Bioinformatics, 32(6), 2016, 867–874
doi: 10.1093/bioinformatics/btv671
Advance Access Publication Date: 14 November 2015
Original Paper

OXFORD

Systems biology

Metabolic network-guided binning of metagenomic sequence fragments

Matthew B. Biggs and Jason A. Papin*

Department of Biomedical Engineering, University of Virginia, Charlottesville, VA, 22903 USA

*To whom correspondence should be addressed.

Associate Editor: Igor Jurisica

Received on July 13, 2015; revised on October 16, 2015; accepted on November 9, 2015

Abstract

Motivation: Most microbes on Earth have never been grown in a laboratory, and can only be studied through DNA sequences. Environmental DNA sequence samples are complex mixtures of fragments from many different species, often unknown. There is a pressing need for methods that can reliably reconstruct genomes from complex metagenomic samples in order to address questions in ecology, bioremediation, and human health.

Results: We present the SOrting by NEtwork Completion (SONEC) approach for assigning reactions to incomplete metabolic networks based on a metabolite connectivity score. We successfully demonstrate proof of concept in a set of 100 genome-scale metabolic network reconstructions, and delineate the variables that impact reaction assignment accuracy. We further demonstrate the integration of SONEC with existing approaches (such as cross-sample scaffold abundance profile clustering) on a set of 94 metagenomic samples from the Human Microbiome Project. We show that not only does SONEC aid in reconstructing species-level genomes, but it also improves functional predictions made with the resulting metabolic networks.

Availability and implementation: The datasets and code presented in this work are available at: https://bitbucket.org/mattbiggs/sorting_by_network_completion/.

Contact: papin@virginia.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Most microbes cannot be cultured using existing techniques (Handelsman, 2004). It is possible to interrogate this vast world of 'unculturables' by analysis of DNA from environmental samples. Metagenomics is a burgeoning field, and databases are accumulating trillions of bases of DNA sequence from complex environmental samples. These DNA fragments contain information about new and interesting microbes. Many approaches for analyzing such complex mixtures of DNA fragments seek to catalog the families of genes contained in the community metagenome, and how those families of genes change over time (Abubucker *et al.*, 2012; Greenblum *et al.*, 2012; Owen *et al.*, 2015). Other approaches seek to assign DNA fragments to known taxonomic groups (Afshinnekoo *et al.*, 2015; Yarza *et al.*, 2014). What is more difficult is the assignment of DNA fragments—genes in particular—to yet undiscovered parent genomes, and as a result, discovering the context in which those genes operate. The goal is not only to know that a given gene exists within the community, but to know also to which species that gene belongs, what other genes that species has, what metabolic capacity that species presents, the regulatory network that controls those genes and so on. Answers to these questions will advance efforts to discover new pathogens, industrially-relevant microbes and drivers of global geo-chemical cycles (Kinross *et al.*, 2011; Rousk and Bengtson, 2014; Smid *et al.*, 2014).

Recent advances in reconstructing species-level genomes from metagenomic samples have relied on several sources of information: nucleotide patterns that differentiate species, such as G/C content and tetranucleotide frequencies (Fig. 1) (Iverson *et al.*, 2012; Teeling *et al.*, 2004); taxonomic assignment based on similar, known genomes (MacDonald *et al.*, 2012); improved fragment assembly

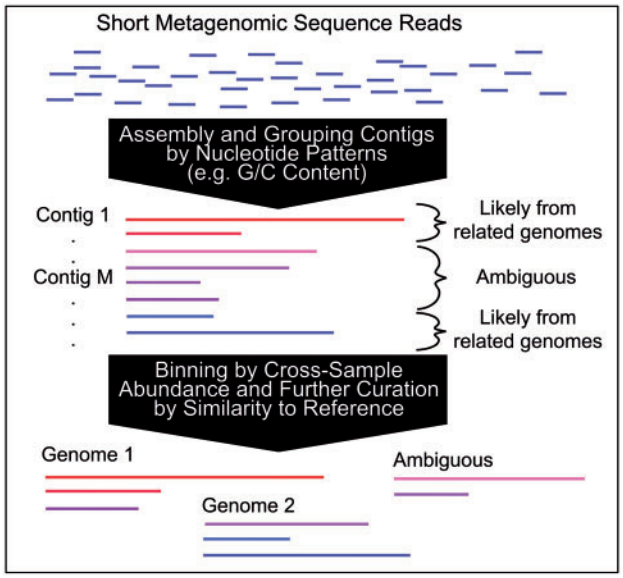

Fig. 1. Current approaches to reassembling species-level genomes from metagenomic data include: assembly, where short reads are assembled into larger fragments ('contigs') by sequence overlap; grouping by sequence composition, where fragment similarity is gauged by nucleotide sequence patterns (e.g. G/C content or tetranucleotide frequencies); clustering by cross-sample abundance profiles, where fragments with strongly correlated abundance across independent samples are grouped together; further curation can include mapping to closely-related reference genomes, taxonomic annotations, or testing that bins contain minimal gene sets common to most organisms

(Namiki *et al.*, 2012); and differential scaffold abundance across multiple samples (Albertsen *et al.*, 2013; Alneberg *et al.*, 2014; Carr *et al.*, 2013; Nielsen *et al.*, 2014; Sharon *et al.*, 2013). The best approaches to-date use all of these sources of information to extract high quality, species- or strain-specific genomes (Nielsen *et al.*, 2014). While the best current approaches have demonstrated the ability to extract hundreds of genomes from a complex community such as the human gut, they still leave a third of the available DNA fragments unassigned (Nielsen *et al.*, 2014).

We propose a new, orthogonal source of information that can be used to further improve species genome reconstruction, in conjunction with existing approaches. Metabolic networks are assumed to be effectively complete (i.e. gapless) (Krumholz and Libourel, 2015; Pitkänen *et al.*, 2014; Satish Kumar *et al.*, 2007; Thiele and Palsson, 2010). This assumption of network completeness is a physiological equivalent of the law of conservation of mass: that is, that mass drawn into a cell must eventually leave or be integrated into biomass. Thus, real metabolic networks do not contain 'dead end' metabolites—reaction substrates or products that are exclusively consumed or produced (Krumholz and Libourel, 2015; Satish Kumar *et al.*, 2007). This fact can be leveraged in the assignment of metagenomic fragments to species bins. Given a set of bins containing genetic fragments (formed using orthogonal sources of information as described above), and a set of unassigned fragments, a metabolic network can be reconstructed based on the gene content of each bin, and new fragments assigned to these bins based on a metabolite connectivity metric. The underlying assumption driving this approach is that genetic fragments containing metabolic genes will tend to fill gaps in the correct host metabolic network, and will be less likely to fill gaps in a foreign network to which they do not belong.

