Physiological niche informs evolution of metabolic function and corresponding drug targets of pathobionts

Emma M. Glass[1], Lillian R. Dillard[2], Andrew S. Warren[3], Jason A. Papin[1,2,4*]

[1] Department of Biomedical Engineering
[2] Department of Biochemistry and Molecular Genetics
[3] Biocomplexity Institute and Initiative
[4] Division of Infectious Diseases & International Health, Department of Medicine
University of Virginia, Charlottesville, Virginia, USA
[*] Corresponding Author

ABSTRACT

Pathogens pose a major risk to human health globally, causing 44% of deaths in low-resource countries. Currently, there are over 500 known bacterial pathobionts, covering a wide range of functional capabilities. Some well-known pathobionts are well characterized computationally and experimentally. However, to gain a deeper understanding of how pathobionts are evolutionarily related to the principles that govern their different functions and ultimately identify possible targeted antimicrobials, we must consider both well-known and lesser-known pathobionts. Here, we developed a database of genome-scale metabolic network reconstructions (GENREs) called PATHGENN, which contains 914 models of pathobiont metabolism to address these questions related to functional metabolic evolution and adaptation. We determined the metabolic phenotypes across all known pathobionts and the role of isolate environment in functional metabolic adaptation. We also predicted novel antimicrobial targets for bacteria specific to their physiological niche. Understanding the functional metabolic similarities between pathobionts is the first step to ultimately developing a precision medicine framework for addressing all infections.

INTRODUCTION

Bacterial pathogens pose a major risk to human health. Globally, pathogens are responsible for 16% of all deaths, and responsible for 44% of deaths in low-resource countries[1]. Financially, global economic losses from pathogenic disease outbreaks amount to tens of billions of dollars in the past 10 years[2]. In recent years, there has been an increase in infectious disease emergence attributed to urbanization, globalization, climate change, population growth, and human/animal interaction [3]. Currently, there are over 500 known human bacterial pathobionts[4]. Pathobionts are microorganisms that have the capacity to be pathogenic[5] and range across many taxonomic classes and genera. Therefore, there exists a wide range in metabolic function, phylogeny, and infection niches (e.g., stomach, wound, lung) across pathobionts.

Due to their imminent danger to human health, some pathobiont species have been well characterized experimentally and computationally [6–8]. However, to gain a deeper understanding of how pathobionts are evolutionarily related and the principles that govern their differential functions and ultimately identify novel targeted antimicrobial therapies, we need to consider both well-characterized and poorly-characterized pathobionts. We can leverage ‘omics approaches to understand the relationship between pathobionts and their physiological environment to shed light on functional metabolic differences between species. A better characterization of governing principles of pathobiont function could enable the development of new approaches to target pathobionts through novel therapies or drug repurposing. Additionally, using antimicrobial therapies to target environment-specific essential genes rather than organism-specific essential genes could reduce the harmful effects of broad-spectrum antimicrobials[9]

Genome-scale metabolic network reconstructions (GENREs) for can be used to elucidate the functional metabolic mechanisms of individual pathobionts[6,10]. Once assembled, GENREs can probe an organism's genotype-phenotype relationship through constraint-based modeling and analysis (COBRA)[11]. Computational modeling through GENREs has proven effective at defining functional metabolism in individual priority pathogens, allowing for interpretation of mechanisms of infection and antibiotic resistance[10].

Here, we determined the evolutionary relatedness of metabolic phenotypes across pathobionts and the role of isolate environment in functional metabolic adaptation. We characterized the correlation of functional metabolism with the physiological niche of a pathobiont. We also predicted novel antimicrobial targets for pathobionts specific to a given physiological niche. To address the above questions, we generated the first database of GENREs of all known bacterial pathobionts (referred to as PATHGENN) with a current total of 914 *in silico* models of pathobiont metabolism, which can serve as a key resource for the community.

RESULTS

The PATHGENN Database

We created PATHGENN, a database of GENREs for all known human bacterial pathobionts through an automated pipeline (Figure S1). PATHGENN utilizes publicly available genome sequences from the Bacterial and Viral Bioinformatics Resource Center (BV-BRC)[12] paired with open-source software including Python and COBRApy [11], and a recently developed GENRE reconstruction algorithm[13]. The PATHGENN database is the first to contain GENREs of all known human bacterial pathobionts and is among the largest publicly available databases of GENREs [14,15]. PATHGENN consists of 914 GENREs, covering 345 species, 94 genera, 36 orders, 17 classes, and 9 phyla (Figure 1a, c) of pathobionts. PATHGENN GENREs account for the function of a sum total of 1.27 million reactions (6,304 unique reactions), 1.22 million genes, and 1.20 million metabolites. Each GENRE contains an average of 1,355 reactions (standard deviation: 344), 1,310 genes (standard deviation: 593), and 1,394 metabolites (standard deviation: 331) (Figure 1b). The relationship between the number of genes and reactions in the reconstructions is logarithmic, which is consistent with the expectation that there are limited evolutionary advantages for bacteria with increasingly large genomes[16](Figure 1d).

KEGG reaction annotations were utilized and reactions across all PATHGENN GENREs were separated into core (present in > 75% of GENREs), accessory (between 25% and 75%), and unique (present in < 25%) metabolism. There are 2,515 annotated unique reactions, 1,044 annotated accessory reactions, and 752 annotated core reactions (Figure 2a). The large number of unique reactions can be attributed to the size of the PATHGENN database and the taxonomic range PATHGENN GENREs represent. Furthermore, we determined notable differences in the unique and core metabolic subsystems through KEGG reaction subsystem annotation. More unique reactions were involved in xenobiotic metabolism (7% more), terpenoid/polyketide metabolism (11% more), and carbohydrate metabolism (4% more). Additionally, more core reactions were involved in nucleotide metabolism (7% more), and cofactor/vitamin metabolism (2% more) (Figure 2b).

