-4)	1.82E-2	mitochondrial matrix	P50866 P0A6H1 P31660 O25926 Q56063 Q8ZRC0
32	GO:0016558 (BP)	0 53 1 37 14	IBA (1.00)	1.64E-1 (1.00)	2.32E-2	protein import into peroxisome matrix	O34703
33	GO:0006122 (BP)	0 71 1 33 24	IEA (1.00)	2.34E-1 (1.00)	2.60E-2	mitochondrial electron transport, ubiquinol to cytochrome c	A0A0H2ZGY7

34	GO:0016573 (BP)	1 366 0 119 80	IEA (1.00)	5.02E-1 (1.00)	3.70E-2	histone acetylation	Q6LWX7
35	GO:0001654 (BP)	0 147 0 1 0	IEA (1.00)	1.00 (1.00)	6.05E-2	eye development	Q59TT8
36	GO:0004402 (BP)	1 311 0 69 68	IEA (1.00)	1.00 (1.00)	6.37E-2	histone acetyltransferase activity	Q6LWX7
37	GO:0006314 (BP)	0 0 1 9 4	IEA (1.00)	1.00 (1.00)	9.53E-2	intron homing	O34479

Table S6. List of potential misannotated GO terms detected by phylum-level analysis for UniProt-GOA release 2018-11-06, ranked in ascending order of annotation rate ratio (fifth column).

[a] "Aspect" refers to three aspects of GO terms: Molecular Function (MF), Biological Process (BP), and Cellular Component (CC). The full definition for each GO term is available at http://purl.obolibrary.org/obo/go.obo.

[b] The two integers are the number of proteins annotated with the specific GO term in vertebrates and invertebrates, respectively, where the shaded number corresponds to the group to which the GO term is potentially misannotated. GO:0021987 and GO:0007565 are potentially misannotated to both invertebrates and one species (*Gallus gallus*) of vertebrate.

[c] "Frequency" is the number of potential misannotations associated with the most frequent evidence code, divided by all potential misannotations.

[d] "P-value" is the *p*-value of rate ratio test (Text S1). In the whole phylum-level analysis, we perform the rate ratio test 650 times for the 650 GO terms with rate ratio $\theta < 0.1$, this column also reports "Q-value", which is the *p*-value after adjusting for multiple testing by controlling false discovery rate (Benjamini and Hochberg, 1995).

#	GO term (Aspect)[a]	Number of proteins for vertebrates, invertebrates[b]	Most frequent evidence code (frequency)[c]	P-value[d] (Q-value)	Rate ratio	GO term name	Examples of potential misannotations
1	GO:0048749 (BP)	5 225	IEA (0.40)	2.22E-16 (4.42E-13)	7.52E-3	compound eye development	F1ND76 Q9BU40 Q920C1
2	GO:0001525 (BP)	1023 3	IEA (1.00)	2.26E-11 (4.49E-08)	8.67E-3	angiogenesis	Q9V3C1 Q9VPX8
3	GO:0006954 (BP)	1411 9	IBA (1.00)	2.70E-11 (5.37E-08)	1.89E-2	inflammatory response	P08953 P15330 P98149 M9NDW9 Q9VLE6 Q9VIA4 Q9VPH1 Q9VJX9 Q9VVJ1
4	GO:0001701 (BP)	755 6	IBA (1.00)	1.03E-05 (2.05E-2)	2.35E-2	in utero embryonic development	Q93212 A8JQY3 R9PY60 Q9VVY7 Q15KK8 Q4V495
5	GO:0001750 (CC)	350 1	IEA (1.00)	2.18E-4 (4.34E-1)	8.45E-3	photoreceptor outer segment	P06002
6	GO:0016028 (CC)	4 60	IEA (0.50)	7.78E-4 (1.00)	2.26E-2	rhabdomere	F1P3V0 Q12866 Q60805
7	GO:0007565 (BP)	261 1	IBA (1.00)	3.23E-3 (1.00)	1.13E-2	female pregnancy	E1C1R3 E1C688 Q9VA76
8	GO:0035003 (CC)	5 53	IEA (0.40)	8.73E-3 (1.00)	3.19E-2	subapical complex	F1NJ32 Q86UT5 Q99MJ6 A0A0R4J0D4

9	GO:0021987 (BP)	282 2	IBA (1.00)	8.90E-3 (1.00)	2.10E-2	cerebral cortex development	G5ECG0 A0A0B4JD97 Q9PTR5 Q5ZMC9 Q4JIM4 P10288 Q91987 P28673 A0A1D5P4G9 F1NTE7 F1NGP6 A0A1D5PT51 F1NI11 A0A1D5PWG7 E1C8S5 A0A1D5P8D7 A0A1D5P8Z8 F1NCL7 F1NLP0 A0A1D5P310 A0A1D5NW61 E1BTY2 A0A1D5PXK9 A0A1D5NZL7 A0A1D5NW33 F1NY57 F1NE63 A0A1D5PHE0 A0A1D5PZD9 A0A1D5PI40 E1BW44 E1BT91 Q6R6I2 R4GFT9 Q9W601
10	GO:0044548 (MF)	75 1	IEA (1.00)	5.70E-1 (1.00)	3.94E-2	S100 protein binding	Q9W3Y3
11	GO:0005118 (MF)	1 7	IEA (1.00)	8.48E-1 (1.00)	4.83E-2	sevenless binding	Q8K3J9
12	GO:0042622 (CC)	46 2	IBA (1.00)	9.38E-1 (1.00)	1.28E-1	photoreceptor outer segment membrane	Q22875 Q9VEK6
13	GO:0001755 (BP)	179 10	IBA (1.00)	1.88E-1 (1.00)	1.64E-1	neural crest cell migration	Q95XP4 Q17330 Q19764 Q9TYS4 Q24323 Q24322 A0A0B4KG38 Q9VTT0 Q7KK54 Q7YU67

Table S7. List of potential misannotated GO terms for nucleus and mitochondrion in the full UniProt-GOA release 2018-11-06 and 2019-06-03. The taxon labels (bacteria, archaea, and viruses) of each protein is assigned by UniProt (ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/taxonomic_divisions/). It should be noted that GO:0005634 "nucleus" and GO:0005739 "mitochondrion" are inappropriate GO terms for proteins from virus, bacteria or archaea even if the annotated proteins can enter the nucleus and mitochondria of eukaryote hosts. Such proteins should be annotated with GO:0042025 "host cell nucleus" or GO:0033650 "host cell mitochondria" instead (Deegan, et al., 2010).

