SCIENTIFIC DATA

OPEN

Data Descriptor: A geographically-diverse collection of 418 human gut microbiome pathway genome databases

Received: 3 August 2016
Accepted: 10 February 2017
Published: 11 April 2017

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]

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 (ePGDBs) 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.

Design Type(s)	data integration objective • database creation objective
Measurement Type(s)	metagenomics analysis
Technology Type(s)	digital curation
Factor Type(s)	Clinical Treatment
Sample Characteristic(s)	

[1]Department of Microbiology and Immunology, University of British Columbia, Vancouver, British Columbia V6T 1Z3, Canada. [2]Koonkie Inc., Menlo Park, California 94025, USA. [3]Biomedical Informatics, Stanford University School of Medicine, Stanford, California 94305, USA. [4]Whole Biome, Inc., 953 Indiana Street, San Francisco, California 94107, USA. [5]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [6]Department of Computer Science, University of British Columbia, Vancouver, British Columbia V6T 1Z3, Canada. [7]Department of Microbiology and Immunology, Stanford University School of Medicine, 299 Campus Drive, Stanford, California 94305, USA. [8]Department of Medicine, Stanford University School of Medicine, Stanford, California 94305, USA. [9]Veterans Affairs Palo Alto Health Care System, Palo Alto, California 94304, USA. [10]Department of Computer Science, Stanford University, Stanford, California 94305, USA. [11]Ecosystem Services, Commercialization and Entrepreneurship (ECOSCOPE), University of British Columbia, Vancouver, British Columbia V6T 1Z3, Canada. *These authors contributed equally to this work. Correspondence and requests for materials should be addressed to S.J.H. (email: shallam@mail.ubc.ca).

Background & Summary

The myriad collections of microorganisms found on and in the human body are known as the human microbiome[1]. Changes in microbiome structure and function have been implicated in numerous disease states including inflammatory bowel disease, cancer, and even cardiovascular disease[2,3]. 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[4], 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)[5], Beijing Genomics Institute (BGI)[6], and Metagenomes of the Human Intestinal Tract project (MetaHIT)[7] 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[8], Metagenome Rapid Annotation using Subsystem Technology (MG-RAST)[9], SmashCommunity[10] and HUMAnN[11]. 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)[12]. Although extant pipelines often provide links to KEGG module and pathways maps[12] (using KEGG ontology (KO) or pathway identifiers) that can be visualized with coverage or gene count information using programs like KEGG Atlas[13], 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[11].

In responding to the deficiencies of existing tools, we recently developed a modular annotation and analysis pipeline enabling reproducible research[14] called MetaPathways, that guides construction of environmental Pathway Genome Databases (ePGDBs) from environmental sequence information[15] using Pathway Tools[16] and MetaCyc[17–19]. Pathway Tools is a production-quality software environment developed at SRI International 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[20]. 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 PGDBs 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[5], the MetaHIT inflammatory bowel disease study[7], and the BGI diabetes study[6]. Relevant pipeline modules are summarized in Fig. 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[21].

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, $n = 148$)[5], a MetaHIT (European inflammatory bowel disease subjects, $n = 125$)[22], and a BGI (Chinese type 2 diabetes subjects, $n = 145$) study[6]. See Supplementary Tables 1 and 2 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-based functional and taxonomic annotation, (3) analyses consisting of tRNA and lowest common ancestor (LCA)[23]

