From genomes to phenotypes: Traitar, the microbial trait analyzer

A. Weimann[1,2,3], J. Frank[4], P. B. Pope[4], A. Bremges[1,2] and A. C. McHardy[1,2,3,†]

[1]Computational Biology of Infection Research, Helmholtz Centre for Infection Research, 38124 Braunschweig, Germany

[2]German Center for Infection Research (DZIF), partner site Hannover-Braunschweig, 38124 Braunschweig, Germany

[3]Department for Algorithmic Bioinformatics, Heinrich Heine University, 40225 Düsseldorf, Germany

[4]Department of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, Ås, 1432 Norway

[†]Correspondence: alice.mchardy@helmholtz-hzi.de

Abstract

The number of sequenced genomes is growing exponentially, profoundly shifting the bottleneck from data generation to genome interpretation. Traits are often used to characterize and distinguish bacteria, and are likely a driving factor in microbial community composition, yet little is known about the traits of most microbes. We present Traitar, the microbial trait analyzer, a fully automated software package for deriving phenotypes from the genome sequence. Traitar accurately predicts 67 traits related to growth, oxygen requirement, morphology, carbon source utilization, antibiotic susceptibility, amino acid degradation, proteolysis, carboxylic acid use and enzymatic activity.

Traitar uses L1-regularized L2-loss support vector machines for phenotype assignments, trained on protein family annotations of a large number of characterized bacterial species, as well as on their ancestral protein family gains and losses. We demonstrate that Traitar can reliably phenotype bacteria even based on incomplete single-cell genomes and simulated draft genomes. We furthermore showcase its application by characterizing two novel *Clostridiales* based on genomes recovered from the metagenomes of commercial biogas reactors, verifying and complementing a manual metabolic reconstruction.

Traitar enables microbiologists to quickly characterize the rapidly increasing number of bacterial genomes. It could lead to models of microbial interactions in a natural environment and inference of the conditions required to grow microbes in pure culture. Our phenotype prediction framework offers a path to understanding the variation in microbiomes. Traitar is available under the GPL 3.0 license at https://github.com/hzi-bifo/traitar.

Introduction

Microbes are often characterized and distinguished by their traits, for instance, in *Bergey's Manual of Systematic Bacteriology* (Goodfellow et al., 2012). A trait or phenotype can vary in complexity; for example, it can refer to the degradation of a specific substrate or the activity of an enzyme inferred in a lab assay, the respiratory mode of an organism, the reaction to Gram staining or antibiotic resistances. Traits are also likely to be a driving factor for microbial community composition; for example, in the cow rumen microbiota, bacteria capable of cellulose degradation influence the ability to process plant biomass material (Hess et al., 2011). In the Tammar wallaby foregut microbiome, the dominant bacterial species is implicated in the lower methane emissions produced by wallaby compared to ruminants (Pope et al., 2011).

Microbial diversity analysis via 16S rRNA amplicon sequencing can reveal the taxonomic composition of microbial communities, for example, in the human gut (Human Microbiome Project Consortium, 2012); however, it remains to be investigated to what degree the differences in composition have an impact on function. The major challenge for analyzing and interpreting the growing amount of microbial community sequencing data is to learn the molecular determinants of microbial phenotypes (Martiny et al., 2015).

The genotype–phenotype relationships for some microbial traits have been well studied; for instance, bacterial motility is attributed to the proteins of the flagellar apparatus (Macnab, 2003). Recently, we have shown that delineating these relationships from microbial genomes with known phenotype information using statistical learning methods allows the *de novo* discovery of novel protein families that are relevant for the realization of the plant biomass degradation phenotype and

to predict the phenotype with high accuracy (Weimann et al., 2013). However, a fully automated software framework for prediction of a broad range of traits from only the genome sequence is missing thus far. Additionally, horizontal gene transfer, a common phenomenon across bacterial genomes, has not been utilized to improve trait prediction before. Traits with their causative genes may be transferred from one bacterium to the other (Ochman et al., 2000; Pal et al., 2005) (e.g. for antibiotic resistances (Martinez, 2008)) and the vertically transferred part of a bacterial genome might be unrelated to the traits under investigation (Barker and Pagel, 2005; Harvey and Pagel, 1991; Martiny et al., 2015) .

Recently, several methods based on e.g. differential coverage and k-mer usage (Alneberg et al., 2014; Cleary et al., 2015; Gregor et al., 2016; Imelfort et al., 2014; Kang et al., 2015; Nielsen et al., 2014) to recover genomes from metagenomes (GFMs) were developed, which allow to recover genomes without the need to obtain microbial isolates in pure cultures (Brown et al., 2015; Hess et al., 2011). In addition, single-cell genomics provides a complementary approach for culture-independent analysis and allows, although often fragmented, genome recovery also for less abundant taxa (Lasken and McLean, 2014; Rinke et al., 2013).

The aim of this study was to develop an easy-to-use, fully automated software framework for the accurate prediction of a large number of microbial phenotypes solely from features derived from the microbial genome sequence. This allows the *in silico* characterization of the growing number of microbial isolates and microbial community members with genomes recovered from single-cell sequencing or metagenomes. We named our software Traitar, the microbial trait analyzer (Figure 1).

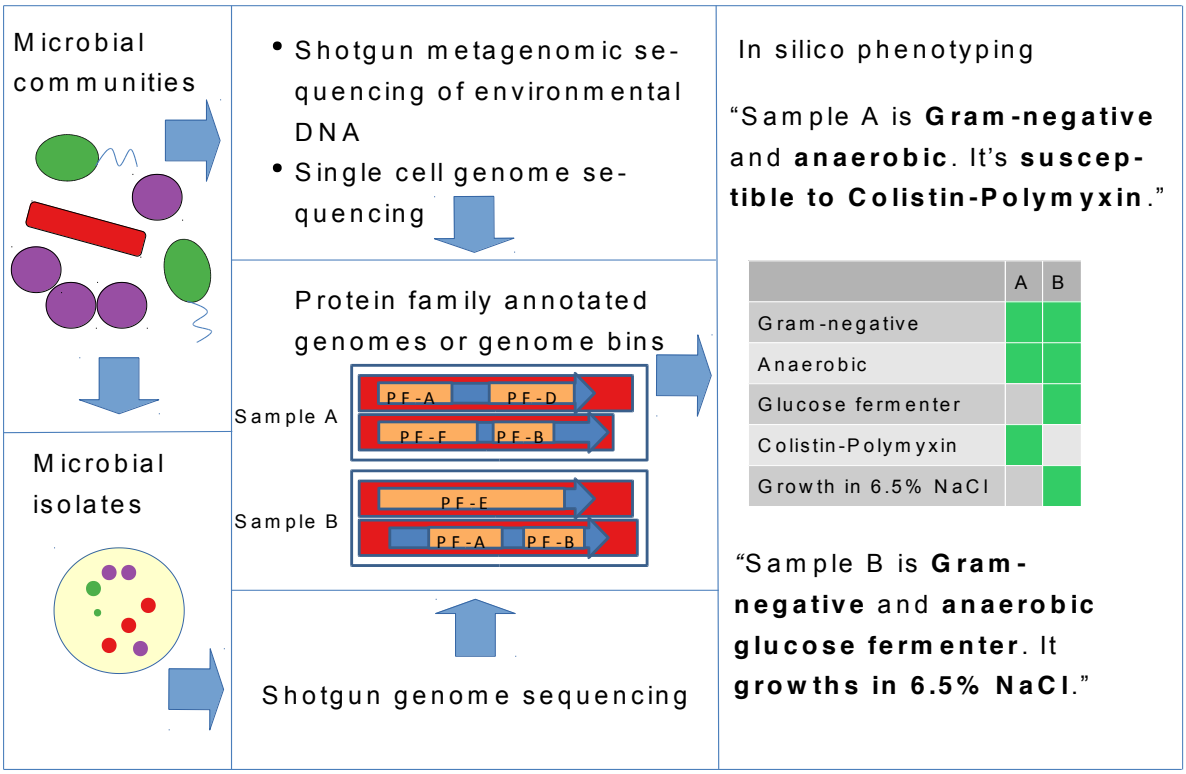

Figure 1: Application scenario: Traitar can be used to phenotype microbial community members based on genomes recovered from single-cell sequencing or (metagenomic) environmental shotgun sequencing data and microbial isolate genomes. Traitar provides classification models based on protein family annotation for a wide variety of different phenotypes related to growth, oxygen requirement, morphology, carbon source utilization, antibiotic susceptibility, amino acid degradation, proteolysis, carboxylic acid use and enzymatic activity.

We used phenotype data from the microbiology section of the Global Infectious Disease and Epidemiology Network (GIDEON), a resource dedicated to the diagnosis, treatment and teaching of infectious diseases and microbiology (Berger, 2005), for training phenotype classification models on the protein family annotation of a large number of sequenced genomes of microbial isolates. We investigated the effect of incorporating ancestral protein family gain and losses into the model inference on classification performance, to allow consideration of horizontal gene transfer events for phenotype-related protein families. We rigorously tested the performance of our software in cross-validation experiments and by applying it to an independent test set. We then used Traitar to phenotype bacteria based on single amplified genomes (SAGs) and simulated incomplete genomes to investigate its

potential for phenotyping microbial samples with incomplete genome sequences. We characterized two novel *Clostridiales* species of a biogas reactor community with Traitar, based on their population genomes recovered with metagenomics. We show how our predictions verified and complemented a manual metabolic reconstruction.

Traitar is implemented in Python 2.7, is freely available under the open-source GPL 3.0 license at https://github.com/hzi-bifo/traitar and can be installed on Linux/Unix via the Python Package Index as a stand-alone program. It is also available as a web service at https://research.bifo.helmholtz-hzi.de/traitar and as a Docker container.

Results

We begin with a description of Traitar software and the training of the phenotype models, first based on phyletic patterns and then including ancestral protein family gains and losses. We then proceed by describing the performance of the phenotype classification obtained in the cross-validation experiments. Subsequently, we provide the classification performance for different taxonomic levels. We then show the accuracy estimates computed for an independent test set. We further evaluate Traitar on simulated GFMs and SAGs, thereby demonstrating the suitability of our approach for incomplete bacterial genomes.

Afterwards, we provide several examples for the underlying genetic components of specific phenotypes, namely i.e. ‘Motility’, ‘Nitrate to nitrite’ conversion and ‘L-arabinose’ fermentation. Finally, we apply Traitar to characterize two uncultured phylotypes that were derived from a commercial biogas reactor and cross-reference our results with a manual metabolic reconstruction.

Traitar software

Traitar classifies 67 microbial traits (Table 1, Supplementary Table 1) for input samples. The software predicts these phenotypes based on models that were trained on protein and phenotype presence–absence data from 234 bacterial training species, as well as on models that were trained by incorporating the ancestral protein family gains and losses. The input to Traitar is either a nucleotide DNA FASTA for every sample, which is first run through gene prediction software, or an amino acid FASTA file for every sample. Traitar then annotates the proteins with protein families. Subsequently, it predicts phenotype labels for the input samples based on the models for the 67 traits. Finally, it cross-references the predicted phenotypes with the Pfam families that are relevant for their classification and then outputs GFF files that allow their inspection in a genome browser (Figure 2). These steps are described in more detail below. Parallel execution is supported by GNU parallel (Tange, 2011).

