Global probabilistic annotation of metabolic networks enables enzyme discovery

Supplementary information

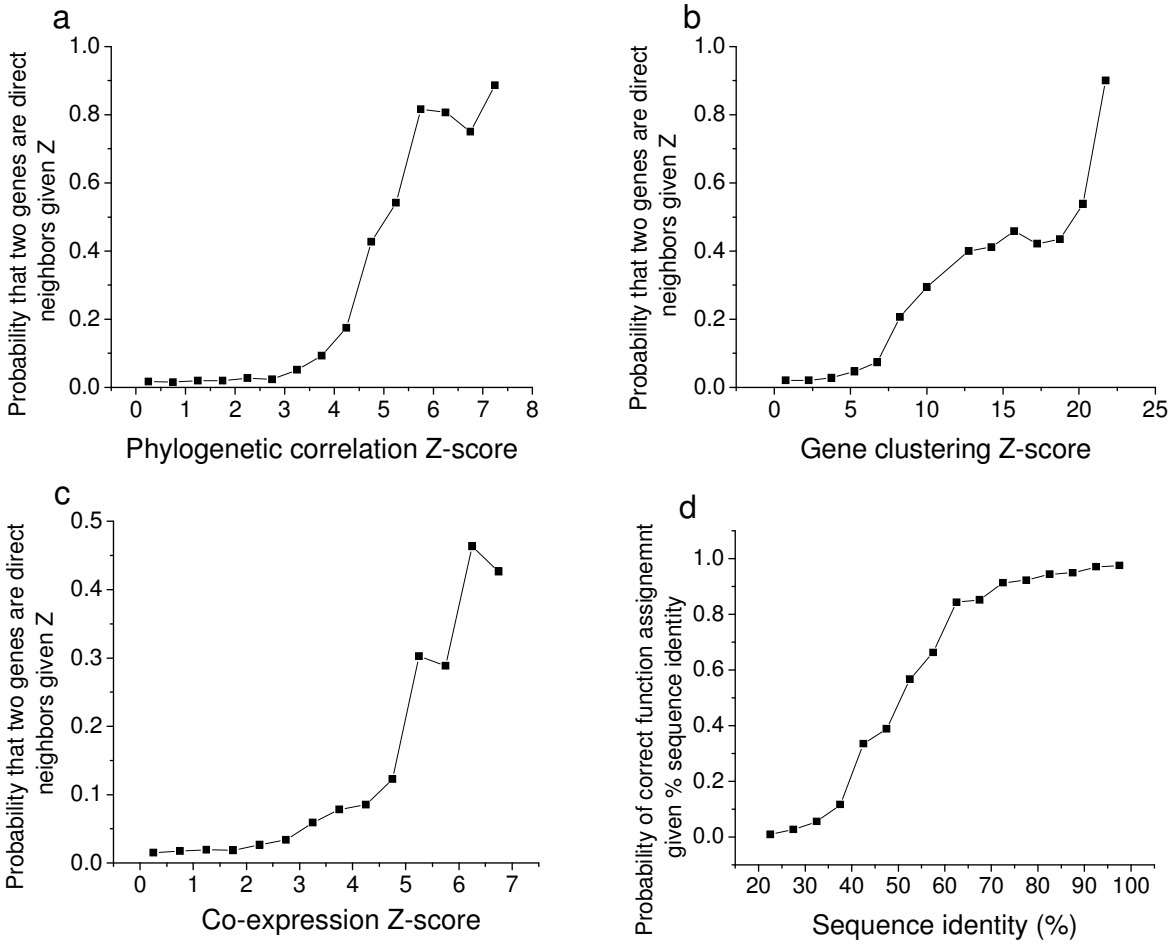
Germán Plata[1,2,*], Tobias Fuhrer[3,*], Tzu-Lin Hsiao[1,4,*], Uwe Sauer[3], Dennis Vitkup[1,4,a]

[1]Center for Computational Biology and Bioinformatics, Columbia University, New York, NY, U.S.A. [2]Integrated Program in Cellular, Molecular, Structural, and Genetic Studies, Columbia University, New York, NY, U.S.A. [3]Institute of Molecular Systems Biology, ETH Zurich, Zurich, Switzerland. [4]Department of Biomedical Informatics, Columbia University, New York, NY, U.S.A.

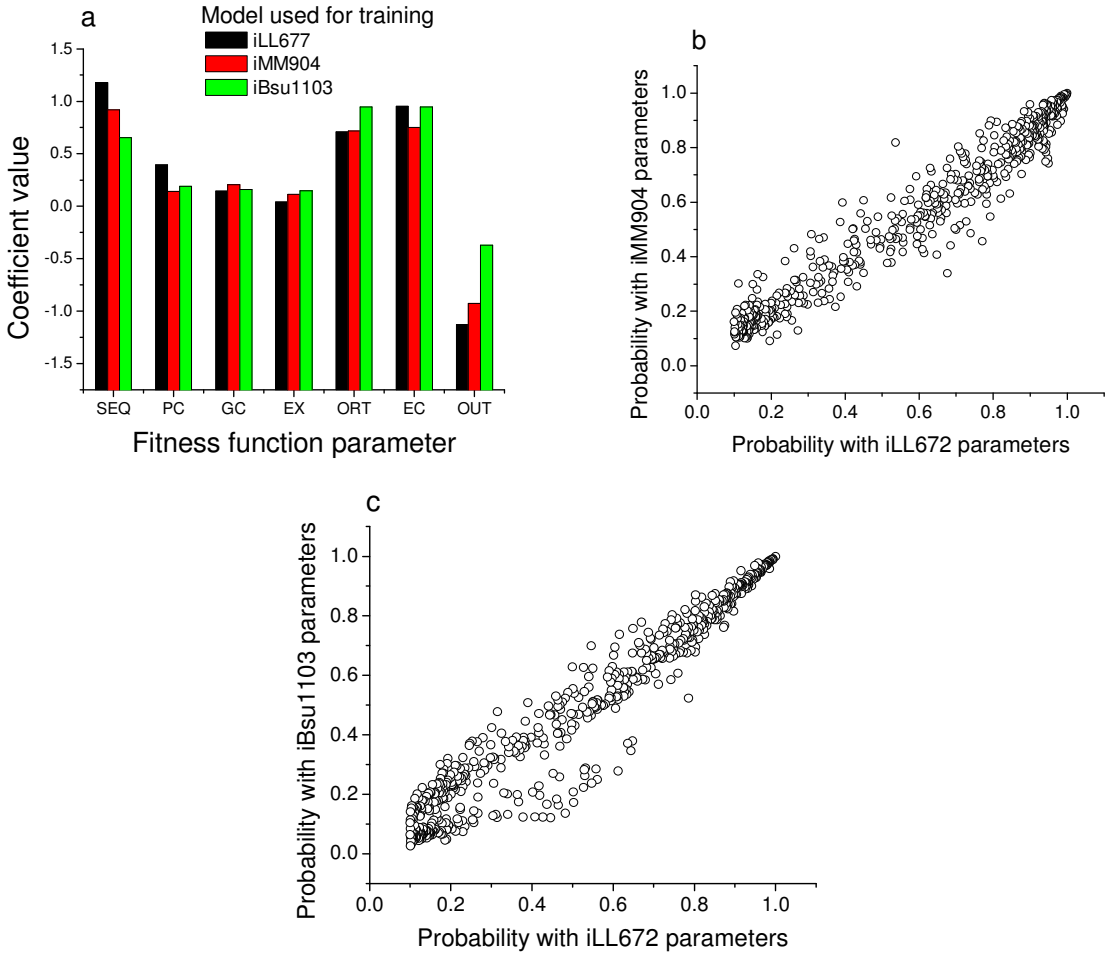
[*]equal contribution
[a]communication should be addressed to DV at dv2121@columbia.edu