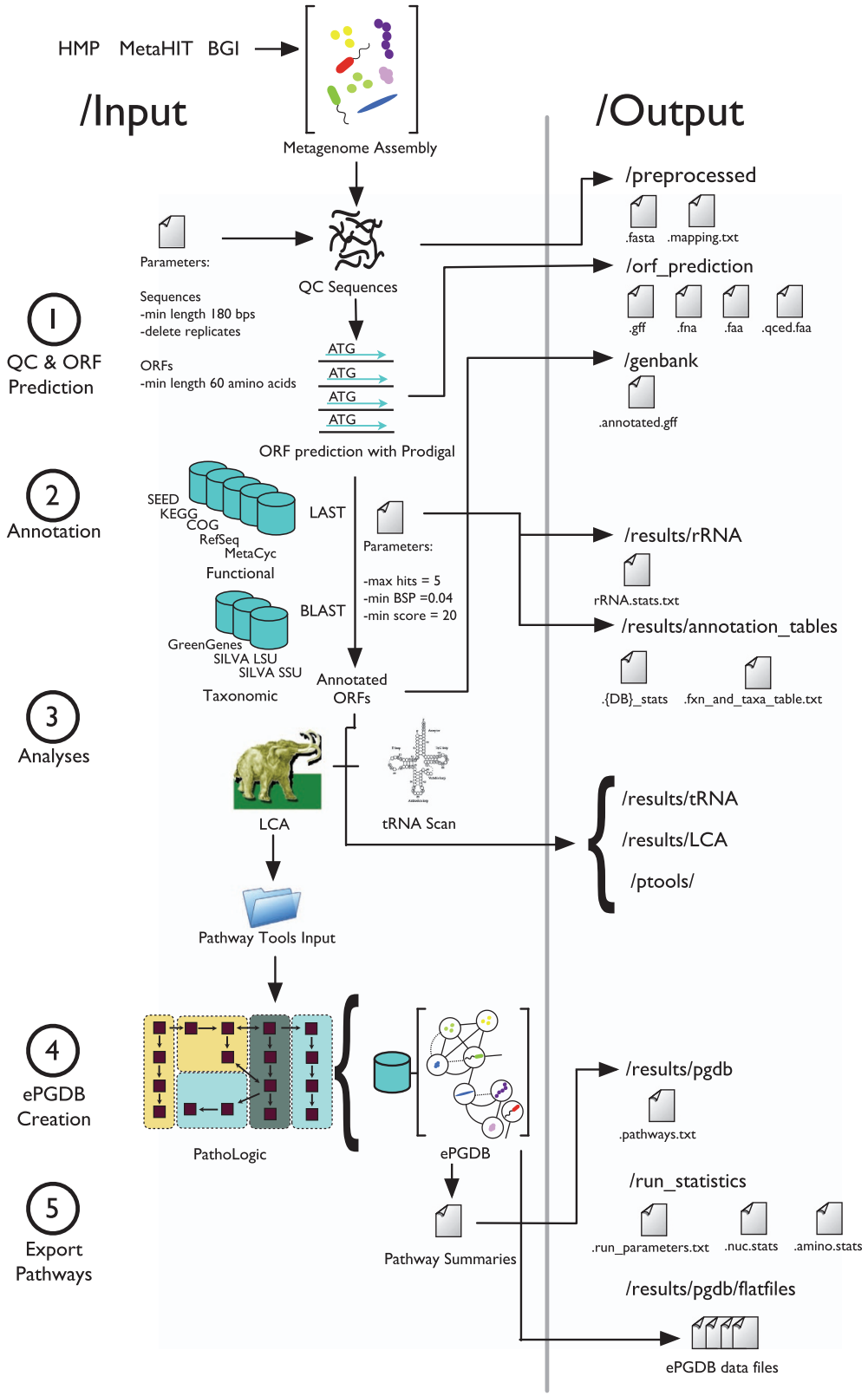

Figure 1. The GutCyc pipeline diagram. The MetaPathways pipeline consists of five modular stages including (1) Quality control (QC) and open reading frame (ORF) prediction (2) Functional and taxonomic annotation, (3) Analysis (4) ePGDB construction, and (5) Pathway export. Inputs and programs are depicted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

identification, (4) construction of ePGDBs using Pathway Tools and, finally, (5) pathway export[24,25] (see Fig. 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[26].

ORF prediction. Sequences passing quality control were scanned for ORFs using MetaProdigal[27], a robust ORF prediction tool for microbial metagenomes considered to be among the most accurate ORF predictors[28]. Resulting ORF sequences were translated to amino acid sequences using NCBI genetic code Table 11 for bacteria, archaea, and plant plastids[29]. 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[30].

Functional annotation. The quality-controlled amino acid sequences were queried against a panel of functionally-annotated protein reference databases used in the validation of the MetaPathways pipeline[31] : KEGG[12] (downloaded 2011-06-18), COG[32] (downloaded 2013-12-27), MetaCyc[19] (downloaded 2011-07-03), RefSeq[33] (downloaded 2014-01-18), and SEED[34] (downloaded 2014-01-30). Protein sequence similarity searches were performed using the program FAST[35] 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[30]; and blast-score ratio (BSR) greater than 0.4 (ref. 36)). The choice of parameter thresholds were selected to maximize annotation accuracy, and were guided based on parameter choices used in previous studies[31,37,38].

