Nucleic Acids Research Advance Access published October 24, 2016

doi: 10.1093/nar/gkw964

The landscape of microbial phenotypic traits and associated genes

Maria Brbić[1], Matija Piškorec[1], Vedrana Vidulin[1], Anita Kriško[2], Tomislav Šmuc[1] and Fran Supek[1,3,4,*]

[1]Division of Electronics, Ruder Boskovic Institute, 10000 Zagreb, Croatia, [2]Mediterranean Institute of Life Sciences, 21000 Split, Croatia, [3]EMBL/CRG Systems Biology Research Unit, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, 08003 Barcelona, Spain and [4]Universitat Pompeu Fabra (UPF), 08002 Barcelona, Spain

Received June 12, 2016; Revised September 21, 2016; Editorial Decision October 06, 2016; Accepted October 11, 2016

ABSTRACT

Bacteria and Archaea display a variety of phenotypic traits and can adapt to diverse ecological niches. However, systematic annotation of prokaryotic phenotypes is lacking. We have therefore developed ProTraits, a resource containing ∼545 000 novel phenotype inferences, spanning 424 traits assigned to 3046 bacterial and archaeal species. These annotations were assigned by a computational pipeline that associates microbes with phenotypes by text-mining the scientific literature and the broader World Wide Web, while also being able to define novel concepts from unstructured text. Moreover, the ProTraits pipeline assigns phenotypes by drawing extensively on comparative genomics, capturing patterns in gene repertoires, codon usage biases, proteome composition and co-occurrence in metagenomes. Notably, we find that gene synteny is highly predictive of many phenotypes, and highlight examples of gene neighborhoods associated with spore-forming ability. A global analysis of trait interrelatedness outlined clusters in the microbial phenotype network, suggesting common genetic underpinnings. Our extended set of phenotype annotations allows detection of 57 088 high confidence gene-trait links, which recover many known associations involving sporulation, flagella, catalase activity, aerobicity, photosynthesis and other traits. Over 99% of the commonly occurring gene families are involved in genetic interactions conditional on at least one phenotype, suggesting that epistasis has a major role in shaping microbial gene content.

INTRODUCTION

Bacteria and Archaea inhabit a wide spectrum of ecological niches, including growth in extreme environments and association to plant or animal hosts, whether mutualistic, commensal or parasitic. This is made possible by a plethora of physiological adaptations observed in prokaryotes, such as the use of different carbon sources and electron acceptors, resistance to stressors and molecular interactions with host cells. Broadly construed, the notion of a microbial phenotypic trait encompasses all of the above facilities – the ability to colonize ecological niches and the underlying physiological features. A deeper characterization of such traits can be obtained by linking them to the genetic makeup of the microbes (1,2). Statistical associations of genes to phenotypes can implicate certain proteins or pathways, providing insight into the mechanistic basis of phenotypic traits (3,4), as demonstrated for adaptation to stress (5,6), host-association (7,8), pathogenesis (9,10), drug resistance (11,12) and relevance to biotechnological applications (13,14).

The amount of prokaryotic genomes is increasing rapidly (15), aided by single-cell sequencing (16) and metagenomics (17). However, efforts to obtain systematic, high-quality phenotype annotation of microbes are not keeping pace, meaning that the potential for comprehensive gene-trait association studies cannot be realized (1,18). A thorough characterization of phenotypes of phylogenetically diverse microbial taxa could implicate genes in phenotypic traits by capturing the evolutionary signal across the prokaryotic tree of life. Moreover, multiple traits could be considered jointly, as they have been for mammalian phenotypes (19,20), thus boosting statistical power and elucidating relationships between the traits (21). Comprehensive resources that link phenotypes to genes and pathways exist for several eukaryotic model organisms (22,23), but not so for prokaryotes.

Current databases of microbial genome sequencing projects (24,25) do supply a certain amount of annotated

*To whom correspondence should be addressed. Tel: +385 1 4561 080; Email: fran.supek@irb.hr

traits, however, these important resources do not focus on phenotype data and thus the coverage is not extensive. In contrast, the scientific literature abounds with trait descriptions stored as unstructured text, which are therefore not directly accessible to automated analyses. Relying on manual curation to organize these data does not scale with the increasing volume of scientific publications. Motivated by the above, we have developed the ProTraits resource, a comprehensive atlas of 424 microbial phenotypes, covering 3046 Bacterial and Archaeal species, each receiving on average 165 novel high-confidence annotations in addition to 23 previously known labels; available online at http://protraits.irb.hr/.

The pipeline behind ProTraits relies on automated text mining of diverse corpora of biological literature, annotating microbes with existing phenotypic traits, while additionally being able to define novel phenotypic concepts from free text in an unsupervised manner. Furthermore, our approach draws extensively on genomic data, including novel methods of automated phenotype inference from conserved gene neighborhoods and from evolution of synonymous codon biases. The broad coverage with phenotype annotations allows us to systematically discover thousands of high-confidence statistical associations that link genes to traits, thereby informing about their genetic basis. Finally, our analyses suggest that epistatic interactions are ubiquitous within bacterial and archaeal gene repertoires, affecting almost all commonly-occurring gene families.

MATERIALS AND METHODS

Definitions

Here, we adopt a broad definition of a phenotypic trait (or, simply, trait) of a microorganism. Firstly, this encompasses the common meanings of this term in the field of microbiology (metabolic capabilities, morphology, growth conditions). Secondly, this also encompasses the microbe's ability to colonize certain ecological niches, which includes the association to a particular host and an organ/tissue thereof, and moreover the type of this association (pathogenic, symbiotic). While a trait is any attribute of an organism which matches the above definition, the term 'phenotype' also implies a value of this attribute. Upon initial data collection (described below), all traits were binarized, meaning that multi-valued categorical traits were split into three or more binary traits. In other words, as defined herein, there are two phenotypes each trait can exhibit: presence or absence of the trait.

Phenotype data collection

The initial set of phenotype assignments was combined from several sources. First, we merged the phenotype data in the NCBI microbial genome projects list ('lproks0' table, now retired) with the largely overlapping set of traits from the BacMap database (24). We collapsed together synonymous traits from the two databases and furthermore the data on causal roles in diseases for various hosts from BacMap was manually categorized (Supplementary Table S1). In rare cases of traits exhibiting discordant phenotype labels between databases, we recorded that trait/species combination as a missing value (n = 147, list in Supplementary Table S1).

Second, we included descriptions of the ecosystems where microbes were isolated from, as provided in the GOLD database of genome sequencing projects (25). Since GOLD provides only positive assertions, we have provisionally annotated as negative examples for a certain ecosystem all those organisms which were not explicitly assigned to that ecosystem type. For instance, organisms annotated as 'marine' were used as provisional negatives for 'soil' or for 'thermal springs'. In GOLD, each organism can have more than one assigned value, e.g. being annotated as both 'marine' and 'freshwater' and thus receiving positive labels for these two ecosystems, and negative labels for all remaining ones. Such provisional negative annotations were used to train the classification models (see below).

Third, we further considered a set of biochemical phenotypes, which here implies the experimental data we manually collected from 265 articles describing new microbial strains, published in the IJSEM journal (26) during the years 2013 and 2014. These publications describe series of laboratory tests meant to discriminate novel species, such as the ability to grow on a particular substrate. Next, we collapsed together synonymous biochemical traits, further retaining the trait only if at least 30 species were covered with annotations (full list of synonymous names in Supplementary Table S1). We merged this data with a smaller set of experimental measurements of growth on various substrates for 40 species (27), measured using Biolog phenotype arrays, a technique with broad application to metabolic modeling (28).

Text sources and preprocessing

We analyzed text documents describing 1640 bacterial and archaeal species from (i) Wikipedia, (ii) MicrobeWiki, (iii) HAMAP proteomes, (iv) PubMed abstracts retrieved upon searching for the name of each species (setting 'sort by relevance'; first 100 items kept); (v) a set of PubMed Central publications that were looked up via the 'Reference' field in the KEGG Organisms repository; and (vi) a mixed collection of smaller resources that were pooled together (Bacmap, Genoscope, Joint Genome Institute, KEGG, NCBI and Karyn's Genomes). From the HAMAP source we downloaded all available organisms, for Wikipedia and MicrobeWiki we searched for bacterial and archaeal species names as defined in NCBI Taxonomy, and for the 'Pubmed abstracts' corpus (which was retrieved last) we searched for all species for which we already had at least one text document. All documents describing bacterial/archaeal strains were mapped to the corresponding species-level taxon, as defined by the NCBI Taxonomy. The documents were preprocessed by removing reference parts, English language stop-words (29) and all words occurring less than four times. Porter stemming (30) was used to reduce words to their root form.

Unsupervised discovery of phenotypic concepts

We applied non-negative matrix factorization (NMF) (31), a statistical method commonly used for text analysis (32,33)

in order to discover novel phenotypic concepts which may not be represented in existing databases. To this end, we constructed a standard 'bag-of-words' representation: a matrix where rows correspond to words and columns to organisms for each of the five corpora separately (excluding the mixed collection). In order to enforce consistency between corpora, for the NMF analysis we used only the words that appeared in all corpora (see Supplementary Methods for an exception). Matrix element values were the term frequency-inverse document frequency (*tf-idf*) weights, often used as a measure of how important a word is to a document, given a collection of documents (34). Briefly, *tf-idf* normalizes data so as to give greater weights to words which occur rarely, and thus are presumably more informative, while common words are down-weighted. In the NMF analysis, we filtered out names of bacterial and archaeal taxa, as well as a custom list of very frequent words (Supplementary Table S1).

We performed NMF in Matlab 2011b using alternating non-negativity constrained least squares (35) on each data matrix separately, with the number of hidden factors set to 50 and to 100 (in two separate NMF runs). We grouped together similar NMF factors across corpora, while requiring that a similar phenotypic concept had to be consistently discoverable in at least three corpora. The similarity between factors was calculated as the Pearson correlation coefficient between the NMF weights of top 20 ranked words. We further summarized the group by finding the median of word weights across factors in the group, so that each group (phenotypic concept) was represented with the top ranked words of the summary. These groups were then manually examined and those that were of high quality, in terms of consistency and content of words relevant for phenotypes, were further retained and treated in the same way as the other phenotypic traits collected from databases (see above). The groups and the constituent NMF factors are listed in Supplementary Table S2. Several runs of the NMF algorithm were performed to maximize coverage of discovered concepts (Supplementary Methods).

Predicting phenotypes by text mining

The training and the unlabeled data sets were generated for each trait and for each of the six text corpora separately, wherein learning examples were species, each described with the document(s) assigned to that species in a given text corpus. Again, we a used a standard bag-of-words representation with *tf-idf* weighting of word frequencies across texts; unlike the NMF analysis, here all words were used. All species with known positive or negative trait labels were part of the training set, while all species without known annotations for that trait were in the unlabeled data set, which later received predictions. A support vector machine (SVM) classifier (36) with a linear kernel was trained for each trait and for each text corpus separately using LibSVM (37). The regularization parameter C was optimized in five runs of 4-fold cross-validation ($C = 2^{-15}, 2^{-14}, \ldots, 2^{5}$), retaining the value yielding the best average area-under-the-curve (AUC) score. We then used a single SVM run of 10-fold cross-validation with the optimal C to estimate confidence scores of predictions, which were converted to precision (or positive predictive value) scores; equivalent to 1-false discovery rate (FDR). The precision scores were obtained separately for the positive and the negative class by thresholding the precision-recall curves obtained via cross-validation; see Supplementary Methods. If a training set corresponding to a trait-corpus pair did not have AUC > 0.6, or did not have at least one example that could be assigned at FDR < 0.25, we considered it as 'unlearnable' and did not use those predictions. We retained 424 of 522 traits that were learnable from at least one text corpus. In case of the GOLD ecosystems, provisionally annotated negatives (see above) were used for SVM training, but they were also included in the unlabeled data set to which the trained models were applied, meaning these provisional annotations could potentially be reassigned.

Inferring phenotypes from genomes and metagenomes

Prokaryotic genome sequences and gene annotations were from NCBI Entrez Genomes, while COG/NOG gene families were from eggNOG 3 (38). We only considered species with available genomes having a quality score ≥ 0.9 in (39). We further complemented inferences from text mining using genomic data, by predicting traits from: (i) the proteome composition, encoded as relative frequencies of amino acids (40,41); (ii) the gene repertoire, encoded as presence/absence indicators of COG gene families in a genome (42,43); (iii) co-occurrence of species across environmental sequencing data sets (44); (iv) gene neighborhoods (45,46), encoded as pairwise chromosomal distances between commonly occurring COGs; and (v) genomic signatures of translation efficiency in gene families (6,47), encoded as the MILC codon bias measure (48). For all five genomic representations, learning examples were species and class labels were the phenotypic traits. For the details of how the genomic features were calculated, see Supplementary Methods. For each representation and each trait, the Random Forest (RF) classification model (49) was trained to learn to distinguish between positive and negative trait annotations, using the FastRandomForest (50) implementation with 500 trees. A single run of 10-fold cross-validation was used to determine the performance and to estimate the precision/FDR scores of individual predictions in the same manner as for the text mining (Supplementary Methods). Again, we required cross-validation AUC > 0.6 and at least one organism labelled at FDR < 0.25 to consider a trait-genomic representation pair as learnable. For the purposes of determining individual genomic features informative of traits, we used the RF Gini importance measure (as implemented in scikit-learn; 200 trees; missing values imputed by the median).

Gene-trait associations and epistatic interactions

The Supplementary Methods describe the logistic regression-based statistical methodology we used to detect associations between the occurrence of 80 576 prokaryotic COG/NOG gene families and each phenotype, as well as the genetic interactions between pairs of COGs conditional on a phenotype, while controlling for phylogenetic relatedness of organisms.

Gene function analyses

The lists of genes known to be involved in endospore formation were from the *B. subtilis* genetic screen in (51) and from the collection of differentially expressed genes during sporulation (52). For the analysis of conserved gene neighborhoods associated with sporulation, we considered only organisms belonging to the *Firmicutes* and *Actinobacteria*, as almost all known spore-forming bacteria are contained within. There is a known confounding effect of genome size with the other genetic determinants of spore-forming ability (52). We have therefore used the propensity score matching methodology (53) to control for this effect, as implemented in the R package *Matching*, with parameters *replace = false* and *caliper = 0.5*. In brief, this algorithm resamples genomes to create two sets, one sporulating and other non-sporulating, which are matched by their genome size distribution. For every pair of examined COGs, we repeated this matching only on the set of genomes containing both COGs. The pairwise distances of the COGs in the sporulating versus the genome size-matched set of non-sporulating bacteria were compared using Mann–Whitney tests on all 44 850 pairs of 300 ubiquitous COGs (see above).

RESULTS

Integrating the known phenotype annotation databases

The ProTraits methodology infers microbial phenotypes from free-text and from genomic data by using supervised machine learning algorithms, in particular the Support Vector Machine (SVM) and Random Forest classifiers (Materials and Methods). In order to operate, such classifiers require an initial set of labelled examples – here, organisms with known phenotypes. These were obtained from four sources (Figure 1A): (i) 97 microbial traits collected from NCBI Genomes and from the BacMap genome atlas (24); (ii) 105 ecosystems where microbes were isolated from, provided by the GOLD database of genome sequencing projects (25); (iii) 113 phenotypic concepts that we inferred *de novo* by a topic modeling of biological literature, see below; and (iv) 109 biochemical phenotypic traits that we manually curated from experimental measurements in scientific publications (Materials and Methods), including phenotype array data from reference (27). Considered together, these sources span a wide range of microbial traits, however, they are often assigned only to a limited number of microbes (Figure 1B). In particular, out of 3046 prokaryotic species with sufficient coverage with text or high-quality genome sequences (Materials and Methods), a trait is currently annotated to a median of 112 species (Q1–Q3: 47–327), counting both the positive and the negative assignments. The modest coverage presents an opportunity for automated inference methods that draw on unstructured text sources, such as the ProTraits pipeline.

Defining novel phenotypic concepts from free text

We collected text documents describing microbes from six sources: (i) Wikipedia, (ii) MicrobeWiki, (ii) HAMAP proteomes (54), (iv) PubMed abstracts, (v) PubMedCentral publications and (vi) a collection of smaller text sources. All six corpora were encoded using a standard 'bag-of-words' approach, wherein each microbe was represented by the normalized word frequencies in the texts describing the organism (Materials and Methods). In order to discover novel phenotypic traits from this extensive collection, we applied non-negative matrix factorization algorithm (NMF (31)) to model phenotypic concepts across the texts (55). Each individual corpus was processed with repeated NMF runs and the results were clustered to uncover trends that are consistent across at least three text corpora (Figure 2A; Materials and Methods). Such novel phenotypic trait candidates (example in Figure 2B; list in Supplementary Table S2) further passed manual curation to ensure that the clustered words conform to a common theme, thus resulting in 113 non-redundant traits discovered from biological literature *de novo*. Expectedly, this method has recovered some of the phenotype annotations in existing databases: based on the overlap in microbes assigned to the known versus NMF-inferred traits, we find 9 traits highly similar to known ones (Spearman correlation of species' NMF weights >0.7), serving to validate the methodology. We have compared this NMF-based approach against a previous methodology that used principal components analysis (PCA) followed by hierarchical clustering (HC) to discover 'word sets' informative of various traits (56). We performed a blinded evaluation, where two human curators consistently judged the NMF concepts to be preferable over the PCA+HC word sets, according to subjective criteria (sign-test *P*-value ranging from 10^{-8} to 10^{-5} across evaluators and text corpora; outputs of the methods and the evaluation results are in Supplementary Table S2).

Automated annotation of microbes with phenotypes

In order to annotate bacterial and archaeal taxa with new phenotypes, we have input the bag-of-words data sets and the known phenotype labels to a SVM classifier. This was done separately for each trait and for each text corpus, and cross-validation was used to estimate the accuracy for the 2272 trained SVM models. Of note, the ecosystem classification of GOLD does not provide explicit negative annotations (Figure 1B) and we thus modeled using provisional negative labels (Materials and Methods). We also considered four other algorithms for this text classification task, but the SVM consistently outperformed them in accuracy on held-out data, across many different traits (Supplementary Figure S1; Supplementary Table S3).