We refer to this approach as SOrting by NEtwork Completion (SONEC). We present proof-of-principle results from the successful application of this method using a set of 100 genome-scale metabolic network reconstructions. Furthermore, we demonstrate the application of this approach to 94 metagenomic samples from the Human Microbiome Project (Huttenhower *et al.*, 2012). These computational experiments highlight the utility of this novel method, and delineate the sensitivity to variables that impact practical applications.

2 Methods

2.1 Obtaining metabolic network reconstructions

All metabolic network reconstructions were generated by the Model SEED server (Overbeek *et al.*, 2005). These were downloaded as spreadsheets and converted to Matlab objects using custom scripts, available in Supplementary Material (MATLAB and Statistics Toolbox, 2012). To generate network reconstructions for each individual cluster in the anterior nares dataset, the set of 9910 assembled contigs was uploaded to the model SEED server. The reactions for each cluster were assigned by mapping the annotated open reading frames to the gene-protein-reaction associations in the meta-reconstruction. All 100 single-species reconstructions are publically available through the model SEED, and our copies of all reconstructions are also available as Matlab objects.

2.2 Proof-of-concept simulations

All simulations were performed using custom scripts in Matlab R2013a on a machine running 64-bit Windows 7, 32 GB RAM and 3.6 GHz processor speed. Confidence intervals were calculated in R (R: A language and environment for statistical computing, 2013). All scripts and data are available in the Supplementary Materials.

2.3 Binary error estimation

We organized errors into the following categories: True Positives (TP) result from the case where reactions were unambiguously assigned to the correct parent network; False Positives (FP) result from the case where reactions were unambiguously assigned to an incorrect network; True but Ambiguous (TA) results from the case where there were one or more ties in the maximum metabolite connectivity score (for definition of 'metabolite connectivity score', see Section 3.1), and included the correct parent network; False and Ambiguous (FA) results from the case where there were one or more ties in the maximum metabolite connectivity score, none of which were the correct parent network; True Rejection (TR) results from the case where there was a metabolite connectivity score of zero for all networks and the rejected reaction originated from a shadow network (and thus, was correctly rejected; for definition of 'shadow network', see Section 3.1); False Rejection (FR) results from the case where there was a metabolite connectivity score of zero for all networks, but the rejected reaction originated from one of the visible networks and so was incorrectly rejected. All error bars represent the 95% confidence interval for the observed accuracy. Because the assignments resulted in binary outputs (correct assignment or not), confidence intervals were estimated using the Wilson score interval (Agresti and Coull, 1998) in R.

2.4 Obtaining metagenomic samples

Illumina whole-genome shotgun reads were obtained from the Human Microbiome Project database (Huttenhower *et al.*, 2012). All 94 samples corresponding to the anterior nares were downloaded, while 49 samples containing more than one million reads were used to estimate coverage of assembled fragments.

These 49 samples were each reduced to one million reads in order to normalize coverage estimates. This was done by randomly selecting one million reads from the total sample using a custom Python script (available in the Supplementary Material). The methods pertaining to the complete analysis of this metagenomic dataset can be found in the Supplementary Materials.

2.5 *In silico* reaction essentiality screen

Reaction essentiality was determined by setting the upper and lower flux bounds to zero for each reaction in turn. Flux Balance Analysis was performed using the COBRA toolbox for Matlab (Schellenberger *et al.*, 2011) and the Gurobi Optimizer (2015). Reactions were considered essential if, when the reaction was prevented from carrying flux, flux through biomass was also reduced to zero. Visualization of the metabolic network and essential reactions was performed using MetDraw (Jensen and Papin, 2014). Our data and code are available in the Supplementary Material.

3 Results

3.1 Algorithm

We define a metabolite connectivity score (MCS) for reaction i with respect to metabolic network j as:

$$MCS_{ij} = \frac{|RS_i \cap NC_j|}{|RS_i|} + \frac{|RP_i \cap NP_j|}{|RP_i|} \quad (1)$$

where RS_i is the set of substrates for reaction i ($|RS_i|$ is the number of substrates for reaction i), RP_i is the set of products for reaction i, NC_j is the set of metabolites that are not consumed by any reaction in network j, and NP_j is the set of metabolites that are not produced by any reaction in network j. $\cap$ indicates the intersection between sets. Given an unassigned reaction and a set of metabolic networks, the metabolite connectivity score is calculated for each network and the reaction is assigned to the network with the maximum metabolite connectivity score (Fig. 2). In the case of a tie, the correct assignment is ambiguous. In this work we chose to only assign reactions with unambiguous metabolite connectivity scores, but the algorithm could be readily adapted to make more liberal assignments.

Additionally, the concepts of 'groups' and 'shadow networks' are important for understanding the proof-of-concept simulations that follow. We define a 'group' as a set of reactions that originate from the same metabolic network. A group can be thought of as a set of metabolic reactions that are obtained from genes on the same contiguous metagenomic sequence fragment (or 'contig'), thus we can be confident that these genes come from the same parent organism. A group metabolite connectivity score is defined as the sum of the scores for each individual reaction: $MCS_{kj} = \sum_i^N MCS_{ij}$ where MCS_{kj} is the metabolite connectivity score for group k (of size N reactions) with respect to metabolic network j, and MCS_{ij} is the metabolite connectivity score for reaction i (within group k) with respect to metabolic network j.

We define 'shadow networks' as a pool of metabolic networks which contribute reactions to the metagenome, but which are not considered as potential bins to which reactions can be assigned. For example, consider a metagenomic dataset with many high-abundance species and several low-abundance species. A bin can be created corresponding to each high-abundance species because there is sufficient signal in the dataset. However, species of very low-abundance in the community are probably not sequenced to sufficient depth to be assigned their own bins (in other words, this is a 'shadow species'). Because the sequence fragments from these low-abundance species cannot be assigned to their own bins, they may be incorrectly assigned to bins of high-abundance species (because bins corresponding to high-abundance species are the only available choices for assignment). Including these 'shadow networks' in our simulations allows us to evaluate the strength of the MCS in differentiating reactions that do not originate from any available choice of reconstruction.

