Supplementary Information

# Automatic policing of functional annotations using genomic correlations

Tzu-Lin Hsiao[1,*], Olga Revelles[2,*], Lifeng Chen[1,*], Uwe Sauer[2], Dennis Vitkup[1]

[1]Center for Computational Biology and Bioinformatics and Department of Biomedical Informatics, Columbia University, 1130 St. Nicholas Ave., Irving Cancer Research Center, New York, NY 10032, USA
[2]Institute of Molecular Systems Biology, ETH Zurich, CH-8093 Zurich, Switzerland

* These authors contributed equally to this manuscript.
(The correspondence should be addressed to DV at dv2121@columbia.edu)

## Supplementary Results

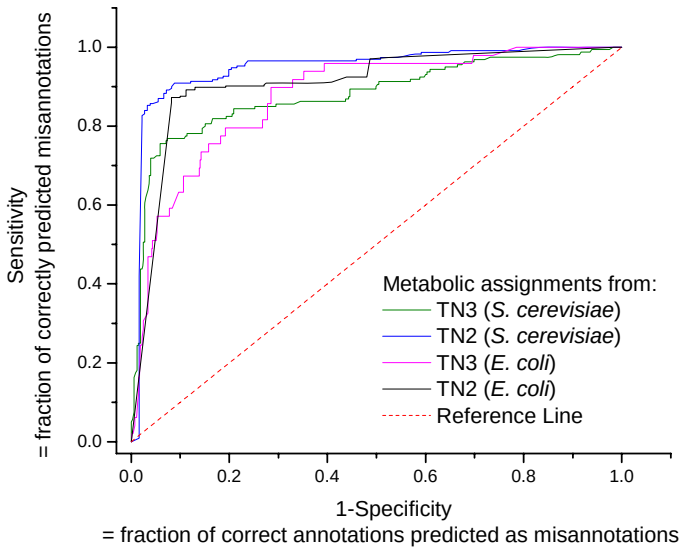
**Supplementary Table 1.** Analysis of potential misannotations listed in Table 1 of the manuscript.

| Gene Name | Comments and evidence supporting the identified misannotation |
|---|---|
| *adhB* | The gene is annotated in KEGG as alcohol dehydrogenase EC 1.1.1.284. It is annotated in Swiss-Prot as "NAD alcohol dehydrogenase". It is marked as "alcohol dehydrogenase" in MetaCyc. No literature/experimental evidence supporting the annotation is available. |
| *alaT/yugH* | The gene is annotated in KEGG as an ortholog of N-succinyldiaminopimelate aminotransferase EC 2.6.1.17. It is annotated in Swiss-Prot as "alanine transaminase". In MetaCyc the gene is marked as "similar to aspartate aminotransferase". No literature/experimental evidence supporting the annotation is available. |
| *bcsA* | The gene is annotated in KEGG/MetaCyc as naringenin-chalcone synthase EC 2.3.1.74. It is marked as "putative chalcone synthase" in Swiss-Prot. The annotation is based on remote sequence homology and not experimental evidence. The enzyme is involved in flavonoid biosynthesis, which occurs primarily in plants[1]. |
| *bsaA* | The gene is annotated in KEGG as glutathione peroxidase EC 1.11.1.9. It is marked in Swiss-Prot as "glutathione peroxidase homolog" and in MetaCyc as "glutathione peroxidase". Several studies suggested that glutathione is probably absent in *B. subtilis*[2, 3]. |
| *Cad/speA* | The gene is annotated in KEGG/Swiss-Prot as arginine decarboxylase EC 4.1.1.19. The gene is annotated in MetaCyc as lysine decarboxylas EC 4.1.1.18. While this gene was previously thought to be for lysine decarboxylase, it was later characterized to be arginine decarboxylase[4]. |
| *dgkA* | The gene is annotated in KEGG/MetaCyc/Swiss-Prot as diacylglycerol kinase EC 2.7.1.107. However, in the recently published paper (July 2007) by Jerga *et* |