Taxonomic annotation. Quality-controlled contigs were also searched against the SILVA[39] (version 115) and Greengenes[40] (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 rRNA gene 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[23] for taxonomic binning. In brief, the LCA in the NCBI Taxonomy Database[33] 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 NCBI Taxonomy Database. The RefSeq taxonomic names often contain multiple synonyms or alternative spellings. Therefore, names that conform to the NCBI Taxonomy Database were selected in preference over unknown synonyms.

tRNA scan. MetaPathways uses tRNAscan-SE version 1.4 (ref. 41) 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[42,43] 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) 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 Fig. 1). The MetaPathways 2.5 graphical user interface (GUI) enables interactive exploration, visualization, and searches of individual sample results along with comparative queries of multiple samples, *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 WestGrid's Bugaboo cluster part of Compute Canada's national platform of Advanced Research Computing resources. `https://www.westgrid.ca/support/systems/bugaboo`. The Hallam lab computers have a configuration profile of 2 × 2.4 GHz Quad-Core Intel Xeon processors with 64 GB 1,066 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 h 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 (`github.com/hallamlab/metapathways2`), licensed under the GNU General Public License, version 3), along with a companion tutorial (`github.com/hallamlab/mp_tutorial/`) 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 (`www.biocyc.org/download-bundle.shtml`). 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[31], and an additional refinement step of running the Transport Inference Parser[44] to predict transport reactions (i.e., `-tip`). FAST is freely available under the GNU General Public License, version 3 on our GitHub page (`github.com/hallamlab/FAST`).

Data Records

A list of each sample, its provenance, location and relevant data processing steps can be found in Supplementary Table 1. All records are available at the GutCyc project website (`www.gutcyc.org`), and at Figshare as described in (Data Citation 1). Each sample's data records are contained within a single directory. Within this directory, sub-directories and files are located as depicted in Fig. 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 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[45] 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`),

	Min	1st quartile	Median	3rd 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.' are functional annotations. 'Trans. Reactions' are transport reactions. 'Compounds' are small molecule metabolites. 'Base Pathways' include all pathways except complex pathways known as Super-Pathways.

(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 small subunit ribosomal RNA (SSU or 16S rRNA) gene is a standard marker gene used for measuring taxonomic diversity[46]. 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[47]) 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[5], a MetaHIT study[22], and a BGI study[6]. The technical validation of third-party software used in MetaPathways may be found in the corresponding publications for MetaProdigal[27], BLAST[48], and tRNAscan-SE[41]. GutCyc functional sequence similarity was computed using FAST, an aligner based on a version of LAST[49], with multi-threading performance improvements and new support for generating BLAST-like E-values, with significant correlation with BLAST output (correlation of the *log*(E-value) outputs of BLAST and LAST: $R^2 = 0.887$, $P < 0.01$)[24]. The protocols undertaken in the MetaCyc project for the ongoing manual curation of new metabolic pathways, and its subsequent implications for accurate pathway prediciton, may be found in the following MetaCyc publications[17–19,50].

Validation of the overall MetaPathways pipeline may be found in previously published reports[31,51] 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[52], which has an accuracy of 91% as evaluated using organism pathway databases with high levels of manual curation[43]). The performance of the Pathway Tools PathoLogic module has also been evaluated using datasets with different complexity and coding potential, including simulated metagenomes, a symbiotic system, and the Hawaii Ocean Time-series[31]. The authors provide detailed information about the effects of read length, coverage and sample diversity on pathway recovery but found that performance specificity was high (>85%) using all three datasets. The authors also provide a list of `prediction hazards' such as the identification of dissimilatory nitrate reduction pathways of which the user should be aware and conclude that despite being imperfect, Pathway Tools provides a powerful means with which to predict metabolic interactions[31].

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[31]. Annotations from each of the protein reference databases can also be explored individually using MetaPathways. In addition, a file for each sample, located at `sample/results/annotation_table/sample.2.txt` provides a detailed overview of the annotation for each ORF in each database as well as score reflecting the confidence of the alignment and annotation. 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[53]. The MetaFlux[54] module of Pathway Tools for performing flux balance analysis (FBA)[55] 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[56], and provides a computational substrate for systems biology approaches to engineering the gut microbiome[57]. 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.*[58], they demonstrated a causal connection between the intestinal gut microbiota's metabolism of red meat and the promotion of atherosclerosis. 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 an arbitrarily selected ePGDB from the GutCyc Collection, SRS015217Cyc, for the biochemical reaction path from *L*-carnitine to TMA, which is not provided explicitly by Koeth *et al.*[58] Utilizing the Pathway Tools Metabolic Route Search feature, we found an optimal path between *L*-carnitine to TMA for this ePGDB, using the MetaCyc *carnitine degradation II* pathway (PWY-3,602, 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.