Potentially misannotated GO terms	Potentially misannotated proteins, 453 in total (UniProt-GOA release 2018-11-06)	Potentially misannotated proteins, 466 in total (UniProt-GOA release 2019-06-03)
GO:0005634 "nucleus" in bacteria, archaea and viruses	A0A2P9AE88, A0A2P9AEW8, A0A2P9AFX3, A0A2P9AGY9, A0A2P9AGZ6, A0A2P9AJC7, A0A2P9AKG6, A0A2P9AQA5, A0A2P9ARQ2, A0A2P9ASD7, A0A2P9AWP5, A9A2H1, A9A2S5, A9A2U7, A9A3M8, A9A4A1, A9A4S8, A9A4Y2, A9A585, A9A5I9, A9A5J6, A9A5V1, A9WFD7, A9WJV9, B1L3G6, B1L4T1, B1L650, B1L6P0, B1L6S7, B1L6U4, B1L6X8, B1L727, B8E084, K1ZNZ6, O67010, O67467, O69822, O74023, O84613, O86365, O93728, P12295, P15009, P39615, P43731, P47343, P54034, P56397, P57697, P57705, P67073, P74153, P9WFQ9, Q2G0J7, Q58356, Q58554, Q58634, Q58924, Q59044, Q59046, Q5JDH5, Q5JDJ9, Q5JE24, Q5JF27, Q5JFZ0, Q5JG19, Q5JGP6, Q5JHL8, Q5JIB3, Q5JIU9, Q60177, Q6ZEA1, Q74AS6, Q74DZ2, Q7LXZ1, Q7NK15, Q7UMR5, Q814M9, Q81AY6, Q83BN8, Q83CW4, Q89WV9, Q8A0C8, Q8A0C9, Q8A0E8, Q8A0M7, Q8A0N0, Q8A0N1, Q8A0Q6, Q8A0Q8, Q8A0Q9, Q8A113, Q8A182, Q8A193, Q8A1X8, Q8A1Y1, Q8A325, Q8A3K6, Q8A4H0, Q8A5N9, Q8A5V6, Q8A5X5, Q8A5Y1, Q8A6K3, Q8A6V7, Q8A8Y5, Q8CJP8, Q8DN32, Q8DPQ4, Q8EB78, Q8EFK2, Q8EHW4, Q8EZN4, Q8P357, Q8P4D7, Q8P5J8, Q8P5S2, Q8P6S1, Q8PBN8, Q8PCT1, Q8R634, Q8TGX6, Q8TJB5, Q8TJB9, Q8TK57, Q8TL35, Q8TMY4, Q8TPW5, Q8TPX5, Q8TRU1, Q8TST2, Q8Y7P6, Q8Y9X7, Q8ZD85, Q8ZST5, Q8ZSU8, Q8ZU12, Q8ZVM1, Q8ZWI0, Q8ZWR6, Q8ZWT5, Q8ZYC3, Q8ZYF2, Q8ZYT6, Q93JJ2, Q97U95, Q97W57, Q97WD6, Q97XC0, Q97ZH0, Q97ZJ8, Q980L4, Q9EX12, Q9HMK5, Q9HPP4, Q9HQ62, Q9HQ94, Q9HR36, Q9HS90, Q9HSK2, Q9HT99, Q9HTW1, Q9HYQ0, Q9HYT8, Q9I5H9, Q9JZA1, Q9K3Z0, Q9KKY9, Q9KPK8, Q9RIU7, Q9RWH9, Q9RWU9, Q9UXA0, Q9UXC6, Q9UXF3, V9H131. (176 in total)	A5I268, A5I2D8, A9A2H1, A9A2S5, A9A2U7, A9A310, A9A3M8, A9A4A1, A9A4S8, A9A4Y2, A9A585, A9A5I9, A9A5J6, A9A5V1, A9WFD7, A9WJV9, B1L3G6, B1L4T1, B1L650, B1L6L9, B1L6P0, B1L6S7, B1L6U4, B1L6X8, B1L727, B5YJR8, B8E084, I6YCF3, O67010, O67467, O69822, O74023, O84559, O84613, O84714, O86365, O93728, P15009, P39615, P43731, P47264, P47343, P54034, P56397, P57697, P57705, P67073, P74153, P74552, P94593, P9WFQ9, Q2FYC2, Q2G0J7, Q57809, Q58356, Q58371, Q58554, Q58634, Q58884, Q58924, Q59044, Q59046, Q5JDH5, Q5JDJ9, Q5JE24, Q5JEJ0, Q5JF27, Q5JFZ0, Q5JG19, Q5JGP6, Q5JGW1, Q5JHL8, Q5JIB3, Q5JIT1, Q5JIU9, Q60177, Q60275, Q6ZEA1, Q74AS6, Q74DZ2, Q7LXZ1, Q7NHW7, Q7NIB7, Q7NK15, Q7UGF2, Q7ULR2, Q7UMR5, Q7UZE8, Q814M9, Q815C0, Q81AY6, Q81F57, Q83BN8, Q83CW4, Q89WV9, Q8A0C8, Q8A0C9, Q8A0E8, Q8A0M7, Q8A0N0, Q8A0N1, Q8A0Q6, Q8A0Q8, Q8A0Q9, Q8A113, Q8A182, Q8A193, Q8A1X8, Q8A1Y1, Q8A2F2, Q8A325, Q8A3K6, Q8A4H0, Q8A5N9, Q8A5V6, Q8A5X5, Q8A5Y1, Q8A6K3, Q8A6V7, Q8A8Y5, Q8CJP8, Q8CK37, Q8DN32, Q8DP39, Q8DPQ4, Q8EB78, Q8EFK2, Q8EGF3, Q8EHW4, Q8EZN4, Q8P357, Q8P4D7, Q8P5J8, Q8P5S2, Q8P6S1, Q8PAM7, Q8PBN8, Q8PCT1, Q8R634, Q8RDV9, Q8REE7, Q8TGX6, Q8TJB5, Q8TJB9, Q8TJF6, Q8TK57, Q8TL35, Q8TMY4, Q8TPW5, Q8TPX5, Q8TRU1, Q8TST2, Q8TSW4, Q8TU84, Q8Y6P0, Q8Y7P6, Q8Y9X7, Q8ZD85, Q8ZST5, Q8ZSU8, Q8ZU12, Q8ZVM1, Q8ZWI0, Q8ZWR6, Q8ZWT5, Q8ZY88, Q8ZYC3, Q8ZYF2, Q8ZYT6, Q93JJ2, Q97U95, Q97W57, Q97WD6, Q97XC0, Q97XQ7, Q97ZH0, Q97ZJ8, Q980L4, Q9EX12, Q9HMK5, Q9HNA5, Q9HPP4, Q9HQ62, Q9HQ94, Q9HR36, Q9HS90, Q9HSK2, Q9HT99, Q9HTW1, Q9HYQ0, Q9HYT8, Q9I5D9, Q9I5H9, Q9JZA1, Q9K3Z0, Q9KKY9, Q9KPK8, Q9RIU7, Q9RKT0, Q9RUX1, Q9RWH9, Q9RWU9, Q9UXA0, Q9UXC6, Q9UXF3, Q9UXG1. (206 in total)

