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

Aria S. Hahn[1,2+] Tomer Altman[3,4+]
Kishori M. Konwar[1,2,5+] Niels W. Hanson[1]
Dongjae Kim[6] David A. Relman[7,8,9] David L. Dill[10]
Steven J. Hallam[1,2,11*]

July 31, 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, USA 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, 299 Campus Drive, 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

Advances in high-throughput sequencing are reshaping how we perceive microbial communities inhabiting the human body, with implications for therapeutic interventions. Several large-scale datasets derived from hundreds of human microbiome samples sourced from multiple studies are now publicly available. However, idiosyncratic data processing methods between studies introduce systematic differences that confound comparative analyses. To overcome these challenges, we developed GutCyc, a compendium of environmental pathway genome databases constructed from 418 assembled human microbiome datasets using MetaPathways, enabling reproducible functional metagenomic annotation. We also generated metabolic network reconstructions for each metagenome using the Pathway Tools software, empowering researchers and clinicians interested in visualizing and interpreting metabolic pathways encoded by the human gut microbiome. For the first time, GutCyc provides consistent annotations and metabolic pathway predictions, making possible comparative community analyses between health and disease states in inflammatory bowel disease, Crohn's disease, and type 2 diabetes. GutCyc

data products are searchable online, or may be downloaded and explored locally using MetaPathways and Pathway Tools.

Background & Summary

The myriad collections of microorganisms found on and in the human body 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 [34, 11]. Increasingly, researchers are using high-throughput sequencing approaches to study the genes and genomes of microbiomes 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 medicine. 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 (HMP) [55], Beijing Genomics Institute (BGI) [58], and Metagenomes of the Human Intestinal Tract project (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 provide the software environment used for generating their annotations. In contrast to these proprietary pipelines, over the past few years a number of metagenomic annotation pipelines available to third parties have emerged including IMG/M [46], Metagenome Rapid Annotation using Subsystem Technology (MG-RAST) [68], SmashCommunity [9] and HUMAnN [6]. Differing pipelines used to process sequence information between studies introduces biases based on idiosyncratic formatting, and alternative annotations or algorithmic methods. Specifically, 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) [26]. Although extant pipelines often provide links to KEGG module and pathways maps [26] (using KEGG Orthology (KO) or pathway identifiers) that 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 implicit metabolic model directly. Moreover, prediction using KEGG results in amalgamated pathways with limited taxonomic resolution, impeding enrichment and association studies [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 [37] using Pathway Tools [27] and MetaCyc [32, 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). 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 compendium of over 418 ePGDBs constructed from public shotgun metagenome datasets generated by the HMP [55], the MetaHIT inflammatory bowel disease study [57], and the BGI diabetes study [58]. 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. This metabolic network reconstruction provides a multidimensional view of the microbiome that invites discovery and collaboration [30].

Methods

Metagenomic Data Sources

We collected 418 assembled human gut shotgun metagenomes from public repositories and supplementary materials sourced from the HMP (American healthy subjects, $n = 148$) [55], a MetaHIT (European inflammatory bowel disease subjects, $n = 125$) [17], and a BGI (Chinese type 2 diabetes, $n = 145$) study [58]. See Supplementary Table 1 for a detailed listing of accession numbers and file descriptors.

Data Processing

Microbiome project sample metadata were manually curated to ensure compatibility with MetaPathways. ePGDBs were created for each sample by running the MetaPathways 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-

