Supporting Information

Borenstein *et al.* 10.1073/pnas.0806162105

SI Text

Sensitivity of Seed Set Identification to Missing or Erroneous Data. The effect of missing or erroneous metabolic data on the composition of the obtained seed set was examined by simulations. We used the metabolic network of *Saccharomyces cerevisiae*, a species whose metabolism has been extensively studied, as a reference "complete" network. Using the metabolic network of *Escherichia coli*, another species for which metabolic information is extensive, produced practically identical results. We perturbed this network by either deleting existing reactions or altering the reactions' directions with varying probabilities. As metabolic map reconstruction that is based on as low as 50% gene coverage still detects >70% of the total number of reactions (1), we analyzed deletion probabilities of up to 30%. The seed sets obtained for the perturbed networks were then compared to the seed set of the original, unperturbed network, measuring the percentage of true positives (the percentage of the original seed set that was still identified) and false positives (the percentage of falsely detected seeds). As evident from Fig. S1*A*, the percentage of correctly identified seed compounds is only slightly lower than the percentage of reactions included in the incomplete network, which could be conceived as an upper bound on the percentage of identified seeds. Furthermore, the size of the obtained seed set in comparison to the size of the seed set of the original network (which can be calculated as the percentage of true positives plus the percentage of false positives) is almost constant for all missing data levels. When the directionality of numerous reactions is erroneously altered (rather than deleting the reactions altogether), the percentage of true positives is higher (i.e., a larger portion of the original seed set is correctly identified), but the percentage of falsely detected seed compounds also increases, causing a slight inflation in the size of the seed set (Fig. S1*B*). In summary, however, it seems that the seed set identification process does not amplify the level of noise or incompleteness in the metabolic data.

Notably, the observed robustness is probably, to some extent, linked to the robust structure of metabolic networks (2–5). Specifically, the existence of alternative pathways in the network—a major contributing factor to metabolic robustness in the face of gene knockouts—guarantees that in many cases a seed compound participates in more than one pathway, making its detection more robust to missing data in one of these pathways.

Strongly Connected Components Statistics. In accordance with previous studies (6, 7), for most species our SCC decomposition results in a bow-tie structure, where a large fraction of the compounds (36.3%, on average, in our analysis) constitutes a giant strongly connected component whereas other compounds are arranged in relatively small components (1.4 compounds per component on average) with short paths to or from the giant component (Fig. S5).

Source components (i.e., components with no incoming edges) that include more than one compound are of special interest. These components entail ambiguity in seed prediction because it cannot be determined, from network topology alone, which of the compounds included in such a component is a seed. Such ambiguities may hinder, for example, the accurate identification of necessary inputs and cofactors in autocatalytic cycles. In the analysis presented in this article, we address this issue by regarding all candidate compounds in these components as seeds. Yet it should be noted that such source components are relatively rare—the majority of source components in each species (>89% on average) are singletons, comprising only a single compound, and therefore render no ambiguity in the composition of the seed set (Fig. S7A). Less than 9% include two compounds (i.e., two interchangeable compounds), and <2.4% include three or more. Because larger source components may contribute, by definition, more compounds to the seed sets, we also calculated the percentage of the seed compounds that are part of larger source components. Again, we find that a high percentage of the seed compounds in each species form singleton components (>77%), and only 8.5% are part of source components with three compounds or more (Fig. S7*B*).

Environmental Attributes. The environmental properties included in our analysis are the ones provided by the NCBI genome project as descriptors of prokaryotes' preferred environments (*Materials and Methods*). Additional motivation for the use of these properties arises from the following considerations: Oxygen was shown not only to play a profound role in shaping the electron transfer "market place" (8) and the synthesis of transmembrane proteins (9), but also to allow a major transition in the evolution of biochemical–metabolic networks and complexity (10). Similarly, the variability of the biochemical environment, which is strongly related to the habitat property used in our analysis (e.g., host-associated, specialized, or terrestrial habitat), was shown to dramatically affect the structure of bacterial metabolic networks (11, 12). These two environmental properties (namely, oxygen requirement and habitat) are therefore expected to directly shape the topology of the network and may consequently affect the composition of the seed set, making them obvious candidates for the analysis. Temperature and salinity, on the other hand, are clearly major environmental features and are of great interest, but their potential effects on metabolic network topology are probably less direct. Hence, taken together, these four properties form an appropriate and well balanced set for studying the relation between seed sets' composition and global environmental features, examining both direct and indirect effects of the environment on metabolic networks.

2-Oxoglutarate Phyletic Pattern. The citric acid (TCA) cycle has been found to be incomplete and requires the exogenous acquisition of 2-oxoglutarate in several obligate intracellular species including *Chlamydia* (13, 14) and *Buchnera aphidicola* (15). Other obligate parasites, such as Mollicutes, demonstrate minimal metabolism and lack the TCA cycle altogether (16). The phyletic occurrence and seed patterns obtained by our analysis for 2-oxoglutarate are in full agreement with the above studies; 2-oxoglutarate appears as a seed in all Chlamydiae and Buchnera, is completely absent in all Mollicutes, and is an occurring compound in all other species (Fig. S8). A table describing the phyletic patterns of all of the compounds included our analysis and several calculated measures of the coherency of each pattern (*SI Materials and Methods*) can be found in Dataset S1 and Dataset S2.

Correlation Between the Probability of Being a Seed and Topological Characteristics. The ratio between the number of species (networks) in which a certain compound is a seed and the number of species in which it occurs, N_s/N_o, provides a measure for the probability that this compound is a seed and can be used to examine which properties of a compound make it more likely to be included in the seed set. Specifically, comparing this measure

and the topological characteristics of the compound within the global metabolic network [corresponding to the collective potential of the biosphere's meta-metabolome (10)], several significant correlations can be found. This probability is correlated with the reach of the compound (the number of other compounds to which it has a path) and its centrality (the average length of these paths) (0.59, $P < 10^{-300}$ and 0.52, $P < 10^{-300}$, respectively; Spearman rank correlation). Furthermore, filtering out some of the noise in the network data by considering only compounds that were never pruned during the network SCC decomposition (*Materials and Methods*), the probability of being a seed is correlated with the outgoing degree of the compound and inversely correlated with its incoming degree (Spearman rank correlation 0.36, $P < 10^{-5}$ and -0.47, $P < 10^{-9}$, respectively). These findings suggest that seed compounds tend to be those that are located on the periphery of the network (i.e., distant from the core metabolism) but are the precursors of many other compounds.

Compounds that have a high probability of being seeds ($N_s/N_o > 0.5$) tend to be associated with certain metabolic pathways (*SI Materials and Methods*) including fatty acid biosynthesis and aminoacyl-tRNA biosynthesis ($P < 0.05$ after multiple testing correction). The enrichment of the fatty acid biosynthesis pathway is in accordance with a recent study of metabolic network evolution, which found that the gain of many novel reactions occurred predominantly in the lipid metabolism pathways (17).

