GutCyc: a Multi-Study Collection of Human Gut Microbiome Metabolic Models

Aria S. Hahn1,2+ Tomer Altman2,3+
Kishori M. Konwar2,5+ Niels W. Hanson1 Dongjae Kim6
David A. Relman7,8,9 David L. Dill10
Steven J. Hallam1,2,11*

May 26, 2016

1. Department of Microbiology and Immunology, University of British Columbia, Vancouver, BC, Canada 2. Koonkie Inc., Menlo Park, CA, USA 3. Biomedical Informatics, Stanford University School of Medicine, Stanford, CA 94305, USA 4. Whole Biome, Inc., 953 Indiana Street, San Francisco, CA 94107 5. Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA 6. Department of Computer Science, University of British Columbia, Vancouver, BC, Canada 7. Department of Microbiology and Immunology, Stanford University School of Medicine, Stanford, CA 94305, USA 8. Department of Medicine, Stanford University School of Medicine, Stanford, CA 94305, USA 9. Veterans Affairs Palo Alto Health Care System, Palo Alto, CA 94304, USA 10. Department of Computer Science, Stanford University, Stanford, CA 94305, USA 11. Ecosystem Services, Commercialization and Entrepreneurship (ECOSCOPE), University of British Columbia, Vancouver, BC, Canada + These authors contributed equally to this work. *Corresponding author: Steven J. Hallam (shallam@mail.ubc.ca).

Abstract

The most complex niche in the human microbiome is found in the distal gut, where communities harbor thousands of strains from across the microbial taxonomy. Despite playing important roles in immune response, nutrition, metabolism of pharmaceuticals, and enteric diseases, little is known of the intricate network of metabolic transformations mediated by the microbiota. Advances in high-throughput sequencing are enabling researchers to explore the metabolic potential of microbial communities inhabiting the human body with unprecedented resolution. Resulting datasets are reshaping how we perceive ourselves and opening new opportunities for prevention and therapeutic intervention. Several large-scale metagenomic datasets derived from hundreds of human microbiome samples and sourced from multiple studies are now publicly available.

However, the different proprietary functional annotation pipelines used to process sequence information from each of these studies, with their own choice of functional and metabolism reference databases, and cut-off parameters for relevant hits, introduce systematic differences that confound comparative analyses. To overcome these challenges, we developed GutCyc, a freely-available compendium of environmental pathway genome
databases (ePGDBs) constructed from metagenome assemblies from 431 human microbiome samples, across three different large-scale studies. The ePGDBs were constructed using the open-source MetaPathways metagenomic annotation pipeline that enables reproducible functional metagenomic annotation. We also generated metabolic network reconstructions for each metagenome using the Pathway Tools software, that empowers researchers and clinicians interested in visualizing and interpreting the metabolic pathways, reactions, compounds, and transporters in the human gut microbiome. For the first time, GutCyc provides consistent annotations and metabolic pathway predictions across these three studies, making possible reproducible comparative community analyses between health and disease states in inflammatory bowel disease, Crohn’s disease, and type 2 diabetes. We demonstrate the utility of GutCyc as a computational model by reconstructing a missing metabolic route from a research study on the role of the microbiome in cardiovascular disease, and statistical enrichment analysis and visualization of high-throughput data from a microbiome metabolomics study.

With GutCyc, the publicly-deposited knowledge about human distal gut microbiotic transport and enzymatic reactions is integrated in a form that is both readily searchable by researchers and easily processed programmatically. GutCyc enables research on drug/target discovery, analysis of pharmaceutical fate in the lumen, and engineering of therapeutic microbiomes. GutCyc data products are searchable online, or may be downloaded and explored locally using the MetaPathways graphical user interface and Pathway Tools.

