IDENTIFICATION AND ANALYSIS OF BACTERIAL GENOMIC METABOLIC SIGNATURES

NATHANIEL BOWERMAN
Department of Biology, Hope College, 35 E 12th St,
Holland, MI 49423 USA
nathaniel.bowerman@hope.edu

NATHAN TINTLE
Department of Mathematics and Statistics, Dordt College, 498 4th Ave NE
Sioux Center, IA 51250, USA
nathan.tintle@dordt.edu

MATTHEW DEJONGH
Department of Computer Science, Hope College, 27 Graves Place,
Holland, MI 49423 USA
dejongh@hope.edu

AARON A. BEST*
Department of Biology, Hope College, 35 E 12th St,
Holland, MI 49423 USA
best@hope.edu

With continued rapid growth in the number and quality of fully sequenced and accurately annotated bacterial genomes, we have unprecedented opportunities to understand metabolic diversity. We selected 101 diverse and representative completely sequenced bacteria and implemented a manual curation effort to identify 846 unique metabolic variants present in these bacteria. The presence or absence of these variants act as a metabolic signature for each of the bacteria, which can then be used to understand similarities and differences between and across bacterial groups. We propose a novel and robust method of summarizing metabolic diversity using metabolic signatures and use this method to generate a metabolic tree, clustering metabolically similar organisms. Resulting analysis of the metabolic tree confirms strong associations with well-established biological results along with direct insight into particular metabolic variants which are most predictive of metabolic diversity. The positive results of this manual curation effort and novel method development suggest that future work is needed to further expand the set of bacteria to which this approach is applied and use the resulting tree to test broad questions about metabolic diversity and complexity across the bacterial tree of life.

* To whom correspondence should be addressed.

1. Introduction

The metabolism of an organism relies on thousands of biochemical reactions, which comprise a network that allows the cell to grow, reproduce, and respond to changing environmental conditions. The set of metabolic reactions are defined by the genes the organism carries and dictate the metabolic properties of the organism. Developing an understanding of the metabolic reactions possible by an organism begins to coalesce into a coherent picture of the metabolic capability of the cell. With thousands of annotated genome sequences of microbial organisms available, it is now possible to analyze not only the metabolic properties of individual organisms, but also the patterns that are seen in metabolic networks across organisms. This includes analyses of the evolution of specific metabolic pathways [*e.g.*, 1,2], analyses based on network topology and properties [*e.g.*, 3–6], analyses of simulated metabolic networks [*e.g.*, 7,8], and combinations of flux balance analysis based modeling of metabolic networks within the context of phylogenies [9–11]. Such analyses can lead to a deeper understanding of the metabolic landscape represented by microbial diversity. Further, sequence-based taxonomic surveys and metagenomic analyses of diverse environments are beginning to allow the systematic exploration of relationships between microbial diversity, functional diversity and environment [12–16].

Accurate annotation of sequenced genomes is foundational to downstream analyses of genomes and metagenome communities. We have reviewed [17] the rapid and accurate subsystem approach to genome annotation implemented in the SEED [18] and RAST [19] frameworks. Achieving highly accurate automated annotations of genomes in RAST is predicated upon a core set of manually curated subsystems in which an expert has catalogued the functional elements of a biological process (*e.g.*, a metabolic pathway) and assigned genes to those functional elements for a large set of sequenced microbes. This ensures high quality annotation of each subsystem and the propagation of knowledge captured in the subsystem to all existing and newly sequenced genomes. One outcome of the subsystems approach is the declaration and discovery of metabolic variants, which are defined as different forms or combinations of forms of a functioning metabolic process [17,20,21]. By identifying patterns of genes comprising a variant, one can quickly assign an organism to a particular variant based on the pattern of genes found during the annotation process. Thus, an organism is assigned a variant code for each subsystem, which yields an abstraction of the metabolic capabilities and the forms of those metabolic functions. Further, a catalogue of functional variants that exist for a particular subsystem captures the diversity with which that biological process is performed among sequenced microbes. Such a catalogue represents a rich data set through which we can gain insight into the complexity and diversity of microbial metabolism.

To enable these types of inquiries and to provide consistent descriptions of metabolic variants among sequenced microbes, we selected a representative set of 101 microbial genomes that were used to manually define and annotate metabolic variants in 139 distinct subsystems covering much of known metabolism. We used this resource to (i.) generate a metabolic signature for each of the 101 organisms comprised of assigned variants for each of the 139 subsystems and (ii.) conduct comparative analyses of metabolic signatures of this diverse set of microbes. These variants and their definitions yield a set of high-confidence metabolic subsystems that have been used to aid the automated generation of genome-scale metabolic reconstructions [22], provide a framework

for automated recognition and propagation of variants to newly sequenced genomes, and allow for comparative studies of metabolic variation observed in sequenced microbes.