| | |
|---|---|
| | *al.*[5], the authors confirmed that *dgkA* is not a diacylglycerol kinase (DagK) but rather an undecaprenol kinase. |
| *hipO/ytnL* | The gene is annotated in KEGG and MetaCyc as hippurate hydrolase EC 3.5.1.32. In Swiss-Prot it is marked as "uncharacterized hydrolase". No literature/experimental evidence supporting the annotation is available. |
| *pps* | The gene is annotated in KEGG and MetaCyc as phosphoenolpyruvate synthase EC 2.7.9.2. In Swiss-Prot the protein is described as "phosphoenolpyruvate synthase". No literature/experimental evidence supporting the annotation is available. |
| *xpt* | The gene is annotated in MetaCyc as xanthine phosphoribosyltransferase and also EC 2.4.2.7 (adenine phosphoribosyltransferase). In Swiss-Prot and KEGG, the gene is annotated only as xanthine phosphoribosyltransferase EC 2.4.2.22. Arent *et al.* purified the protein xpt and showed that it is a highly xanthine specific enzyme without detectable activity using adenine as substrate[6]. |
| *ybbD* | The gene is annotated in KEGG as an ortholog of beta-N-acetylhexosaminidase EC 3.2.1.52. It is marked in MetaCyc as "similar to beta-hexosaminidase". No EC annotation is available in Swiss-Prot. No literature/experimental evidence supporting the annotation is available. |
| *ycgT* | The gene is annotated in KEGG as an ortholog of thioredoxin reductase (NADPH) EC 1.8.1.9. In MetaCyc the protein is marked as "similar to thioredoxin reductase". No EC annotation is available in Swiss-Prot. No literature/experimental evidence supporting the annotation is available. |
| *yhcV* | The gene is annotated in KEGG as an ortholog of IMP dehydrogenase EC 1.1.1.205. No EC annotation is available in Swiss-Prot/MetaCyc. No literature/experimental evidence supporting the annotation is available. |
| *yhdR* | The gene is annotated in KEGG as aspartate aminotransferase EC 2.6.1.1. In MetaCyc the protein is marked as "similar to aspartate aminotransferase". No EC annotation is available in Swiss-Prot. No literature/experimental evidence supporting the annotation is available. |
| *yhfR* | The gene is annotated in KEGG as an ortholog of phosphoglycerate mutase (PGM) EC 5.4.2.1. In MetaCyc the protein is marked as "similar to phosphoglycerate mutase". No EC annotation is available in Swiss-Prot. Pearson *et al.*[7] demonstrated that *yhfR* is non-essential for growth, sporulation, and spore germination. They also purified the gene, expressed it in *E. coli* and *B. subtilis* but were not able to detect PGM activity in *B. subtilis*. |
| *yisP* | The gene is annotated in KEGG as an ortholog of phytoene synthase EC 2.5.1.32. No EC annotation is available in Swiss-Prot/MetaCyc. No literature/experimental evidence supporting the annotation is available. |
| *yitC* | The gene is annotated in KEGG as ortholog of 2-phosphosulfolactate phosphatase EC 3.1.3.71. In Swiss-Prot the protein is marked as "probable 2-phosphosulfolactate phosphatase". No EC annotation is available in MetaCyc. No literature/experimental evidence supporting the annotation is available. |
| *yjmC* | The gene is annotated in KEGG as an ortholog of malate dehydrogenase EC 1.1.1.37. It is marked but is marked as "uncharacterized oxidoreductase" (EC 1.1.1.-) in Swiss-Prot. In MetaCyc it is marked as "similar to malate dehydrogenase". As the paper by Mekjian *et al.*[8] suggests this gene is more likely to be involved in the glucuronate pathway (for which EC 1.1.1.37 is not a member), no literature/experimental evidence supporting the annotation is available. |
| *yktC* | The gene is annotated in Swiss-Prot and KEGG as myo-inositol-1(or 4)-monophosphatase EC 3.1.3.25. In MetaCyc it is annotated as "similar to myo-inositol-1(or 4)-monophosphatase". No literature/experimental evidence supporting the annotation is available. |
| *ykuR* | The gene is annotated in KEGG as an ortholog of N-acetyldiaminopimelate deacetylase EC 3.5.1.47. In MetaCyc it is marked as "similar to hippurate |