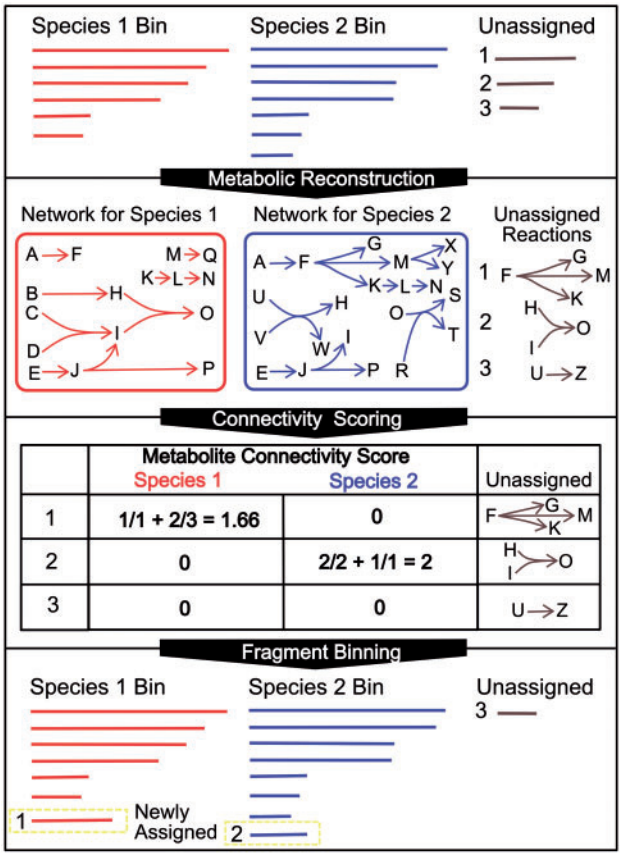

Fig. 2. The SONEC Algorithm. The algorithm is initialized with bins of contigs (where the bins correspond to species from the metagenome), and a set of unassigned sequence fragments. These initial bins can be produced using existing methods, and the unassigned fragments would be those which existing methods cannot confidently assign to a species bin. A metabolic network reconstruction is produced for each bin and all unassigned contigs. To determine the correct parent bin to which unassigned contigs should be assigned, a metabolite connectivity score is calculated for each pair of unassigned reaction and parent network. This metabolite connectivity score quantifies the number of dead-end metabolites in the parent network which would no longer be dead-end with the addition of the unassigned reaction. Unassigned reactions will remove more dead-end metabolites, on average, from the correct parent network than from other, off-target networks. If there is a single maximum metabolite connectivity score for a given reaction, the contig associated with that reaction is assigned to the parent bin indicated by the metabolite connectivity score (e.g. unassigned contig 1 is assigned to species bin 1, and unassigned contig 2 is assigned to species bin 2, while unassigned contig 3 is ambiguous and cannot be assigned)

While the analysis below demonstrates the value of SONEC, we provide here specific examples of binning based on the MCS to highlight the functionality and caveats of this scoring scheme (Supplemental Fig. S1). Beginning with a set of 10 draft-quality metabolic network reconstructions, we randomly removed reactions from each and used the MCS to assign these reactions back to a metabolic network. As an example of a true positive result, the MCS was calculated for a reaction catalyzed by a 5-phosphomevalonate

phosphotransferase drawn from *Enterococcus* sp. GMD1E (Supplemental Fig. S1A). The metabolic network for *Enterococcus* sp. GMD1E was the only network of 10 that contained dead-end metabolites that overlapped with products of the reaction. In this case, diphosphomevalonate was not produced by any reaction in the *Enterococcus* network, and the MCS captured this complementary overlap with the reaction products, resulting in a correct assignment.

In contrast, an example of a false positive result is informative (Supplemental Fig. S1B). The reaction catalyzed by a nicotinate-nucleotide dimethylbenzimidazole phosphoribosyltransferase overlapped with dead-end metabolites in several networks. In the correct parent network of *Shigella flexneri*, the reaction product—alpha-ribazole 5'-phosphate—was an unproduced metabolite. Because there were three products in the reaction, the MCS is 0.33. Conversely, in the network for *Pelagibacter ubique*, a reaction substrate nicotinate ribonucleotide was an unconsumed metabolite. Because there are only two substrates in the reaction, the MCS is 0.5, and because this was the maximum, the reaction was incorrectly assigned to *P. ubique*. These specific examples of true and false positives exhibit how the MCS works in practice. We performed further simulations which help to elucidate the role of variables that influence reaction assignment accuracy using the MCS.

3.2 Proof-of-concept simulations

We simulated the problem of binning metagenomic samples into appropriate species bins. For each independent simulation, we started with a set of draft-quality metabolic network reconstructions randomly drawn from among 100 bacterial networks obtained from the Model SEED. We randomly removed reactions from each. We used the MCS to assign these reactions back to a metabolic network reconstruction. Unless otherwise noted, accuracy of assignment was evaluated over 1000 independent simulations for every unique combination of parameters. While the simulations presented here are performed with networks from the Model SEED, the same analysis could be performed with networks derived from other resources such as KEGG or Pathway Tools (Kanehisa *et al.*, 2010; Karp *et al.*, 2010), ideally with more coverage of known microbial taxa.

We first evaluated the effect of parent network completeness on reaction assignment accuracy (Fig. 3A). We randomly removed increasingly large subsets of reactions from each of 10 metabolic networks. These reactions were then assigned to a network. More complete parent networks (fewer reactions removed from the original) produced more true positive, and fewer false positive, reaction assignments. As expected, as parent networks become less complete, assignment accuracy diminishes with decreases in true positives and increases in false positives. Each simulation was repeated 1000 times, with a new set of 10 parent networks being selected randomly each time from the pool of 100 networks. We display results from group sizes (number of reactions annotated from the same sequence fragment or contig) of 1, 20 and 40 (Fig. 3A).