2. Results

2.1 *Defining Metabolic Variants for Sequenced Bacteria*

A metabolic variant can be described as a particular version of a metabolic process performed by an organism [21]. We will use the synthesis of isoprenoids (terpenoids) to illustrate the concept of metabolic variants and how particular variants are assigned to an organism. Isoprenoids (*e.g.,* chlorophyll and cholesterol) are found in all organisms and are essential to survival. Key isoprenoid precursors, isopentenyl diphosphate (IPP) and dimethylallyl diphosphate (DMAPP) are produced via two known biosynthetic pathways, the so-called mevalonate and non-mevalonate (DOXP) pathways [1]. The reactions in each pathway are catalyzed by non-homologous proteins, and represent two distinct routes to IPP and DMAPP for organisms. In considering the simplest case of defining metabolic variants for this metabolic process, each of these routes represent separate variants – alternative ways to accomplish the same function of producing precursors to isoprenoids. A third variant exists in the case of an organism containing the necessary genes for both of these pathways. A fourth variant indicates absence of this function through known metabolic pathways in an organism. For each variant (defined in this case as A, B, C and -1, respectively), the possible patterns of metabolic steps involved in each variant is generated, and a brief verbal description of the variant is given. Assignment of any one organism to a known variant of the pathway is accomplished by identifying genes in the organism's genome that encode functions corresponding to the area of metabolism and matching the pattern of metabolic steps the organism is predicted to be capable of to one of the defined variants (see Supplemental Figure 1 for additional details).

We have implemented the approach of identifying variants, defining variants, and assigning variants to organisms in the framework of SEED subsystems [18]. This represents a significant, multi-year manual curation effort on the part of SEED annotators through the capture of known metabolic diversity described in the literature and the analysis of patterns seen in sequenced microbial genomes. We chose a set of 101 bacterial genomes, representing 14 bacterial divisions, and 139 subsystems in the SEED that maximized our coverage of metabolism represented in major metabolic databases (*e.g.*, KEGG) and that facilitated the automated generation of metabolic models for bacteria [22]. We characterized a total of 846 metabolic variants in these subsystems that our set of organisms are capable of based on known information of each subsystem and the annotated function of genes in each genome. The outcome of this curation effort is a metabolic variant catalogue comprising descriptions of naturally occurring variations of central and intermediate metabolism for a phylogenetically diverse group of bacteria. Supplemental Figure 2 and Supplemental Files 1-4 give detailed information on the organisms and variants selected and defined.

2.2 *Analyses of Bacterial Metabolic Signatures*

In order to gain a more thorough understanding of metabolic diversity and how metabolic functions are distributed throughout Bacteria, we devised a measure of the metabolic distance, D_{FM}, between two organisms based on the curated metabolic variant catalogue. For a given organism *i*, it is possible to summarize the metabolic capabilities as a binary vector, v_i, of 846 0's and 1's, representing the absence and presence, respectively, of each of the 846 metabolic variants. In effect, *v* is a metabolic signature (or barcode) describing the metabolic capabilities of an organism. D_{FM} measures the metabolic distance between two given organisms *i* and *j* by comparing the similarity of v_i and v_j to the likelihood of observing the similarity between the two vectors by chance. We utilized complete linkage hierarchical clustering of all pairwise D_{FM} of organisms in our dataset to produce a dendrogram summarizing the relationships of the organisms based on metabolic distances (Figure 1). We used a false discovery rate (FDR) of 1×10^{-15} to identify 5 distinct clusters of organisms (Clusters A through E in Figure 1). Each cluster represents a group of organisms with highly similar metabolic signatures. To assess the face validity of the resulting metabolic signature tree, we sought to confirm that the ordering seen in the tree met reasonable biological expectations. For instance, one would expect that closely related organism pairs are likely to be closely paired on the dendrogram – *E. coli* and *Salmonella* are nearest neighbors in the tree as are two representatives of the genus *Shewanella*. Furthermore, the four oxygenic photosynthetic organisms in the set form a tight cluster (FDR $<1 \times 10^{-60}$, Supplemental Figure 3a, organism names colored green). These observations, and many others not detailed here (for example, Supplemental Figures 3b and 3c), indicate that the metabolic distance metric reveals biologically meaningful patterns and gave us confidence that we could use the tree to address additional biological questions of interest.