Table of contents

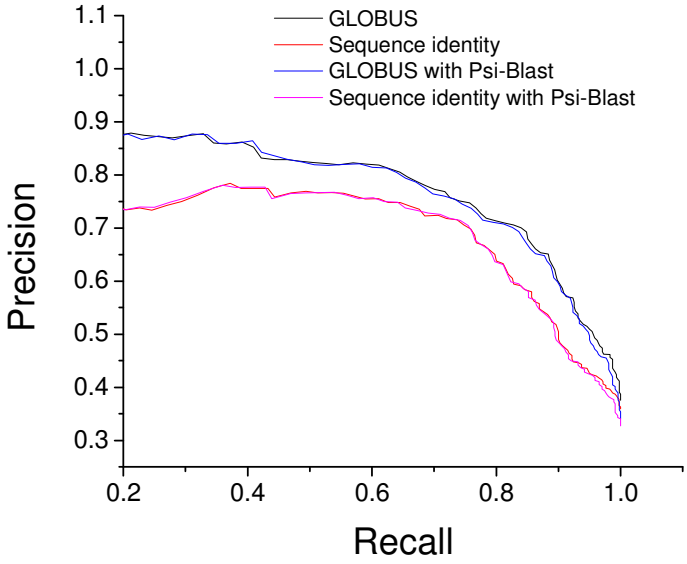
Supplementary Results



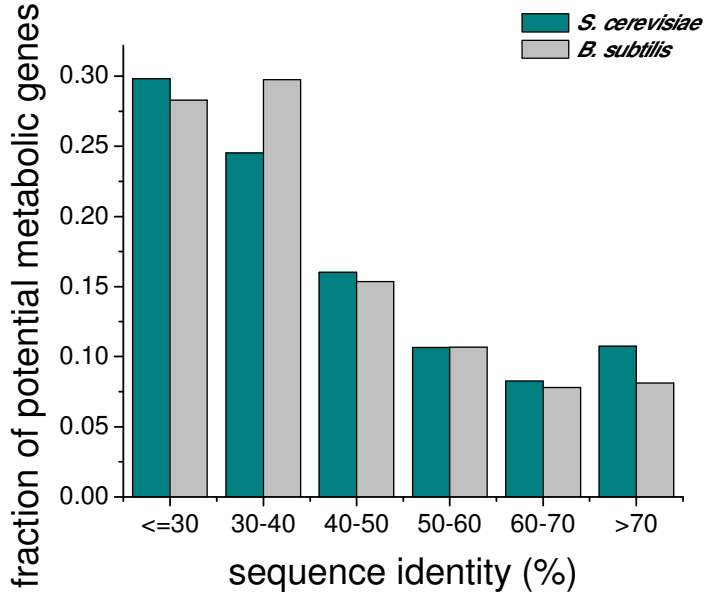
Supplementary Figure 1 | Conditional probabilities used in the GLOBUS fitness function. The context correlations used in GLOBUS (a-c) were first transformed into Z-scores[1] using the distribution of correlations for all pairs of candidate metabolic genes. Then we estimated the conditional probability that two genes are direct network neighbors given their context association Z-score. The greater the context correlation Z-score, the more likely the two genes are network neighbors. The conditional probabilities were estimated based on the iLL672 yeast metabolic model[2]. **(a)** The conditional probabilities for phylogenetic profiles, **(b)** The conditional probabilities for chromosomal gene clustering, **(c)** The conditional probabilities for mRNA co-expression. **(d)** As a sequence homology term in GLOBUS we used the conditional probability that a gene performs the assigned function, given the highest sequence identity to a Swiss-Prot[3] protein annotated to catalyze the target activity. The conditional probabilities for sequence homology were estimated using the well-curated yeast iLL672 metabolic model[2].



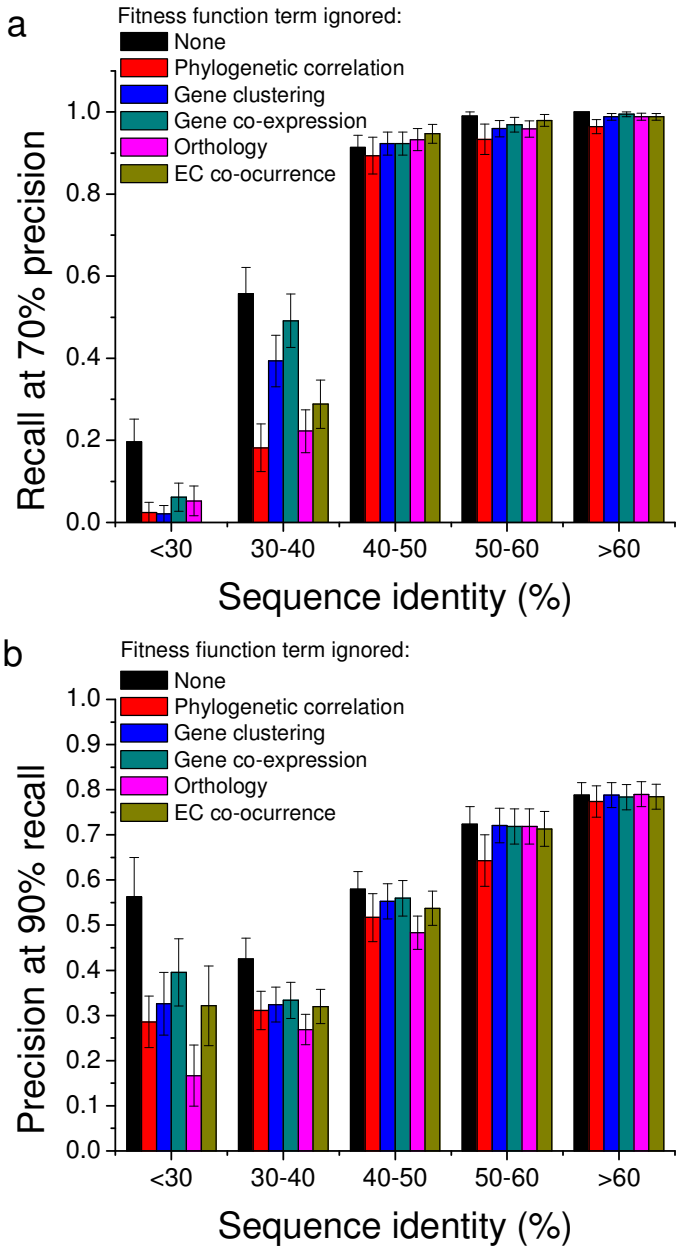
Supplementary Figure 2 | Obtaining GLOBUS parameters using different models. (a) Maximum likelihood values of the context weight coefficients derived using the iLL672[2] and iMM904[4] *S. cerevisiae* models and the iBsu1103[5] model for *B. subtilis*. SEQ: sequence identity; PC: phylogenetic correlation; GC: gene clustering; EX: co-expression; ORT: Orthology; EC: EC co-occurrence; OUT: not in the network **(b)** The correlation of probabilities with values higher than 0.1 in *S. aureus* based on GLOBUS parameters obtained by training with the two different yeast models (Pearson's r = 0.94, median probability difference = 0.04, maximum probability difference = 0.33). **(c)** The correlation of probabilities with values higher than 0.1 in *S. aureus* based on GLOBUS parameters obtained by training with the yeast iLL672 and the iBsu1103 metabolic models (Pearson's r = 0.96, median probability difference = 0.05, maximum probability difference = 0.35).



