Extensive Complementarity between Gene Function Prediction Methods

(Supplementary materials)

Vedrana Vidulin[1], Tomislav Šmuc[1] and Fran Supek[1,2]

[1]Division of Electronics, Ruđer Bošković Institute, Bijenička cesta 54, 10000 Zagreb, Croatia, [2]EMBL/CRG Systems Biology Unit, Centre for Genomic Regulation, Dr. Aiguader 88, 08003 Barcelona, Spain.

S1 IMPLEMENTATION DETAILS

Data collection and fusion. We collected the microbial genome data from NCBI (ftp://ftp.ncbi.nih.gov/genomes/) directories Bacteria and ASSEMBLY_BACTERIA on October 2014 and kept 2,071 high-quality genomes (≥0.95 according to Land *et al.*, 2014) covered by eggNOG v4 (Powell *et al.*, 2013). GO was downloaded from Uniprot-GOA database (Camon *et al.*, 2005) from December, 2013. Databases were fused on gene level using gene ID cross-references (Huang *et al.*, 2011). Where genes could not be assigned to OGs, their sequences were searched against eggNOG 4 database (eggnogv4.proteins.core_periphery.fa.gz) using the Lambda v0.4.7 (Hauswedell *et al.*, 2014) local aligner tool in blastp mode with an e-value threshold of 10^{-5}. A gene was assigned to the lowest e-value OG with percentage identity ≥90% and with alignment coverage ≥90% of both the query and the database sequence.

Data sets construction. Starting from the genomic data, five data sets were constructed, which were composed of the same set of instances that represent OGs occurring in ≥20 genomes, but with distinct sets of features (Fig. S1):

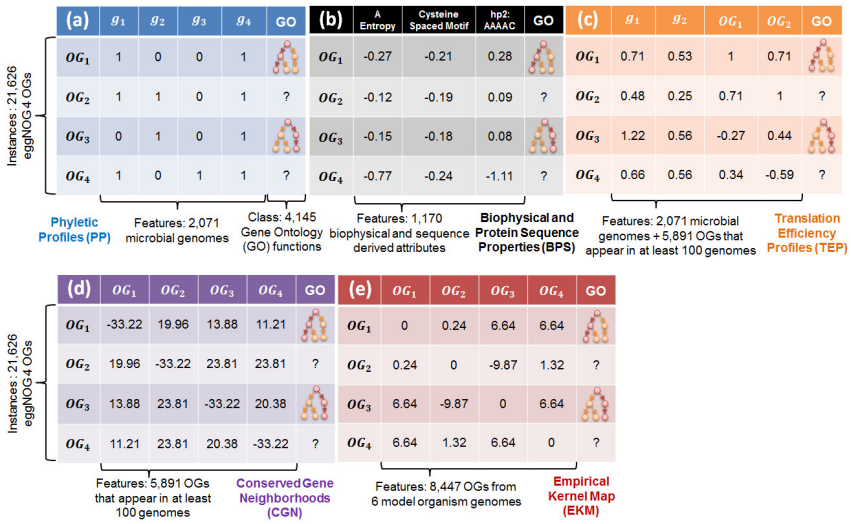

Fig. S1. A schematic overview of the five **genome-based representations used for gene function prediction in this work.**

Biophysical and protein sequence properties. BPS features can be divided into six categories (Ofer and Linial, 2015): biophysical quantitative properties (e.g., sequence length, molecular weight, net charge at various pH), word-based features (e.g., single and di-peptide amino acid composition, overlapping K-mers), local potential features (potential post-translational modification sites, potential disorder), information-based statistics (e.g., total entropy per letter, binary autocorrelation), amino acids scales-based features (amino acids are mapped to quantitative values representing their physicochemical or biochemical properties, such as hydrophobicity; scales are then used to represent a protein sequence as a time series and statistics like maximum, minimum and averages are computed for sliding windows of different sizes) and transformed CDT features (e.g., protein sequence composition regarding amino acids of different polarity, hydrophobicity, charge or solvent accessibility (Dubchak *et al.*, 1995)). Feature values were computed by inputting at most 20 (randomly selected when ≥20) protein sequences per OG into the ProFET Python script (https://github.com/ddofer/ProFET) and by averaging the resulting gene-level values.

Conserved gene neighborhoods. A feature value represents a log-distance (in the number of nucleotides) between an instance (row) OG and a feature (column) OG, averaged over 2,071 genomes. It is computed by: first, sampling at most 20 pairs of distinct genes (selecting randomly when ≥20) from a genome, such that one gene represents the instance OG and another represents the feature OG; second, computing for each pair of genes a distance as explained below and computing its $\log_2$; third, averaging the log-distances; fourth, collecting those averages for all genomes in which the instance and feature OGs co-occur; finally, averaging through genomes if the OGs co-occur in ≥3 genomes or otherwise taking the $\log_2$ of maximum distance (equal to the length of the longest observed genome).

$$D\left(g_i, g_j\right) = \min\left(s_j - e_i,\ L - e_j + s_i\right)$$

A distance between the two genes g_i and g_j, where g_i precedes g_j, was computed as:

This measure accounts for shorter distance between those measured in forward or reverse directions, assuming that microbial genomes are circular. In particular, g_i ending position e_i is first subtracted from g_j starting position s_j and then the number of nucleotides that follow g_j is subtracted from the genome length L and the result added to the number of nucleotides that precede g_i.

Empirical kernel map. Features represent OGs from the six model microorganisms: *Escherichia coli* K12 MG1655, *Streptomyces coelicolor* A3 2, *Bacteroides fragilis* NCTC 9343, *Bacillus subtilis* 168, *Pseudomonas aeruginosa* PAO1 and *Staphylococcus aureus* NCTC 8325. Feature values are logarithms (base 2) of the lowest e-values observed between distinct pairs of genes (max. 20 per OG), one from the instance (row) and another from the feature (column) OG. E-values were computed with Lambda v.0.4.7 in blastp mode by searching genes from the rest of the 2,065 genomes against the genes of the six model organisms. To avoid possible circularities with the transfer of gene function *via* homology within OGs (see below), the Lambda e-values below 10^{-2} were represented as missing values.

Translation efficiency profiles. Values of the first group of features are predicted OG expression levels through 2,071 genomes. When a gene is present in a genome its expression was predicted from codon usage biases using the 'Measure independent of length and composition' (MILC; Supek and Vlahoviček, 2005). MILC compares a gene against highly expressed genes in which codon usage is biased towards efficiently translated codons (ribosomal genes, translation initiation/elongation factors, and chaperones). OG-level MILC is maximal MILC observed across the genes in an OG. Since the OG presence matrix is sparse, we substituted missing values with random values sampled from the first quartile of the known MILC values for the genome. Values of the second group of features are the predicted OG co-expression levels through multiple genomes, i.e., Spearman's rank correlation coefficients computed between pairs of MILC vectors (without random values) one representing an instance OG and another a feature OG. Coefficients were computed for OGs that co-occur in ≥3 genomes. For other OG-pairs it is assumed that they are not correlated (feature value is zero).

Functional annotation scheme. Each OG was annotated with a set of GO terms that were originally assigned to ≥50% of OG member genes, counting only across genes that initially had any GO term assigned. Annotations with evidence codes denoting both the experimental and the electronic annotations from all three GO domains were assigned to OGs, while propagating upwards to the GO root.