Next, we investigated the impact of increasing the number of parent networks from which unassigned reactions were derived (Fig. 3B). For these simulations, the fraction of reactions removed was fixed at 0.15 and the group size fixed at 25. As the number of networks increased, we observed corresponding decreases in the number of true positives and increases in the number of false positive reaction assignments.

We further evaluated the effect of increasing group size (Fig. 3C). The fraction of reactions removed was fixed at 0.15, and the number of parent networks was fixed at 10. We observed that increasing the group size improved assignment accuracy. Group

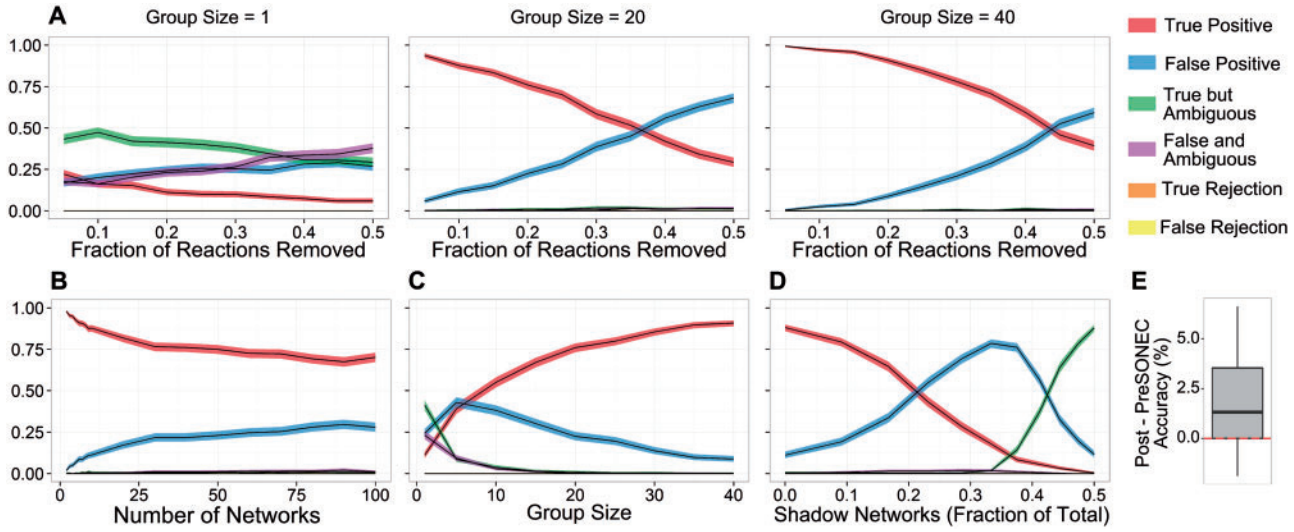

Fig. 3. Reaction assignment accuracy from simulations of the SONEC approach. (**A**) The accuracy is displayed as a function of the number of reactions that were removed from the parent networks (shown as a fraction of parent network size). An increasing number of reactions were removed from the total reaction content of 10 randomly-selected metabolic networks. Results for group sizes (the number of reactions being assigned together) of 1, 20 and 40 are shown. (**B**) Accuracy is displayed as a function of the number of parent bins to which a reaction could potentially be assigned. The fraction of reactions removed was fixed at 0.15, and the group size fixed at 25. (**C**) Accuracy is displayed as a function of reaction group size. The fraction of reactions removed was fixed at 0.15, and the number of parent networks was fixed at 10. **(D)** Accuracy is displayed as a function of the number of shadow networks. Shadow networks are a pool of network reconstructions from which reactions are contributed to the unassigned pool, but which are not available as bins to which those reactions can be assigned. The fraction of reactions removed was fixed at 0.15, the number of visible parent networks fixed at 10 and the group size fixed at 25. **(E)** Reaction essentiality predictions from gap-filled networks pre- and post-SONEC were compared to predictions from the full, reference networks. The difference in accuracy between paired experiments (post-SONEC—pre-SONEC) is shown here as a boxplot, with the null hypothesis (zero, no difference) indicated by the dashed, red line. SONEC improved the average accuracy by 1.8%, with a *P*-value of 2.7×10^{-6} (by paired, one-sided Wilcoxon Rank Sum test on 50 replicates). For A–D, all results are from 1000 independent replicates and shaded areas represent a 95% confidence interval around the mean, determined by the Wilson score interval (see Section 2)

sizes less than five tended to produce true but ambiguous assignments. Group sizes of six or greater produced mostly true positive assignments, with a steadily improving margin between true and false positives as group size increased. The interaction between network completeness and group size (or any other combination of parameters) can be evaluated extensively through further simulations (Supplemental Fig. S2).

We also explored the impact of shadow networks (networks which contribute reactions to the unassigned pool, but do not have a corresponding bin to which reactions can be assigned) (Fig. 3D). The fraction of reactions removed was fixed at 0.15, the number of parent networks fixed at 10 and the group size fixed at 25. Reactions drawn from shadow networks were included for assignment, but the shadow networks were not included as candidates to which reactions could be assigned. We observed an interesting pattern of assignment accuracy as the number of shadow networks increased (displayed as a fraction of the total population of networks). True positive assignments account for the majority, up until the number of shadow networks is equivalent ~0.2 of the population. Between 0.23 and 0.44, false positives account for the majority and from 0.44 to 0.5, true but ambiguous assignments form the majority. We also observed an interesting increase in false and ambiguous assignments that peaked at ~0.3.

Finally, we evaluated the impact of SONEC on functional network predictions by comparing reaction essentiality predictions from pre- and post-SONEC networks to the predictions from the full, parent network (Fig. 3E). In this set of simulations, the fraction of reactions removed was fixed at 0.15, the number of parent networks fixed at 10 and the group size fixed at 25, over 50 replicates. In each replicate, one network was chosen for evaluation. After reactions were removed, a copy of the incomplete network (the pre-SONEC network) was gap-filled (Reed *et al.*, 2006). Subsequently, all reactions assigned by SONEC were added to a separate copy of the imcomplete network (the post-SONEC network), which was then gap-filled. Reaction essentiality for the pre-SONEC, post-SONEC and full networks were all evaluated using the same biomass function and exchange flux bounds. The post-SONEC reaction essentiality predictions achieved accuracies 1.8% greater, on average, than the pre-SONEC predictions (P-value $= 2.7 \times 10^{-6}$ by paired, one-sided Wilcoxon Rank Sum test). The post-SONEC predictions were the same or better 80% of the time, and strictly better 68% of the time. In this case, we evaluated reaction essentiality rather than gene essentiality (a more common measure) due to the draft-quality status of the gene-protein-reaction relationships.