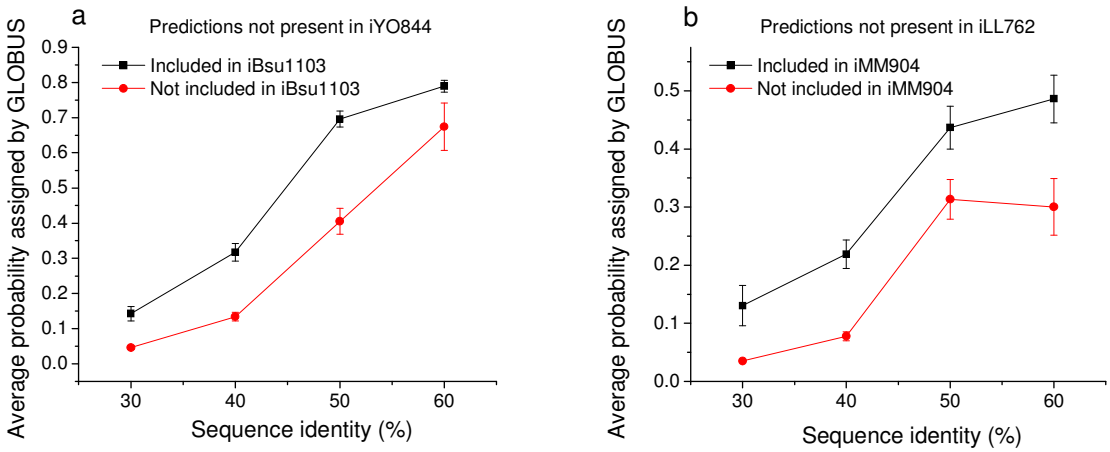
Supplementary Figure 3 | Comparison of the precision-recall relationships obtained using homology information established by BLAST and PSI-BLAST. Using PSI-BLAST, instead of regular BLAST, does not improve the performance significantly because additional sequences with low identity (detected by PSI-BLAST) only rarely have the target function.



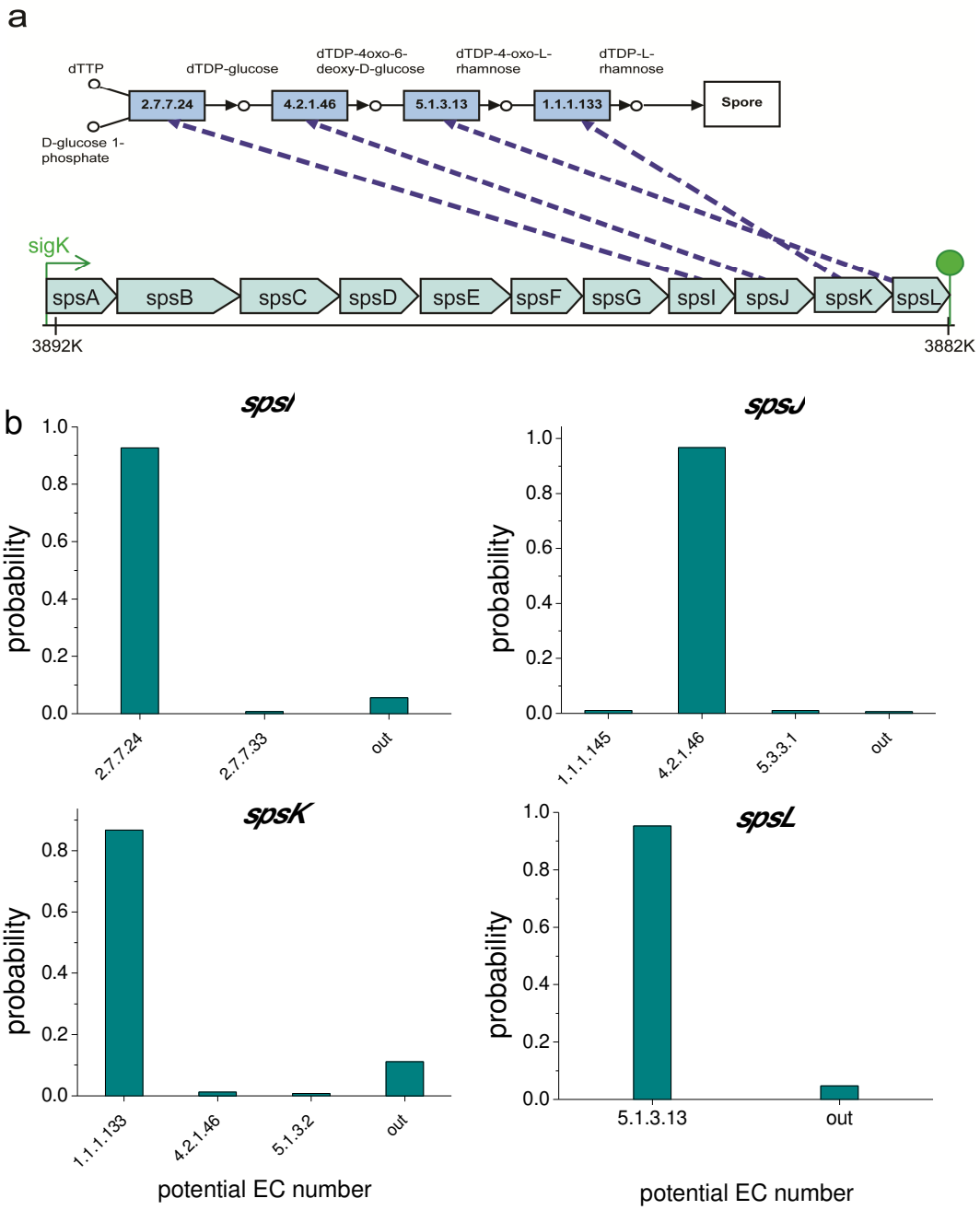
Supplementary Figure 4 | Distribution of sequence identities. Fractions of potential metabolic genes in *S. cerevisiae* (green) and *B. subtilis* (grey) are shown as a function of sequence identity to annotated enzymes in other species. Over half of potential metabolic genes have relatively small sequence identity (<40%) to known enzymes in both model organisms.



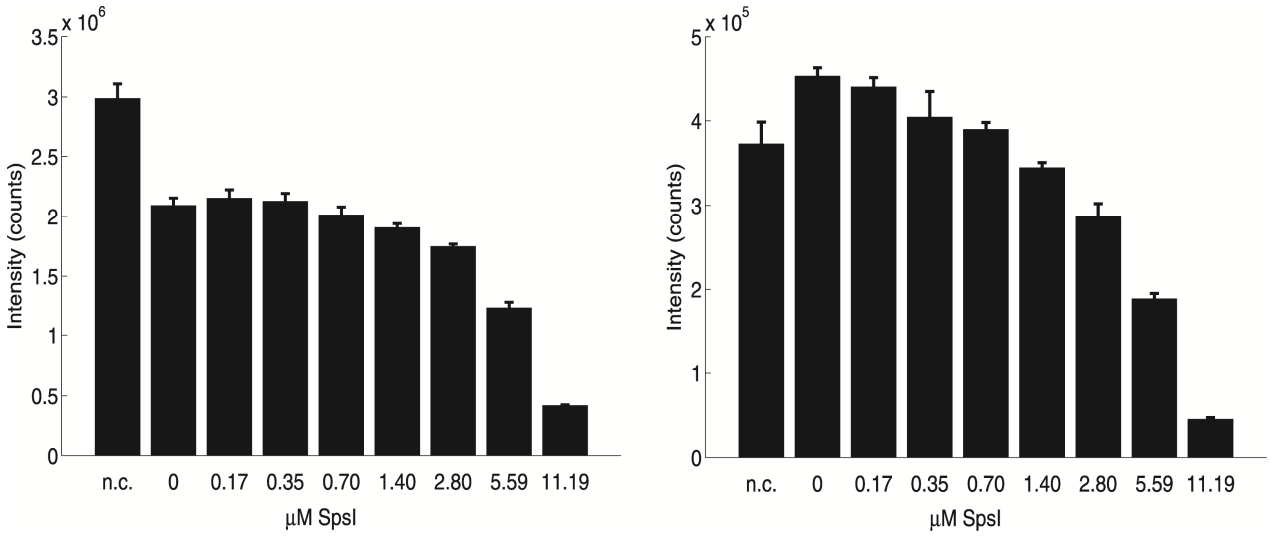
Supplementary Figure 5 | Contribution of individual context correlations to the GLOBUS performance. Different columns in the figure represent precision/recall values - across sequence identity bins - achieved by GLOBUS without using individual context correlations. The corresponding GLOBUS parameters were determined by simulated annealing optimizations performed without using each of the context correlations. The results show that at the same level of precision (70%) **(a)** and recall (90%) **(b)**, there is a marked reduction in performance compared to the results using the full fitness function; such reduction is most apparent for cases with low sequence identity. Error bars represent the s.e.m.