In the analysis we differentiate GO terms by their *generality* and *information accretion*. Generality is expressed though Shannon Information Content (IC) that assigns high scores to infrequently used terms (Bourne, 2009):

$$\mathrm{IC}(GO_i) = -\log_2 \mathrm{frequency}(GO_i)$$

Information accretion (IA) assigns high scores to GO terms that contribute with new information when added as a specialization of a parent or a set of parent terms (Clark and Radivojac, 2013):

$$\mathrm{IA}(GO_i) = -\log_2 \mathrm{P}(GO_i|T)$$

T is a set of parent terms in GO and P denotes conditional probability.

IC and IA were measured among UniProt-GOA genes of the 2,071 genomes that received at least one annotation. IA was computed using the *SemDist* R package (Gonzalez and Clark, 2014).

Hierarchical multi-label classification. A separate classifier was constructed from each of the five data sets using CLUS-HMC (https://dtai.cs.kuleuven.be/clus/) with default parameters, except for these settings: decision tree pre-pruning to prevent the algorithm to form a leaf node when the number of instances in the node is <5; forest size to 200 trees; size of feature subsets to square root of the total number of features. Classifiers were constructed from 15,318 OGs with at least one GO term assigned. Predictions were collected for both annotated (from the out-of-bag crossvalidation procedure) and unannotated OGs. For each OG, a classifier outputs a vector of confidence scores ranging from zero to one, which indicate classifier's confidences in assigning each of the GO terms to the OG.

Late fusion schemes. Two basic approaches to fusion of multiple classification models are early and late fusion. The former concatenates distinct feature sets into a single data set from which one classifier is constructed, while the later fuses confidence scores output by the separate classifiers, each constructed from a distinct feature set. In practice, for the late fusion schemes that work across OGs, we took as input a set of vectors with confidence scores output by the five individual classifiers. We then combined them into a new vector with fused confidence scores – the 'one vote' scheme, for example, means taking the maximum confidence score observed between the individual classifiers for each OG-GO pair; the 'best precision' scheme takes the maximum Pr score (see below).

Converting confidence scores into precision (Pr) scores. The confidence scores of the individual classifiers and the fused confidence scores were converted into Pr scores which, unlike the confidences, have a probabilistic interpretation: they are equivalent to 1 - false discovery rate. First, for each classifier/scheme pair, the mapping between confidence and Pr scores were computed separately for each GO category by constructing a precision-recall (P-R) curve. In particular, this entails: varying confidence thresholds from 1.0 to 0.0, with the step of 0.001, consequently increasing the number of OGs annotated with the GO. At each threshold, we computed the number of true positives (TP) that represent the correctly predicted true annotations, false positives (FP) that represent the number of incorrectly predicted true annotations and Pr score that represent a proportion of predictions known to be true: TP/(TP+FP). Then, for each OG-GO pair, the confidence score was rounded to three decimals and substituted with the Pr score corresponding to that specific confidence threshold and the GO of interest. The first step was performed on the training set OGs, while the second step was applied to all OGs.

Evaluation measures in cross-validation. Classifier/scheme performance in cross-validation (i.e., out-of-bag) was evaluated using P-R curves and the 'area under the P-R curve' (AUPRC) scores. P-R curves were computed separately for each GO category by varying a Pr threshold from one to zero and collecting at each threshold TP, FP, false negatives (FN) that represent the number of missed true annotations, precision (TP/(TP+FP)) and recall (TP/(TP+FN)) that represents a proportion of true annotations that were successfully predicted. Intermediate P-R points were estimated using linear interpolation. P-R curves for individual GO terms were averaged and presented on a

graph where recall is plotted on x and precision on y-axis. AUPRC was computed as area enclosed between x-axis and the curve (when min. observed recall was >0, the precision computed at this min. point was estimated at recall =0 point in order to close the curve). Curves shifted to the left and upwards (AUPRC closer to one) denote better performance of the classifier or late fusion scheme.

Validation using CAFA 2 benchmark. We downloaded the benchmark from http://biofunctionprediction.org/node/12, which included: 70 *E. coli* 'no-knowledge' benchmark genes (with no previous annotations in all three domains), 406 experimentally-verified GO annotations assigned to them (232 Biological process, 139 Molecular function and 35 Cellular component GOs), and the results of 129 automated function prediction (AFP) methods and BLAST baseline on that benchmark (full evaluation mode).

Classifier/scheme performance on CAFA 2 was evaluated by measuring F_{max} according to the CAFA 2 rules:

$$Precision(t) = \frac{1}{m(t)} \sum_{i=1}^{m(t)} \frac{\sum_{f \in Functions} 1(f \in P_i(t) \wedge f \in E_i)}{\sum_{f \in Functions} 1(f \in P_i(t))}$$

$$Recall(t) = \frac{1}{n} \sum_{i=1}^{n} \frac{\sum_{f \in Functions} 1(f \in P_i(t) \wedge f \in E_i)}{\sum_{f \in Functions} 1(f \in E_i)}$$

$$F_{max} = \max_{t} \left(\frac{2 \times Precision(t) \times Recall(t)}{Precision(t) + Recall(t)} \right)$$

where t denotes the Pr threshold, which varies from 0.01 to 1 with the step of 0.01, n represents the number of *E. coli* benchmark genes, while $m(t)$ the number of those genes having score $\geq t$ for at least one GO term. $P_i(t)$ is a set of GOs having score $\geq t$ for gene i, and E_i denotes the set of experimentally-verified GOs for the same gene. Standard deviation of F_{max} was computed using bootstrapping with 10,000 iterations on the set of genes in evaluation (Efron and Tibshirani, 1994). F_{max} was computed separately for each GO domain.

Note on comparison of results from cross-validation and CAFA 2 benchmark. The comparison of results from cross-validation experiments and CAFA 2 benchmark has to be interpreted in light of the following points:

- Different metrics are used: cross-validation experiments were assessed using AUPRC while the CAFA 2 benchmark used F_{max} measure. AUPRC entails comparing classifiers over the whole range of Pr scores/recall thresholds, while F_{max} is based on a single Pr/recall point, the one at which F is maximal for the particular classifier/scheme. F_{max} measure in that sense under-emphasizes the differences between classifiers/schemes, in contrast to the AUPRC which does not depend on choosing the optimal cutoff point for individual classifiers.
- Number and the distribution of GO functions over which the results are aggregated is very different for the CAFA 2 benchmark and for our cross-validation experiments, especially for the Molecular function and Cellular component domains.

Despite the above, the results demonstrate that fusion schemes generally improve over individual classifiers, with some exceptions in Cellular component domain, which can be explained as a combined effect of very small number of GO functions (in terms of number of predictions made), and the nature of F_{max} measure, which is determined on different Pr levels for each classifier/scheme. Moreover, despite the combined models being based on overall precision weighting, high F_{max} scores were achieved in comparison with best methods on CAFA 2 challenge, especially on the Biological process domain.

References

Bourne, P. E. (2009) The Gene Ontology's Reference Genome Project: a unified framework for functional annotation across species. PLoS computational biology, 5(7), e1000431.

Clark, W. T., & Radivojac, P. (2013) Information-theoretic evaluation of predicted ontological annotations. Bioinformatics, 29(13), i53-i61.

Dubchak, I., Muchnik, I., Holbrook, S. R., & Kim, S. H. (1995) Prediction of protein folding class using global description of amino acid sequence. Proceedings of the National Academy of Sciences, 92(19), 8700-8704.