Metabolic Phenotype Evolution

To understand the evolutionary relationship between pathobionts and their essential genes and network structure (two important attributes of functional metabolism), we calculated predicted essential genes, genetic distance between all pairs of pathobionts, and delineated differences in the reactions present in each organism. For each strain, essential gene profiles were determined

by using an FBA single-gene-knockout method in COBRApy. Given gene essentiality is a function of the organism's physiological environment, for this analysis all exchange reactions were open which results in the minimum number of essential genes for a given organism. Reaction presence profiles were created by probing the model in COBRApy (see Methods). These analyses produced binary profiles describing the presence of all essential genes and reactions in each model, which were subsequently used to calculate pairwise dissimilarity. The evolution of essential gene and reaction presence profiles is shown in Figures 3 and S2, respectively. Both relationships can be approximated with a three-parameter logarithmic growth function. Additionally, the logarithmic function reaches a saturation point ($x \mid y = 1.0$) for essential gene dissimilarity and reaction presence dissimilarity.

The saturation points observed in Figure 3 are indicative of conserved essential genes and reactions, respectively, across bacterial strains. That is, even at genetic difference of 100%, a pair of pathobionts will be only 18% different with respect to the essential gene profiles, and 34% different with respect to the reaction presence profiles. A previous study[17] determined a similar relationship between essential gene profiles and genetic distance across bacteria (not specifically pathobionts), but determined a saturation point of ~53% essential gene difference. This discrepancy in essential gene saturation point could be attributed to possible inherent pathobiont similarities that are not shared across all genera of bacteria. With host infection as a shared functional process of pathobionts, this result could suggest a shared functional signature associated with infection regardless of the specific niche which is not shared with non-pathobiont bacteria.

Additionally, the logarithmic trends shown in Figure 3 suggests there is adaptive pressure for closely related pairs of organisms to evolve to occupy their own distinct metabolic niche. As pathobionts begin to occupy distinct metabolic niches, they continually adapt their metabolic capabilities to better take advantage of their new environments, suggesting metabolic composition of the environment as a major governing principle of the evolution of functional metabolism.

Essential Gene Metabolic Subsystem Analysis

We further explored the relationship between physiological environment and metabolic function by essential gene subsystem analysis. We pooled the essential genes for all isolates of a given environment, and determined the metabolic subsystem distribution through KEGG genes annotation. Figure 4 shows the metabolic subsystem distribution of essential genes in eight of the most represented isolate environments: throat, respiratory, lung, stool, ear, stomach, mouth, and blood. There is significantly different subsystem representation across physiological environments as determined by an ANOVA test for each subsystem ($p < 0.05$ for all subsystems).

Some of the most notable differences in metabolic subsystem representation were amongst stomach isolates. There was evident lack of nucleotide metabolism, energy metabolism, and glycan metabolism in the essential genes of stomach isolates. Additionally, there was a clear enrichment of amino acid and lipid metabolic subsystems compared to essential gene subsystems in other isolate environments. The clear differences in metabolic subsystem utilization by organisms in different environments provides strong evidence for differential metabolic functional adaptation according to environment.

Influence of Environment on Functional Metabolism

Previous studies have delineated a relationship between functional metabolism and taxonomic class [15,18,19]. While it is clear that taxonomy is a driver for metabolic function, functional metabolism could also be attributed to physiological environment because an organism's environment

influences adaptation. To determine if there is a significant association between functional metabolism and physiological envirionment in addition to taxonomic class in pathobionts, we utilized flux balance analysis (FBA)[20] for each strain (n = 10 samples per strain). t-SNE was used to reduce the dimensionality of the flux output across strains and for subsequent visualization (see methods). We colored the t-SNE output on both taxonomic class (Figure 5a) and isolate environment (Figure 5b). Significant clustering was exhibited in Figure 5a and b (PERMANOVA: $p < 0.01$), suggesting functional metabolism is related to both taxonomic class and isolate environment.

Gammaproteobacteria is the class of bacteria with the largest number of models in PATHGENN (Figure 5a). However, Gammaproteobacteria isolates came from a variety of sources including stool, urine, lung, and blood among others (Figure 5b). Gammaproteobacteria is the most genera-rich taxon of Prokaryotes, containing over 250 genera[21]. This diversity in bacterial genera within the Gammaproteobacteria suggests a broader range of functional capabilities than other taxa, providing reasoning for the diverse environments from which Gammaproteobacteria were isolated. Another notable cluster, Actinomycetia, contains isolates from lung, respiratory, sputum, and throat sites. *Mycobacterium tuberculosis* and *Actinomyces* species belong to this class and are known to infect the lungs and throat respectively[22,23]. Clustering of *M. tuberculosis* and *Actinomyces* suggests organisms in similar environments across the respiratory tract exhibit similar functional capabilities.

A prominent cluster in Figure 5b is associated with bacteria isolated from the stomach. The stomach environment is highly acidic (pH 1.5 to 2.0)[24], allowing for only a few key bacteria to take up residence, one of which is *Helicobacter pylori. H. pylori* has adapted to this extremely unique environment by utilizing differential metabolic pathways[25]. The evident separation of the stomach cluster from others and the uniqueness of the stomach environment suggests[25] bacteria with highly unique functional metabolism. This result suggests genes essential to growth in stomach isolates are uniquely essential compared to pathobionts from other isolation sites. We can leverage these uniquely essential genes to identify novel antimicrobial targets that are specific to stomach pathobionts.