In total, 424 (of 522) phenotypic traits could produce one or more highly-confident novel predictions (Methods), and this 'learnable' set of traits was considered further. In many cases, the text-based SVM models were highly predictive: median AUC scores for the traits in NCBI+Bacmap, GOLD and NMF phenotypic concepts of 0.87, 0.86 and 0.93, respectively (Figure 3C). Expectedly, the novel concepts appear easiest to predict, since they were discovered from the same text corpora using a different method (NMF). Biochemical phenotypes were challenging to infer from text in an automated manner (Figure 3C), but there were individual examples with good accuracy.

Next, we applied the trained SVM models to predict phenotypes for microbes that had no previous label for the

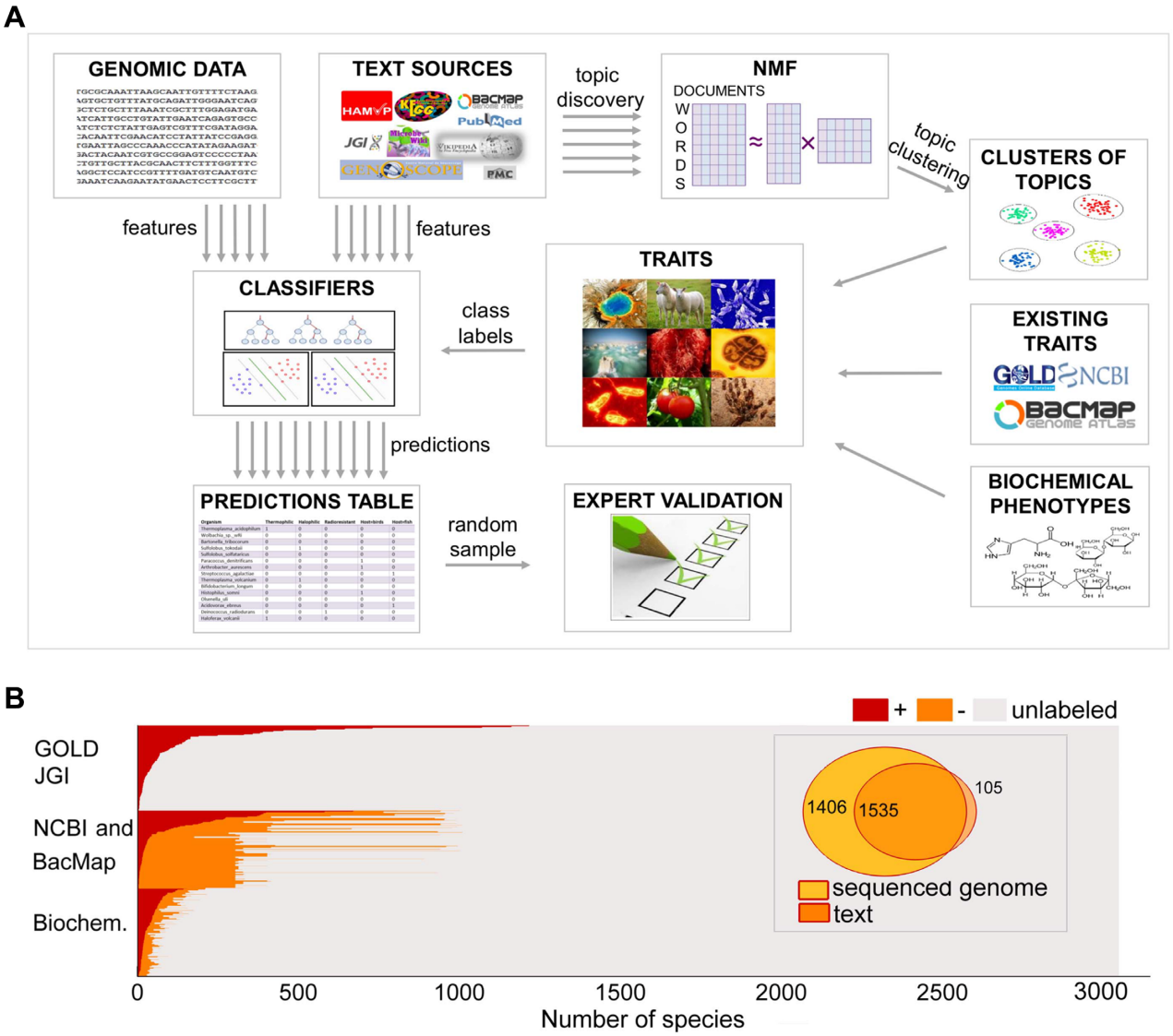


Figure 1. Inferring microbial phenotypic traits from free text and genomic data. (**A**) A schematic overview of the ProTraits pipeline. (**B**) The coverage with currently available trait labels in various databases. Inlay shows number of bacterial and archaeal species with available free-text and high-quality genome data we used for phenotype inference.

given trait, while obtaining FDR estimates for each prediction using crossvalidation precision-recall curves (57). An example for the halophilic phenotype is shown in Figure 3A, where progressively more relaxed FDR thresholds allow better coverage with phenotypic annotations, while trading off precision. Importantly, even the less-accurate SVM models can still annotate some microbes confidently (at low FDR), although the coverage is limited in such cases (Supplementary Figure S2A and B). The AUC scores for the individual classification models and the estimated coverage with annotations at several FDR thresholds is provided in Supplementary Table S3.

Orthogonal predictions from comparative genomics

We next complement these text-mining inferences by using an orthogonal approach, which relies on genome sequences and encompasses five independent data sources. These include three data types known to be informative of microbial phenotypes, in particular (i) the amino acid content of the proteome, e.g. (41), (ii) the gene repertoire of the genome, e.g. (42), here encoded as the presence/absence profile of COG/NOG gene families; and (iii) the patterns of co-occurrence of prokaryotic taxa across environmental sequencing data sets (44). Such classification models have previously been applied to predict small sets of up to 10 phenotypic traits, e.g. (40,43). Here, we extend this approach to a comprehensive set of 424 traits, thereby creating many broadly accurate models (median cross-validation AUC 0.83, 0.85 and 0.69, for the proteome composition, gene repertoires and metagenome co-occurrence, respectively; Figure 4A). While the biochemical phenotypes were, overall, challenging to infer by text mining (see above; Figure 3C), the genome data is more informative in this respect, thus complementing the text-based annotations (me-

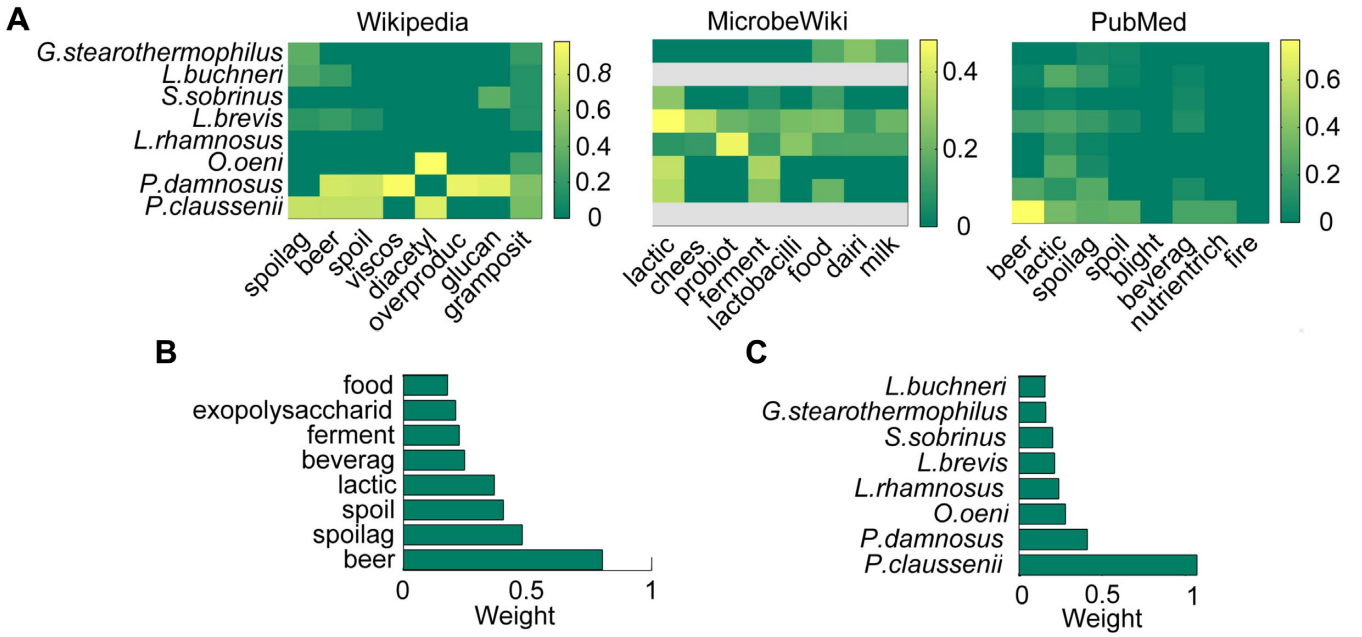


Figure 2. Unsupervised discovery of phenotypic concepts from text by non-negative matrix factorization (NMF). (**A**) An example phenotype 'beer/spoilage/spoil/lactic/beverage' captures lactic acid bacteria that cause beer spoilage. NMF weights for selected species (listed left) are shown separately for various keywords from three text sources: Wikipedia, MicrobeWiki and PubMed abstracts. Gray squares are missing data. (**B** and **C**) The final NMF-inferred phenotype consists of (B) keywords and of (C) species that were consistently highly weighted by NMF across the text corpora; shown weights are from medoids of clusters of NMF factors.

dian AUC from text is 0.54 versus 0.72 from genomes; Figure 4A). For this task of phenotype inference from comparative genomics, we used a Random Forest classifier, which we found to outperform the SVM and three other algorithms, providing more accurate predictions on a held-out data set (Supplementary Figure S1; Supplementary Table S3).

Gene synteny patterns accurately predict phenotypes

Evolutionarily conserved gene neighborhoods in prokaryotes (45) are known to reflect gene functional interactions (58,59). Given that a phenotype emerges from a network of many such interactions, we hypothesized that presence of certain gene neighborhoods in a genome may be used to infer the phenotype. One known example are the syntenic regions associated with cold tolerance in the genus *Shewanella* (46). Here, we systematically look for links between gene neighborhoods and various traits. We have thus described each genome using the pairwise chromosomal distances of 300 COGs that commonly occur across organisms (Materials and Methods), and indeed found that this can predict phenotypes accurately (median AUC = 0.80, Q1–Q3: 0.72–0.87), comparable to the well-established gene repertoire method. In some cases, the conserved gene neighborhoods were particularly accurate, e.g. for predicting chemolithotrophic organisms (Figure 4D). Thus, gene synteny patterns appear broadly associated with many microbial phenotypic traits.

We examined the spore-forming phenotype in more depth, since the involved genes were well-characterized experimentally (51,60,61) and *via* analyses of genomic occurrence (62–64). A core set of genes with homologs in nearly all sporulating *Firmicutes* bacteria was outlined previously (52); curiously, many of these genes also have homologs in their non-sporulating relatives, suggesting they may be retained because of roles in other biological processes (52,62). We hypothesized that their involvement in sporulation, or lack thereof, may be reflected in the genomic neighborhoods of such genes. In our data, the ability to form spores was highly predictable from genomic clustering of commonly occurring COGs (AUC = 0.98), prompting us to examine the association of the genomic proximity between the individual COG pairs with the sporulation phenotype. This yielded 3010 significant COG pairs (Mann–Whitney test; Bonferroni-adjusted $P < 0.01$; Figure 4C). The best-supported results were enriched in COG pairs involving a known sporulation gene (51) and a ribosomal protein (RP) gene or a translation factor ($P < 2 \times 10^{-16}$, Mann–Whitney test on distribution of *P*-values; Figure 4C). RP gene operons are well-known to cluster in prokaryotes (65,66) with other genes of apparently unrelated function 'hitchhiking' in the vicinity (45). We find many sporulation gene COGs that are proximal to RPs specifically in sporulating, but not in non-sporulating bacteria. For instance, COGs with *spoVFB*, *spoIVB* and *spoIIIE* genes form statistically significant clusters with eight RPs and translation factors (Figure 4B, Supplementary Figure S2C; median FDR across all pairs <1e-6). The known sporulation gene *spoIIIJ* clusters with a tRNA modification enzyme *trmE*—which exhibits a strong sporulation knockout phenotype (51)—and with RP genes S18, L9 and L31 (Supplementary Figure S2D). These examples illustrate how gene neighborhoods can be conserved preferentially in organisms exhibiting a particu-

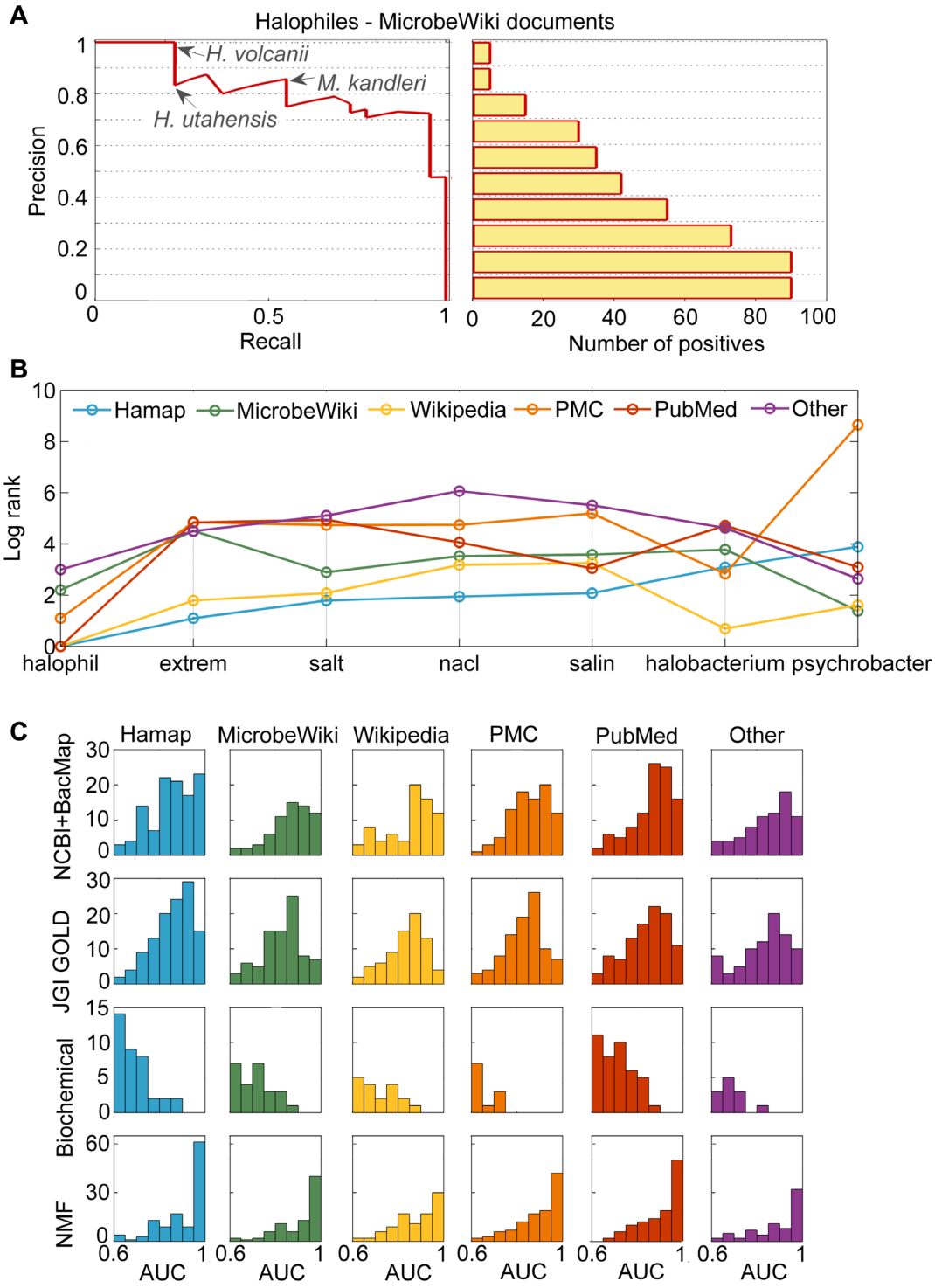


Figure 3. Annotating microbes with phenotypic traits from free text. (**A**) An example precision-recall curve for predicting halophilicity from text of MicrobeWiki pages using a Support Vector Machine (SVM) classifier (left). Arrows highlight microbes recovered at certain threshold values of precision (equivalent to 1-false discovery rate, FDR). The proportion of recovered halophilic microbes (recall) increases with more permissive precision thresholds (shown right). (**B**) The keywords important for predicting halophilicity in the six text corpora, as estimated by the SVM. (**C**) Accuracy of 2272 SVM models trained on all combinations of text corpora (columns) and phenotypic trait groups (rows), measured as the area-under-curve (AUC) score, where 1.0 is perfect performance and 0.5 indicates random guessing. Plots are histograms, showing the total number of classification models (traits) in a certain accuracy range. Data shown in all panels are from cross-validation.