Efron, B., & Tibshirani, R. J. (1994) An introduction to the bootstrap. CRC press.

Gonzalez, I. & Clark, W. (2014) SemDist: Information accretion-based function predictor evaluation. R package version 1.5.0, http://github.com/iangonzalez/SemDist.

Hauswedell, H., Singer, J., & Reinert, K. (2014) Lambda: the local aligner for massive biological data. Bioinformatics, 30(17), i349-i355.

Huang, H., McGarvey, P. B., Suzek, B. E., Mazumder, R., Zhang, J., Chen, Y., & Wu, C. H. (2011) A comprehensive protein-centric ID mapping service for molecular data integration. Bioinformatics, 27(8), 1190-1191.

Land, M. L., Hyatt, D., Jun, S. R., Kora, G. H., Hauser, L. J., Lukjancenko, O., & Ussery, D. W. (2014) Quality scores for 32,000 genomes. Stand Genomic Sci, 9, 20.

Ofer, D., & Linial, M. (2015) ProFET: Feature engineering captures high-level protein functions. Bioinformatics, btv345.

Powell, S., Forslund, K., Szklarczyk, D., Trachana, K., Roth, A., Huerta-Cepas, J., ... & Bork, P. (2013) eggNOG v4.0: nested orthology inference across 3686 organisms. Nucleic acids research, gkt1253.

Supek, F., & Vlahoviček, K. (2005) Comparison of codon usage measures and their applicability in prediction of microbial gene expressivity. BMC bioinformatics, 6(1), 182.

S2 SUPPLEMENTARY RESULTS

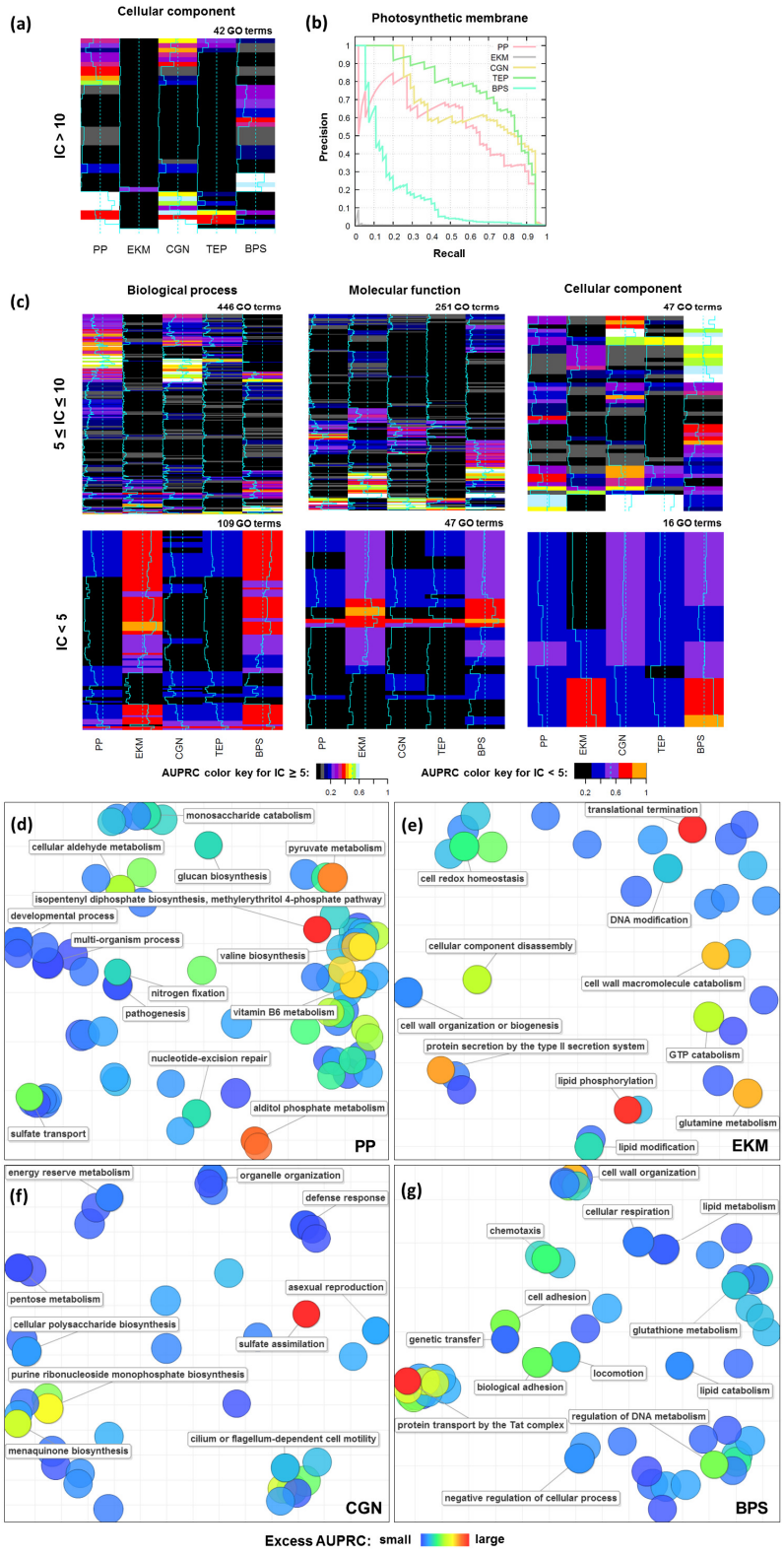

Fig. S2. Complementarity between the individual AFP methods. Heatmaps (**a** and **c**) represent complementarity patterns related to GO terms of different domains and generality levels: rows are GO terms, columns are classifiers and brighter colors (as well as higher histogram bars) indicate higher AUPRC. Hierarchical clustering was applied to rows. (**b**) Precision-recall curve for a selected GO term where the TEP method performs well. Examples of GO terms with positive excess AUPRC, meaning they are learned by one classifier better than by the rest of the classifiers, are presented for distinct classifiers in Revigo plots (**d**-**g**). Excess AUPRC for a classifier and a GO term is computed by subtracting GO AUPRC of the second best-performing classifier from the classifier's GO AUPRC. Large excess (red) indicates high complementarity, i.e., high level of classifier's specialization in learning a GO term. PP, phyletic profiles; EKM, empirical kernel map; CGN, conserved gene neighborhoods; TEP, translation efficiency profiles; BPS, biophysical and protein sequence properties.

