1 Workflow of sparse representation

Fig. S1 illustratively shows the usage of sparse representation to compute the semantic similarity between genes based on GO annotations of them. In Fig. S1, A is a gene-term association matrix which contains 8 genes and 10 GO terms. By minimizing the first Equation in Fig. S1, we can obtain \( \{ \gamma_i \}_{i=1}^8 \), which make \( \gamma_i A_i \) close to \( A(i, \cdot) \) and \( ||\gamma_i||_1 \) is minimized. For example, \( \gamma_1 \) means that \( A(1, \cdot) \) is reconstructed by \( 0.3074 \times A(3, \cdot) \), \( 0.3327 \times A(5, \cdot) \) and \( 0.2069 \times A(7, \cdot) \), so the similarity between \( A(1, \cdot) \) and \( \{ A(i, \cdot) \}_{i=2}^8 \) are 0, 0.3074, 0, 0.3327, 0, 0.2069 and 0, respectively. Since we will not use the similarity between a gene and itself, we set the diagonal entry of \( S \) as 0. Finally, we set \( S = (S^T + S) / 2 \) to ensure \( S \) be a symmetric matrix and store the semantic similarity between genes in \( S \).

Figure S1: A workflow of using sparse representation to compute semantic similarity between genes

2 Performance of noisy annotations prediction

In the main text, we reported the results of eight comparing methods, LF, SR, EC, NtN, NoisyGOA, NtN+EC, NoisyGOA+EC and NoGOA in predicting the noisy annotations of genes on achieved GOA files of *H. Sapiens* (archived date: May 2016). Here, the results on *A. thaliana* and *S. cerevisiae* and *H. Sapiens*, *G. gallus*, *B. Taurus* and *M. musculus* (archived date: May 2015 or May 2016) are included in Tables S1-S11. In these tables, the statistical best performer is highlighted in boldface and significance
Table S1: Performance of predicting noisy annotations in *A. thaliana* on archived GOA file (archived date: May, 2016). The numbers in **boldface** denote the best performance.

| Dataset | GOA + GOA-EC | GOA + GOA-EC | GOA + GOA-EC | GOA + GOA-EC | GOA + GOA-EC | GOA + GOA-EC | GOA + GOA-EC |
|---------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|
| BP      | 36.92 ± 1.07  | 39.63 ± 1.00  | 39.13 ± 1.02  | 39.63 ± 1.00  | 39.13 ± 1.02  | 39.63 ± 1.00  | 39.13 ± 1.02  |
| Recall  | 39.52 ± 1.07  | 40.34 ± 1.00  | 39.12 ± 1.02  | 39.52 ± 1.07  | 39.12 ± 1.02  | 39.52 ± 1.07  | 39.12 ± 1.02  |
| Coverage | 37.77 ± 1.07  | 36.53 ± 1.00  | 35.86 ± 1.02  | 37.77 ± 1.07  | 35.86 ± 1.02  | 37.77 ± 1.07  | 35.86 ± 1.02  |

Table S2: Performance of predicting noisy annotations in *S. cerevisiae* on archived GOA file (archived date: May, 2016). The numbers in **boldface** denote the best performance.

| Dataset | GOA + GOA-EC | GOA + GOA-EC | GOA + GOA-EC | GOA + GOA-EC | GOA + GOA-EC | GOA + GOA-EC | GOA + GOA-EC |
|---------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|
| BP      | 23.12 ± 1.07  | 23.53 ± 1.00  | 23.53 ± 1.02  | 23.53 ± 1.00  | 23.53 ± 1.02  | 23.53 ± 1.00  | 23.53 ± 1.02  |
| Recall  | 33.05 ± 1.07  | 33.05 ± 1.00  | 33.05 ± 1.02  | 33.05 ± 1.07  | 33.05 ± 1.02  | 33.05 ± 1.07  | 33.05 ± 1.02  |
| Coverage | 29.77 ± 1.07  | 29.77 ± 1.00  | 29.77 ± 1.02  | 29.77 ± 1.07  | 29.77 ± 1.02  | 29.77 ± 1.07  | 29.77 ± 1.02  |

Table S3: Performance of predicting noisy annotations in *B. taurus* on archived GOA file (archived date: May, 2016). The numbers in **boldface** denote the best performance.