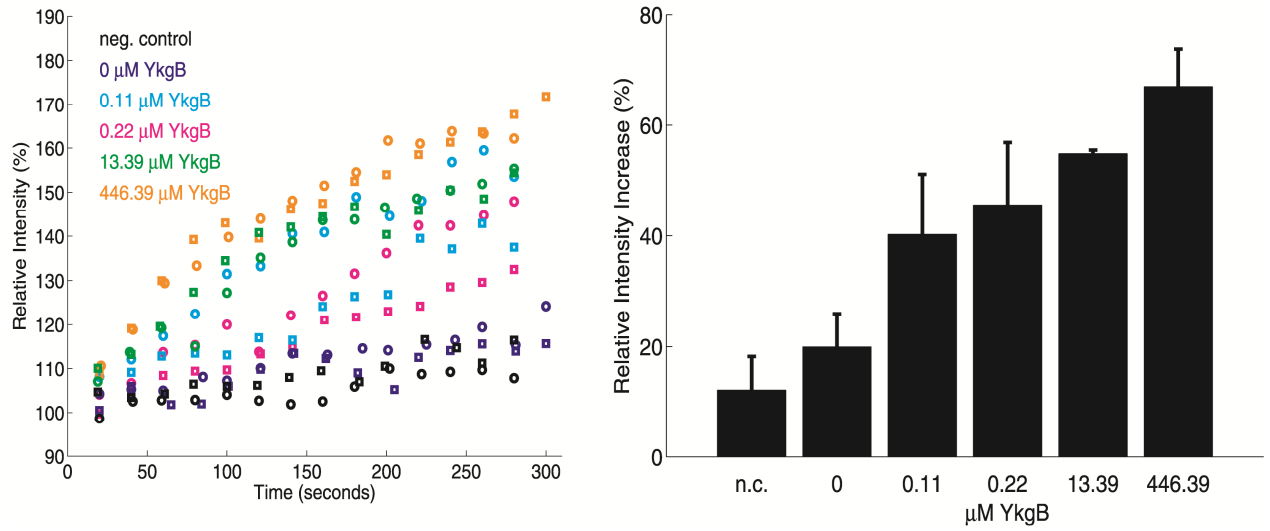
Supplementary Figure 6 | Potential utility of GLOBUS in refining manually curated metabolic models. (a) Annotations with non-zero GLOBUS probabilities that were not included in the older iYO844[6] *B. subtilis* model were subdivided into those included (black) and those that were not included (red) in the newer iBsu1103 model. Results show that, for different bins of sequence identity, higher GLOBUS probabilities correspond to higher likelihoods that annotations were included in the newer model. **(b)** A similar result is observed for yeast. Annotations with non-zero GLOBUS probabilities that were not included in the older iLL672[2] *S. cerevisiae* model were subdivided into those included (black) or not included (red) in the updated iMM904[4] model. Error bars represent the s.e.m.



Supplementary Figure 7 | GLOBUS predictions for *sps* genes in *B. subtilis*. **(a)** Genomic positions of the *sps* genes, as well as gene mapping (dashed arrows) to the dTDP-L-rhamnose biosynthesis pathway. The expression of *sps* genes is controlled by the σ^K transcription factor[7]. **(b)** GLOBUS-derived probabilities for potential functions of *spsI*, *spsJ*, *spsK*, and *spsL*.



Supplementary Figure 8 | Substrate consumption at different spsI concentrations. Mass spectrometry intensities of α-D-glucose-1-phosphate (left panel) and dTTP (right panel) are shown as a function of the SpsI concentration. As negative control (n.c.), the protein free filtrate of 6.99 µM spsI solution was used. Error bars represent standard deviations from two independent assays.



Supplementary Figure 9 | Product accumulation at different YkgB concentrations. Mass spectrometry intensities of 6-Phosphogluconic acid (left panel) and relative intensity increase (right panel) comparing final to initial values are shown as a function of the YkgB concentration. As negative control (n.c.), the protein free filtrate of 232 µM YkgB solution was used. Error bars represent standard deviations from two independent assays.

Supplementary Table 1 | Prediction of gene function in *S. aureus*.

Gene	EC number	Enzyme name	Probability	Identity (%)	Average Context Z-score
bioD	6.3.3.3	dethiobiotin synthase	0.99	31.2	7.9
hisG	2.4.2.17	ATP phosphoribosyltransferase	0.99	39.6	6.3
murE	6.3.2.13	UDP-N-acetylmuramoyl-L-alanyl-D-glutamate-2,6-diaminopimelate ligase	0.96	39	7.0
thrB	2.7.1.39	homoserine kinase	1.00	42.4	7.0
mvaA	1.1.1.34	hydroxymethylglutaryl-CoA reductase (NADPH)	0.95	40.1	5.9
hemD	4.2.1.75	uroporphyrinogen-III synthase	0.76	27.4	6.4
SA2374	1.3.3.1	dihydroorotate oxidase	0.84	37.8	2.2
murF	6.3.2.10	UDP-N-acetylmuramoyl-tripeptide-D-alanyl-D-alanine ligase	1.00	46	7.2
mvaK1	2.7.1.36	mevalonate kinase	0.73	35.1	6.7
ribB	2.5.1.9	riboflavin synthase	0.91	43.3	8.3
ribC	2.7.1.26	riboflavin kinase	0.96	45.5	2.3
lysC	2.7.2.4	aspartate kinase	0.86	41.6	5.4
scrB	3.2.1.26	beta-fructofuranosidase	0.82	40.5	5.9
folA	1.5.1.3	dihydrofolate reductase	0.89	42.8	9.6
SA1288	6.3.4.15	biotin-[acetyl-CoA-carboxylase] ligase	0.56	33.1	2.0
aroK	2.7.1.71	shikimate kinase	0.66	34.9	2.0
coaW	2.7.1.33	pantothenate kinase	0.71	36.6	2.1
nagA	3.5.1.25	N-acetylglucosamine-6-phosphate deacetylase	0.95	45.5	8.1
ansA	3.5.1.1	asparaginase	0.66	36	2.2
SA2317	4.3.1.17	L-Serine ammonia-lyase	0.90	43.9	2.8
bioA	2.6.1.62	adenosylmethionine-8-amino-7-oxononanoate transaminase	0.97	48.2	9.3
asd	1.2.1.11	aspartate-semialdehyde dehydrogenase	0.98	48.9	5.8
gcvPB	1.4.4.2	glycine dehydrogenase (decarboxylating)	0.81	42.3	9.2
thiD	2.7.4.7	phosphomethylpyrimidine kinase	0.89	44	8.9
hemY	1.3.3.4	protoporphyrinogen oxidase	0.94	47	4.1
trpG	4.1.3.27	anthranilate synthase	0.68	40.6	6.6
SA2006	4.1.1.5	acetolactate decarboxylase	0.90	46.6	4.9
alr1	5.1.1.1	alanine racemase	0.82	43.6	3.3
SA0511	1.1.1.103	L-threonine 3-dehydrogenase	0.78	43.5	2.2
SA2318	4.3.1.17	L-Serine ammonia-lyase	0.83	45.6	9.3

In the table we show predictions without experimental validation that have GLOBUS-assigned probabilities above 0.5 and protein sequence identity to known enzymes below 50%. The annotations in the table are ordered by averaging the prediction ranks sorted by decreasing annotation probability and the prediction ranks sorted by decreasing identity distance to known enzymes. The last column shows the average Z-score of phylogenetic correlations, gene clustering, and gene co-expression when all sequences are assigned to their most probable locations.