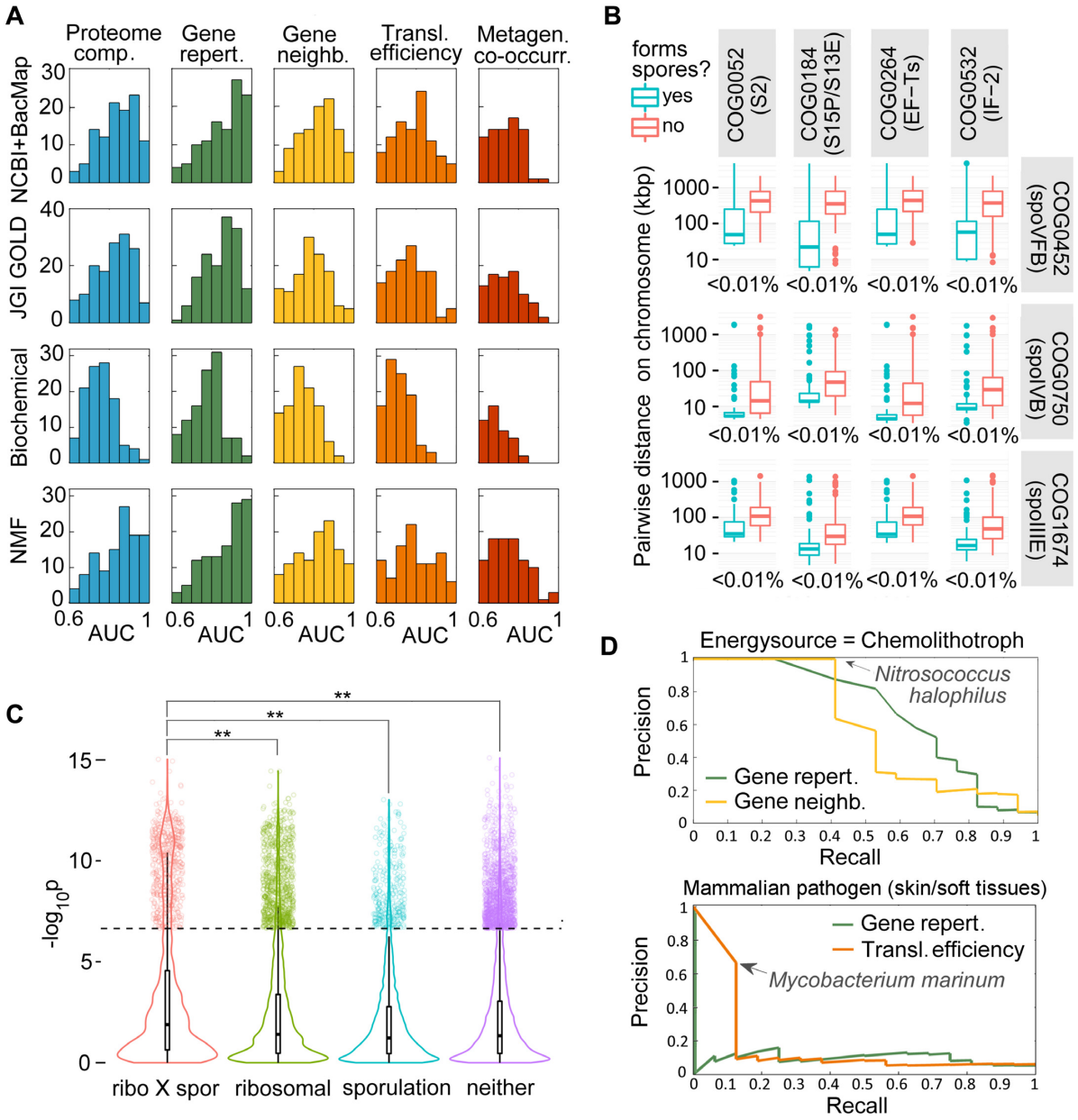


Figure 4. Inferring microbial phenotypes from comparative genomics. (**A**) Accuracy of RF classification models in predicting groups of phenotypic traits (rows) from five independent representations of genomic data (columns). This includes known methods: proteome composition (blue), gene repertoires (green) and metagenomic co-occurrence (red), as well as two novel methods: gene neighborhoods (yellow) and translation efficiency (estimated via codon usage biases; orange). Histograms show frequency of RF models of certain accuracy, estimated as the cross-validation AUC score, ranging from 0.5 (random guessing) to 1.0 (perfect performance). (**B**) Examples of gene synteny patterns associated with the spore-forming phenotype, which involve ribosomal proteins or translation factors (columns) and known sporulation factors (rows). Overlaid percentages are FDRs for significant shift in distribution of gene distances in sporulating microbes by Mann–Whitney test. (**C**) Gene neighborhoods involving a ribosomal protein and a sporulation gene are more often associated with spore-forming ability than other pairs ($P < 0.001$ for the first distribution of P-values versus each other distribution, Mann–Whitney test). Dashed line is the FDR = 1% threshold. (**D**) Precision-recall curves, in cross-validation, for two example phenotypes where the gene neighborhoods (top) and codon usage biases (bottom) have prediction accuracy that compares favorably to the gene repertoire method.

lar phenotype, conceivably due to co-regulation (45) or differential gene essentiality (67).

Systematic inference of phenotypes from codon adaptation

Next, we considered the possibility that evolution of codon usage biases within gene families may be predictive of microbial phenotypes. Highly expressed genes are known to be enriched with optimal codons across many bacterial and archaeal genomes (68,69), facilitating efficient and accurate protein translation. This 'codon adaptation' of orthologous genes may however change in evolution (70), driving phenotypic divergence (47,71). Such genomic signatures of translation efficiency within certain gene families have been used to infer the adaptive value of individual genes to various

environmental niches (6) and we thus hypothesized that the overall pattern of codon adaptation across many genes of an organism can predict its phenotype (Materials and Methods). Indeed, such translation efficiency profiles predicted phenotypes with median AUC = 0.75 (Q1–Q3: 0.67–0.81), when used as features in Random Forest models. Again, a number of models based on codon biases were similarly informative as gene repertoires (Figure 4B; example in Figure 4D), consistent with past work linking evolution of translation efficiency to a broad range of gene functions and phenotypic traits; reviewed in (18).

For each of the five comparative genomics approaches described above, we supply the sets of most important features (Materials and Methods) as Supplementary Table S4, thereby providing information about the functioning of our genome-based models. Importantly, these models draw on statistical associations that are robust and predictive on held-out data, but may not necessarily be representative of the biological mechanisms underlying particular phenotypes.

Validating the accuracy of novel annotations

In summary, we provide predictions for a comprehensive set of 424 microbial traits using six independent text corpora, and five independent sources of genomic data; available at http://protraits.irb.hr/. Of these ∼545 000 novel annotations (at FDR < 10%, including both positive and negative annotations; Supplementary Figure S3), ∼308 000 are supported in two or more independent data sources (Figure 5A). Next, we evaluated a random sample of 2489 phenotype predictions by a literature search performed by two curators. Overall, our FDR estimates (Materials and Methods) appear trustworthy, reflecting researcher judgement well, particularly when requiring agreement of any two independent predictions ('two-votes') for each phenotype (Figure 5C; Supplementary Figure S4; detailed statistics in Supplementary Table S5). Of note, combining the 11 individual predictors using the two-votes scheme consistently provided increased coverage over each individual predictor (Supplementary Figure S5), without trading off accuracy.

When examining sets of traits individually, in case of the phenotype labels obtained via the BacMap/NCBI databases, the two-votes annotations with nominal FDR < 10% will have an actual FDR of 5.8%, and similarly so for the GOLD ecosystem annotations (FDR = 7.5%; Supplementary Figure S4A). The phenotypic concepts we inferred from free text are also supported in validation, with FDR = 12.9%. These estimates are highly consistent between the two curators (Supplementary Figure S4B). Furthermore, upon breaking down the annotations by the prediction methodology, we find that the five comparative genomics methods have the actual FDR ranging from 6.6% for the proteome composition to 8.9% for the metagenome co-occurrence, and similarly so for the six text mining corpora (6.8–8.2%; Figure 5C); all values given for the nominal FDR < 10%. At more permissive thresholds that afford a broader coverage with annotations, the nominal FDRs again match the observed FDR found via curation (Figure 5C; Supplementary Figure S4; Supplementary Table S5).

Importantly, we have also applied the models to organisms that already had known labels for a particular phenotypic trait (in crossvalidation; Materials and Methods), following the rationale that some of these initial labels may be incorrect. Indeed, we found that the apparent false positives for models that predict existing NCBI/BacMap phenotypes with high confidence (<10% nominal FDR) are, in fact, in many cases genuine positives (55.1% accuracy from curation; Supplementary Figure S4C). We suggest that automated phenotype prediction methods that use free text and genomic data can be useful in highlighting misannotated phenotypes in existing databases.

Additionally, we examined the infrequent cases of annotations that differed between the source databases (listed in Supplementary Table S1), and found the predictive accuracy to be roughly comparable to the general set of annotations (Supplementary Figure S4C).

The network of co-occurring microbial phenotypic traits

The ProTraits atlas of 424 phenotypic traits annotated to thousands of taxa presents an opportunity for a global, unbiased analysis of the interrelatedness among microbial phenotypic traits. Here, we estimate the distance between any two traits as the overlap between the sets of species that display the one or the other trait, accounting for both the positive and the high-confidence negative phenotype annotations (Supplementary Methods). For instance, free-living bacteria are more often flagellated (odds ratio, OR = 4.2, 95% CI: [3.4, 5.2], $P = 10^{-44}$) while host-associated ones are associated with ability to grow on sucrose (OR = 4.8, 95% CI: [1.0, 23.5], $P = 0.045$). We further represented all such pairwise similarities as a phenotype network that describes the convergence between groups of microbial traits (Figure 6, Supplementary Figure S6; Supplementary Methods).

Clustering the network highlights well-known associations between phenotypes, such as a densely interconnected cluster of human and animal pathogenicity-related traits (Figure 6, top), or grouping of traits related to food fermentation, lactic acid production and vaginal microflora (Figure 6, right). This serves as a validation of our approach, and suggests that some less-obvious associations between phenotypes might be informative. For instance, secondary metabolite production appears strongly associated with the aerobic lifestyle (Figure 6, right; OR = 124.5, 95% CI: [17.3, 896.6], $P = 10^{-30}$), as was noted in the literature, even though the basis of this association is elusive (72,73). Moreover, the microaerophilic and the facultative aerobic phenotypes are associated with mentions of bacterial infectivity and pathogenicity (from NMF concept discovery; Figure 6, top): OR = 8.2, $P = 0.01$; and OR = 8.9, $P = 10^{-58}$, respectively, consistent with prior evidence linking the transition to microoxic conditions with expression of virulence genes (74). Of note, this phenotypic trait network pertains to the ∼3000 genomes covered by ProTraits. An important caveat is that this network may not necessarily reflect universal trends of microbial trait co-occurrence, since the representation of prokaryotic phyla among the currently sequenced set of genomes is not unbiased (15).

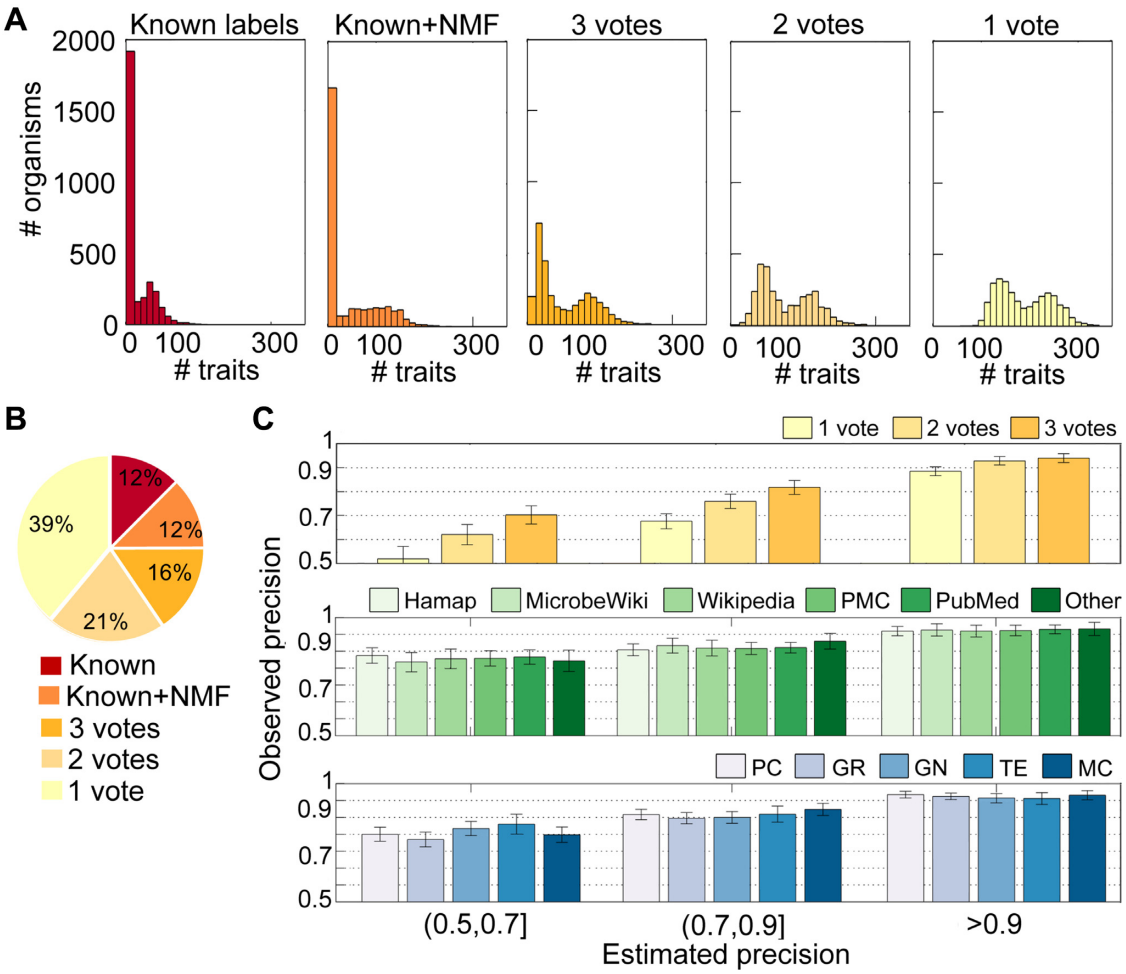


Figure 5. The number and accuracy of inferred annotations at different confidence levels. (**A**) Histograms show the number of organisms covered with traits, tallying both positive and negative labels. Shown for (from left to right): known phenotype labels, knows labels including NMF-inferred traits; using predicted annotations at different stringency levels, requiring agreement in three, two or only a single data source (including both text and genomic data sources). All predictions are at FDR < 10%. (**B**) The relative proportions of predictions accessible only with certain integration schemes, from the most stringent ("Known") to the least stringent ("1 vote"). (**C**) The precision (equivalent to 1-FDR) estimated by manual curation of 2489 predictions, obtained at different nominal precision thresholds. Bins by nominal precision on x-axis; observed precision calculated using Wilson point estimate on y-axis. Error bars are 95% C.I. (adjusted Wald). Top panel compares schemes for integrating predictions across the individual data sources; middle panel shows accuracy of the six text sources; bottom panel shows accuracy of the five genomic methods (PC stands for proteome composition, GR for gene repertoires, GN for gene neighborhoods, TE for translation efficiency and MC for metagenome co-occurrence). In all cases, predictions for microbes without previously assigned phenotypes are evaluated (see Supplementary Figure S4 for data on previously known labels, indicating putative mis-annotations); detailed statistics are in Supplementary Table S5.

Increased power to detect genes associated with traits

An important application of a comprehensive atlas of phenotypic traits would be to search for genetic underpinnings of the traits of interest. Indeed, past analyses have sought associations between patterns of presence/absence of homologous genes across microbial genomes, and certain phenotypes (3,4,75–79), while focusing on selected sets of traits ($n \leq 10$) and genomes ($n \leq 300$). Here, we extend each of these dimensions by approximately an order of magnitude, greatly increasing statistical power to detect associations, which is particularly important for rarely occurring phenotypes or gene families.

We tested for statistically significant associations of ∼80 000 COG or NOG gene families to 332 traits, while controlling for confounding effects of the phylogeny (principal components of the 16S rRNA evolutionary tree were included as covariates in logistic regression; see Supplementary Methods) and of genomic %G+C and genome size (44,80). Previously known phenotypic labels (Figure 1) can retrieve 20 348 associations across all COGs, while requiring FDR < 10% (t-test for significance of regression coefficient; Figure 7A) and OR ≥ 4 or ≤ 0.25. However, when combining the previously known with the newly annotated phenotypes, this increases approximately 6-fold, to 116 639 gene-trait associations. About half of those ($n = 57\,088$) are also supported at a stringent FDR < 1%. Of note, phenotype labels used in this test were inferred only from text sources and not from genomics (Materials and Methods). This new, broader set of associations includes, among others, 648 and 1187 very high-confidence associations of genes to pathogenicity to plants or to mammals, respec-

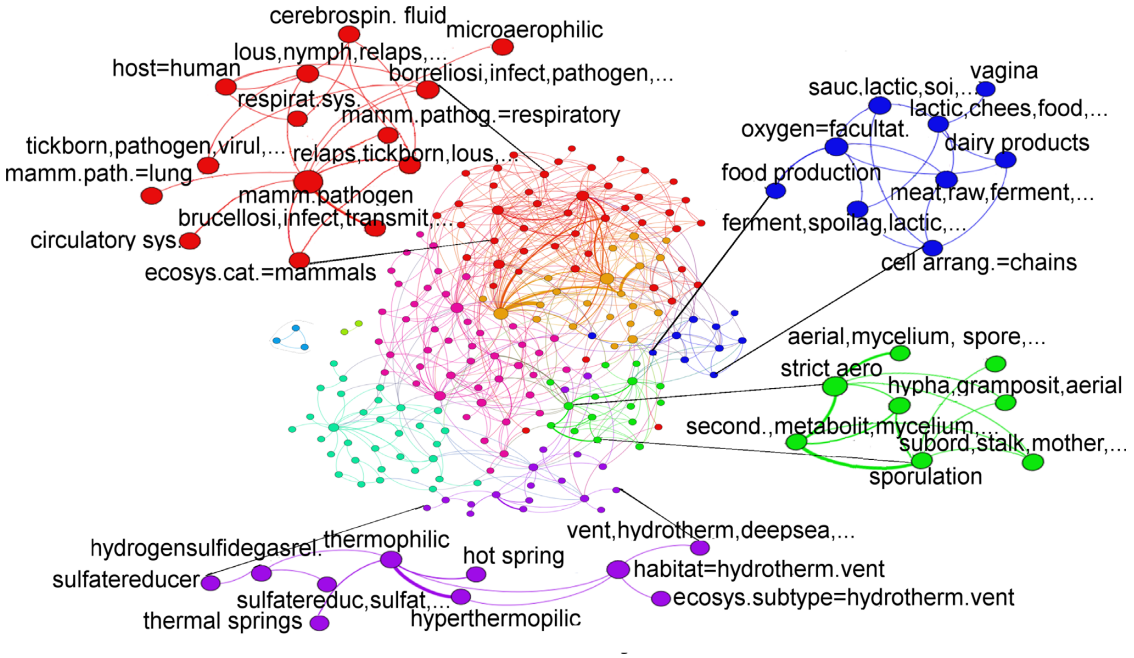


Figure 6. A network of microbial phenotypic traits. Similarity of trait pairs was estimated using the F_1-measure (shown as line thickness) that accounts for overlap in both the positive and the negative labels. A clustering was used to partition the network (colors), and selected sections are highlighted. The nodes labeled with comma-separated keywords are the NMF-inferred phenotypic concepts. In Supplementary Figure S6, the full network is shown with labels for all nodes.

tively, while original phenotypic labels would recover less than one-tenth of these links (61 and 57, respectively; at FDR $\leq$ 1%). Other traits of high general interest are also represented in the list of significant associations (Supplementary Table S6).