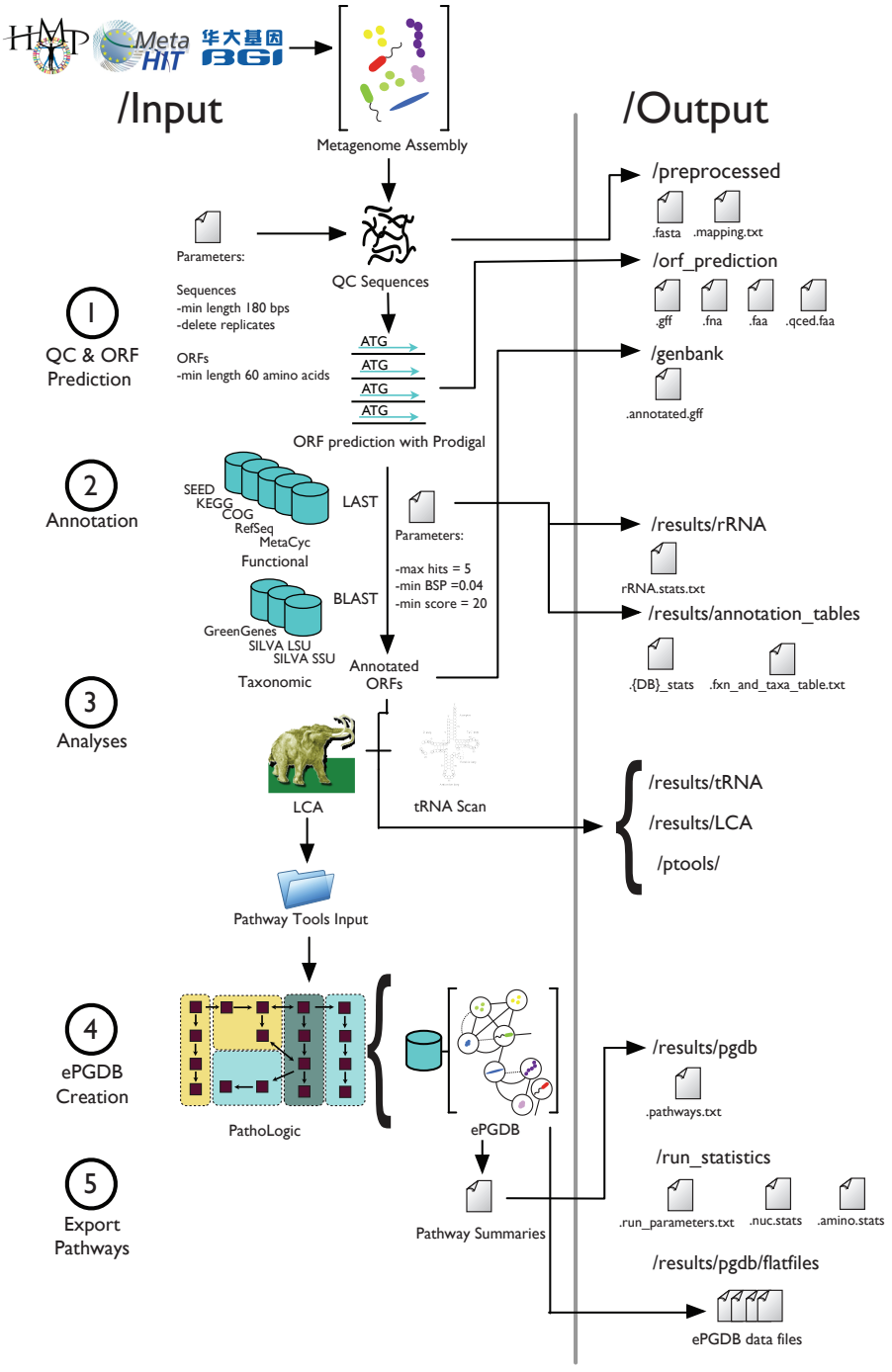

Figure 1: The MetaPathways pipeline consists of five modular stages including (1) Quality control (QC) and ORF prediction (2) Functional and taxonomic annotation, (3) Analysis (4) ePGDB construction, and (5) Pathway export. Inputs and programs are depicted on the left with corresponding output directories and exported files on the right.

based functional and taxonomic annotation, (3) analyses consisting of tRNA and lowest common ancestor (LCA) [24] identification, (4) construction of ePGDBs using PATHWAY TOOLS and, finally, (5) pathway export [38, 33] (see Figure 1). The following paragraphs describe the individual processing steps required to construct an ePGDB for each sample, starting with assembled contigs in FASTA format.

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 [39].

ORF Prediction Sequences passing quality control were scanned for ORFs using METAPRODIGAL [25], 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 NCBI genetic code table 11 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 [26] (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 FAST [42] with standard alignment result cutoffs (E-value less than 1×10^{-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 FAST, and the smaller reference database sizes make the relatively larger computational requirement justifiable.

Additionally, predicted ORFs were taxonomically annotated using the LCA algorithm [24] for taxonomic binning. In brief, the LCA in the NCBI Taxonomy Database (TaxonDB) [62] was selected based on the previously calculated FAST hits from the RefSeq database. This effectively sums the number of FAST 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 (`0.pf`), (2) a catalog of contigs and scaffolds (`genetic-elements.dat`), and (3) a PGDB parameters file (`organism-params.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 superpathways) 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 (see 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`) 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.

	Min	1^{st} Quartile	Median	3^{rd} Quartile	Max
Bases	0.98	54.75	81.35	113.75	370.51
Contigs	2,506	27,788	47,486.5	76,275.75	399,331
ORFs	2,448	61,703.5	95.531	139,690	550,312
Func. Annots.	2,176	57,102.25	86,054.5	123,747.25	425,033
Reactions	1,635	2,385.5	3,438	3,667.75	4,881
Trans. Reactions	12	26	31	34	46
Compounds	1,052	1,678	2,008.5	2,119.5	2,676
Base Pathways	257	350	616	654	832

Table 1: Summary statistics for the GutCyc Collection across 418 samples. The statistics for the number of bases processed is in units of Megabases. "Func. Annots.": functional annotations. "Trans. Reactions" are transport reactions. "Compounds" are small molecule metabolites. "Base Pathways" include all pathways except complex pathways known as Super-Pathways.

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 [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`). FAST is freely available under a (licensed under the GNU General Public License, version 3) software license on our GitHub page [1].

Data Records

A list of each sample, provenance, and relevant data processing steps can be found in Supplementary Table 1. 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. A full set of summary data for each ePGDB may be found in Supplementary Table 2.

`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 28^{th} 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 [18] 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 a ePGDB using Pathway Tools: (1) `genetic-elements.dat`, (2) `organism-params.dat`, 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 [28]) for individual ePGDBs.

`run_statistics` This directory contains three files: (1) `<sample_ID>.run.stats`, the parameters used to process the sample; (2) `<sample_ID>.nuc.stats`, the number and length of nucleic acid sequences before and after quality control filtering; and (3) `<sample_ID>.amino.stats`, the number and length of amino acid sequences before and after 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 [25], BLAST [10], and tRNAscan-SE [44]. GutCyc functional sequence similarity was computed using FAST, an aligner based on a version of LAST [35], with multi-threading performance improvements and new support for generating BLAST-like E-values, with significant correlation with BLAST output ($R^2 = 0.887$, $P < 0.01$) [38]. Validation of the overall MetaPathways pipeline may be found in previously published reports [22, 23] with specific emphasis on how changes in taxonomic pruning, read length and metagenomic assembly coverage impact the accuracy and sensitivity of pathway recovery. In brief, pathway prediction 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. Summary data generated for each ePGDB as presented in Supplementary Table 2 was reviewed to detect samples with unusual statistics, such as a lack of 16S gene annotations. The metabolic reconstruction pathways were computationally predicted using the Pathway Tools PathoLogic module [53], which has an accuracy of 91% [15]).