Principal Components Analysis of the Seed Sets Data. A principal components analysis (PCA) of the seed sets data was used to examine the distribution of the different taxa in the seed set space. The clear partition of the various taxonomic groups by the first two principal components suggests that the seed set composition is a good characteristic of each species (Fig. S9*A*). This partition is improved by correcting for the considerably larger number of bacterial taxa included in the analysis (Fig. S9*B*).

SI Materials and Methods

Metabolic Networks Reconstruction. A list of the main reactions in the database was retrieved from the file *reaction_mapformula.lst* in the KEGG LIGAND database. This file also lists for each reaction its definition (i.e., the substrates and product compounds) and directionality (if known) for each pathway it participates in. The chemical compounds are limited to main reactants. For each species, the list of reactions present in each pathway was retrieved from the *rn* files in the PATHWAY database. Using this reaction–pathway pairs list, along with the reactions' definition and directionality obtained above, the metabolic network of each species was reconstructed. The network is represented as a directed graph where nodes denote compounds and edges denote reactions. A directed edge from compound a to compound b indicates that compound a is a substrate in some reaction that produces compound b (i.e., for each given reaction, all of the nodes that represent its substrates are connected by directed edges to all of the nodes that represent its products). Glycans were omitted from the graph. Reversible reactions or reactions for which the directionality is unknown were represented as directed edges in both directions. We also recorded the number of reactions and compounds that appear in each species' network. For each compound, the metabolic pathways in which it participates was also retrieved from the *compound* file in the LIGAND database.

Because of inherent noise and incomplete reaction data, the reconstructed networks contain a large number of small disconnected components (i.e., groups of nodes that are not connected to the main part of the network) that may markedly interfere with the detection of meaningful seed compounds. Any such component, containing 10 compounds or fewer, was dropped from the network before the rest of the analysis was performed. We refer to the compounds included in these dropped components as *pruned* compounds and in the analysis regard them as compounds whose seed status is unknown.

Strongly Connected Components Decomposition. Given a network G, the strongly connected components (SCC) decomposition is performed by Kosaraju's algorithm (18), which works as follows:

(i) Run a Depth-First Search (DFS) on G (19) to compute finishing times $f[v]$ for each node v.
(ii) Calculate the transposed network G' (the network G with the direction of every edge reversed).
(iii) Run DFS on G', traversing the nodes in decreasing order of $f[v]$.

Each tree in the DFS forest created by the second DFS run forms a separate SCC.

Phyletic Occurrence Patterns and Phyletic Seed Patterns. A phyletic pattern represents the presence and absence pattern of a specific trait across the species analyzed. For example, considering sets of orthologous genes (20, 21), the phyletic pattern of a certain gene can be conceived as a Boolean vector, indicating the set of species in which an ortholog can be found. In the context of our article, phyletic patterns are associated with each compound (Fig. S2): The phyletic occurrence pattern of each compound is a binary vector, indicating in which species this compound occurs. Similarly, the phyletic seed pattern of each compound is a binary vector, indicating the species in which this compound is a seed (Fig. S2). Considering the phyletic patterns of a specific compound of interest allows us to examine its state (e.g., seed vs. non-seed) across the extant species and to trace the state of the compound in the internal nodes of the tree (representing ancestral species) using maximum parsimony or maximum likelihood approaches.

Detecting Coherent Phyletic Patterns. We wish to detect compound or seed patterns that correspond to major environmental changes or shifts in the metabolism of living organisms. The phyletic patterns of these compounds should both demonstrate high consistency with the phylogenetic tree topology and induce a meaningful partition of the species into those in which the compound/seed is present and those in which it is absent. However, considering the noisy nature of the data (stemming from the inherent noise involved in the classification and annotation of orthologous genes) and, specifically, the potential effect of this noise on the resulting occurrence and seed phyletic patterns, a simple parsimony analysis may be misleading. To detect these patterns we thus used a novel method based on information gain.

Formally, assume that in a given phyletic pattern covering L species, P of the species have a 1 ("present") state, and Q have a 0 ("absent") state. Let $p = P/L$ and $q = Q/L$ denote the relative frequencies. The entropy of this pattern is then given by $H = -p\log(p) - q\log(q)$. For each internal node in the tree (assuming the tree is rooted) the species can be partitioned into L_1 species that are the descendants of that internal node and L_2 species that are not. Denote P_1,Q_1 and P_2,Q_2 the presence/absence counts in each of these two groups, respectively. Again, denote the relative frequencies by $p_1 = P_1/L_1$, $q_1 = Q_1/L_1$, $p_2 = P_2/L_2$ and $q_2 = Q_2/L_2$. The entropies within each group are given by $H_1 = -p_1\log(p_1) - q_1\log(q_1)$ and $H_2 = -p_2\log(p_2) - q_2\log(q_2)$. The information gain of this node is thus $IG = H - [H_1(L_1/L) + H_2(L_2/L)]$. We traversed all of the internal nodes in the tree and found the one with the maximal information gain. This information gain value (and the partition it induces) was assigned to the compound and its phyletic patterns. We sorted all compounds according to their

information gain and examined those with the highest information gain values. A table describing all of the obtained measures for each compound can be found in Dataset S2.

Pathway Enrichment. The metabolic pathways in which each compound participates were retrieved from KEGG. Given a set of compounds (e.g., those for which $N_s/N_o > 0.5$), we counted the number of compounds from this set that participate in each pathway. We repeated the same procedure for 10,000 random compound sets (of the same size) to calculate which pathways are significantly overrepresented or underrepresented in the given set. The resulting P values were further corrected for multiple testing via the false discovery rate procedure (22).