| Dataset | GOA + GOA-EC | GOA + GOA-EC | GOA + GOA-EC | GOA + GOA-EC | GOA + GOA-EC | GOA + GOA-EC | GOA + GOA-EC |
|---------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|
| BP      | 39.81 ± 1.07  | 40.34 ± 1.00  | 39.52 ± 1.07  | 39.52 ± 1.07  | 39.52 ± 1.07  | 39.52 ± 1.07  | 39.52 ± 1.07  |
| Recall  | 39.52 ± 1.07  | 40.34 ± 1.00  | 39.52 ± 1.07  | 39.52 ± 1.07  | 39.52 ± 1.07  | 39.52 ± 1.07  | 39.52 ± 1.07  |
| Coverage | 37.77 ± 1.07  | 36.53 ± 1.00  | 35.86 ± 1.02  | 37.77 ± 1.07  | 35.86 ± 1.02  | 37.77 ± 1.07  | 35.86 ± 1.02  |

Table S4: Performance of predicting noisy annotations in *G. gallus* on archived GOA file (archived date: May, 2016). The numbers in **boldface** denote the best performance.

| Dataset | GOA + GOA-EC | GOA + GOA-EC | GOA + GOA-EC | GOA + GOA-EC | GOA + GOA-EC | GOA + GOA-EC | GOA + GOA-EC |
|---------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|
| BP      | 43.74 ± 1.07  | 43.74 ± 1.00  | 43.74 ± 1.02  | 43.74 ± 1.07  | 43.74 ± 1.02  | 43.74 ± 1.07  | 43.74 ± 1.02  |
| Recall  | 43.74 ± 1.07  | 43.74 ± 1.00  | 43.74 ± 1.02  | 43.74 ± 1.07  | 43.74 ± 1.02  | 43.74 ± 1.07  | 43.74 ± 1.02  |
| Coverage | 43.74 ± 1.07  | 43.74 ± 1.00  | 43.74 ± 1.02  | 43.74 ± 1.07  | 43.74 ± 1.02  | 43.74 ± 1.07  | 43.74 ± 1.02  |

Table S5: Performance of predicting noisy annotations in *M. musculus* on archived GOA file (archived date: May, 2016). The numbers in **boldface** denote the best performance.

| Dataset | GOA + GOA-EC | GOA + GOA-EC | GOA + GOA-EC | GOA + GOA-EC | GOA + GOA-EC | GOA + GOA-EC | GOA + GOA-EC |
|---------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|
| BP      | 23.07 ± 0.49  | 23.07 ± 0.47  | 22.80 ± 0.46  | 22.80 ± 0.46  | 22.80 ± 0.46  | 22.80 ± 0.46  | 22.80 ± 0.46  |
| Recall  | 30.65 ± 0.49  | 30.65 ± 0.47  | 30.65 ± 0.46  | 30.65 ± 0.46  | 30.65 ± 0.46  | 30.65 ± 0.46  | 30.65 ± 0.46  |
| Coverage | 27.63 ± 0.49  | 27.63 ± 0.47  | 27.63 ± 0.46  | 27.63 ± 0.46  | 27.63 ± 0.46  | 27.63 ± 0.46  | 27.63 ± 0.46  |

Table S6: Performance of predicting noisy annotations in *A. thaliana* on archived GOA file (archived date: May, 2015). The numbers in **boldface** denote the best performance.

| Dataset | GOA + GOA-EC | GOA + GOA-EC | GOA + GOA-EC | GOA + GOA-EC | GOA + GOA-EC | GOA + GOA-EC | GOA + GOA-EC |
|---------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|
| BP      | 47.90 ± 1.17  | 47.90 ± 1.17  | 47.90 ± 1.17  | 47.90 ± 1.17  | 47.90 ± 1.17  | 47.90 ± 1.17  | 47.90 ± 1.17  |
| Recall  | 52.84 ± 1.17  | 52.84 ± 1.17  | 52.84 ± 1.17  | 52.84 ± 1.17  | 52.84 ± 1.17  | 52.84 ± 1.17  | 52.84 ± 1.17  |
| Coverage | 57.87 ± 1.17  | 57.87 ± 1.17  | 57.87 ± 1.17  | 57.87 ± 1.17  | 57.87 ± 1.17  | 57.87 ± 1.17  | 57.87 ± 1.17  |
Table S7: Performance of predicting noisy annotations in S. sapiens on archived GOA file (date: May, 2015). The numbers in boldface denote the best performance.

| Method | F1-Score | Precision | Recall |
|--------|----------|-----------|--------|
| BP     |          |           |        |
| F1-Score | 56.55±0.72 | 30.90±0.45 | 49.27±0.79 |
| CC     | 18.41±0.54 | 34.97±0.84 | 20.01±0.57 |
| MF     | 24.22±0.86 | 35.10±0.80 | 24.02±0.58 |

Table S8: Performance of predicting noisy annotations in B. cerevisiae on archived GOA file (date: May, 2015). The numbers in boldface denote the best performance.