Usage Notes

Once a set of data such as GutCyc Collection has been crafted into a format that is both comprehensible to domain experts, and interpretable by machines, there are myriads of uses that can be explored. For example, 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. Furthermore, the Pathway Tools software allows for sophisticated comparative analyses between ePGDBs, at the level of compounds, reactions, enzymes, and pathways [31]. The MetaFlux [40] 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.

In a 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 [36]. 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],

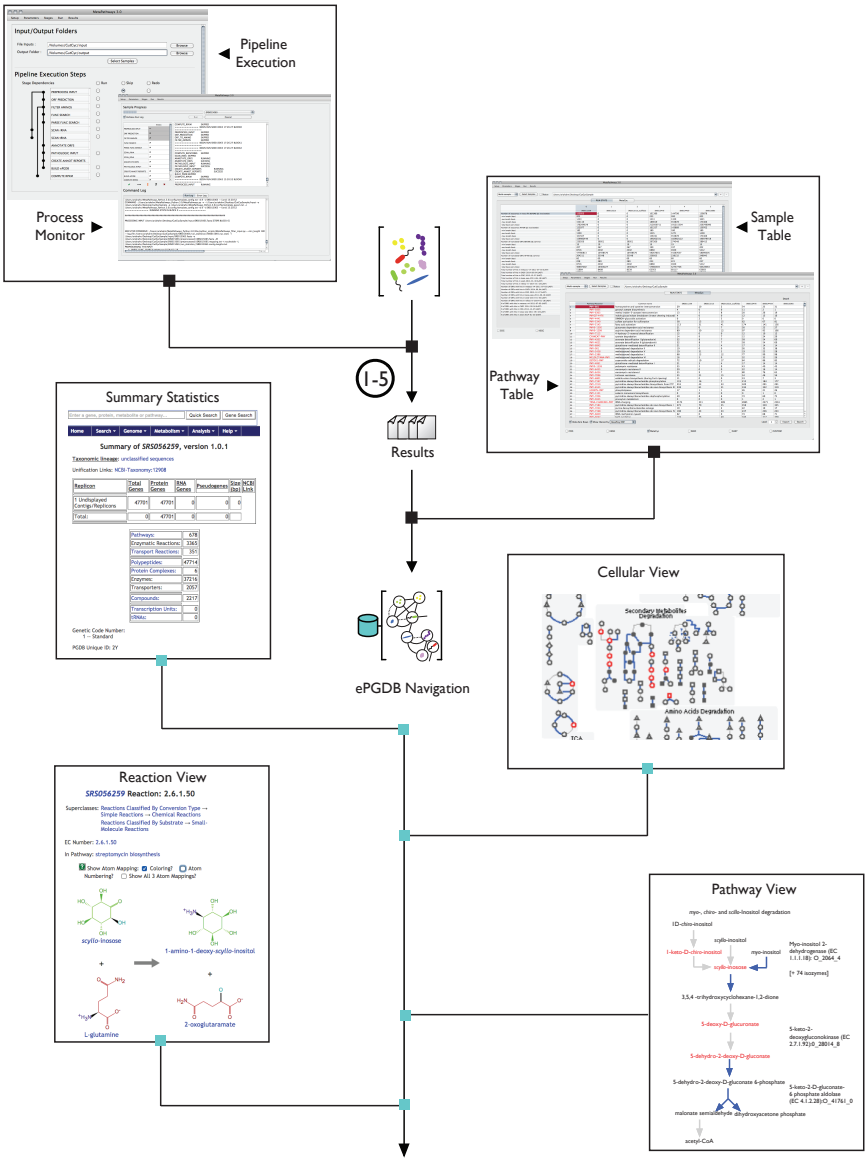

Figure 2: GutCyc ePGDB use cases. In the upper left and upper right insets, a GutCyc ePGDB is opened in MetaPathways. In the upper left, we display the Pipeline execution step, and the Process Monitor interfaces. In the upper right, we display the Summary Table (with exportable sample statistics), and the Pathway Table (with exportable pathway abundances) interfaces. In the lower for inset images, a GutCyc ePGDB is opened in Pathway Tools. Clockwise from the upper left, we display the ePGDB summary statistics, interative metabolic network visualization, the Pathway View, and the biochemical Reaction View.