Background & Summary

The myriad collections of microorganisms found on and in the human gastrointestinal tract, on other mucus membranes, and skin are known as the human microbiome \[60\]. Changes in microbiome structure and function have been implicated in numerous disease states including inflammatory bowel disease, cancer, and even cardiovascular disease \[35, 11\]. Increasingly, researchers are using high-throughput sequencing approaches to study entire metagenomes and characterize diversity and metabolic potential in relation to health and disease states \[69\] opening new opportunities for prevention and therapeutic intervention at the interface of microbial ecology, bioinformatics and medical science. The most densely colonized human habitat is the distal gut, inhabited by thousands of diverse microorganisms, as differentiated at the strain level. Despite providing essential ecosystem services, including nutritional provisioning, detoxification and immunological conditioning, the metabolic network driving matter and energy transformations by the distal gut microbiome remains largely unknown. Several large-scale metagenomic datasets (derived from hundreds of microbiome samples) from the Human Microbiome Project \[55\], Beijing Genomics Institute (BGI) \[58\], and MetaHIT \[57\] are now available on-line, creating an opportunity for large-scale metabolic network comparisons.

While the studies cited above provide the sequencing data, they do not pro-
vide the software environment used for generating their annotations. Over the past few years a number of available metagenomic annotation pipelines have emerged including IMG/M [46], Metagenome Rapid Annotation using Subsystem Technology MG-RAST [68], SMASHCommunity [9] and HUMAnN [6]. Each pipeline used to process sequence information between studies introduces biases based on idiosyncratic formatting, and alternative annotations or algorithmic methods. In particular, support for metabolic pathway annotation varies significantly among pipelines due to differences in reference database selection with resulting impact on metabolic network comparisons. The most common metabolism reference database currently in use is Kyoto Encyclopedia of Genes and Genomes (KEGG) [27]. To the extent that previous metagenomic annotation pipelines provided metabolic reconstructions, they are usually limited to links to KEGG module and pathways maps [27] based on KEGG Orthology (KO) or pathway identifiers, and do not provide explicit representations of compounds, reactions, enzymes, pathways, and their interconnections [20][6]. Although these identifier lists can be visualized with coverage or gene count information using programs like KEGG Atlas [50], they do so using often incompatible formats. Such mapping is limited because there is no simple way to query, manipulate, or visualize the underlying metabolic model directly. Moreover, prediction using KEGG results in hits to large (often including dozens or over a hundred reactions), amalgamated, chimeric pathways [7], which impedes enrichment and association studies by drowning out [6].

In responding to the deficiencies of existing tools, we recently developed a modular annotation and analysis pipeline enabling reproducible research [12] called METAPathways, that guides construction of Environmental Pathway/Genome Database (ePGDB)s from environmental sequence information [38] using Pathway Tools [28] and MetaCyc [33][13][14]. Pathway Tools is a production-quality software environment developed at SRI that supports metabolic inference and flux balance analysis based on the MetaCyc database of metabolic pathways and enzymes representing all domains of life. Unlike KEGG, MetaCyc emphasizes smaller, evolutionarily conserved or co-regulated units of metabolism and contains the largest collection (over 2,400) of experimentally validated metabolic pathways [7]. Navigable and extensively commented pathway descriptions, literature citations, and enzyme properties combined within an ePGDB provide a coherent structure for exploring and interpreting predicted metabolic networks from the human microbiome across multiple levels of biological information (DNA, RNA, protein and metabolites). In addition to browsing and advanced search functions, ePGDB navigation via Pathway Tools enables metabolite tracing through reaction networks, dead-end or choke point identification, enrichment/depletion analysis, and an Application Programming Interface (API) for programmatic development. Over 9,800 Pathway/Genome Database (PGDB)s have been developed by researchers around the world, and thus ePGDBs represent a data format for metabolic reconstructions that exhibit a potential for reusability in further studies.

Here we present GUTCyc a freely-available compendium of over 430 ePGDBs constructed from public shotgun metagenome datasets generated by the Human
Microbiome Project (HMP) [55], a MetaHIT Project (MetaHIT) inflammatory bowel disease study [57], and a BGI diabetes study [58]. The compendium was constructed using META PATHWAYS version 2.5 and Pathway Tools version 17.5. Relevant pipeline modules are summarized in Figure 1. GUTCyc provides consistent taxonomic and functional annotations, facilitates large-scale and reproducible comparisons between ePGDBs, and directly links into robust software and database resources for exploring and interpreting metabolic networks. For example, Figure 3 depicts a metabolic wall chart at three different semantic zoom levels for the metabolic network reconstruction of Human Microbiome Project sample SRS056219, as generated by META PATHWAYS (discussed further in Section ). This metabolic network reconstruction provides a multidimensional view of the microbiome that invites discovery and collaboration [31].