| Method | F1-Score | Precision | Recall |
|--------|----------|-----------|--------|
| BP     |          |           |        |
| F1-Score | 49.51±0.75 | 30.92±0.45 | 49.27±0.79 |
| CC     | 18.41±0.54 | 34.97±0.84 | 20.01±0.57 |
| MF     | 24.22±0.86 | 35.10±0.80 | 24.02±0.58 |

Table S9: Performance of predicting noisy annotations in B. taurus on archived GOA file (date: May, 2015). The numbers in boldface denote the best performance.

| Method | F1-Score | Precision | Recall |
|--------|----------|-----------|--------|
| BP     |          |           |        |
| F1-Score | 49.51±0.75 | 30.92±0.45 | 49.27±0.79 |
| CC     | 18.41±0.54 | 34.97±0.84 | 20.01±0.57 |
| MF     | 24.22±0.86 | 35.10±0.80 | 24.02±0.58 |

Table S10: Performance of predicting noisy annotations in G. gallus on archived GOA file (date: May, 2015). The numbers in boldface denote the best performance.

| Method | F1-Score | Precision | Recall |
|--------|----------|-----------|--------|
| BP     |          |           |        |
| F1-Score | 49.51±0.75 | 30.92±0.45 | 49.27±0.79 |
| CC     | 18.41±0.54 | 34.97±0.84 | 20.01±0.57 |
| MF     | 24.22±0.86 | 35.10±0.80 | 24.02±0.58 |

3 Parameter sensitivity analysis

In order to study the effectiveness of NoGOA under different input values of $\theta$, we vary $\theta$ from 0.1 to 0.9 while fixing $\alpha$ as 0.2. The results are shown in Fig. [S2]. From the results in Fig. [S2], we can find that F1-Score remains stable when $\theta \in [0.1,0.7]$ and then decreases when $\theta \in [0.7,0.9]$ in most cases. These results suggest that the weight for evidence codes whose estimated ratios of noisy annotations above the threshold $\tau$, and for those whose estimated ratios below $\tau$ should be differently specified. The sensitivity study of $\theta$ again justifies the rationality of setting different weights to different annotations based on the estimated ratio of noisy annotations per evidence code.

We also study the effectiveness of NoGOA under different specifications of $\tau$, we report the results of NoGOA in predicting noisy annotations when $\tau$ is the average (or median) of $r_{mc}$. In fact, we also test $\tau = 0$ for NoGOA, since annotations with respect to more than half evidence codes in the historical GOA files do not disappear in the recent GOA files, the results of NoGOA with $\tau = 0$ are the same as NoGOA with $\tau$ equal to the median of $r_{mc}$. The results are shown in Tables [S12][S17]. From these tables, we can find that $\tau$ influences the prediction of measuring predictive annotations. In most cases, when $\tau = 0$, NoGOA gets lower performance. That is because the evidence codes which do not have noisy annotations always have few annotations. NoGOA assigns larger weights to annotations tagged with these evidence codes, and smaller weights to others. As a result, using 0 to differentiate noisy
Figure S2: Performance of NoGOA in predicting noisy annotations under different input values of $\theta$.

Table S12: Performance of NoGOA in predicting noisy annotation under different values of $\tau$ in *A. thaliana* on archived GOA files (May, 2016). The numbers in **boldface** denote the better performance.

|       | Median($\theta$) | Average |
|-------|------------------|---------|
|       | Precision        | Recall  | F1-Score |
| **BP**| 54.04±0.76       | 61.68±0.81 |
| **Recall** | 51.94±0.77 | 61.40±0.80 |
| **F1-Score** | 51.46±0.76 | 61.02±0.80 |
| **CC** | Precision        | Recall  | F1-Score |
|       | 55.54±1.28       | 59.44±1.34 |
| **Recall** | 57.29±1.32 | 58.67±1.32 |
| **F1-Score** | 56.47±1.30 | 58.00±1.27 |
| **MF** | Precision        | Recall  | F1-Score |
|       | 61.29±1.33       | 59.02±1.31 |
| **Recall** | 62.15±1.34 | 58.85±1.29 |

Table S13: Performance of predicting noisy annotations on different $\tau$ in *H. sapiens* on archived GOA files (May, 2016). The numbers in **boldface** denote the best performance.