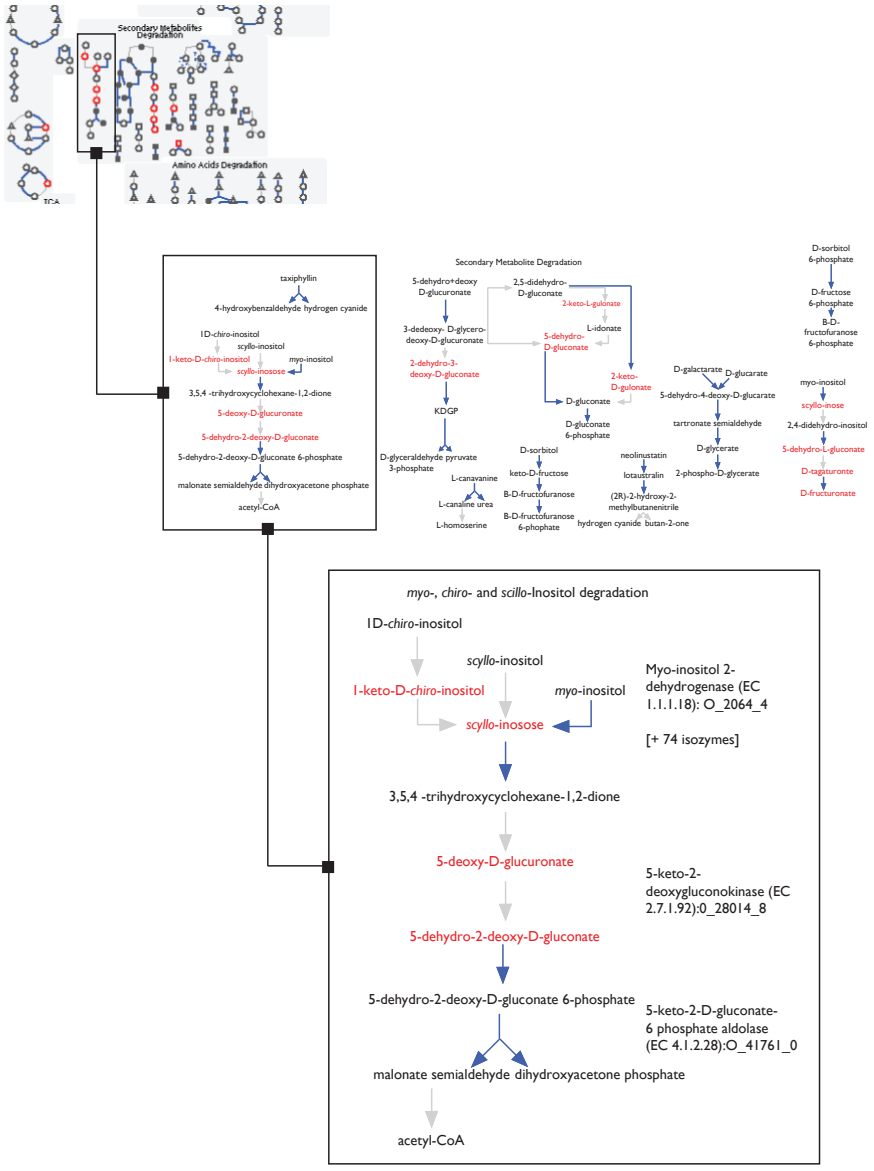

Figure 3: The Cellular Overview for the SRS056259Cyc ePGDB, at three different zoom levels, with compounds highlighted in red if identified from a mass spectrometry analysis of the gut microbiome [45]. Compounds with no mass spectrometry highlights appear as grey icons. Reactions with enzyme data in SRS056259Cyc are drawn in blue. The top left inset shows a fraction of the full metabolic map. The middle inset shows a zoom-in of the "Secondary Metabolite Degradation" pathway class. Bottom right inset shows zoom-in on Pathway P562-PWY, "myo-, chiro-, and scillo-inositol degradation pathway", showing four mass-spectrometry identified compounds in red.

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 [36] in the Supplementary Materials is that *all statistically-significant correlations (positive or negative) found between plasma TMAO levels and species abundance, involved species affiliated wth the order Clostridiales*, which is the subsuming taxon of the betaine reductase reaction's taxonomic range, as 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 "Omics 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 [29]. 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 MetaCycby performing standard adduct monoisotopic mass manipulations [64], followed by monoisotopic mass search using Pathway Tools. We mapped the identified compounds on the GutCyc Cellular Overview [41], 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 [29], 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 Relman Lab 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.

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. 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`. See below for the figshare DOI Data Citation.

Glossary

BGI Beijing Genomics Institute. 2, 3, 9

BSR Blast-score ratio. 5

ePGDB environmental Pathway/Genome Database. 3–13

FBA flux balance analysis. 10

GFF General Feature Format. 8

GUI graphical user interface. 6, 8

HMP Human Microbiome Project. 2, 3, 9

KAUST King Abdullah University of Science and Technology. 14

KEGG Kyoto Encyclopedia of Genes and Genomes. 2, 3

KO KEGG Orthology. 2

LCA lowest common ancestor. 5, 6

MetaHIT Metagenomes of the Human Intestinal Tract project. 2, 3, 9

ORF open reading frame. 3–6, 8

PGDB Pathway/Genome Database. 3, 6, 9, 10, 13

TaxonDB NCBI Taxonomy Database. 6

TMA trimethylamine. 10, 13

TMAO trimethylamine *N*-oxide. 10, 13

References

[1] GitHub: FAST. `https://github.com/hallamlab/FAST`. Accessed on: 2016-02.

[2] GitHub: MetaPathways. `https://github.com/hallamlab/metapathways2`. Accessed on: 2016-02.

[3] GitHub: MetaPathways Tutorial. `https://github.com/hallamlab/mp_tutorial/`. Accessed on: 2016-02.

[4] Licensing Pathway Tools. `http://www.biocyc.org/download-bundle.shtml`. Accessed on: 2016-02.

[5] Bugaboo Cluster on ComputeCanada WestGrid Resource. `https://www.westgrid.ca/support/systems/bugaboo`, 2016.

[6] Sahar Abubucker, Nicola Segata, Johannes Goll, Alyxandria M. Schubert, Jacques Izard, Brandi L. Cantarel, Beltran Rodriguez-Mueller, Jeremy Zucker, Mathangi Thiagarajan, Bernard Henrissat, Owen White, Scott T. Kelley, Barbara Methé, Patrick D. Schloss, Dirk Gevers, Makedonka Mitreva, and Curtis Huttenhower. Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput Biol*, 8(6):e1002358, 2012.