2.3 *Contribution of Organism Characteristics to Bacterial Metabolic Signatures*

To provide a quantitative estimate of the ability of organism characteristics to explain the clustering observed in the metabolic signature tree, we produced a data set capturing 19 characteristics for each of the 101 organisms, covering attributes such as phylogenetic grouping, environment classification, and oxygen utilization (Supplemental File 1). We performed a multiple regression analysis, using the 19 phenotypic characteristics to predict metabolic distance. The variables in our data set were able to explain 50% of the variance of metabolic distance ($r^2 = 0.50$). The top four characteristics contributing to the clustering are genome size, metabolic mode, host association, and ability to survive in an intracellular environment, uniquely explaining 19.7%, 9.6%, 7.5% and 7.2% of the overall variation in metabolic distance, respectively. All other characteristics contribute to ~5% or less of the overall r^2. Phylogenetic distance ranked 11th of the 19 characteristics, indicating that only a small fraction of the metabolic distance variance could be attributed to phylogeny. A phylogenetic tree of the organisms in this study annotated with metabolic signature cluster membership shows the clear mixing of related organisms throughout the 5 clusters (Supplemental Figure 2). A follow-up analysis which removed 14 organisms with small genomes (Cluster B), showed that there is a slight decrease in the ability to explain the overall variation in metabolic distance ($r^2 = 0.48$) with the 19 phenotypes combined, and less predictive

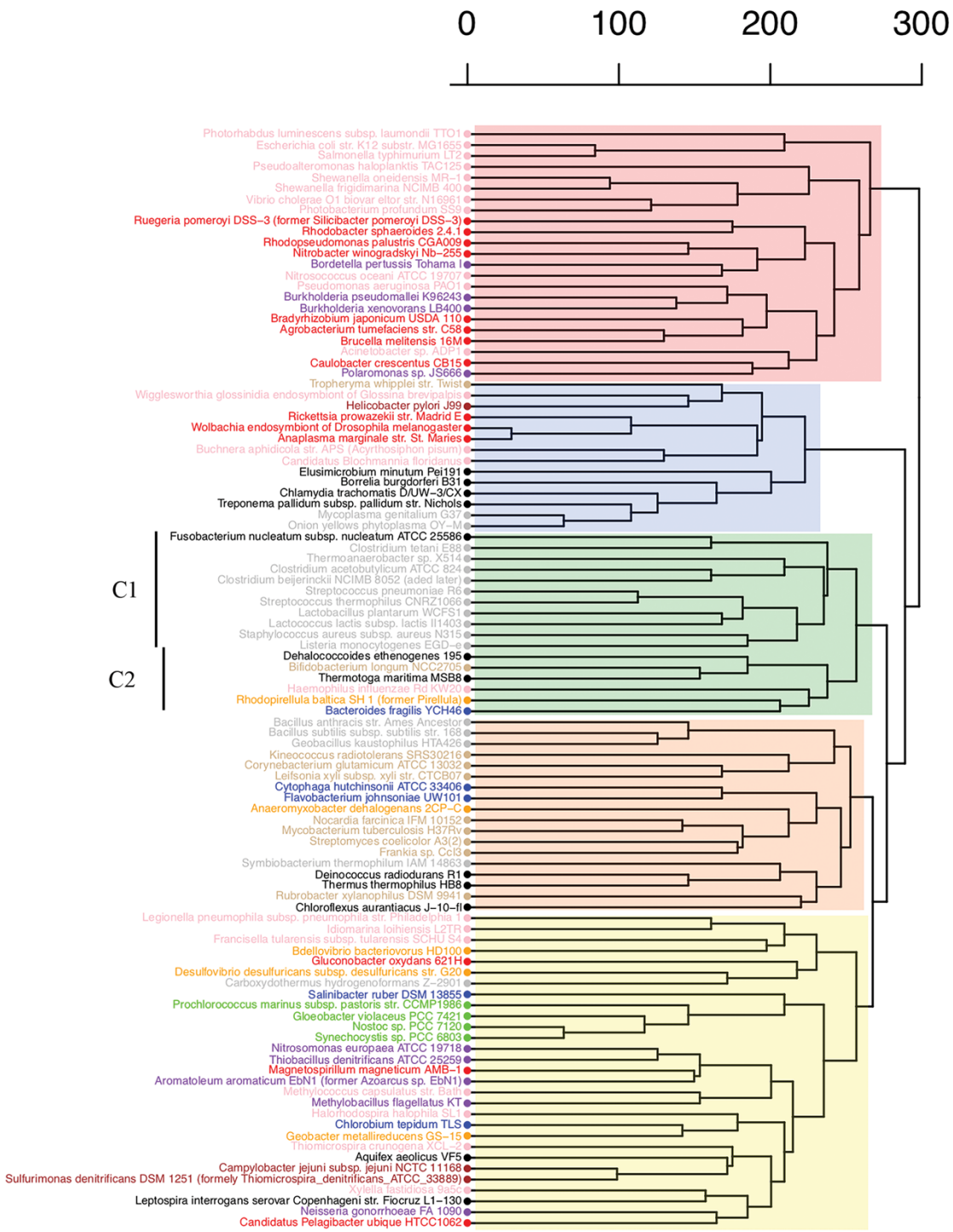

Figure 1. Metabolic Signature Tree from complete linkage hierarchical clustering of D_{FM} of organisms. Five clusters corresponding to an FDR of 1 x 10^{-15} are highlighted by shading – Clusters A-E; pink, blue, green, orange and yellow, respectively. Subclusters C1 and C2 are indicated by black bars. Organism names are colored according to phylogenetic classification: Actionomycetes, Tan; Firmicutes, Gray; Cyanobacteria, Green; Bacteroides/Chloribi, Blue; Other, Black; Proteobacteria: Alpha, Red; Beta, Purple; Delta, Orange; Epsilon, Brown; Gamma, Pink.