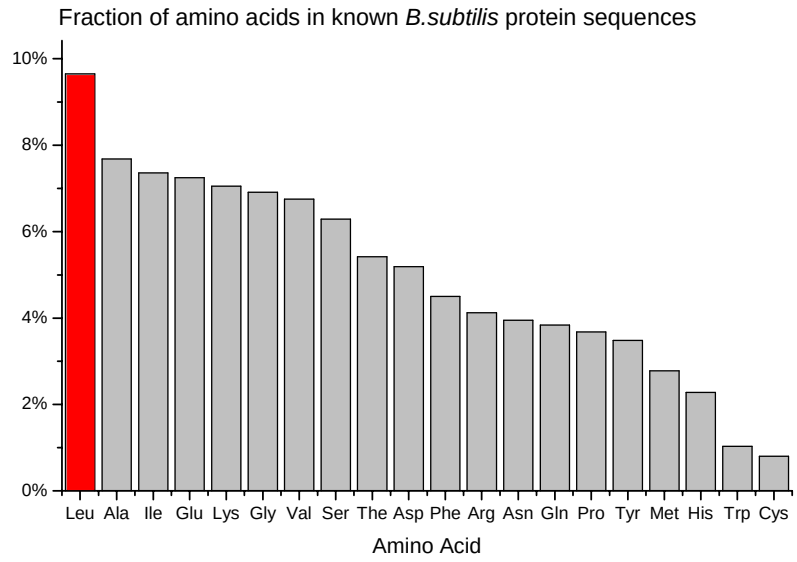
| | |
|---|---|
| | hydrolase". No EC annotation is available in Swiss-Prot. No literature/experimental evidence supporting the annotation is available. |
| *yngE* | The gene is annotated in KEGG as an ortholog of propionyl-CoA carboxylase beta chain EC 6.4.1.3. In MetaCyc it is marked as "similar to propionyl-CoA carboxylase" (EC 4.1.1.70). No EC annotation is available in Swiss-Prot. |
| *yngF* | The gene is annotated in KEGG as enoyl-CoA hydratase EC 4.2.1.17. It is listed in MetaCyc as "similar to 3-hydroxbutyryl-CoA dehydratase" (EC 4.2.1.17 EC 4.2.1.55). No EC annotation is available in Swiss-Prot. |
| *yngI* | The gene is annotated in KEGG as fatty-acyl-CoA synthase EC 2.3.1.86. Until recently this gene was annotated in KEGG to EC 6.2.1.3. No EC annotation is available in Swiss-Prot. In MetaCyc the protein is marked as "similar to long-chain acyl-CoA synthetase". No literature/experimental evidence supporting the annotation is available. |
| *yoaD* | The gene is annotated in KEGG as an ortholog of D-3-phosphoglycerate dehydrogenase EC 1.1.1.95. No EC annotation is available in Swiss-Prot. In MetaCyc the protein is marked as "similar to phosphorglycerate dehydrogenase". No literature/experimental evidence supporting the annotation is available. |
| *yogA* | The gene is annotated in KEGG as an ortholog of alcohol dehydrogenase EC 1.1.1.1. No annotation is available in Swiss-Prot. The protein is marked in MetaCyc as "similar to alcohol dehydrogenase". No literature/experimental evidence supporting the annotation is available. |
| *yqhT* | The gene is annotated in KEGG as an ortholog of EC 3.4.11.9. It is marked as "similar to Xaa-Pro dipeptidase" in MetaCyc. No literature/experimental evidence supporting the annotation is available. |
| *yrhE* | The gene is annotated in KEGG as an ortholog of formate dehydrogenase EC 1.2.1.2. In Swiss-Prot the protein is named "formate dehydrogenase chain A". In MetaCyc the protein is characterized as "similar to formate dehydrogenase". No literature/experimental evidence supporting the annotation is available. |
| *ysfC* | The gene is annotated in KEGG as an ortholog of (S)-2-hydroxy-acid oxidase EC 1.1.3.15. No EC annotation is available in Swiss-Prot. In MetaCyc, it is annotated as "similar to glycolate oxidase subunit". No literature/experimental evidence supporting the annotation is available. |
| *yumB* | The gene is annotated in KEGG as an ortholog of NADH dehydrogenase EC 1.6.99.3. No EC annotation is available in Swiss-Prot. In MetaCyc it is marked as "similar to NADH dehydrogenase". In the paper by Gyan *et al.*[9], the authors tested the growth of three *B. subtilis* genes potentially responsible for EC 1.6.99.3 (*yumB*, *yjlD*, and *yutJ*). Only *yjlD*- mutant showed growth slower growth on the LB media, while the other two grew as well as wild type. |
| *yumC* | The gene is annotated in KEGG as an ortholog of thioredoxin reductase EC 1.8.1.9. In Swiss-Prot the protein is named "thioredoxine reductase". In MetaCyc the protein is listed as "similar to thioredoxin reductase", In the study by Seo *et al.*[10] the protein *yumC* was purified and characterized as ferredoxin-NADP+ reductase (EC 1.18.1.2). |
| *yvcN* | The gene is annotated in KEGG as an ortholog of EC 2.3.1.5 and is marked in Swiss-Prot as "uncharacterized acetyltransferase" (EC 2.3.1.-). No EC annotation is available in MetaCyc. No literature/experimental evidence supporting the annotation is available. |
| *yvcT* | The gene is annotated in KEGG as gluconate 2-dehydrogenase EC 1.1.1.215. The gene is marked "similar to glycerate dehydrogenase" in MetaCyc and marked as probable EC 1.1.1.215 in Swiss-Prot. No literature/experimental evidence supporting the annotation is available. |
| *ywrD* | The gene is annotated in KEGG as an ortholog of gamma-glutamyltranspeptidase EC 2.3.2.2, This enzyme is necessary to utilize glutathione (GSH) as the sulfur source. No EC annotation for this gene is |