The metabolic route identified may also help generate mechanistic hypotheses from microbiome study observations. *L*-carnitine and glycine betaine have known transporter families that facilitate their movement across the cell membrane[59], as do TMA and TMAO[60], and thus the metabolic route in this ePGDB may be a distributed pathway[31]. 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[61]. 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[62]. As a demonstration of this capability, we analyzed mass-spectrometry data from a metabolomic study of humanized mice microbiomes[63]. The dataset contained 867 unique masses, of which 453 masses were identified using MetaCyc by performing standard adduct monoisotopic mass manipulations[64], followed by monoisotopic mass search using Pathway Tools. We mapped the identified compounds on the Cellular Overview[65] of an arbitrarily-selected ePGDB from the GutCyc Collection for illustrative purposes, as seen in Fig. 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[62], 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.

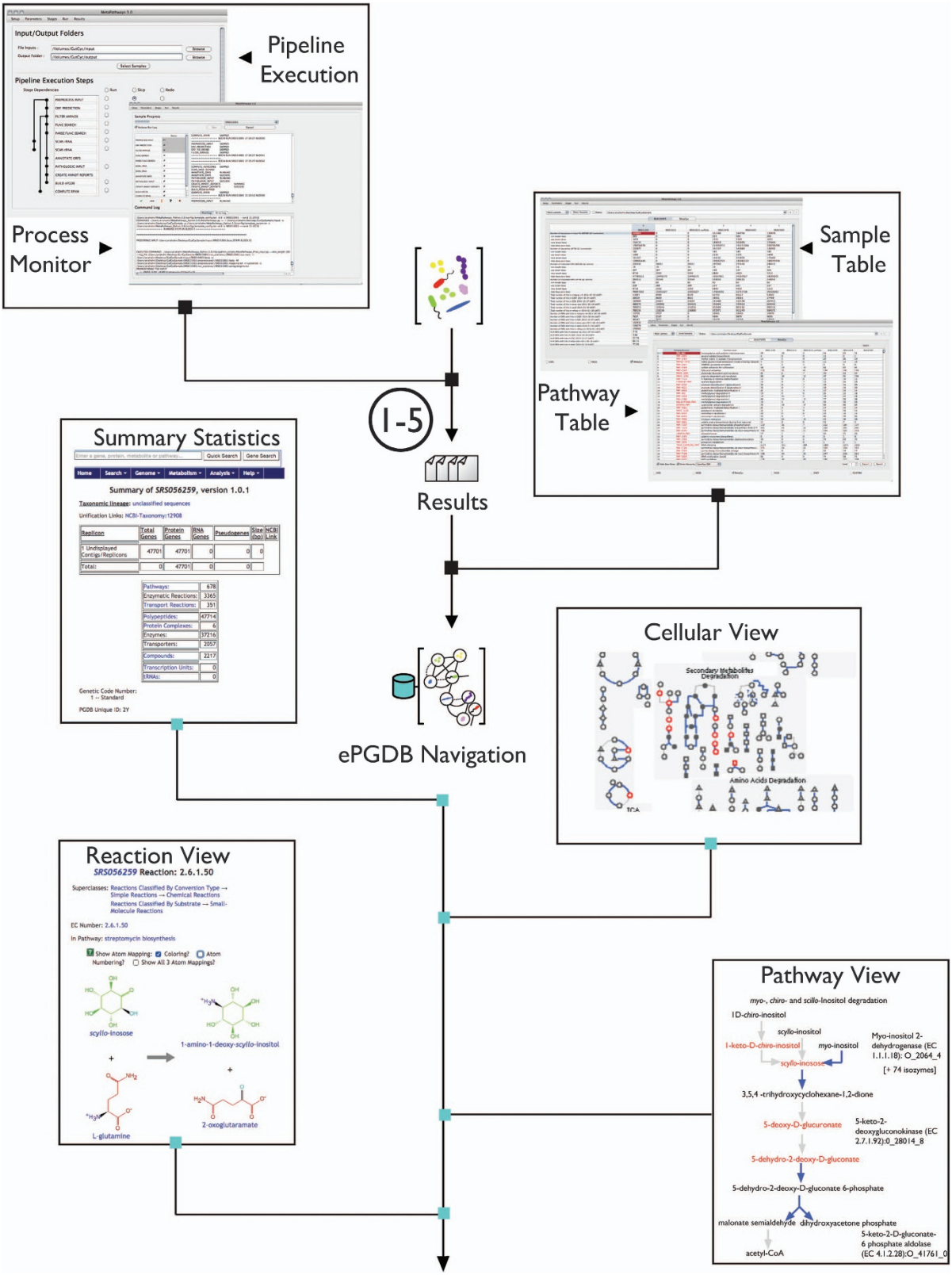

Figure 2. GutCyc ePGDB use cases. In the upper left and upper right insets, a GutCyc ePGDB is opened in MetaPathwaysIn the upper left, we display the Pipeline Execution step, and the Process Monitor interfaces. In the upper right, we display the Summary Table (with exportable sample statistics), and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