Overall, the ProTraits resource yields a 6.6-fold increase in the coverage of microbial gene-trait associations (at FDR $<$ 1%). This suggests that further efforts toward systematically annotating phenotypes are yet to uncover many novel genes underlying the corresponding traits. To investigate this, we simulated sets of phenotype annotations with varying coverage (Materials and Methods). The tally of highly confident gene-trait associations can be approximated by a linear fit to number of available phenotype labels (Supplementary Figure S7) and does not generally appear to saturate when considering all currently available annotations. The slope of the linear fits suggests how many genes might be linked to traits with future increases in phenotype coverage: for every additional microbial species labeled, we estimate that a median of 0.67 novel gene families will be associated with the trait (Q1–Q3: 0.18–1.06; at <1% FDR and odds ratio $\geq$ 4), thus extending our knowledge of the genetic basis of microbial phenotypes.

Validation of novel gene-trait associations

As a validation, we considered a set of genes with sporulation phenotypes in knockout strains of *B. subtilis* (51) and also a broader set of genes regulated during sporulation (52). We additionally considered a set of genes previously found to co-occur with the sporulation phenotype via phylogenetic profiling (64). Since sporulation is well-investigated experimentally, these known genes can serve to benchmark our methodology for predicting whether an organism sporulates (S) or does not sporulate (NS). In particular, if the ProTraits pipeline infers this phenotype correctly, its predictions should enhance the statistical power to recover known genes. Gene-phenotype associations using original phenotypic labels (104 S and 524 NS species) can capture 18, 24 and 10 genes from the three validation sets. This increases to 25, 34 and 13 genes (Figure 7C), respectively, upon including additional phenotype annotations predicted from text (FDR $<$ 10%; 164 S and 1156 NS). Importantly, the enrichment with known genes was not significantly different between original and extended sets of phenotype labels (P = 0.7–0.8, Z-test for difference of log OR, given for the three validation sets; Supplementary Figure S8A). This implies that the newly obtained gene-trait associations are similarly accurate as the original ones, while affording broader coverage. Next, we evaluated associations of the gene families known to encode the catalase enzymes, with the 'catalase activity' biochemical phenotype. In COG0376 (*katG*) and COG0753 (*katE*), the statistical support for the associations increased upon annotating a broader set of microbes as being catalase positive or negative (Figure 7B). Finally, we also examined the COG gene families that appear significantly associated with the 'flagellated' phenotype only upon introducing the additional annotations obtained herein, by validating against known *B. subtilis* flagellar genes (81) (Figure 7C). Again, the enrichment was similar for the original and the extended set of labels: OR = 16.1 and OR = 14.6, respectively (Supplementary Figure S8A). The novel labels allowed us to recover an additional flagellar gene (the chaperone *motE*) and moreover two flagella-related genes not given in the validation set: the regulator of flagellin synthesis *flbT* and a chemotaxis

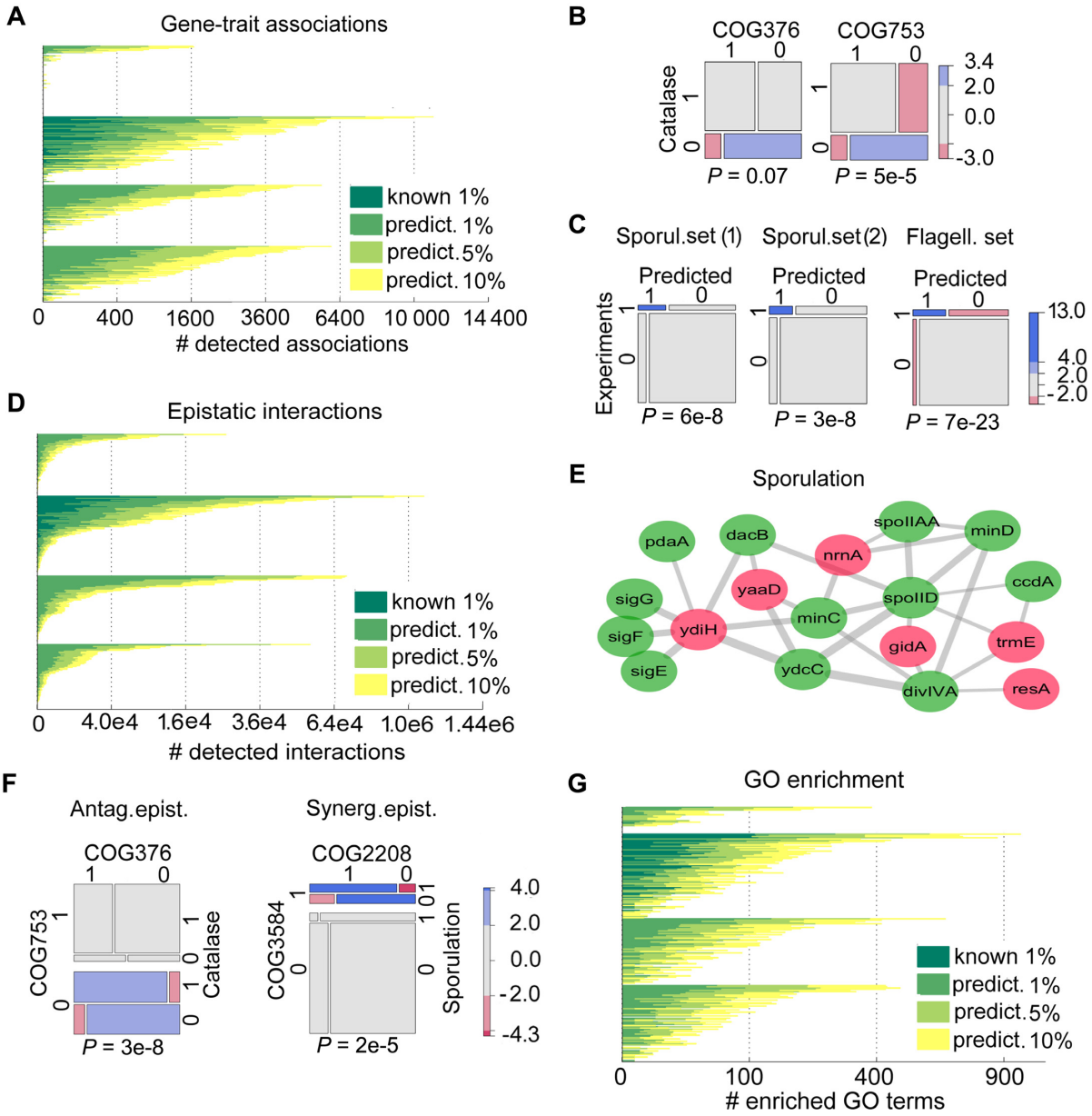


Figure 7. The extended collection of phenotypes reveals many novel associated genes. (**A**) The number of significant gene-trait associations for various traits (rows), using the previously known phenotypes (dark green) and the newly predicted phenotypes at different stringency levels (lighter shades; broken down by FDR cutoffs). (**B**) The two gene families encoding catalase genes *katG* (COG376) and *katE* (COG753) are associated with the biochemical phenotype describing catalase activity. Size of tiles in the mosaic plot is proportional to the number of organisms with (1) or without (0) a catalase homolog, and those exhibiting (1) or not exhibiting (0) catalase activity. (**C**) Validation of the predicted gene-phenotype interactions using two known sets of sporulation genes (in *B. subtilis*) and a known set of flagellar genes. Mosaic plots as above. (**D**) Same as in (A), but counting the number of epistatic interactions involving pairs of genes and a phenotype. (**E**) A network of epistatic interactions involving the sporulating phenotype, known *B. subtilis* sporulation genes (green) and novel sporulation genes (red) that were recently discovered in a *B. subtilis* genetic screen (51). Edges represent synergistic epistatic interactions, in all cases FDR < 1% and $OR_{inter} > 9.2$. Edge thickness denotes log OR. (**F**) Examples of an antagonistic epistatic interaction between the pair of catalase genes *katG* (COG376) and *katE* (COG753) and the 'catalase activity' phenotype (left), and a synergistic epistatic interaction between the pair of sporulation genes *spoIIE* (COG3854) and *spoIIIAA* (COG2208) and the sporulation phenotype (right). Mosaic plots as above. (**G**) Same as in (A), but counting the number of significantly enriched Gene Ontology (GO) terms for various traits.

signal transduction gene *cheF* (82). Using only the original set of phenotypic labels, these genes would not have been retrieved.

Next, the genes we linked to individual phenotypic traits were examined for consistency of biological function. For instance, the trait 'strictly aerobic' has phenotypic annotations that we found to be associated with 1187 gene families (t-test on logistic regression coefficient, FDR < 10%), which are enriched in the Gene Ontology (GO) terms *aerobic respiration* (FDR = 6×10^{-5}, Fisher's exact test) and the more specific functions *tricarboxylic acid cycle* (0.3%) and *electron transport chain* (3%). Next, the gene families associated with the label 'energy source = photosynthetic' in our extended set of annotations are enriched in the GO term *carbon fixation* (FDR = 6%). These expected functional enrichments validate our general approach. Importantly, using previously known labels, the linked genes were enriched in average of 14.7 GO categories per trait (FDR < 10% by Fisher's exact test), increasing to 22.7 GO categories with the extended set of phenotypic labels (Figure 7G). In other words, many previously inaccessible trait-GO associations were revealed. We highlight some examples in Table 1, and moreover provide a comprehensive set of 8180 statistically supported links between microbial phenotypic traits and gene functions or pathways (Supplementary Table S6), which may boost further efforts to elucidate the mechanisms underlying particular microbial phenotypes.

Widespread epistasis in microbial gene repertoires

Evolutionary innovation in prokaryotes often occurs by gene losses and by gains *via* horizontal transfer, which are subject to constraints wherein the ability to gain a gene depends on the prior gene content (83). We thus examine genetic interactions involving presence/absence patterns of gene families across genomes, focusing on those which become apparent only in the context of a particular phenotypic trait. This bears a certain analogy to large-scale yeast experiments that systematically examined phenotypes of double gene knockouts (84), but instead relies on comparative genomics to gauge the relevance of epistasis for complex traits (85). Crucially, our extended set of annotations for many microorganisms affords more statistical power to search for gene interactions associated with diverse traits. We focus on pairs of 2663 commonly occurring COG/NOG gene families, in relation to traits defined above, while controlling for confounding effects of phylogeny (covariates in logistic regression; Supplementary Methods). Overall, we find that epistatic interactions are pervasive in prokaryotic gene repertoires and appear broadly associated with microbial traits. In particular, we find 3.9×10^6 significant three-way (gene-gene-trait) interactions (at FDR < 1%) after having included our phenotypic annotations found by text mining; this is a 6.8-fold increase over the coverage that could be obtained with original annotations only (Figure 7D). Many of the tested traits (76%, 254 out of 332) have at least one significant epistatic interaction associated with them, and 230 have five or more interactions (Supplementary Figure S8B). Conversely, almost all (99.4%) of the tested COGs are involved in an epistatic interaction with respect to at least one trait. Importantly, the phenotypic concepts we inferred from free text *de novo* also provided a context for the genetic interactions: 53 of 73 tested NMF concepts (Materials and Methods) have ≥5 interactions; see above. This suggests that the bottleneck in determining the genetic underpinnings of microbial traits is not only the number of microbes annotated with common phenotypes, but also the incompleteness of our current dictionary of phenotypic traits.

Epistatic interactions associate genes with diverse phenotypes

Of the detected genetic interactions, 56% are instances of antagonistic epistasis, where the phenotype is associated with gene mutual exclusivity, while the remaining 44% are synergistic epistasis, here meaning that the phenotype is associated with gene co-occurrence. An example of the former are the gene families of the catalases *katG* and *katE*, where their joint occurrence is less strongly associated with the catalase phenotype than expected from individual effects (Figure 7F; $P = 2 \times 10^{-5}$, t-test for significance of regression coefficient of gene interaction term). This can be interpreted as either one or the other COG being sufficient to attain a measurable catalase activity, with little additional benefit in having both. Next, many prominent examples of synergistic epistasis were observed among sporulation genes, where, for instance, the *spoIIE* and *spoIIIAA* gene families are jointly associated with the spore-forming phenotype much more strongly than expected from their individual effects (Figure 7F, $P = 3 \times 10^{-8}$, t-test on interaction term; logistic regression OR = 6.8 and 3.7 for individual COGs and OR_{inter} = 45.2 for the interaction term).

Such synergistic relationships suggest that the two proteins may function as subsequent steps in the same molecular pathway. We found further instances of epistatic interactions that involve one gene known to be involved in sporulation, and other genes where such a role was not previously described. In particular, the gene *minC* has a known role in septum formation during *B. subtilis* sporulation, and it is strongly linked with, among others, the gene families containing the transcription factor *ydiH* (or *rex*), the pyridoxine biosynthesis enzyme *yaaD* (or *pdxS*) and the oligoribonuclease *nrnA*; in all three cases, $P < 5 \times 10^{-5}$ and $OR_{inter} > 28$. A recent transposon mutagenesis screen in *B. subtilis* has indeed found *ydiH*, *yaaD* and *nrnA* to exhibit very strong sporulation phenotypes (51). We also detected epistatic relationships between well-known sporulation genes and three other novel genes (Figure 7E) that were validated in experimental data (51). These examples suggest that signatures of epistasis in gene repertoires may be broadly useful in assigning novel roles to poorly characterized prokaryotic genes, similarly as large-scale experimental genetic interaction screens have proven to be for model eukaryotes (86,87). Moreover, the comparative genomics approach can in principle screen many phenotypes at once, thus highlighting functional links between genes and additionally associating the linked genes to a particular phenotype.

DISCUSSION

There is an unmet need for automated methods that systematically collect and compute over phenotypic data (1,88).

Table 1. Example GO terms significantly enriched with phenotype-associated gene families

Phenotype	GO term	FDR original	FDR inferred	OR original	OR inferred
gelatin hydrolysis	peptidase activity	n.s.	5%	n.s.	86.4
iron-reducer	iron ion binding	22%	1%	12.5	11.2
halophilic	sodium ion transport	n.s.	7×10^{-5}	n.s.	38.9
energy source = photosynthetic	pigment biosynthesis	20%	1%	5.0	9.3
pathogenic in mammals	pathogenesis	29%	3%	4.7	8.6
lactic/cheese/food/ferment/milk	galactose metabolic process	NA	1%	NA	52.6
legum/symbiosis/rhizobia/nod/stem	nitrogen fixation	NA	3×10^{-6}	NA	1362.6

Association of genes to phenotypes is determined by a *t*-test on the logistic regression coefficient, requiring FDR < 10%. Table shows the examples in which the GO-phenotype association was not significant in the sets of gene families retrieved using the original set of phenotypic labels, but became significant as the inferred phenotypic labels were also considered. The FDR column is by Fisher's exact test for COG-GO association by Fisher's exact test, one-tailed. 'OR', odds ratio. 'n.s.' denotes associations that did not have statistically significant gene-trait associations; Supplementary Methods). 'NA' denotes that NMF phenotypic concepts were not available in the original data.

Microbes represent a particular challenge in that respect, given the staggering diversity of traits observed across the bacterial and archaeal domains of life. A wealth of phenotype information is contained within the text of scientific articles and other online sources, which is not directly amenable to statistical analysis – implying, most prominently, elucidating the genetic determinants of various traits. Indeed, mining of free-text shows great promise for establishing gene-trait associations, as previously demonstrated by linking individual gene families to PubMed keywords (56). We aimed to further exploit these free-text resources by employing machine learning to accurately assign known phenotypes to novel microbes, as well as to infer phenotypic concepts *de novo* from free text, while focusing on robust trends that replicate across individual text corpora.

In addition to the abundance of text, a further opportunity for systematic inference of microbial phenotypes is presented by the availability of many complete genomes, which can provide independent support for the text-based predictions. An example is the Genome Properties system (89,90), which applies curated rules to detect activity of biochemical pathways or other molecular subsystems, based on the occurrence of critical gene families. The ability of such rule-based systems to serve as accurate predictors for certain (mechanistically well-understood) traits provides the rationale for more generally attempting to predict a broader spectrum of phenotypes from gene content. To this end, we have employed a general-purpose statistical learning method, the Random Forest. The algorithm chooses the combination of gene families that best explains a particular phenotype, in terms of high statistical support, and uses that combination to annotate the phenotype—or absence thereof—in novel genomes. Since this does not require human input, it scales well with the increase in number of phenotypic traits and sequenced genomes. The ProTraits pipeline can thus generate new predictive models in an automated manner, not requiring expert knowledge on the genetic basis of the phenotypes, which is still lacking for many complex traits.