Table 1: The 67 phenotypes from the GIDEON database annotated as being present or absent for at least 10 microbial species each. We group each of these phenotypes into a microbiological or biochemical category, and provide information about the general type of microbiological assay required for testing the phenotype according to the GIDEON database.

Phenotype[(a)]	Category[(b)]	Test type[(c)]
Arginine dihydrolase	Amino Acid	General test
Indole	Amino Acid	General test
Lysine decarboxylase	Amino Acid	General test
Ornithine decarboxylase	Amino Acid	General test
Acetate utilization	Carboxylic Acid	General test
Citrate	Carboxylic Acid	General test
Malonate	Carboxylic Acid	General test
Tartrate utilization	Carboxylic Acid	General test
Alkaline phosphatase	Enzyme	General test
Beta hemolysis	Enzyme	General test
Coagulase production	Enzyme	General test
DNase	Enzyme	General test
Lipase	Enzyme	General test
Nitrate to nitrite	Enzyme	General test
Nitrite to gas	Enzyme	General test
Pyrrolidonyl-beta-naphthylamide	Enzyme	General test

Bile-susceptible	Growth	General test
Growth at 42°C	Growth	General test
Growth in 6.5% NaCl	Growth	General test
Growth in KCN	Growth	General test
Growth on MacConkey agar	Growth	Basic test
Growth on ordinary blood agar	Growth	Basic test
Mucate utilization	Growth	General test
Colistin-Polymyxin susceptible	Growth: Antibiotic	General test
Bacillus or coccobacillus	Morphology	Basic test
Coccus	Morphology	Basic test
Coccus - clusters or groups predominate	Morphology	Basic test
Coccus - pairs or chains predominate	Morphology	Basic test
Gram negative	Morphology	Basic test
Gram positive	Morphology	Basic test
Motile	Morphology	General test
Spore formation	Morphology	Basic test
Yellow pigment	Morphology	General test
Aerobe	Oxygen	Basic test
Anaerobe	Oxygen	Basic test
Capnophilic	Oxygen	General test
Facultative	Oxygen	Basic test
Catalase	Oxygen:Enzyme	Basic test
Oxidase	Oxygen:Enzyme	Basic test
Hydrogen sulfide	Product	General test
Casein hydrolysis	Proteolysis	General test
Gelatin hydrolysis	Proteolysis	General test
Cellobiose	Sugar	Fermentation or acidification
D-Mannitol	Sugar	Fermentation or acidification
D-Mannose	Sugar	Fermentation or acidification
D-Sorbitol	Sugar	Fermentation or acidification
D-Xylose	Sugar	Fermentation or acidification
Esculin hydrolysis	Sugar	General test
Glycerol	Sugar	Fermentation or acidification
Lactose	Sugar	Fermentation or acidification
L-Arabinose	Sugar	Fermentation or acidification
L-Rhamnose	Sugar	Fermentation or acidification
Maltose	Sugar	Fermentation or acidification
Melibiose	Sugar	Fermentation or acidification
myo-Inositol	Sugar	Fermentation or acidification
ONPG (beta galactosidase)$_{(d)}$	Sugar	General test
Raffinose	Sugar	Fermentation or acidification
Salicin	Sugar	Fermentation or acidification
Starch hydrolysis	Sugar	General test
Sucrose	Sugar	Fermentation or acidification
Trehalose	Sugar	Fermentation or acidification
Urea hydrolysis	Sugar	General test
Gas from glucose	Sugar:Glucose	General test

Glucose fermenter	Sugar:Glucose	Basic test
Glucose oxidizer	Sugar:Glucose	Basic test
Methyl red	Sugar:Glucose	General test
Voges Proskauer	Sugar:Glucose	General test

(a) GIDEON phenotypes with at least 10 presence and 10 absence labels
(b) Phenotypes assigned to microbiological / biochemical categories
(c) Type of test required for wet lab phenotype determination according to GIDEON
(d) ONPG: o-Nitrophenyl-β-D-galatopyranosid

Annotation

In the case of nucleotide DNA sequence input, Traitar uses Prodigal (Hyatt et al., 2010) for gene prediction prior to Pfam family annotation. The amino acid sequences are then annotated in Traitar with protein families (Pfams) from the Pfam database (version 27.0) (Finn et al., 2014) using the hmmsearch command of HMMER 3.0 (Finn et al., 2011).

Each Pfam family has a hand-curated threshold for the bit score, which is set in such a way that no false positive is included (Punta et al., 2012). A fixed threshold of 25 is then applied to the bit score (the log-odds score) and all Pfam domain hits with an E-value above 10^{-2} are discarded. The resulting Pfam family counts (phyletic patterns) are turned into presence or absence values, as we found this representation to yield a favorable classification performance (Weimann et al., 2013).

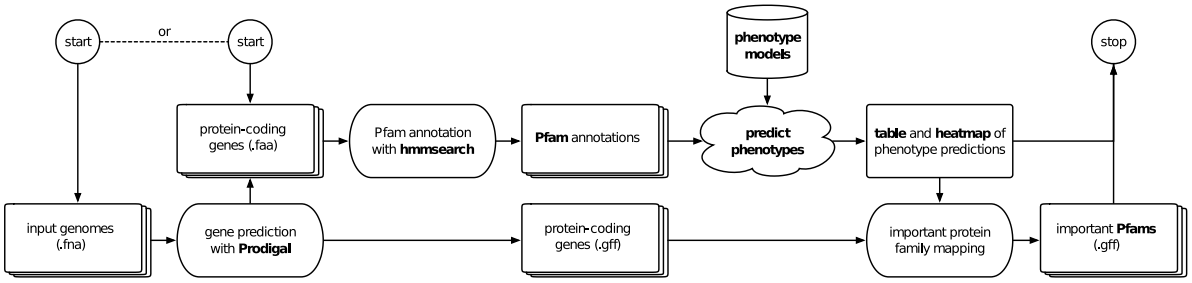

Figure 2: Work flow of Traitar. Input to the software can be genomic samples in nucleotide or amino acid FASTA format. Traitar predicts phenotypes based on pre-computed models and provides graphical and tabular output. In the case of nucleotide input, the protein families that are important for the phenotype predictions will be further mapped to the predicted protein-coding genes

Traitar phenotype models

We represented each phenotype from the set of GIDEON phenotypes across all genomes as a vector ***yp***, and solved a binary classification problem using the matrix of Pfam phyletic patterns *XP* across all genomes as input features and ***yp*** *a*s the binary target variable (Figure 3). For classification, we relied on support vector machines (SVMs), which are a well-established machine learning method (Boser et al., 1992). Specifically, we used a linear L1-regularized L2-loss SVM for classification as implemented in the LIBLINEAR library (Fan et al., 2008). For many datasets, linear SVMs achieve comparable accuracy to SVMs with a non-linear kernel but allow faster training. The weight vector of the separating hyperplane provides a direct link to the Pfam families that are relevant for the classification. L1-regularization enables feature selection, which is useful when applied to highly correlated and high-dimensional datasets, as used in this study (Zou and Hastie, 2005). We used the interface to LIBLINEAR implemented in scikit-learn (Pedregosa et al., 2011). For classification of unseen data points—genomes without available phenotype labels supplied by the user—Traitar uses a voting committee of five SVMs with the best single cross-validation accuracy (see Materials & Methods – Nested cross-validation). Traitar then assigns each unseen data point to the majority class (phenotype presence or absence class) of the voting committee.

Ancestral protein family and phenotype gains and losses

We constructed an extended classification problem by including ancestral protein family gains and losses, as well as the ancestral phenotype gains and losses in our analysis, as implemented in GLOOME (Cohen and Pupko, 2011). Barker *et al.* report that common methods for inferring functional links between genes, that do not take

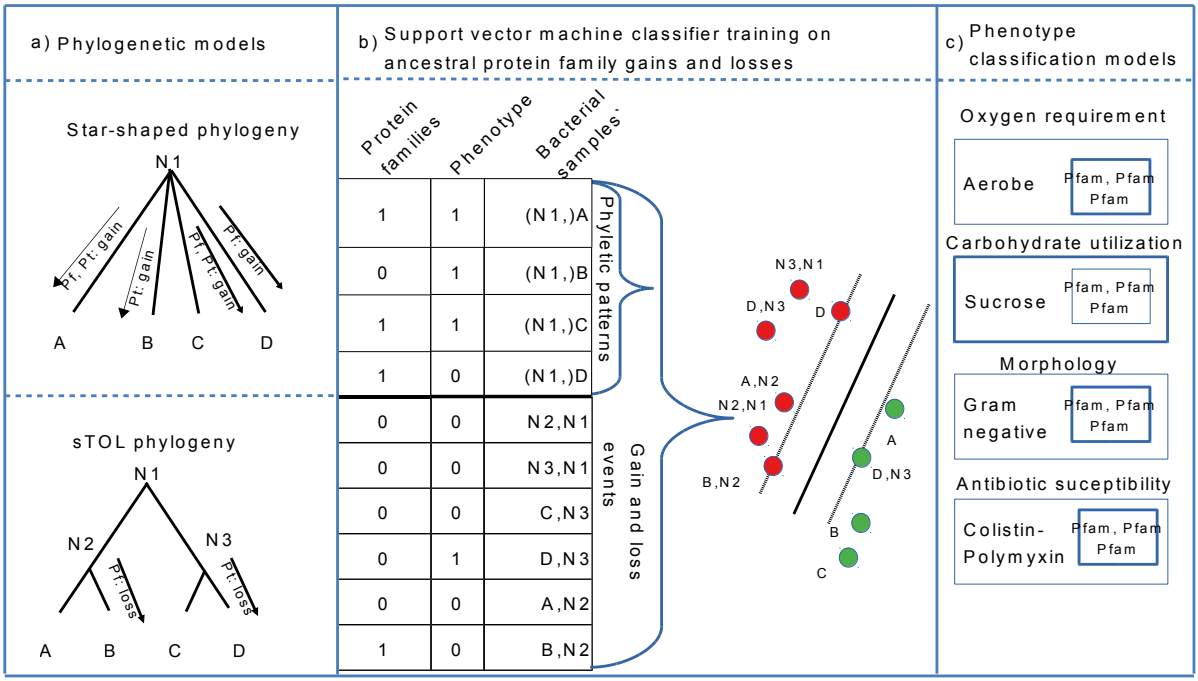

Figure 3: Schematic overview of the Traitar phenotype model training. (a) The phenotype and Pfam protein family phyletic patterns correspond to gain events on a star-shaped phylogenetic tree. Alternatively, we reconstructed the ancestral Pfam family and phenotype gain and loss events on the sequenced Tree of Life. (b) We trained a support vector machine classifier either on the phyletic patterns and on the ancestral gain and loss events, or solely on the phyletic patterns. (c) In this way, we inferred classification models for all available phenotypes.

the phylogeny into account, suffer from high rates of false positives (Barker and Pagel, 2005). Here, we jointly derived the classification models from the observable phyletic patterns and phenotype labels, and from phylogenetically unbiased ancestral protein family and phenotype gains and losses, that we inferred via a maximum likelihood approach from the observable phyletic patterns on a phylogenetic tree, showing the relationships among the samples. In this case, the phyletic patterns correspond to gain events on the branches of a star-shaped phylogeny (Figure 3). Ancestral character state evolution in GLOOME is modeled via a continuous-time Markov process with exponential waiting times. The gain and loss rates are sampled from two independent gamma distributions (Cohen and Pupko, 2010).