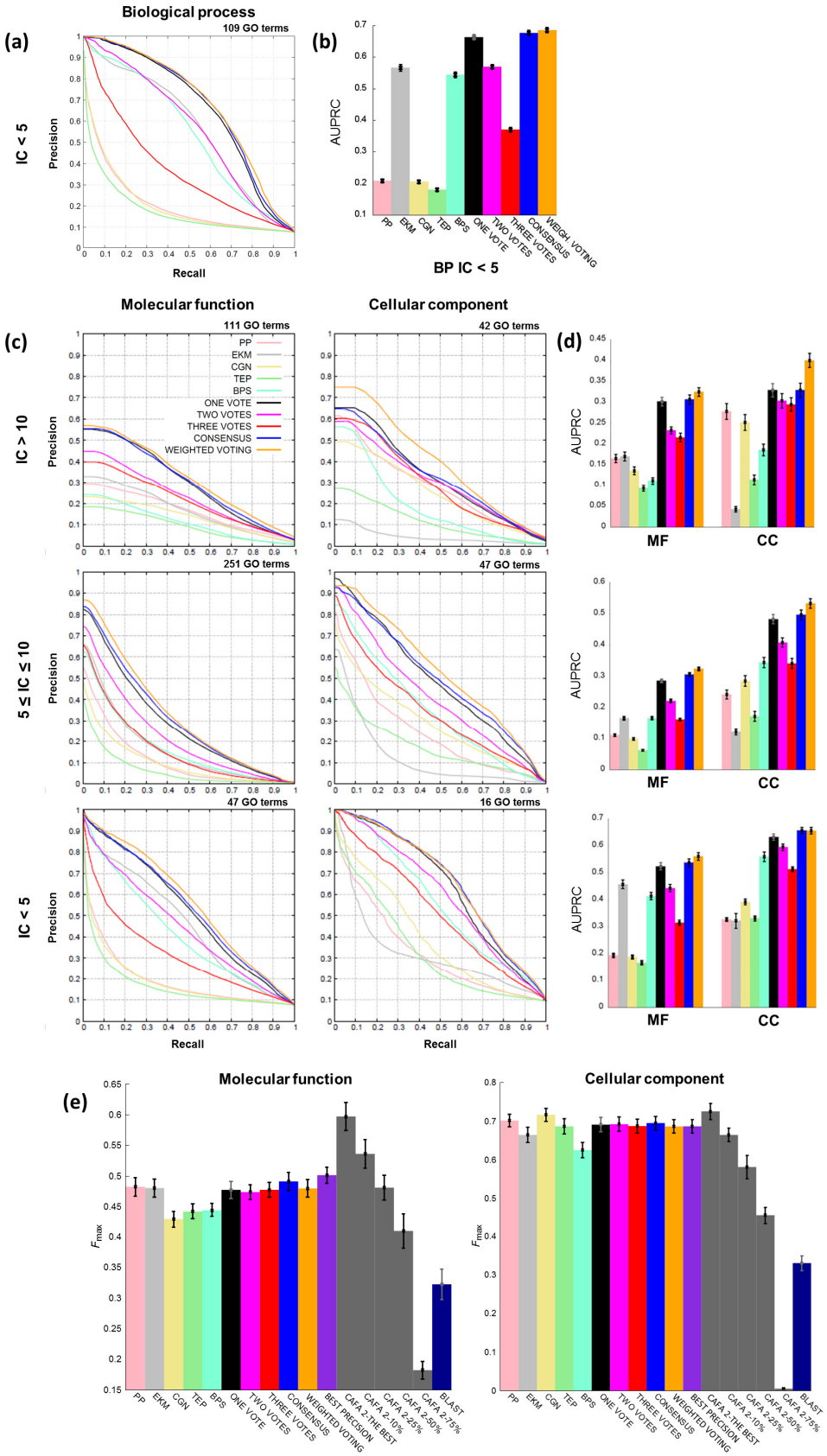

Fig. S3. Comparisons of predictive performance of individual methods and fusion schemes, measured in cross-validation and on the CAFA 2 benchmark for the Molecular function and Cellular component GO domains. (**a**, **c**) Precision-recall (P-R) curves are computed by averaging P-R curves of individual GO terms, computed in cross-validation. (**b**, **d**) Bars represent the average AUPRCs computed from the P-R curves in (**a**, **c**) and error bars represent standard error of the mean. (**e**) Bars represent F_{max} on CAFA 2 *E. coli* benchmark and error bars represent standard deviation obtained by bootstrapping the set of benchmark genes. PP, phyletic profiles; EKM, empirical kernel map; CGN, conserved gene neighborhoods; TEP, translation efficiency profiles; BPS, biophysical and protein sequence properties.

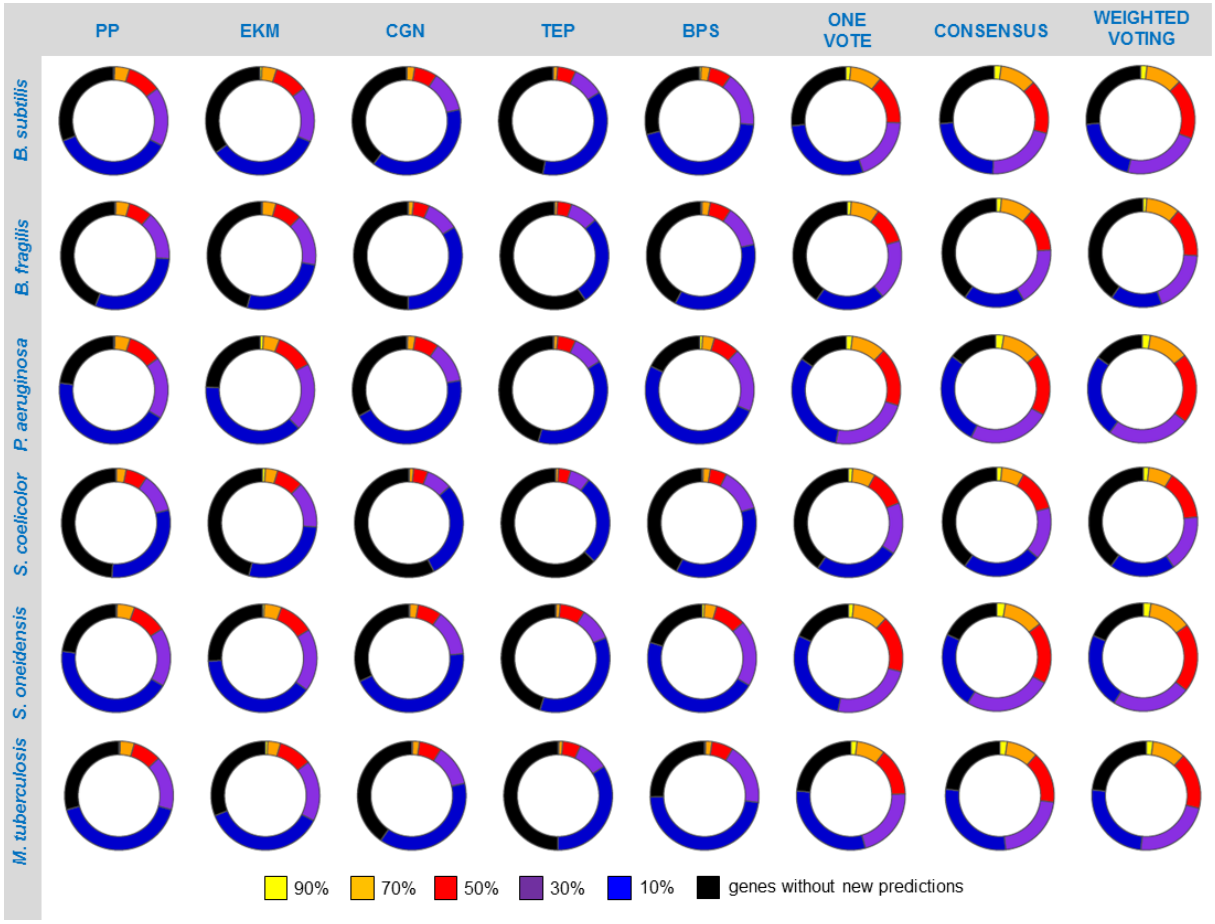

Fig. S4. Proportion of genes in six representative microorganisms that received at least one novel specific GO prediction (IC≥5) at different Pr thresholds.