Methods

Metagenomic Data Sources

We gathered a collection of 431 assembled human gut shotgun metagenomes from public repositories and supplementary materials sourced from the HMP (American healthy subjects, n = 161) [55], a MetaHIT (European inflammatory bowel disease subjects, n = 125) [17], and a BGI (Chinese type 2 diabetes, n = 145) study [58].

Data Processing

Microbiome project sample metadata were manually curated to ensure compatibility with META PATHWAYS. ePGDBs were created for each sample by running the META PATHWAYS 2.5 pipeline and the PATHWAY TOOLS version 17.5, using the assembled metagenomes described above. The pipeline consists of five modular steps, including (1) quality control and ORF prediction, (2) homology-based functional and taxonomic annotation, (3) analyses consisting of tRNA and lowest common ancestor (LCA) [25] identification, (4) calculation of ePGDB using PATHWAY TOOLS and finally (5) ePGDB creation [39, 34] (Figure 1). The following paragraphs describe the individual processing steps followed to construct an ePGDB for each sample, starting with assembled contigs in FASTA format as initial file input.

Quality Control  Contigs from each sample were collected from their respective repositories and curated locally. The MetaPathways pipeline performs a number of quality control steps. First, each contig was checked for the presence of ambiguous base pairs and homopolymer runs, splitting contigs into smaller sequences by removing such problematic regions. Next, the contigs were screened for duplicates. Finally, a length cutoff of 180 base pairs was applied to the remaining sequences to ensure that very short sequences were removed from downstream processing steps [40].
Figure 1: The MetaPathways pipeline consists of five modular stages including (1) Quality control (QC) and ORF prediction (2) Functional and taxonomic annotation, (3) Analysis (4) ePGDB construction, and (5) Pathway export. Inputs and programs are depicted on the left with corresponding output directories and exported files on the right.
**ORF Prediction**  Sequences passing quality control were scanned for ORFs using MetaProdigal [26], a robust ORF prediction tool for microbial metagenomes considered to be among the most accurate ORF predictors [65]. Resulting ORF sequences were translated to amino acid sequences using the NCBI genetic code table for bacteria, archaea, and plant plastids [8]. Translated amino acid sequences shorter than 30 amino acids were removed as these sequences approached the so-called functional homology search “twilight zone”, where it becomes difficult to detect true homologs [61].

**Functional Annotation**  The quality controlled amino acid sequences were queried against a panel of functionally-annotated protein reference databases including KEGG [27] (downloaded 2011-06-18), COG [63] (downloaded 2013-12-27), MetaCyc [14] (downloaded 2011-07-03), RefSeq [62] (downloaded 2014-01-18), and SEED [52] (downloaded 2014-01-30). Protein sequence similarity searches were performed using the program LAST+ [1] version 475, with standard alignment result cutoffs (E-value less than $1 \times 10^{-5}$, bit-score greater than 20, and alignment length greater than 40 amino acids [61]; and BLAST-score ratio (BSR) greater than 0.4 [59]). The choice of parameter thresholds were selected to maximize annotation accuracy, and were guided based on parameter choices used in previous studies [22, 70, 67].

**Taxonomic Annotation**  Quality-controlled contigs were also searched against the SILVA [56] (version 115) and Greengene [16] (downloaded 2012-11-06) ribosomal RNA (rRNA) gene databases using BLAST+ version 2.2.25, with the same post-alignment thresholds applied as was previously described for the functional annotation. BLAST+ was applied for 16S annotation because it has greater sensitivity for nucleotide-nucleotide searches than LAST+, and the smaller reference database sizes make the relatively larger computational requirement justifiable.