Further generalizing this well-known gene content-based methodology, our work demonstrates how other comparative genomics approaches normally used to predict gene function—here, synteny patterns (58,59) and codon adaptation (6,70)—can be efficiently repurposed into phenotype predictors. Here, given that the resulting models may be complex and thus challenging to interpret, it is critical to ensure that the predictions they provide are trustworthy and that they robustly classify novel data. We systematically gauged the accuracy of the FDR estimates provided for the inferences in ProTraits by manually evaluating a large sample of the predictions by literature review (Figure 5C; Supplementary Figure S4, Supplementary Table S5). Since the supplied confidence estimates have a probabilistic interpretation which we have validated, they enable the users of the ProTraits resource to make informed decisions about how best to use this massive data set in their work.

We provide a summary of the ∼545 000 phenotypic labels assigned to 3046 microbial species *via* the ProTraits pipeline by outlining the structure of the microbial phenotype co-occurrence network. Furthermore, we perform a systematic search for the genetic underpinnings of various prokaryotic traits. The ∼57 000 significant gene family-trait links represent a 6.6-fold increase over the genes that could be implicated by using only the previously available databases. Our data suggest that the availability of phenotypic labels is, in many cases, still limiting for elucidating the genetic basis of the traits, which commonly involves epistatic interactions between genes. Thus, it is imperative to direct further effort toward systematically annotating microbial phenotypes, thereby matching the strides made toward systematic sequencing of prokaryotic genomes (15). We anticipate that future developments in the natural language processing methods that can annotate semi-structured or free-text (91–94) will make important contributions towards organizing the legacy phenotype data scattered throughout the scientific literature into structured, computable formats.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

Authors contributions: M.B. collected and processed the data sets and performed all statistical analyses. M.P. developed the ProTraits Web application. V.V. contributed genomic data. A.K. and F.S. curated phenotypic predictions for validation. F.S. conceived and supervised the project.

T.S. supervised the project. F.S. and M.B. interpreted the data and wrote the manuscript.

FUNDING

This work was funded by the Croatian Science Foundation grants HRZZ-9623 (DescriptiveInduction) and HRZZ-5660 (MultiCaST) and by the European Union FP7 grants ICT-2013-612944 (MAESTRA) and REGPOT-2012-2013-1-316289 (InnoMol). FS acknowledges the support of the Spanish Ministry of Economy and Competitiveness, 'Centro de Excelencia Severo Ochoa 2013-2017' (SEV-2012-0208) and the FP7 project 4DCellFate (277899). Funding for open access charge: EU FP7 FET ICT-2013-612944 (MAESTRA).
Conflict of interest statement. None declared.

REFERENCES

1. Dutilh,B.E., Backus,L., Edwards,R.A., Wels,M., Bayjanov,J.R. and van Hijum,S.A.F.T. (2013) Explaining microbial phenotypes on a genomic scale: GWAS for microbes. *Brief. Funct. Genomics*, **12**, 366–380.
2. Read,T.D. and Massey,R.C. (2014) Characterizing the genetic basis of bacterial phenotypes using genome-wide association studies: a new direction for bacteriology. *Genome Med.*, **6**, 109.
3. Liu,Y., Li,J., Sam,L., Goh,C.-S., Gerstein,M. and Lussier,Y.A. (2006) An integrative genomic approach to uncover molecular mechanisms of prokaryotic traits. *PLOS Comput. Biol.*, **2**, e159.
4. Slonim,N., Elemento,O. and Tavazoie,S. (2006) Ab initio genotype–phenotype association reveals intrinsic modularity in genetic networks. *Mol. Syst. Biol.*, **2**, 2006.0005.
5. Singh,A.H., Wolf,D.M., Wang,P. and Arkin,A.P. (2008) Modularity of stress response evolution. *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 7500–7505.
6. Krisko,A., Copic,T., Gabaldón,T., Lehner,B. and Supek,F. (2014) Inferring gene function from evolutionary change in signatures of translation efficiency. *Genome Biol.*, **15**, R44.
7. Sheppard,S.K., Didelot,X., Meric,G., Torralbo,A., Jolley,K.A., Kelly,D.J., Bentley,S.D., Maiden,M.C.J., Parkhill,J. and Falush,D. (2013) Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in Campylobacter. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 11923–11927.
8. Chaston,J.M., Newell,P.D. and Douglas,A.E. (2014) Metagenome-Wide Association of Microbial Determinants of Host Phenotype in Drosophila melanogaster. *mBio*, **5**, e01631–e01614.
9. Holt,K.E., Wertheim,H., Zadoks,R.N., Baker,S., Whitehouse,C.A., Dance,D., Jenney,A., Connor,T.R., Hsu,L.Y., Severin,J. *et al.* (2015) Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in Klebsiella pneumoniae, an urgent threat to public health. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, E3574–E3581.
10. Vidovic,A., Supek,F., Nikolic,A. and Krisko,A. (2014) Signatures of conformational stability and oxidation resistance in proteomes of pathogenic bacteria. *Cell Rep.*, **7**, 1393–1400.
11. Salipante,S.J., Roach,D.J., Kitzman,J.O., Snyder,M.W., Stackhouse,B., Butler-Wu,S.M., Lee,C., Cookson,B.T. and Shendure,J. (2015) Large-scale genomic sequencing of extraintestinal pathogenic escherichia coli strains. *Genome Res.*, **25**, 119–128.
12. Baines,S.L., Holt,K.E., Schultz,M.B., Seemann,T., Howden,B.O., Jensen,S.O., van Hal,S.J., Coombs,G.W., Firth,N., Powell,D.R. *et al.* (2015) Convergent adaptation in the dominant global hospital clone ST239 of methicillin-resistant staphylococcus aureus. *mBio*, **6**, e00080–e00015.
13. Bayjanov,J.R., Molenaar,D., Tzeneva,V., Siezen,R.J. and van Hijum,S.A.F.T. (2012) PhenoLink - a web-tool for linking phenotype to ~omics data for bacteria: application to gene-trait matching for Lactobacillus plantarum strains. *BMC Genomics*, **13**, 170.
14. Konietzny,S.G., Pope,P.B., Weimann,A. and McHardy,A.C. (2014) Inference of phenotype-defining functional modules of protein families for microbial plant biomass degraders. *Biotechnol. Biofuels*, **7**, 124.
15. Kyrpides,N.C., Hugenholtz,P., Eisen,J.A., Woyke,T., Göker,M., Parker,C.T., Amann,R., Beck,B.J., Chain,P.S.G., Chun,J. *et al.* (2014) Genomic encyclopedia of bacteria and archaea: sequencing a myriad of type strains. *PLOS Biol.*, **12**, e1001920.
16. Rinke,C., Schwientek,P., Sczyrba,A., Ivanova,N.N., Anderson,I.J., Cheng,J.-F., Darling,A., Malfatti,S., Swan,B.K., Gies,E.A. *et al.* (2013) Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, **499**, 431–437.
17. Albertsen,M., Hugenholtz,P., Skarshewski,A., Nielsen,K.L., Tyson,G.W. and Nielsen,P.H. (2013) Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. Biotechnol.*, **31**, 533–538.
18. Supek,F. (2015) The code of silence: widespread associations between synonymous codon biases and gene function. *J. Mol. Evol.*, **82**, 65–73.
19. Kim,S. and Xing,E.P. (2009) Statistical estimation of correlated genome associations to a quantitative trait network. *PLOS Genet.*, **5**, e1000587.
20. Casale,F.P., Rakitsch,B., Lippert,C. and Stegle,O. (2015) Efficient set tests for the genetic analysis of correlated traits. *Nat. Methods*, **12**, 755–758.
21. Bulik-Sullivan,B., Finucane,H.K., Anttila,V., Gusev,A., Day,F.R., Loh,P.-R., Duncan,L. and ReproGen Consortium, Psychiatric Genomics Consortium, Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium 3 et al. (2015) An atlas of genetic correlations across human diseases and traits. *Nat. Genet.*, **47**, 1236–1241.
22. Groth,P., Pavlova,N., Kalev,I., Tonov,S., Georgiev,G., Pohlenz,H.-D. and Weiss,B. (2007) PhenomicDB: a new cross-species genotype/phenotype resource. *Nucleic Acids Res.*, **35**, D696–D699.
23. Sam,L.T., Mendonça,E.A., Li,J., Blake,J., Friedman,C. and Lussier,Y.A. (2009) PhenoGO: an integrated resource for the multiscale mining of clinical and biological data. *BMC Bioinformatics*, **10**, 1–8.
24. Stothard,P., Domselaar,G.V., Shrivastava,S., Guo,A., O'Neill,B., Cruz,J., Ellison,M. and Wishart,D.S. (2005) BacMap: an interactive picture atlas of annotated bacterial genomes. *Nucleic Acids Res.*, **33**, D317–D320.
25. Reddy,T.B.K., Thomas,A.D., Stamatis,D., Bertsch,J., Isbandi,M., Jansson,J., Mallajosyula,J., Pagani,I., Lobos,E.A. and Kyrpides,N.C. (2015) The genomes OnLine database (GOLD) v.5: a metadata management system based on a four level (meta)genome project classification. *Nucleic Acids Res.*, **43**, D1099–D1106.
26. Oren,A. (2015) 70th anniversary collection for the microbiology society: international journal of systematic and evolutionary microbiology. *Int. J. Syst. Evol. Microbiol.*, **65**, 4291–4293.
27. Plata,G., Henry,C.S. and Vitkup,D. (2015) Long-term phenotypic evolution of bacteria. *Nature*, **517**, 369–372.
28. Barua,D., Kim,J. and Reed,J.L. (2010) An automated phenotype-driven approach (Gene Force) for refining metabolic and regulatory models. *PLOS Comput. Biol.*, **6**, e1000970.
29. Google Code Archive - Collection of stop words in 29 languages. Available at: https://code.google.com/archive/p/stop-words/.
30. Porter,M.F. (2006) An algorithm for suffix stripping. *Program*, **40**, 211–218.
31. Lee,D.D. and Seung,H.S. (1999) Learning the parts of objects by non-negative matrix factorization. *Nature*, **401**, 788–791.
32. Arora,S., Ge,R. and Moitra,A. (2012) Learning topic models – going beyond SVD. In: *Proceedings of the 2012 IEEE 53rd Annual Symposium on Foundations of Computer Science, FOCS '12*. IEEE Computer Society, Washington, DC, pp. 1–10.
33. Nielsen,F.A., Balslev,D. and Hansen,L.K. (2005) Mining the posterior cingulate: segregation between memory and pain components. *NeuroImage*, **27**, 520–532.
34. Salton,G., Fox,E.A. and Wu,H. (1983) Extended Boolean Information Retrieval. *Commun. ACM*, **26**, 1022–1036.
35. Kim,J. and Park,H. (2008) Toward faster nonnegative matrix factorization: a new algorithm and comparisons. In: *2008 Eighth IEEE International Conference on Data Mining*. pp. 353–362.
36. Cortes,C. and Vapnik,V. (1995) Support-vector networks. *Mach. Learn.*, **20**, 273–297.
37. Chang,C.-C. and Lin,C.-J. (2011) LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, **2**, 1–27.
38. Powell,S., Szklarczyk,D., Trachana,K., Roth,A., Kuhn,M., Muller,J., Arnold,R., Rattei,T., Letunic,I., Doerks,T. *et al.* (2012) eggNOG

v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Res.*, **40**, D284–D289.
39. Land,M.L., Hyatt,D., Jun,S.-R., Kora,G.H., Hauser,L.J., Lukjancenko,O. and Ussery,D.W. (2014) Quality scores for 32,000 genomes. *Stand. Genomic Sci.*, **9**, 20.
40. Brbić,M., Warnecke,T., Kriško,A. and Supek,F. (2015) Global shifts in genome and proteome composition are very tightly coupled. *Genome Biol. Evol.*, **7**, 1519–1532.
41. Smole,Z., Nikolic,N., Supek,F., Šmuc,T., Sbalzarini,I.F. and Krisko,A. (2011) Proteome sequence features carry signatures of the environmental niche of prokaryotes. *BMC Evol. Biol.*, **11**, 26.
42. MacDonald,N.J. and Beiko,R.G. (2010) Efficient learning of microbial genotype–phenotype association rules. *Bioinformatics*, **26**, 1834–1840.
43. Feldbauer,R., Schulz,F., Horn,M. and Rattei,T. (2015) Prediction of microbial phenotypes based on comparative genomics. *BMC Bioinformatics*, **16**, 1–8.
44. Chaffron,S., Rehrauer,H., Pernthaler,J. and von Mering,C. (2010) A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome Res.*, **20**, 947–959.
45. Rogozin,I.B., Makarova,K.S., Murvai,J., Czabarka,E., Wolf,Y.I., Tatusov,R.L., Szekely,L.A. and Koonin,E.V. (2002) Connected gene neighborhoods in prokaryotic genomes. *Nucleic Acids Res.*, **30**, 2212–2223.
46. Karpinets,T.V., Obraztsova,A.Y., Wang,Y., Schmoyer,D.D., Kora,G.H., Park,B.H., Serres,M.H., Romine,M.F., Land,M.L., Kothe,T.B. *et al.* (2010) Conserved synteny at the protein family level reveals genes underlying shewanella species' cold tolerance and predicts their novel phenotypes. *Funct. Integr. Genomics*, **10**, 97–110.
47. Man,O. and Pilpel,Y. (2007) Differential translation efficiency of orthologous genes is involved in phenotypic divergence of yeast species. *Nat. Genet.*, **39**, 415–421.
48. Supek,F. and Vlahovicek,K. (2005) Comparison of codon usage measures and their applicability in prediction of microbial gene expressivity. *BMC Bioinformatics*, **6**, 182.
49. Breiman,L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.
50. Supek,F. (2015) The FastRandomForest Weka Extension. http://fast-random-forest.googlecode.com.
51. Meeske,A.J., Rodrigues,C.D.A., Brady,J., Lim,H.C., Bernhardt,T.G. and Rudner,D.Z. (2016) High-throughput genetic screens identify a large and diverse collection of new sporulation genes in bacillus subtilis. *PLOS Biol.*, **14**, e1002341.
52. Galperin,M.Y., Mekhedov,S.L., Puigbo,P., Smirnov,S., Wolf,Y.I. and Rigden,D.J. (2012) Genomic determinants of sporulation in bacilli and clostridia: towards the minimal set of sporulation-specific genes. *Environ. Microbiol.*, **14**, 2870–2890.
53. Austin,P.C. (2011) An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivar. Behav. Res.*, **46**, 399–424.
54. Lima,T., Auchincloss,A.H., Coudert,E., Keller,G., Michoud,K., Rivoire,C., Bulliard,V., de Castro,E., Lachaize,C., Baratin,D. *et al.* (2009) HAMAP: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot. *Nucleic Acids Res.*, **37**, D471–D478.
55. Kuang,D., Choo,J. and Park,H. (2015) Nonnegative matrix factorization for interactive topic modeling and document clustering. In: Celebi,ME (ed). *Partitional Clustering Algorithms*. Springer International Publishing, Cham, pp.215–243.
56. Korbel,J.O., Doerks,T., Jensen,L.J., Perez-Iratxeta,C., Kaczanowski,S., Hooper,S.D., Andrade,M.A. and Bork,P. (2005) Systematic association of genes to phenotypes by genome and literature mining. *PLOS Biol.*, **3**, e134.
57. Davis,J. and Goadrich,M. (2006) The Relationship Between Precision-Recall and ROC Curves. In: *Proceedings of the 23rd International Conference on Machine Learning, ICML '06*. ACM, NY, pp. 233–240.
58. Wolf,Y.I., Rogozin,I.B., Kondrashov,A.S. and Koonin,E.V. (2001) Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. *Genome Res.*, **11**, 356–372.
59. Overbeek,R., Fonstein,M., D'Souza,M., Pusch,G.D. and Maltsev,N. (1999) The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. U.S.A.*, **96**, 2896–2901.
60. Steil,L., Serrano,M., Henriques,A.O. and Völker,U. (2005) Genome-wide analysis of temporally regulated and compartment-specific gene expression in sporulating cells of Bacillus subtilis. *Microbiology*, **151**, 399–420.
61. Arrieta-Ortiz,M.L., Hafemeister,C., Bate,A.R., Chu,T., Greenfield,A., Shuster,B., Barry,S.N., Gallitto,M., Liu,B., Kacmarczyk,T. *et al.* (2015) An experimentally supported model of the Bacillus subtilis global transcriptional regulatory network. *Mol. Syst. Biol.*, **11**, 839–839.
62. Rigden,D.J. and Galperin,M.Y. (2008) Sequence analysis of GerM and SpoVS, uncharacterized bacterial 'sporulation' proteins with widespread phylogenetic distribution. *Bioinformatics*, **24**, 1793–1797.
63. Browne,H.P., Forster,S.C., Anonye,B.O., Kumar,N., Neville,B.A., Stares,M.D., Goulding,D. and Lawley,T.D. (2016) Culturing of 'unculturable' human microbiota reveals novel taxa and extensive sporulation. *Nature*, **533**, 543–546.
64. Traag,B.A., Pugliese,A., Eisen,J.A. and Losick,R. (2013) Gene conservation among endospore-forming bacteria reveals additional sporulation genes in Bacillus subtilis. *J. Bacteriol.*, **195**, 253–260.
65. Lathe,W.C. III, Snel,B. and Bork,P. (2000) Gene context conservation of a higher order than operons. *Trends Biochem. Sci.*, **25**, 474–479.
66. Yang,Q. and Sze,S.-H. (2008) Large-scale analysis of gene clustering in bacteria. *Genome Res.*, **18**, 949–956.
67. Fang,G., Rocha,E.P. and Danchin,A. (2008) Persistence drives gene clustering in bacterial genomes. *BMC Genomics*, **9**, 4.
68. Hershberg,R. and Petrov,D.A. (2009) General rules for optimal codon choice. *PLOS Genet.*, **5**, e1000556.
69. Supek,F., Škunca,N., Repar,J., Vlahoviček,K. and Šmuc,T. (2010) Translational selection is ubiquitous in prokaryotes. *PLOS Genet*, **6**, e1001004.
70. Fraser,H.B., Hirsh,A.E., Wall,D.P. and Eisen,M.B. (2004) Coevolution of gene expression among interacting proteins. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 9033–9038.
71. Karlin,S., Brocchieri,L., Campbell,A., Cyert,M. and Mrázek,J. (2005) Genomic and proteomic comparisons between bacterial and archaeal genomes and related comparisons with the yeast and fly genomes. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 7309–7314.
72. Lincke,T., Behnken,S., Ishida,K., Roth,M. and Hertweck,C. (2010) Closthioamide: An unprecedented polythioamide antibiotic from the strictly anaerobic bacterium clostridium cellulolyticum. *Angew. Chem. Int. Ed.*, **49**, 2011–2013.
73. Pidot,S., Ishida,K., Cyrulies,M. and Hertweck,C. (2014) Discovery of clostrubin, an exceptional polyphenolic polyketide antibiotic from a strictly anaerobic bacterium. *Angew. Chem. Int. Ed.*, **53**, 7856–7859.
74. Morris,R.L. and Schmidt,T.M. (2013) Shallow breathing: bacterial life at low O2. *Nat. Rev. Microbiol.*, **11**, 205–212.
75. Levesque,M., Shasha,D., Kim,W., Surette,M.G. and Benfey,P.N. (2003) Trait-to-Gene: A Computational Method for Predicting the Function of Uncharacterized Genes. *Curr. Biol.*, **13**, 129–133.
76. Goh,C.-S., Gianoulis,T.A., Liu,Y., Li,J., Paccanaro,A., Lussier,Y.A. and Gerstein,M. (2006) Integration of curated databases to identify genotype-phenotype associations. *BMC Genomics*, **7**, 257.
77. Tamura,M. and D'haeseleer,P. (2008) Microbial genotype–phenotype mapping by class association rule mining. *Bioinformatics*, **24**, 1523–1529.
78. Gonzalez,O. and Zimmer,R. (2008) Assigning functional linkages to proteins using phylogenetic profiles and continuous phenotypes. *Bioinformatics*, **24**, 1257–1263.
79. Kastenmüller,G., Schenk,M.E., Gasteiger,J. and Mewes,H.-W. (2009) Uncovering metabolic pathways relevant to phenotypic traits of microbial genomes. *Genome Biol.*, **10**, R28.
80. Konstantinidis,K.T. and Tiedje,J.M. (2004) Trends between gene content and genome size in prokaryotic species with larger genomes. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 3160–3165.
81. Rajagopala,S.V., Titz,B., Goll,J., Parrish,J.R., Wohlbold,K., McKevitt,M.T., Palzkill,T., Mori,H., Finley,R.L. and Uetz,P. (2007) The protein network of bacterial motility. *Mol. Syst. Biol.*, **3**, 128.
82. Schlesner,M., Miller,A., Streif,S., Staudinger,W.F., Müller,J., Scheffer,B., Siedler,F. and Oesterhelt,D. (2009) Identification of Archaea-specific chemotaxis proteins which interact with the flagellar apparatus. *BMC Microbiol.*, **9**, 56.
83. Press,M.O., Queitsch,C. and Borenstein,E. (2016) Evolutionary assembly patterns of prokaryotic genomes. *Genome Res.*, **26**, 826–833.