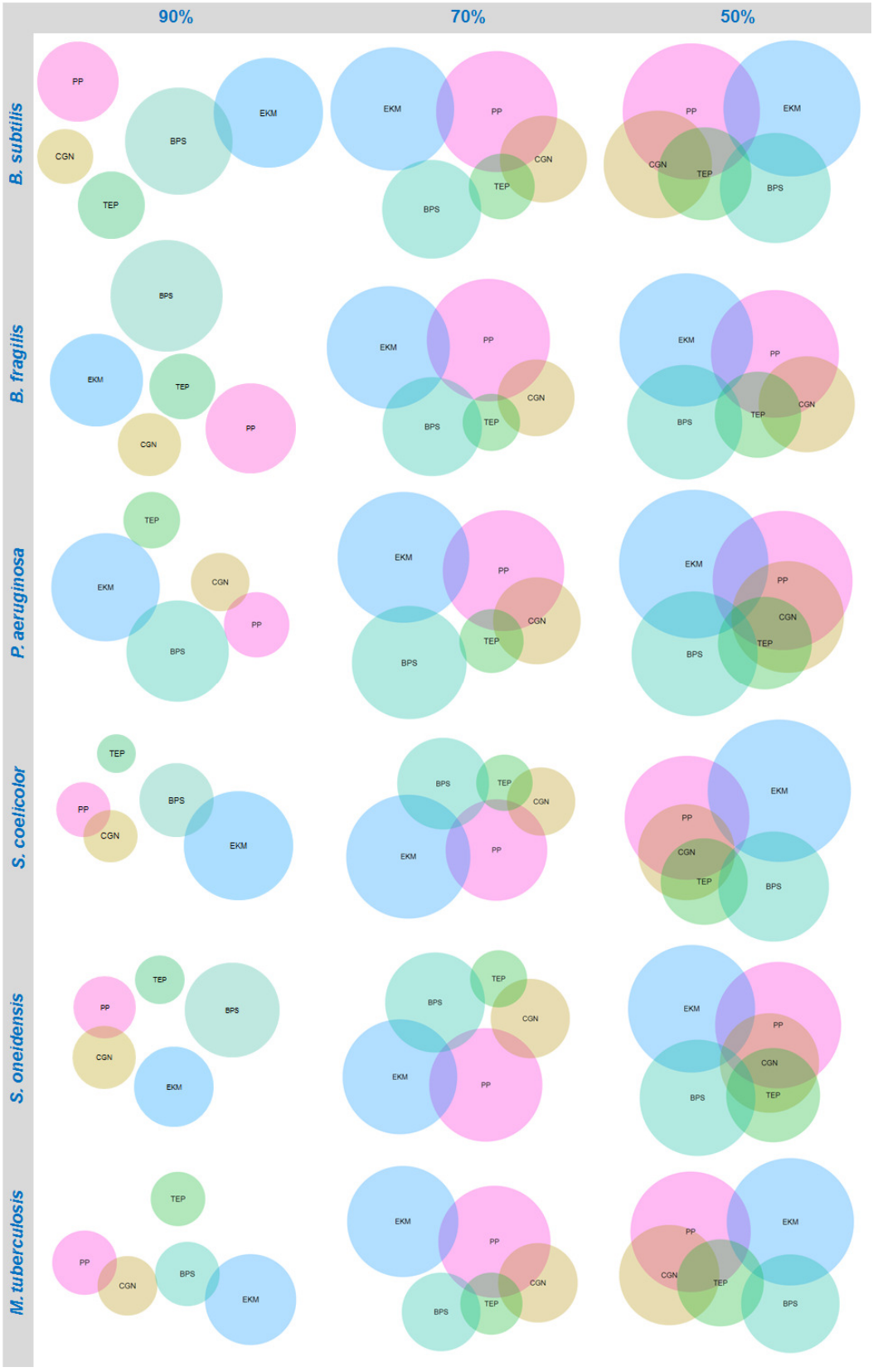

Fig. S5. The overlap between prediction methods in terms of genes of representative organisms that received at least one novel prediction at three Pr thresholds. Venn diagrams are approximate and may omit minor in order to emphasize major overlaps; exact data is given in Table S1.

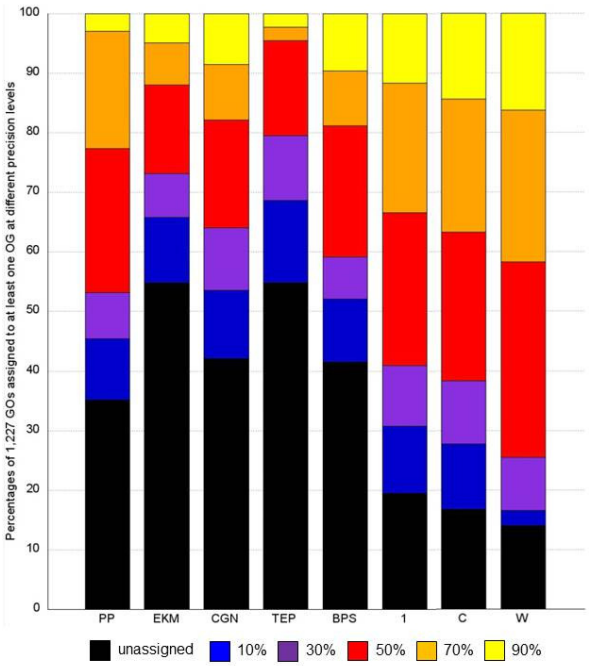

Fig. S6. Percentages of 1,227 GOs assigned by individual prediction methods and fusion schemes to at least one OG at different Pr thresholds. PP, phyletic profiles; EKM, empirical kernel map; CGN, conserved gene neighborhoods; TEP, translation efficiency profiles; BPS, biophysical and protein sequence properties; 1, one vote; C, consensus; W, weighted voting.