GLOOME needs a binary phylogenetic tree with branch lengths as input. The taxonomy of the National Center for Biointechnology Information (NCBI) and other

taxonomies are not suitable, because they provide no branch length information. We used the sequenced tree of life (sTOL) (Fang et al., 2013), which is bifurcating and was inferred with a maximum likelihood approach based on unbiased sampling of structural protein domains from whole genomes of all sequenced organisms (Gough et al., 2001). We employed GLOOME with standard settings to infer posterior probabilities for the phenotype and Pfam family gains and losses from the Pfam phyletic patterns of all NCBI bacteria represented in the sTOL and the GIDEON phenotypes. Each GIDEON phenotype p is available for a varying number of bacteria. Therefore, for each phenotype, we pruned the sTOL to those bacteria that were both present in the NCBI database and had a label for the respective phenotype in GIDEON. The posterior probabilities of ancestral Pfam gains and losses were then mapped onto this GIDEON phenotype-specific tree (Gps-sTOL, Figure 4).

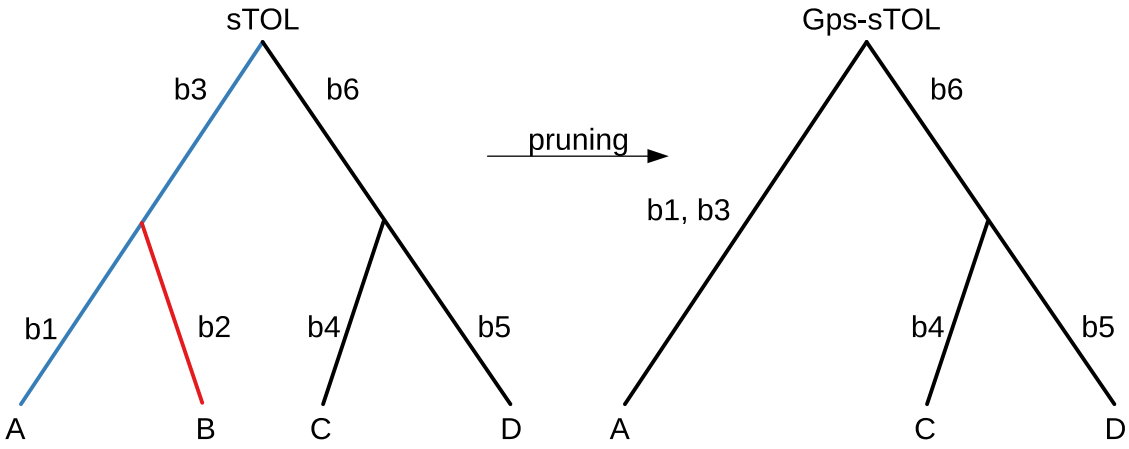

Figure 4: Sequenced Tree of Life (sTOL) and GIDEON phenotype-specific Tree of Life (Gps-sTOL) correspondence. The phenotype label for Sample B is not available. Consequently, only branches b_4, b_5 and b_6 are also found in the Gps-sTOL. The posterior probabilities for a Pfam gain or loss are the same for b_4, b_5 and b_6 in both trees. Branches b_1 and b_3 (blue) are collapsed into a single branch. The posterior probability for a gain on branch $b_{1,3}$, $g_{b_{1,3}}$ is computed from the posterior probability for a Pfam gain for b_1 and b_3, g_{b_1} and g_{b_3} as follows: $g_{b_{1,3}} = g_{b_1} + (1 - g_{b_1}) \cdot g_{b_3}$. Branch b_2 (red) in the sTOL does not have an analog in the Gps-sTOL.

Let *B* be the set of all branches in the sTOL and *P* be the set of all Pfam families. We then denote the posterior probability g_{ij} of an event a for a Pfam family pf to be a gain event on branch b in the sTOL computed with GLOOME as:

$$g_{ij} = P(a = gain | i = b, j = pf) \; \forall \; i \; \in B, \forall \; j \; \in P,$$

and the posterior probability of a to be a loss event for a Pfam family p on branch b as:

$$l_{ij} = P(a = loss | i = b, j = pf) \; \forall \; i \; \in B, \forall \; j \; \in \; P.$$

We established a mapping $f: B' \rightarrow B$ between the branches of the sTOL B and the set of branches B' of the Gps-sTOL (Figure 4). This was achieved by traversing the tree from the leaves to the root.

There are two different scenarios for a branch b' in B' to map to the branches in B:

a) Branch b' in the Gps-sTOL derives from a single branch b in the sTOL: $f(b') = \{b\}$. The posterior probability of a Pfam gain inferred in the Gps-sTOL on branch b' consequently is the same as that on branch b in the sTOL $g_{b'j} = g_{bj} \forall \; j \epsilon P$.

b) Branch b' in the Gps-sTOL derives from *m* branches $b_1, \ldots, b_m$ in the sTOL: $f(b') = \{b_1, \ldots, b_m\}$ (Figure 4). In this case, we iteratively calculated the posterior probabilities for at least one Pfam gain g' on branch b' from the posterior probabilities for a gain $g'_{b_1 j}$ from the posterior probabilities $g_1, \ldots, g_m$ of a gain on branches $b_1, \ldots, b_m$ with the help of h:

$$\begin{aligned} h_1 &= g_{b_1 j} \\ h_{n+1} &= (1 - h_n) \cdot g_{b_{n+1} j} \\ g'_{b_1 j} &= h_m \; \forall j \in P. \end{aligned}$$

Inferring the Gps-sTOL Pfam posterior loss probabilities l'_{ij} from the sTOL posterior Pfam loss probabilities is analogous to deriving the gain probabilities. The posterior probability for a phenotype p to be gained g_{ip}' or lost l_{ip}' can be directly defined for the Gps-sTOL in the same way as for the Pfam probabilities.

For classification, we did not distinguish between phenotype or Pfam gains or losses, assuming that the same set of protein families gained with a phenotype will also be lost with the phenotype. This assumption simplified the classification problem. Specifically, we proceeded in the following way:

1. We computed the joint probability x_{ij} of a Pfam family gain or loss on branch b' and the joint probability y_j of a phenotype gain or loss on branch b':

$$\begin{aligned} x_{ij} &= g'_{ij} l'_{ij} + \left(1 - g'_{ij}\right) \cdot l'_{ij} + \left(1 - l'_{ij}\right) \cdot g'_{ij} \; \forall \, i \in B', \forall \, j \in P \\ &= g'_{ij} + \left(1 - g'_{ij}\right) \cdot l'_{ij} \end{aligned}$$

$$y_i = g'_{ip} + \left(1 - g_{ip}'\right) \cdot \; l'_{ip} \quad \forall \, i \in B'.$$

2. Let $\boldsymbol{x_i}$ be a vector representing the probabilities x_{ij} for all Pfam families $j \in P$ on branch b_i. We discarded any samples $(\boldsymbol{x}_i, y_i)$ that had a probability for a phenotype gain or loss y_i above the reporting threshold of GLOOME but below a threshold t. We set the threshold t to 0.5.

This defines the matrix X and the vector **y** as:

$$(X, \boldsymbol{y}) = \{(\boldsymbol{x}_{\mathbf{i}}, y_{\mathrm{i}}) \,|y_i = 0 \vee y_i \geq t, i \in B'\} \ ,$$

By this means, we avoided presenting the classifier with samples corresponding to uncertain phenotype gain or loss events and used only confident labels in the subsequent classifier training instead.

3. We inferred discrete phenotype labels $\boldsymbol{y}'$ by applying this threshold t to the joint probability y_i for a phenotype gain or loss to set up a well-defined classification problem with a binary target variable. Whenever the probability for a phenotype to be gained or lost on a specific branch was larger than t, the event was considered to have happened:

$$\boldsymbol{y}' = \begin{cases} 1, \text{if } y_i \geq t \\ 0, \quad \text{otherwise} \end{cases} \forall i \in B'.$$

4. Finally, we formulated a joint binary classification problem for each target phenotype $\boldsymbol{yp}$ and the corresponding gain and loss events $\boldsymbol{y}'$, the phyletic patterns XP, and the Pfam gain and loss events X, which we solved again with a linear L1-regularized L2-loss SVM. We applied this procedure for all GIDEON phenotypes under investigation.

Software requirements

Traitar can be run on a standard laptop with Linux/Unix. The runtime for phenotyping a typical microbial genome with 3 Mbp is 9 minutes (3 min/Mbp) on an Intel(R) Core(TM) i5-2410M dual core processor with 2.30 GHz, requiring only a few megabytes of memory.

Evaluation

We considered two sets of classifiers trained on:

a) Binary phyletic patterns of Pfam families (phypat) inferred from 234 species-level bacterial genomes (see Traitar phenotype models).
b) Binary phyletic patterns of Pfam families inferred from 234 species-level genomes augmented with ancestral Pfam gains and losses (phypat+PGL) (see Ancestral protein family and phenotype gains and losses).

We determined the cross-validated macro-accuracy for the 67 GIDEON phenotypes as 82.6% for the phypat classifier and 85.5% for the phypat+PGL classifier; the accuracy for phypat was 88.1% compared with 89.8% for phypat+PGL (see Materials & Methods – Evaluation metrics). Notably, we could classify 53 phenotypes with at least 80% macro-accuracy and 26 phenotypes with at least 90% macro-accuracy in one of the two classification settings (Figure 5, Supplementary Table S2). We received a perfect classification assignment when predicting ‘Spore formation’ and the outcome of a ‘Methyl red’ test with the phypat+PGL classifier, and ‘Gram positive’ with the phypat classifier. Both ‘Spore formation’ and the phenotype ‘Methyl red’ test clearly profited from the ancestral protein family and phenotype gains and losses. The phypat classifier exhibited a slightly lower classification accuracy of around 90% for those phenotypes. Other phenotypes that could be predicted with very high confidence with both types of classifiers included oxygen requirement (i.e. ‘Anaerobe’ and ‘Aerobe’), ‘Growth on MacConkey agar’ or ‘Catalase’. Only a few phenotypes proved to be difficult to predict correctly, which included ‘DNAse’, ‘myo-Inositol’ or ‘Yellow pigment’ and ‘Tartrate utilization’, regardless of which classifier was used. For all these phenotypes, only a relatively small number (<20) of positive (phenotype present) examples were available.