Prediction of Exogenously Acquired Amino Acids and Cofactors in Ehrlichiosis Agents. Data concerning amino acid and cofactor biosynthesis in ehrlichiosis agents were retrieved from ref. 23; see table 5 therein. These data span three newly sequenced agents (*Anaplasma phagocytophilum*, *Ehrlichia chaffeensis*, and *Neorickettsia sennetsu*), as well as other species from the Rickettsias order (*Anaplasma marginale*, *Ehrlichia ruminantium*, *Wolbachia pipientis wMel*, and *Rickettsia prowazekii*) and several insect symbionts (two *Buchnera* species, *Candidatus Blochmannia floridanus*, and *Wigglesworthia glossinidia*). We discarded the *Buchnera* species (to avoid reuse of *Buchnera* related data for validation), leaving a total of nine species. In each species, the ability to synthesize 20 amino acids and 10 vitamins/cofactors was reported. We examined the seed sets obtained by our analysis and retrieved a corresponding dataset, describing the seed state of each of these 30 compounds in the same nine species (Table S3). To obtain maximum information for these newly sequenced species we used a more recent KEGG compilation (Release 45.0, January 1, 2008) and did not prune the networks. Because our focus here is the ability of the seed-detection algorithm to correctly distinguish externally acquired seeds from synthesized (non-seed) compounds, we limited our analysis to compounds that were found in the network. We compared the two datasets and examined whether compounds reported in ref. 23 not to be synthesized in a specific species are correctly identified by our algorithm as seeds. To evaluate the accuracy of our seed prediction, we regarded it as a binary classification problem, wherein the seed detection algorithm aims to classify synthesized (non-seed) vs. non synthesized (seed) compounds. Classification accuracy is therefore defined as $(TP + TN)/(TP + FP + FN + TN)$, where TP denotes the number of true positives, TN denotes the number of true negatives, FP denotes the number of false positives, and FN denotes the number of false negatives. Statistical significance of the resulting accuracy measure was computed by shuffling the species' and compounds' labels 10,000,000 times and calculating the probability to achieve an equal or higher accuracy by chance. Similarly, focusing on correct seed prediction, precision was calculated as $TP/(TP + FP)$ and recall as $TP/(TP + FN)$.

Tamura and Nei's Method for Substitution Rate Estimation. We follow Tamura and Nei (24) in estimating the number and rate of substitutions across a phylogenetic tree (see also ref. 25). Their method was originally developed for nucleotide substitution estimates in DNA sequences but can be applied in an analogous manner to transitions between states of any discrete trait. In the context of our study, it is convenient to imagine that each species is associated with a sequence of states, wherein locus l describes the state of compound l in this species, as an analog to a DNA sequence of a species. The state of the trait in all of the ancestral species (internal nodes of the tree) is inferred from the states in the extant species, using the maximum parsimony principle. The number of substitutions of each type (from state i to state j) is counted by comparing the state of the trait in each species with its immediate ancestor. In cases where the most parsimonious assignment is ambiguous (i.e., there are two equality parsimonious assignments), each of the two states (and pertaining substitutions) is considered with probability 0.5. The states of compounds that were pruned during network reconstruction (see *Metabolic Networks Reconstruction*) are also considered ambiguous and can take either a seed or a non-seed state. Compounds for which the most parsimonious assignment in some node comprises all three states are dropped from our analysis (following ref. 24). To estimate the relative frequencies of each substitution type, the number of substitutions is divided by the frequency of the original (ancestor) state across all of the sequences that are included in the analysis. The relative frequencies are then expressed such that the total sum of all of the frequencies is 100%.

Phylogenetic Tree Reconstruction and Evaluation. We again restricted our analysis to the species that can be matched to those appearing in the reference phylogenetic tree of ref. 26. We calculated the Jaccard distance (27) matrix between the seed sets of each pair of species and applied both the neighbor-joining algorithm (28) and the Fitch–Margoliash algorithm (29) (implemented by the programs NEIGHBOR and FITCH from the PHYLIP package, respectively) to this matrix to reconstruct the phylogenetic tree relating the species. Similarly, we reconstructed phylogenetic trees based on distances between sets of occurring compounds and random seeds sets (i.e., random subsets of the occurring compounds with the same size as the real seed sets). The distance between each of these trees and the reference tree was evaluated by both the Branch Score Distance measure (30) and the Symmetric Difference measure (31) (implemented by the TREEDIST program from the PHYLIP package). MEGA Tree Explorer (32) was used to draw the phylogenetic tree.

Environmental Attributes. We provide here a detailed description of each category used in the environmental attributes data. This description is adapted from the NCBI genome project help and can be found online (www.ncbi.nlm.nih.gov/genomes/static/gprj_help.html).

"Salinity" describes the salinity requirements of the bacterium (percentage of salt as sodium chloride equivalent in the growth medium): nonhalophilic, 0–2% NaCl; mesophilic, 2–5% NaCl; moderate halophile, 5–20% NaCl; extreme halophile, 20–30% NaCl.

"Oxygen" describes the ability of the organism to live at various levels of oxygen: null, unknown oxygen requirements; aerobic, the organism can grow in the presence of oxygen and probably uses oxygen as an electron acceptor; microaerophilic, the organism can tolerate low levels of oxygen and probably does not use oxygen as an electron acceptor; facultative, the organism can grow both aerobically or anerobically; anaerobic, the organism grows in the absence of oxygen and utilizes alternative electron acceptors.

"Temperature range" describes the basic category of temperature range (in Celsius) at which the organism grows. Organisms that grow at ranges that overlap multiple categories are classified based on with which category the majority of their temperature range overlapped: unknown, it is not known at what temperature this organism grows; cryophilic, the organism grows at −30 to −2; psychrophilic, the organism grows at −1 to + 10; mesophilic, the organism grows at +11 to +45; thermophilic, the organism grows at +46 to +75; hyperthermophilic, the organism grows above +75.

"Habitat" describes the basic environments in which the organism is found: unknown, it is not known where this organism grows; host-associated, this organism is often or obligately associated with a host organism; aquatic, this organism is often

or obligately associated with either fresh or seawater environments; terrestrial, this organism is often or obligately associated with a terrestrial environment such as soil; specialized, this organism lives in a specialized environment like a marine thermal vent; multiple, the organism can be found in more than one of the above environments.