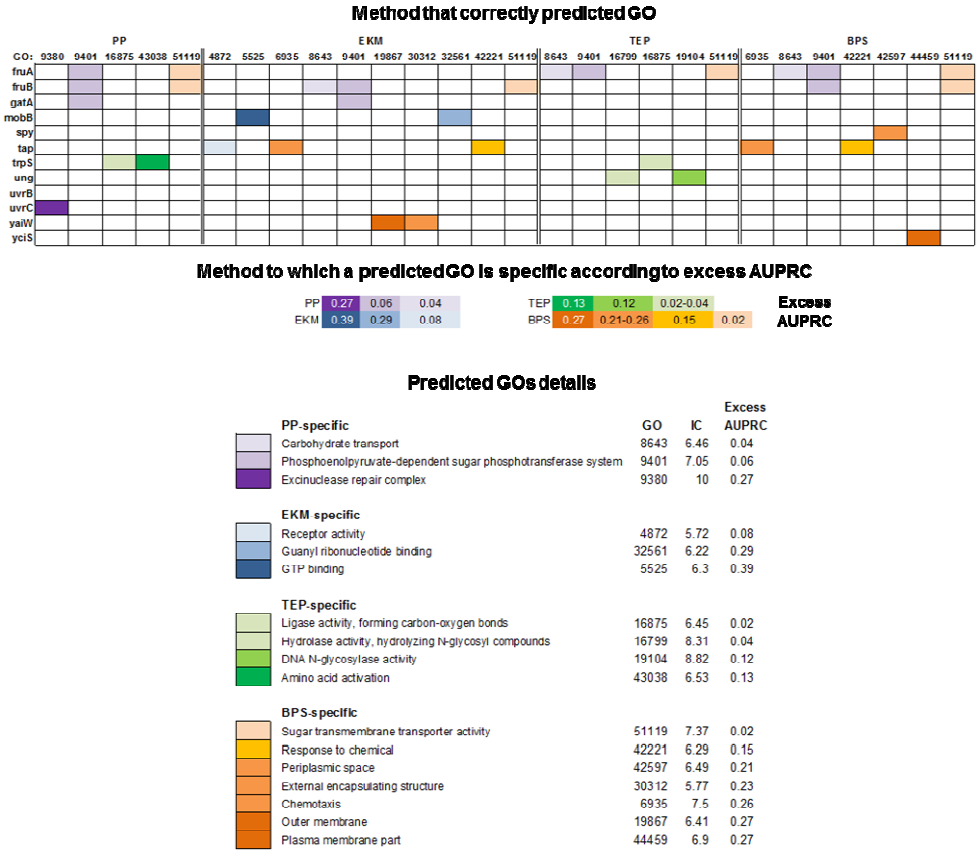

Fig. S7. Overlap between methods in terms of *E. coli* genes that received CAFA 2-validated novel predictions at Pr thresholds corresponding to F_{max}. Rows represent genes, columns are GOs with IC≥5 and colored cells represent validated predictions. Colors are grouped to represent methods with positive excess AUPRC for the predicted GOs (e.g., purple for PP). Color intensity represents excess AUPRC value. The table represents a reduced subset of predictions that still maintains overlap patterns. F_{max} thresholds: Biological process: PP 0.44, EKM 0.39, CGN 0.5, TEP 0.51, BPS 0.56; Molecular function: PP 0.45, EKM 0.64, CGN 0.41, TEP 0.39, BPS 0.43; Cellular component: PP 0.26, EKM 0.14, CGN 0.26, TEP 0.21, BPS 0.12.

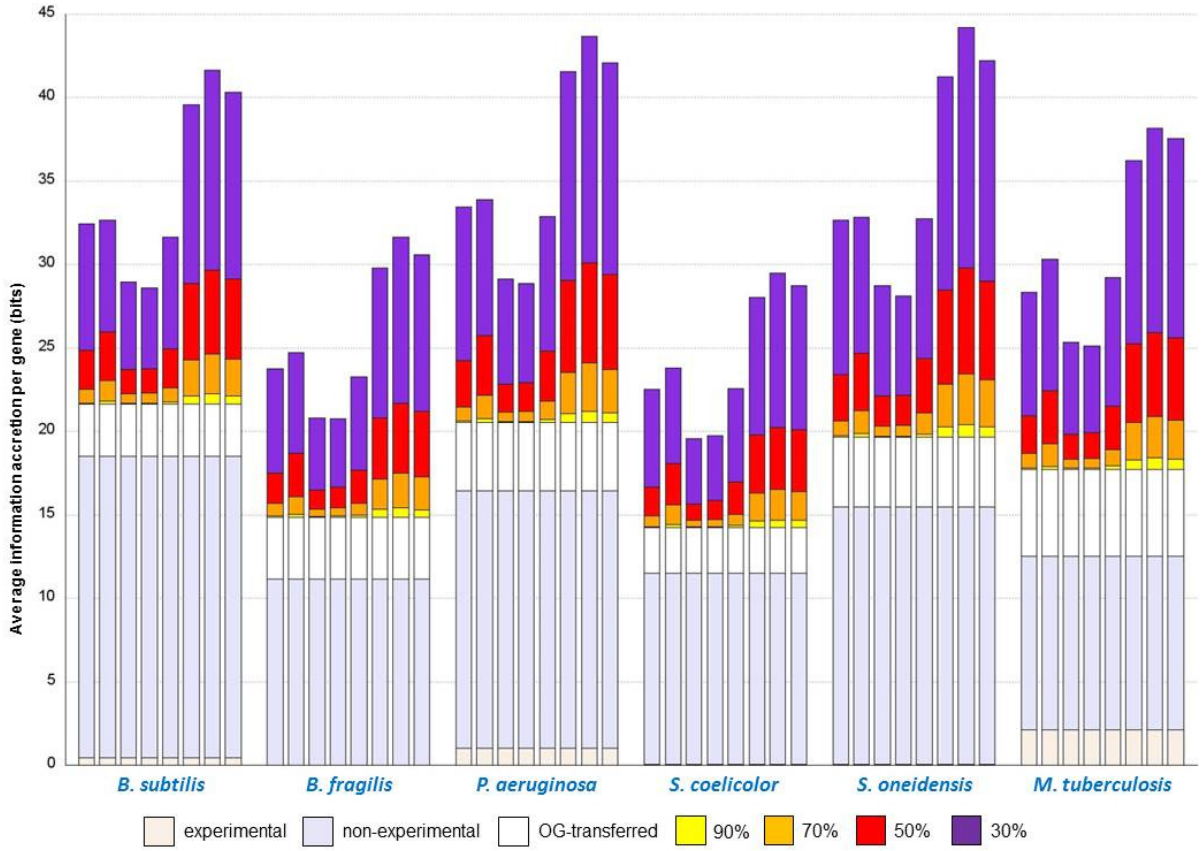

Fig. S8. Average information accretion per gene of the known annotations and the novel annotations assigned by methods and fusion schemes at different Pr thresholds. For a microorganism, the bars represent methods/schemes in the following order: PP, EKM, CGN, TEP, BPS, ONE VOTE, CONSENSUS and WEIGHTED VOTING. PP, phyletic profiles; EKM, empirical kernel map; CGN, conserved gene neighborhoods; TEP, translation efficiency profiles; BPS, biophysical and protein sequence properties.

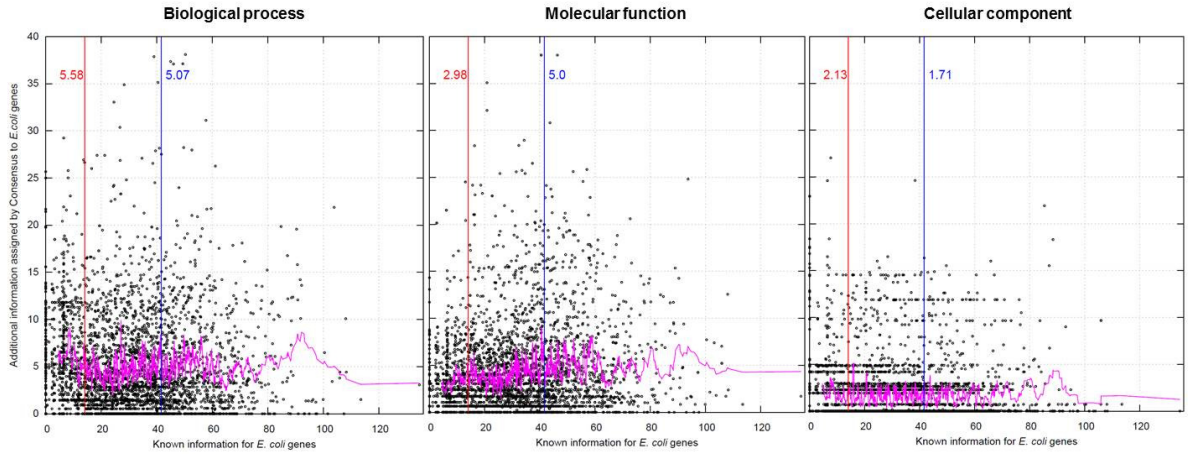

Fig. S9. Known *vs.* novel information for *E. coli* genes. Scatterplot points represent *E. coli* genes, while their positions represent known information accretion (x-axis) *vs.* new information accretion that was assigned by the 'consensus' fusion scheme (y-axis). Each scatterplot presents information that stems only from the domain-specific GO annotations assigned to the genes. Information is expressed in bits. Red line represents the first quartile with genes having the least known information and blue line the third quartile. The numbers beside the lines represent average information accretion (in bits/gene) per first/third quartile genes contributed by Consensus at Pr≥50% to the specific domain. Magenta line represents a moving average over novel information.