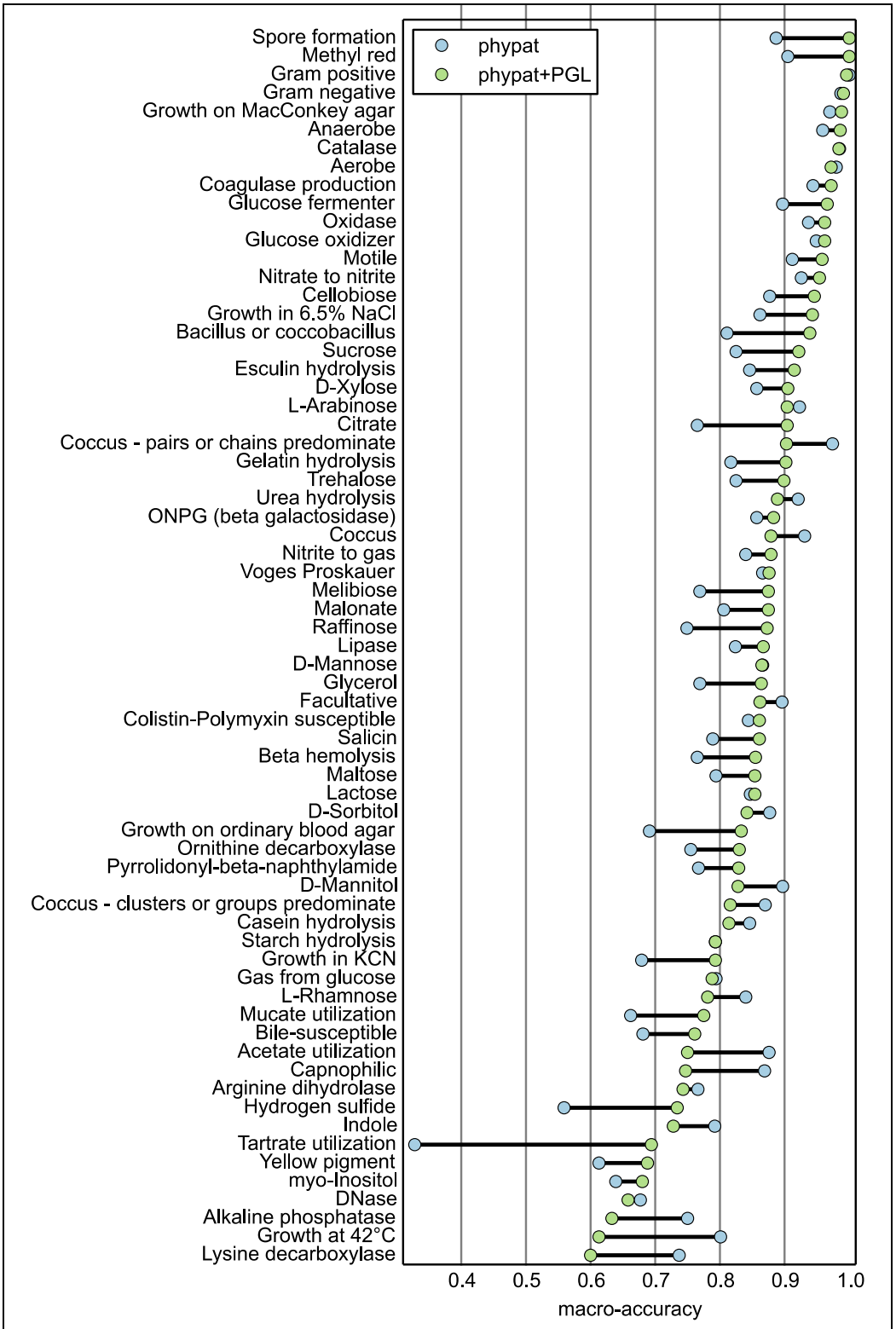

Figure 5: Macro-accuracy for each phenotype for the Traitar phypat and phypat+PGL phenotype classifiers. The macro-nested cross-validation accuracy (Material and Methods – Evaluation metrics) is shown for the individual GIDEON phenotypes (Table 1) for the phypat and the phypat+PGL classifiers.

We used the bacterial species that had phenotype information available in GIDEON, but had no representative in the sTOL tree, for an independent assessment of Traitar's classification accuracy. In total, this dataset comprised 42 unique species with 58 corresponding sequenced bacterial strains (Supplementary Table S3). We obtained an additional 1836 phenotype labels for these bacteria, consisting of 574 positive labels and 1262 negative labels. Traitar predicted the phenotypes in this dataset with a macro- accuracy of 85.3% using the phypat classifier and 86.7% using the phypat+PGL classifier, and accuracies of 87.5% and 87.9%, respectively. For calculation of the macro-accuracy, we considered only phenotypes with at least five positive and five negative labels.

Performance per clade at different depths of the taxonomy

We were interested in studying the potential taxonomic biases of the predictors. For this purpose, we evaluated the nested cross-validation performance of the phypat and phypat+PGL classifiers at different levels of the NCBI taxonomy. For a given GIDEON taxon, we pooled all instances of classification as bacterial species that are ancestors of this taxon. Figure 6 shows the accuracy estimates projected on the NCBI taxonomy from the domain level down to taxa at the genus level. We report the overall accuracy, as the macro-accuracy requires us to calculate the accuracies for all individual phenotypes, which cannot be computed for some phenotypes, since low-ranking taxa may have few labels or no labels at all. The accuracy of the phypat+PGL (phypat) classifier for the phyla covered by at least five bacterial species showed low variance and was high across all phyla, ranging from 84% (81%) for Actinobacteria over 90% (89%) for Bacteroidetes, 89% (90%) for Proteobacteria, 91% (90%) for Firmicutes to 91% (86%) for Tenericutes. Of the 39 genera, only the genus *Treponema* (with the representatives Treponema pallidum and T. denticola)

was classified with 69% (70%) accuracy, whereas all other genera were classified at least 78% (76%) correctly. This is probably because the genus Treponema was the only representative of the phylum Spirochaetes.

Figure 6: Classification accuracy for each clade depicted at different depths of the NCBI taxonomy. The nested cross-validation accuracy obtained for the GIDEON dataset was projected onto the NCBI taxonomy up to the genus level. Colored circles at the tree nodes depict the performance of the phypat+PGL classifier (left-hand circles) and the phypat classifier (right-hand circles). The size of the circles reflects the number of species per taxon.

Phenotyping incomplete genomes

GFMs or SAGs are often incomplete and thus we analyzed the impact of missing parts in the genome assemblies on the performance of Traitar. As an initial benchmark, we phenotyped bacteria based on SAGs and compared these results to the phenotypes inferred from joint genome assemblies. Rinke *et al.* used a single-cell sequencing approach to analyze poorly characterized parts of the bacterial and arachaeal tree of life, the so-called 'microbial dark matter' (Rinke et al., 2013). They pooled 20 SAGs from the candidate phylum Cloacimonetes, formerly known as WWE1, to generate joint—more complete—genome assemblies that had at least a genome-wide average nucleotide identity of 97% and belonged to a single 16S-based operational taxonomic unit, namely *Cloacamonas acidaminovorans* (Figure 7 a,b).

According to our predictions based on the joint assembly, *C. acidaminovorans* is Gram-negative and is adapted to an anaerobic lifestyle, which agrees with the description by Rinke *et al.* (Figure 7). Traitar further predicted 'Arginine dihydrolase' activity, which is in line with the characterization of the species as an amino acid degrader (Rinke et al., 2013). Remarkably, the prediction of a bacillus or coco-bacillus shape agrees with the results of Limam *et al.* (Limam et al., 2014), who used a WWE1-specific probe and characterized the samples with fluorescence *in situ* hybridization. They furthermore report that members of the Cloacimonetes candidate phylum are implicated in anaerobic digestion of cellulose, primarily in early hydrolysis, which is in line with the very limited carbohydrate degradation spectrum found by Traitar.

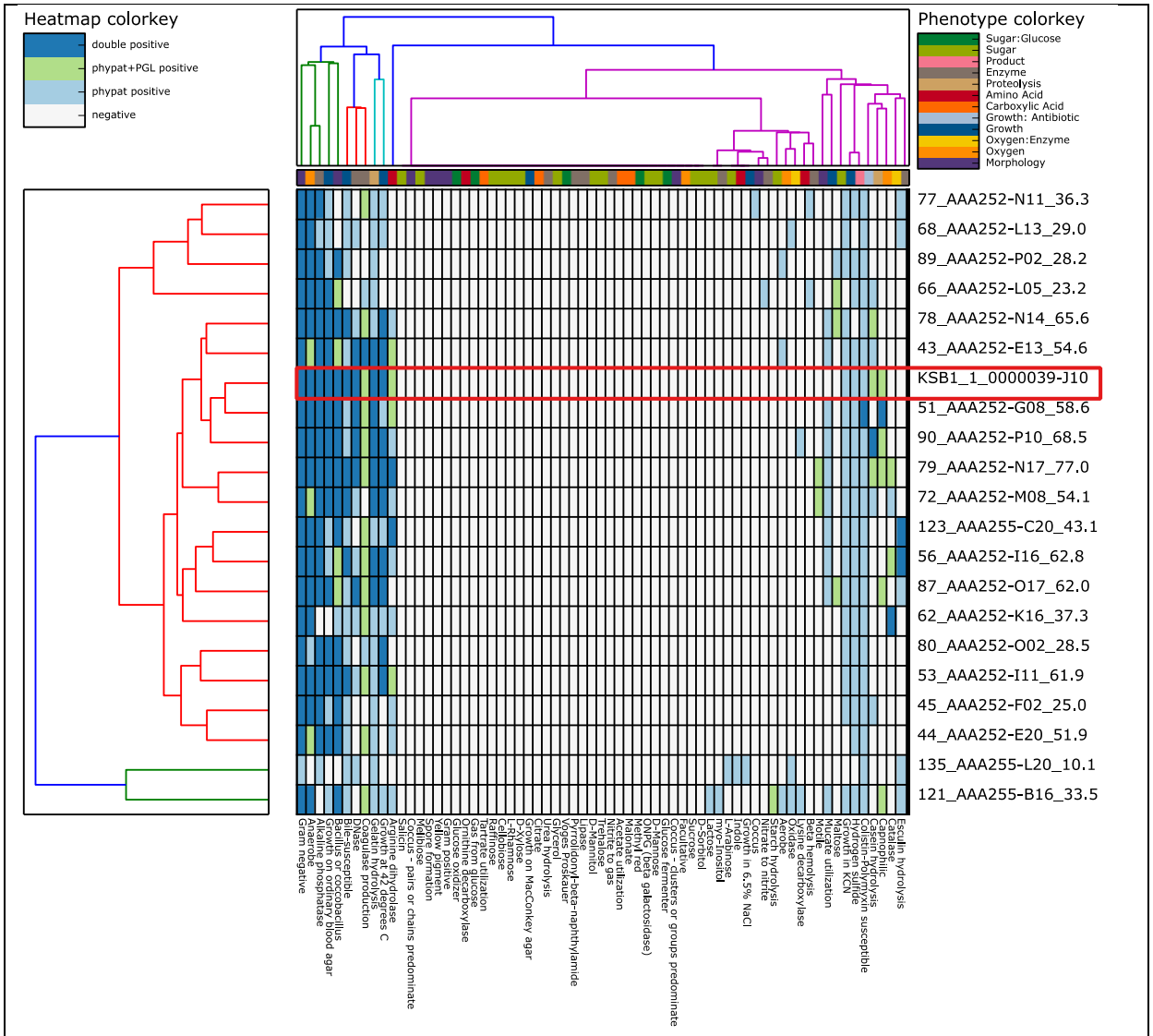

Figure 7: Single-cell phenotyping. We used 20 genome assemblies with varying degrees of completeness from single cells of the Cloacimonetes candidate phylum and a joint assembly for phenotyping with Traitar. We show a heatmap of the phenotype predictions (samples vs. phenotypes), which is the standard visualization for phenotype predictions in Traitar. The predicted phenotypes for the joint assembly are surrounded by a red box. The origin of the phenotype's prediction (Traitar phypat and/or Traitar phypat+PGL classifier) determines the color of the heatmap entries. The sample labels have their genome completeness estimates as suffixes. The different colors of the dendrograms code the groupings in the clustering.