1. Ahren D, Ouzounis C (2004) Robustness of metabolic map reconstruction. *J Bioinform Comput Biol* 2:589–593.
2. Jeong H, Tombor B, Albert R, Oltvai Z, Barabasi A (2000) The large-scale organization of metabolic networks. *Nature* 407:651–654.
3. Edwards J, Palsson B (2000) Robustness analysis of the Escherichia coli metabolic network. *Biotechnol Progr* 16:927–937.
4. Stelling J, Klamt S, Bettenbrock K, Schuster S, Gilles E (2002) Metabolic network structure determines key aspects of functionality and regulation. *Nature* 420:190–193.
5. Deutscher D, Meilijson I, Kupiec M, Ruppin E (2006) Multiple knockout analysis of genetic robustness in the yeast metabolic network. *Nat Genet* 38:993–998.
6. Zeng A, Ma H (2003) The connectivity structure, giant strong component and centrality of metabolic networks. *Bioinformatics* 19:1423–1430.
7. Csete M, Doyle J (2004) Bow ties, metabolism and disease. *Trends Biotechnol* 22:446–450.
8. Falkowski P (2006) Tracing oxygen's imprint on earth's metabolic evolution. *Science* 311:1724–1725.
9. Acquisti C, Kleffe J, Collins S (2007) Oxygen content of transmembrane proteins over macroevolutionary time scales. *Nature* 445:47–52.
10. Raymond J, Segre D (2006) The effect of oxygen on biochemical networks and the evolution of complex life. *Science* 311:1764–1767.
11. Parter M, Kashtan N, Alon U (2007) Environmental variability and modularity of bacterial metabolic networks. *BMC Evol Biol* 7:169.
12. Kreimer A, Borenstein E, Gophna U, Ruppin E (2008) The evolution of modularity in bacterial metabolic networks. *Proc Natl Acad Sci USA* 105:6976–6981.
13. Stephens R, *et al.* (1998) Genome sequence of an obligate intracellular pathogen of humans: Chlamydia trachomatis. *Science* 282:754–759.
14. Kubo A, Stephens R (2001) Substrate-specific diffusion of select dicarboxylates through chlamydia trachomatis PorB. *Microbiology* 147:3135–3140.
15. Shigenobu S, Watanabe H, Hattori M, Sakaki Y, Ishikawa H (2000) Genome sequence of the endocellular bacterial symbiont of aphids Buchnera sp. APS. *Nature* 407:81–86.
16. Pollack J, Williams M, McElhaney R (1997) The comparative metabolism of the mollicutes (Mycoplasmas): The utility for taxonomic classification and the relationship of putative gene annotation and phylogeny to enzymatic function in the smallest free-living cells. *Crit Rev Microbiol* 23:269–354.
17. Tanaka T, Ikeo K, Gojobori T (2006) Evolution of metabolic networks by gain and loss of enzymatic reaction in eukaryotes. *Gene* 365:88–94.
18. Aho A, Hopcroft J, Ullman J (1974) *The Design and Analysis of Computer Algorithms* (Addison-Wesley, Reading, MA).
19. Tarjan R (1972) Depth-first search and linear graph algorithms. *SIAM J Comput* 1:146–160.
20. Tatusov RL, Galperin MY, Natale DA, Koonin EV (2000) The COG database: A tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* 28:33–36.
21. Tatusov R, *et al.* (2003) The COG database: An updated version includes eukaryotes. *BMC Bioinformatics* 4:41.
22. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Stat Soc* 57:289–300.
23. Dunning Hotopp J, *et al.* (2006) Comparative genomics of emerging human ehrlichiosis agents. *PLoS Genet* 2:e21.
24. Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial dna in humans and chimpanzees. *Mol Biol Evol* 10:512–526.
25. Imanishi T, Gojobori T (1992) Patterns of nucleotide substitutions inferred from the phylogenies of the class I major histocompatibility complex genes. *J Mol Evol* 35:196–204.
26. Ciccarelli FD, *et al.* (2006) Toward automatic reconstruction of a highly resolved tree of life. *Science* 311:1283–1287.
27. Jaccard P (1908) Nouvelles recherches sur la distribution florale. *Bull Soc Vaudoise Sci Nat* 44:223–270.
28. Saitou N, Nei M (1987) The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol Biol Evol* 4:406–425.
29. Fitch W, Margoliash E (1967) Construction of phylogenetic trees. *Science* 155:279–284.
30. Kuhner MK, Felsenstein J (1994) A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol Biol Evol* 11:459–468.
31. Robinson D, Foulds LR (1981) Comparison of phylogenetic trees. *Math Biosci* 53:131–147.
32. Kumar S, Tamura K, Nei M (2004) MEGA3: Integrated software for molecular evolutionary genetics. *Anal Sequence Alignment Brief Bioinformatics* 5:150–163.