84. Costanzo,M., Baryshnikova,A., Bellay,J., Kim,Y., Spear,E.D., Sevier,C.S., Ding,H., Koh,J.L.Y., Toufighi,K., Mostafavi,S. *et al.* (2010) The genetic landscape of a cell. *Science*, **327**, 425–431.
85. Mackay,T.F.C. (2014) Epistasis and quantitative traits: Using model organisms to study gene-gene interactions. *Nat. Rev. Genet.*, **15**, 22–33.
86. Tong,A.H.Y. (2004) Global mapping of the yeast genetic interaction network. *Science*, **303**, 808–813.
87. Fievet,B.T., Rodriguez,J., Naganathan,S., Lee,C., Zeiser,E., Ishidate,T., Shirayama,M., Grill,S. and Ahringer,J. (2013) Systematic genetic interaction screens uncover cell polarity regulators and functional redundancy. *Nat. Cell Biol.*, **15**, 103–112.
88. Deans,A.R., Lewis,S.E., Huala,E., Anzaldo,S.S., Ashburner,M., Balhoff,J.P., Blackburn,D.C., Blake,J.A., Burleigh,J.G., Chanet,B. *et al.* (2015) Finding our way through phenotypes. *PLOS Biol.*, **13**, e1002033.
89. Selengut,J.D., Haft,D.H., Davidsen,T., Ganapathy,A., Gwinn-Giglio,M., Nelson,W.C., Richter,A.R. and White,O. (2007) TIGRFAMs and Genome Properties: tools for the assignment of molecular function and biological process in prokaryotic genomes. *Nucleic Acids Res.*, **35**, D260–D264.
90. Haft,D.H., Selengut,J.D., Brinkac,L.M., Zafar,N. and White,O. (2005) Genome Properties: a system for the investigation of prokaryotic genetic content for microbiology, genome annotation and comparative genomics. *Bioinformatics*, **21**, 293–306.
91. Pajić,V.S., Lažetić,G.M.P., Beljanski,M.V., Brandt,B.W. and Pajić,M.B. (2013) Towards a database for genotype-phenotype association research: mining data from encyclopaedia. *Int. J. Data Min. Bioinform.*, **7**, 196.
92. Ratkovic,Z., Golik,W. and Warnier,P. (2012) Event extraction of bacteria biotopes: a knowledge-intensive NLP-based approach. *BMC Bioinformatics*, **13**, 1–11.
93. Bossy,R., Golik,W., Ratkovic,Z., Valsamou,D., Bessières,P. and Nédellec,C. (2015) Overview of the gene regulation network and the bacteria biotope tasks in BioNLP'13 shared task. *BMC Bioinformatics*, **16**, 1–16.
94. Ananiadou,S., Thompson,P., Nawaz,R., McNaught,J. and Kell,D.B. (2015) Event-based text mining for biology and functional genomics. *Brief. Funct. Genomics*, **14**, 213–230.

 doi:10.1093/nar/gkw964

SUPPLEMENTARY METHODS

Performing multiple runs of NMF text analysis

Since NMF has a stochastic component and can therefore yield different solutions depending on the initial values, we ran the 50-factor NMF five times, and additionally the 100-factor NMF three times, with a different random seed, in order to maximize the coverage with discovered phenotypic concepts. We also performed another run of NMF with 100 factors for the matrix in which we allowed word to be absent in one of the five text corpora, which allowed a broader set of words to be considered at the expense of consistency across texts. For each of these variants, we repeated the procedure of grouping similar topics described in the Methods. We calculated the Pearson correlation coefficient between the centroids of these new groups with the ones already chosen (by manual inspection) as traits, and after further manual curation kept only those describing new traits.

Finally, we assigned organisms to phenotypes according to the weights in the NMF H-matrix across all factors in the group describing that trait. For each organism with at least three assigned weights, we determined the median of the weights in the group of factors. Organisms with exactly two NMF weights, if both >0, received the smaller weight. In order to empirically determine a threshold for the provisional phenotypic labels, we examined examples of well-known traits (hyperthermophiles, plant pathogen, human oral cavity bacteria) to determine the following rule-of-thumb: the top-ranked organisms that made up 60% of the total sum of NMF weights, or the 10 top-ranked organisms (whichever number is higher) were labelled as positives, while all organisms with exactly zero NMF weight were negative examples.

Definition of classification accuracy measures

In a classification task, the precision is defined as $TP/(TP+FP)$, where TP denotes the number of true positives and FP denotes the number of false positives. Recall is defined as $TP/(TP+FN)$, where FN denotes the number of false negatives. $AUPRC$ score is calculated as the area under the precision-recall curve. AUC score is the area under the receiver operating characteristic (ROC) curve defined by false positive rate (FPR) on the x axis and true positive rate (TPR) on the y axes. TPR is equivalent to recall, while FPR is defined as $FP/(FP+TN)$, where TN denotes the number of true negatives. The F_1 measure is the harmonic mean of precision (P) and recall (R): $F_1 = 2PR/(P+R)$.

Estimating false discovery rates using precision-recall curves

The precision score for the positive class of each prediction (phenotype assignment to a microbe) was calculated as described above (the number of true positives divided by the total number of examples classified as positives), at the confidence score threshold that was assigned to that prediction by the SVM. Conversely, for the negative class, the precision score for the negative assignment of a particular phenotype to an organism was determined as the number of true negatives divided by the total number of examples classified as negatives. Precision scores for organisms in the initially unlabeled set of organisms were calculated *via* linear interpolation between the neighboring confidence points in the cross-validation precision-recall curve, which was previously determined using known examples.

Furthermore, we adjusted the precision score estimates to account for difference in class sizes. In particular, the estimates of the precision score (or, equivalently, FDR) depend on the relative proportion of positive/negative labels for the particular phenotype: the minimal precision cannot fall below the percentage of true positives/negatives in the learning set. This is a particularly evident issue in highly unbalanced classes, where by default, the precision for the majority label (typically, the negative one) will always be large, even for inaccurate classifiers. Thus, we adjusted the precision scores by subtracting the percentage of true positives and dividing with the (1-percentage of true positives). This ensured that the minimum precision is 0 (or equivalently that the maximum FDR is 100%), regardless of the number of positively/negatively labelled examples for the

phenotype. We used the adjusted FDRs in all further analysis. Importantly, this adjustment is always conservative i.e. the FDRs are always adjusted upwards.

Assignment of positive and negative classes.

The adjusted precision scores (and, equivalently, FDRs, as described above) were determined separately for both the positive and the negative class of each phenotypic trait. The overall precision was calculated as 'n votes', meaning that we took the n^{th} highest score for that class. Then, for a chosen FDR threshold we assigned the value 1 (presence of a trait) if the 'n votes' FDR for a positive class was greater than a chosen FDR threshold and the value 0 (absence of a trait) if the 'n votes' FDR for a negative class was greater than a chosen threshold. In the cases where the 'n votes' FDRs for both positive and negative class were greater than a threshold, we assigned the value of a minority class. Only for the purposes of visualization in the Figure S3, which shows cumulative coverage with annotations at different precision thresholds, we employed the following rule: we always assigned the value of the class with higher FDR value and only in the case of ties assign the value of a minority class. Instances that did not have the 'n votes' FDR greater than a chosen threshold did not receive a label for positive or negative class and remained unannotated.

In most cases the minority class was positive class, but for some phenotypes such as 'mesophilic' or 'free-living' the minority class was negative. Therefore, for the results calculated only for the minority class, we also report the information about the sign (positive or negative) of the minority class for that phenotypic trait (Supplementary Table S3).

Constructing features used to predict of phenotypes from genomic data

For the prediction of traits from the amino acid content of the proteome (1), we used amino acid and di-amino acid frequencies of a proteome as features, yielding 420 features. If there was more than one sequenced strain for a species, we took the strain with the highest genome quality score (2).

The gene repertoire of the genome was encoded as the presence/absence of the clusters of orthologous groups (COG) of proteins resulting in the total of 80576 binary valued features. For those species containing more than one high quality sequenced strain, we took the more frequent value; in the case of equal frequencies we gave advantage to the gene presence.

Pairwise co-occurrences of species in metagenomes were calculated as previously (3). We compared 16S rRNA sequences of species from our database against OTUs representative sequences using BLAST, with all parameters as in (3), except that ≥95% sequence identity was required, thus resulting in 1240 mapped species and 1240 features in the learning data.

The gene neighborhood representation covers the 300 COGs occurring in at least 80% (2205/2756) of species with high-quality sequenced genomes (allowed number of scaffolds ≤50). Features were encoded as the log pairwise chromosomal distance in nucleotides between each pair of COGs, in total 44850 features. Distances were measured from closest end of gene coding region. If either member of the COG pair was absent in a genome (or was located on distinct scaffolds in draft genomes), a missing value was recorded. If a COG was assigned to more than one gene in a genome, the minimal distance to the genes in the other COG was recorded. If COGs in the pair were found on different chromosomes, we set the distance to half-length of the larger chromosome. Additionally, all species with less than 100 non-missing values were removed from the data set. For species with multiple sequenced strains, the COG pairwise distance was calculated as the median across strains.

In translation efficiency profiles (4), the data set features quantify codon usage biases of COG/NOC gene families across genomes, measured using the MILC method (5). MILC is a normalized chi-square statistic that compares the relative codon frequencies in a protein-coding gene against a reference set of highly expressed genes, here encompassing ribosomal protein genes, translation initiation factors, translation elongation factors and chaperones (as in (6)). The OG-level score is the maximal observed MILC of genes assigned to that COG in

one genome. The features describe a set of 990 COGs occurring in at least 50% of examined species. If a COG was absent in a species, the feature value was set to missing value. For species containing more than one strain with high-quality genomes, we took the average MILC across all such strains.

Overlap between phenotypic traits

Network analysis was performed on binary phenotypic trait labels at a FDR<20% requiring agreement of two independent predictions ('two-votes'). The binary overall precision was 1 if the 'two-votes' FDR for the positive class was <20% and less than the FDR for the negative class, and 0 otherwise; equivalently, for the negative class. The network was visualized using Gephi (7). Edge weights between nodes were calculated using the F_1 measure and only the 533 edges with highest weights were retained prior to visualization. The nodes were arranged using ForceAtlas2 visual layout. We resized the nodes based on their degree and filtered out all nodes without neighbors. We partitioned the nodes based on modularity that uses a community detection algorithm proposed in (8).

Covariates describing phylogenetic relatedness

We obtained a microbial phylogeny from the Living Tree Project (LTP), release 123 (9) reconstructed using 16S rRNA sequences from SILVA (10). Out of 3046 microbial species, 2017 could be matched to LTP exactly, and we cross-referenced a further 713/136/103 organisms to the LTP by finding matching species at the genus/family/order-level; the remaining 76 microbes in our data could not be matched to LTP and were not used in the association analyses. The LTP tree was converted to a pairwise distance matrix of all LTP species and processed using principal components (PC) analysis, wherein the first 8 PCs retained 96.2% of the variance from the original species distance matrix and were included as covariates in logistic regression (see below).

Gene-trait associations

We used the binary phenotypic trait annotations predicted from text mining to search for associations between the occurrence of each gene family and each of the phenotypes, while considering 80,576 prokaryotic COG/NOG gene families from eggNOG 3 and 1640 (of 3046) microbial species that had textual data. We required the phenotypic labels to have FDR<10% for the positive class in at least one text corpus to be annotated as positive examples; equivalently for negative labels. In cases where FDR <10% for both the positive and the negative class, the minority class label was assigned. We considered phenotypes having ≥10 labelled examples, resulting in 166/424 phenotypes for known phenotypic annotations and 332/424 phenotypes for the known plus novel annotations. As a first-pass filter, we tested all COG-phenotype pairs with odds ratio (OR) ≥2 or OR≤0.5 and significant at nominal p<=0.01 using Fisher exact tests, performed separately for bacterial and for archaeal species.

This resulted in $2.8*10^5$ associations for the known annotations, and $1.0*10^6$ for the known plus novel annotations, which were further tested using logistic regression to control for confounding of evolutionary relatedness. In particular, we included 8 covariates derived from a known 16s rRNA phylogenetic tree (principal components of the species' pairwise distance matrix; Supplementary Methods). In addition, we also adjusted for confounding of genome size and G+C content (3, 11). Confounders were normalized to [0,1] and logistic regression in Matlab 2011b was then run on these 8+2 covariates and the presence/absence pattern of one COG as the independent variables, and one phenotypic trait annotation as the dependant variable. This was repeated for each COG-trait combination, and significant results re-tested using R-3.2.4 (*glm* function, setting *family=binomial*). The coefficient β of the COG variable and its standard error were used to find the OR adjusted for covariates, and its confidence interval. The p-values from a t-test on the β coefficient were FDR-corrected, pooling tests across all COGs and all phenotypes. Conservatively, the total number of tests for the FDR correction also included those that failed the first-pass Fisher's exact tests, if the effect size was sufficient. Finally, we report the COG/trait combinations for which the OR (adjusted for covariates) was >4 or <0.25.

Epistatic interactions

We used the same binary phenotype annotations as for finding gene-phenotype associations (only text predictions, FDR<10%), while focusing on 2663 COG/NOG gene families appearing in ≥200 (of 1640) species. As a first-pass filter, we required COG-COG-phenotype combinations to have the ratio of ORs ≥2 (for positive class) or ≤0.5 (for negative class) and tested them using a *Z*-test for the difference of log odds ratios, requiring p≤0.0001 (unadjusted) in either bacteria or archaea. This resulted in $3.3*10^6$ and $12.8*10^6$ tests, for the known and the known plus novel annotations, respectively, which were further tested using logistic regression while controlling for 8+2 confounders as described above. In addition, the presence/absence patterns of both COGs were also included as covariates, while the genetic interaction variable to be tested was represented as the product (equivalent to a logical AND) between COGs in a pair. The logistic regression was repeated for each COG-COG-phenotype combination, where the covariate-adjusted ORs of the interaction variable (OR_{inter}) and its confidence intervals were determined from the β coefficient. The p-values from a t-test on the β coefficients were FDR-corrected (total number of tests included also the COG-COGs-phenotype combinations that did not pass the first-pass filter). We impose effect size thresholds to require $OR_{inter}<0.25$ to call antagonistic epistasis and $OR_{inter}>4$ for synergistic epistasis.

Simulation studies of prevalence of gene-phenotype associations

The effect that coverage with phenotypic labels affects has the number of recovered gene-phenotype associations was examined in an analysis of 18 representative phenotypes. Herein, we considered the set of gene-phenotype associations that were significant using the full set of phenotypic annotations (including the annotations we predicted from text mining at FDR<10%), while requiring FDR<10% and OR>4 in the initial association analysis (test on the β coefficient of logistic regression; see above). Then, we created random samples of this data by choosing 100%, 98%, 96%, 94%, ... 50% of the labelled organisms from the initial analysis, and repeated the logistic regression test on the COGs that were initially significant at FDR<10%. We recorded the number of highly confident (FDR<1%, OR>4) significant COGs for each sampling, and fit a linear function between the number of organisms annotated with a phenotype versus the number of discovered significant relationships.