GO:0005739 "mitochondrion" in bacteria, archaea and viruses	A0A2P9A9V0, A0A2P9A9X9, A0A2P9A9Y2, A0A2P9AA00, A0A2P9AAB8, A0A2P9AAI9, A0A2P9AAN2, A0A2P9ACG1, A0A2P9ACU1, A0A2P9ACW5, A0A2P9AEN5, A0A2P9AER5, A0A2P9AER6, A0A2P9AET0, A0A2P9AET9, A0A2P9AFK8, A0A2P9AFT0, A0A2P9AGY9, A0A2P9AHK1, A0A2P9AI60, A0A2P9AIF0, A0A2P9AIG8, A0A2P9AJA8, A0A2P9AJB7, A0A2P9AJI2, A0A2P9AKA2, A0A2P9AKD0, A0A2P9AKD6, A0A2P9AKG6, A0A2P9AKJ0, A0A2P9ALM2, A0A2P9ALR4, A0A2P9ALU5, A0A2P9AP18, A0A2P9AP39, A0A2P9APA4, A0A2P9ASD7, A0A2P9ASG2, A0A2P9ASI0, A0A2P9ASK3, A0A2P9ASL6, A0A2P9ASL7, A0A2P9ASM8, A0A2P9ASS6, A0A2P9ASY7, A0A2P9ATS8, A0A2P9ATV8, A0A2P9AU47, A0A2P9AU57, A0A2P9AV08, A0A2P9AV43, A0A2P9AVE5, A0A2P9AVH2, A0A2P9AVI6, A0A2P9AWL7, A0A2P9AWX6, A0A2P9AWY7, A0A2P9AX04, A0A2P9AX27, A5I3W7, A5I4J0, A5I554, A5I5S8, A5I6S8, A5I713, A5I7Q0, A9A423, A9WC41, A9WGQ0, A9WI67, A9WIG2, A9WKF8, B1L5Q3, B1L5U2, B5YJI4, B5YJJ5, B8E2Z1, H7C6H6, I6Y3Q0, I6YCF5, K1YIU2, K1YMB2, K1YWP4, K1ZNE0, K2AGS1, K2AWD2, K2BL15, K2BNG3, K2BUF1, K2DYD3, K4PSE0, O32176, O34893, O66922, O68560, O84096, O84613, O84849, P0AGL5, P0C6Q0, P12295, P32099, P32732, P36430, P37686, P39615, P41795, P43731, P45142, P45867, P47343, P47352, P54225, P56397, P59883, P60340, P60344, P67073, P72154, P73505, P73885, P74696, P76553, P94317, P95281, P95968, P96414, P96608, P96845, P9WFQ9, P9WHP7, P9WQG1, Q0WHU3, Q2FVW7, Q2FXH2, Q2G0J7, Q2G2Q3, Q4AAX7, Q58337, Q58772, Q5JH16, Q5JIP5, Q7AKM9, Q7CPZ2, Q7MBB8, Q7NEL3, Q7NF75, Q7NJV3, Q812X9, Q814M9, Q814S9, Q815X2, Q816T0, Q81D89, Q81DR7, Q83C30, Q83CW4, Q83CX8, Q89CI0, Q89CJ6, Q89E18, Q89E40, Q89EH7, Q89FZ0, Q89GR2, Q89IM8, Q89LE2, Q89LR5, Q89LX4, Q89P26, Q89PP3, Q89Q26, Q89R25, Q89RT4, Q89VM3, Q89VT7, Q89WB1, Q89WU0, Q89Y17, Q8A2U2, Q8A5V6, Q8A9D3, Q8CWR2, Q8CXU1, Q8DP22, Q8DPQ4, Q8DRB6, Q8E7Z9, Q8E8M3, Q8EB78, Q8EDK8, Q8EFR9, Q8EFS0, Q8EGV1, Q8EGW7, Q8EYU6, Q8F718, Q8F8J5, Q8F8Y4, Q8P4D7, Q8P5Y6, Q8P7U9, Q8P949, Q8PAG9, Q8PAL6, Q8PBS3, Q8PCN7, Q8PDT9, Q8R5V8, Q8R5X8, Q8R634, Q8R674, Q8RIN6, Q8TJX5, Q8TM42, Q8TT52, Q8Y6M4, Q8Y7F3, Q8Y7P6, Q8Y9X7, Q8ZBC4, Q8ZD85, Q8ZDY2, Q8ZJV1, Q8ZKS2, Q8ZLT2, Q8ZU33, Q8ZUR4, Q8ZVY0, Q8ZXA3, Q93JE3, Q93LE9, Q97UW6, Q97V73, Q9EX12, Q9FCA4, Q9HP80, Q9HPT2, Q9HQF0, Q9HQI1, Q9HRB3, Q9HRI6, Q9HS75, Q9HU39, Q9HUI6, Q9HX77, Q9HZT2, Q9I0T2, Q9I1C3, Q9I296, Q9I2R6, Q9I4I9, Q9I4L3, Q9I4V4, Q9I567, Q9I5H9, Q9I714, Q9JYY1, Q9JZA1, Q9JZL9, Q9K028, Q9K3Z0, Q9KPK8, Q9KS71, Q9KSA9, Q9KU78, Q9RD27, Q9RJX3, Q9RKY7, Q9RL06, Q9RU50, Q9RUQ9, Q9RUX5, Q9RWH9, Q9RY41, Q9WZW0, Q9Z528. (277 in total)	A5I386, A5I3W7, A5I4J0, A5I554, A5I5S8, A5I6S8, A5I713, A5I7Q0, A9A2B1, A9A4I7, A9A5U1, A9WB73, A9WC41, A9WD13, A9WGQ0, A9WI67, A9WIG2, A9WKB1, A9WKF8, B1L3C1, B1L494, B1L5F5, B1L5Q3, B1L6S1, B5YJI4, B5YJJ5, B8E0G5, B8E2Z1, H7C6H6, I6Y3Q0, I6YCF5, K4PSE0, O25598, O32176, O34893, O66922, O68560, O84096, O84613, O84849, P0C6Q0, P32732, P36430, P39615, P41795, P43731, P45142, P45867, P47343, P47352, P52035, P54225, P56397, P59883, P60344, P61414, P67073, P72154, P73505, P73824, P73842, P73885, P74250, P74696, P94317, P95281, P96414, P96608, P96845, P9WFQ9, P9WHP7, P9WQG1, Q0WHU3, Q2FUZ6, Q2FVW7, Q2FXH2, Q2FYZ0, Q2G0J7, Q2G2Q3, Q4AAX7, Q55806, Q58130, Q58337, Q58597, Q58991, Q5JDB8, Q5JEP5, Q5JFV8, Q5JGC8, Q5JIP5, Q60363, Q7AKM9, Q7CPZ2, Q7MBB8, Q7NCU0, Q7NE37, Q7NEL3, Q7NF75, Q7NJV3, Q7NLP7, Q7UW63, Q812X9, Q814M9, Q814S9, Q815X2, Q816T0, Q81D89, Q81DR7, Q81E75, Q83C30, Q83CW4, Q83CX8, Q89CI0, Q89CJ6, Q89E18, Q89E40, Q89EH7, Q89FG8, Q89FZ0, Q89GM4, Q89GR2, Q89IM8, Q89J27, Q89LE2, Q89LR5, Q89LX4, Q89P26, Q89PP3, Q89Q26, Q89R25, Q89RJ1, Q89RM2, Q89RT4, Q89VM3, Q89VT7, Q89WB1, Q89WU0, Q89Y17, Q8A0Q0, Q8A2U2, Q8A5V6, Q8A838, Q8A9D3, Q8CWR2, Q8CXU1, Q8DP22, Q8DPQ4, Q8DRB6, Q8E7Z9, Q8E8M3, Q8EB78, Q8EDK8, Q8EFR9, Q8EFS0, Q8EGN7, Q8EGV1, Q8EGW7, Q8EH27, Q8EYB7, Q8EYP2, Q8EYU6, Q8F718, Q8F8J5, Q8F8Y4, Q8P4D7, Q8P5Y6, Q8P7U9, Q8P949, Q8PAG9, Q8PAL6, Q8PAR5, Q8PBS3, Q8PCN7, Q8PDT9, Q8R5V8, Q8R5X8, Q8R634, Q8R674, Q8RIN6, Q8TJN2, Q8TJX5, Q8TKX4, Q8TLX7, Q8TM42, Q8Y6M4, Q8Y7F3, Q8Y7P6, Q8Y8C5, Q8Y9X7, Q8ZBC4, Q8ZD85, Q8ZDY2, Q8ZJV1, Q8ZKS2, Q8ZLT2, Q8ZU06, Q8ZUR4, Q8ZVY0, Q8ZW34, Q8ZWK4, Q8ZXA3, Q8ZYM4, Q93JE3, Q97U19, Q97UW6, Q97V73, Q97VW8, Q980D1, Q980V1, Q9EX12, Q9FCA4, Q9HNH7, Q9HP27, Q9HP80, Q9HPT2, Q9HQF0, Q9HQR6, Q9HRB3, Q9HRI6, Q9HS75, Q9HU39, Q9HUI6, Q9HX77, Q9HZT2, Q9I017, Q9I0T2, Q9I1C3, Q9I296, Q9I2R6, Q9I4I9, Q9I4L3, Q9I4V4, Q9I567, Q9I5A2, Q9I5H9, Q9I714, Q9JYY1, Q9JZA1, Q9JZL9, Q9K028, Q9K3Z0, Q9KPK8, Q9KS71, Q9KSA9, Q9KU78, Q9KZY2, Q9RD27, Q9RJX3, Q9RKY7, Q9RL06, Q9RTT2, Q9RU50, Q9RUQ9, Q9RUX5, Q9RWH9, Q9RY41, Q9UXB2, Q9WY58, Q9WZW0, Q9Z528. (260 in total)