ability of genome size (from 19.7% to 6.7%). Full results are provided in Supplemental Figures 4a and 4b.

2.4 *Specific Phenotypes Associated with Individual Clusters*

In addition to characterizing the influence of phenotype on the global topology of the metabolic tree, it is possible to associate specific phenotypic characteristics with individual clusters of organisms in the metabolic tree. We assessed the distribution of each phenotypic characteristic within a cluster and compared this to the distribution of that phenotypic character in the other clusters to yield a statistical measure of the differential distribution of any one phenotypic trait among clusters (see Methods). Each of the clusters is characterized by a particular set of phenotypes as summarized in Table 1 that are over or underrepresented at a conservative measure of statistical confidence ($p < 0.0006$). As expected, the phenotypic characters with the lowest p-values for each cluster correspond to initial observations seen with the overlay of phenotypic characters on the metabolic tree, while providing more specificity to the observations and highlighting characters that may not be otherwise apparent. Cluster A consists completely of Gram negative organisms that also tend to have large genomes (5.2 Mb vs 3.1 Mb average for entire dataset). All organisms are phylogenetically related, being members of the α, β, and γ Proteobacteria. However, these taxonomic groups are not identified as statistically significant due to the broad distribution of other members of these taxonomic groups throughout the clusters (*i.e*, B and E). This result is consistent with the diverse habitats and lifestyles associated with Proteobacteria. Cluster B contains organisms that tend to have small genome sizes, are classified as intracellular and obligate host associated, and have a low GC%. Obligate intracellular parasites tend to have smaller genomes as they require fewer genes due to obtaining resources from the host cell and smaller genomes tend to have lower GC content to facilitate evolution through an increased mutation rate. Cluster C consists of organisms that tend to be in the phylum Firmicutes, families Bacillales or Lactobacillales, are Gram positive, and are anaerobic. Cluster D contains an over-representation of Actinomycetes, Gram positive bacteria, and sporulating bacteria. Cluster E contains many phylogenetically unrelated organisms, a majority of organisms that have preferred metabolic modes other than chemoheterotrophy, and also contains a disproportionate number of Gram negative bacteria.

2.5 *Metabolic Variants Associated with Specific Clusters*

As a complementary approach to exploring organism characteristics associated with specific clusters, it is also possible to explore whether particular metabolic variants are over- or under-represented in the specific metabolic clusters. As an example, we observed that Cluster C could be divided into two subgroups, C1 and C2. The organisms in Cluster C1 are low-GC Gram positive organisms in the phylum Firmicutes with the exception of *Fusobacterium*; the subcluster can be further divided by the oxygen requirement characteristic – the organisms in class Clostridia and *Fusobacterium* are all obligate anaerobes, whereas the organisms in class Bacilli are facultative (Figure 1, Supplemental File 1). We hypothesized that there should be specific metabolic variants (likely related to respiratory systems) that would distinguish these two groups. To investigate this and similar hypotheses, we used an approach that compared the frequencies of metabolic variants

in two groups of organisms (*e.g.,* subgroups of Cluster C1) to highlight those variants that were the most different between the groups (see Methods for details). In the case of Cluster C1, there