Subsequently, we compared the predicted phenotypes based on the SAGs to the predictions for the joint assembly. We observed that the phypat classifier recalled more of the phenotype predictions of the joint assembly than the phypat+PGL classifier for the SAGs. However, the phypat+PGL classifier made fewer false positive predictions (Figure 8 a). By analyzing the protein families with assigned

weights and the bias terms of the two classifiers, we found the phypat+PGL classifier to base its predictions primarily on the presence of protein families that were typical for the phenotypes. In contrast, the phypat classifier also took typically absent protein families from phenotype-positive genomes into account in its decision. More technically, the positive weights in models of the phypat classifier are balanced out by negative weights, whereas for the phypat+PGL classifier, they are balanced out by the bias term. This explains the higher number of false positive predictions in sparse scenarios for the phypat classifier in comparison to the phypat+PGL classifier, which, in turn, made more false negative predictions.

In the next experiment, we inferred phenotypes based on simulated GFMs, by subsampling from the coding sequences of each of the 42 bacterial genomes that we already had used for an independent assessment of the overall accuracy. We started with the complete set of coding sequences and randomly deleted genes from the genomes. Note that we did not sub-sample from the Pfam annotation directly, as protein coding sequences can be annotated with more than one Pfam family. For draft genomes with different degrees of completeness, we re-ran the classification and computed the accuracy measures, as before. Once again, we observed that the performance of the phypat+PGL classifier dropped more quickly with more missing coding sequences than the performance of the phypat classifier (Figure 8 b). However, the recall of the positive class of the phypat+PGL classifier improved with a decreasing number of coding sequences (i.e. fewer but more reliable predictions were made). These tradeoffs in the recall of the phenotype-positive and the phenotype-negative classes were reflected in the similar macro-accuracy of the two classifiers across the range of tested genome completeness.

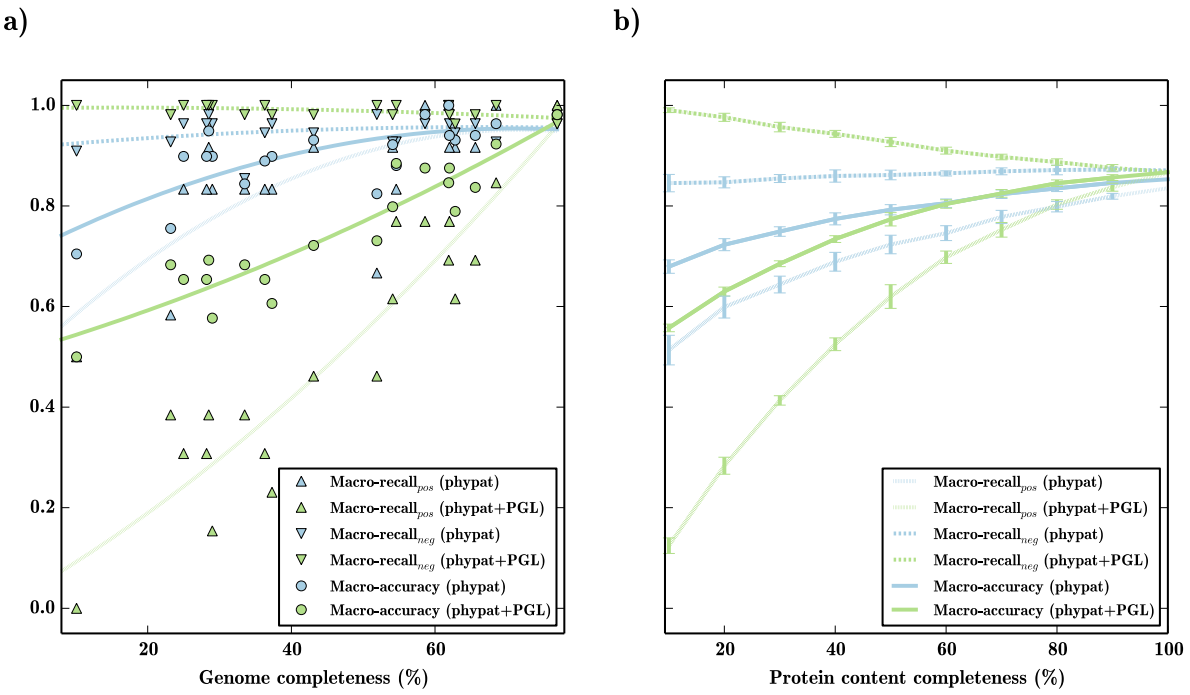

Figure 8: Phenotyping simulated draft genomes and single cell genomes. In (a) we used 20 genome assemblies with varying degrees of completeness from single cells of the Cloacimonetes candidate phylum and a joint assembly for phenotyping with the Traitar phypat and the Traitar phypat+PGL classifiers. We depict the performance of the phenotype prediction vs. the genome completeness of the single cells with respect to the joint assembly. In (b) we simulated draft genomes based on an independent test set of 42 microbial (pan)genomes. The coding sequences of these genomes were down-sampled and the resulting simulated draft genomes were used for phenotyping with the Traitar phypat and the Traitar phypat+PGL classifiers. We plotted various performance estimates against the protein content completeness.

By down-weighting the bias term for the phypat+PGL models by the protein content completeness, we could show that the accuracy of the phypat classifier could be exceeded by the phypat+PGL model, regardless of the protein content completeness (data not shown). However, this requires knowledge of the protein content completeness for each genomic sample, which could be indirectly estimated using methods such as checkM (Parks et al., 2015).

Depending on the intended usage, the classifiers can be chosen: we expect that the reliable predictions inferred with the phypat+PGL classifier and the more abundant,

but less reliable predictions made with the phypat classifier will complement one another in cases such as application to partial genomes recovered from metagenomic data.

Relevant protein families for selected phenotypes

We selected the three GIDEON phenotypes 'Motile', 'Nitrate to nitrite' conversion and 'L-arabinose' metabolism for a more detailed analysis of the protein families that contributed most to the phenotype classification from the Traitar phenotype models (see Materials & methods – Majority feature selection, Table 2). The selected phenotypes represent relevant phenotypes from the broader phenotype categories morphology, enzymatic activity and carbon source utilization. The protein families important for classification can be seen to be gained and lost jointly with the respective phenotypes within the microbial phylogeny, demonstrating the candidate link inferred by Traitar for each of these (Figure 9). Note that L1-regularized models tend to select the most informative features for classification, rather than all protein families that are involved in realization of the phenotype (Zou and Hastie, 2005).

Among the selected Pfam families that are important for classifying the motility phenotype were proteins of the flagellar apparatus and chemotaxis-related proteins (Figure 9a, Table 2). Motility allows bacteria to colonize their preferred environmental niches. Genetically, it is mainly attributed to the flagellum, which is a molecular motor, and is closely related to chemotaxis, a process that lets bacteria sense chemicals in their surroundings. Motility also plays a role in bacterial pathogenicity, as it enables bacteria to establish and maintain an infection. For example, pathogens can use flagella to adhere to their host and they have been reported to be less virulent if they lack flagella (Josenhans and Suerbaum, 2002). Of 48 flagellar proteins described in (Liu and Ochman, 2007), four proteins (FliS, MotB, FlgD and FliJ) were sufficient for

accurate classification of the motility phenotype and were selected by our classifier, as well as FlaE, which was not included in this collection. FliS (PF02561) is a known export chaperone that inhibits early polymerization of the flagellar filament FliC in the cytosol (Lam et al., 2010). MotB (PF13677), part of the membrane proton-channel complex, acts as the stator of the bacterial flagellar motor (Hosking et al., 2006). We also identified protein families related to chemotaxis, such as CZB (PF13682), a family of chemoreceptor zinc-binding domains found in many bacterial signal transduction proteins involved in chemotaxis and motility (Draper et al., 2011), and the P2 response regulator-binding domain (PF07194). The latter is connected to the chemotaxis kinase CheA and is thought to enhance the phosphorylation signal of the signaling complex (Dutta et al., 1999).

Nitrogen reduction in nitrate to nitrite conversion is an important step of the nitrogen cycle and has a major impact on agriculture and public health. Two types of nitrate reductases are found in bacteria: the membrane-bound Nar and the periplasmic Nap nitrate reductase (Moreno-Vivian et al., 1999), which we found both to be relevant for the classification of the phenotype: we identified all subunits of the Nar complex as being relevant for the ‘Nitrate to nitrite’ conversion phenotype (i.e. the gamma and delta subunit (PF02665, PF02613)), as well as Fer4_11 (PF13247), which is in the iron–sulfur center of the beta subunit of Nar. The delta subunit is involved in the assembly of the Nar complex and is essential for its stability, but probably is not directly part of it (Pantel et al., 1998) (Figure 9b, Table 2). We also identified the Molybdopterin oxidoreductase Fe4S4 domain (PF04879), which is bound to the alpha subunit of the oxidoreductase Fe4S4 domain (PF04879), which is bound to the alpha