Identifying Site-Specific Antimicrobial Targets

To determine genes that are uniquely essential to stomach bacteria, essential genes were determined for all strains in PATHGENN using an FBA single-gene-knockout method in COBRApy (see Methods). If a gene was considered essential if >= 80% of strains in an isolate environment requires the gene to produce biomass. Two genes were identified as uniquely essential to stomach pathobionts (not essential in any other environment), *fabF* and *tktA*. *fabF* encodes the beta-ketoacyl-ACP synthase (KAS), implicated in the chain elongation step of fatty acid synthesis[26], and *tktA* encodes transketolase (TK), the most critical enzyme in the non-oxidative pentose phosphate pathway[27.] While neither of these genes are currently known antimicrobial targets specific to stomach pathobionts, there already exist several antimicrobials that target these gene products. According to DrugBank[28], *fabF* is a target of lauric acid. Lauric acid has been shown to have bactericidal effects against the stomach pathogen *H. pylori* and was cited to have a lower propensity to develop resistance compared to metronidazole or tetracycline[29]. Other drugs that target *fabF* and *tktA* are Cerulenin (*fabF*, currently used as an antifungal antibiotic), Platensimycin (*fabF*, currently in preclinical trials as a MRSA antibiotic), and Cocarboxylase (*tktA*, currently used to target *tktA* in *E. coli*), although there is no published literature regarding their use to treat stomach specific infections. The ability to predict lauric acid as a possible stomach-targeted antimicrobial with indirect literature validation demonstrates the value of PATHGENN to enable clinical hypothesis generation.

Additionally, we visualized the pathway structure that the genes *tktA* and *fabF* are implicated in across three stomach isolates that were captured in the PATHGENN database: *Helicobacter pylori*, *Arcobacter butzleri*, and *Campylobacter coli* using fluxer[30] and adapted the generated pathways in Figure 6. There are clear differences in pathway structure between the three different species of stomach isolates.

DISCUSSION

Here, we present a novel pipeline for generating GENREs of human bacterial pathobionts and apply it to create 914 GENREs representing all known bacterial pathobionts, a resource called PATHGENN. PATHGENN is among the largest databases of GENREs[14,15], and the first specific to pathobionts. PATHGENN GENREs adhere to the community benchmarking standards (MEMOTE, see Methods) and utilizes the ModelSEED namespace. These standards allow PATHGENN GENREs to be easily used in conjunction with existing models from other sources. All PATHGENN models are publicly available, and we encourage others to utilize the database to probe biological and clinically relevant questions not explored here. While the models in PATHGENN are not manually curated, they were all developed using the same pipeline utilizing an automated gap-filling process, allowing for a large number of GENREs in PATHGENN to be directly compared.

There are a total of 2,515 reactions that were unique to less than 25% of GENREs (unique reactions) in PATHGENN, while there were 752 reactions that were common in greater than or equal to 75% of GENREs (core reactions). There is an evident enrichment of nucleotide metabolic subsystems in core reactions (7% more). This result is consistent with the ubiquitous role of nucleotide metabolism across bacterial species[31]. Additionally, it has been shown that the nucleotide metabolism pathway plays a role in pathogenesis, further providing evidence that the GENREs in PATHGENN accurately capture and represent the biochemical processes in pathobionts[32]. Furthermore, there was an enrichment of xenobiotic metabolic subsystems in unique reactions (7% more). Bacterial species evolve to utilize differential xenobiotic pathways to best make use of ingested compounds through the utilization of different enzymes and hydrolytic/reduction reactions[33]. The evolution of unique xenobiotic metabolic reaction pathways allows bacteria to occupy their own metabolic niches and take advantage of their environment.

Understanding the evolution of metabolic phenotypes can provide important insight into fitness and adaptation of pathobionts. We used PATHGENN to better understand metabolic evolution in the context of adaptation through changes in functional metabolism over generational time. Results presented in Figure 3 (and Figure S2) suggest that there is adaptive pressure for closely related organisms to occupy their own distinct metabolic niche, which could occur through possible mechanisms of horizontal gene transfer, random mutation, or other methods. Closely related pathobionts experience pressure to adapt and quickly occupy a distinct metabolic niche to avoid competition and ensure the survival of the species. In more distantly related species, organisms have already adapted to occupy their own unique metabolic niches. It is evident that organisms continue to specialize after finding their niche, adapting further to gain fitness in their given environment. This observation suggests a two-phase evolutionary process. First, an initial diversification of both essential genes and reaction network due to adaptive pressure, followed by further diversification over generations. Additionally, by definition, pathobionts share a common function with host infection. Consequently, that shared activity could limit functional differences even if genetic history of the pathobiont is quite distinct. This concept could explain results in the logarithmic nature of the relationship between essential gene/reaction similarity and genetic distance (Figure 3).

It is important to note that in Figure 3 there is one group of pathobiont pairs that are more genetically distant from each other. For every pair in this group, one bacterium in the pair is *Mycolicibacterium fortuitum,* which is an opportunistic pathogen that is responsible for skin and bone infections belonging to the *actinomycetia* taxonomic class[34]. In this group, the bacteria paired with *Mycolicibacterium fortuitum* are: seven different *Bacillus* species, two *Vibrio* species, two *Acinetobacter* species, two *Burkholderia* species, and one *Providencia, Enterobacter,* and *Stenotrophomonas* species. This result suggests that these species are genetically distant from *Mycolicibacterium fortuitum*, but have more similar essential gene profiles to *Mycolicibacterium fortuitum* than expected according to the log fit function. Additionally, there is a high density of pathobiont pairs with genetic distances between 0.2 and 0.3. This result suggests that the average genetic distance between pairs of pathobionts is between 0.2 and 0.3, which is consistent with what has been found in another study examining pairwise genetic distances (determined by 16S rRNA sequence alignment) across pairs of bacteria[35].