[7] Tomer Altman, Michael Travers, Anamika Kothari, Ron Caspi, and Peter D. Karp. A systematic comparison of the MetaCyc and KEGG pathway databases. *BMC Bioinformatics*, 14:112, 2013.

[8] Elzanowski Andrzej and Ostell Jim. The Bacterial, Archaeal and Plant Plastid Code. www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi#SG11, 2013.

[9] Manimozhiyan Arumugam, Eoghan D. Harrington, Konrad U. Foerstner, Jeroen Raes, and Peer Bork. SmashCommunity: a metagenomic annotation and analysis tool. *Bioinformatics*, 26(23):2977–2978, Dec 2010.

[10] Grzegorz M Boratyn, Christiam Camacho, Peter S Cooper, George Coulouris, Amelia Fong, Ning Ma, Thomas L Madden, Wayne T Matten, Scott D McGinnis, Yuri Merezhuk, Yan Raytselis, Eric W Sayers, Tao Tao, Jian Ye, and Irena Zaretskaya. BLAST: a more efficient report with usability improvements. *Nucleic Acids Research*, 41(Web Server issue):W29–33, 2013.

[11] Scott J. Bultman. Emerging roles of the microbiome in cancer. *Carcinogenesis*, 35(2):249–255, Feb 2014.

[12] Benjamin Callahan, Diana Proctor, David Relman, Julia Fukuyama, and Susan Holmes. Reproducible research workflow in R for the analysis of personalized human microbiome data. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 21:183–94, 2016.

[13] Ron Caspi, Tomer Altman, Richard Billington, Kate Dreher, Hartmut Foerster, Carol A. Fulcher, Timothy A. Holland, Ingrid M. Keseler, Anamika Kothari, Aya Kubo, Markus Krummenacker, Mario Latendresse, Lukas A. Mueller, Quang Ong, Suzanne Paley, Pallavi Subhraveti, Daniel S. Weaver, Deepika Weerasinghe, Peifen Zhang, and Peter D. Karp. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res*, 42(Database issue):D459–D471, Jan 2014.

[14] Caspi, Ron and Billington, Richard and Ferrer, Luciana and Foerster, Hartmut and Fulcher, Carol A and Keseler, Ingrid M and Kothari, Anamika and Krummenacker, Markus and Latendresse, Mario and Mueller, Lukas A and Ong, Quang and Paley, Suzanne and Subhraveti, Pallavi and Weaver, Daniel S and Karp, Peter D. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic acids research*, 44(D1):D471–80, 2016.

[15] Joseph M. Dale, Liviu Popescu, and Peter D. Karp. Machine learning methods for metabolic pathway prediction. *BMC Bioinformatics*, 11:15, 2010.

[16] DeSantis, T. Z. and Hugenholtz, P. and Larsen, N. and Rojas, M. and Brodie, E. L. and Keller, K. and Huber, T. and Dalevi, D. and Hu, P. and Andersen, G. L. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology*, 72(7):5069–5072, 2006.

[17] S. Dusko Ehrlich and MetaHIT Consortium . [Metagenomics of the intestinal microbiota: potential applications]. *Gastroenterol Clin Biol*, 34 Suppl 1:S23–S28, Sep 2010.

[18] Eilbeck, Karen and Lewis, Suzanna E and Mungall, Christopher J and Yandell, Mark and Stein, Lincoln and Durbin, Richard and Ashburner, Michael. The Sequence Ontology: a tool for the unification of genome annotations. *Genome biology*, 6(5):R44, 2005.

[19] M. L. Green and P. D. Karp. A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Bioinformatics*, 5:76, Jun 2004. [PubMed Central:PMC446185] [DOI:10.1186/1471-2105-5-76] [PubMed:15189570].

[20] Sharon Greenblum, Peter J. Turnbaugh, and Elhanan Borenstein. Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. *Proc Natl Acad Sci U S A*, 109(2):594–599, Jan 2012.

[21] Henry J Haiser and Peter J Turnbaugh. Developing a metagenomic view of xenobiotic metabolism. *Pharmacological research*, 69(1):21–31, 2013.

[22] Niels W. Hanson, Kishori M. Konwar, Alyse K. Hawley, Tomer Altman, Peter D. Karp, and Steven J. Hallam. Metabolic pathways for the whole community. *BMC Genomics*, 15:619, 2014.

[23] Niels W. Hanson, Kishori M. Konwar, Shang-Ju Wu, and Steven J. Hallam. Metapathways v2.0: A master-worker model for environmental pathway/genome database construction on grids and clouds. *2014 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology*, May 2014.

[24] Daniel H. Huson and Nico Weber. Microbial community analysis using MEGAN. *Methods Enzymol*, 531:465–485, 2013.

[25] Hyatt, D. and LoCascio, P. F. and Hauser, L. J. and Uberbacher, E. C. Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics*, 28(17):2223–2230, 2012.

[26] Minoru Kanehisa, Susumu Goto, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res*, 42(Database issue):D199–D205, Jan 2014.

[27] P. D. Karp, S. M. Paley, M. Krummenacker, M. Latendresse, J. M. Dale, T. J. Lee, P. Kaipa, F. Gilham, A. Spaulding, L. Popescu, T. Altman, I. Paulsen, I. M. Keseler, and R. Caspi. Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. *Brief. Bioinformatics*, 11(1):40–79, Jan 2010. [PubMed Central:PMC2810111] [DOI:10.1093/bib/bbp043] [PubMed:19955237].