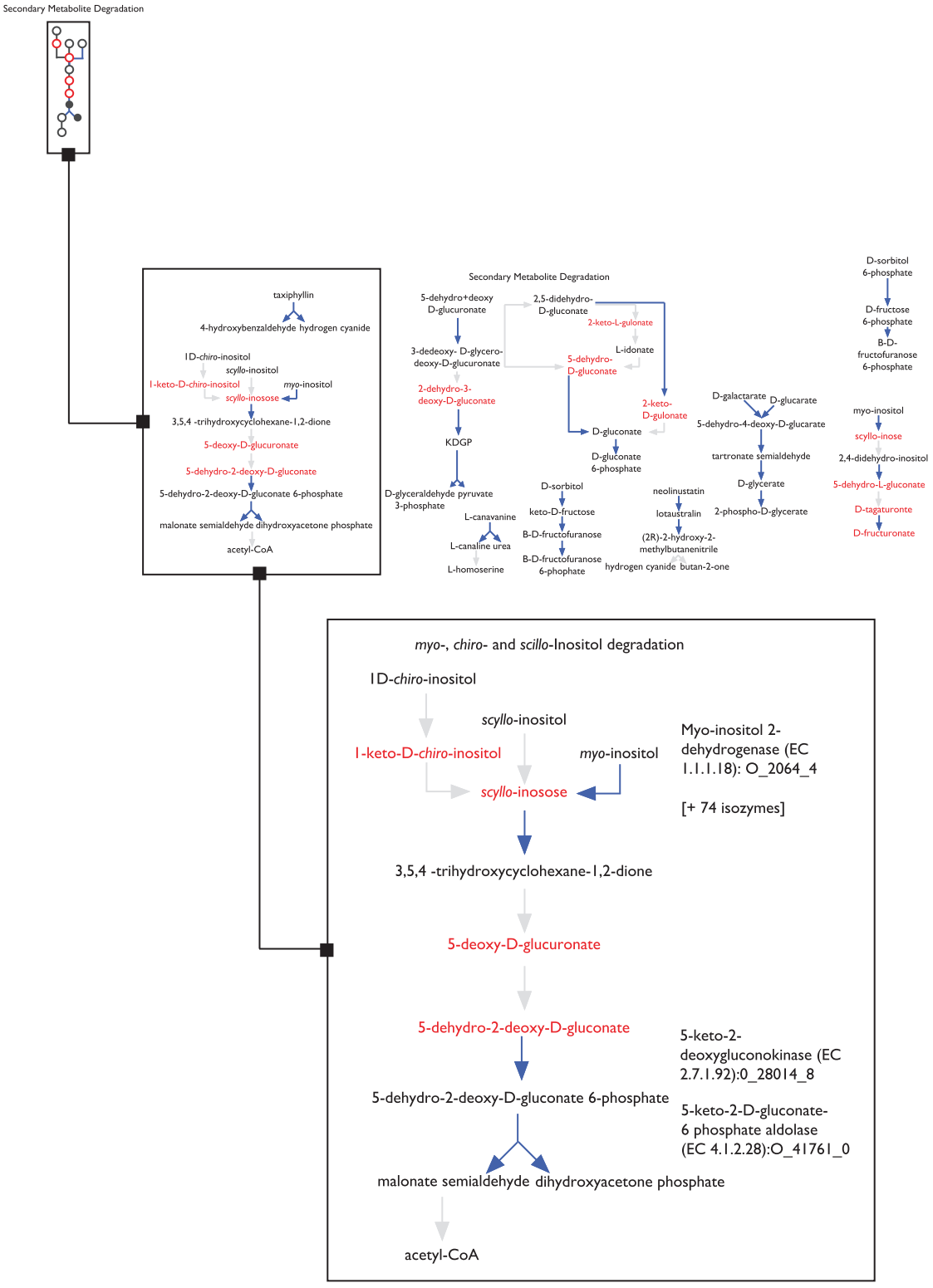

Figure 3. The Cellular Overview for the SRS056259Cyc ePGDB, at three different zoom levels. Compounds are highlighted in red if identified from a mass spectrometry analysis of the gut microbiome[63], and otherwise appear in grey. Reactions with enzyme data in SRS056259Cyc are drawn in blue. The top left 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.