Table 1. Over- and under-representation of characteristics by cluster

Cluster	Characteristic*	Present Inside Cluster	Present Outside Cluster	p-value
A	Genome Size	Mean = 5.2	Mean = 3.1	5.88×10^{-6}
	Gram Stain Negative	23/23 (100%)	43/78 (55%)	1.22×10^{-5}
	Gram Stain Positive	0/23 (0%)	26/78 (33%)	6.79×10^{-4}
B	Genome Size	Mean = 1.1	Mean = 4	3.71×10^{-24}
	Intracellular Survival - Obligate Intracellular	8/14 (57%)	0/87 (0%)	1.49×10^{-8}
	Free Living/Host Associated - Obligate Host Association	12/14 (86%)	10/87 (11%)	3.96×10^{-8}
	Host Type - Arthropod/Insect	8/14 (57%)	2/87 (2%)	5.94×10^{-7}
	GC Content	35.5	51.6	1.47×10^{-5}
	Free Living/Host Associated - Free Living	0/14 (0%)	50/87 (57%)	4.35×10^{-5}
	Intracellular Survival - Not Applicable	0/14 (0%)	50/87 (57%)	4.35×10^{-5}
	Habitat Outside Host - Soil	0/14 (0%)	38/87 (44%)	8.39×10^{-4}
C	Taxonomic Class - Mixed Firmicutes	10/17 (59%)	7/84 (8%)	1.17×10^{-5}
	Gram Stain - Positive	12/17 (71%)	14/84 (17%)	2.25×10^{-5}
	Oxygen Requirement - Aerobe	1/17 (6%)	47/84 (56%)	1.09×10^{-4}
	GC Content	Mean = 39.5	Mean = 51.4	1.24×10^{-4}
	Oxygen Requirement - Anaerobe	9/17 (53%)	8/84 (1%)	1.44×10^{-4}
	Gram Stain - Negative	4/17 (24%)	62/84 (74%)	1.47×10^{-4}
	Bacillales, Lactobacillales	6/17 (35%)	3/84 (4%)	5.98×10^{-4}
D	Taxonomic Class - Actinomycetes	8/18 (44%)	2/83 (2%)	7.96×10^{-6}
	Gram Stain - Positive	12/18 (67%)	14/83 (17%)	5.73×10^{-5}
	Sporulation - Sporulating	8/18 (44%)	5/83 (6%)	1.67×10^{-4}
	Gram Stain - Negative	5/18 (28%)	61/83 (73%)	5.86×10^{-4}
	Sporulation - Nonsporulating	10/18 (56%)	76/83 (92%)	6.77×10^{-4}
E	Prefered Metabolic Mode - Chemoorganoheterotroph	14/29 (48%)	69/72 (96%)	1.29×10^{-7}
	Prefered Metabolic Mode - Photolithoautotroph	6/29 (21%)	0/72 (0%)	3.75×10^{-4}
	Gram Stain - Positive	1/29 (3%)	25/72 (35%)	7.92×10^{-4}

*Genome Size given as average number of megabases in group; GC Content given as average percentage in group

are 12 metabolic variants that are unequally distributed between anaerobic and facultative organisms (p-value ≤ 0.05) within the cluster (Supplemental File 5). These 12 variants represent 7 unique subsystems associated with the synthesis of cofactors, vitamins, and isoprenoids. Three of these subsystems are associated with respiratory functions (heme and siroheme biosynthesis, sulfur related anaerobic respiratory reductases, and sodium translocating oxidoreductases). There is differential distribution between the anaerobic (4 of 5) and facultative (0 of 6) cluster members for the presence of sulfur reductases. Likewise, 4 of 5 anaerobic cluster members have an operon of *rnf* like genes encoding putative electron transport complexes associated with nitrogen fixation,

whereas none of the facultative cluster members have this operon, which is consistent with the classical differentiation of *Clostridia* from *Bacilli* organisms in the low-GC Gram positive group.

3. Discussion

We have described a novel approach to examining the metabolic relationships among bacterial genomes that focuses on the collection of metabolic variants associated with an organism. The vector of metabolic variants succinctly describes the organism's metabolic capabilities and allows for statistical comparison of vectors between organisms that is scalable to thousands of genomes. In the current study, we have provided a proof of concept with a phylogenetically diverse set of 101 bacterial genomes, comprising 846 variants and covering much of known metabolism. The variant definitions are the result of a targeted manual curation effort in the framework of the SEED database [18], which breaks down bacterial metabolism into subsystems (defined as collections of functional roles necessary to perform a cellular function). In this study, 139 subsystems were individually examined to define the possible metabolic variants. The outcome of the manual curation effort is a set of curated metabolic variants that can be rapidly assigned to bacterial genomes and used to compare the metabolic capabilities present in the genomes.

Many of the approaches to understanding the breadth, conservation and evolution of metabolic networks found in the bacterial domain have focused on properties of network architecture such as scale, network path length, network motifs, centrality, modularity and connectedness [3,4,23]. Common themes are observed in that metabolic networks have been shown to be scale-free and highly modular for most organisms. It has been shown that the complexity of a metabolic network can be associated with particular lifestyles/habitats. For example, obligate symbionts that experience relatively stable environments have less complex networks than organisms that are free-living and exposed to many environments. These approaches are highly granular in that they connect networks on the level of individual reactions, compounds and enzymes. An extension to network based approaches was introduced by Mazurie *et al.* [5] that compares higher level functional units called networks of interacting pathways. These were used to classify organisms into phenotypic categories. They observed similar trends with respect to the nature of the networks as seen with other network-based approaches and were able to assign functional pathways to organisms of particular phenotypes. For instance, free-living and host-associated organisms differed with respect to frequency of observed carbohydrate and energy metabolism pathways; motile and non-motile organisms differed with respect to xenobiotic degradation pathways. More recently, Pearcy *et al.* [6] introduced a method that produced vectors for an organism whose elements described individual network motifs. They analyzed 3 and 4 node motifs that are abstractions of specific compound and reaction connections and identified network motifs that were enriched for organisms with different habitats/lifestyles, such as aerobic/facultative vs. anaerobic. By looking at the reactions and compounds that made up the enriched motifs, it was possible to identify specific metabolites associated with the different lifestyles. Patterns such as these supported the assertion that environmental conditions shape the properties of metabolic networks that occur in organisms. In a departure from analyzing network properties, Poot-Hernandez *et al.* [24] calculated linear enzymatic step sequences (ESS) found in metabolic maps in KEGG and defined core and peripheral metabolic pathways for 40 gamma proteobacteria