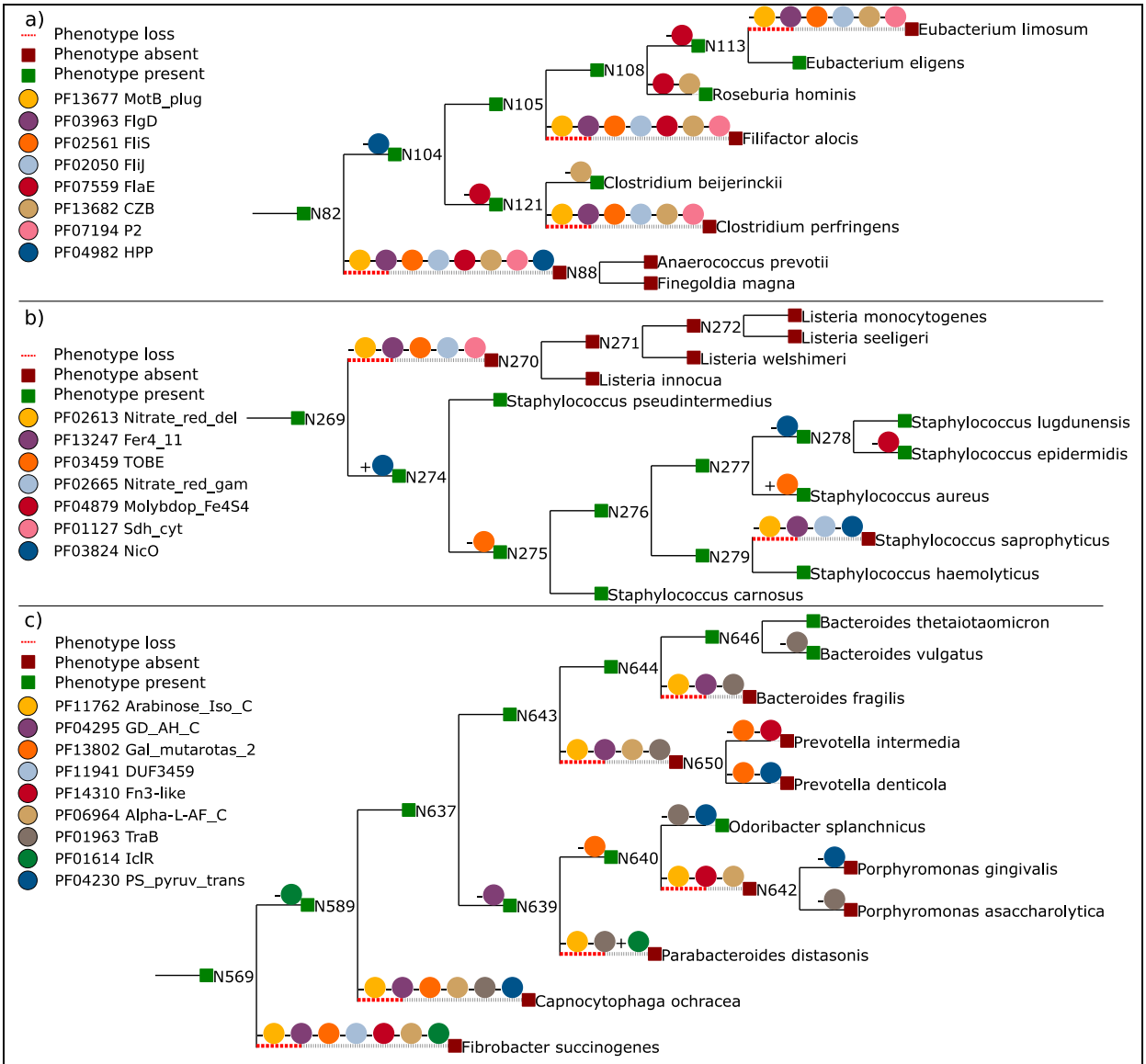

Figure 9: Phenotype gain and loss dynamics match protein family dynamics. Three phenotypes were selected for more detailed analysis: (a) 'Motile', (b) 'Nitrate to nitrite' conversion and (c) 'L-arabinose' metabolism. We show the phenotype–protein family gain and loss dynamics of subtrees of the sequenced tree of life with a substantial number of gain or loss events for these phenotypes. Colored circles along the tree branches depict protein family gains (+) or losses (-). The taxa are colored according to their (ancestral) phenotype state.

subunit of the nitrate reductase complex (Pantel et al., 1998). We furthermore found NapB (PF03892), which is a subunit of the periplasmic Nap protein and NapD (PF03927), which is an uncharacterized protein implicated in forming Nap (Moreno-Vivian et al., 1999).

Table 2: The most relevant Pfam families for classification of three important phenotypes: 'Nitrate to Nitrite', 'Motility' and 'L-Arabinose'. We ranked the Pfam families with positive weights in the SVM weight vector by the correlation of the phenotype with the Pfam families and present the 10 highest ranking Pfam families along with their descriptions and a link to the phenotype, where we found one.

Accession(a)	Phenotype(b)	Pfam description(c)	Remarks(d)
PF13677	Motile	Membrane MotB of proton-channel complex MotA/MotB	Flagellar protein
PF03963	Motile	Flagellar hook capping protein N-terminal region	Flagellar protein
PF02561	Motile	Flagellar protein FliS	Flagellar protein
PF02050	Motile	Flagellar FliJ protein	Flagellar protein
PF07559	Motile	Flagellar basal body protein FlaE	Flagellar protein
PF13682	Motile	Chemoreceptor zinc-binding domain	Chemotaxis-related
PF03350	Motile	Uncharacterized protein family, UPF0114	
PF05226	Motile	CHASE2 domain	Chemotaxis-related
PF07194	Motile	P2 response regulator binding domain	Chemotaxis-related
PF04982	Motile	HPP family	
PF03927	Nitrate to nitrite	NapD protein	Involved in Nar formation
PF13247	Nitrate to nitrite	4Fe-4S dicluster domain	Iron-sulfur cluster center of the beta subunit of Nar
PF03892	Nitrate to nitrite	Nitrate reductase cytochrome c-type subunit (NapB)	Periplasmic Nap subunit
PF02613	Nitrate to nitrite	Nitrate reductase delta subunit	Nap subunit
PF01127	Nitrate to nitrite	Succinate dehydrogenase/Fumarate reductase transmembrane subunit	
PF01292	Nitrate to nitrite	Prokaryotic cytochrome b561	
PF03459	Nitrate to nitrite	TOBE domain	
PF03824	Nitrate to nitrite	High-affinity nickel transport protein	
PF04879	Nitrate to nitrite	Molybdopterin oxidoreductase Fe4S4 domain	Bound to the alpha subunit of Nar
PF02665	Nitrate to nitrite	Nitrate reductase gamma subunit	Nar subunit
PF11762	L-Arabinose	L-arabinose isomerase C-terminal domain	Catalyzes first reaction in L-arabinose metabolism
PF04295	L-Arabinose	D-galactarate dehydratase / Altronate hydrolase, C terminus	
PF13802	L-Arabinose	Galactose mutarotase-like	
PF11941	L-Arabinose	Domain of unknown function (DUF3459)	
PF14310	L-Arabinose	Fibronectin type III-like domain	
PF06964	L-Arabinose	Alpha-L-arabinofuranosidase C-terminus	Acts on L-arabinose side chains in pectins
PF01963	L-Arabinose	TraB family	
PF01614	L-Arabinose	Bacterial transcriptional regulator	
PF06276	L-Arabinose	Ferric iron reductase FhuF-like transporter	
PF04230	L-Arabinose	Polysaccharide pyruvyl transferase	

(a) Pfam families that are relevant for one of three phenotypes that were selected for detailed investigation

(b) Phenotypes that were selected for detailed investigation

(c) Description of the Pfam family in the Pfam database

(d) Link from the Pfam family to the phenotype

L-arabinose is major constituent of plant polysaccharides, which is located, for instance, in pectin side chains and is an important microbial carbon source (Martinez et al., 2008). We could identify the L-arabinose isomerase *C*-terminal domain (PF11762), which catalyzes the first step in L-arabinose metabolism—the conversion of L-arabinose into L-ribulose (Sa-Nogueira et al., 1997), as being important for realizing the L-arabinose metabolism. We furthermore found the *C*-terminal domain of Alpha-L-arabinofuranosidase (PF06964), which cleaves nonreducing terminal alpha-L-arabinofuranosidic linkages in L-arabinose-containing polysaccharides (Gilead and Shoham, 1995) and is also part of the well-studied L-arabinose operon in *Escherichia coli* (Sa-Nogueira et al., 1997) (Figure 9c, Table 2).

Phenotyping biogas reactor population genomes

We used Traitar to phenotype two novel *Clostridiales* species (unClos_1, unFirm_1) based on their genomic information reconstructed from metagenome samples. These were taken from a commercial biogas reactor operating with municipal waste (Frank et al., 2015). The genomes of unClos_1 and unFirm_1 were estimated to be 91% complete and 60% complete based on contigs ≥5 kb, respectively. Traitar predicted unClos_1 to utilize a broader spectrum of carbohydrates than unFirm_1 (Table 3). We cross-referenced our predictions with a metabolic reconstruction conducted by Frank *et al.* (submitted). This reconstruction and predictions inferred with Traitar agreed to a great extent (Table 3). We considered all phenotype predictions that Traitar inferred with either the phypat or the phypat+PGL classifier. Traitar recalled 87.5% (6/7) of the phenotypes inferred via the metabolic reconstruction and also agreed to 81.8% (9/11) on the absent phenotypes. Notable exceptions were that Traitar only found a weak signal for ‘D-xylose’ utilization. A weak signal means that only a minority of the classifiers in the voting committee assigned these samples to

the phenotype-positive class (see Traitar phenotype models). However, the metabolic reconstruction was also inconclusive with respect to xylose fermentation. Furthermore, Traitar only found a weak signal for 'Glucose fermentation' for unFirm_1. Whilst genomic analysis of unFirm_1 revealed the Embden–Meyerhof–Parnas (EMP) pathway, which would suggest glucose fermentation, gene-centric and metaproteomic analysis of this phylotype indicated that the EMP pathway was probably employed in an anabolic direction (gluconeogenesis); therefore unFirm_1 is also unlikely to ferment D-Mannose (Frank *et al.* submitted). The authors of this study conclude that unFirm_1 is unlikely to ferment sugars and instead metabolizes acetate (predicted by Traitar, Table 3) via a syntrophic interaction with hydrogen-utilizing methanogens.

Traitar predicted further phenotypes for both species that were not targeted by the manual reconstruction. One of these predictions was an anaerobic lifestyle, which is likely to be accurate, as the genomes were isolated from an anaerobic bioreactor environment. It also predicted a Gram-positive stain, which is probably correct based on the presence of the Gram-positive sortase protein family found in both genomes—

Table 3: Phenotype predictions for two novel *Clostridiales* species with genomes reconstructed from a commercial biogas reactor metagenome. Traitar output (yes, no, weak) was cross-referenced with phenotypes manually reconstructed based on Kyoto Encyclopedia of Genes and Genomes orthology annotation (Frank *et al.* submitted), which are primarily the fermentation phenotypes of various sugars. We considered all phenotype predictions that Traitar inferred with either the phypat or the phypat+PGL classifier. A weak prediction means that only a minority of the classifiers in the Traitar voting committee assigned this sample to the phenotype-positive class (Traitar phenotype). Table entries colored in red show a difference between the prediction and the reconstruction, whereas green denotes an overlap; yellow is inconclusive.

	Glucose	Acetate	Mannitol	Starch hydrolysis	Xylose	L-Arabinose	Capnophilic	Sucrose	D-Mannose	Maltose	Arginine dihydrolase
unClos_1	yes	no	yes	no	weak	yes	yes	yes	yes	yes	No
unFirm_1	weak	yes	no	no	no	no	no	no	no	no	Yes

a Gram-positive biomarker (Paterson and Mitchell, 2004) and because all Firmicutes are known to be Gram-positive (Goodfellow et al., 2012). Furthermore, Traitar assigned 'Motile' and 'Spore formation' to unFirm_1, based on the presence of several flagellar proteins (e.g. FliM, MotB, FliS and FliJ) and the sporulation proteins CoatF and YunB.