Additionally, predicted ORFs were taxonomically annotated using the LCA algorithm [25] for taxonomic binning. In brief, the LCA in the NCBI Taxonomy Database (TaxonDB) [62] was selected based on the previously calculated LAST+ hits from the RefSeq database. This effectively sums the number of LAST+ hits at the lowest shared position of the TaxonDB. The RefSeq taxonomic names often contain multiple synonyms or alternative spellings. Therefore, names that conform to the TaxonDB were selected in preference over unknown synonyms.

**tRNA Scan**  MetaPathways uses TRNASCAN-SE version 1.4 [44] to identify relevant tRNAs from quality-controlled sequences. Resulting tRNA identifications are appended as additional functional annotations.

**ePGDB Creation**  Functional annotations were parsed and separated into three files that serve as inputs to PATHWAY TOOLS, namely: (1) an annotation file containing gene product information (.pf), (2) a catalog of con-
tigs and scaffolds (genetic-elements.dat), and (3) a PGDB parameters file (organism-init.dat). The PathoLogic module[19][15] in the PATHWAY TOOLS software, was used to build the ePGDB and predict the presence of metabolic pathways based on functional annotations. Following ePGDB construction, the base pathways (i.e., pathways that are not contained within super-pathways) in the sample were extracted from ePGDBs to generate a summary table of predicted metabolic pathways for each sample.

**Accessibility and Flexibility**  MetaPATHWAYS 2.5 generates data in a consistent file and directory structure. The output for each sample is contained within a single directory, which in turn is organized into sub-directories containing relevant files (Figure 1). The MetaPATHWAYS 2.5 graphical user interface (GUI) enables interactive exploration of individual sample results along with comparative queries of multiple samples, and is designed for fast and interactive data visualization and searches via a custom knowledge engine data structure. Input and output files are available for download from the GutCyc website [www.gutcyc.org](http://www.gutcyc.org) and may be readily explored in the MetaPATHWAYS GUI or PATHWAY TOOLS on Linux, Mac OS X and Windows machines.

**Computational Environment**  Computational processing was performed using a local cluster of machines in the Hallam laboratory and on the Bugaboo cluster on the Canadian WestGrid computation resource [5]. The Hallam lab computers have a configuration profile of 2x2.4 GHz Quad-Core Intel Xeon processors with 64 GB 1066 MHz DDR3 RAM. The Bugaboo cluster provides 4,584 cores with 2 GB of RAM per core on average. The average sample took 7-8 hours to process on a single thread, and the span of the processing required to generate the GutCyc COLLECTION was 135 days.

**Software Availability**  MetaPATHWAYS 2.5, including integrated third party software, is available on GitHub, including both software [2] (licensed under the GNU General Public License, version 3), and a tutorial [3] released under the Creative Commons Attribution License (allows reuse, distribution, and reproduction given proper citation). PATHWAY TOOLS is available under a free license for academic use, and may be commercially licensed as well [4]. METAPATHWAYS outputs were processed using PATHWAY TOOLS version 17.5 under default settings except for disabling of the PathoLogic taxonomic pruning step (i.e., -no-taxonomic-pruning) as was described previously [22], and an additional refinement step of running the Transport Inference Parser [43] to predict transport reactions (i.e., -tip). LAST+ is freely available under a (licensed under the GNU General Public License, version 3) software license on our GitHub page [1].

---

[1]: http://dx.doi.org/10.1101/055574
[2]: https://github.com/GutCyc
[3]: https://creativecommons.org/licenses/by-nc/4.0/
[4]: https://www.gutcyk.org
[5]: http://www-westgrid.ca
Table 1: Summary statistics for the GutCyc Collection across 431 samples. The statistics for the number of bases processed is in units of Megabases.