References

1. Relman, D. A. The human microbiome: ecosystem resilience and health. *Nutr Rev* **70**(Suppl 1): S2–S9 (2012).
2. Khanna, S. & Tosh, P. K. A clinician's primer on the role of the microbiome in human health and disease. *Mayo Clin Proc* **89,** 107–114 (2014).
3. Bultman, S. J. Emerging roles of the microbiome in cancer. *Carcinogenesis* **35,** 249–255 (2014).
4. Wilson, M. *Bacteriology of humans: an ecological perspective* (Blackwell Pub., 2008).
5. Peterson, J. *et al.* The NIH Human Microbiome Project. *Genome Res.* **19,** 2317–2323 (2009).
6. Qin, J. *et al.* A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490,** 55–60 (2012).
7. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464,** 59–65 (2010).
8. Markowitz, V. M. *et al.* IMG/M 4 version of the integrated metagenome comparative analysis system. *Nucleic Acids Res* **42,** D568–D573 (2014).
9. Wilke, A. *et al.* A metagenomics portal for a democratized sequencing world. *Methods Enzymol* **531,** 487–523 (2013).
10. Arumugam, M., Harrington, E. D., Foerstner, K. U., Raes, J. & Bork, P. SmashCommunity: a metagenomic annotation and analysis tool. *Bioinformatics* **26,** 2977–2978 (2010).
11. Abubucker, S. *et al.* Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput Biol* **8,** e1002358 (2012).
12. Kanehisa, M. *et al.* Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res* **42,** D199–D205 (2014).
13. Okuda, S. *et al.* KEGG Atlas mapping for global analysis of metabolic pathways. *Nucleic Acids Res* **36,** W423–W426 (2008).
14. Callahan, B., Proctor, D., Relman, D., Fukuyama, J. & Holmes, S. Reproducible research workflow in R for the analysis of personalized human microbiome data. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* **21,** 183–194 (2016).
15. Konwar, K. M., Hanson, N. W., Pagé, A. P. & Hallam, S. J. MetaPathways: a modular pipeline for constructing pathway/genome databases from environmental sequence information. *BMC Bioinformatics* **14,** 202 (2013).
16. Karp, P. D. *et al.* Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. *Brief. Bioinformatics* **11,** 40–79 (2010).
17. Karp, P. D. *et al.* Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. *Brief Bioinform* **11,** 40–79 (2010).
18. Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res* **42,** D459–D471 (2014).
19. Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res* **44,** D471–D480 (2016).
20. Altman, T., Travers, M., Kothari, A., Caspi, R. & Karp, P. D. A systematic comparison of the MetaCyc and KEGG pathway databases. *BMC Bioinformatics* **14,** 112 (2013).
21. Karp, P. D. *et al.* Multidimensional annotation of the Escherichia coli K-12 genome. *Nucleic Acids Res* **35,** 7577–7590 (2007).
22. Dusko Ehrlich, S. & MetaHIT consortium. Metagenomics of the intestinal microbiota: potential applications. *Gastroenterol Clin Biol* **34**(Suppl 1): S23–S28 (2010).
23. Huson, D. H. & Weber, N. Microbial community analysis using MEGAN. *Methods Enzymol* **531,** 465–485 (2013).
24. Konwar, K. M. *et al.* MetaPathways v2.5: quantitative functional, taxonomic and usability improvements. *Bioinformatics* **31,** 3345–3347 (2015).