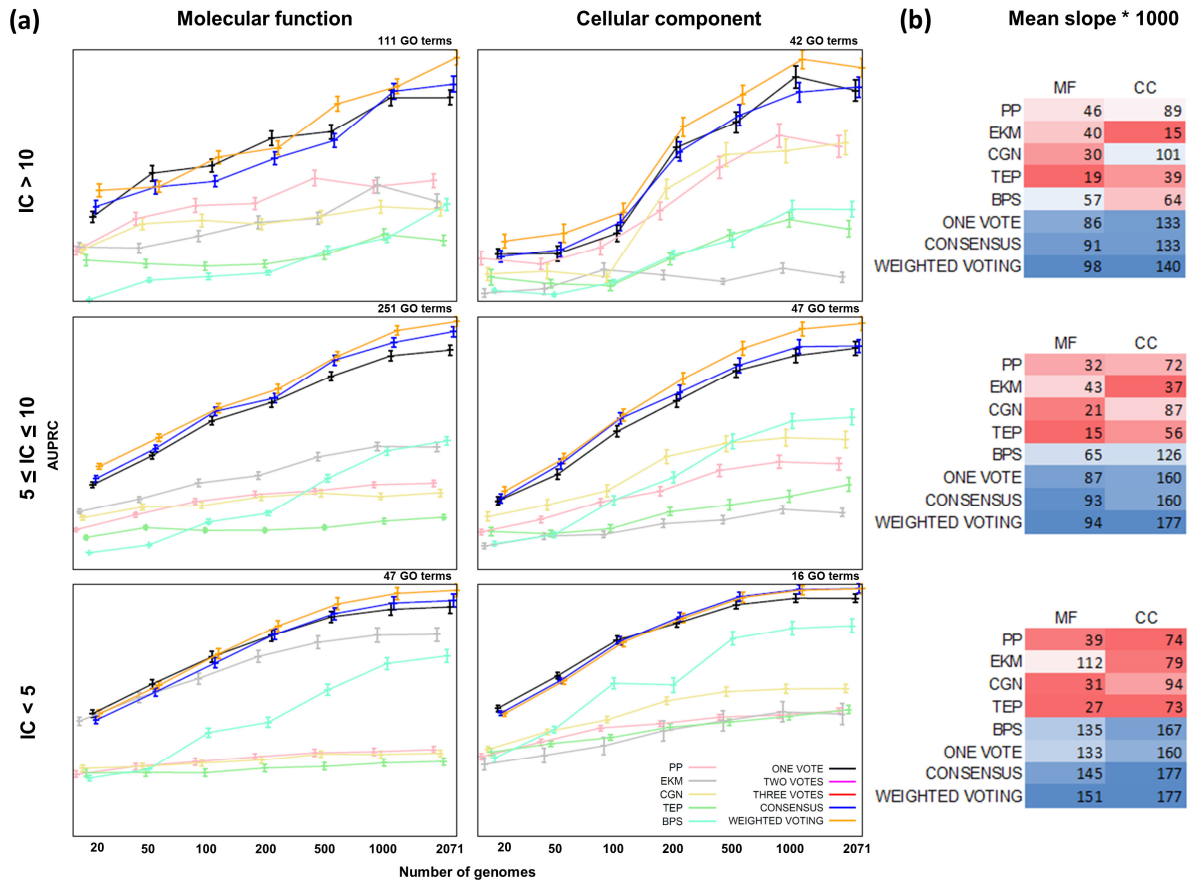

Fig. S10. Accuracy of classifiers increases with addition of genomic data. (a) X-axes represent the number of randomly sampled genomes (of the 2,071 total), shown in approximate log scale. Y-axes represents classifiers' AUPRCs (in cross-validation) averaged over the selected subset of GO terms from the Molecular function and Cellular component domains and error bars standard error of the mean. IC, information content. (b) Approximate slopes of the regression lines for a prediction method/integration scheme, as average over the slopes of segments connecting points in plot; complete table with slopes in Table S3.

Table S1. The number of microorganisms' genes that received at least one novel prediction from different combinations of methods at three Pr thresholds. This data was used to draw Venn diagrams in Fig. 4b and Fig. S5.

	E. coli			*S. coelicolor*			*B. subtilis*			*B. fragilis*			*P. aeruginosa*			*S. aureus*			*S. oneidensis*			*M. tuberculosis*		
Combination/Precisions:	90	70	50	90	70	50	90	70	50	90	70	50	90	70	50	90	70	50	90	70	50	90	70	50
PP	13	181	262	14	172	254	14	139	214	12	135	157	19	182	283	10	100	158	12	166	214	13	117	156
EKM	18	180	375	69	309	622	23	178	356	13	148	292	54	260	536	5	100	226	22	182	339	27	149	331
CGN	16	89	155	12	64	107	7	56	109	7	45	72	15	80	125	9	57	92	11	69	94	11	50	83
TEP	13	34	65	11	37	75	8	29	53	6	21	39	15	41	82	5	27	36	9	28	115	10	26	51
BPS	47	215	419	35	151	332	22	108	196	19	96	200	48	210	378	13	75	151	32	137	298	13	72	180
PP, EKM	0	20	79	0	19	91	0	13	64	0	19	59	0	30	97	0	11	40	0	20	74	0	17	55
PP, CGN	1	37	108	4	22	117	0	26	83	0	17	69	1	37	145	0	22	74	1	24	99	1	23	91
PP, TEP	0	12	53	0	15	64	0	10	38	0	6	43	0	12	57	0	8	32	0	11	55	0	11	39
PP, BPS	0	4	45	0	8	40	0	6	41	0	5	38	0	3	32	0	3	21	0	4	37	0	3	20
EKM, CGN	0	9	35	0	4	38	0	4	14	0	2	12	0	4	30	0	4	15	0	5	18	0	3	21
EKM, TEP	0	1	21	0	0	17	0	0	13	0	2	9	0	0	18	0	0	16	0	0	22	0	0	17
EKM, BPS	1	15	127	2	35	104	1	11	65	0	12	77	1	23	138	0	8	40	0	29	109	1	5	54
CGN, TEP	0	7	30	0	13	33	0	11	21	0	6	22	0	11	27	0	7	20	0	8	23	0	8	24
CGN, BPS	0	5	48	0	1	14	0	3	17	0	1	11	0	3	23	0	1	15	0	3	19	0	1	14
TEP, BPS	0	3	12	0	3	12	0	3	15	0	3	10	0	3	15	0	3	12	0	4	14	0	3	10
PP, EKM, CGN	0	0	19	0	1	21	0	0	23	0	0	15	0	2	31	0	0	14	0	0	30	0	0	18
PP, EKM, TEP	0	2	22	0	0	20	0	2	18	0	0	14	0	3	19	0	0	19	0	2	15	0	0	17
PP, EKM, BPS	0	0	15	0	3	14	0	0	13	0	0	10	0	0	11	0	0	4	0	0	14	0	0	8
PP, CGN, TEP	0	3	64	0	5	56	0	2	57	0	0	33	0	3	64	0	0	38	0	3	51	0	2	44
PP, CGN, BPS	0	2	26	0	1	12	0	0	9	0	1	6	0	1	17	0	1	11	0	1	15	0	1	10
PP, TEP, BPS	0	1	13	0	2	7	0	1	6	0	1	5	0	1	9	0	1	5	0	1	9	0	3	8
EKM, CGN, TEP	0	0	6	0	0	7	0	0	6	0	0	6	0	0	8	0	0	7	0	0	6	0	0	7
EKM, CGN, BPS	0	0	9	0	0	8	0	0	1	0	0	1	0	0	5	0	0	1	0	0	6	0	0	1
EKM, TEP, BPS	0	1	6	0	0	3	0	0	4	0	1	4	0	1	3	0	0	1	0	1	4	0	0	2
CGN, TEP, BPS	0	0	4	0	0	5	0	0	4	0	0	3	0	0	4	0	0	4	0	0	4	0	0	4
PP, EKM, CGN, TEP	0	0	25	0	2	23	0	0	22	0	0	9	0	1	29	0	0	12	0	0	22	0	2	15
PP, EKM, CGN, BPS	0	0	5	0	0	4	0	0	2	0	0	2	0	0	4	0	0	2	0	0	5	0	0	2
PP, EKM, TEP, BPS	0	0	13	0	0	35	0	0	13	0	0	11	0	0	14	0	0	5	0	0	9	0	0	9
PP, CGN, TEP, BPS	0	1	18	0	0	7	0	1	14	0	0	7	0	1	16	0	0	7	0	1	14	0	0	8
EKM, CGN, TEP, BPS	0	0	3	0	0	4	0	0	1	0	0	0	0	0	2	0	0	1	0	0	2	0	0	2
PP, EKM, CGN, TEP, BPS	0	0	7	0	0	8	0	0	3	0	0	5	0	0	10	0	0	5	0	0	7	0	0	7