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>1st Quartile</th>
<th>Median</th>
<th>3rd Quartile</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bases</td>
<td>0.98</td>
<td>52.63</td>
<td>79.28</td>
<td>113</td>
<td>3,705</td>
</tr>
<tr>
<td>Contigs</td>
<td>2,506</td>
<td>28,340</td>
<td>48,360</td>
<td>77,480</td>
<td>399,300</td>
</tr>
<tr>
<td>ORFs</td>
<td>2,448</td>
<td>65,570</td>
<td>102,300</td>
<td>149,900</td>
<td>636,000</td>
</tr>
<tr>
<td>Pathways</td>
<td>256</td>
<td>571</td>
<td>629</td>
<td>672</td>
<td>831</td>
</tr>
</tbody>
</table>

Data Records

All records are available at the GutCyc project website (www.gutcyc.org). Each sample's data records are contained within a single directory. Within this directory, sub-directories and files are located as depicted in Figure 1. A summary of the data present in the GutCyc Collection is presented in Table 1.

preprocessed For a sample with an identifier of <sample_ID>, this directory contains two files: (1) <sample_ID>.fasta, which contains the renamed, quality-controlled sequences, and (2) <sample_ID>.mapping.txt, which maps the original sequence names to the new names assigned by MetaPathways. Sequences are renamed to <sample_ID>_X where X is the zero-indexed contig number pertaining to the order in which the contig appears in the input file (e.g., a contig identified as DLF001_27 is interpreted as the 28th contig listed in the FASTA file for sample DLF001’s assembly).

orf_prediction This directory contains four files, (1) <sample_ID>.fna which contains nucleic acid sequences of all predicted ORFS, (2) <sample_ID>.faa which contains amino acid sequences of all predicted ORFs, (3) <sample_ID>.qced.faa which contains amino acid sequences of all predicted ORFs meeting user-defined quality thresholds (in this study, a minimum length of 60 amino acids), and (4) <sample_ID>.gff, a General Feature Format (GFF) file containing all quality-controlled sequences and information about the strand (- or +) on which the ORF was predicted. ORFs are named <sample_ID>_X_Y, where X is the contig number pertaining to the order in which the contig appears and Y represents the order in which the ORFs were predicted on the contig.

results This directory contains four sub-directories: (1) annotation_table, (2) rRNA, (3) tRNA, and (4) pgdb. The annotation_table sub-directory contains (1) statistics files for each functional database used to annotate the ORFs (<sample_ID>.<DB>_stats_<index>.txt), (2) <sample_ID>.functional_and_taxonomic_table.txt detailing the length, location, strand and annotation (functional and taxonomic) of each ORF, and (3) a file listing all ORFs and their functional annotations (<sample_ID>.ORF_annotation_table.txt). The prokaryotic 16S ribosomal
RNA gene is a standard marker gene used for measuring taxonomic diversity [66]. The rRNA sub-directory contains files detailing statistics for each taxonomic database used to annotate the ORFs (named as <sample_ID>.<DB>.rRNA.stats.txt). The tRNA sub-directory contains (1) <sample_ID>.trna.stats.txt, detailing the type, anticodon, location and strand of each predicted tRNA and (2) <sample_ID>.tRNA.fasta containing all predicted tRNA sequences. The pgdb sub-directory contains a <sample_ID>.pwy.txt file describing metabolic pathways predicted in the ePGDB, specifically, each predicted pathway, the ORF identities involved in each pathway, the enzyme abundance, and the pathway coverage in a tabular format navigable via the MetaPATHWAYS GUI.

genbank This directory contains a file named <sample_ID>.annotated.gff, a GFF file containing all quality-controlled sequences with their annotations.

ptools This directory contains the three files necessary to build ePGDBs, (1) genetic-elements.dat and (2) organism-params.dat both of which contain sample-naming and species information needed to build ePGDBs, and (3) 0.pf which contains all functional annotations to be processed by PATHWAY TOOLS. A sub-directory called flat-files contains flat files describing database objects such as compounds, reactions, pathways (each of which is described in more detail in [20]) for individual ePGDB.

run_statistics This directory contains three files, (1) <sample_ID>.run.stats, which details the parameters used to process the sample, (2) <sample_ID>.nuc.stats, which details the number and length of nucleic acid sequences before and after the quality control filtering, and (3) <sample_ID>.amino.stats, which details the number and length of amino acid sequences before and after the quality control filtering.