Example values of SONEC parameters (e.g. group size, network completeness, etc.) in existing metagenomic datasets are described in the supplemental materials.

3.3 Pathway enrichment

Pathway enrichment was performed to evaluate the contribution of different families of metabolic reactions to assignment accuracy (Supplemental Fig. 3 and Supplemental Methods). For these simulations, 10 parent networks were available for assignment, the fraction of reactions removed was fixed at 0.15, the group size fixed at 25, and there were no reactions from shadow networks. Sulfur metabolism, nicotinate and nicotinamide metabolism, galactose metabolism, porphyrin and chlorophyll metabolism, biosynthesis of steroids, and terpenoid biosynthesis contributed to true positive assignments more than expected by chance alone. Propanoate metabolism, pyruvate metabolism, amino sugars metabolism, and others contributed to more false positive assignments than expected by chance alone. Several pathways, including ubiquinone biosynthesis, D-glutamine and D-glutamate metabolism, were all enriched in both true positive and false positive assignments.

3.4 SONEC applied to metagenomic samples

We applied the SONEC approach to metagenomic sequences from 94 samples sourced from the human anterior nares as part of the Human Microbiome Project (Fig. 4) (Huttenhower *et al.*, 2012).

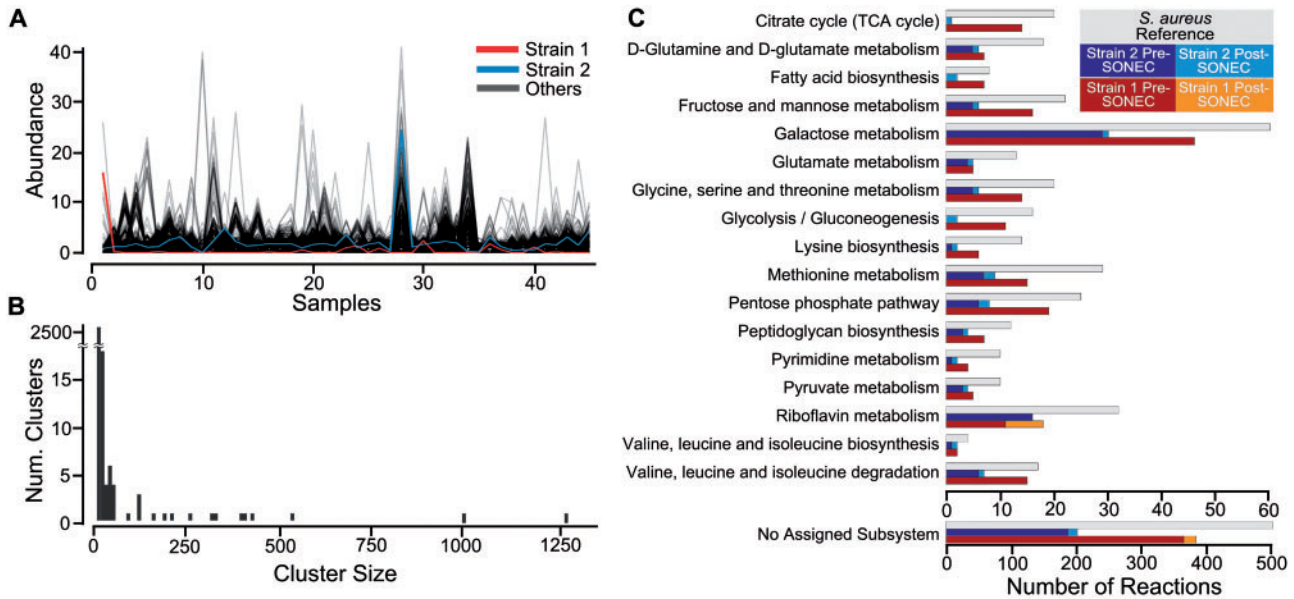

Fig. 4. Generating species-level clusters from the anterior nares metagenome data set. (**A**) Average cross-sample abundance profiles for all 2849 clusters after application of the canopy algorithm. The profiles for the two clusters which we refer to as Strain 1 and Strain 2 are highlighted in red and blue, respectively. (**B**) Histogram of cluster size (number of contigs) for all clusters after application of the canopy algorithm (Nielsen *et al.*, 2014). Note that most clusters are very small (2500 clusters with fewer than 10 contigs), while there are few very large clusters. (**C**) Reaction content of metabolic network reconstructions, organized by subsystem, for Strain 1 and Strain 2, before and after the application of SONEC. Reaction content from a reference network for *S.aureus* is provided

The short reads were assembled into contigs, and the abundance of each contig was estimated across the subset of 49 samples containing more than one million reads. The assembly process resulted in 1 543 959 contigs, with an N50 of 261. The N50 indicates the contig length at which all contigs of that length or greater contribute 50% of the cumulative length of the dataset. We continued the analysis with the 9910 contigs with length of 800 base pairs (bp) or greater and which had a non-zero abundance in at least 3 samples (Fig. 4A). These contigs were clustered into 2849 clusters, such that contigs within a cluster likely originated from the same organism (Fig. 4B). Metabolic network reconstructions were obtained for all clusters by uploading the corresponding contigs to the Model SEED server resulting in 14 083 annotations including open reading frames and RNA elements (Overbeek *et al.*, 2005). We observed 14 clusters with 90 or more annotated reactions and an average cumulative length of 682 232 bp. The remaining smaller clusters contained 44 or fewer annotated reactions and an average cumulative length of 2171 bp. The taxonomic content of each cluster was estimated, and the two large clusters with the most consistent taxonomic identity (Clusters 614 and 1357) corresponded to strains of *Staphylococcus aureus*. Cluster 614 (labeled 'Strain 1') contained 1001 assembled fragments with a cumulative length of 1 623 468 bp, 742 assigned metabolic reactions, and 100% of fragments aligned well to *S.aureus* genomes. Cluster 1357 (labeled 'Strain 2') contained 396 assembled fragments with a cumulative length of 479 515 bp, 326 assigned metabolic reactions, and 90% of fragments aligned well to *S.aureus* genomes. For reference, the complete genome for *S.aureus* Newman is 2.9 million bp long (Baba *et al.*, 2008). These two clusters were not correlated and thus, likely originated from different strains of *S.aureus* (Fig. 4A).