Supplementary Methods

Supplementary Table 2 | Highly connected metabolites that were not used in establishing connections between metabolic activities (EC numbers).

Metabolite	Number of connected EC numbers
H_2O	1224
H+	703
NADP+	435
NADPH	433
NAD+	422
NADH	412
Oxygen	379
ATP	375
Orthophosphate	306
ADP	294
CO_2	254
CoA	230
Pyrophosphate	185
NH_3	183
UDP	150
S-Adenosyl-L-methionine	115
Reduced acceptor	115
AMP	111
Pyruvate	109
S-Adenosyl-L-homocysteine	107
Acetyl-CoA	103
H_2O_2	102
L-Glutamate	100
2-Oxoglutarate	96
UDP-glucose	76
Acetate	73
D-Glucose	56
Carboxylate	48
Succinate	43
Oxaloacetate	41
Glycine	41

1. The fitness (energy) function

The fitness (energy) function over all metabolic genes, $E(g_1, g_2, \ldots, g_n)$, was defined to reflect the hypotheses that a particular global assignment of genes into their network locations will be more probable if genes have significant homologies to the assigned locations, and also exhibit strong context correlations with their network neighbors. Accordingly, in GLOBUS calculations we used the following fitness function for genes included in the network:

$$E(g_1, g_2, \ldots, g_n) = -b_{\text{homology}} f_{\text{homology}} - b_{\text{orthology}} f_{\text{orthology}} - b_{\text{context}} f_{\text{context}} - b_{\text{ECco-occurrences}} f_{\text{ECco-occurrences}}$$

where, b(s) are positive coefficients representing weights of each functional feature, and f(s) are various functional features described below.

A. Sequence homology

The term, f_{homology}, represents the descriptor of sequence homology. The higher the sequence identity between a protein and enzymes in other species known to catalyze the assigned activity, the more likely is the assignment to be correct[8,9]. As the sequence homology descriptor we used the logarithm of the conditional probability that the gene performs the assigned function, given the highest sequence identity to a Swiss-Prot[3] protein annotated to catalyze the target activity:

$$f_{homology} = \sum_{i=1}^{n} \log P(\text{gene performs target function} | \text{highest sequence identity to annotated SwissProt protein})$$

We only considered Swiss-Prot sequences with protein-based BLAST[10] E-values < $5*10^{-2}$ to the target. We also excluded from consideration Swiss-Prot proteins that were: 1) from the query or closely related genomes (from species in the same taxonomic genus) or 2) likely to be annotated based exclusively on computational methods, i.e., genes with annotation keywords such as *probable*, *like*, *by similarity*, *hypothetical*, or *putative.* The conditional probabilities were estimated using the well-curated yeast iLL672 metabolic model[2] (Supplementary Fig. 1d). If non-overlapping regions of a considered gene had homologies to separate Swiss-Prot sequences performing different enzyme activities, these regions were treated independently and assigned to different locations of the EC network using f_{homology} as defined above.

B. Gene orthology

An additional binary descriptor related to sequence homology was the possible gene orthology to a gene from another species annotated with the target activity. The orthology descriptor was based on bi-directional best hits by protein BLAST; in these calculations we used the bi-directional best hits in the KEGG SSDB database[11] (http://www.genome.jp/kegg/ssdb/). For each gene, the orthology term was either 1, if at least one possible ortholog was annotated in Swiss-Prot to perform the target activity, or 0, if no orthologs with the target activity could be identified. Again we excluded annotations based exclusively on computational methods, and treated separately non-overlapping regions with homology to different activities (see above).

C. Genomic context correlations

Gene pairs that share similar biological functions tend to be either present or absent together in genomes of sequenced species (phylogenetic profiles), tend to be co-localized on

chromosomes across multiple genomes (gene clustering), and tend to be co-regulated. These context-based correlations were initially developed to infer gene functions and provide complementary information to sequence homology data[12,13]. Multiple studies have also demonstrated that genes located close to each other in a metabolic network tend to have significantly stronger context associations[14,15]. Previously, we and others used context associations in combination with local structure of the metabolic network to identify genes responsible for orphan metabolic activities[16-19].

In GLOBUS we used the context correlations by first transforming them into Z-scores[1] using the distribution of correlations between all pairs of candidate metabolic genes, and then estimating the conditional probability that two genes are direct network neighbors given their context association Z-score. The conditional probabilities were derived based on the iLL672 yeast metabolic model (Supplementary Fig. 1a-c). In the GLOBUS fitness function for each assigned gene we considered the maximum log probability among all network neighbors of the gene:

$$f_{context} = \sum_{i=1}^{n} max(\log P(\text{two genes are network neighbors}|\text{context correlation Z-score between the genes}))$$

C.1 Phylogenetic correlation

Phylogenetic correlation[20,21] measures the co-occurrence (co-presence) of homologues for a pair of genes across genomes. Phylogenetic profiles were constructed using protein BLAST searches against a collection of 70 diverged genomes[16]. We used the binary phylogenetic profiles, i.e. 70-dimensional binary vectors representing the presence or absence of homologues in the target genomes. Pearson's correlation between the profile vectors was calculated using the following equation:

$$r = \frac{Nz - xy}{\sqrt{(Nx - x^2)(Ny - y^2)}}$$

, where N is the total number of target genomes. For genes X and Y, x is the number of genomes in which any homologue of X is present, y is the number of genomes in which any homologue of Y is present, and z in the number of genomes in which homologues of both X and Y are present.

C.2 Gene chromosomal clustering

For a pair of genes, chromosomal gene clustering[22-24] measures the degree of co-localization of their orthologues across a set of genomes. We considered gene order statistics instead of the exact nucleotide positions of genes, i.e. we defined a gene order distance d(X,Y) as the minimum number of genes separating genes X and Y. Under the null hypothesis that genes are distributed randomly within a genome, $P(d_\gamma(\text{X, Y}))$ is the probability of observing gene order distance equal or less than $d_\gamma(\text{X, Y})$ between a pair of genes X and Y in a genome γ. $P(d_\gamma(\text{X, Y}))$ can be calculated directly as the fraction of gene pairs in genome γ that are separated by gene order distance $d_\gamma(\text{X, Y})$ or smaller. Assuming gene order distances are independent across a set of 108 evolutionary divergent organisms Γ, and given that X and Y are orthologues of genes A and B from the target genome, we calculated the clustering of genes A and B:

$$C_\Gamma(A,B) = -log \prod_{\gamma \in \Gamma} \left(P\left(d_\gamma(X,Y)_\gamma\right)\right)$$

For a given set of genomes, this clustering measure can be biased by the variable phylogenetic proximity between different organisms. Therefore we deliberately filtered the genome set to eliminate species closely related to the target genome using a mutual information threshold of 0.9 for ortholog occurrences[17]. Orthology mapping required for the chromosomal clustering calculations was established using best bi-directional hits in the KEGG SSDB database[11] (http://www.genome.jp/kegg/ssdb/).

C.3 Co-expression

Numerous studies[25,26] demonstrated that genes with similar mRNA expression profiles usually have related biological functions. Descriptors of mRNA co-expression used in GLOBUS were calculated as Spearman's rank correlation between expression profiles obtained from the Rosetta "compendium" dataset[27] for *S. cerevisiae* and the GEO database[28] for *B. subtilis and S. aureus*. In all calculations Log_{10} intensity ratio values were used.