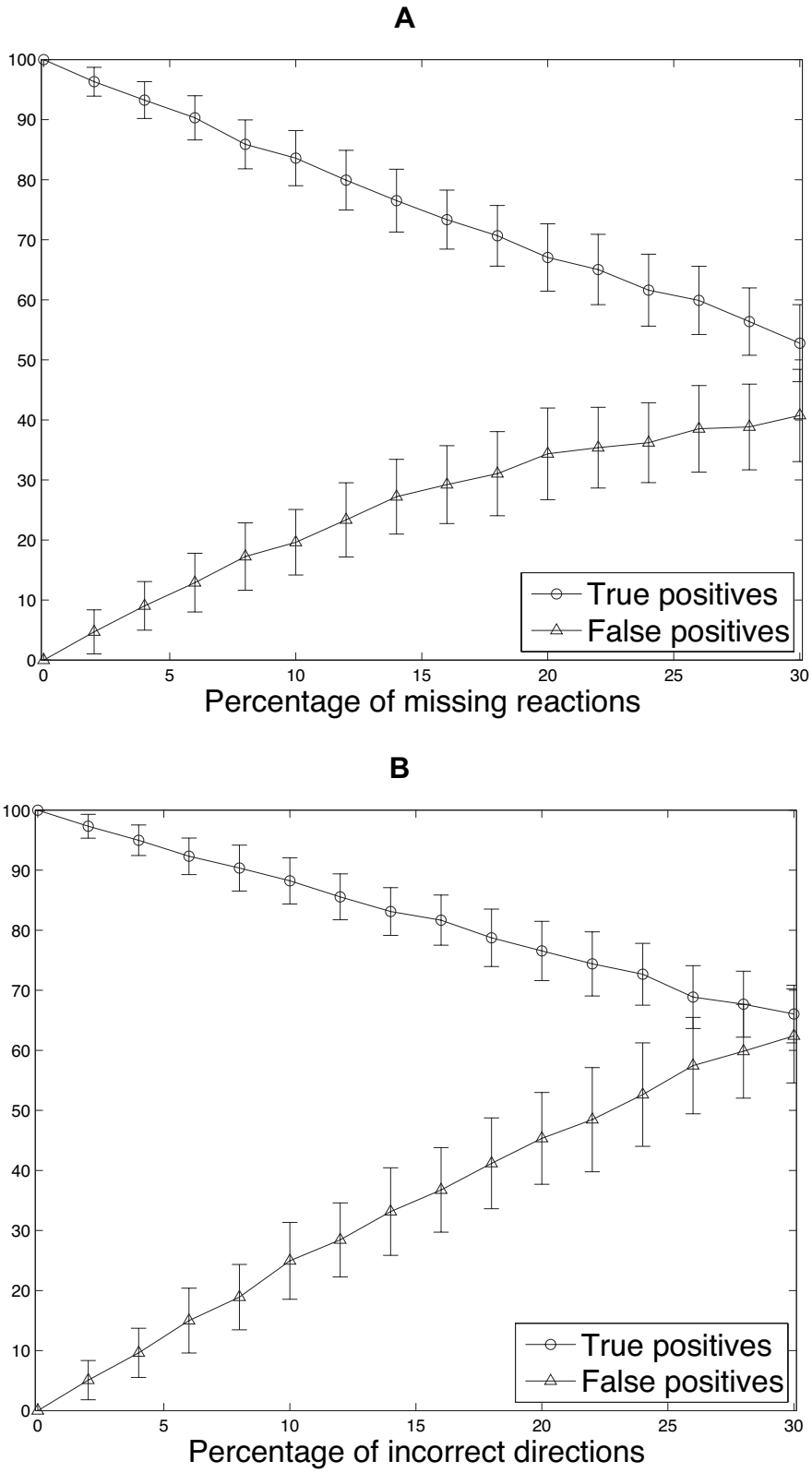

Fig. S1. The effect of missing or erroneous data on seed set identification. The results were obtained by perturbing the metabolic network of *Saccharomyces cerevisiae* and comparing the seed set of the perturbed network to that of the original one. The curves illustrate the percentage of true positive and false positive seed compounds (compared with the size of the original seed set) as a function of the percentage of missing reactions (*A*) or erroneous reaction directionality (*B*). Each data point represents the average of 100 simulations.

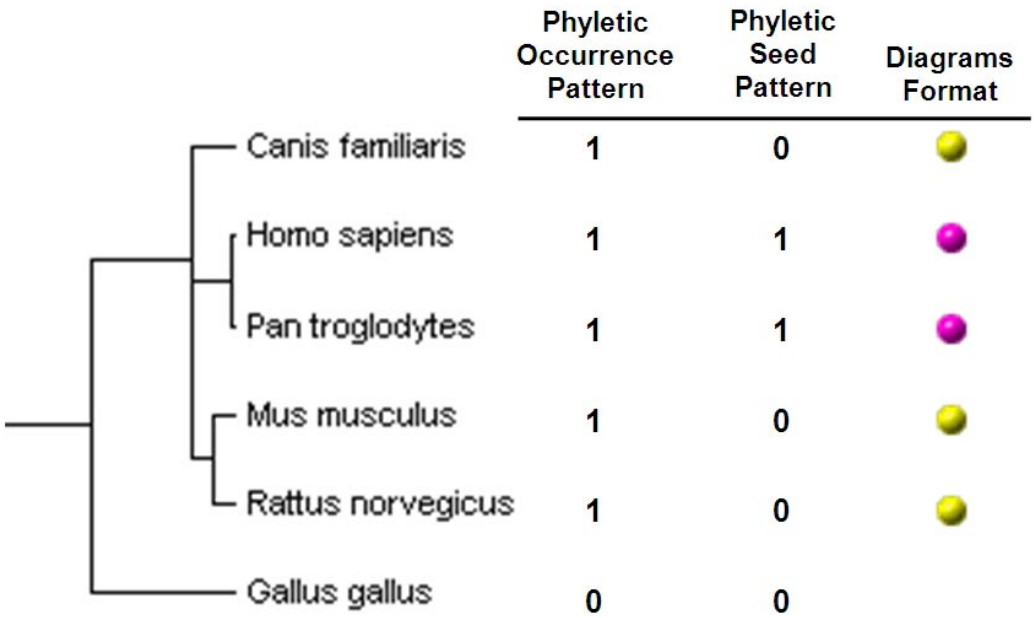

Fig. S2. Phyletic occurrence patterns and phyletic seed patterns. In this illustrative example, a certain compound of interest can be found as an occurring compound in all species but *Gallus gallus* yet is included in the seed sets of only *Homo sapiens* and *Pan troglodytes*. The phyletic occurrence pattern and the phyletic seed pattern accordingly represent the presence/absence pattern in each set. In our diagram format representation that illustrates the phyletic patterns of various key compounds, yellow bullets represent species in which the compound occurs and purple bullets represent species in which the compound is a seed.

Fig. S3. The phyletic patterns of phenylalanine (*A*) and glutamate (*B*). Yellow bullets indicate species in which this compound occurs. Purple bullets indicate species in which this compound is included in the seed set. Question marks indicate species in which this compound was pruned and hence its status is unknown.

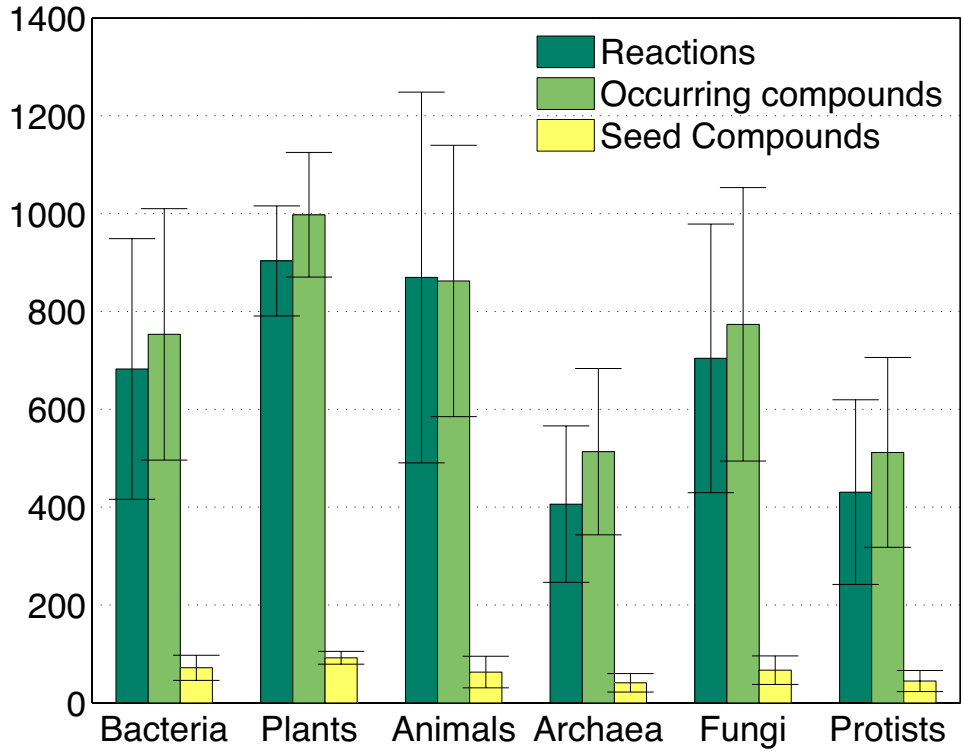

Fig. S4. The average number of reactions, occurring compounds, and seed compounds within different taxonomic groups. The number of seed compounds is estimated by the number of source components.

Fig. S5. The Strongly Connected Component graph of *Saccharomyces cerevisiae* with the source components marked in red. The number within each node denotes the number of compounds (from the original metabolic network) included in this component.