Comparing Strain 1 to a reference metabolic network for *S.aureus* N315 (obtained from the Model SEED) revealed 692 shared reactions of a possible 1118. Strain 1 contained 50 unique reactions that were not found in the reference metabolic network. These unique reactions were found in the following pathways: biosynthesis of steroids; butanoate metabolism; glycine, serine and threonine metabolism; pentose and glucuronate interconversions; pentose phosphate pathway. Strain 2 shared 319 reactions with the *S.aureus* N315 reference. Strain 2 contained a further 7 unique reactions in the following pathways: glutathione metabolism; pentose phosphate pathway; purine metabolism; pyrimidine metabolism; pyruvate metabolism.

After using established techniques (Albertsen *et al.*, 2013; Nielsen *et al.*, 2014) to identify Strain 1 and Strain 2, we applied SONEC to further complete these two clusters. Seven smaller clusters were identified as also originating from strains of *S.aureus*. These smaller clusters had cumulative lengths from 920 to 59 378 bp, were annotated with 10–42 metabolic reactions, and 100% of fragments aligned well to *S.aureus* genomes. The SONEC MCS was utilized to assign these smaller clusters to one of the larger *S.aureus* clusters. Five of the seven clusters produced non-zero metabolite connectivity scores and could be assigned unambiguously. Two were assigned to Strain 1 and three to Strain 2, which increased reaction overlap with the reference metabolic network for *S.aureus* N315 by 3.8% and 7.8% respectively. Many of the newly assigned reactions expanded core subsystems such as glycolysis and amino acid metabolism (Fig. 4C). The addition of these smaller clusters increased the total genetic content of Strain 1 by 2968 bp and Strain 2 by 69 913 bp. To determine the impact of these SONEC assignments on functional predictions of the resulting metabolic networks, we performed an *in silico* reaction essentiality screen on Strain 1 before and after the application of SONEC (Fig. 5). To begin, we identified an *S.aureus* minimal medium and a biomass function (Becker and Palsson, 2005). We performed gap filling based on the identified

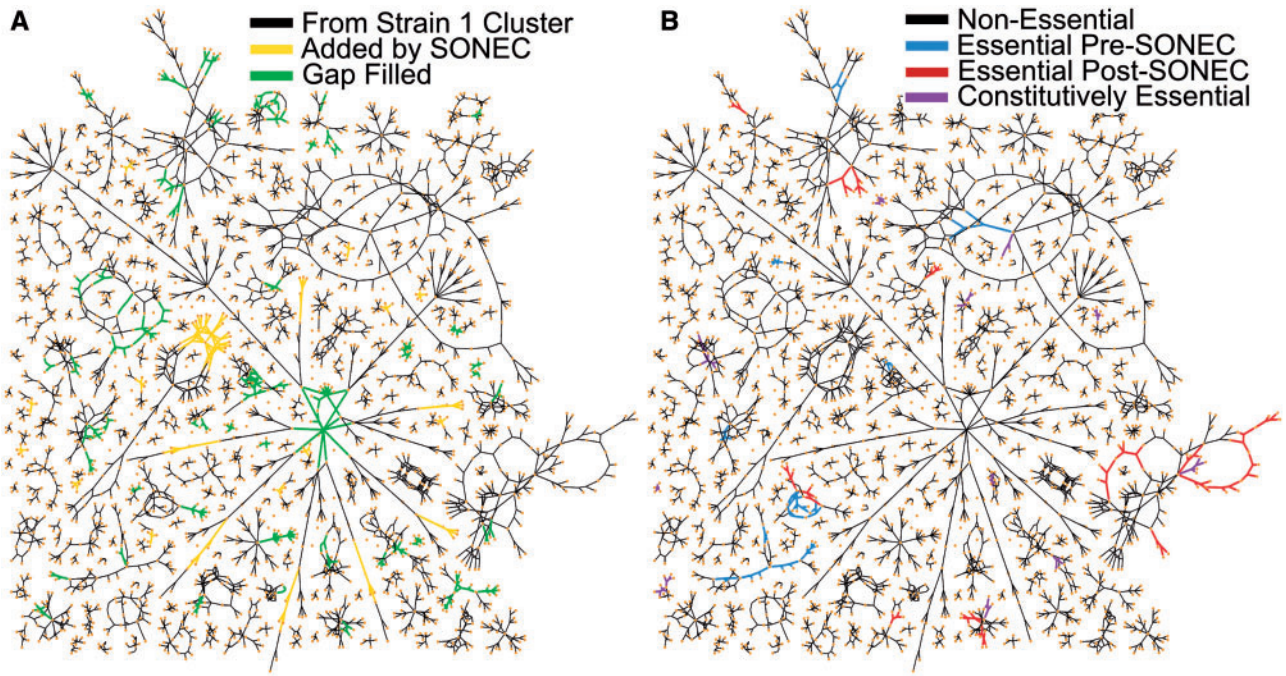

Fig. 5. Application of SONEC alters functional predictions of metabolic network. In both panels the metabolic network for Strain 1 is represented with metabolites as nodes (orange circles) and reactions as edges between metabolites. (**A**) Reactions are colored by source: black indicates reactions from the original cluster for Strain 1; yellow indicates reactions added by SONEC; and green reactions were added during the gap filling process. (**B**) Reactions are colored to indicate essentiality: black reactions are non-essential in all conditions; blue reactions were essential before the application of SONEC, but not after; red reactions were essential after the application of SONEC, but not before; and purple indicates reactions which were essential before and after SONEC

medium and biomass formulation using a custom implementation of a previously described gap fill algorithm (code available in the Supplementary Material) (Reed *et al.*, 2006). We chose candidate reactions for gap filling from the complete Model SEED reaction database (Overbeek *et al.*, 2005). We gap filled Strain 1 before and after the application of SONEC (Fig. 5A), and evaluated the essentiality of all reactions (excluding the reactions added during the gap filling process). There were 14 reactions which were essential before SONEC, but not after (Fig. 5B). An example of these is a nucleosidase classified under methionine metabolism. There were 18 reactions which became essential after SONEC but were not essential beforehand. An example of these is an oxidoreductase found in glutamate and arginine metabolic pathways. There were 14 reactions which were constitutively essential. Interestingly, no reaction added by SONEC was essential.