Functional annotation of COG gene families

Gene Ontology (GO) terms were assigned to COGs/NOGs by propagating the GO annotations of the underlying genes across the gene families. In particular, genes were mapped to OGs in the eggNOG 3 database (31) using Lambda v0.4.7 (41) in *blastp* mode with e-value threshold of 10^{-5}, and assigning a gene to the lowest e-value OG if the hit had sequence identity>30%. Then, each COG was annotated with a set of GO terms that appear in >=50% of its constituent genes, as recommended previously (12) , tallying only the genes that had any GO term assigned. Both the experimentally verified and electronic annotations in the UniProt GOA database (13) from all three GO domains were assigned, and propagated upwards to their parent GO terms, following the structure of the GO graph.

SUPPLEMENTARY METHODS REFERENCES

1. Brbić,M., Warnecke,T., Kriško,A. and Supek,F. (2015) Global Shifts in Genome and Proteome Composition Are Very Tightly Coupled. *Genome Biol. Evol.*, **7**, 1519–1532.

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

3. Chaffron,S., Rehrauer,H., Pernthaler,J. and Mering,C. von (2010) A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome Res.*, **20**, 947–959.

4. Krisko,A., Copic,T., Gabaldón,T., Lehner,B. and Supek,F. (2014) Inferring gene function from evolutionary change in signatures of translation efficiency. *Genome Biol.*, **15**, R44.

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

6. Supek,F., Škunca,N., Repar,J., Vlahoviček,K. and Šmuc,T. (2010) Translational Selection Is Ubiquitous in Prokaryotes. *PLOS Genet*, **6**, e1001004.

7. Bastian,M., Heymann,S. and Jacomy,M. (2009) Gephi: An Open Source Software for Exploring and Manipulating Networks. In *Third International AAAI Conference on Weblogs and Social Media*.

8. Blondel,V.D., Guillaume,J.-L., Lambiotte,R. and Lefebvre,E. (2008) Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.*, **2008**, P10008.

9. Munoz,R., Yarza,P., Ludwig,W., Euzéby,J., Amann,R., Schleifer,K.-H., Oliver Glöckner,F. and Rosselló-Móra,R. (2011) Release LTPs104 of the All-Species Living Tree. *Syst. Appl. Microbiol.*, **34**, 169–170.

10. Quast,C., Pruesse,E., Yilmaz,P., Gerken,J., Schweer,T., Yarza,P., Peplies,J. and Glöckner,F.O. (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.*, **41**, D590–D596.

11. Konstantinidis,K.T. and Tiedje,J.M. (2004) Trends between gene content and genome size in prokaryotic species with larger genomes. *Proc. Natl. Acad. Sci. U. S. A.*, **101**, 3160–3165.

12. Škunca,N., Bošnjak,M., Kriško,A., Panov,P., Džeroski,S., Šmuc,T. and Supek,F. (2013) Phyletic Profiling with Cliques of Orthologs Is Enhanced by Signatures of Paralogy Relationships. *PLoS Comput Biol*, **9**, e1002852.

13. Barrell,D., Dimmer,E., Huntley,R.P., Binns,D., O'Donovan,C. and Apweiler,R. (2009) The GOA database in 2009--an integrated Gene Ontology Annotation resource. *Nucleic Acids Res.*, **37**, D396-403.

14. Demšar,J. (2006) Statistical Comparisons of Classifiers over Multiple Data Sets. *J. Mach. Learn. Res.*, **7**, 1–30.

15. Meeske,A.J., Rodrigues,C.D.A., Brady,J., Lim,H.C., Bernhardt,T.G. and Rudner,D.Z. (2016) High-Throughput Genetic Screens Identify a Large and Diverse Collection of New Sporulation Genes in Bacillus subtilis. *PLOS Biol*, **14**, e1002341.

16. Galperin,M.Y., Mekhedov,S.L., Puigbo,P., Smirnov,S., Wolf,Y.I. and Rigden,D.J. (2012) Genomic determinants of sporulation in Bacilli and Clostridia: towards the minimal set of sporulation-specific genes. *Environ. Microbiol.*, **14**, 2870–2890.

17. Traag,B.A., Pugliese,A., Eisen,J.A. and Losick,R. (2013) Gene conservation among endospore-forming bacteria reveals additional sporulation genes in Bacillus subtilis. *J. Bacteriol.*, **195**, 253–260.

18. Rajagopala,S.V., Titz,B., Goll,J., Parrish,J.R., Wohlbold,K., McKevitt,M.T., Palzkill,T., Mori,H., Finley,R.L. and Uetz,P. (2007) The protein network of bacterial motility. *Mol. Syst. Biol.*, **3**, n/a-n/a.

 doi:10.1093/nar/gkw964

SUPPLEMENTARY FIGURES

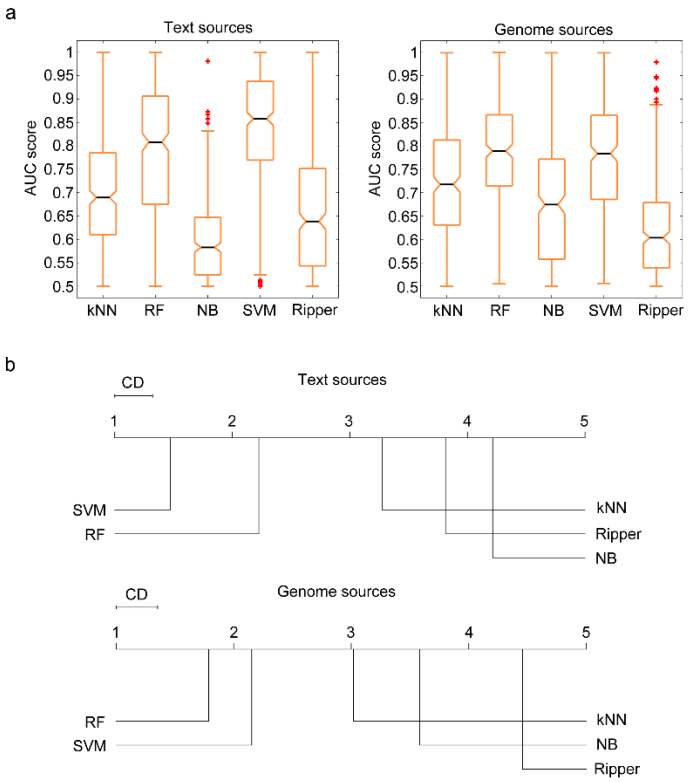


Figure S1. Benchmarking the accuracy of five machine learning algorithms for phenotype prediction. A selection of 60 phenotypic traits was used to determine their prediction accuracy, quantified as the AUC score on a held-out data set that consisted of 1/3 of the original data points (species). Shown separately for the six text corpora, and for the five genome representations. (**a**) Distributions of AUC scores. (**b**) Critical difference diagram (14) showing the average relative ranks of the five classifiers, where 1 denotes the best-ranking and 5 denotes the worst-ranking algorithm. The performance of two classifiers is significantly different ($p<0.05$) if the corresponding average ranks differ by at least the critical difference (denoted as “CD” in the plot).

 doi:10.1093/nar/gkw964

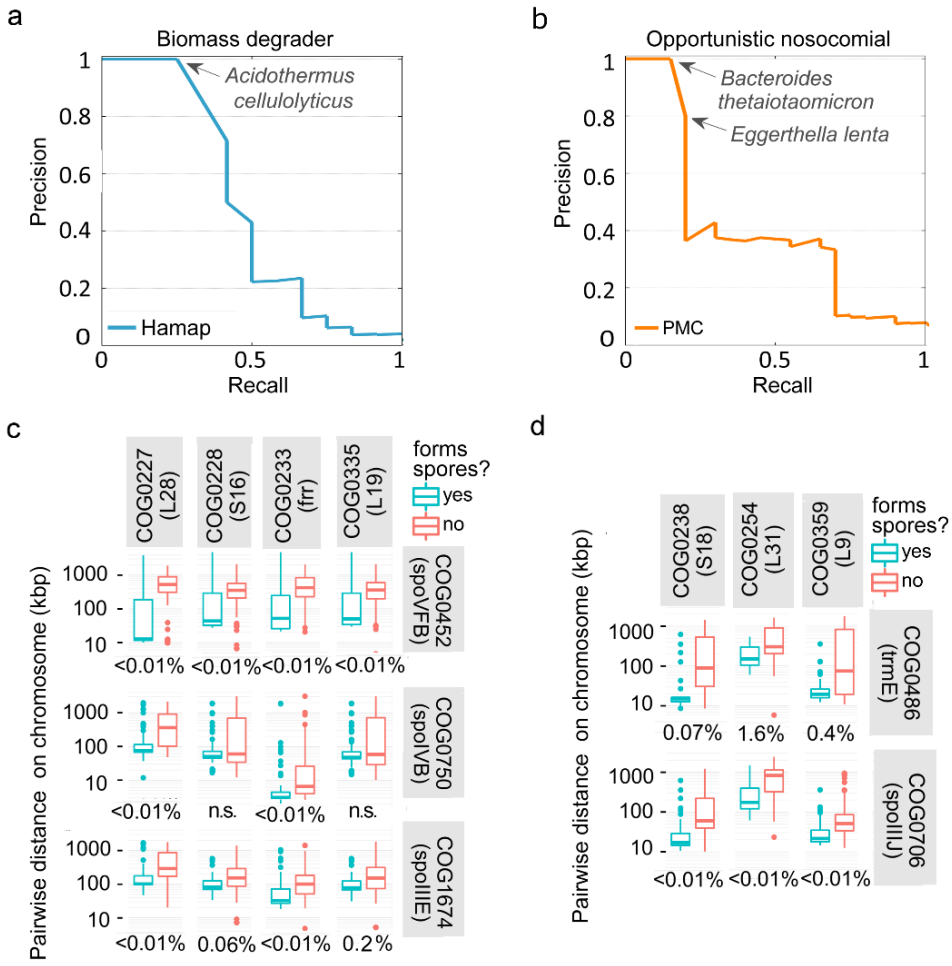


Figure S2. Predicting various phenotypes from text and genomic data. (**a**, **b**) Precision-recall curves of SVM classification models that were not broadly accurate in predicting phenotypes, but could still annotate a certain number of organisms at high precision thresholds. Shown for the biomass degrading (**a**) and the opportunistic/nosocomial pathogen (**b**) phenotype. Both curves are in cross-validation. (**c**, **d**) Gene neighborhoods involving a ribosomal gene or translation factor (columns) and a known sporulation gene (rows). Overlaid numbers are FDRs, for difference in pairwise distances between sporulating and non-sporulating bacteria, by Mann-Whitney test. The gene neighbourhood involving *spoVFB*, *spoIVB* and *spoIIIE* genes (**c**) shown separately from the gene cluster with *spoIIIJ* and *trmE* genes (**d**).

 doi:10.1093/nar/gkw964

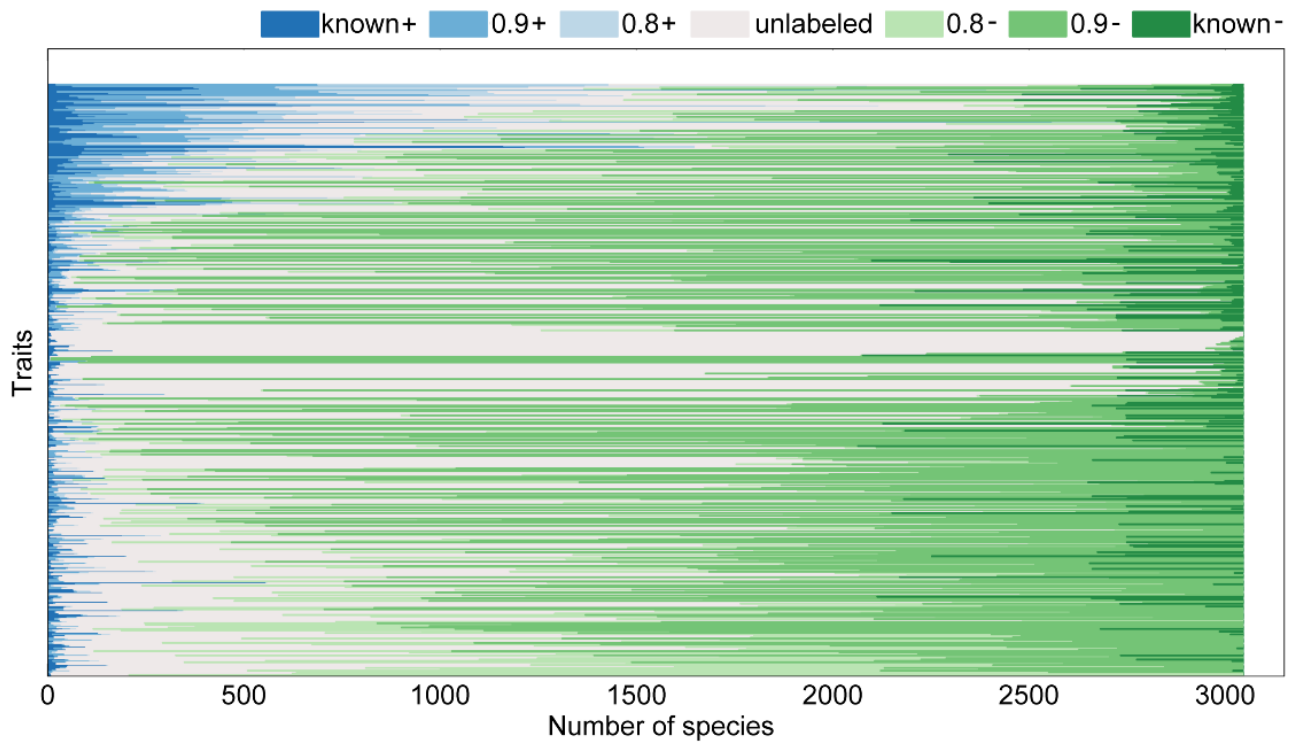


Figure S3. Coverage of microbial species with positive and negative annotations at various precision thresholds. "0.8" denotes a precision of >80% and thus a FDR of <20%; "0.9" a precision of >90% and a FDR of <10%. "+" denotes positive labels and "-" negative labels. Known stands for the previously known labels. All predictions are from "one-vote" scheme.

 doi:10.1093/nar/gkw964

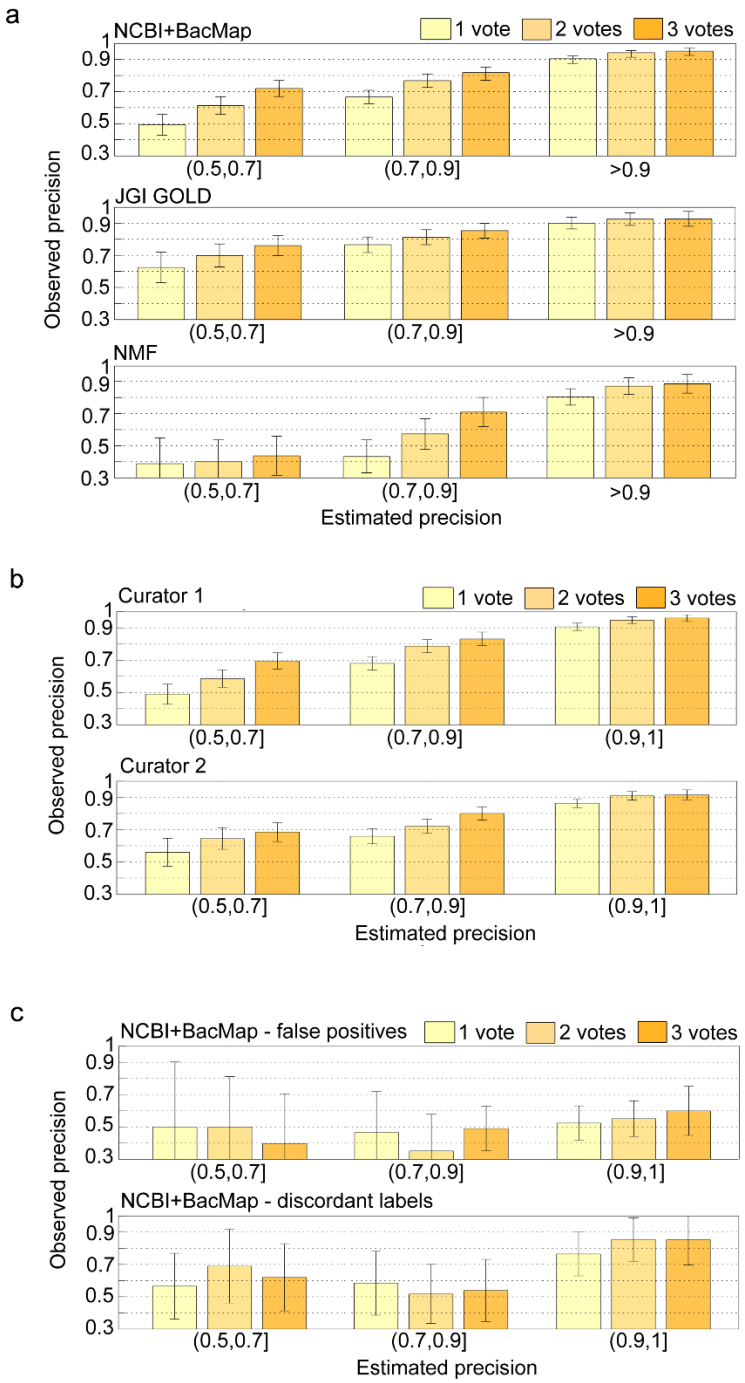


Figure S4. Validating a sample of phenotype inferences by manual curation. All panels show the actual precision scores (equivalent to 1-FDR) determined *via* literature curation, where the x axis shows data points binned by the nominal precision. (**a**) Accuracy of the FDR estimates for the three sets of phenotypic traits. (**b**) The evaluation results are consistent between the two curators. (**c**) Top panel shows the validation of cases where the predicted label contradicted the known label. Approx. ½ of such predictions were ultimately correct, while the initial labels appeared to be incorrect. Bottom panel shows the rare cases where the same annotation was supplied in the two source databases, but having opposite sense. Error bars are 95% C.I. (adjusted Wald).

 doi:10.1093/nar/gkw964

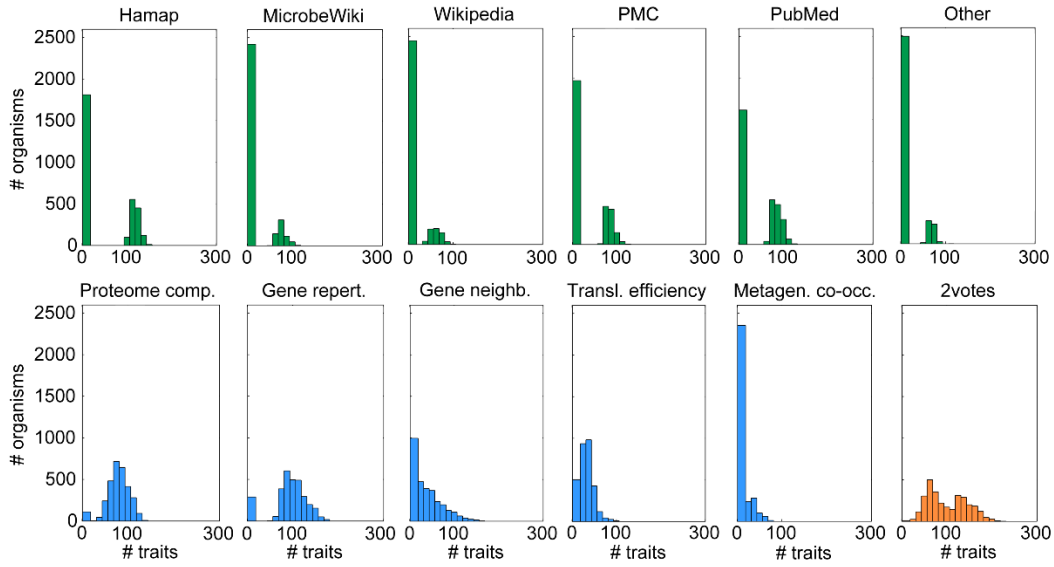


Figure S5. The distributions of the number of traits annotated per species, shown for individual prediction methods. The top row shows coverage histograms pertaining to the text-mining predictions from the six text corpora. The bottom row (blue bars) describes the coverage with the predictions from the five comparative genomics methods. The bottom right plot (orange bars) shows the coverage after integrating the predictions over the eleven methods by using the 'two-votes' scheme. All panels show numbers of annotations at a FDR<10%.