Supporting Figures

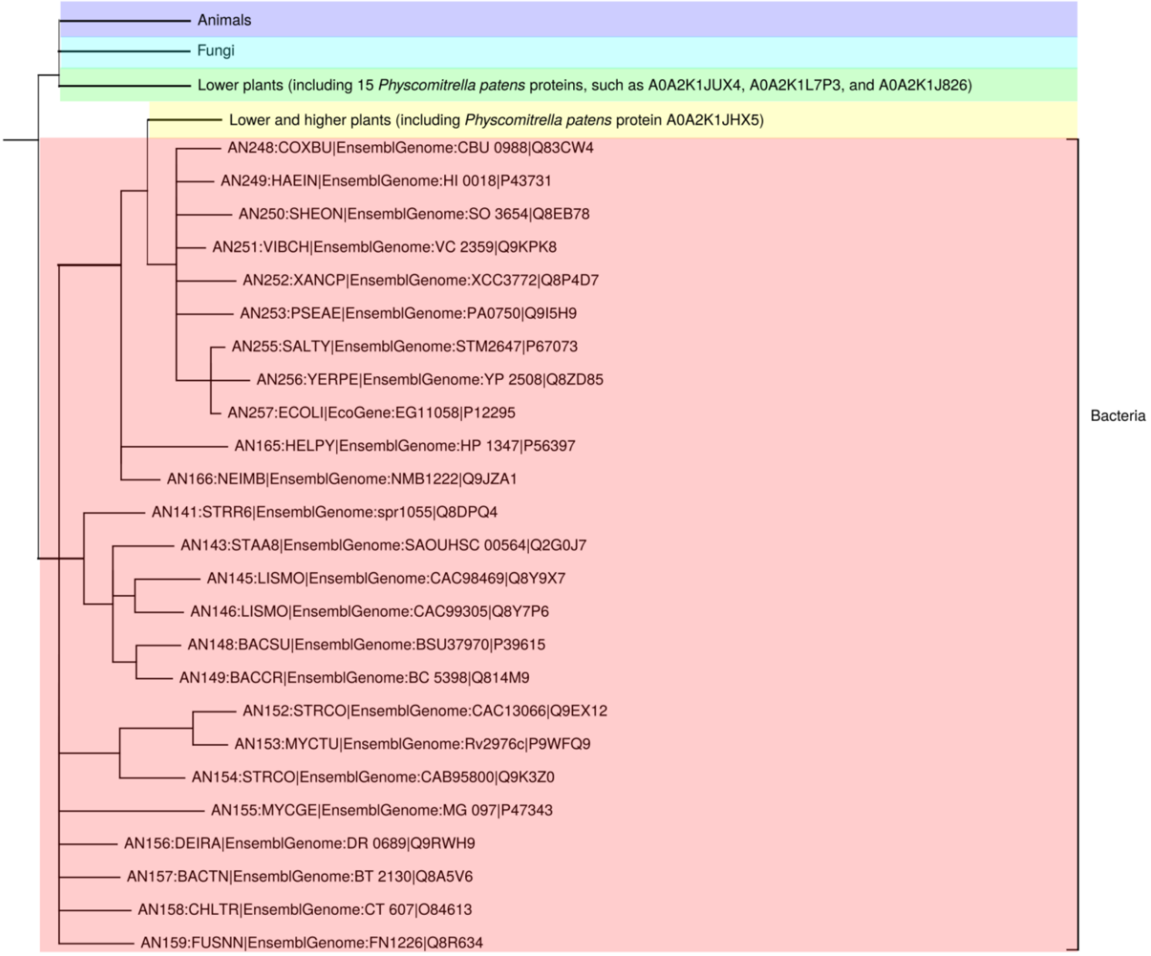
Animals
Fungi
Lower plants (including 15 *Physcomitrella patens* proteins, such as A0A2K1JUX4, A0A2K1L7P3, and A0A2K1J826)
Lower and higher plants (including *Physcomitrella patens* protein A0A2K1JHX5)
AN248:COXBU|EnsemblGenome:CBU 0988|Q83CW4
AN249:HAEIN|EnsemblGenome:HI 0018|P43731
AN250:SHEON|EnsemblGenome:SO 3654|Q8EB78
AN251:VIBCH|EnsemblGenome:VC 2359|Q9KPK8
AN252:XANCP|EnsemblGenome:XCC3772|Q8P4D7
AN253:PSEAE|EnsemblGenome:PA0750|Q9I5H9
AN255:SALTY|EnsemblGenome:STM2647|P67073
AN256:YERPE|EnsemblGenome:YP 2508|Q8ZD85
AN257:ECOLI|EcoGene:EG11058|P12295
AN165:HELPY|EnsemblGenome:HP 1347|P56397
AN166:NEIMB|EnsemblGenome:NMB1222|Q9JZA1
AN141:STRR6|EnsemblGenome:spr1055|Q8DPQ4
AN143:STAA8|EnsemblGenome:SAOUHSC 00564|Q2G0J7
AN145:LISMO|EnsemblGenome:CAC98469|Q8Y9X7
AN146:LISMO|EnsemblGenome:CAC99305|Q8Y7P6
AN148:BACSU|EnsemblGenome:BSU37970|P39615
AN149:BACCR|EnsemblGenome:BC 5398|Q814M9
AN152:STRCO|EnsemblGenome:CAC13066|Q9EX12
AN153:MYCTU|EnsemblGenome:Rv2976c|P9WFQ9
AN154:STRCO|EnsemblGenome:CAB95800|Q9K3Z0
AN155:MYCGE|EnsemblGenome:MG 097|P47343
AN156:DEIRA|EnsemblGenome:DR 0689|Q9RWH9
AN157:BACTN|EnsemblGenome:BT 2130|Q8A5V6
AN158:CHLTR|EnsemblGenome:CT 607|O84613
AN159:FUSNN|EnsemblGenome:FN1226|Q8R634
Bacteria

Figure S1. Phylogenetic tree of the Udg protein family (PANTHER database ID: PTN000137400). Each leaf node shows the PANTHER tree node ID, species ID, database ID, and UniProt ID of the protein. Four branches containing eukaryotic species of animals (purple), fungi (cyan), plants (green and yellow) are shown in Figure S2 to Figure S5.

From the point view of protein evolution, this phylogenetic tree was probably correctly constructed to group some plant proteins together with proteobacterial ones while other plant proteins with animal and fungal ones. In fact, the moss *Physcomitrella patens subsp. patens*, a lower plant, has 16 copies of Udg paralogs scattered across 11 chromsomes. 1 copy (UniProt ID: A0A2K1JHX5, Figure S5) is located in the "Lower and higher plant" branch (yellow), which is a branch within the bacteria sub-tree (red), and the other 15 copies (UniProt IDs: A0A2K1JUX4, A0A2K1L7P3, A0A2K1J826, A0A2K1JYM2, A0A2K1KFR1, A0A2K1JDR7, A0A2K1K3K7, A0A2K1L6F5, A0A2K1KLV6, A0A2K1K2Z2, A0A2K1IY43, A0A2K1KPV2, A9RL80, A0A2K1K656, and A0A2K1I9L1, Figure S4) are in the "Lower plant" branch, which is a branch in the eukaryote-only sub-tree consisting of "Animals", "Fungi", and "Lower plant". Since characterized proteins in this family are known to be associated with both mitochondrion and nucleus, it is possible that Udg proteins in plant cells have two origins: one plant Udg gene copy was originally incorporated from the mitochondrion, which is from exogeneous prokaryotes according to the endosymbiotic theory (Sagan, 1967); other Udg paralogs are endogenous. Compared to the exogenous paralog, the endogenous Udg paralogs are evolutionarily closer to other eukaryotic orthologs in animals and fungi. Through the course of plant evolution, the endogenous paralog may be lost through time, and only the mitochondrion-originated prokaryote-like Udg gene become the dominant Udg in higher plants (Figure S5) such as *Arabidopsis thaliana*, whose Udg protein (TAIR:2086904) was one of the orthologs used to annotate other Udg proteins.