| | |
|---|---|
| | available in Swiss-Prot/MetaCyc. It has been shown by Minami *et al.*[11] that *ywrD* mutant grows well on minimal media supplied with GSH as the sole sulfur source. In addition, His-tag purified *ywrD* cannot hydrolyze GSH. |

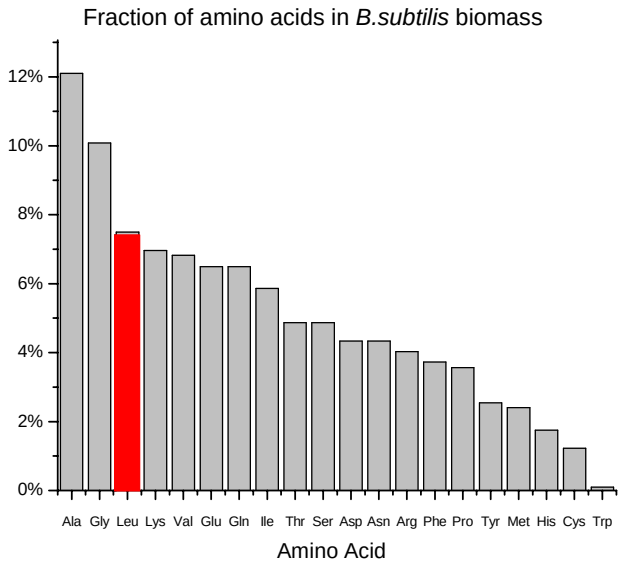



**Supplementary Figure 1.** The algorithm performance on the *S. cerevisiae* and *E. coli* metabolic networks. The ROC curves are based on TN2 and TN3 sets (see text for details). The performance of the method on the *E. coli* and *S. cerevisiae* networks was very similar, although the algorithm was optimized on the *S. cerevisiae* network and applied to *E.coli* without further modification. The areas under ROC curves are: 0.91 (95% CI: 0.88-0.93) and 0.93 (95% CI: 0.90 – 0.95) for TN2, and 0.88 (95% CI: 0.83-0.92) and 0.87 (95% CI: 0.86-0.88) for TN3 in *E. coli* and *S. cerevisiae*, respectively.