D. Co-occurrence of metabolic activities (EC number) across species

In addition to phylogenetic gene profiles, we used in GLOBUS a functional descriptor based on likely co-occurrence between different metabolic activities (EC numbers) across species. This descriptor measures the correlation between the occurrence vectors for different activities across a set of organisms, without considering genes assigned to the activities. To calculate the correlation between different metabolic activities (EC numbers) we used a 70-dimentional binary vector for each EC number representing its presence or absence in a set of 70 genomes (see section C.1) according to the KEGG database[11] (http://www.genome.jp/kegg). For every pair of EC numbers the Pearson's correlation between their profile vectors was calculated (see section C.1).

In the GLOBUS fitness function for each assigned gene we considered the EC co-occurrence descriptor equal to the average correlation between the EC activity of the assigned gene and the EC activities for all its network neighbors.

2. Calculating marginal probabilities using Gibbs sampler

The marginal probability, $P(g_i)$, represents the probability that a gene is responsible for a metabolic activity (EC number) consistent with all possible assignments of other genes into the network. Formally, given all parameters of the GLOBUS fitness function, b_{homology}, $b_{\text{orthology}}$, b_{context}, and $b_{\text{EC co-occurrence}}$, $P(g_i)$ can be calculated by summation:

$$P(g_i) = \sum_{g_1} \ldots \sum_{g_{i-1}} \sum_{g_{i+1}} \ldots \sum_{g_n} \frac{1}{Z} EXP\{-E(g_1,\ldots,g_{i-1},g_i,g_{i+1},\ldots,g_n)\}$$

, where Z is normalizing partition function. Suppose that there are n metabolic genes in the genome and each metabolic gene has m potential network assignments, obtaining $P(g_i)$ then requires summing over m^n possible terms. Because a typical genome contains many hundreds to thousands of metabolic genes, this summation is computationally intractable. Nevertheless, the success of the GLOBUS approach is due to the fact that the vast majority of all possible gene

assignments have very low probabilities. Consequently, it is possible to recover correct marginal probabilities for each gene using an efficient sampling of high probability configurations (assignments).

To sample probable gene assignments we applied a widely used algorithm, the Gibbs sampler[29-31]. The Gibbs sampler is a special case of Markov Chain Monte Carlo (MCMC) and the Metropolis-Hasting algorithm[32,33]. The Gibbs sampler allows obtaining marginal probabilities using sampling based on conditional probabilities. Starting with a random initial assignment of n metabolic genes to a network, a Gibbs chain of t steps: $G^1,G^2,\ldots,G^t$ is obtained iteratively by selecting a random gene i and re-assigning to a location g_i according to:

$$G_i^{k+1} \sim P(g_i \mid g_1 = G_1^k, \ldots, g_{i-1} = G_{i-1}^k, g_{i+1} = G_{i+1}^k, g_n = G_n^k)$$

Where G_i^k represents the location of gene i at step k of the Gibbs chain G. If at each iteration the location of every gene was recorded; it can be proven that the distribution of G_i converges to $P(g_i)$ as the number of iterations $t \to \infty$.

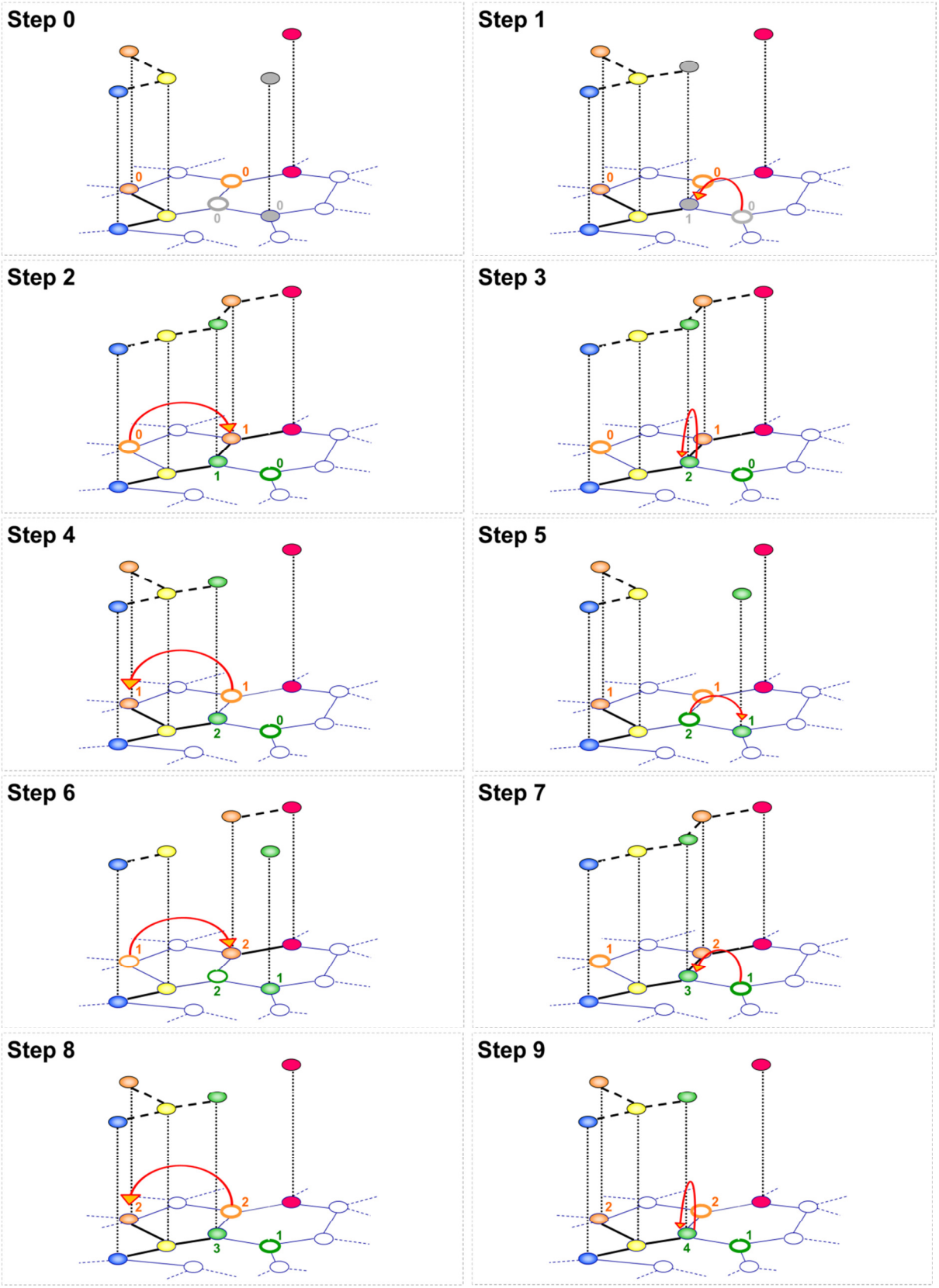
The conditional probability used in the iterative sampling, $P(g_i \mid g_1,\ldots,g_{i-1}, g_{i+1},\ldots,g_n)$, is:

$$P(g_i \mid g_1,\ldots,g_{i-1},g_{i+1},\ldots,g_n) = \frac{P(g_1,\ldots,g_{i-1},g_i,g_{i+1},\ldots g_n)}{P(g_1,\ldots,g_{i-1},g_{i+1},\ldots,g_n)}$$

Since in each iteration the denominator of the equation and Z (the partition function) are constant, the conditional probability can be derived from the fitness function, $E(g_1, g_2, \ldots, g_n)$:

$$\begin{aligned} P(g_i \mid g_1,\ldots,g_{i-1},g_{i+1},\ldots,g_n) &\propto P(g_1,\ldots,g_{i-1},g_i,g_{i+1},\ldots,g_n) \\ &\propto EXP\{-E(g_1,\ldots,g_{i-1},g_i,g_{i+1},\ldots,g_n)\} \end{aligned}$$

A schematic illustration of a Gibbs sampler chain generating iterative gene assignments is shown in Supplementary Figure 10.