Discussion

We have developed Traitar, a software framework for predicting phenotypes from the protein family profiles of bacterial genomes. We showed that Traitar provides a quick and fully automated way of characterizing microbiota based on their genome protein family content alone. Traitar includes classification models for an unprecedented number of phenotypes. Microbial trait prediction from phyletic patterns has been proposed in previous studies for a limited number of phenotypes (Feldbauer et al., 2015; Kastenmuller et al., 2009; Konietzny et al., 2014; Lingner et al., 2010; MacDonald and Beiko, 2010; Weimann et al., 2013). To our knowledge, the only currently available software for microbial genotype-phenotype inference is PICA, which is based on learning associations of clusters of orthologous genes (Tatusov et al., 2001) with traits (MacDonald and Beiko, 2010). Recently, PICA was extended by Feldbauer *et al.* for predicting eleven traits overall, optimized for large datasets and tested on incomplete genomes (Feldbauer et al., 2015). Traitar allows prediction of 67 phenotypes, including 60 entirely novel ones. It includes different prediction modes, one based on phyletic patterns, one additionally including a statistical model of protein family evolution for its predictions.

In cross-validation experiments with 234 bacterial species, we showed that the Traitar phypat classifier can phenotype samples with a macro-accuracy of 82.6%. Considering ancestral protein family gains and losses in the classification, which is implemented in the Traitar phypat+PGL classifier, we could significantly improve the accuracy compared to prediction from phyletic patterns only, with respect to both individual phenotypes and overall. We demonstrated the reliability of these performance estimates by phenotyping, with a similar accuracy, an independent test dataset that comprised an additional 42 genomes and pangenomes of species which had not been used in the cross-validation. Barker *et al.* (2005) were first to note the phylogenetic dependence of genomic samples and how this can lead to biased conclusions (Barker and Pagel, 2005). Beiko *et al.* selected protein families based on correlations with a phenotype and corrected for the taxonomy (MacDonald and Beiko, 2010). However, we are unaware of an approach that directly incorporates the phylogeny and accounts for horizontal gene transfer to improve phenotype classification. Incorporation of the protein family and phenotype gains and losses did not only improve the classification accuracy, but we could also showcase the co-evolutionary dynamics of genotype and phenotype gains and losses for selected phenotypes. We thus add evolutionary evidence to the relevance of the identified protein families for establishment of the investigated phenotypes. Several of these were already known to be associated with those phenotypes, whereas others were uncharacterized, representing targets for further experiments. Note that even though for neither type of classifier, we observed any evidence for a phylogenetic bias towards specific taxa, some of the phenotypes might be realized with different protein families in taxa that are less well represented in GIDEON.

We found that Traitar can accurately characterize the metabolic capabilities of microbial community members even from partial genomes, which is a very common scenario nowadays for genomes recovered from single cells or metagenomes. The analysis of both the SAGs and simulated genomes for phenotyping led us to the same conclusions: the phypat classifier is more suitable for exploratory analysis, as it assigned more phenotypes to incomplete genomes, at the price of more false positive predictions. In contrast, the phypat+PGL classifier assigned fewer phenotypes, but also fewer false assignments.

For the phenotyping of novel microbial species, a detailed (manual) metabolic reconstruction such as the one by Frank *et al.* (submitted) is time-intensive, as such reconstructions are usually focused on specific pathways and are dependent on the research question. This is not an option for studies with 10–50+ genomes, which are becoming more and more common in microbiology (Brown et al., 2015; Hess et al., 2011; Rinke et al., 2013). In that sense, our method is more suitable for the analysis of multi-genome studies. It furthermore may pick up on things outside of the original research focus and could serve as a seed for a detailed metabolic reconstruction in future studies.

As data for more phenotypes (e.g. further antibiotic resistance phenotypes) become available, Traitar could be easily extended to include such models into its framework. We also expect that the accuracy of the phenotype classification models already available in Traitar will profit from additional data points. It should be noted that genotype–phenotype inference with our method only takes into account the presence and absence of protein families of the bacteria analyzed. This information can be readily computed from the genomic and metagenomic data, but future research could focus also on integration of other ‘omics’ data to allow even more accurate phenotype

assignments. Additionally, expert knowledge of the biochemical pathways that are used in manual metabolic reconstructions, for example, could be integrated as prior knowledge into the model in future studies.

We have developed a highly accurate prediction framework for a large number of phenotypes from the protein content of bacterial genomes. Our approach provides a novel and fully automated way for microbiologists to characterize the rapidly increasing number of microbial genomes originating from genome and metagenome studies. Revealing the catabolic and anabolic capabilities and other phenotype-related information of bacteria with Traitar, such as the presence of antibiotic resistances, the ability to survive at high temperatures or motility, could provide insights into bacterial lifestyles, could lead to models of microbial interactions in a natural environment and inference of the conditions required to obtain bacteria in pure cultures. Furthermore, the present phenotype prediction framework offers a path to understanding the variation in microbiomes and may help to discover the traits and the associated protein families under selection.

Materials & Methods

We downloaded the coding sequences of all complete bacterial genomes that were available via the NCBI FTP server under ftp://ftp.ncbi.nlm.nih.gov/genomes/ as of 11 May 2014. These were annotated with the Traitar annotation procedure. For bacteria with more than one sequenced strain available, we chose the union of the Pfam family annotation of the single genomes to represent the pangenome Pfam family annotation as proposed in (Liu et al., 2006). We obtained our phenotype data from the GIDEON database (Berger, 2005). The GIDEON traits can be grouped into growth, oxygen requirement, morphology, carbon source utilization, antibiotic susceptibility, amino acid degradation, proteolysis, carboxylic acid use and enzymatic

activity (Table 1, Supplementary Table 1). We only considered phenotypes that were available in GIDEON for at least 20 bacteria, with a minimum of 10 bacteria annotated as positive (phenotype presence) for a given phenotype and 10 as negative (phenotype absence) to enable a robust and reliable analysis of the respective phenotypes. Furthermore, to be included in the analysis, we required each bacterial sample to have:

a) at least one annotated phenotype,
b) at least one sequenced strain,
c) a representative in the sTOL.

In total, we extracted 234 species-level bacterial samples with 67 phenotypes with sufficient total, positive and negative labels from GIDEON. GIDEON associates these bacteria with 9305 individual phenotype labels, 2971 being positive and 6334 negative (Supplementary Table 4). GIDEON species that had at least one sequenced strain available but were not part of the sTOL tree were set aside for a later independent assessment of the classification accuracy. All sequenced strains were annotated with Pfam protein families (Finn et al., 2014) (see Results - Annotation). The number of unique Pfam families per species-level (pan)genome ranged from 547 to 3610 (see Results – Annotation).

Cross-validation

We employed cross-validation to assess the performance of the classifiers individually for each phenotype. For a given phenotype, we divided the bacterial samples that were annotated with that phenotype into ten folds. Each fold was selected once for testing the model, which was trained on the remaining folds. The optimal regularization parameter *C* needed to be determined independently in each

step of the cross-validation; therefore, we employed a further inner cross-validation using the following range of values for the parameter *C*: 10^{-3}, $10^{-2} \cdot 0.7$, $10^{-2} \cdot 0.5$, $10^{-2} \cdot 0.2$, $10^{-2} \cdot 0.1$, …, 1. In other words, for each fold kept out for testing in the outer cross-validation, we determined the value of the parameter *C* that gave the best accuracy in an additional tenfold cross-validation on the remaining folds. This value was then used to train the SVM model in the current outer cross-validation step. Whenever we proceeded to a new cross-validation fold, we re-computed the ancestral character state reconstruction of the phenotype with only the training samples included (see Ancestral protein family and phenotype gains and losses). This procedure is known as nested cross-validation (Ruschhaupt et al., 2004).

The bacterial samples in the training folds imply a Gps-sTOL in each step of the inner and outer cross-validation without the samples in the test fold. We used the same procedure as before to map the Pfam gains and losses inferred previously on the Gps-sTOL onto the tree defined by the current cross-validation training folds. Importantly, the test error is only estimated on the observed phenotype labels rather than on the inferred phenotype gains and losses.

Evaluation metrics

We used evaluation metrics from multi-label classification theory for performance evaluation (Manning et al., 2008). We determined the performance for the individual phenotype-positive and the phenotype-negative classes based on the confusion matrix of true positive (*TP*), true negative (*TN*), false negative (*FN*) and false positive (*FP*) samples from their binary classification equivalents by averaging over all n phenotypes. We utilized two different accuracy measures for assessing multi-class classification performance (i.e. the accuracy pooled over all classification decisions and the macro-accuracy). Macro-accuracy represents an average over the accuracy

of the individual binary classification problems and we computed this from the macro-recall of the phenotype-positive and the phenotype-negative classes as follows:

$$Macro\text{-}recall_{Pos} = \left(\sum_{i=1}^{n} \frac{TP_i}{TP_i + FN_i}\right)/n$$

$$Macro\text{-}recall_{Neg} = \left(\sum_{i=1}^{n} \frac{TN_i}{FP_i + TN_i}\right)/n$$

$$Macro\text{-}accuracy = \left(Macro\text{-}recall_{Pos} + Macro\text{-}recall_{Neg}\right)/2.$$

However, if there are only few available labels for some phenotypes, the variance of the macro-accuracy will be high and this measure cannot be reliably computed anymore; it cannot be computed at all if no labels are available. The accuracy only assesses the overall classification performance without consideration of the information about specific phenotypes. Large classes dominate small classes (Manning et al., 2008).

$$Recall_{Pos} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FN_i}$$

$$Recall_{Neg} = \frac{\sum_{i=1}^{n} TN_i}{\sum_{i=1}^{n} TN_i + \sum_{i=1}^{n} FP_i}$$

$$Accuracy = \left(Recall_{Pos} + Recall_{Neg}\right)/2.$$

Majority feature selection

The weights in linear SVMs can directly be linked to features that are relevant for the classification. We identified the most important protein families used as features from the voting committee of SVMs consisting of the five most accurate models, which

were also used for classifying new samples. If the majority, which is at least three predictors, included a positive value for a given protein family, we added this feature to the list of important features. We further ranked these protein families features by their correlation with the phenotype using Pearson's correlation coefficient.

Acknowledgements

We would like to thank Andreas Klötgen, David Lähnemann, Susanne Reimering and Alex Sczyrba for providing helpful comments on the manuscript; Johannes Dröge and Jens Loers for reviewing the Traitar software and Garry Robertson for helping to set up the Traitar web service. JAF and PBP are supported by a grant from the European Research Council (336355-MicroDE).

Referencess

Alneberg, J., Bjarnason, B.S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U.Z., Lahti, L., Loman, N.J., Andersson, A.F., and Quince, C. (2014). Binning metagenomic contigs by coverage and composition. Nat Methods doi:10.1038/nmeth.3103.

Barker, D., and Pagel, M. (2005). Predicting functional gene links from phylogenetic-statistical analyses of whole genomes. PLoS computational biology doi:10.1371/journal.pcbi.0010003.

Berger, S.A. (2005). GIDEON: a comprehensive Web-based resource for geographic medicine. International journal of health geographics doi:10.1186/1476-072X-4-10.