species. An analysis of the relationships of ESS vectors among organisms was not conducted. Mithani *et al.* [4] analyzed the presence/absence of enzymatic reactions in pathways of *Pseudomonas* species based on KEGG map reaction mining. They found interesting patterns of gains and losses associated with niche specific adaptations to host association. Their approach is limited by the restriction to KEGG maps and boundary effects (reactions that appear in more than one map do not get connected). Further, the authors noted that other information such as genome context could improve understanding of evolutionary processes. The approach that we describe here is fundamentally different than those employed to date in that the unit being analyzed – variants associated with an organism – is non-network based; implicitly incorporates genome context, paralogs, isoenzymes and non-orthologous replacements through manual curation; and allows for coverage of metabolic capabilities across the modular nature of networks and their representation as disconnected metabolic maps. Further, each variant represents a functioning biological process, allowing the succinct assertion of organism capabilities (both positive and negative attribution). The analysis of variant vectors and the patterns observed therein give rise to clusters of metabolic forms comprised of the organisms and their individual variants. It is then possible to attribute the influence of phenotypic characters and phylogenetic relationships to these clusters through standard statistical approaches. It would be instructive to map individual variants to data types analyzed previously (e.g., networks of interacting pathways, individual network motifs, and ESS) to enable systematic comparison of each of these approaches to the variant approach.

We identified five main clusters of metabolically related organisms in our analysis (A-E in Fig. 1), each of which share some phenotypic traits (Table 1). We also described a complementary approach to evaluate which variants are most differentially distributed between clusters on the tree. These analyses yield patterns that are consistent with the approaches mentioned above. For instance, Cluster B is comprised of organisms that are host-associated and found in relatively stable environments; the 144 variants that are significantly differentially distributed ($p < 0.05$) include the absence of functions in amino acid, purine and pyrimidine, and vitamin/co-factor biosynthesis pathways (Supplemental File 5). There are other cases where there are hints at what drives the members of a cluster together in metabolic space (*e.g.*, *Neisseria*, *Pelagibacter*, *Xylella*, *Leptospira* – amino acid usage; *Gluconobacter*, *Desulfovibrio*, *Carboxydothermus* – extreme environments), but the current sampling of 101 organisms limits the statistical analysis of small clusters such as these.

These types of problems will become tractable with the inclusion of new genomes that begin to fill out metabolic signature clusters. Importantly, the fundamental structure of the dendrogram will not change as new genomes are added given a constant set of defined variants (*e.g.*, Cluster B will continue to contain organisms with small genomes/symbionts, Cluster C will contain most low GC Gram positive organisms, Cluster D will contain high GC organisms, and Cluster E will likely expand and subdivide as representation of metabolically diverse organisms increases). Organisms that do not follow these expectations may yield insight into novel combinations of metabolic capabilities; the current metabolic clusters represent a framework of hypotheses about relationships between suites of metabolic variants associated with any one organism. In contrast, as additional metabolic variants are identified, curated and assigned, the nature of the metabolic clusters may change. In short, as more well-annotated genomes are included, the statistical power

for this type of analysis increases, enhancing our ability to examine the metabolic relationships between organisms and what factors impact these metabolic commonalities.

The proof of concept described in this work serves as a foundation for identifying metabolic signatures for all sequenced bacteria and associating those signatures with specific organism characteristics and metabolic variants. Analyses of correlations between metabolic variants observed across bacterial life will enhance our understanding of the nature of the metabolic space occupied by diverse organisms.

4. Methods

4.1 *Organisms and Features*

The 101 organisms chosen were representatives of 14 phylogenetic divisions of eubacteria (Supp Fig. 2), which provides a reasonable coverage of sequenced microbial diversity with complete genomes. Each of the 101 organisms were classified on 19 different phenotypic features based on information already present in the SEED and via literature review. The features considered here and summary statistics are provided as Supplemental File 1. In order to generate maximum likelihood phylogenetic distances for each pair of organisms, we selected a representative 16S rRNA sequence of each organism from the Silva SSU Reference Set Release 106 using the ARB environment [25]. RaxML 7.0.4 [26] was then used to generate a set of maximum likelihood pairwise distances. Pairwise phylogenetic distances are included as Supplemental File 2.