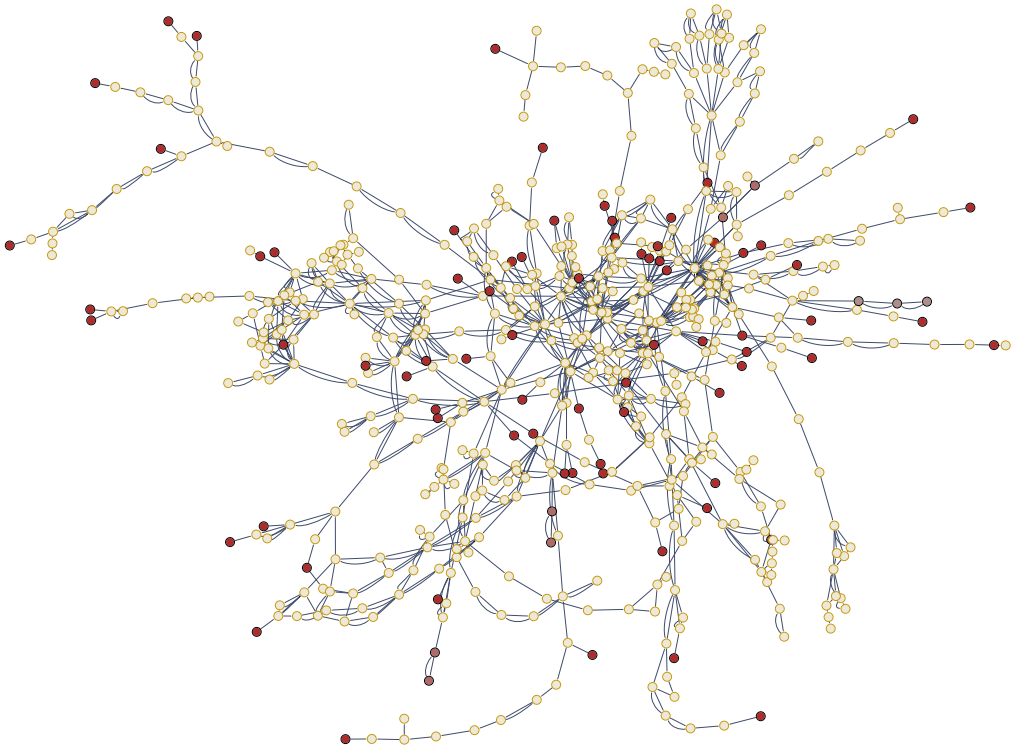

Fig. S6. The metabolic network of *Saccharomyces cerevisiae* with the seed compounds colored in red (as in Fig. 1). The color saturation denotes the seed's confidence level, *C* (see *Materials and Methods*), with a darker red indicating a higher confidence level.

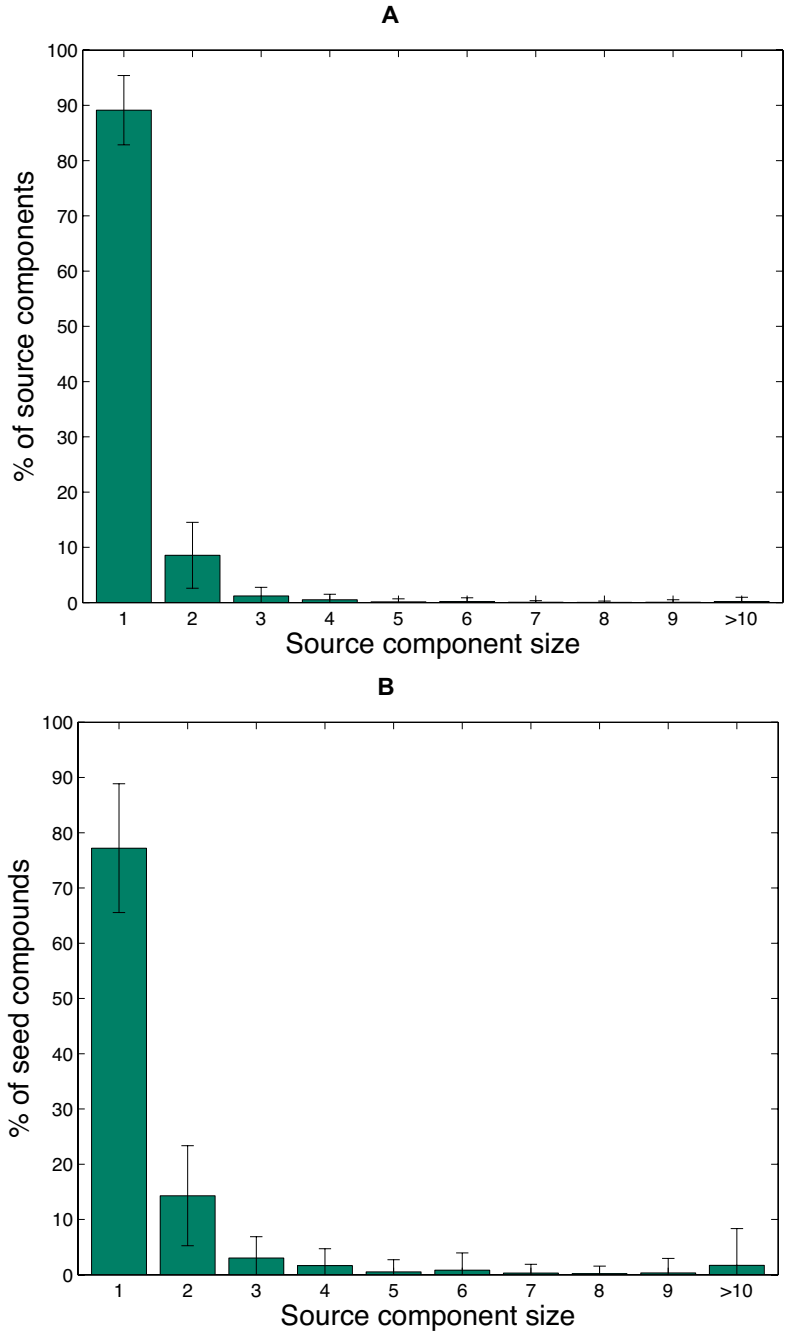

Fig. S7. Source component statistics. (*A*) The average percentage of source components as a function of their size (and the standard deviation across the different species). Evidently, most (>89%) of the source components are singletons. The largest source component found in our analysis includes 24 compounds; however, such giant source components are extremely rare. (*B*) The percentage of seed compounds that are part of source components of varying size.