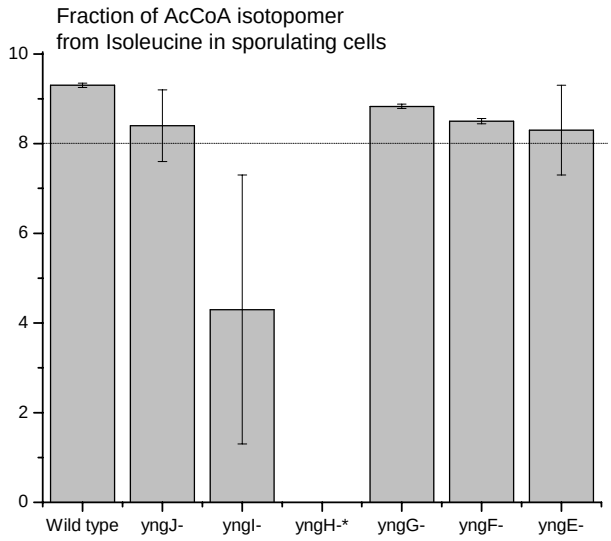
**a**



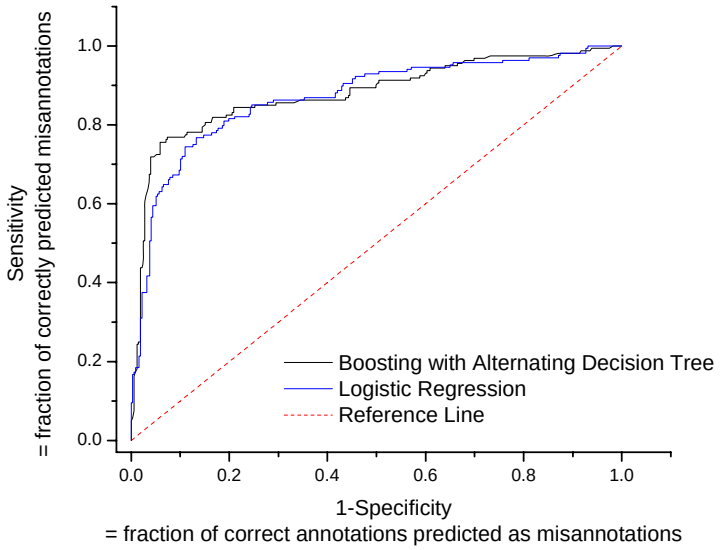

**b**




**Supplementary Figure 2.** The amino acid composition of *B. subtilis* protein sequences and biomass. a) The fractions of amino acids in all known *B. subtilis* protein sequences. Leucine is the most frequent amino acid, responsible for about 10% of residues in *B. subtilis* proteins. b) The amino acid composition of *B. subtilis* biomass determined in logarithmically growing wild-type strains[12]. Leucine is one of the most abundant amino acid, responsible for about 8% of all amino acids. Combined concentrations of asparagine/aspartate and glutamine/glutamate were measured; for display purposes equal fractions were assumed (asparagine=aspartate, glutamine=glutamate). Cysteine was measured as cystic acid.




**Supplementary Figure 3.** The fractional labeling of the Acetyl-CoA m2 isotopomer from [U$^{13}$C]-L-Isoleucine. Cells were grown under sporulation conditions supplemented with [U$^{13}$C]-L-Isoleucine; metabolites were extracted 2.5 hours after the sporulation onset (see Methods). The standard error data were calculated based on two independent experiments. * *yngH*- not determined.




**Supplementary Figure 4.** The ROC curves for multivariable logistic regression (blue) and AdaBoost classifier (black) on the *S. cerevisiae* TN3 set (see main text for details). The AdaBoost algorithm tends to slightly outperform logistic regression (70% true positives for AdaBoost versus 60% true positives for logistic regression, at 5% false positives rate).

## Supplementary Methods

### Alternative location ratio (ALR)

For display purposes only (used in Table 1), we calculate the alternative location ratio (ALR) to indicate the existence of a good alternative location using the following equation.:

$$ALR = \frac{s - s_a}{\frac{1}{2} * | s + s_a |} \quad (2)$$

where $s$ is the AdaBoost classification score at the database assigned network location calculated using all available sequence and context-based descriptors, $s_a$ is the best classification score among all possible alternative locations. A negative ALR ratio indicates the existence of a better alternative network location; the smaller the ratio, the better fitness in the alternative location.