The analysis of the evolution of metabolic phenotypes suggests that isolate environment could be a major evolutionary driver of metabolic function. This idea was further confirmed by metabolic subsystem annotation of essential genes via KEGG orthologs. There was a clear difference in metabolic subsystem representation of essential genes in different isolate environments (ANOVA with $p < 0.05$ for each subsystem). This difference in metabolic subsystem utilization could also suggest isolates from different isolate environments are functionally different, thereby occupying distinct metabolic niches.

Functional metabolic similarities have been tied to taxonomic class in many studies[14,15,18,19], but the underlying importance of isolate environment and its role in driving adaptation is often underappreciated. We determined that functional metabolism is related to both taxonomic class and isolation source through FBA, dimensionality reduction and visualization (t-SNE), and subsequent PERMANOVA ($p < 0.01$). This result provides more support for the hypothesis that functional metabolism is related to metabolic niche, which has been suggested in previous work [15]. Additionally, within taxonomic classes, there are distinct clusters of flux samples based on isolate environment. There are visibly distinct clusters of throat, respiratory, lung, ear, stomach, blood, and stool, which were also shown to have distinct metabolic subsystem utilization in the essential gene and metabolic subsystem analysis (Figure 4). The corroboration of results in these two different analyses provides further evidence that isolate environment is a strong factor in the evolution of metabolic phenotypes.

Additionally, within the class of *Epsilonproteobacteria* there are two distinct clusters: a stomach cluster and a stool cluster. This result further implies that closely related organisms develop distinct functional metabolic capabilities related to their specific environment to outcompete related organisms and ensure the survival of the distinct population or species. These results suggest similarities between organisms that occupy the same environment and not only because they are phylogenetically related. While phylogeny is undoubtedly related to metabolic phenotype, it is clear that environment is also a driving factor for the evolution of functional metabolic characteristics.

The most distinct cluster of metabolic flux samples is the stomach cluster, implying these isolates exhibit strong similarities in functional metabolism. Additionally, this suggests these isolates are functionally distinct from isolates of different environments. These functional metabolic differences could be driven by the extreme environment of the stomach, pressuring adaptation. Distinct metabolic phenotypes in the stomach environment were also shown in Figure 4, with a visible enrichment of amino acid and lipid metabolism subsystems and a lack of nucleotide, energy, and glycan metabolic subsystems in the essential genes of stomach isolates.

Stomach infection with *H. pylori* can cause a variety of adverse effects including chronic gastritis leading to complications (peptic ulcer, gastric cancer, lymphoma)[36,37]. Additionally, *H. pylori* infection is incredibly difficult to treat, requiring multi-antimicrobial regimens and acid suppressants[36]. Given that stomach isolates are functionally different from isolates in other environments, we identified two genes, *fabF* and *tktA*, that are uniquely essential to stomach isolates. Creating antimicrobial therapies specifically targeting these genes could eliminate the need for multi-antimicrobial regimens and broad-spectrum antibiotics which are associated with adverse health effects[9]. Additionally, targeted antimicrobial therapies would allow for more rapid response to infection, since all organisms in an environment can be treated unilaterally with one antimicrobial so species identification is not necessary. We identified four drugs that target these genes: lauric acid (*fabF*), Cerulenin (*fabF*), Platensimycin (*fabF*), and Carboxylase (*tktA*). Lauric acid has been cited to have antimicrobial properties against *H. pylori*, and a lower propensity to cause the development of resistance than if *H. pylori* were treated with metronidazole or tetracycline[29]. Since the GENREs in PATHGENN were able to correctly predict lauric acid as an antimicrobial target, the other three identified drugs could be tested. Additionally, we visualized the pathways that *fabF* and *tktA* are a part of in three different stomach isolate species (*H. pylori*, *A. butzleri*, and *C. coli*) (Figure 6). There are clear differences in pathway structure between the three different species despite *tktA* and *fabF* being essential genes in stomach isolates. This finding further highlights the importance of investigating unique metabolic functional capabilities that develop due to adaptive pressures for antimicrobial discovery and drug repurposing.

The GENREs in PATHGENN were generated through an automated pipeline, first generating genome-informed draft network reconstructions then a curation of the reconstructions through an automated gapfilling process based on parsimony principles. Generating all models through the same pipeline with the same level of automated curation allows for comparison across all GENREs for a high-level, cross-genome, analysis of bacterial pathobionts. However, the strength of the models is dependent on the accuracy and detail of genome annotations. The analyses presented in this paper could be enhanced by further manual curation of poorly annotated species.

We successfully generated a database of 914 GENREs of all human bacterial pathobionts (PATHGENN) which we used to investigate the role of environment in adaptation and generation of unique functional metabolism. Additionally, we were able to use uniquely essential metabolic genes in pathobionts isolated from the stomach to predict possible targeted antimicrobial options for treating stomach-specific bacterial infection. We can continue to investigate questions related to functional metabolism by curating the isolate environment to simulate metabolism in more specific contexts. This effort will allow for better understanding of the functional metabolic differences in pathobionts in the context in which they grow as infections. Furthermore, we can begin to integrate environment-specific functional metabolism and other pertinent metadata to identify drug targets that are relevant to patient-specific infections. Identifying unique metabolic functions across pathobiont species is the first step to developing a framework for a personalized medicine approach to addressing infection in the clinic.