4 Discussion

Here we present the SONEC approach for the assignment of metabolic reactions (and as an extension, metagenomic sequence fragments annotated with metabolic genes) back to a parent metabolic network. This work is motivated by the fact that current approaches are still unable to group complete metagenomic samples into member genomes, leaving, in a recent study, 32% of metagenomic sequence fragments unaccounted for (Nielsen *et al.*, 2014).

We propose that information about the metabolic network can be used to improve metagenomic fragment binning. It is commonly assumed that metabolic networks are gapless, and gap filling of metabolic network reconstructions is used regularly as a source of new biological knowledge (Bartell *et al.*, 2013; Krumholz and Libourel, 2015; Reed *et al.*, 2006; Satish Kumar *et al.*, 2007). Here, we demonstrate that gap filling can similarly be used to assign reactions to the correct parent metabolic network by using a metabolite connectivity score, and thus improve metagenome sequence annotation (Fig. 2).

We observe that more complete networks (reconstructed from bins of metagenomic sequence fragments) initially lead to improved reaction assignment accuracy (Fig. 3A). As parent networks degrade and lose more and more reaction content, accuracy is lost. This observation aligns with intuition, as more complete networks provide context in which to place new reactions. Similarly, as the number of parent networks increases, accuracy is lost (Fig. 3B). This observation also makes sense, recognizing that the presence of more networks increases the opportunity to mis-assign a reaction. Encouragingly, increasing group (a set of metabolic reactions known to originate from the same organism) size significantly improves reaction assignment accuracy under all conditions (Fig. 3C). Group size can be increased by improved assembly or fragment clustering—anything that will increase the number of genes that can be confidently associated with each other. While any given reaction may fill gaps in several possible networks, the likelihood is low of an entire group of reactions filling gaps in the same, incorrect, network. In other words, for large groups of reactions, the error is diluted over the many possible wrong choices, while the metabolite connectivity score accrues for the correct parent network. The presence of shadow networks takes a toll on accuracy (Fig. 3D). Shadow networks can be thought of as the set of organisms in the community that contributed metagenomic sequence, but were not assigned bins. Therefore, any attempt to assign those reactions to existing bins will be incorrect. Finally, simulations showed that functional network predictions (Fig. 3E) are generally improved by SONEC, an outcome that has significant implications for application of subsequent metabolic network analysis (Fig. 5). Interestingly, none of the reactions added by SONEC were essential. However, by adding them, the network structure changed in such a way as to make some previously essential reactions non-essential, and vice versa. One possible explanation for this improvement is that SONEC assigns reactions in a relatively unbiased way (based on metabolite connectivity) compared to traditional gap-filling, which adds reactions to allow flux through a biomass function. Future applications which require functional predictions of the impact of genome engineering or drug targeting within microbial communities can benefit from SONEC. In the end, the goal of SONEC is to improve the reconstruction of individual genomes from metagenomic data. More complete genomes will improve any downstream analyses.

Future work may improve assignment accuracy by modifying the metabolite connectivity score. The example in Supplemental Figure S1 highlights a weakness of the metabolite connectivity score, wherein two models may contain a single dead-end metabolite that overlaps with a reaction, but depending on whether it is a substrate or product, the final gap score may be different. Maintaining the ratios in the metabolite connectivity score is prudent from a parsimony standpoint, because they ensure that the smallest reaction (with the fewest participating metabolites) that can fill a gap will be used. However, future work could explore alternative metabolite connectivity scores that address the weaknesses with the scoring framework presented here. One possibility would be to penalize the addition of new metabolites that do not exist in the network, which would have improved the outcome for the false positive example in Supplemental Figure S1. This may prohibit filling larger gaps consisting of more than one reaction, or filling gaps in less complete networks. Another approach is to apply a global optimization-based gap fill algorithm based on existing methods (Reed *et al.*, 2006). We chose not to pursue this approach because it would be sensitive to the choice of optimization function and exchange constraints, which are difficult to determine for uncharacterized microbes in complex environments.

Enrichment analysis highlights the families of reactions that tend to provide better assignment accuracy (Supplemental Fig. S3). The underlying driver may be that reactions that contain uncommon metabolites are more likely to be assigned to the correct parent network. Within the selection of 100 prokaryotic reconstructions used here, porphyrin and chlorophyll metabolism are uncommon. Given this hypothesis, future work may improve assignment accuracy by selectively weighting reactions that are unique within the environment being studied. For example, in the anterior nares dataset explored here, the rarest pathways include lipoic acid metabolism, inositol metabolism and caprolactam degradation. To improve group assignment accuracy, metabolite connectivity scores corresponding to reactions from these rare subsystems would be weighted more heavily (as they would be expected to increase accuracy disproportionately).

To demonstrate how the SONEC approach can be applied to real metagenomic data, we analyzed 94 metagenomic samples sourced from the human anterior nares (Fig. 4). It is important to note that these samples were not sequenced very deeply, and as a result, the N50 we could achieve after assembly was quite low (250 bp). As a comparison, a recent study assembled DNA fragments from stool samples to achieve an N50 of more than 40 000 bp (Jeraldo *et al.*, 2015). This observation simply indicates that in applications with deeper sequencing, contigs will tend to be much longer. Knowing that larger group size—which is a function of longer contigs—improves SONEC performance, it is likely that SONEC performance will improve with deeper sequencing and more complete assembly. While it is clear that the performance of SONEC is highly dependent on the existing tools used to create the initial bins,

the simulations we performed demonstrate that SONEC can add value and improve predictions even with imperfect data.