[28] Peter Karp. Pathway Tools Data File Formats. `http://bioinformatics.ai.sri.com/ptools/flatfile-format.html`, 2016.

[29] Peter D. Karp, Richard Billington, Timothy A. Holland, Anamika Kothari, Markus Krummenacker, Daniel Weaver, Mario Latendresse, and Suzanne Paley. Computational Metabolomics Operations at BioCyc.org. *Metabolites*, 5(2):291–310, 2015.

[30] Peter D. Karp, Ingrid M. Keseler, Alexander Shearer, Mario Latendresse, Markus Krummenacker, Suzanne M. Paley, Ian Paulsen, Julio Collado-Vides, Socorro Gama-Castro, Martin Peralta-Gil, Alberto Santos-Zavaleta, Mónica I. Peñaloza-Spínola, César Bonavides-Martinez, and John Ingraham. Multidimensional annotation of the Escherichia coli K-12 genome. *Nucleic Acids Res*, 35(22):7577–7590, 2007.

[31] Peter D. Karp, Mario Latendresse, Suzanne M. Paley, Markus Krummenacker, Quang D. Ong, Richard Billington, Anamika Kothari, Daniel Weaver, Thomas Lee, Pallavi Subhraveti, Aaron Spaulding, Carol Fulcher, Ingrid M. Keseler, and Ron Caspi. Pathway Tools version 19.0 update: software for pathway/genome informatics and systems biology. *Brief Bioinform*, Oct 2015.

[32] Peter D. Karp, Suzanne M. Paley, Markus Krummenacker, Mario Latendresse, Joseph M. Dale, Thomas J. Lee, Pallavi Kaipa, Fred Gilham, Aaron Spaulding, Liviu Popescu, Tomer Altman, Ian Paulsen, Ingrid M. Keseler, and Ron Caspi. Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. *Brief Bioinform*, 11(1):40–79, Jan 2010.

[33] Karp, P. D. and Paley, S. and Romero, P. The Pathway Tools software. *Bioinformatics*, 18 Suppl 1:S225–32, 2002.

[34] Sahil Khanna and Pritish K. Tosh. A clinician's primer on the role of the microbiome in human health and disease. *Mayo Clin Proc*, 89(1):107–114, Jan 2014.

[35] Kiełbasa, Szymon M and Wan, Raymond and Sato, Kengo and Horton, Paul and Frith, Martin C. Adaptive seeds tame genomic sequence comparison. *Genome research*, 21(3):487–93, 2011.

[36] R. A. Koeth, Z. Wang, B. S. Levison, J. A. Buffa, E. Org, B. T. Sheehy, E. B. Britt, X. Fu, Y. Wu, L. Li, J. D. Smith, J. A. DiDonato, J. Chen, H. Li, G. D. Wu, J. D. Lewis, M. Warrier, J. M. Brown, R. M. Krauss, W. H. Tang, F. D. Bushman, A. J. Lusis, and S. L. Hazen. Intestinal microbiota metabolism of L-carnitine, a nutrient in red meat, promotes atherosclerosis. *Nat. Med.*, 19(5):576–585, May 2013. [PubMed Central:PMC3650111] [DOI:10.1038/nm.3145] [PubMed:23563705].

[37] Kishori M. Konwar, Niels W. Hanson, Antoine P. Pagé, and Steven J. Hallam. MetaPathways: a modular pipeline for constructing pathway/genome databases from environmental sequence information. *BMC Bioinformatics*, 14:202, 2013.

[38] Konwar, K. M. and Hanson, N. W. and Bhatia, M. P. and Kim, D. and Wu, S. and Hahn, A. S. and Morgan-Lang, C. and Cheung, H. K. and Hallam, S. J. MetaPathways v2.5: quantitative functional, taxonomic and usability improvements. *Bioinformatics*, pages 1–3, 2015.

[39] Konwar, K. M. and Hanson, N. W. and Page, A. P. and Hallam, S. J. MetaPathways: a modular pipeline for constructing pathway/genome databases from environmental sequence information. *BMC Bioinformatics*, 14:1–3, 2013.

[40] M. Latendresse, M. Krummenacker, M. Trupp, and P. D. Karp. Construction and completion of flux balance models from pathway databases. *Bioinformatics*, 28(3):388–396, Feb 2012. [PubMed Central:PMC3268246] [DOI:10.1093/bioinformatics/btr681] [PubMed:22262672].

[41] Mario Latendresse and Peter D. Karp. Web-based metabolic network visualization with a zooming user interface. *BMC Bioinformatics*, 12:176, 2011.

[42] Dongjae Kim, Aria S Hahn, Niels W Hanson, Kishori M Konwar and Steven J Hallam FAST: Fast Annotation with Synchronized Threads *IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology*, In Press, 2016.

[43] Thomas J Lee, Ian Paulsen, and Peter Karp. Annotation-based inference of transporter function. *Bioinformatics (Oxford, England)*, 24(13):i259–67, 2008.

[44] Lowe, T M and Eddy, S R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic acids research*, 25(5):955–64, 1997.

[45] Marcobal, A and Kashyap, P C and Nelson, T A and Aronov, P A and Donia, M S and Spormann, A and Fischbach, M A and Sonnenburg, J L. A metabolomic view of how the human gut microbiota impacts the host metabolome using humanized and gnotobiotic mice. *The ISME journal*, 7(10):1933–43, 2013.