Table S2. The number of GO terms that different combinations of methods simultaneously assigned to at least one OG at different Pr thresholds.

Combination/Precisions:	90	70	50
PP	11	79	99
EKM	38	41	71
CGN	39	48	23
TEP	0	2	9
BPS	52	84	143
PP, EKM	0	14	1
PP, CGN	7	68	132
PP, TEP	0	1	22
PP, BPS	0	12	25
EKM, CGN	0	0	2
EKM, TEP	0	7	0
EKM, BPS	18	46	82
CGN, TEP	7	3	2
CGN, BPS	28	6	19
TEP, BPS	7	2	1
PP, EKM, CGN	0	5	8
PP, EKM, TEP	0	2	1
PP, EKM, BPS	0	5	17
PP, CGN, TEP	10	17	57
PP, CGN, BPS	8	41	45
PP, TEP, BPS	0	2	8
EKM, CGN, TEP	0	0	0
EKM, CGN, BPS	1	0	8
EKM, TEP, BPS	0	0	11
CGN, TEP, BPS	1	3	2
PP, EKM, CGN, TEP	0	2	23
PP, EKM, CGN, BPS	0	16	25
PP, EKM, TEP, BPS	0	4	20
PP, CGN, TEP, BPS	0	5	35
EKM, CGN, TEP, BPS	3	0	3
PP, EKM, CGN, TEP, BPS	0	5	57

Table S3. Slopes of the lines from Fig. 5a and S10. Large numbers (green boxes) represent steeper slopes, which indicate bigger improvements in accuracy with arrival of more sequenced genomes. In contrast, small numbers (red boxes) represent less steep and negative slopes, which indicate saturation and suggest no further improvement from additional genomic data. Slope values are determined as in Fig S10 and are multiplied by 1000.

		Biological process							Molecular function							Cellular component						
		≤ 20	≤ 50	≤ 100	≤ 200	≤ 500	≤ 1000	≤ 2071	≤ 20	≤ 50	≤ 100	≤ 200	≤ 500	≤ 1000	≤ 2071	≤ 20	≤ 50	≤ 100	≤ 200	≤ 500	≤ 1000	≤ 2071
IC > 10	PP	75	35	143	91	69	32	36	56	114	63	9	91	-40	28	55	-24	93	202	181	177	-58
	EKM	15	-13	26	51	15	45	-34	60	-4	56	65	15	157	-72	11	18	104	-28	-27	72	-46
	CGN	35	198	-9	114	76	-28	51	57	92	17	-18	28	46	-13	35	11	-30	485	143	21	41
	TEP	27	14	-15	19	-12	122	0	45	-12	-11	9	32	99	-27	31	-27	-11	147	101	85	-51
	BPS	7	13	54	-2	50	89	9	2	69	17	18	79	61	151	14	-17	66	151	64	169	-1
	ONE VOTE	74	164	38	146	119	65	27	93	156	36	130	24	161	1	60	2	113	473	105	251	-72
	CONSENSUS	94	176	4	72	87	58	50	104	70	27	109	64	234	32	57	28	152	390	152	127	27
	WEIGHTED VOTING	104	151	-13	251	44	64	63	122	13	141	44	159	81	130	77	33	117	469	136	192	-44
5 ≤ IC ≤ 10	PP	54	143	86	74	27	17	8	39	48	58	31	12	27	6	63	68	130	79	114	62	-13
	EKM	36	26	-11	34	39	9	15	57	42	69	32	61	46	-3	39	55	12	78	21	78	-24
	CGN	102	15	11	25	43	2	-25	51	36	6	39	12	-15	16	88	67	103	249	73	40	-11
	TEP	38	45	-21	15	43	25	12	32	32	-12	1	8	28	15	63	-13	37	124	39	64	82
	BPS	11	35	47	32	74	64	27	16	25	100	38	115	119	40	43	48	255	166	197	148	26
	ONE VOTE	97	122	77	126	63	68	28	84	95	149	80	85	89	24	115	146	311	222	168	110	51
	CONSENSUS	119	102	71	133	63	81	-3	91	98	159	58	124	76	44	119	190	329	191	151	135	4
	WEIGHTED VOTING	132	90	72	126	56	70	28	103	94	126	82	108	111	37	131	181	314	266	167	142	38
IC < 5	PP	107	57	54	47	24	7	16	106	46	35	38	24	9	15	161	99	125	37	44	19	34
	EKM	255	179	187	176	88	58	13	200	135	144	166	81	56	5	145	61	85	133	66	80	-18
	CGN	110	48	49	40	27	14	11	117	11	26	23	34	-2	8	176	120	103	168	61	23	5
	TEP	101	14	20	31	18	40	23	109	4	-3	36	9	27	11	170	57	53	93	34	51	56
	BPS	123	1	469	64	286	276	75	100	50	280	77	185	201	55	158	190	409	-13	311	90	22
	ONE VOTE	290	213	282	163	104	62	20	213	168	205	163	107	57	18	260	216	317	151	123	58	-2
	CONSENSUS	277	223	297	207	123	67	15	202	161	219	219	119	80	18	250	218	346	216	136	62	8
	WEIGHTED VOTING	285	206	321	199	121	66	22	213	163	238	215	126	80	22	249	220	344	213	138	64	9