We first applied established approaches to create initial clusters of metagenomic sequence fragments, including short read assembly and clustering by cross-sample abundance and nucleotide composition patterns. A BLAST-based estimate of cluster taxonomic consistency (that is, the percentage of fragments within the cluster that map to the same taxonomy) revealed that of the large clusters, only two clusters were >90% consistent. This consistency can be compared to a larger-scale study which analyzed 396 microbiome samples from the human gut, in which 115 large clusters were found to be >95% consistent (Nielsen *et al.*, 2014). Clearly, it is possible to improve the initial clustering and conditions before applying SONEC. Given the two large clusters which mapped consistently to strains of *S.aureus*, we demonstrated how SONEC can be used to assign smaller, orphan clusters to these larger clusters. This practical demonstration on real data shows that by including metabolic information, ambiguous fragments can be assigned to the parent genomes. As a quality check, the resulting metabolic networks after applying SONEC are more consistent with a reference *S.aureus* metabolic network reconstruction.

Acknowledgements

The authors would like to thank Phillip Yen for his help utilizing the computational resources at UVA.

Funding

This work was supported by the National Institutes of Health [grant number R01 GM108501 to J.A.P.], a Jefferson Trust Big Data Fellowship [to M.B.B], and a National Institutes of Health Training Grant [project number 2T32GM008715-16] through the University of Virginia [to M.B.B.].

Conflict of Interest: none declared.

References

Abubucker,S. *et al.* (2012) Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput. Biol.*, **8**, e1002358.

Afshinnekoo,E. *et al.* (2015) Geospatial resolution of human and bacterial diversity with city-scale metagenomics. *Cell Syst.*, **1**, 72–87.

Agresti,A. and Coull,B.A. (1998) Approximate is better than 'Exact' for interval estimation of binomial proportions. *Am. Stat.*, **52**, 119–126.

Albertsen,M. *et al.* (2013) Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. Biotechnol.*, **31**, 533–538.

Alneberg,J. *et al.* (2014) Binning metagenomic contigs by coverage and composition. *Nat. Methods*, **11**, 1144–1146.

Baba,T. *et al.* (2008) Genome sequence of *Staphylococcus aureus* strain newman and comparative analysis of staphylococcal genomes: polymorphism and evolution of two major pathogenicity islands. *J. Bacteriol.*, **190**, 300–310.

Bartell,J.A. *et al.* (2014) Comparative metabolic systems analysis of pathogenic Burkholderia. *J. Bacteriol.*, **196**, 210–226.

Becker,S.A. and Palsson,B.Ø. (2005) Genome-scale reconstruction of the metabolic network in *Staphylococcus aureus* N315: an initial draft to the two-dimensional annotation. *BMC Microbiol.*, **5**, 8–19.

Carr,R. *et al.* (2013) Reconstructing the genomic content of microbiome taxa through shotgun metagenomic deconvolution. *PLoS Comput. Biol.*, **9**, e1003292.

Greenblum,S. *et al.* (2012) Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. *Proc. Natl. Acad. Sci.*, **109**, 594–599.

Gurobi Optimization. (2015) Gurobi Optimizer Reference Manual.

Handelsman,J. (2004) Metagenomics: application of genomics to uncultured microorganisms. *Microbiol. Mol. Biol. Rev.*, **68**, 669–685.

Huttenhower,C. *et al.* (2012) Structure, function and diversity of the healthy human microbiome. *Nature*, **486**, 207–214.

Iverson,V. *et al.* (2012) Untangling genomes from metagenomes: revealing an uncultured class of marine euryarchaeota. *Science*, **335**, 587–590.

Jensen,P.A. and Papin,J.A. (2014) MetDraw: Automated visualization of genome-scale metabolic network reconstructions and high-throughput data. *Bioinformatics*, **30**, 1327–1328.

Jeraldo,P. *et al.* (2015) Draft Genome sequences of 24 microbial strains assembled from direct sequencing from 4 stool samples. *Genome Announc.*, **3**, e00526–e00515.

Kanehisa,M. *et al.* (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res., **38**, D355–D360.

Karp,P.D. *et al.* (2010) Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. *Brief. Bioinf.*, **11**, 40–79.

Kinross,J.M. *et al.* (2011) Gut microbiome-host interactions in health and disease. *Genome Med.*, **3**, 14–25.

Krumholz,E.W. and Libourel,I.G.L. (2015) Sequence-based network completion reveals the integrality of missing reactions in metabolic networks. *J. Biol. Chem.*, **290**, 19197–19207.

MacDonald,N.J. *et al.* (2012) Rapid identification of high-confidence taxonomic assignments for metagenomic data. *Nucleic Acids Res.*, **40**, e111.

MATLAB and Statistics Toolbox (2012).

Namiki,T. *et al.* (2012) MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Res.*, **40**, e155.

Nielsen,H.B. *et al.* (2014) Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat. Biotechnol.*, **32**, 822–828.

Overbeek,R. *et al.* (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.*, **33**, 5691–5702.

Owen,J.G. *et al.* (2015) Multiplexed metagenome mining using short DNA sequence tags facilitates targeted discovery of epoxyketone proteasome inhibitors. *Proc. Natl. Acad. Sci.*, **112**, 4221–4226.

Pitkänen,E. *et al.* (2014) Comparative genome-scale reconstruction of gapless metabolic networks for present and ancestral species. *PLoS Comput. Biol.*, **10**, e1003465.

Qin,J., *et al.* (2010) A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, **464**, 59–65.

R: A language and environment for statistical computing (2013).

Reed,J.L. *et al.* (2006) Systems approach to refining genome annotation. *Proc. Natl. Acad. Sci. U. S. A.*, **103**, 17480–17484.

Rousk,J. and Bengtson,P. (2014) Microbial regulation of global biogeochemical cycles. *Front. Microbiol.*, **5**, 305–307.

Satish Kumar,V. *et al.* (2007) Optimization based automated curation of metabolic reconstructions. *BMC Bioinformatics*, **8**, 212–227.

Schellenberger,J. *et al.* (2011) Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nat. Protoc.*, **6**, 1290–1307.

Sharon,I. *et al.* (2013) Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome Res.*, **23**, 111–120.

Smid,E.J. *et al.* (2014) Functional implications of the microbial community structure of undefined mesophilic starter cultures. *Microb. Cell Fact.*, **13**, S2.

Teeling,H. *et al.* (2004) Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environ. Microbiol.*, **6**, 938–947.

Thiele,I. and Palsson,B.Ø. (2010) A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat. Protoc.*, **5**, 93–121.

Yarza,P. *et al.* (2014) Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat. Rev. Microbiol.*, **12**, 635–645.