Technical Validation

GUTCyc was derived from third-party sequence data from three publicly-available human gut microbiome sampling projects with metagenomic assemblies, with published details on their own technical validation steps: the HMP [55], a MetaHIT study [17], and a BGI study [58]. The technical validation of third-party software used in MetaPATHWAYS may be found in the corresponding publications for MetaProdigal [20], BLAST+ [10], and tRNAscan-SE [44]. GUTCyc functional sequence similarity was computed using LAST+, our improved version of LAST [30], with multi-threading performance improvements and new support for generating BLAST-like E-values, with significant correlation with BLAST+ output ($R^2 = 0.887, P < 0.01$) [39]. Validation of the overall MetaPATHWAYS pipeline may be found in previously published reports [22, 24, 23] with specific emphasis on how changes in taxonomic pruning, read length and assembly coverage impact the accuracy and sensitivity
of pathway recovery. In brief, pathway recovery is affected by taxonomic distance, sequence coverage and sample diversity, nearing an asymptote of maximum accuracy for metagenomes with increasing coverage. Additionally, like any alignment-based analysis, annotation quality is a function of both the level of errors in the input sequence data and the selection of reference databases. The metabolic reconstruction pathways were computationally predicted using the Pathway Tools PathoLogic module [53], which has an accuracy of 91% [15]).

Usage Notes

The once a set of data such as GutCyc Collection has been crafted into a format that is both comprehensible to domain experts, and also interpretable by machine, there are myriads of uses that can be explored. The ePGDBs can be used for visualizing meta-transcriptomics and meta-metabolomics data. Comparing ePGDBs with sets of microbial PGDBs from the same environment can aid in identifying “distributed pathways” present in the metagenome metabolic reconstruction, but absent from any individual genomic metabolic reconstruction [22]. The predicted transport proteins can be used to predict trophism patterns within a community. The PATHWAY TOOLS software allows for sophisticated comparative analyses between ePGDBs, at the level of compounds, reactions, enzymes, and pathways [32]. The METAFLUX [41] module of PATHWAY TOOLS for performing flux balance analysis (FBA) [51] can be used with GutCyc ePGDBs to generate quantitative simulations of microbiome growth and pathway flux. A set of microbiome metabolic models also facilitates the exploration of the impact of xenobiotics [21], and provides a computational substrate for systems biology approaches to engineering the gut microbiome [47]. Figure 2 demonstrates the user interface for MetaPathways and PATHWAY TOOLS, along with example data analysis use cases.

In this section we motivate further two specific use cases for GutCyc. In the first case, we demonstrate how to use a GutCyc ePGDB to determine the metabolic path between two small molecules of interest. In the second case, we use GutCyc to visualize different levels of biological information, e.g. metabolomics data, in the context of a microbiome metabolic network.

Optimal Metabolite Tracing

The PATHWAY TOOLS software provides advanced biochemical querying capabilities for both PGDBs and ePGDBs. One such capability is energy-optimal metabolite tracing. To summarize, given both a starting and a terminal/target compound within an ePGDB, PATHWAY TOOLS is able to compute the shortest and most energetically-favorable route through the metabolic reaction network. While there is no guarantee that, in a complex milieu such as the gut microbiome, the syntrophic flux will necessarily follow a short and minimal energy path, these criteria allow us to narrow down a multiplicity of possible paths to a single parsimonious candidate path.
Figure 2: **GutCyc ePGDB use cases.** In (A), a **GutCyc ePGDB** is opened in **MetaPathways**. Clockwise from the upper left, we display the Pipeline execution step, Exportable summary sample statistics, Exportable pathway abundance table, and Monitor sampling processing interfaces. In (B), a **GutCyc ePGDB** is opened in **Pathway Tools**. Clockwise from the upper left, we display the ePGDB summary statistics, Pathway viewer, Biochemical reaction information, Interactive metabolic network visualization, and Metabolic Route Search interfaces.
In a remarkable study by Koeth et al., they demonstrated a causal connection between the intestinal gut microbiota’s metabolism of red meat and the promotion of atherosclerosis [37]. In brief, the gut microbiome is capable of transforming excess L-carnitine into trimethylamine (TMA), which is further processed by the liver to form the cardiovascular disease-associated metabolite trimethylamine N-oxide (TMAO). Using this biotransformation as a motivating case, we queried the GutCyc SRS015217Cyc ePGDB for the biochemical reaction path from L-carnitine to TMA, which is not provided explicitly by Koeth et al. Utilizing the PATHWAY TOOLS Metabolic Route Search feature, we found an optimal path between L-carnitine to TMA, using the MetaCyc carnitine degradation II pathway (PWY-3602, expected in Proteobacteria) along with a betaine reductase reaction (EC 1.21.4.4; found in Clostridium sticklandii and Eubacterium acidaminophilum, both species affiliated with the order Clostridiales), minimizing the number of enzymes involved and chemical bond rearrangements. PATHWAY TOOLS found the optimal path in seconds, displayed in Figure 2.