[46] Victor M. Markowitz, I-Min A. Chen, Ken Chu, Ernest Szeto, Krishna Palaniappan, Manoj Pillay, Anna Ratner, Jinghua Huang, Ioanna Pagani, Susannah Tringe, Marcel Huntemann, Konstantinos Billis, Neha Varghese, Kristin Tennessen, Konstantinos Mavromatis, Amrita Pati, Natalia N. Ivanova, and Nikos C. Kyrpides. IMG/M 4 version of the integrated metagenome comparative analysis system. *Nucleic Acids Res*, 42(Database issue):D568–D573, Jan 2014.

[47] Katherine D. McMahon, Héctor García Martín, and Philip Hugenholtz. Integrating ecology into biotechnology. *Curr Opin Biotechnol*, 18(3):287–292, Jun 2007.

[48] Jamie A. Meadows and Matthew J. Wargo. Carnitine in bacterial physiology and metabolism. *Microbiology*, Mar 2015.

[49] Lindsay Murdock, Tangi Burke, Chelsea Coumoundouros, Doreen E. Culham, Charles E. Deutch, James Ellinger, Craig H. Kerr, Samantha M. Plater, Eric To, Geordie Wright, and Janet M. Wood. Analysis of strains lacking known osmolyte accumulation mechanisms reveals contributions of osmolytes and transporters to protection against abiotic stress. *Appl Environ Microbiol*, 80(17):5366–5378, Sep 2014.

[50] Shujiro Okuda, Takuji Yamada, Masami Hamajima, Masumi Itoh, Toshiaki Katayama, Peer Bork, Susumu Goto, and Minoru Kanehisa. KEGG Atlas mapping for global analysis of metabolic pathways. *Nucleic Acids Res*, 36(Web Server issue):W423–W426, Jul 2008.

[51] Jeffrey D. Orth, Ines Thiele, and Bernhard Ø. Palsson. What is flux balance analysis? *Nat Biotechnol*, 28(3):245–248, Mar 2010.

[52] Ross Overbeek, Tadhg Begley, Ralph M. Butler, Jomuna V. Choudhuri, Han-Yu Chuang, Matthew Cohoon, Valérie de Crécy-Lagard, Naryttza Diaz, Terry Disz, Robert Edwards, Michael Fonstein, Ed D. Frank, Svetlana Gerdes, Elizabeth M. Glass, Alexander Goesmann, Andrew Hanson, Dirk Iwata-Reuyl, Roy Jensen, Neema Jamshidi, Lutz Krause, Michael Kubal, Niels Larsen, Burkhard Linke, Alice C. McHardy, Folker Meyer, Heiko Neuweger, Gary Olsen, Robert Olson, Andrei Osterman, Vasiliy Portnoy, Gordon D. Pusch, Dmitry A. Rodionov, Christian Rückert, Jason Steiner, Rick Stevens, Ines Thiele, Olga Vassieva, Yuzhen Ye, Olga Zagnitko, and Veronika Vonstein. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res*, 33(17):5691–5702, 2005.

[53] Suzanne M. Paley and Peter D. Karp. Evaluation of computational metabolic-pathway predictions for *Helicobacter pylori*. *Bioinformatics*, 18(5):715–724, May 2002.

[54] Suzanne M. Paley and Peter D. Karp. The Pathway Tools Cellular Overview Diagram and Omics Viewer. *Nucleic Acids Res*, 34(13):3771–3778, 2006.

[55] J. Peterson, S. Garges, M. Giovanni, P. McInnes, L. Wang, J. A. Schloss, V. Bonazzi, J. E. McEwen, K. A. Wetterstrand, C. Deal, C. C. Baker, V. Di Francesco, T. K. Howcroft, R. W. Karp, R. D. Lunsford, C. R. Wellington, T. Belachew, M. Wright, C. Giblin, H. David, M. Mills, R. Salomon, C. Mullins, B. Akolkar, L. Begg, C. Davis, L. Grandison, M. Humble, J. Khalsa, A. R. Little, H. Peavy, C. Pontzer, M. Portnoy, M. H. Sayre, P. Starke-Reed, S. Zakhari, J. Read, B. Watson, and M. Guyer. The NIH Human Microbiome Project. *Genome Res.*, 19(12):2317–2323, Dec 2009. [PubMed Central:PMC2792171] [DOI:10.1101/gr.096651.109] [PubMed:19819907].

[56] Pruesse, E. and Quast, C. and Knittel, K. and Fuchs, B. M. and Ludwig, W. G. and Peplies, J. and Glockner, F. O. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research*, 35(21):7188–7196, 2007.

[57] Junjie Qin, Ruiqiang Li, Jeroen Raes, Manimozhiyan Arumugam, Kristoffer Solvsten Burgdorf, Chaysavanh Manichanh, Trine Nielsen, Nicolas Pons, Florence Levenez, Takuji Yamada, Daniel R. Mende, Junhua Li, Junming Xu, Shaochuan Li, Dongfang Li, Jianjun Cao, Bo Wang, Huiqing Liang, Huisong Zheng, Yinlong Xie, Julien Tap, Patricia Lepage, Marcelo Bertalan, Jean-Michel Batto, Torben Hansen, Denis Le Paslier, Allan Linneberg, H Bjørn Nielsen, Eric Pelletier, Pierre Renault, Thomas Sicheritz-Ponten, Keith Turner, Hongmei Zhu, Chang Yu, Shengting Li, Min Jian, Yan Zhou, Yingrui Li, Xiuqing Zhang, Songgang Li, Nan Qin, Huanming Yang, Jian Wang, Søren Brunak, Joel Doré, Francisco Guarner, Karsten Kristiansen, Oluf Pedersen, Julian Parkhill, Jean Weissenbach, MetaHIT Consortium,

Peer Bork, S Dusko Ehrlich, and Jun Wang. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464(7285):59–65, Mar 2010.