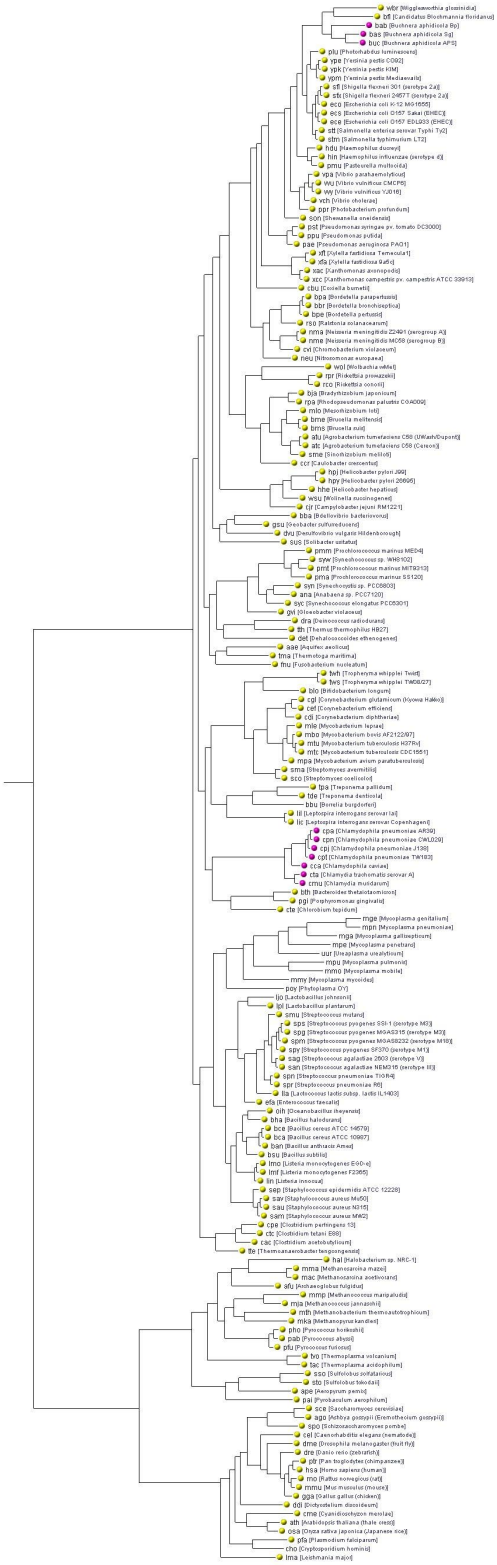

Fig. S8. The phyletic pattern of 2-oxoglutarate. Yellow bullets indicate species in which this compound occurs. Purple bullets indicate species in which this compound is included in the seed set.

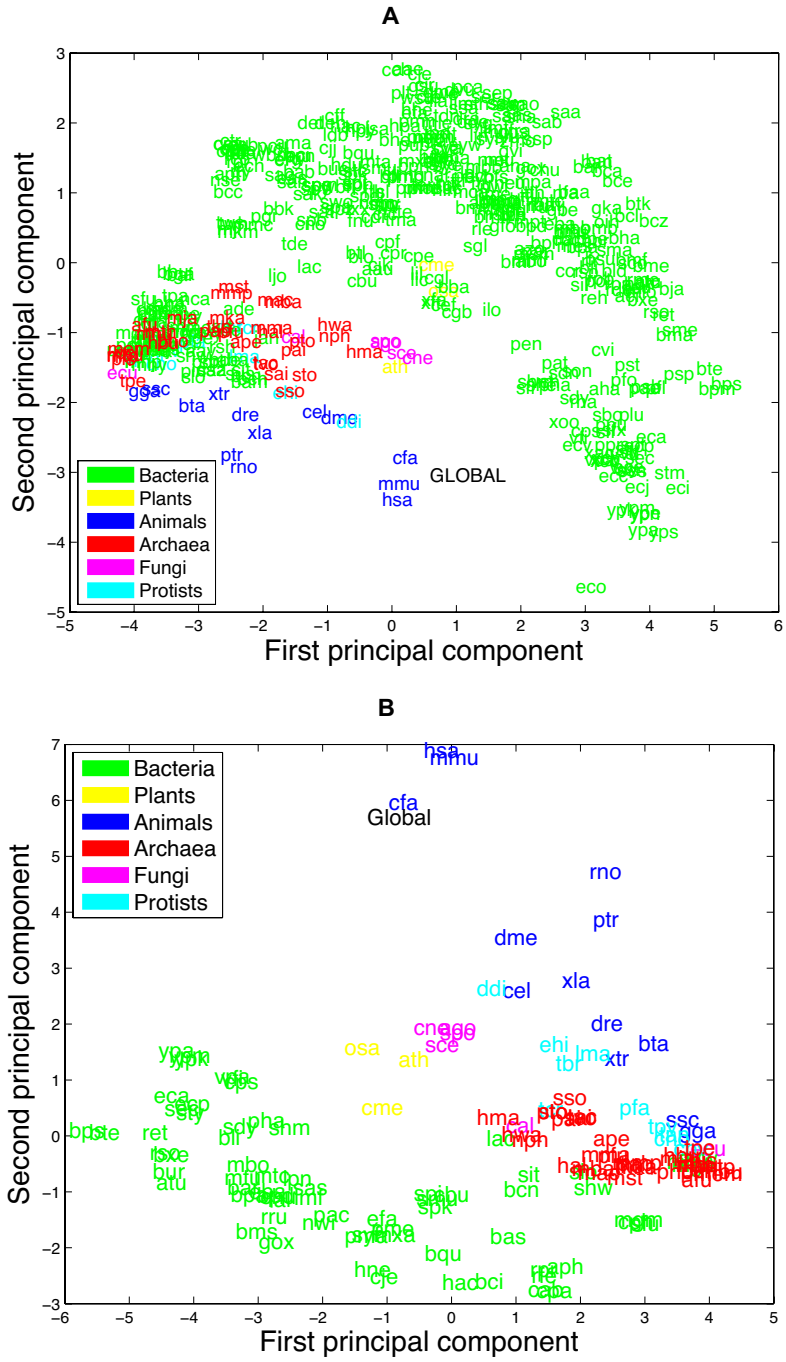

Fig. S9. Principal components analysis of the seed sets data. (*A*) Projection of the various taxa on the first two principal components of the seed sets. All species were used in this analysis. (*B*) Principal components analysis of seed sets composition using only a subset of the bacterial taxa. We randomly picked bacterial taxa to be included in this analysis so that they constitute approximately half the total number of species. Evidently, this correction improves the partition between bacteria and other taxonomic groups.

Table S2.