L-carnitine and glycine betaine have known transporter families that facilitate their movement across the cell membrane [48], as do TMA and TMAO [49], and thus this metabolic route may be a distributed pathway [22]. In fact, no single PGDB in the BioCyc Collection of over 5,500 microbial genomes (release 19.0 [14]), has both the carnitine degradation II pathway and the betaine reductase reaction, which suggests that there is no single microbe capable of completing this entire metabolic route.

The metabolic route identified may also help generate mechanistic hypotheses from microbiome study observations. Among the findings reported in [37] in the Supplementary Materials is that all statistically-significant correlations (positive or negative) found between plasma TMAO levels and the proportion of various taxa, involved species affiliated with the order Clostridiales, which is the subsuming taxon of the betaine reductase reaction curated in MetaCyc. This indicates that Clostridiales are integral to understanding the regulation of TMA and TMAO concentrations in the gut, which in turn affects plasma concentrations. This demonstrates the power of ePGDBs in computing connections between nutritional or pharmaceutical inputs (such as L-carnitine) to identify potential interactions with known disease biomarkers (as TMAO is to cardiovascular disease).

High-Throughput Data Visualization

Another capability of PATHWAY TOOLS is to visualize the results of high-throughput experiments mapped onto the Cellular, Genome, and Regulation Overviews, or as “Oomics Pop-Ups” when viewing a particular pathway [54]. Specifically, PATHWAY TOOLS provides support for the analysis of mass spectrometry data, by automatically mapping a list of monoisotopic masses to matching entries in MetaCyc, or in specific ePGDBs [30]. As a demonstration of this capability, we analyzed mass-spectrometry data from a metabolomic study of humanized mice microbiomes [45]. The dataset contained 867 unique
masses, of which 453 masses were identified using MetaCyc by performing standard adduct monoisotopic mass manipulations, followed by monoisotopic mass search using Pathway Tools. We mapped the identified compounds on the GutCyc Cellular Overview, as seen in Figure 3. This facilitates turning a massive table of data into a more intuitive construct based on the community metabolic interaction network, enabling more efficient pattern matching. For example, using the enrichment analysis tools in Pathway Tools, we identified the pathway class of “Secondary Metabolites Degradation” as enriched for identified compounds ($p = 2.0 \times 10^{-2}$, Fisher Exact Test with Benjamini-Hochberg multiple testing correction). By visually inspecting the pathways in the class, we can see that pathway P562-PWY, “myo-, chiro-, and scillo-inositol degradation pathway”, has four matched compounds from the metabolomics dataset.

Acknowledgments

We would like to thank Peter D. Karp for feedback on the MetaPathways software and the GutCyc project; Robert Pesich for orchestrating our sneakernet transfer of data; and Les Dethlefsen for assisting in loading the data onto the server. A special thanks to the members of the Hallam, Relman, and Dill labs, and Whole Biome, for constructive feedback on the GutCyc project. Thank you to Pallavi Subhraveti of SRI International for help with exporting GutCyc data using Pathway Tools. Thank you to the Stanford FarmShare computation resource, for aiding in the development of an early version of GutCyc.