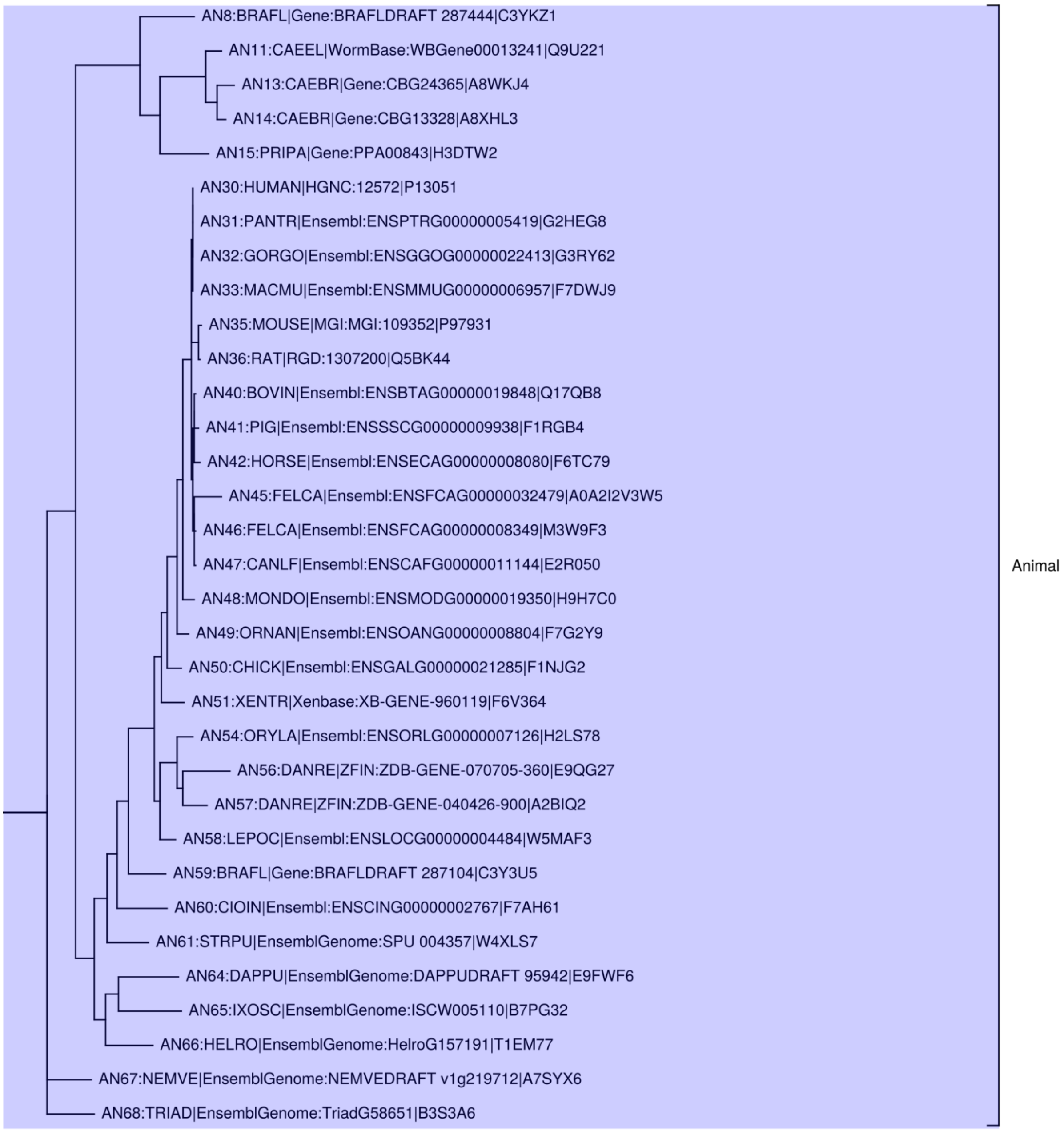
AN8:BRAFL|Gene:BRAFLDRAFT 287444|C3YKZ1
AN11:CAEEL|WormBase:WBGene00013241|Q9U221
AN13:CAEBR|Gene:CBG24365|A8WKJ4
AN14:CAEBR|Gene:CBG13328|A8XHL3
AN15:PRIPA|Gene:PPA00843|H3DTW2
AN30:HUMAN|HGNC:12572|P13051
AN31:PANTR|Ensembl:ENSPTRG00000005419|G2HEG8
AN32:GORGO|Ensembl:ENSGGOG00000022413|G3RY62
AN33:MACMU|Ensembl:ENSMMUG00000006957|F7DWJ9
AN35:MOUSE|MGI:MGI:109352|P97931
AN36:RAT|RGD:1307200|Q5BK44
AN40:BOVIN|Ensembl:ENSBTAG00000019848|Q17QB8
AN41:PIG|Ensembl:ENSSSCG00000009938|F1RGB4
AN42:HORSE|Ensembl:ENSECAG00000008080|F6TC79
AN45:FELCA|Ensembl:ENSFCAG00000032479|A0A2I2V3W5
AN46:FELCA|Ensembl:ENSFCAG00000008349|M3W9F3
AN47:CANLF|Ensembl:ENSCAFG00000011144|E2R050
AN48:MONDO|Ensembl:ENSMODG00000019350|H9H7C0
AN49:ORNAN|Ensembl:ENSOANG00000008804|F7G2Y9
AN50:CHICK|Ensembl:ENSGALG00000021285|F1NJG2
AN51:XENTR|Xenbase:XB-GENE-960119|F6V364
AN54:ORYLA|Ensembl:ENSORLG00000007126|H2LS78
AN56:DANRE|ZFIN:ZDB-GENE-070705-360|E9QG27
AN57:DANRE|ZFIN:ZDB-GENE-040426-900|A2BIQ2
AN58:LEPOC|Ensembl:ENSLOCG00000004484|W5MAF3
AN59:BRAFL|Gene:BRAFLDRAFT 287104|C3Y3U5
AN60:CIOIN|Ensembl:ENSCING00000002767|F7AH61
AN61:STRPU|EnsemblGenome:SPU 004357|W4XLS7
AN64:DAPPU|EnsemblGenome:DAPPUDRAFT 95942|E9FWF6
AN65:IXOSC|EnsemblGenome:ISCW005110|B7PG32
AN66:HELRO|EnsemblGenome:HelroG157191|T1EM77
AN67:NEMVE|EnsemblGenome:NEMVEDRAFT v1g219712|A7SYX6
AN68:TRIAD|EnsemblGenome:TriadG58651|B3S3A6
Animal

Figure S2. The animal branch in the phylogenetic tree of Udg protein family (PANTHER database ID: PTN000137400).