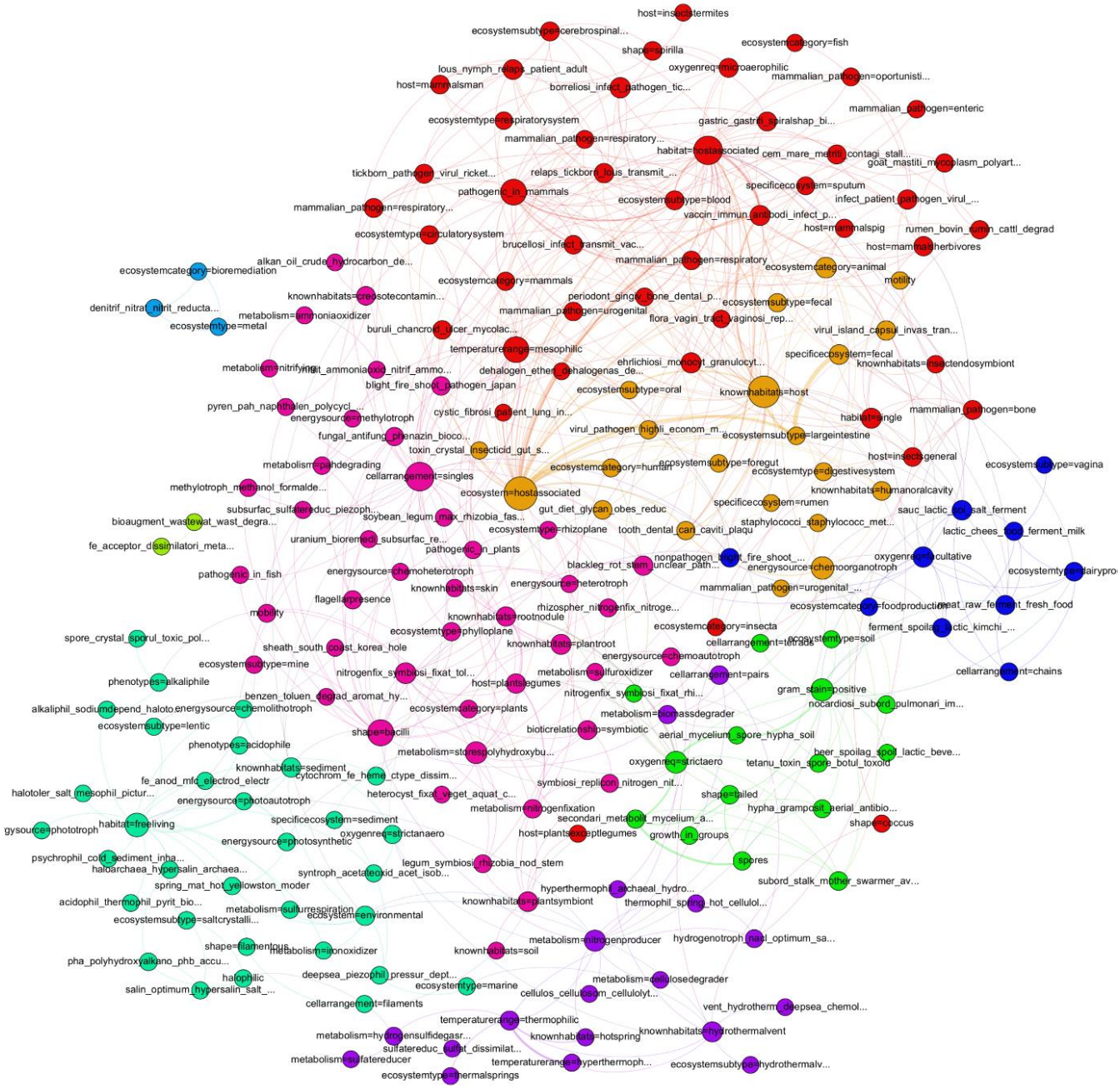
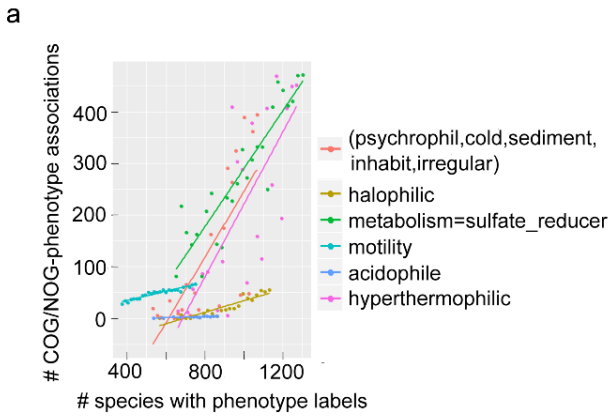
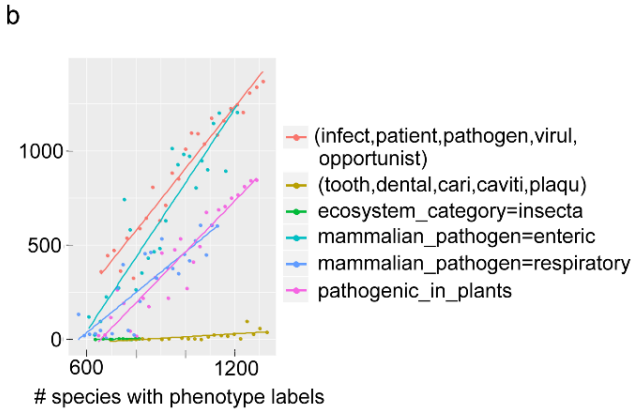
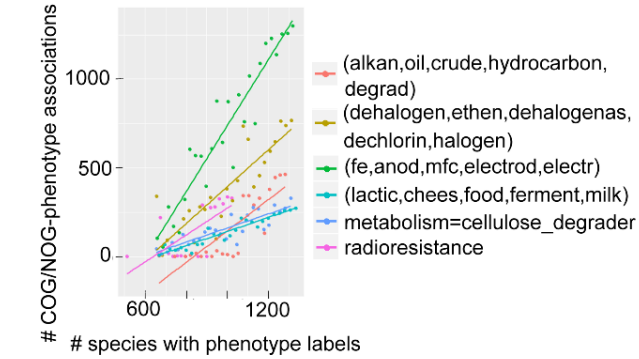


Figure S6. A network of microbial phenotypes. Similarity between pairs of phenotypes was estimated using the F_1-measure, which accounts for the overlap both in the organisms receiving positive labels and in those receiving negative labels. Edges show the 533 edges with highest F_1 similarities; width of edge reflects degree of similarity. Colors show results of the modularity based partitioning run on the network. Size of the nodes corresponds to the nodes degree. Nodes are arranged according to the ForceAtlas2 visual layout.

 doi:10.1093/nar/gkw964







d

phenotype	R^2	slope
hyperthermophilic	0.57	0.703
motility	0.92	0.090
radioresistance	0.47	0.758
halophilic	0.84	0.111
acidophile	0.70	0.011
(psychrophil,cold,sediment,inhabit,irregular)	0.57	0.634
metabolism=sulfatereducer	0.84	0.560
pathogenic_in_plants	0.89	1.346
(infect,patient,pathogen,virul, opportunist)	0.94	1.637
mammalian_pathogen=enteric	0.84	1.961
mammalian_pathogen=respiratory	0.74	1.065
(tooth,dental,cari,caviti,plaqu)	0.45	0.079
ecosystemcategory=insecta	0.23	0.014
(fe,anod,mfc,electrod,electr)	0.90	1.862
(dehalogen,ethen,dehalogenas, dechlorin,halogen)	0.74	1.040
(alkan,oil,crude,hydrocarbon,degrad)	0.70	0.884
metabolism=cellulosedegrader	0.75	0.397
(lactic,chees,food,ferment,milk)	0.90	0.398

Figure S7. Number of discovered gene-trait associations increases approximately linearly with the number of labelled organisms. (**a**, **b**, **c**) Simulations were used to estimate the number of gene-phenotype relationships for different levels of coverage with trait labels. Experiments were run on 18 representative phenotypes. Rightmost point of every phenotype (color) is the full set of phenotypic labels, including known and inferred (at FDR<10%; text sources only) labels. Points to the left are obtained by progressively reducing the number of microbes by random sampling, down to 50% of the original coverage. Y axis is the number of associations significant at FDR<1% (t-test for significance of logistic regression coefficient). (**d**) The R^2 and slopes of the fitted linear functions.

 doi:10.1093/nar/gkw964

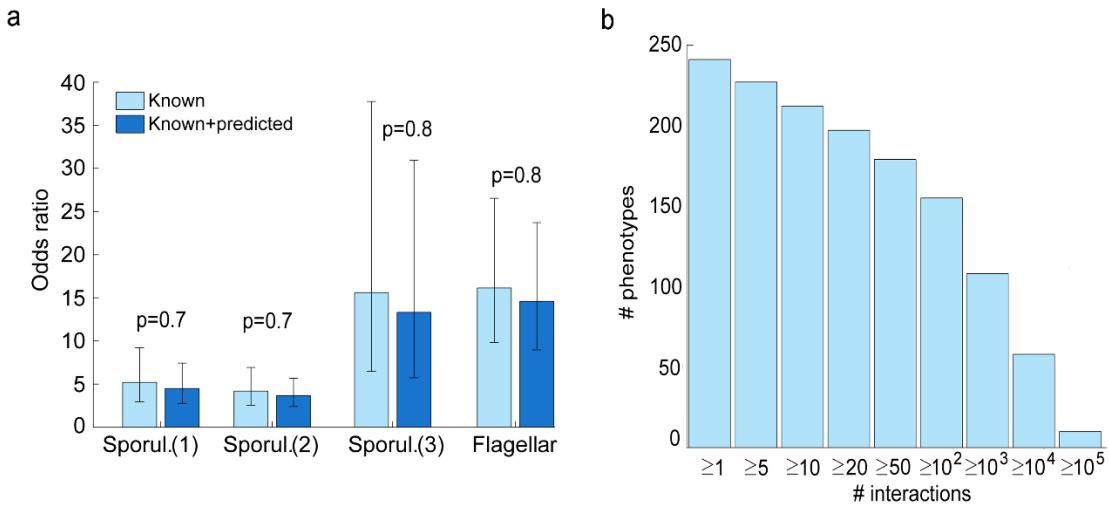


Figure S8. Coverage and validation of predicted gene-phenotype associations and epistatic interactions. (**a**) Odds ratios are denote relative enrichments of the genes associated to the phenotypes in the known sporulation and flagella gene sets (18). "Sporul(1)" denotes the set of sporulation genes found in reference (15), "Sporul(2)" the set of genes in reference (16), and "Sporul(3)" in reference (17). The known *versus* the known+inferred (<10% FDR, text mining only) set of annotated microbes are compared. Error bars are 95% C.I. of the odds ratio. Shown p-values by Z-test for difference of log odds ratios. (**b**) A histogram showing the amount of significant epistatic interactions detected per phenotypic trait.

SUPPLEMENTARY TABLE LEGENDS

Supplementary tables are available for download from the NAR website.

http://nar.oxfordjournals.org/content/early/2016/10/24/nar.gkw964/suppl/DC1

Table S1. Supporting information regarding the curation of the known phenotype labels from existing databases. Contains: (**a**) the list of matched and non-matched traits between the NCBI and BacMap databases and, in addition, the categorization of diseases and hosts used in these databases; (**b**) all instances of Discordant annotations between the NCBI and the BacMap databases; (**c**) the list of the biochemical phenotypes and all synonymous names thereof, used in manual curation from journal articles; and (**d**) a curated list of frequent keywords that were filtered out from texts prior to the NMF analysis.

Table S2. Discovery of phenotypic concepts from free-text using non-negative matrix factorization (NMF). Contains: **(a)** Top 20 keywords and their weights are shown for each concept (group of NMF factors), as well as its constituent NMF factors; and (**b**) a benchmark of the NMF concepts *versus* a methodology based on hierarchical clustering of keywords.

Table S3. Accuracy of phenotype prediction from text and genomic data sources. Contains: (**a**) benchmarks of the accuracy of five machine learning algorithms in predicting a sample of 60 phenotypic traits; (**b**) Accuracy (as AUC and AUPRC scores) for all combinations of trait-data source; (**c**) recall scores at two FDR thresholds for each classification model; and (**d**) same, but providing the numbers of false positive and of false negative examples.

Table S4. The sets of comparative genomics features with positive Random Forest feature importance scores, broken down by individual phenotypic traits.

Table S5. Detailed statistics describing the validation of the inferred phenotypes via literature searches by two curators.

Table S6. Gene-trait associations detected after controlling for confounders (phylogenetic relatedness, genome size and G+C content) and their enrichment in Gene Ontology functional categories. Contains: (**a**) Statistically significant associations of COG/NOG gene families to phenotypic traits. The “odds ratio” column is O.R., adjusted for covariates using logistic regression. Significance calls were by a t-test on the β coefficient, reported as false discovery rates (“FDR” column). The “data set” column shows whether the association was more confidently detected in the known set of phenotypic labels, or in the extended set of labels. (**b**) Gene functional categories significantly enriched with COG/NOG gene families that were associated to particular phenotypic traits. Significance calls were by Fisher’s exact test (one-tailed, enrichment only), and reported as false discovery rates (“FDR” column). The “data set” column shows whether the association was more confidently detected in the known set of phenotypic labels, or in the extended set of labels.

ONLINE DATA SETS

Additional online data sets are available for download from the ProTraits web interface. http://protraits.irb.hr/

"**ProTraits_precisionScores.txt**" contains predictions of 424 phenotypic traits for 3,046 bacterial and archaeal species. Table contains precision scores (equivalent to 1-FDR) obtained using 11 individual text mining and comparative genomics data sources, used to train Support Vector Machines (text) or Random Forest (genomics) classifiers. The precision scores were obtained by calibrating the classifier confidence scores using precision-recall curves obtained in cross-validation; see Methods in our publication for details.

Precision is provided separately for the positive **(+)** and the negative **(-)** class of a phenotypic trait. These scores need not add up to 1.0, since separate precision-recall curves were used to calibrate the scores for the positive and the negative phenotype for each trait. For instance, in cases where there is substantial uncertainty about the prediction, both the **(+)** and the **(-)** score reported will be low.

These are the precision scores browsable on http://protraits.irb.hr/ (web site reports only the prediction for the minority class of a given phenotype). We validated these scores by extensive manual curation; please see Brbic et al. publication for details.

The last two columns provide an integrated score obtained using the 'two-votes' scheme, meaning that two independent classifiers must support the given inference at that level of confidence. We recommend these scores for general use, based on their high coverage (~308,000 predictions) and excellent support in validation data (actual precision of the 11 data sources was 0.911-0.934 at nominal precision ≥ 90%).

"ProTraits_adjustedWaldConfInt.txt" - same as above, but reports precision scores and their 95% confidence intervals. These are obtained by applying the adjusted Wald method (Agresti & Coull, 1998) to the appropriate cut-off points in the crossvalidation precision-recall curves. The Wilson point estimate of the precision score is provided here.

"ProTraits_binaryIntegratedPr0.90.txt" A convenient tab-separated table with binarized predictions, requiring precision ≥ 0.9 (equivalent to FDR ≤ 10%) using only the integrated score (obtained via the two-votes scheme, as above). The value "**1**" denotes that a positive label was assigned to that phenotypic trait, while "**0**" denotes that a negative label was assigned. A "**?**" denotes that neither positive nor negative label could be assigned at precision ≥ 0.9.

In the extremely rare cases where both precision scores were greater than 0.9, the value of the class with the higher precision was assigned; in the case of ties the value of a minority class was assigned.

"ProTraits_binaryIntegratedPr0.95.txt" As above, but requires the a more stringent threshold of precision ≥ 0.95 (FDR ≤ 5%).