|       | Median($\theta$) | Average |
|-------|------------------|---------|
|       | Precision        | Recall  | F1-Score |
| **BP**| 35.65±0.67       | 41.14±0.76 |
| **Recall** | 35.84±0.67 | 41.28±0.76 |
| **F1-Score** | 35.84±0.67 | 41.28±0.76 |
| **CC** | Precision        | Recall  | F1-Score |
|       | 41.55±0.95       | 41.34±0.97 |
| **Recall** | 42.06±0.95 | 41.85±0.98 |
| **F1-Score** | 41.77±0.94 | 41.56±0.97 |
| **MF** | Precision        | Recall  | F1-Score |
|       | 59.02±0.60       | 59.47±0.60 |
| **Recall** | 47.41±0.54 | 59.14±0.60 |
| **F1-Score** | 47.18±0.54 | 59.14±0.60 |

Table S14: Performance of predicting noisy annotations on different $\tau$ in *S. cerevisiae* on archived GOA files (May, 2016). The numbers in **boldface** denote the best performance.

|       | Median($\theta$) | Average |
|-------|------------------|---------|
|       | Precision        | Recall  | F1-Score |
| **BP**| 40.39±0.61       | 65.15±0.76 |
| **Recall** | 40.78±0.61 | 65.50±0.76 |
| **F1-Score** | 41.92±1.04 | 59.30±1.20 |
| **CC** | Precision        | Recall  | F1-Score |
|       | 46.36±1.07       | 59.23±1.23 |
| **Recall** | 45.59±1.05 | 58.48±1.21 |
| **F1-Score** | 45.38±0.56 | 74.80±0.69 |
| **MF** | Precision        | Recall  | F1-Score |
|       | 44.95±0.55       | 74.60±0.69 |
| **Recall** | 45.08±0.55 | 74.69±0.69 |
| **F1-Score** | 45.08±0.55 | 74.69±0.69 |
Table S15: Performance of predicting noisy annotations on different $\tau$ in *A. thaliana* on archived GOA files (May, 2015). The numbers in **boldface** denote the best performance.

| Term | Median(0) | Average |
|------|-----------|---------|
| BP   | Precision | 39.56±1.14 | 39.39±1.09 |
|      | Recall    | 40.08±1.16 | 39.91±1.11 |
|      | F1-Score  | 39.79±1.15 | 39.62±1.10 |
| CC   | Precision | 73.39±3.54 | 71.51±3.58 |
|      | Recall    | 74.05±3.57 | 72.20±3.60 |
|      | F1-Score  | 73.66±3.55 | 71.80±3.59 |
| MF   | Precision | 57.75±0.98 | 52.53±0.96 |
|      | Recall    | 57.96±0.98 | 52.67±0.96 |
|      | F1-Score  | 57.84±0.98 | 52.59±0.96 |

Table S16: Performance of predicting noisy annotations on different $\tau$ in *H. sapiens* on archived GOA files (May, 2015). The numbers in **boldface** denote the best performance.

| Term | Median(0) | Average |
|------|-----------|---------|
| BP   | Precision | 33.62±0.54 | 42.74±0.64 |
|      | Recall    | 34.21±0.55 | 43.45±0.65 |
|      | F1-Score  | 33.87±0.55 | 43.04±0.64 |
| CC   | Precision | 46.21±1.06 | 61.77±1.34 |
|      | Recall    | 46.93±1.07 | 62.23±1.35 |
|      | F1-Score  | 46.52±1.06 | 61.98±1.35 |
| MF   | Precision | 45.27±0.89 | 64.40±1.22 |
|      | Recall    | 45.63±0.89 | 64.77±1.22 |
|      | F1-Score  | 45.42±0.89 | 64.59±1.22 |

Table S17: Performance of predicting noisy annotations on different $\tau$ in *S. cerevisiae* on archived GOA files (May, 2016). The numbers in **boldface** denote the best performance.

| Term | Median(0) | Average |
|------|-----------|---------|
| BP   | Precision | 34.04±1.67 | 34.13±1.68 |
|      | Recall    | 34.22±1.68 | 34.34±1.69 |
|      | F1-Score  | 34.11±1.67 | 34.23±1.69 |
| CC   | Precision | 52.61±5.47 | 59.77±6.28 |
|      | Recall    | 52.99±5.51 | 59.98±6.30 |
|      | F1-Score  | 52.79±5.49 | 59.87±6.29 |
| MF   | Precision | 35.19±2.07 | 69.22±3.24 |
|      | Recall    | 35.71±2.09 | 69.82±3.25 |
|      | F1-Score  | 35.38±2.08 | 69.44±3.24 |

4 Evaluation metrics for gene function prediction

Here, we introduce six evaluation metrics, *MacroAvgF1*, *MicroAvgF1*, *AvgAUC*, *AvgPrec* *Fmax* and *Smin* to measure the performance of gene function prediction. In fact, the first four metrics are widely used to measure the performance of multi-label classification ([Zhang et al., 2014](#)), which targets at predicting a set of related labels to an instance instead of a single label. From this perspective, gene function prediction can also be viewed as a multi-label learning problem. Therefore, these four metrics are used to measure the performance of gene function prediction ([Fu et al., 2016](#) [Yu et al., 2015](#)). The latter two metrics (*Smin* and *Fmax*), along with *AvgAUC*, are suggested and used in the Second Critical Assessment of protein Function Annotation algorithms (CADATA) ([Jiang et al., 2016](#)). The definitions of these metrics are detailed in ([Zhang et al., 2014](#) [Jiang et al., 2016](#)). To be self-inclusive, we list the definitions of these metrics here.