METHODS

GENRE Creation From Genome Sequences

We first filtered all genome sequences in the BV-BRC 3.6.12 database to only include those that were considered “good” quality and “complete”. BV-BRC guidelines define “good” as “a genome that is sufficiently complete (80%), with sufficiently low contamination (10%)”, and amino acid sequences that are at least 87% consistent with known protein sequence. “Complete” means that replicons were completely assembled.

There are 538 species of bacterial pathobionts[4], some of which either do not have publicly available genome sequences via BV-BRC or do not have "good" and "complete" genome sequences in BV-BRC. There is at least one NCBI taxid for each pathobiont species, with some species having multiple unique NCBI taxids. Multiple genome sequences are available in BV-BRC for each NCBI taxid, so sequences were selected based on the presence of metadata in a hierarchical nature. Sequences with the most associated metadata were prioritized. If multiple sequences had the same amount of metadata, we selected the sequence that had isolate environment-associated metadata. If multiple sequences fulfilled the previous requirements, the strain that had host health-associated metadata was selected. This hierarchical selection was continued for metadata categories of isolation country, collection date, and host age, in that order of priority. The resulting list contained 914 unique genome sequences. This procedure was automated with a python script.

All amino acid sequences were then automatically annotated with RAST 2.0[38,39], and GENREs were created for each strain using the Reconstructor[13] algorithm. All models are publicly available (see Data Availability section). We benchmarked all GENREs using the community standard, MEMOTE[40], and have included all scores in stable .html files on GitHub.

Genetic Distance and Essential Gene Profile/ Reaction Presence profile distance

All sequences used to create GENREs in PATHGENN were re-annotated to determine the rRNA genome features. All 16S rRNA sequences were extracted from the annotation output, for a total of 245 16S rRNA sequences, each from a unique PATHGENN strain (still representing the same 9 phyla represented in all 914 PATHGENN GENREs). The 16s rRNA sequences were then aligned using Clustal Omega and the resulting Percent Identity Matrix was downloaded. Identity percentages were converted to values between 0 and 1, 0 being the most similar and 1 being the most different. This value was then converted to a percentage. This metric was defined as the genetic distance for subsequent analyses.

Essential gene profiles for each of the corresponding 245 GENREs (those with available 16s rRNA sequences) using an FBA-based, single-gene-knockout method in COBRApy (cobra.flux_analysis.variability.find_essential_genes()). Essential genes were then converted to KEGG Orthologs, and a binary matrix was created indicating essential gene presence in each strain (1 = presence, 0 = absence). The pairwise essential gene distance was defined as the calculated hamming distances[41] between each strain's essential gene profile.

Reaction presence was determined for each of 245 GENREs via model probing in COBRApy. A binary matrix was created indicating reaction presence or absence in each strain (1 = presence, 0 = absence). The pairwise reaction presence distance was defined as the calculated hamming distances between each strain's reaction presence profile.

Genetic distance vs essential gene distance, and genetic distance vs reaction presence distance were plotted for each pair of pathobionts. Logarithmic functions were fit to both plots using the scipy.optimize.curve_fit function in the python scipy toolbox.

FBA and t-SNE Dimensionality Reduction/Visualization

For each of the 914 models, Flux Balance Analysis (FBA) was performed using the COBRApy toolbox for each model in PATHGENN to capture metabolic flux through all model reactions. 10 flux samples were taken per model for a total of 9,140 flux samples.

t-distributed stochastic neighbor embedding (t-SNE)[42] was used for dimensionality and subsequent visualization of the FBA output. The perplexity parameter was optimized to preserve local and global relationships in the data using $P = N^{\frac{1}{2}}$, where P = perplexity, and N = number of points. Points were colored based on taxonomic class, and subsequently colored on isolation source for visualization purposes. Significant clusters in both taxonomic class and isolation site t-SNE outputs were determined using a PERMANOVA[43] test.

To ensure that 10 flux samples was sufficient to capture the flux solution space as well as 100 flux samples per model would, we ran paired-down t-SNE analyses. We randomly sampled 100 GENREs from the 914 total GENREs in PATHGENN. Then, for each of those 100 GENREs we used 100 flux samples to perform dimensionality reduction and subsequent visualization via t-SNE (Figure S3). We performed this analysis three times, to ensure that the results would hold true for multliple randomly selected subsets of GENREs.

Through this subsequent t-SNE analysis, we still see clustering by taxonomic class in figure S3. Specifically, we still see large clusters of *Gammaproteobacteria* and *Actinomycetia*. Additionally, we still see the separation of *Epsilonproteobacteria* into distinct clusters, one of which is completely comprised of stomach isolates.

Determination of Novel Antibiotics to Target Stomach Isolates

Essential genes for all 914 models were determined using an FBA based single-gene-knockout method in COBRApy (cobra.flux_analysis.variability.find_essential_genes()). All essential genes were translated to KEGG orthologs. Strains and their corresponding essential genes were grouped by isolation site. Essential genes present in >= 80% of strains in a given isolation source were defined as uniquely essential to that isolation source. Uniquely essential genes present in stomach isolates that are not uniquely essential to other isolation sites were selected. DrugBank[28] was used to identify drugs that target uniquely essential genes of stomach isolates.