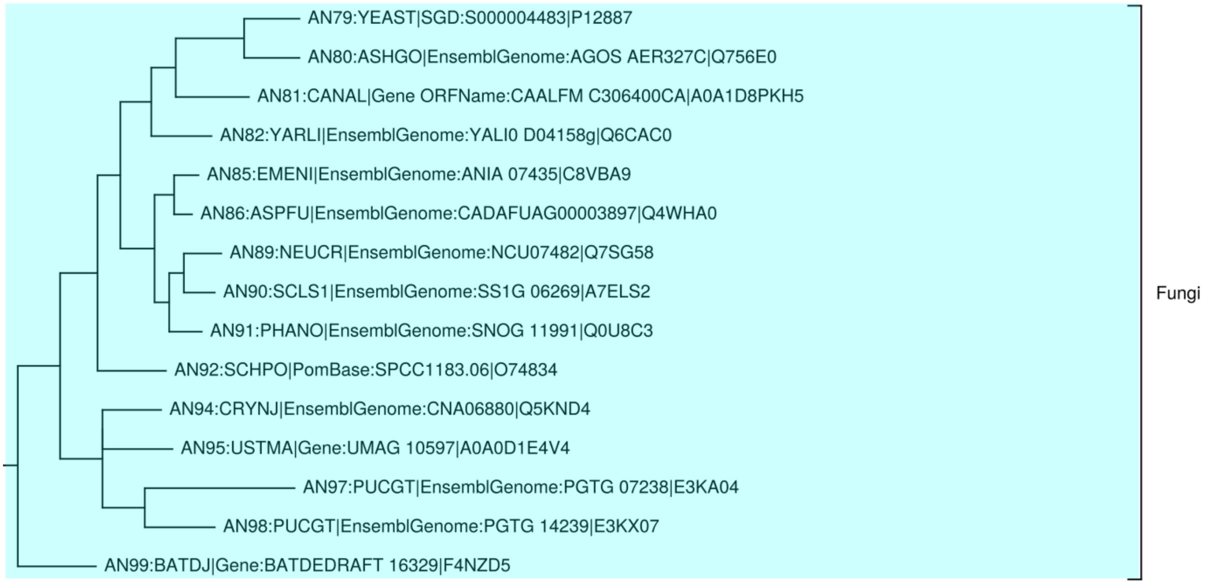
AN79:YEAST|SGD:S000004483|P12887
AN80:ASHGO|EnsemblGenome:AGOS AER327C|Q756E0
AN81:CANAL|Gene ORFName:CAALFM C306400CA|A0A1D8PKH5
AN82:YARLI|EnsemblGenome:YALI0 D04158g|Q6CAC0
AN85:EMENI|EnsemblGenome:ANIA 07435|C8VBA9
AN86:ASPFU|EnsemblGenome:CADAFUAG00003897|Q4WHA0
AN89:NEUCR|EnsemblGenome:NCU07482|Q7SG58
AN90:SCLS1|EnsemblGenome:SS1G 06269|A7ELS2
AN91:PHANO|EnsemblGenome:SNOG 11991|Q0U8C3
AN92:SCHPO|PomBase:SPCC1183.06|O74834
AN94:CRYNJ|EnsemblGenome:CNA06880|Q5KND4
AN95:USTMA|Gene:UMAG 10597|A0A0D1E4V4
AN97:PUCGT|EnsemblGenome:PGTG 07238|E3KA04
AN98:PUCGT|EnsemblGenome:PGTG 14239|E3KX07
AN99:BATDJ|Gene:BATDEDRAFT 16329|F4NZD5
Fungi

Figure S3. The fungi branch in the phylogenetic tree of Udg protein family (PANTHER database ID: PTN000137400).

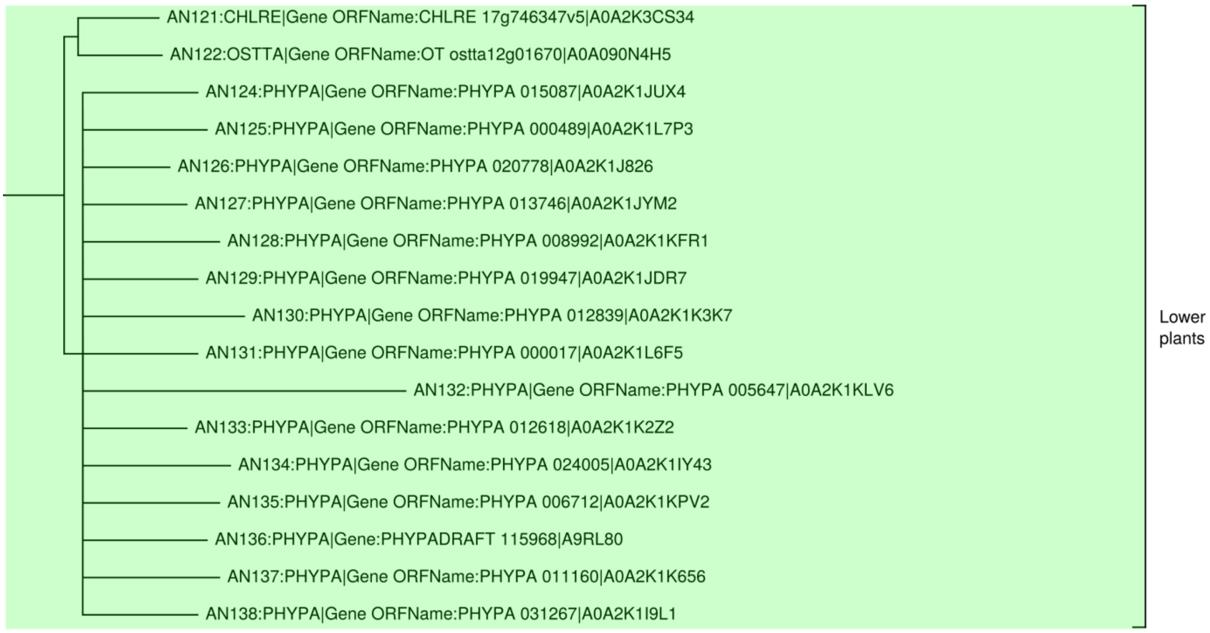
AN121:CHLRE|Gene ORFName:CHLRE 17g746347v5|A0A2K3CS34
AN122:OSTTA|Gene ORFName:OT ostta12g01670|A0A090N4H5
AN124:PHYPA|Gene ORFName:PHYPA 015087|A0A2K1JUX4
AN125:PHYPA|Gene ORFName:PHYPA 000489|A0A2K1L7P3
AN126:PHYPA|Gene ORFName:PHYPA 020778|A0A2K1J826
AN127:PHYPA|Gene ORFName:PHYPA 013746|A0A2K1JYM2
AN128:PHYPA|Gene ORFName:PHYPA 008992|A0A2K1KFR1
AN129:PHYPA|Gene ORFName:PHYPA 019947|A0A2K1JDR7
AN130:PHYPA|Gene ORFName:PHYPA 012839|A0A2K1K3K7
AN131:PHYPA|Gene ORFName:PHYPA 000017|A0A2K1L6F5
AN132:PHYPA|Gene ORFName:PHYPA 005647|A0A2K1KLV6
AN133:PHYPA|Gene ORFName:PHYPA 012618|A0A2K1K2Z2
AN134:PHYPA|Gene ORFName:PHYPA 024005|A0A2K1IY43
AN135:PHYPA|Gene ORFName:PHYPA 006712|A0A2K1KPV2
AN136:PHYPA|Gene:PHYPADRAFT 115968|A9RL80
AN137:PHYPA|Gene ORFName:PHYPA 011160|A0A2K1K656
AN138:PHYPA|Gene ORFName:PHYPA 031267|A0A2K1I9L1
Lower
plants

Figure S4. The lower plant branch, which is grouped with other eukaryotes (animals and fungi), in the phylogenetic tree of Udg protein family (PANTHER database ID: PTN000137400) shown in Figure S1.