*MacroAvgF1* is the average $F1$ scores of different terms:

$$MacroAvgF1 = \frac{1}{|T|} \sum_{t\in T} \frac{2p_tr_t}{p_t + r_t}$$

where $|T|$ is the number of terms, $p_t$ and $r_t$ are the precision and recall for GO term $t$, defined as:

$$p_t = \frac{TP_t}{TP_t + FP_t},\quad r_t = \frac{TP_t}{TP_t + FN_t}$$

$TP_t$, $FP_t$, and $FN_t$ are the true positive, false positive, and false negative with respect to term $t$. From the definition, it can be observed MacroAvgF1 first calculates the $F1$ measures for each term, and then averages over all the terms. MacroAvgF1 is more affected by the performance on the specific terms, which annotated to few genes.

*MicroAvgF1* calculates the $F1$ measure on the predictions of different terms as a whole:

$$MicroAvgF1 = \frac{\sum_{t=1}^{\mid T \mid} 2p_tr_t}{\sum_{t=1}^{\mid T \mid} p_t + r_t}$$

MicroAvgF1 is more affected by the performance on general terms than on specific terms.
Average Precision (AvgPrec) evaluates the average fraction of terms ranked above a particular term $t \in T_i$ ($T_i$ stores all the GO terms annotated to the $i$-th gene).

$$\text{AvgPrec} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{|T_i|} \sum_{t_1 \in T_i} \{ t_2 \in T_i | \text{rank}(V(i,t_2)) \leq \text{rank}(V(i,t_1)) \}$$

where \( \text{rank}(V(i, \cdot)) \) is a rank function, which ranks the largest \( V(i, \cdot) \in \mathbb{R}^{|T_i|} \) as 1 and the smallest \( V(i, \cdot) \) as \( |T_i| \). The larger the value of AvgPrec, the better the performance is.

Average AUC (AvgAUC) is a term-centric evaluation metric, it averages the area under the receiver operating curve (AUC) of each term across all terms in \( T \). The receiver operating curve plots the true positive rate (sensitivity) as a function of the false positive rate (1-specificity) under different classification thresholds. The curve measures the overall quality of the ranking induced by the classifier, instead of the quality of a single value of the threshold in that ranking.

Fmax is a gene-centric evaluation metric used in the community-based critical assessment of protein function annotation (CAFA) [Radivojac et al. 2013]. Fmax is an F-measure computed as:

$$F_{\text{max}} = \max_{\theta} \frac{2p(\theta)r(\theta)}{p(\theta) + r(\theta)}$$

where \( p(\theta) = \frac{1}{m(\theta)} \sum_{i=1}^{m(\theta)} p_i(\theta) \) is the the precision at threshold \( \theta \in [0, 1] \), \( p_i(\theta) \) is the precision on the \( i \)-th gene, \( m(\theta) \) is the number of genes on which at least one prediction was made above the threshold \( \theta \), \( r(\theta) = \frac{1}{N} \sum_{i=1}^{N} r_i(\theta) \) is the recall across \( N \) genes at threshold \( \theta \).

Smin is an evaluation metric used in the second critical assessment of functional annotation (CAFA2) [Jiang et al. 2016] and defined as:

$$S_{\text{min}} = \min_{\theta} \sqrt{ru(\theta)^2 + mi(\theta)^2}$$

where the remaining uncertainty \( ru(\theta) = \frac{1}{N} \sum_{i=1}^{N} \sum_{t} ic(t) \cdot \mathbf{1}(t \in P_i(\theta) \land t \in T_i) \), misinformation \( mi(\theta) = \frac{1}{N} \sum_{i=1}^{N} \sum_{t} ic(t) \cdot \mathbf{1}(t \notin P_i(\theta) \land t \in T_i) \), \( P_i(\theta) \) denotes the set of terms whose predicted scores greater than or equal to \( \theta \) for the \( i \)-th gene, \( T_i \) denotes the ground-truth set of terms for that gene, \( \mathbf{1}(\cdot) \) is an indicator function, and \( ic(t) \) is the information content of the ontology term \( t \). It is estimated in a maximum likelihood manner as the negative binary logarithm of the conditional probability that the term \( t \) is present in a gene’s annotations given that all its parent terms are also present.