REFERENCES

1. Wang, H., Naghavi, M., Allen, C., Barber, R.M., Bhutta, Z.A., Carter, A., Casey, D.C., Charlson, F.J., Chen, A.Z., Coates, M.M., et al. (2016). Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: a systematic analysis for the Global Burden of Disease Study 2015. The Lancet *388*, 1459–1544. 10.1016/S0140-6736(16)31012-1.
2. Smith, K.M., Machalaba, C.C., Seifman, R., Feferholtz, Y., and Karesh, W.B. (2019). Infectious disease and economics: The case for considering multi-sectoral impacts. One Health *7*. 10.1016/j.onehlt.2018.100080.
3. Bloom, D.E., and Cadarette, D. (2019). Infectious disease threats in the twenty-first century: Strengthening the global response. Front Immunol *10*. 10.3389/fimmu.2019.00549.
4. Taylor, L.H., Latham, S.M., and Woolhouse, M.E.J. (2001). Risk factors for human disease emergence. Philosophical Transactions of the Royal Society B: Biological Sciences *356*, 983–989. 10.1098/rstb.2001.0888.
5. Chandra, H., Sharma, K.K., Tuovinen, O.H., Sun, X., and Shukla, P. (2021). Pathobionts: mechanisms of survival, expansion, and interaction with host with a focus on Clostridioides difficile. Gut Microbes *13*. 10.1080/19490976.2021.1979882.
6. Dunphy, L.J., Yen, P., and Papin, J.A. (2019). Integrated Experimental and Computational Analyses Reveal Differential Metabolic Functionality in Antibiotic-Resistant Pseudomonas aeruginosa. Cell Syst *8*, 3-14.e3. 10.1016/j.cels.2018.12.002.

7. Dinges, M.M., Orwin, P.M., and Schlievert, P.M. (2000). Exotoxins of Staphylococcus aureus.
8. Kellogg, D.S., Peacock, W.L., Deacon, W.E., Brown, L., and Pirkle, C.I. NEISSERIA GONORRHOEAE I. VIRULENCE GENETICALLY LINKED TO CLONAL VARIATION (Communicable Disease Center).
9. Wiens, J., Snyder, G.M., Finlayson, S., Mahoney, M. v., and Celi, L.A. (2018). Potential adverse effects of broad-spectrum antimicrobial exposure in the intensive care unit. Open Forum Infect Dis *5*. 10.1093/ofid/ofx270.
10. Sertbas, M., and Ulgen, K.O. (2020). Genome-Scale Metabolic Modeling for Unraveling Molecular Mechanisms of High Threat Pathogens. Front Cell Dev Biol *8*. 10.3389/fcell.2020.566702.
11. Ebrahim, A., Lerman, J.A., Palsson, B.O., and Hyduke, D.R. (2013). COBRApy: COnstraints-Based Reconstruction and Analysis for Python. BMC Syst Biol *7*. 10.1186/1752-0509-7-74.
12. Davis, J.J., Wattam, A.R., Aziz, R.K., Brettin, T., Butler, R., Butler, R.M., Chlenski, P., Conrad, N., Dickerman, A., Dietrich, E.M., et al. (2020). The PATRIC Bioinformatics Resource Center: Expanding data and analysis capabilities. Nucleic Acids Res *48*, D606–D612. 10.1093/nar/gkz943.
13. Jenior+, M.L., Glass+, E.M., and Papin, J.A. Title Reconstructor: A COBRApy compatible tool for automated genome-scale metabolic network reconstruction with parsimonious flux-based gap-filling. 10.1101/2022.09.17.508371.
14. Magnúsdóttir, S., Heinken, A., Kutt, L., Ravcheev, D.A., Bauer, E., Noronha, A., Greenhalgh, K., Jäger, C., Baginska, J., Wilmes, P., et al. (2017). Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. Nat Biotechnol *35*, 81–89. 10.1038/nbt.3703.
15. Carey, M.A., Medlock, G.L., Stolarczyk, M., Petri, W.A., Guler, J.L., and Papin, J.A. (2022). Comparative analyses of parasites with a comprehensive database of geno-scale metabolic models. PLoS Comput Biol *18*. 10.1371/journal.pcbi.1009870.
16. Lefébure, T., Morvan, C., Malard, F., François, C., Konecny-Dupré, L., Guéguen, L., Weiss-Gayet, M., Seguin-Orlando, A., Ermini, L., Sarkissian, C. der, et al. (2017). Less effective selection leads to larger genomes. Genome Res *27*, 1016–1028. 10.1101/gr.212589.116.
17. Plata, G., Henry, C.S., and Vitkup, D. (2015). Long-term phenotypic evolution of bacteria. Nature *517*, 369–372. 10.1038/nature13827.
18. Lee, C.C., Lo, W.C., Lai, S.M., Chen, Y.P.P., Tang, C.Y., and Lyu, P.C. (2012). Metabolic classification of microbial genomes using functional probes. BMC Genomics *13*. 10.1186/1471-2164-13-157.
19. Luo, G., Fotidis, I.A., and Angelidaki, I. (2016). Comparative analysis of taxonomic, functional, and metabolic patterns of microbiomes from 14 full-scale biogas reactors by metagenomic sequencing and radioisotopic analysis. Biotechnol Biofuels *9*. 10.1186/s13068-016-0465-6.
20. Orth, J.D., Thiele, I., and Palsson, B.O. (2010). What is flux balance analysis? Nat Biotechnol *28*, 245–248. 10.1038/nbt.1614.
21. Williams, K.P., Gillespie, J.J., Sobral, B.W.S., Nordberg, E.K., Snyder, E.E., Shallom, J.M., and Dickerman, A.W. (2010). Phylogeny of gammaproteobacteria. J Bacteriol *192*, 2305–2314. 10.1128/JB.01480-09.
22. Smith, I. (2003). Mycobacterium tuberculosis pathogenesis and molecular determinants of virulence. Clin Microbiol Rev *16*, 463–496. 10.1128/CMR.16.3.463-496.2003.
23. Hasan, M., and Kumar, A. (2011). Actinomycosis and tonsillar disease. BMJ Case Rep. 10.1136/bcr.01.2011.3750.