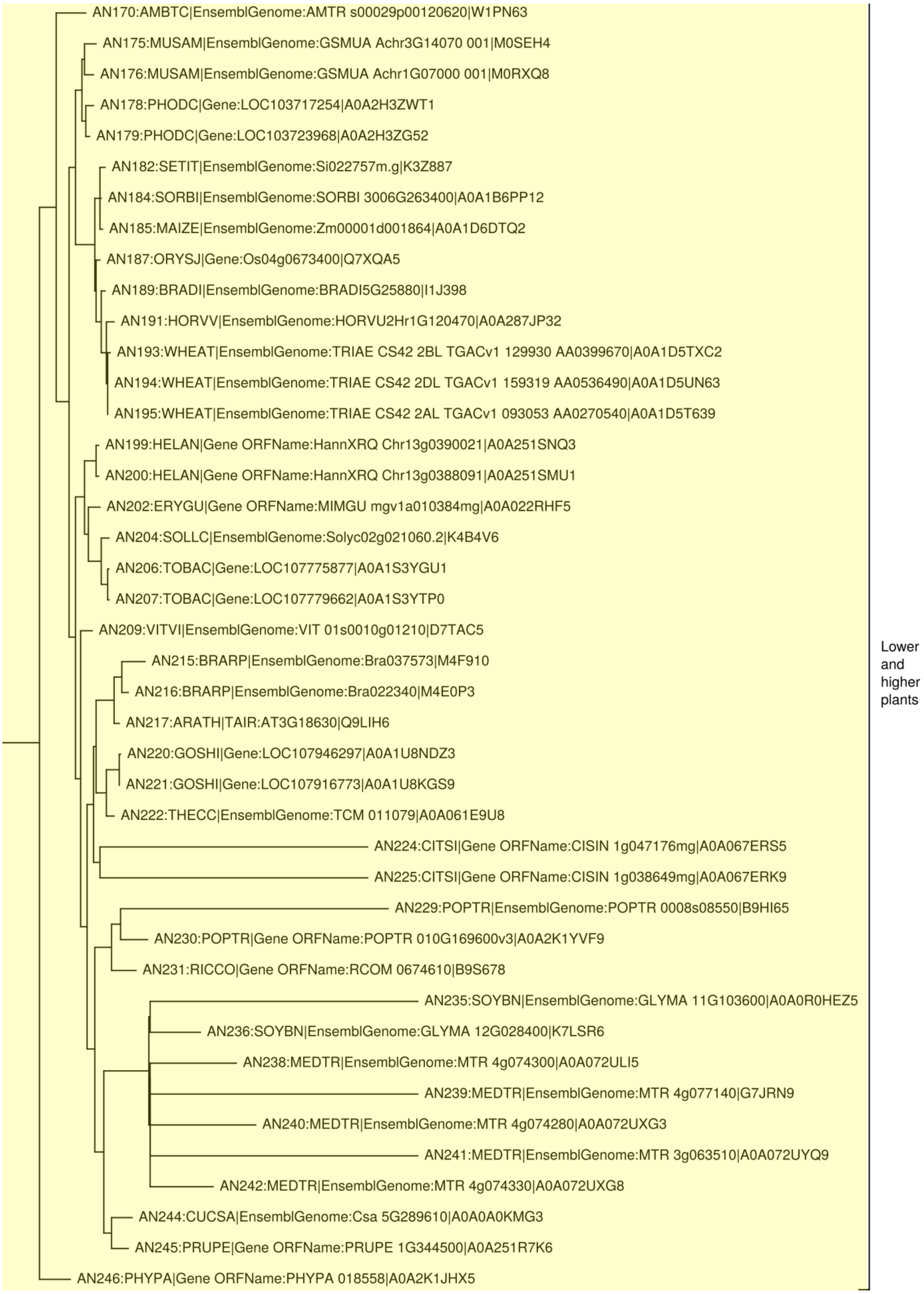
AN170:AMBTC|EnsemblGenome:AMTR s00029p00120620|W1PN63
AN175:MUSAM|EnsemblGenome:GSMUA Achr3G14070 001|M0SEH4
AN176:MUSAM|EnsemblGenome:GSMUA Achr1G07000 001|M0RXQ8
AN178:PHODC|Gene:LOC103717254|A0A2H3ZWT1
AN179:PHODC|Gene:LOC103723968|A0A2H3ZG52
AN182:SETIT|EnsemblGenome:Si022757m.g|K3Z887
AN184:SORBI|EnsemblGenome:SORBI 3006G263400|A0A1B6PP12
AN185:MAIZE|EnsemblGenome:Zm00001d001864|A0A1D6DTQ2
AN187:ORYSJ|Gene:Os04g0673400|Q7XQA5
AN189:BRADI|EnsemblGenome:BRADI5G25880|I1J398
AN191:HORVV|EnsemblGenome:HORVU2Hr1G120470|A0A287JP32
AN193:WHEAT|EnsemblGenome:TRIAE CS42 2BL TGACv1 129930 AA0399670|A0A1D5TXC2
AN194:WHEAT|EnsemblGenome:TRIAE CS42 2DL TGACv1 159319 AA0536490|A0A1D5UN63
AN195:WHEAT|EnsemblGenome:TRIAE CS42 2AL TGACv1 093053 AA0270540|A0A1D5T639
AN199:HELAN|Gene ORFName:HannXRQ Chr13g0390021|A0A251SNQ3
AN200:HELAN|Gene ORFName:HannXRQ Chr13g0388091|A0A251SMU1
AN202:ERYGU|Gene ORFName:MIMGU mgv1a010384mg|A0A022RHF5
AN204:SOLLC|EnsemblGenome:Solyc02g021060.2|K4B4V6
AN206:TOBAC|Gene:LOC107775877|A0A1S3YGU1
AN207:TOBAC|Gene:LOC107779662|A0A1S3YTP0
AN209:VITVI|EnsemblGenome:VIT 01s0010g01210|D7TAC5
AN215:BRARP|EnsemblGenome:Bra037573|M4F910
AN216:BRARP|EnsemblGenome:Bra022340|M4E0P3
AN217:ARATH|TAIR:AT3G18630|Q9LIH6
AN220:GOSHI|Gene:LOC107946297|A0A1U8NDZ3
AN221:GOSHI|Gene:LOC107916773|A0A1U8KGS9
AN222:THECC|EnsemblGenome:TCM 011079|A0A061E9U8
AN224:CITSI|Gene ORFName:CISIN 1g047176mg|A0A067ERS5
AN225:CITSI|Gene ORFName:CISIN 1g038649mg|A0A067ERK9
AN229:POPTR|EnsemblGenome:POPTR 0008s08550|B9HI65
AN230:POPTR|Gene ORFName:POPTR 010G169600v3|A0A2K1YVF9
AN231:RICCO|Gene ORFName:RCOM 0674610|B9S678
AN235:SOYBN|EnsemblGenome:GLYMA 11G103600|A0A0R0HEZ5
AN236:SOYBN|EnsemblGenome:GLYMA 12G028400|K7LSR6
AN238:MEDTR|EnsemblGenome:MTR 4g074300|A0A072ULI5
AN239:MEDTR|EnsemblGenome:MTR 4g077140|G7JRN9
AN240:MEDTR|EnsemblGenome:MTR 4g074280|A0A072UXG3
AN241:MEDTR|EnsemblGenome:MTR 3g063510|A0A072UYQ9
AN242:MEDTR|EnsemblGenome:MTR 4g074330|A0A072UXG8
AN244:CUCSA|EnsemblGenome:Csa 5G289610|A0A0A0KMG3
AN245:PRUPE|Gene ORFName:PRUPE 1G344500|A0A251R7K6
AN246:PHYPA|Gene ORFName:PHYPA 018558|A0A2K1JHX5
Lower
and
higher
plants

Figure S5. The second plant branch consisting of both lower and higher plants. This branch is grouped with bacteria, in the phylogenetic tree of Udg protein family (PANTHER database ID: PTN000137400) shown in Figure S1.