5 Removing noisy annotations improves gene function prediction

In the main text, we provide the experimental results of gene function prediction by a majority vote model on PPI networks with/without removing noisy annotations in archived GOA file of \( H. \ sapiens \). Here, we report the corresponding results on archived GOA files of \( A. \ thaliana \) and \( S. \ cerevisiae \) in Tables S18 and S19. From these tables, we can clearly see that removing the noisy annotations provides improved performance for gene function prediction.

Table S18: Results of gene function prediction on \( A. \ thaliana \) (archived date: May, 2016). The data in boldface denote the better result. ‘Original’ directly uses annotations in the historical GOA file to predict gene function; ‘NoGOA’ removes predicted noisy annotations from the historical GOA file and then predicts gene function. ↓ means the lower the value, the better the performance is.

|         | BP  | CC  | MP  |
|---------|-----|-----|-----|
|         | Original | NoGOA | Original | NoGOA | Original | NoGOA |
| MicroAvgF1 | 82.10 | 82.15 | 89.28 | 89.28 | 85.65 | 85.66 |
| MacroAvgF1 | 74.43 | 74.11 | 81.59 | 81.60 | 74.24 | 74.01 |
| AvgPrec  | 76.70 | 76.73 | 88.22 | 88.22 | 82.77 | 82.79 |
| AvgROC   | 91.40 | 91.83 | 93.22 | 93.22 | 91.72 | 92.11 |
| Fmax     | 85.79 | 85.80 | 92.27 | 92.27 | 89.80 | 89.84 |
| Smin ↓   | 8.08 | 7.98 | 1.66 | 1.66 | 2.78 | 2.75 |

6 Examples of predicted noisy annotations

Tables S20-S25 count the number of direct noisy annotations, predicted noisy annotations, and correctly predicted direct noisy annotations on \( H. \ sapiens \), \( A. \ thaliana \) and \( S. \ cerevisiae \) archived in different years. Table S26 lists some examples of correctly (or wrongly) predicted direct noisy annotations.
Table S19: Results of gene function prediction on *S. cerevisiae* (archived date: May, 2016). The data in **boldface** denote the better result. ‘Original’ directly uses annotations in the historical GOA file to predict gene function; ‘NoGOA’ removes predicted noisy annotations from the historical GOA file and then predicts gene function.  ↓ means the lower the value, the better the performance is.

|       | BP | CC | MF |
|-------|----|----|----|
|       | Original | NoGOA | Original | NoGOA | Original | NoGOA |
| MacroAvgF1 | 94.00 | 94.68 | 93.74 | 94.77 | 93.36 | 95.39 |
| MacroAvgF1 | 90.24 | 90.19 | 92.28 | 92.19 | 93.36 | 93.35 |
| AvgPrec | 92.26 | 92.31 | 94.99 | 95.02 | 94.87 | 94.89 |
| AvgGCOC | 95.60 | 95.24 | 95.71 | 95.42 | 97.53 | 97.54 |
| Fmax | 95.19 | 95.20 | 94.91 | 94.92 | 96.21 | 96.21 |
| Smin ↓ | 1.52 | 1.47 | 0.99 | 0.99 | 0.60 | 0.58 |

Table S20: Statistics of predicted noisy annotations for different evidence codes on *A. thaliana* (archived date: May, 2016). \(N_c\) is the number of noisy annotations, \(\hat{N}_c\) is the number of predicted and \(\bar{N}_c\) is the number of correctly predicted noisy annotations by NoGOA, respectively.

|       | BP | CC | MF |
|-------|----|----|----|
|       | \(N_c\) | \(\hat{N}_c\) | \(\bar{N}_c\) | \(N_c\) | \(\hat{N}_c\) | \(\bar{N}_c\) | \(N_c\) | \(\hat{N}_c\) | \(\bar{N}_c\) |
| MicroAvgF1 | 125 | 65 | 25 | 152 | 13 | 7 | 17 | 111 | 7 | 3 |
| MacroAvgF1 | 21 | 66 | 0 | 275 | 24 | 103 | 0 | 219 | 0 | 0 |
| AvgPrec | 8 | 38 | 0 | 178 | 16 | 54 | 0 | 53 | 0 | 0 |
| AvgGCOC | 3 | 281 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| Fmax | 40 | 77 | 57 | 13 | 2 | 0 | 0 | 0 | 0 | 0 |
| Smin ↓ | 3 | 72 | 53 | 8 | 1 | 0 | 0 | 0 | 0 | 0 |