24. Fujimori, S. (2020). Gastric acid level of humans must decrease in the future. World J Gastroenterol *26*, 6706–6709. 10.3748/wjg.v26.i43.6706.
25. Lee, W.C., Goh, K.L., Loke, M.F., and Vadivelu, J. (2017). Elucidation of the Metabolic Network of Helicobacter pylori J99 and Malaysian Clinical Strains by Phenotype Microarray. Helicobacter *22*. 10.1111/hel.12321.
26. Edwards, P., Nelsen, J.S., Metz, J.G., and Dehesh, K. (1997). Cloning of the fabF gene in an expression vector and in vitro characterization of recombinant fabF and fabB encoded enzymes from Escherichia coli. FEBS Lett *402*, 62–66. 10.1016/S0014-5793(96)01437-8.
27. Jung, Y.-M., Lee, J.-N., Shin, H.-D., and Lee, Y.-H. (2004). Role of tktA gene in pentose phosphate pathway on odd-ball biosynthesis of poly-β-hydroxybutyrate in transformant Escherichia coli harboring phbCAB operon. J Biosci Bioeng *98*, 224–227.
28. Wishart, D.S., Knox, C., Guo, A.C., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z., and Woolsey, J. (2006). DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res *34*. 10.1093/nar/gkj067.
29. Jung, S.W., and Lee, S.W. (2016). The antibacterial effect of fatty acids on Helicobacter pylori infection. Korean Journal of Internal Medicine *31*, 30–35. 10.3904/kjim.2016.31.1.30.
30. Hari, A., and Lobo, D. (2020). Fluxer: A web application to compute, analyze and visualize genome-scale metabolic flux networks. Nucleic Acids Res *48*, W427–W435. 10.1093/NAR/GKAA409.
31. Lopatkin, A.J., and Yang, J.H. (2021). Digital Insights Into Nucleotide Metabolism and Antibiotic Treatment Failure. Front Digit Health *3*. 10.3389/fdgth.2021.583468.
32. Goncheva, M.I., Chin, D., and Heinrichs, D.E. (2022). Nucleotide biosynthesis: the base of bacterial pathogenesis. Trends Microbiol *30*, 793–804. 10.1016/j.tim.2021.12.007.
33. Koppel, N., Rekdal, V.M., and Balskus, E.P. (2017). Chemical transformation of xenobiotics by the human gut microbiota. Science (1979) *356*, 1246–1257. 10.1126/science.aag2770.
34. Morgado, S., Ramos, N. de V., Freitas, F., da Fonseca, É.L., and Vicente, A.C. (2021). Mycolicibacterium fortuitum genomic epidemiology, resistome and virulome. Mem Inst Oswaldo Cruz *116*. 10.1590/0074-02760210247.
35. Bukin, Y.S., Galachyants, Y.P., Morozov, I. v., Bukin, S. v., Zakharenko, A.S., and Zemskaya, T.I. (2019). The effect of 16s rRNA region choice on bacterial community metabarcoding results. Sci Data *6*. 10.1038/sdata.2019.7.
36. Jenks, P.J. (2002). Causes of failure of eradication of Helicobacter pylori. Br Med J *325*, 3–4. 10.1136/bmj.325.7354.3.
37. Buzás, G.M. (2014). Metabolic consequences of Helicobacter pylori infection and eradication. World J Gastroenterol *20*, 5226–5234. 10.3748/wjg.v20.i18.5226.
38. Brettin, T., Davis, J.J., Disz, T., Edwards, R.A., Gerdes, S., Olsen, G.J., Olson, R., Overbeek, R., Parrello, B., Pusch, G.D., et al. (2015). RASTtk: A modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. Sci Rep *5*. 10.1038/srep08365.
39. Aziz, R.K., Bartels, D., Best, A., DeJongh, M., Disz, T., Edwards, R.A., Formsma, K., Gerdes, S., Glass, E.M., Kubal, M., et al. (2008). The RAST Server: Rapid annotations using subsystems technology. BMC Genomics *9*. 10.1186/1471-2164-9-75.
40. Lieven, C., Beber, M.E., Olivier, B.G., Bergmann, F.T., Ataman, M., Babaei, P., Bartell, J.A., Blank, L.M., Chauhan, S., Correia, K., et al. MEMOTE for standardized genome-scale metabolic model testing. 10.5281/zenodo.2636858.
41. Hamming, R.W. (1950). Error detecting and error correcting codes. Comput. Arith.
42. van der Maaten, L., and Hinton, G. (2008). Visualizing Data using t-SNE.