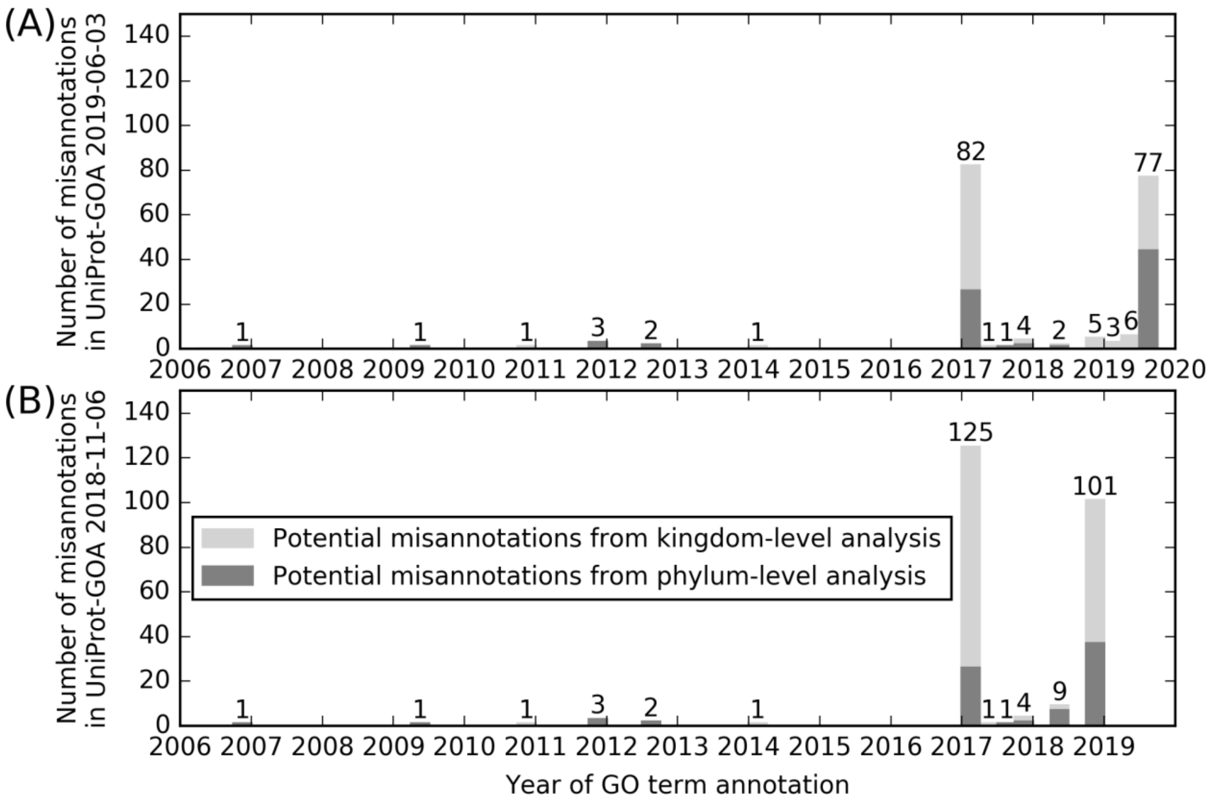

Figure S6. Stacked bar plots for the date of potential misannotations flagged by our taxon-specific rate ratio analysis in UniProt-GOA release 2019-06-03 (A) and 2018-11-06 (B). The *x*-axes correspond to the annotation date recorded by UniProt-GOA, and could be either the date of the initial GO term assignment or the date of the latest revision. The "date" record is particularly ambiguous for IBA GO terms, all of which have "date" record no earlier than 2017-02-28. For example, in UniProt-GOA release 2018-07-16 (ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/old/UNIPROT/goa_uniprot_all.gaf.180.gz), human PGM1 protein is annotated with an IBA GO term GO:0005829 "cytosol" dating back to 2010-04-15, using PANTHER family PTN000501326. However, in the subsequent release (UniProt-GOA release 2018-09-10, ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/old/UNIPROT/goa_uniprot_all.gaf.181.gz), the date of this GO term annotation for PGM1 is changed to 2017-02-28 even though the source of the annotation (PTN000501326) is essentially unchanged. Therefore, the lifetime for a GO term misannotation to persist in the database calculated from the "date" record can be taken only as a lower bound.

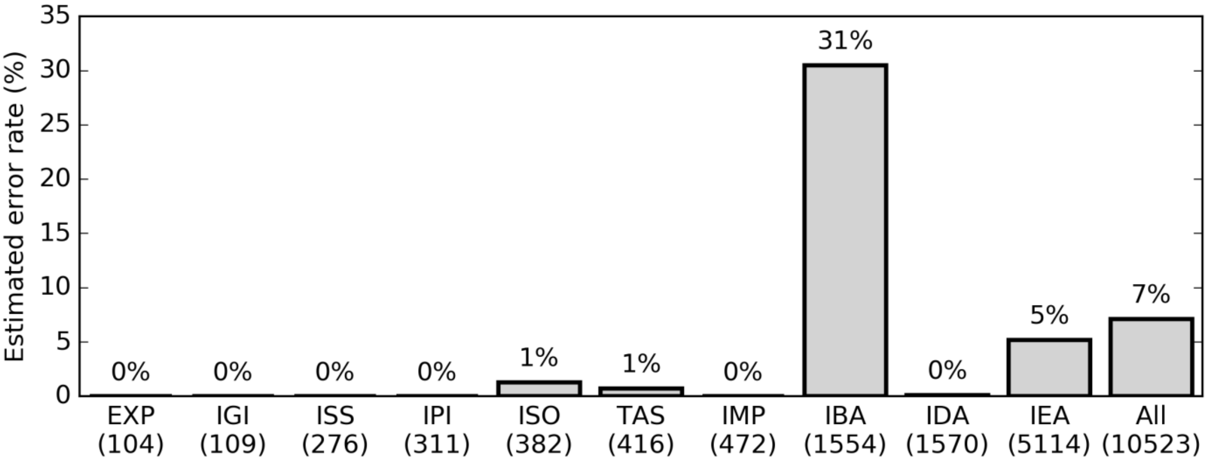

Figure S7. Estimated error rates of GO terms with different evidence codes. These error rates are estimated by checking UniProt-GOA 2018-11-06 annotations that are either rejected or confirmed by new low-throughput experimental evidence (evidence codes: EXP, IDA, IPI, IMP, IGI, and IEP) in UniProt-GOA 2019-06-03. Here, "experimental evidence" excludes author statements (evidence codes TAS and NAS) because these GO terms may be imported from secondary databases instead of from primary literature, especially TAS annotations. Due to the hierarchical structure of Gene Ontology, a GO annotation is considered confirmed if either the exact GO term or its child term is confirmed by new experimental annotations. Similarly, a GO annotation is considered rejected if either the same term or at least one of its parents is rejected. We only consider the 29120 GO terms common to both UniProt-GOA versions 2018-11-06 and 2019-06-03, excluding annotations for obsolete UniProt proteins, for GO:0005515 "protein binding" and for the three root terms, as explained in main text section 3.1. Redundant annotations (multiple entries for the same GO term annotated to the same protein with the same evidence code) are also excluded. In the end, 392550471 (98.45%) of the 398728612 GO annotations in the old release are neither confirmed nor rejected in the new release; 6167618 (1.54%) are simply removed without new conflicting experimental evidence. These annotations could be removed due to, for example, change of GO consortium policy requiring more stringent cutoff of computational prediction, and are therefore not necessarily real errors. For the remaining GO annotations in the old release, 9773 annotations are confirmed by new low-throughput experiments, while 750 are rejected by a "NOT" qualifier. These 10543 GO annotations are used to make this plot, with different evidences shown in ascending order of the number of rejected/confirmed annotations used to estimate error rates (number in parenthesis for *x*-axis labels). Only evidence codes with sufficient statistics (>100 confirmed/rejected annotations) are shown in the plot. “All” means all confirmed/rejected annotations for all evidence codes.

Reference

Benjamini, Y. and Hochberg, Y. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *J R Stat Soc B* 1995;57(1):289-300.

Deegan, J.I., Dimmer, E.C. and Mungall, C.J. Formalization of taxon-based constraints to detect inconsistencies in annotation and ontology development. *Bmc Bioinformatics* 2010;11.

Fay, M.P. Two-sided Exact Tests and Matching Confidence Intervals for Discrete Data. *R J* 2010;2(1):53-58.

Sagan, L. On the origin of mitosing cells. *Journal of theoretical biology* 1967;14(3):225-IN226.