Table S21: Statistics of predicted noisy annotations for different evidence codes on *H. sapiens* (archived date: May, 2016). \(N_c\) is the number of noisy annotations, \(\hat{N}_c\) is the number of predicted and \(\bar{N}_c\) is the number of correctly predicted noisy annotations by NoGOA, respectively.

|       | BP | CC | MF |
|-------|----|----|----|
|       | \(N_c\) | \(\hat{N}_c\) | \(\bar{N}_c\) | \(N_c\) | \(\hat{N}_c\) | \(\bar{N}_c\) | \(N_c\) | \(\hat{N}_c\) | \(\bar{N}_c\) |
| MicroAvgF1 | 742 | 37 | 0 | 105 | 50 | 2 | 0 | 165 | 0 | 0 |
| MacroAvgF1 | 7 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| AvgPrec | 105 | 263 | 2 | 8 | 2 | 0 | 0 | 0 | 0 | 0 |
| AvgGCOC | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Fmax | 3 | 33 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Smin ↓ | 3 | 72 | 53 | 8 | 1 | 0 | 0 | 0 | 0 | 0 |

Table S22: Statistics of predicted noisy annotations for different evidence codes on *S. cerevisiae* (archived date: May, 2016). \(N_c\) is the number of noisy annotations, \(\hat{N}_c\) is the number of predicted and \(\bar{N}_c\) is the number of correctly predicted noisy annotations by NoGOA, respectively.

|       | BP | CC | MF |
|-------|----|----|----|
|       | \(N_c\) | \(\hat{N}_c\) | \(\bar{N}_c\) | \(N_c\) | \(\hat{N}_c\) | \(\bar{N}_c\) | \(N_c\) | \(\hat{N}_c\) | \(\bar{N}_c\) |
| MicroAvgF1 | 147 | 95 | 19 | 331 | 81 | 5 | 1 | 0 | 1 | 0 |
| MacroAvgF1 | 112 | 86 | 2 | 176 | 49 | 5 | 0 | 0 | 0 | 0 |
| AvgPrec | 102 | 269 | 38 | 19 | 0 | 0 | 0 | 0 | 0 | 0 |
| AvgGCOC | 93 | 95 | 7 | 6 | 0 | 0 | 0 | 0 | 0 | 0 |
| Fmax | 6 | 6 | 102 | 153 | 27 | 9 | 0 | 0 | 0 | 0 |
| Smin ↓ | 6 | 4 | 130 | 146 | 47 | 15 | 0 | 0 | 0 | 0 |

Table S23: Statistics of predicted noisy annotations for different evidence codes on *A. thaliana* (archived date: May, 2015). \(N_c\) is the number of noisy annotations, \(\hat{N}_c\) is the number of predicted and \(\bar{N}_c\) is the number of correctly predicted noisy annotations by NoGOA, respectively.

|       | BP | CC | MF |
|-------|----|----|----|
|       | \(N_c\) | \(\hat{N}_c\) | \(\bar{N}_c\) | \(N_c\) | \(\hat{N}_c\) | \(\bar{N}_c\) | \(N_c\) | \(\hat{N}_c\) | \(\bar{N}_c\) |
| MicroAvgF1 | 2 | 22 | 0 | 40 | 16 | 11 | 0 | 27 | 85 | 97 |
| MacroAvgF1 | 0 | 10 | 0 | 1 | 1 | 0 | 0 | 27 | 12 | 13 |
| AvgPrec | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 13 | 0 | 1 |
| AvgGCOC | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Fmax | 9 | 80 | 52 | 11 | 3 | 1 | 0 | 0 | 0 | 0 |
| Smin ↓ | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

References

Fu, G. *et al.* (2016) NegGOA: negative GO annotations selection using ontology structure. *Bioinformatics*, 32(19): 2996-3004.

Radivojac, P. *et al.* (2013) A large-scale evaluation of computational protein function prediction. *Nature Methods*, 10(3), 221-229.

Jiang, Y. *et al.* (2016) An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome Biology*, 17(1), 184.
Table S24: Statistics of noisy annotations prediction for different evidence codes on *H. sapiens* (archived date: May, 2015). \(N_c\) is the number of noisy annotations, \(\hat{N}_c\) is the number of predicted and \(\tilde{N}_c\) is the number of correctly predicted noisy annotations by NoGOA, respectively.