25. Karp, P. D., Paley, S. & Romero, P. The Pathway Tools software. *Bioinformatics* **18**(Suppl 1): S225–S232 (2002).
26. Konwar, K. M., Hanson, N. W., Page, A. P. & Hallam, S. J. MetaPathways: a modular pipeline for constructing pathway/genome databases from environmental sequence information. *BMC Bioinformatics* **14,** 1–3 (2013).
27. Hyatt, D., LoCascio, P. F., Hauser, L. J. & Uberbacher, E. C. Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics* **28,** 2223–2230 (2012).
28. Trimble, W. L. *et al.* Short-read reading-frame predictors are not created equal: sequence error causes loss of signal. *BMC Bioinformatics* **13,** 183 (2012).
29. Andrzej, E. & Jim, O. The Bacterial, Archaeal and Plant Plastid Code. Available at www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi#SG11 (2013).
30. Rost, B. Twilight zone of protein sequence alignments. *Protein Eng.* **12,** 85–94 (1999).
31. Hanson, N. W. *et al.* Metabolic pathways for the whole community. *BMC Genomics* **15,** 619 (2014).
32. Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* **28,** 33–36 (2000).
33. Sayers, E. W. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **37,** D5–15 (2009).
34. Overbeek, R. *et al.* The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res* **33,** 5691–5702 (2005).
35. Kim, D., Hahn, A. S., Hanson, N. W., Konwar, K. M. & Hallam, S. J. In *2016 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology*, 1–8 (IEEE, 2016).
36. Rasko, D. A., Myers, G. S. A. & Ravel, J. Visualization of comparative genomic analyses by BLAST score ratio. *BMC Bioinformatics* **6,** 7188–7196 (2005).
37. Wright, J. J. *et al.* Genomic properties of Marine Group A bacteria indicate a role in the marine sulfur cycle. *The ISME Journal* **8,** 455–468 (2014).
38. White, R. A., Power, I. M., Dipple, G. M., Southam, G. & Suttle, C. A. Metagenomic analysis reveals that modern microbialites and polar microbial mats have similar taxonomic and functional potential. *Frontiers in Microbiology* **6,** 966 (2015).
39. Pruesse, E. *et al.* SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* **35,** 7188–7196 (2007).
40. DeSantis, T. Z. *et al.* Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology* **72,** 5069–5072 (2006).
41. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25,** 955–964 (1997).
42. Green, M. L. & Karp, P. D. A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Bioinformatics* **5,** 76 (2004).
43. Dale, J. M., Popescu, L. & Karp, P. D. Machine learning methods for metabolic pathway prediction. *BMC Bioinformatics* **11,** 15 (2010).
44. Lee, T. J., Paulsen, I. & Karp, P. Annotation-based inference of transporter function. *Bioinformatics (Oxford, England)* **24,** i259–i267 (2008).
45. Eilbeck, K. *et al.* The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biology* **6,** R44 (2005).
46. Tringe, S. G. & Hugenholtz, P. A renaissance for the pioneering 16S rRNA gene. *Curr Opin Microbiol* **11,** 442–446 (2008).
47. Karp, P. Pathway Tools Data File Formats. Available at http://bioinformatics.ai.sri.com/ptools/flatfile-format.html (2016).