Step 0
Step 1
Step 2
Step 3
Step 4
Step 5
Step 6
Step 7
Step 8
Step 9

Supplementary Figure 10 | Illustration of Gibbs sampler used to derive marginal probabilities, $P(g_i)$, in GLOBUS. A marginal probability for a gene assignment reflects the likelihood of the gene to catalyze the corresponding activity consistent with (i.e. integrated or summed over) all possible assignment of other genes in the network. Starting from a random assignment of candidate metabolic genes, every gene is then iteratively re-assigned to one of its possible locations; the likelihood for re-assignment at each location is proportional to the global fitness function with the gene assigned to the corresponding EC number, given the current state for the rest of the network. Each time a gene is re-assigned a record is kept, such that probabilities can be calculated directly from the Gibbs chain counts.

3. Cloning, purification and protein identification of SpsI, SpsJ and YkgB.

The sequences of *B. subtilis* genes *spsI*, *spsJ* and *ykgB* were retrieved from SubtiList[34] and synthesized by Geneart (http://www.invitrogen.com) with codon usage specifically optimized for *E. coli*. The genes were then amplified from the plasmids provided by Geneart by attaching 6x His-Tag to the C-terminal end using following primer pairs:

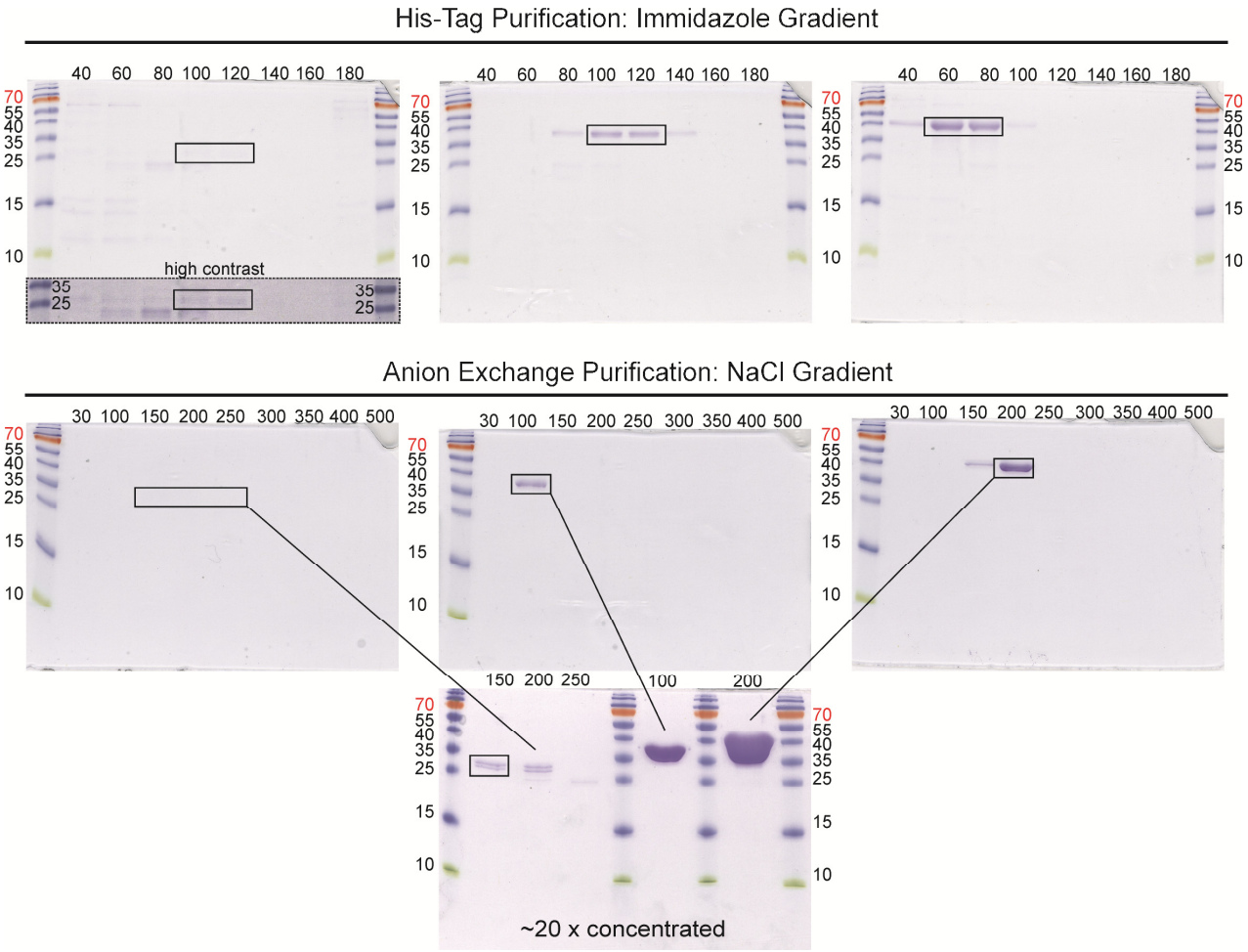
Gene	Strand	Primer sequence (5'->3')
spsI	Forward	ATCCGCTCTAGAATGAAAGGTGTTATTCTGGCAGGCGG
	Reverse	CATGATAAGCTTTTAATGATGATGATGATGATGTTTTTCATCCTGACCTTTACG
spsJ	Forward	GCGCGTCTAGAATGGCAAAAAGCTATCTGATTACCGGTGG
	Reverse	ATATGAAGCTTTTAATGATGATGATGATGATGACGATCATTATCGGTATACCACTGAATGG
ykgB	Forward	GCCGCCTCTAGAATGACCAAATATATTGGTTATGTGGGCACC
	Reverse	ATTATTAAGCTTTTAATGATGATGATGATGATGCACCTGATGCAGAAATTTAACACAAACC

PCR products were digested by XbaI/HindIII and ligated into IPTG-inducible pTrc99α protein expression vector[35].

The genes were overexpressed in *E. coli* BL21 by induction with 0.1 mM IPTG. After growing for 16 hours in 200 ml LB medium at 25°C and 250 rpm, cells were harvested by centrifugation, washed with 0.9% NaCl and 10 mM $MgSO_4$, resuspended in 20 mM sodium phosphate buffer (pH 7.4) with 2 mM DTT, 4 mM PMSF, 0.5 M NaCl, 20 mM imidazole, 1% Triton X-100, 0.2 mg/ml lysozyme, 20 µg/ml DNAse, and lysed by freeze-thaw cycles. Cell debris was separated from lysate by centrifugation at 15'000 rpm and 4°C for 30 minutes, and clear cell extract was loaded onto His Gravi-Trap columns (GE Healthcare Life Sciences). Fractions from imidazole gradient elution containing desired proteins were identified by SDS-PAGE (Supplementary Fig. 11), elution buffer was replaced by 20 mM potassium phosphate buffer (pH 7.4) with 30 mM NaCl using filter columns with 10 kD cutoff (Millipore) and fractions were subsequently loaded onto a 1.5 ml Q Sepharose High Performance anion exchange column (Amersham Biosciences Limited). Fractions from NaCl gradient elution containing desired proteins were identified by SDS-PAGE (Supplementary Fig. 11) and elution buffer was replaced by 50 mM potassium phosphate buffer (pH 7.4) with 2.5 mM MgCl2 using filter columns with 10 kD cutoff (Millipore). Protein concentrations of selected fractions were determined in 20 x concentrated samples by Bradford assay for SpsJ and YkgB. The low abundant SpsI was below detection limit and therefore its concentration was estimated based on comparison of the band intensities on the SDS-PAGE gels.