| BP   | TAS | EXP | IDA | IPI | IMP | IGI | IEP | ISA | ISO | ISM | RCA | ICA | IDA | IBA | IBD | IKR | IRD | ISS | NAS | RLC | IHE |
|------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| \(N_c\) | 989 | 0   | 10  | 0   | 10  | 21  | 0   | 0   | 0   | 0   | 0   | 253 | 0   | 0   | 9   | 2   | 0   | 4087 |
| \(\hat{N}_c\) | 62  | 0   | 4   | 0   | 2   | 0   | 0   | 0   | 0   | 2   | 0   | 0   | 42  | 0   | 0   | 4   | 1   | 0   | 2588 |
| \(\tilde{N}_c\) | 2   | 0   | 8   | 0   | 5   | 0   | 0   | 0   | 0   | 0   | 11  | 0   | 0   | 6   | 0   | 0   | 733  |
| \(N_c\) | 4   | 0   | 24  | 0   | 2   | 0   | 0   | 0   | 0   | 0   | 11  | 0   | 0   | 6   | 0   | 0   | 657  |
| \(\hat{N}_c\) | 0   | 0   | 3   | 0   | 2   | 0   | 0   | 0   | 0   | 0   | 3   | 0   | 0   | 3   | 0   | 0   | 351  |
| \(\tilde{N}_c\) | 11  | 0   | 17  | 1   | 0   | 0   | 0   | 0   | 0   | 0   | 20  | 0   | 0   | 1   | 0   | 0   | 2588 |

Table S25: Statistics of noisy annotations prediction for different evidence codes on *S. cerevisiae* (archived date: May, 2015). \(N_c\) is the number of noisy annotations, \(\hat{N}_c\) is the number of predicted and \(\tilde{N}_c\) is the number of correctly predicted noisy annotations by NoGOA, respectively.

| BP   | TAS | EXP | IDA | IPI | IMP | IGI | IEP | ISA | ISO | ISM | RCA | ICA | IDA | IBA | IBD | IKR | IRD | ISS | NAS | RLC | IHE |
|------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| \(N_c\) | 13  | 0   | 25  | 0   | 20  | 9   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 17  | 0   | 6   | 82  |
| \(\hat{N}_c\) | 23  | 0   | 21  | 10  | 95  | 41  | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 15  | 0   | 7   | 79  |
| \(\tilde{N}_c\) | 9   | 0   | 7   | 0   | 9   | 4   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 8   | 0   | 6   | 52  |
| \(N_c\) | 6   | 0   | 6   | 3   | 5   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 10  |
| \(\hat{N}_c\) | 5   | 0   | 5   | 0   | 5   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 9   |
| \(\tilde{N}_c\) | 7   | 0   | 0   | 0   | 1   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 110 |
| \(N_c\) | 4   | 0   | 0   | 0   | 1   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 88  |

Table S26: Examples of correctly (√) and wrongly(×) predicted direct noisy annotations by NoGOA in MF branch of *S. cerevisiae*.

| Protein | GO term                        | Evidence Codes | Details |
|---------|--------------------------------|----------------|---------|
| AAC1    | GO:0003735(structural constituent of ribosome) | IBA GO_REF:00000033 |         |
| AAC3    | GO:0003735(structural constituent of ribosome) | IBA GO_REF:00000033 |         |
| AAD14   | GO:0005515(protein binding) | IPI PMID:11805826 |         |
| AAP1    | GO:0005155(protein binding) | IPI PMID:11805826 |         |
| AAP1    | GO:00042277(peptide binding) | IBA GO_REF:0000033 |         |
|         | GO:00076006(metalloaminopeptidase activity) | IBA GO_REF:0000033 |         |

Table S27: Examples of correctly (√) and wrongly(×) predicted direct noisy annotations by NoGOA in BP branch of *S. cerevisiae*.

| Protein | GO term                        | Evidence Codes | Details |
|---------|--------------------------------|----------------|---------|
| AAC1    | GO:0006412(translation) | IBA GO_REF:00000033 |         |
| AAC3    | GO:0006412(translation) | IBA GO_REF:00000033 |         |
| AAD14   | GO:00062677(pre-replicative complex assembly involved in nuclear cell cycle DNA replication) | IPI PMID:16824194 |         |
|         | GO:0006270(DNA replication initiation) | IMP PMID:17053779 |         |
|         | GO:0004666(chromatin silencing at silent mating-type cassette) | IBA PMID:12897051 |         |
| AAP1    | GO:0005097(glycogen metabolic process) | IMP PMID:8100228 |         |
|         | GO:0006508(proteolysis) | IMP PMID:8100228 |         |
|         | GO:0043171(ribosome) | IBA GO_REF:0000033 |         |

Yu, G. et al. (2015) Predicting protein function via downward random walks on a gene ontology. *BMC Bioinformatics*, 15, 271.

Zhang, M.L. et al. (2014) A review on multi-label learning algorithms. *IEEE Trans. Knowledge & Data Engineering*, 99, 1–1.