4.2 *Creating a Metabolic Distance Measure*

We calculated a measure of metabolic distance, D_{FM}, between organisms based on the vector v_i, where i is the i^{th} organism, of 0's and 1's, indicating the presence/absence of the 846 subsystem variants. In general, the metabolic distance between organisms i and j, will be a function that measures the dissimilarity of vectors v_i and v_j. While there are numerous options for measuring dissimilarity or similarity between two vectors (e.g., Euclidean distance, Pearson correlation), we chose to use a novel method based on Fisher's exact test because of its robustness to the widely varying numbers of 0's and 1's observed in vector v, along with its ability to directly integrate a measure of statistical confidence into the distance measure, making D_{FM} an indirect measurement of the likelihood of two organisms possessing the observed degree of overlap in metabolism 'by chance.' To generate D_{FM}, first, for each of the 5050 ($101 * 100/2$) pairs of organisms, a 2x2 cross tabulation table was created and a Fisher's exact test p-value was generated. The Fisher's exact test p-value (that is, the likelihood of observing pattern of metabolic consistency by chance) acts as a measure of metabolic similarity and is available for all pairs of organisms in Supplemental File 6. We transformed p-values using: D_{FM}=300+ln(p) to yield a metric of metabolic distance, D_{FM}, which is always greater than 0 in our dataset.

4.3 *Statistical Analyses*

Four main statistical analyses were performed on D_{FM}. First, hierarchical clustering with complete linkage was conducted on the 101 organisms using D_{FM} as computed between all 5050 pairs of organisms. A dendrogram was created and phenotypic features were overlaid on it to aid in

interpretation of subsequent analyses. Clusters of interest on the dendrogram were determined using a false discovery rate (FDR) based on the Fisher's exact test p-values. Second, a multiple regression analysis was conducted to investigate the extent to which the metabolic distance, D_{FM}, could be explained by the 19 phenotype features. We used a dataset comprising 19 phenotype features and metabolic distance for each of the 5050 pairs of organisms (supplemental file #2). Models regressed metabolic distance on each of the 19 phenotype features. Third, we conducted analyses designed to answer the question "Which phenotypes explain why this cluster (on the dendrogram) exists?" After 'cutting' the dendrogram by looking at all of the mutually exclusive clusters for which all pairs of organisms within the cluster have a certain level of association, we wish to compare two mutually exclusive clusters of organisms to attempt to identify phenotypic differences in the clusters which are likely candidates for why the organisms separated into two mutually exclusive clusters. For categorical phenotypes, a Fisher's exact test is conducted which compares the proportion of organisms in cluster #1 with the phenotypic characteristic to the proportion of organisms in cluster #2 with the characteristic. For quantitative phenotypes, a two-sample t-test is used. Full results for all phenotypes and clusters A, B, C, D, E1 and E2 are provided in Supplemental File 7. Lastly, we conducted the same analysis as just described to answer the question "Which metabolic variants associate with specific clusters?" by using the Fisher's exact test approach on mutually exclusive clusters, evaluating association between metabolic variants and cluster memberships. Unless otherwise indicated, all analyses were conducted using R (www.r-project.org).

Supplemental Files

All supplemental files are available online at the following URL: http://homepages.dordt.edu/ntintle/metsig.zip

5. Acknowledgments

This work is supported by NSF MCB-1330734. We gratefully acknowledge discussions with Andrei Osterman, Ross Overbeek and other members of the Fellowship for the Interpretation of the Genome in early phases of this project.

References

[1] Y. Boucher and W. F. Doolittle, *Mol. Microbiol.* **37**, 703 (2000).
[2] G. Xie, C. A. Bonner, T. Brettin, R. Gottardo, N. O. Keyhani, and R. A. Jensen, *Genome Biol.* **4**, R14 (2003).
[3] A. Kreimer, E. Borenstein, U. Gophna, and E. Ruppin, *Proc. Natl. Acad. Sci.* **105**, 6976 (2008).
[4] A. Mithani, G. M. Preston, and J. Hein, *PLoS Comput. Biol.* **6**, (2010).
[5] A. Mazurie, D. Bonchev, B. Schwikowski, and G. A. Buck, *BMC Syst. Biol.* **4**, 59 (2010).
[6] N. Pearcy, J. J. Crofts, and N. Chuzhanova, *Mol. BioSyst.* **11**, 77 (2015).
[7] A. Barve and A. Wagner, *Nature* **500**, 203 (2013).
[8] J. Raymond, *Science* **311**, 1764 (2006).
[9] R. Braakman and E. Smith, *PLoS Comput. Biol.* **8**, e1002455 (2012).
[10] R. Braakman and E. Smith, *Phys. Biol.* **10**, 11001 (2013).