The correct identity of the protein-bands in the used fractions was confirmed by tryptic in-gel digestion as described by a published protocol[36]. The resulting peptide mixes were purified using Ultra Micro Spin Columns (C18, The Nest Group Inc.) and subsequently analyzed on a

LTQ-Orbitrap XL instrument (Thermo Fisher) using published settings[37]. Peptides were assigned to MS/MS spectra by SEQUEST[38] search and subsequent protein assignments were validated by PeptideProphet[39] with 1% error rate cutoff (Supplementary Dataset 1).



Supplementary Figure 11 | SDS-PAGE. Gels show purification of the His-tagged SpsI, SpsJ and YkgB by a imidazole gradient on a Ni Sepharose column followed by a NaCl gradient on a Q Sepharose column. Numbers above lanes indicate increasing imidazole or NaCL concentrations (mM). **Left panel:** Purified SpsI with predicted molecular mass, including His-Tag, of 28.6 kDa. Imidazole elution fractions of 100 and 120 mM were selected for subsequent anion exchange purification. NaCl elution fractions of 150, 200 and 250 mM were identified by eye due to the very low amount (not visible by gel scanner pictures, well visible after concentration). The 150 mM elution fraction was finally used for further experiments. **Middle panel:** Purified SpsJ with predicted molecular mass, including His-Tag, of 36.4 kDa. Imidazole elution fractions of 100 and 120 mM were selected for subsequent anion exchange purification. The 100 mM elution fraction was used for further experiments. **Right panel:** Purified YkgB with predicted molecular mass, including His-Tag, of 39.2 kDa, imidazole elution fractions of 60 and 80 mM were selected for subsequent anion exchange purification. The 100 mM NaCl elution fraction was used for further experiments.

References

1. Faith, J.J. et al. Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol.* **5**, e8 (2007).
2. Kuepfer, L., Sauer, U. & Blank, L.M. Metabolic functions of duplicate genes in Saccharomyces cerevisiae. *Genome Res.* **15**, 1421-30 (2005).
3. Boeckmann, B. et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**, 365-70 (2003).
4. Mo, M.L., Palsson, B.O. & Herrgard, M.J. Connecting extracellular metabolomic measurements to intracellular flux states in yeast. *BMC Syst. Biol.* **3**, 37 (2009).
5. Henry, C.S., Zinner, J.F., Cohoon, M.P. & Stevens, R.L. iBsu1103: a new genome-scale metabolic model of Bacillus subtilis based on SEED annotations. *Genome Biol.* **10**, R69 (2009).
6. Oh, Y.K., Palsson, B.O., Park, S.M., Schilling, C.H. & Mahadevan, R. Genome-scale reconstruction of metabolic network in Bacillus subtilis based on high-throughput phenotyping and gene essentiality data. *J. Biol. Chem.* **282**, 28791-9 (2007).
7. Eichenberger, P. et al. The program of gene transcription for a single differentiating cell type during sporulation in Bacillus subtilis. *PLoS Biol.* **2**, e328 (2004).
8. Tian, W. & Skolnick, J. How well is enzyme function conserved as a function of pairwise sequence identity? *J. Mol. Biol.* **333**, 863-82 (2003).
9. Rost, B. Enzyme function less conserved than anticipated. *J. Mol. Biol.* **318**, 595-608 (2002).
10. Altschul, S.F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-402 (1997).
11. Kanehisa, M. et al. KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* **36**, D480-4 (2008).
12. Eisenberg, D., Marcotte, E.M., Xenarios, I. & Yeates, T.O. Protein function in the post-genomic era. *Nature* **405**, 823-6 (2000).
13. Bork, P. et al. Protein interaction networks from yeast to human. *Curr. Opin. Struct. Biol.* **14**, 292-9 (2004).
14. Kharchenko, P., Church, G.M. & Vitkup, D. Expression dynamics of a cellular metabolic network. *Mol. Syst. Biol.* **1**, 2005 0016 (2005).
15. Spirin, V., Gelfand, M.S., Mironov, A.A. & Mirny, L.A. A metabolic network in the evolutionary context: multiscale structure and modularity. *Proc. Natl. Acad. Sci. USA* **103**, 8774-9 (2006).
16. Chen, L. & Vitkup, D. Predicting genes for orphan metabolic activities using phylogenetic profiles. *Genome Biol.* **7**, R17 (2006).
17. Kharchenko, P., Chen, L., Freund, Y., Vitkup, D. & Church, G.M. Identifying metabolic enzymes with multiple types of association evidence. *BMC Bioinformatics* **7**, 177 (2006).
18. Kharchenko, P., Vitkup, D. & Church, G.M. Filling gaps in a metabolic network using expression information. *Bioinformatics* **20 Suppl 1**, i178-85 (2004).
19. Green, M.L. & Karp, P.D. A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Bioinformatics* **5**, 76 (2004).
20. Huynen, M.A. & Bork, P. Measuring genome evolution. *Proc. Natl. Acad. Sci. U.SA* **95**, 5849-56 (1998).

21. Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D. & Yeates, T.O. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. USA* **96**, 4285-8 (1999).
22. Dandekar, T., Snel, B., Huynen, M. & Bork, P. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.* **23**, 324-8 (1998).
23. Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D. & Maltsev, N. The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. USA* **96**, 2896-901 (1999).
24. Lee, J.M. & Sonnhammer, E.L. Genomic gene clustering analysis of pathways in eukaryotes. *Genome Res.* **13**, 875-82 (2003).
25. DeRisi, J.L., Iyer, V.R. & Brown, P.O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680-6 (1997).
26. Wu, L.F. et al. Large-scale prediction of Saccharomyces cerevisiae gene function using overlapping transcriptional clusters. *Nat. Genet.* **31**, 255-65 (2002).
27. Hughes, T.R. et al. Functional discovery via a compendium of expression profiles. *Cell* **102**, 109-26 (2000).
28. Barrett, T. et al. NCBI GEO: mining millions of expression profiles--database and tools. *Nucleic Acids Res.* **33**, D562-6 (2005).
29. Geman, S. & Geman, D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE T. Pattern Anal.* **6**, 721-741 (1984).
30. Gelfand, A.E. & Smith, A.F.M. Sampling-based approches to calculating marginal densities. *J. Am. Stat. Assoc.* **85**, 398-409 (1990).
31. Casella, G. & George, E.I. Explaining the Gibbs sampler. *Am. Stat.* **46**, 167-174 (1992).
32. Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., H., T.A. & E., T. Equations of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087-1091 (1953).
33. Hastings, W.K. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97-109 (1970).
34. Moszer, I., Jones, L.M., Moreira, S., Fabry, C. & Danchin, A. SubtiList: the reference database for the Bacillus subtilis genome. *Nucleic Acids Res.* **30**, 62-5 (2002).
35. Amann, E., Ochs, B. & Abel, K.J. Tightly regulated tac promoter vectors useful for the expression of unfused and fused proteins in Escherichia coli. *Gene* **69**, 301-15 (1988).
36. Shevchenko, A., Tomas, H., Havlis, J., Olsen, J.V. & Mann, M. In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nat. Protoc.* **1**, 2856-60 (2006).
37. Leitner, A. et al. Expanding the chemical cross-linking toolbox by the use of multiple proteases and enrichment by size exclusion chromatography. *Mol. Cell Proteomics.* **11**, M111 014126 (2012).
38. Yates, J.R., 3rd, Eng, J.K., McCormack, A.L. & Schieltz, D. Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal. Chem.* **67**, 1426-36 (1995).
39. Keller, A., Nesvizhskii, A.I., Kolker, E. & Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **74**, 5383-92 (2002).