48. Boratyn, G. M. *et al.* BLAST: a more efficient report with usability improvements. *Nucleic Acids Res* **41,** W29–W33 (2013).
49. Kiełbasa, S. M., Wan, R., Sato, K., Horton, P. & Frith, M. C. Adaptive seeds tame genomic sequence comparison. *Genome research* **21,** 487–493 (2011).
50. Caspi, R., Dreher, K. & Karp, P. D. The challenge of constructing, classifying, and representing metabolic pathways. *FEMS Microbiology Letters* **345,** 85–93 (2013).
51. Hanson, N. W., Konwar, K. M., Wu, S.-J. & Hallam, S. J. 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* (2014).
52. Paley, S. M. & Karp, P. D. Evaluation of computational metabolic-pathway predictions for Helicobacter pylori. *Bioinformatics* **18,** 715–724 (2002).
53. Karp, P. D. *et al.* Pathway tools version 19.0 update: software for pathway/genome informatics and systems biology. *Brief Bioinform* **17,** 877–890 (2015).
54. Latendresse, M., Krummenacker, M., Trupp, M. & Karp, P. D. Construction and completion of flux balance models from pathway databases. *Bioinformatics* **28,** 388–396 (2012).
55. Orth, J. D., Thiele, I. & Palsson, B. Ø. What is flux balance analysis? *Nat Biotechnol* **28,** 245–248 (2010).
56. Haiser, H. J. & Turnbaugh, P. J. Developing a metagenomic view of xenobiotic metabolism. *Pharmacological Research* **69,** 21–31 (2013).
57. McMahon, K. D., Garca Martn, H. & Hugenholtz, P. Integrating ecology into biotechnology. *Curr Opin Biotechnol* **18,** 287–292 (2007).
58. Koeth, R. A. *et al.* Intestinal microbiota metabolism of L-carnitine, a nutrient in red meat, promotes atherosclerosis. *Nat. Med* **19,** 576–585 (2013).
59. Meadows, J. A. & Wargo, M. J. Carnitine in bacterial physiology and metabolism. *Microbiology* **161,** 1161–1174 (2015).
60. Murdock, L. *et al.* Analysis of strains lacking known osmolyte accumulation mechanisms reveals contributions of osmolytes and transporters to protection against abiotic stress. *Appl Environ Microbiol* **80,** 5366–5378 (2014).
61. Paley, S. M. & Karp, P. D. The Pathway Tools cellular overview diagram and Omics Viewer. *Nucleic Acids Res* **34,** 3771–3778 (2006).
62. Karp, P. D. *et al.* Computational Metabolomics Operations at BioCyc.org. *Metabolites* **5,** 291–310 (2015).
63. Marcobal, A. *et al.* A metabolomic view of how the human gut microbiota impacts the host metabolome using humanized and gnotobiotic mice. *The ISME Journal* **7,** 1933–1943 (2013).
64. Tony, T. & Kevin, S. Qualitative Aspects of Electrospray Ionization, Fragmentation and Adduct Formation. Available at http://www.chromacademy.com/Electrospray-Ionization-ESI-for-LC-MS.html (2011).
65. Latendresse, M. & Karp, P. D. Web-based metabolic network visualization with a zooming user interface. *BMC Bioinformatics* **12,** 176 (2011).

Data Citation

1. Hahn, A. S. *et al. Figshare* https://dx.doi.org/10.6084/m9.figshare.c.3283562 (2016).

Acknowledgements

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 S.J.H. A.S.H. was supported by the Alexander Graham Bell Canada Graduate Scholarships-Doctoral Program (CGS D) administered by NSERC. K.M.K. was supported by the Tula Foundation funded Centre for Microbial Diversity and Evolution (CMDE) at UBC. N.W.H. was supported by a four year doctoral fellowship (4YF) administered through the UBC Faculty of Graduate and Postdoctoral Studies. T.A. 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). T.A. and D.L.D. were partially supported by a King Abdullah University of Science and Technology (KAUST) research grant under the KAUST Stanford Academic Excellence Alliance program. D.A.R. 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

T.A. and S.J.H. 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. N.W.H., K.M.K., A.S.H. and D.K. developed the MetaPathways software pipeline with direction from S.J.H. and assistance from T.A. and others at SRI International. A.S.H. and K.M.K. compiled the microbiome sequence datasets, constructed GutCyc ePGDBs and created figures for the manuscript. T.A. generated validation datasets and drafted an early version of the manuscript with A.S.H. and S.J.H. D.K. developed the GutCyc website. All authors contributed to the final preparation of the manuscript. S.J.H., D.L.D. and D.A.R. supervised the project. All authors reviewed and approved the final manuscript.

Additional Information

Supplementary Information accompanies this paper at http://www.nature.com/sdata

Competing interests: Authors A.H., K.K., and S.J.H. 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. The remaining authors declare no competing financial interests.

How to cite this article: Hahn, A. S. *et al.* A geographically-diverse collection of 418 human gut microbiome pathway genome databases. *Sci. Data* 4:170035 doi: 10.1038/sdata.2017.35 (2017).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0

Metadata associated with this Data Descriptor is available at http://www.nature.com/sdata/ and is released under the CC0 waiver to maximize reuse.

© The Author(s) 2017