[58] Junjie Qin, Yingrui Li, Zhiming Cai, Shenghui Li, Jianfeng Zhu, Fan Zhang, Suisha Liang, Wenwei Zhang, Yuanlin Guan, Dongqian Shen, Yangqing Peng, Dongya Zhang, Zhuye Jie, Wenxian Wu, Youwen Qin, Wenbin Xue, Junhua Li, Lingchuan Han, Donghui Lu, Peixian Wu, Yali Dai, Xiaojuan Sun, Zesong Li, Aifa Tang, Shilong Zhong, Xiaoping Li, Weineng Chen, Ran Xu, Mingbang Wang, Qiang Feng, Meihua Gong, Jing Yu, Yanyan Zhang, Ming Zhang, Torben Hansen, Gaston Sanchez, Jeroen Raes, Gwen Falony, Shujiro Okuda, Mathieu Almeida, Emmanuelle LeChatelier, Pierre Renault, Nicolas Pons, Jean-Michel Batto, Zhaoxi Zhang, Hua Chen, Ruifu Yang, Weimou Zheng, Songgang Li, Huanming Yang, Jian Wang, S Dusko Ehrlich, Rasmus Nielsen, Oluf Pedersen, Karsten Kristiansen, and Jun Wang. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, 490(7418):55–60, Oct 2012.

[59] Rasko, D. A. and Myers, G. S. A. and Ravel, J. Visualization of comparative genomic analyses by BLAST score ratio. *BMC Bioinformatics*, 6:7188–7196, 2005.

[60] David A. Relman. The human microbiome: ecosystem resilience and health. *Nutr Rev*, 70 Suppl 1:S2–S9, Aug 2012.

[61] B. Rost. Twilight zone of protein sequence alignments. *Protein Eng.*, 12(2):85–94, Feb 1999. [PubMed:10195279].

[62] Eric W. Sayers, Tanya Barrett, Dennis A. Benson, Stephen H. Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M. Church, Michael DiCuccio, Ron Edgar, Scott Federhen, Michael Feolo, Lewis Y. Geer, Wolfgang Helmberg, Yuri Kapustin, David Landsman, David J. Lipman, Thomas L. Madden, Donna R. Maglott, Vadim Miller, Ilene Mizrachi, James Ostell, Kim D. Pruitt, Gregory D. Schuler, Edwin Sequeira, Stephen T. Sherry, Martin Shumway, Karl Sirotkin, Alexandre Souvorov, Grigory Starchenko, Tatiana A. Tatusova, Lukas Wagner, Eugene Yaschenko, and Jian Ye. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, 37(Database issue):D5–15, Jan 2009.

[63] R. L. Tatusov, M. Y. Galperin, D. A. Natale, and E. V. Koonin. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res*, 28(1):33–36, Jan 2000.

[64] Taylor Tony and Schug Kevin. Qualitative Aspects of Electrospray Ionization, Fragmentation and Adduct Formation. `http://www.chromacademy.com/Electrospray-Ionization-ESI-for-LC-MS.html`, 2011.

[65] W. L. Trimble, K. P. Keegan, M. D'Souza, A. Wilke, J. Wilkening, J. A. Gilbert, and F. Meyer. Short-read reading-frame predictors are not created

equal: sequence error causes loss of signal. *BMC Bioinformatics*, 13(1):183, 2012.

[66] Susannah G. Tringe and Philip Hugenholtz. A renaissance for the pioneering 16S rRNA gene. *Curr Opin Microbiol*, 11(5):442–446, Oct 2008.

[67] Richard Allen White, Ian M Power, Gregory M Dipple, Gordon Southam, and Curtis A Suttle. Metagenomic analysis reveals that modern microbialites and polar microbial mats have similar taxonomic and functional potential. *Frontiers in microbiology*, 6:966, 2015.

[68] Andreas Wilke, Elizabeth M. Glass, Daniela Bartels, Jared Bischof, Daniel Braithwaite, Mark D'Souza, Wolfgang Gerlach, Travis Harrison, Kevin Keegan, Hunter Matthews, Renzo Kottmann, Tobias Paczian, Wei Tang, William L. Trimble, Pelin Yilmaz, Jared Wilkening, Narayan Desai, and Folker Meyer. A metagenomics portal for a democratized sequencing world. *Methods Enzymol*, 531:487–523, 2013.

[69] Michael Wilson. *Bacteriology of humans : an ecological perspective*. Blackwell Pub., Malden, MA, 2008.

[70] Jody J Wright, Keith Mewis, Niels W Hanson, Kishori M Konwar, Kendra R Maas, and Steven J Hallam. Genomic properties of Marine Group A bacteria indicate a role in the marine sulfur cycle. *The ISME Journal*, 8(2):455–68, 2014.

Data Citations

Hahn, Aria S; Altman, Tomer; Konwar, Kishori M; Hanson, Niels W; Kim, Dongjae; Relman, David A; Dill, David L; Hallam, Steven J (2016): GutCyc. figshare. https://dx.doi.org/10.6084/m9.figshare.c.3283562. Retrieved: 16:04, Jul 12, 2016 (GMT).