**Supplementary Table 2.** The 40 most commonly used metabolites that were removed before connectivity calculations.

| Metabolite | Number of reactions (EC numbers) connected |
|---|---|
| H2O | 1139 |
| H+ | 651 |
| NAD+ (**8**) | 432 |
| NADPH (**9**) | 414 |
| NADP+ (**9**) | 413 |
| NADH (**8**) | 411 |
| ATP (**10**) | 384 |
| Oxygen | 366 |
| ADP (**11**) | 288 |
| Orthophosphate (**12**) | 271 |
| CO2 | 244 |
| CoA (**13**) | 222 |
| Pyrophosphate (**14**) | 195 |
| NH3 | 182 |
| FAD (**15**) | 166 |
| UDP (**16**) | 154 |
| S-Adenosyl-L-methionine (**17**) | 142 |
| S-Adenosyl-L-homocysteine (**18**) | 130 |
| AMP (**19**) | 120 |
| Pyridoxal phosphate (**20**) | 115 |
| Pyruvate (**21**) | 113 |
| Acceptor | 112 |
| Reduced acceptor | 110 |
| Iron | 103 |

| | |
|---|---|
| Acetyl-CoA | 102 |
| H2O2 | 101 |
| 2-Oxoglutarate (**22**) | 94 |
| L-Glutamate (**23**) | 92 |
| Zinc | 78 |
| UDPglucose (**24**) | 77 |
| Acetate | 70 |
| D-Glucose (**25**) | 57 |
| Carboxylate | 51 |
| Manganese | 47 |
| Succinate (**26**) | 45 |
| Heme (**27**) | 42 |
| Oxaloacetate (**28**) | 41 |
| GDP (**29**) | 41 |
| Glycine (**30**) | 39 |
| Acyl-CoA | 38 |

## Supplementary References


1. Winkel-Shirley, B. Flavonoid biosynthesis. A colorful model for genetics, biochemistry, cell biology, and biotechnology. *Plant Physiol* **126**, 485-493 (2001).
2. Fahey, R.C., Brown, W.C., Adams, W.B. & Worsham, M.B. Occurrence of glutathione in bacteria. *J Bacteriol* **133**, 1126-1129 (1978).
3. Newton, G.L. *et al.* Distribution of thiols in microorganisms: mycothiol is a major thiol in most actinomycetes. *J Bacteriol* **178**, 1990-1995 (1996).
4. Sekowska, A., Bertin, P. & Danchin, A. Characterization of polyamine synthesis pathway in Bacillus subtilis 168. *Mol Microbiol* **29**, 851-858 (1998).
5. Jerga, A., Lu, Y.J., Schujman, G.E., de Mendoza, D. & Rock, C.O. Identification of a soluble diacylglycerol kinase required for lipoteichoic acid production in Bacillus subtilis. *J Biol Chem* **282**, 21738-21745 (2007).
6. Arent, S., Kadziola, A., Larsen, S., Neuhard, J. & Jensen, K.F. The extraordinary specificity of xanthine phosphoribosyltransferase from Bacillus subtilis elucidated by reaction kinetics, ligand binding, and crystallography. *Biochemistry* **45**, 6615-6627 (2006).
7. Pearson, C.L., Loshon, C.A., Pedersen, L.B., Setlow, B. & Setlow, P. Analysis of the function of a putative 2,3-diphosphoglyceric acid-dependent phosphoglycerate mutase from Bacillus subtilis. *J Bacteriol* **182**, 4121-4123 (2000).
8. Mekjian, K.R., Bryan, E.M., Beall, B.W. & Moran, C.P., Jr. Regulation of hexuronate utilization in Bacillus subtilis. *J Bacteriol* **181**, 426-433 (1999).
9. Gyan, S., Shiohira, Y., Sato, I., Takeuchi, M. & Sato, T. Regulatory loop between redox sensing of the NADH/NAD(+) ratio by Rex (YdiH) and oxidation of NADH by NADH dehydrogenase Ndh in Bacillus subtilis. *J Bacteriol* **188**, 7062-7071 (2006).

10. Seo, D., Kamino, K., Inoue, K. & Sakurai, H. Purification and characterization of ferredoxin-NADP+ reductase encoded by Bacillus subtilis yumC. *Arch Microbiol* **182**, 80-89 (2004).
11. Minami, H., Suzuki, H. & Kumagai, H. Gamma-glutamyltranspeptidase, but not YwrD, is important in utilization of extracellular blutathione as a sulfur source in Bacillus subtilis. *J Bacteriol* **186**, 1213-1214 (2004).
12. Sauer, U. *et al.* Physiology and metabolic fluxes of wild-type and riboflavin-producing Bacillus subtilis. *Appl Environ Microbiol* **62**, 3687-3696 (1996).