[11] R. Braakman and E. Smith, *PLoS One* **9**, e87950 (2014).
[12] J. Raes, I. Letunic, T. Yamada, L. J. Jensen, and P. Bork, *Mol. Syst. Biol.* **7**, 473 (2014).
[13] C. von Mering, P. Hugenholtz, J. Raes, S. G. Tringe, T. Doerks, L. J. Jensen, N. Ward, and P. Bork, *Science* **315**, 1126 (2007).
[14] T. A. Gianoulis, J. Raes, P. V. Patel, R. Bjornson, J. O. Korbel, I. Letunic, T. Yamada, A. Paccanaro, L. J. Jensen, M. Snyder, P. Bork, and M. B. Gerstein, *Proc. Natl. Acad. Sci.* **106**, 1374 (2009).
[15] S. Chaffron, H. Rehrauer, J. Pernthaler, and C. von Mering, *Genome Res.* **20**, 947 (2010).
[16] N. Fierer, J. W. Leff, B. J. Adams, U. N. Nielsen, S. T. Bates, C. L. Lauber, S. Owens, J. A. Gilbert, D. H. Wall, and J. G. Caporaso, *Proc. Natl. Acad. Sci.* **109**, 21390 (2012).
[17] C. S. Henry, R. Overbeek, F. Xia, A. A. Best, E. Glass, J. Gilbert, P. Larsen, R. Edwards, T. Disz, F. Meyer, V. Vonstein, M. Dejongh, D. Bartels, N. Desai, M. D'Souza, S. Devoid, K. P. Keegan, R. Olson, A. Wilke, J. Wilkening, and R. L. Stevens, *Biochim. Biophys. Acta* **1810**, 967 (2011).
[18] R. Overbeek, T. Begley, R. M. Butler, J. V Choudhuri, H. Y. Chuang, M. Cohoon, V. de Crecy-Lagard, N. Diaz, T. Disz, R. Edwards, M. Fonstein, E. D. Frank, S. Gerdes, E. M. Glass, A. Goesmann, A. Hanson, D. Iwata-Reuyl, R. Jensen, N. Jamshidi, L. Krause, M. Kubal, N. Larsen, B. Linke, A. C. McHardy, F. Meyer, H. Neuweger, G. Olsen, R. Olson, A. Osterman, V. Portnoy, G. D. Pusch, D. A. Rodionov, C. Ruckert, J. Steiner, R. Stevens, I. Thiele, O. Vassieva, Y. Ye, O. Zagnitko, and V. Vonstein, *Nucleic Acids Res.* **33**, 5691 (2005).
[19] R. K. Aziz, D. Bartels, A. A. Best, M. DeJongh, T. Disz, R. a Edwards, K. Formsma, S. Gerdes, E. M. Glass, M. Kubal, F. Meyer, G. J. Olsen, R. Olson, A. L. Osterman, R. a Overbeek, L. K. McNeil, D. Paarmann, T. Paczian, B. Parrello, G. D. Pusch, C. Reich, R. Stevens, O. Vassieva, V. Vonstein, A. Wilke, and O. Zagnitko, *BMC Genomics* **9**, 75 (2008).
[20] A. Osterman and R. Overbeek, Curr. Opin. Chem. Biol. **7**, 238 (2003).
[21] Y. Ye, A. Osterman, R. Overbeek, and A. Godzik, Bioinformatics **21**, i478 (2005).
[22] C. S. Henry, M. DeJongh, A. A. Best, P. M. Frybarger, B. Linsay, and R. L. Stevens, *Nat. Biotechnol.* **28**, 977 (2010).
[23] A.-L. Barabasi and Z. N. Oltvai, Nat Rev Genet **5**, 101 (2004).
[24] A. C. Poot-Hernandez, K. Rodriguez-Vazquez, and E. Perez-Rueda, *BMC Genomics* **16**, 957 (2015).
[25] W. Ludwig, O. Strunk, R. Westram, L. Richter, H. Meier, Yadhukumar, A. Buchner, T. Lai, S. Steppi, G. Jobb, W. Förster, I. Brettske, S. Gerber, A. W. Ginhart, O. Gross, S. Grumann, S. Hermann, R. Jost, A. König, T. Liss, R. Lüssmann, M. May, B. Nonhoff, B. Reichel, R. Strehlow, A. Stamatakis, N. Stuckmann, A. Vilbig, M. Lenke, T. Ludwig, A. Bode, and K.-H. Schleifer, *Nucleic Acids Res.* **32**, 1363 (2004).
[26] A. Stamatakis, *Bioinformatics* **22**, 2688 (2006).