Boser, B.E., Guyon, I.M., and Vapnik, V.N. (1992). A training algorithm for optimal margin classifiers. Paper presented at: Proceedings of the fifth annual workshop on computational learning theory (Association for Computing Machinery).

Brown, C.T., Hug, L.A., Thomas, B.C., Sharon, I., Castelle, C.J., Singh, A., Wilkins, M.J., Wrighton, K.C., Williams, K.H., and Banfield, J.F. (2015). Unusual biology across a group comprising more than 15% of domain Bacteria. Nature doi:10.1038/nature14486.

Cleary, B., Brito, I.L., Huang, K., Gevers, D., Shea, T., Young, S., and Alm, E.J. (2015). Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning. Nat Biotechnol doi:10.1038/nbt.3329.

Cohen, O., and Pupko, T. (2010). Inference and characterization of horizontally transferred gene families using stochastic mapping. Molecular biology and evolution doi:10.1093/molbev/msp240.

Cohen, O., and Pupko, T. (2011). Inference of gain and loss events from phyletic patterns using stochastic mapping and maximum parsimony--a simulation study. Genome biology and evolution doi:10.1093/gbe/evr101.

Draper, J., Karplus, K., and Ottemann, K.M. (2011). Identification of a chemoreceptor zinc-binding domain common to cytoplasmic bacterial chemoreceptors. Journal of bacteriology doi:10.1128/JB.05140-11.

Dutta, R., Qin, L., and Inouye, M. (1999). Histidine kinases: diversity of domain organization. Molecular microbiology *34*, 633-640

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). LIBLINEAR: A library for large linear classification. J Mach Learn Res *9*, 1871-1874

Fang, H., Oates, M.E., Pethica, R.B., Greenwood, J.M., Sardar, A.J., Rackham, O.J., Donoghue, P.C., Stamatakis, A., de Lima Morais, D.A., and Gough, J. (2013). A daily-updated tree of (sequenced) life as a reference for genome research. Scientific reports doi:10.1038/srep02015.

Feldbauer, R., Schulz, F., Horn, M., and Rattei, T. (2015). Prediction of microbial phenotypes based on comparative genomics. BMC bioinformatics doi:10.1186/1471-2105-16-S14-S1.

Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J.*, et al.* (2014). Pfam: the protein families database. Nucleic acids research doi:10.1093/nar/gkt1223.

Finn, R.D., Clements, J., and Eddy, S.R. (2011). HMMER web server: interactive sequence similarity searching. Nucleic acids research doi:10.1093/nar/gkr367.

Frank, J.A., Pan, Y., Tooming-Klunderud, A., Eijsink, V.G.H., McHardy, A.C., Nederbragt, A.J., and Pope, P.B. (2015). Improved metagenome assemblies and taxonomic binning using long-read circular consensus sequence data. bioRxiv doi:10.1101/026922.

Gilead, S., and Shoham, Y. (1995). Purification and characterization of alpha-L-arabinofuranosidase from Bacillus stearothermophilus T-6. Applied and environmental microbiology *61*, 170-174

Goodfellow, M., Kämpfer, P., Busse, H.-J., Trujillo, M.E., Suzuki, K.-i., Ludwig, W., and Whitman, W.B. (2012). Bergey's manual of systematic bacteriology (Springer New York).

Gough, J., Karplus, K., Hughey, R., and Chothia, C. (2001). Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. Journal of molecular biology doi:10.1006/jmbi.2001.5080.

Gregor, I., Droge, J., Schirmer, M., Quince, C., and McHardy, A.C. (2016). PhyloPythiaS+: a self-training method for the rapid reconstruction of low-ranking taxonomic bins from metagenomes. PeerJ doi:10.7717/peerj.1603.

Harvey, P.H., and Pagel, M.D. (1991). The comparative method in evolutionary biology, Vol 239 (Oxford University Press Oxford).

Hess, M., Sczyrba, A., Egan, R., Kim, T.W., Chokhawala, H., Schroth, G., Luo, S., Clark, D.S., Chen, F., Zhang, T., *et al.* (2011). Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. Science doi:10.1126/science.1200387.

Hosking, E.R., Vogt, C., Bakker, E.P., and Manson, M.D. (2006). The Escherichia coli MotAB proton channel unplugged. Journal of molecular biology doi:10.1016/j.jmb.2006.09.035.

Human Microbiome Project Consortium (2012). Structure, function and diversity of the healthy human microbiome. Nature doi:10.1038/nature11234.

Hyatt, D., Chen, G.L., Locascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC bioinformatics doi:10.1186/1471-2105-11-119.

Imelfort, M., Parks, D., Woodcroft, B.J., Dennis, P., Hugenholtz, P., and Tyson, G.W. (2014). GroopM: an automated tool for the recovery of population genomes from related metagenomes. PeerJ doi:10.7717/peerj.603.

Josenhans, C., and Suerbaum, S. (2002). The role of motility as a virulence factor in bacteria. International journal of medical microbiology : IJMM doi:10.1078/1438-4221-00173.

Kang, D.D., Froula, J., Egan, R., and Wang, Z. (2015). MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ doi:10.7717/peerj.1165.

Kastenmuller, G., Schenk, M.E., Gasteiger, J., and Mewes, H.W. (2009). Uncovering metabolic pathways relevant to phenotypic traits of microbial genomes. Genome biology doi:10.1186/gb-2009-10-3-r28.

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. Biotechnology for biofuels doi:10.1186/s13068-014-0124-8.

Lam, W.W., Woo, E.J., Kotaka, M., Tam, W.K., Leung, Y.C., Ling, T.K., and Au, S.W. (2010). Molecular interaction of flagellar export chaperone FliS and cochaperone HP1076 in Helicobacter pylori. FASEB journal : official publication of the Federation of American Societies for Experimental Biology doi:10.1096/fj.10-155242.

Lasken, R.S., and McLean, J.S. (2014). Recent advances in genomic DNA sequencing of microbial species from single cells. Nat Rev Genet doi:10.1038/nrg3785.

Limam, R.D., Chouari, R., Mazeas, L., Wu, T.D., Li, T., Grossin-Debattista, J., Guerquin-Kern, J.L., Saidi, M., Landoulsi, A., Sghir, A., *et al.* (2014). Members of the uncultured bacterial candidate division WWE1 are implicated in anaerobic digestion of cellulose. MicrobiologyOpen doi:10.1002/mbo3.144.

Lingner, T., Muhlhausen, S., Gabaldon, T., Notredame, C., and Meinicke, P. (2010). Predicting phenotypic traits of prokaryotes from protein domain frequencies. BMC bioinformatics doi:10.1186/1471-2105-11-481.

Liu, R., and Ochman, H. (2007). Stepwise formation of the bacterial flagellar system.

Proceedings of the National Academy of Sciences of the United States of America doi:10.1073/pnas.0700266104.

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 computational biology doi:10.1371/journal.pcbi.0020159.

MacDonald, N.J., and Beiko, R.G. (2010). Efficient learning of microbial genotype-phenotype association rules. Bioinformatics doi:10.1093/bioinformatics/btq305.

Macnab, R.M. (2003). How bacteria assemble flagella. Annual review of microbiology doi:10.1146/annurev.micro.57.030502.090832.

Manning, C.D., Raghavan, P., and Schütze, H. (2008). Introduction to information retrieval, Vol 1 (Cambridge university press Cambridge, UK).

Martinez, D., Berka, R.M., Henrissat, B., Saloheimo, M., Arvas, M., Baker, S.E., Chapman, J., Chertkov, O., Coutinho, P.M., Cullen, D.*, et al.* (2008). Genome sequencing and analysis of the biomass-degrading fungus Trichoderma reesei (syn. Hypocrea jecorina). Nat Biotechnol doi:10.1038/nbt1403.

Martinez, J.L. (2008). Antibiotics and antibiotic resistance genes in natural environments. Science doi:10.1126/science.1159483.

Martiny, J.B., Jones, S.E., Lennon, J.T., and Martiny, A.C. (2015). Microbiomes in light of traits: A phylogenetic perspective. Science doi:10.1126/science.aac9323.

Moreno-Vivian, C., Cabello, P., Martinez-Luque, M., Blasco, R., and Castillo, F. (1999). Prokaryotic nitrate reduction: molecular properties and functional distinction among bacterial nitrate reductases. Journal of bacteriology *181*, 6573-6584

Nielsen, H.B., Almeida, M., Juncker, A.S., Rasmussen, S., Li, J., Sunagawa, S., Plichta, D.R., Gautier, L., Pedersen, A.G., Le Chatelier, E.*, et al.* (2014). Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat Biotechnol doi:10.1038/nbt.2939.

Ochman, H., Lawrence, J.G., and Groisman, E.A. (2000). Lateral gene transfer and the nature of bacterial innovation. Nature doi:10.1038/35012500.

Pal, C., Papp, B., and Lercher, M.J. (2005). Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. Nature genetics doi:10.1038/ng1686.

Pantel, I., Lindgren, P.E., Neubauer, H., and Gotz, F. (1998). Identification and characterization of the Staphylococcus carnosus nitrate reductase operon. Molecular & general genetics : MGG *259*, 105-114

Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., and Tyson, G.W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome research doi:10.1101/gr.186072.114.

Paterson, G.K., and Mitchell, T.J. (2004). The biology of Gram-positive sortase enzymes. Trends in microbiology doi:10.1016/j.tim.2003.12.007.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., and Dubourg, V. (2011). Scikit-learn: Machine learning in Python. J Mach Learn Res *12*, 2825-2830

Pope, P.B., Smith, W., Denman, S.E., Tringe, S.G., Barry, K., Hugenholtz, P., McSweeney, C.S., McHardy, A.C., and Morrison, M. (2011). Isolation of Succinivibrionaceae implicated in low methane emissions from Tammar wallabies. Science doi:10.1126/science.1205760.

Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., *et al.* (2012). The Pfam protein families database. Nucleic acids research doi:10.1093/nar/gkr1065.

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 doi:10.1038/nature12352.

Ruschhaupt, M., Huber, W., Poustka, A., and Mansmann, U. (2004). A compendium to ensure computational reproducibility in high-dimensional classification tasks. Statistical applications in genetics and molecular biology doi:10.2202/1544-6115.1078.

Sa-Nogueira, I., Nogueira, T.V., Soares, S., and de Lencastre, H. (1997). The Bacillus subtilis L-arabinose (ara) operon: nucleotide sequence, genetic organization and expression. Microbiology *143 (Pt 3)*, 957-969

Tange, O. (2011). GNU parallel-the command-line power tool. USENIX *36*, 42-47

Tatusov, R.L., Natale, D.A., Garkavtsev, I.V., Tatusova, T.A., Shankavaram, U.T., Rao, B.S., Kiryutin, B., Galperin, M.Y., Fedorova, N.D., and Koonin, E.V. (2001). The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic acids research doi:10.1093/nar/29.1.22.

Weimann, A., Trukhina, Y., Pope, P.B., Konietzny, S.G., and McHardy, A.C. (2013). De novo prediction of the genomic components and capabilities for microbial plant biomass degradation from (meta-)genomes. Biotechnology for biofuels doi:10.1186/1754-6834-6-24.

Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. J R Stat Soc A *67*, 301-320