The GutCyc project at UBC was carried out under the auspices of Compute/Calcul Canada, Genome Canada, Genome British Columbia, Genome Alberta, the Natural Science and Engineering Research Council (NSERC) of Canada, Ecosystem Services, Commercialization Platforms and Entrepreneurship (ECOSCOPE) program, the Canadian Foundation for Innovation (CFI), and the Canadian Institute for Advanced Research (CIFAR) through grants awarded to SJH. ASH was supported by the Alexander Graham Bell Canada Graduate Scholarships-Doctoral Program (CGS D) administered by NSERC. KMK was supported by the Tula Foundation funded Centre for Microbial Diversity and Evolution (CMDE) at UBC. NWH was supported by a four year doctoral fellowship (4YF) administered through the UBC Faculty of Graduate and Postdoctoral Studies. TA was partially supported by the Stanford University School of Medicine Dean’s Funds and the NIH Biotechnology Training Grant at Stanford (grant number 5T32 GM008412). TA and DLD were partially supported by a King Abdullah University of Science and Technology (KAUST) research grant under the KAUST Stanford Academic Excellence Alliance program. DAR was supported by NIH/NIGMS 5R01GM099534 and by the Thomas C. and Joan M. Merigan Endowment at Stanford University. Additional computational resources were provided gratis through the Stanford FarmShare resource.
Figure 3: The Cellular Overview for the SRS056259CYC ePGDB, at three different zoom levels (i.e., metabolic map, pathway class, and individual pathway), with compounds highlighted in red if identified from a mass spectrometry analysis of the gut microbiome [45]. Compounds with no mass spectrometry highlights appear as grey icons. Reactions with enzyme data in SRS056259CYC are drawn in blue. Bottom right inset shows a fraction of the full metabolic map. Top inset shows zoom-in of the “Secondary Metabolite Degradation” pathway class. Bottom left inset shows zoom-in on Pathway P562-PWY, “myo-, chiro-, and scillo-inositol degradation pathway”, showing four mass-spectrometry identified compounds in red.
Author Contributions

Tomer Altman and Steven Hallam conceived of the GutCyc project as part of a movement to develop the Environmental Genome Encyclopedia (EngCyc) a compendium of microbial community metabolic blueprints supported by high performance software tools on grids and clouds. Niels Hanson, Kishori Konwar, Aria Hahn and Dongjae Kim developed the MetaPathways software pipeline with direction from Steven Hallam and assistance from Tomer Altman and others at SRI International. Dongjae Kim, Aria Hahn and Kishori Konwar compiled the microbiome sequence datasets, constructed GutCyc ePGDBs and created figures for the manuscript. Tomer Altman generated validation datasets and drafted an early version of the manuscript with Aria Hahn and Steven Hallam. Dongjae Kim developed the GutCyc website. All authors contributed to the final preparation of the manuscript. Steven Hallam, David L. Dill and David A. Relman supervised the project. All authors reviewed and approved the final manuscript.

Competing financial interests

Authors AH, KK, and SJH are founders of Koonkie Cloud Services, a company offering commercial support for MetaPathways. The authors offer licensed support for customized use of the GutCyc Collection.

Data Availability

The GutCyc Collection, along with metadata for all samples, is freely available at www.gutcyc.org.

Glossary

**API** Application Programming Interface. 3

**BGI** Beijing Genomics Institute. 2, 4, 9

**BSR** BLAST-score ratio. 6

**ePGDB** environmental Pathway/Genome Database. 3–5, 7, 9–12, 14

**FBA** flux balance analysis. 10

**GFF** General Feature Format. 8, 9

**GUI** graphical user interface. 7, 9
HMP Human Microbiome Project. 3, 4, 9

KAUST King Abdullah University of Science and Technology. 13

KEGG Kyoto Encyclopedia of Genes and Genomes. 3

KO KEGG Orthology. 3

LCA lowest common ancestor. 4, 6

MetaHIT MetaHIT Project. 4, 9

ORF open reading frame. 4–6, 8, 9

PGDB Pathway/Genome Database. 3, 7, 10, 12

TaxonDB NCBI Taxonomy Database. 6

TMA trimethylamine. 12

TMAO trimethylamine N-oxide. 12

References