KEGG code	Compound	Confidence (*C*)
C04272	(*R*)-2,3-Dihydroxy-3-methylbutanoate	0.333333333
C02612	(*R*)-2-Methylmalate	1
C06010	(*S*)-2-Acetolactate	0.333333333
C00026	2-Oxoglutarate	1
C04181	3-Hydroxy-3-methyl-2-oxobutanoic acid	0.333333333
C01259	3-Hydroxy-*N*6,*N*6,*N*6-trimethyl-L-lysine	1
C03688	Apo-[acyl-carrier protein]	1
C04246	But-2-enoyl-[acyl-carrier protein]	0.5
C05745	Butyryl-[acp]	0.5
C15811	C15811	1
C00993	D-Alanyl-D-alanine	1
C00857	Deamino-NAD+	0.25
C05755	Decanoyl-[acp]	0.5
C05512	Deoxyinosine	1
C00882	Dephospho-CoA	0.5
C00031	D-Glucose	1
C00217	D-Glutamate	0.5
C00921	Dihydropteroate	1
C00235	Dimethylallyl diphosphate	1
C05223	Dodecanoyl-[acyl-carrier protein]	0.5
C00288	HCO3-	1
C05749	Hexanoyl-[acp]	0.5
C00826	L-Arogenate	1
C00025	L-Glutamate	0.5
C00064	L-Glutamine	1
C00155	L-Homocysteine	1
C01209	Malonyl-[acyl-carrier protein]	1
C00392	Mannitol	1
C00253	Nicotinate	0.25
C05841	Nicotinate D-ribonucleoside	0.25
C01185	Nicotinate D-ribonucleotide	0.25
C05752	Octanoyl-[acp]	0.5
C01260	P1,P4-Bis(5′-adenosyl) tetraphosphate	1
C01134	Pantetheine 4′-phosphate	0.5
C00134	Putrescine	1
C05684	Selenite	1
C00750	Spermine	1
C00059	Sulfate	1
C05761	Tetradecanoyl-[acp]	0.5
C01081	Thiamin monophosphate	1
C00320	Thiosulfate	1
C05754	trans-Dec-2-enoyl-[acp]	0.5
C05758	trans-Dodec-2-enoyl-[acp]	0.5
C05748	trans-Hex-2-enoyl-[acp]	0.5
C05751	trans-Oct-2-enoyl-[acp]	0.5
C05760	trans-Tetradec-2-enoyl-[acp]	0.5
C01636	tRNA(Arg)	1
C01638	tRNA(Asp)	1
C01639	tRNA(Cys)	1
C01640	tRNA(Gln)	1
C01641	tRNA(Glu)	1
C01642	tRNA(Gly)	1
C01643	tRNA(His)	1
C01646	tRNA(Lys)	1
C01647	tRNA(Met)	1
C01648	tRNA(Phe)	1
C01650	tRNA(Ser)	1
C01651	tRNA(Thr)	1
C01652	tRNA(Trp)	1
C04700	UDP-*N*-acetylmuramoyl-L-alanyl-D-glutamyl-L-lysine	1
C05892	UDP-*N*-acetylmuramoyl-L-alanyl-γ-D-glutamyl-L-lysine	1

Table S3.

KEGG	Name	aph	ama	ech	eru	wol	nse	rpr	bfl	wbr
C00041	L-Alanine	N	N	N	N	N	N	N	N	N
C00062	L-Arginine	N	N	N	N	N	N	N	N	N
C00152	L-Asparagine	—	S	—	—	—	—	—	S	N
C00049	L-Aspartate	N	N	N	N	N	N	N	N	N
C00097	L-Cysteine	N	N	N	N	N	N	N	N	N
C00037	Glycine	N	N	N	N	N	N	N	N	N
C00025	L-Glutamate	N	N	N	N	N	N	N	N	N
C00064	L-Glutamine	N	N	N	N	N	N	N	N	N
C00135	L-Histidine	S	S	S	S	S	S	S	N	S
C00123	L-Leucine	S	S	S	S	S	S	S	S	S
C00047	L-Lysine	S	S	N	N	S	S	S	N	S
C00407	L-Isoleucine	S	S	S	S	S	S	S	N	S
C00073	L-Methionine	S	S	S	S	S	S	S	N	S
C00079	L-Phenylalanine	S	S	S	S	S	S	S	N	S
C00148	L-Proline	S	S	S	N	S	S	S	S	N
C00065	L-Serine	N	N	N	N	N	N	N	N	N
C00188	L-Threonine	S	S	S	S	S	S	S	N	S
C00078	L-Tryptophan	S	S	S	S	S	S	S	N	S
C00082	L-Tyrosine	S	S	S	S	S	S	S	N	S
C00183	L-Valine	S	S	S	S	S	S	S	N	S
C00120	Biotin	N	N	N	N	—	N	—	S	N
C00016	FAD	N	N	N	N	N	N	—	N	N
C00504	Folate	—	—	—	—	N	—	—	N	N
C00725	Lipoate	—	—	—	—	—	—	—	S	—
C00003	NAD+	N	N	N	N	S	N	N	—	N
C00010	CoA	N	N	N	N	N	N	N	S	N
C00032	Heme	N	N	N	N	N	N	N	S	N
C00627	Pyridoxine phosphate	N	N	N	N	N	N	—	N	N
C00378	Thiamine	—	—	—	N	—	—	—	—	—
C00399	Ubiquinone	N	N	N	N	N	N	N	N	N

S, seed; **N**, non-seed (synthesized) compound; —, absent.
aph, *Anaplasma phagocytophilu*; ama, *Anaplasma marginalem*; ech, *Ehrlichia chaffeensis*; eru, *Ehrlichia ruminantium*; wol, *Wolbachia wMel*; nse, *Neorickettsia sennetsu*; rpr, *Rickettsia prowazekii*; bfl, *Candidatus Blochmannia floridanus*; wbr, *Wigglesworthia glossinidia*.

Table S4. The relative frequencies of transitions between the various states of a compound

	New state		
Original state	Absent	Non-seed	Seed
Absent	—	13.2587	8.3539
Non-seed	30.7564	—	3.7524
Seed	37.9341	5.9445	—

The values presented are based on maximum parsimony reconstruction of the internal states of each compound, based on its state in the extant species (Mat*erials and Methods*, first assay). These frequencies describe the expected number of conversions from one state to the other among 100 conversion events in a random set of compounds with equal number in each state.

Table S5. The rate matrix, *Q*, describing conversions rates between the various states

	New state		
Original state	Absent	Non-seed	Seed
Absent	—	0.4046	0.1879
Non-seed	1.3521	—	0.2948
Seed	4.7131	1.2222	—

The presented values are based on maximum likelihood estimation of the internal states and conversion rates [see Yang Z (2007) and were obtained with PAML (*Materials and Methods*, third assay). The matrix of transition probabilities is given by $P(t) = exp(Qt)$.

Other Supporting Information Files

Table S1 (PDF)
Dataset S1 (TXT)
Dataset S2 (XLS)