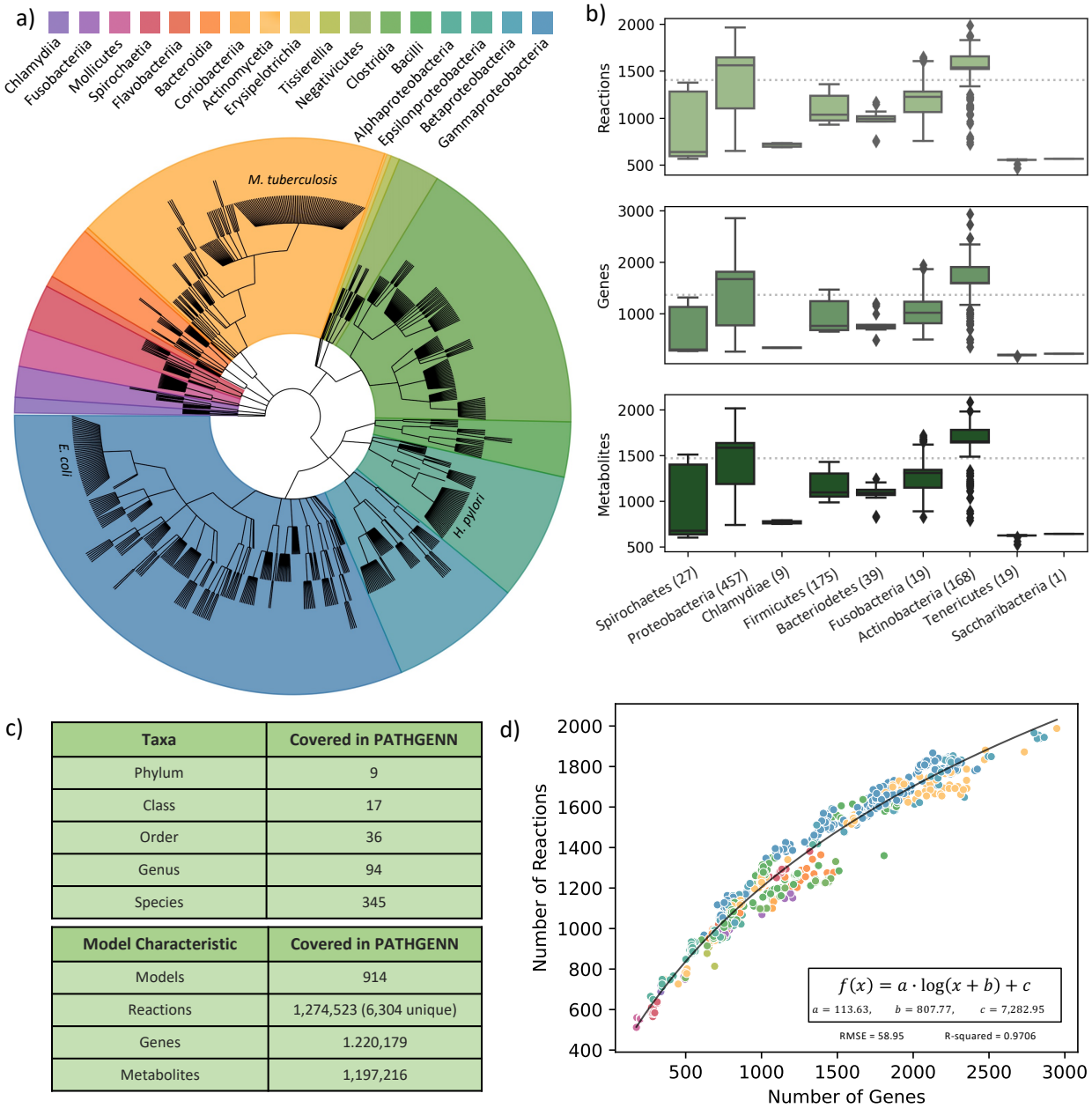
43. Anderson, M.J. (2017). Permutational Multivariate Analysis of Variance (PERMANOVA) . In Wiley StatsRef: Statistics Reference Online (Wiley), pp. 1–15. 10.1002/9781118445112.stat07841.
44. Asnicar, F., Weingart, G., Tickle, T.L., Huttenhower, C., and Segata, N. (2015). Compact graphical representation of phylogenetic data and metadata with GraPhlAn. PeerJ *2015*. 10.7717/peerj.1029.
45. Wattam, A.R., Abraham, D., Dalay, O., Disz, T.L., Driscoll, T., Gabbard, J.L., Gillespie, J.J., Gough, R., Hix, D., Kenyon, R., et al. (2014). PATRIC, the bacterial bioinformatics database and analysis resource. Nucleic Acids Res *42*. 10.1093/nar/gkt1099.

ACKNOWLEDGEMENTS

Funding: This work was supported by the NSF GRFP award number 1842490, the University of Virginia NIH Systems and Biomolecular Data Sciences Training Grant (grant number 1 T32 GM 145443-1), NIH R01s (R01-AI154242 and R01-AT010253) .

Author Contributions: E.M.G and J.A.P conceived of the project. E.M.G generated the PATHGENN collection and performed subsequent analyses. E.M.G wrote the initial manuscript draft. L.R.D aided in data analysis. A.S.W assisted with model annotation. E.M.G, L.R.D, A.S.W, and J.A.P edited and approved the manuscript for final submission.

Data availability: All PATHGENN GENRE models are publicly available on GitHub along with MEMOTE benchmarking scores and all pertinent code to this study: https://github.com/emmamglass/PATHGENN.



c)

Taxa	Covered in PATHGENN
Phylum	9
Class	17
Order	36
Genus	94
Species	345

Model Characteristic	Covered in PATHGENN
Models	914
Reactions	1,274,523 (6,304 unique)
Genes	1.220,179
Metabolites	1,197,216

Figure 1 | Scope of the PATHGENN database. (a) Phylogenetic tree depicting the diversity of 914 considered bacterial pathobionts in PATHGENN. It is important to note there are many strains of *E. coli, H. pylori,* and *M. tuberculosis* included in the database. This cladogram was created using the GraPhlAn[44] python tool. (b) Boxplots representing the spread of genes, reactions, and metabolites in each model, classified by phylum. The number in parentheses after the phylum name represents how many models are in that respective phylum. (c) PATHGENN represents 9 phyla, 17 classes, 36 orders, 94 genera, and 345 species of pathobionts. Across the 914 models, there are a sum total of 1.27 million reactions, 1.22 million genes, and 1.20 metabolites. (d) The relationship between the number of genes and the number of reactions in each model displays a positive trend and heteroscedasticity similar to other model ensembles[15]. Colors correspond to taxonomic class of pathobiont represented by each point (same legend as Figure 1 a)

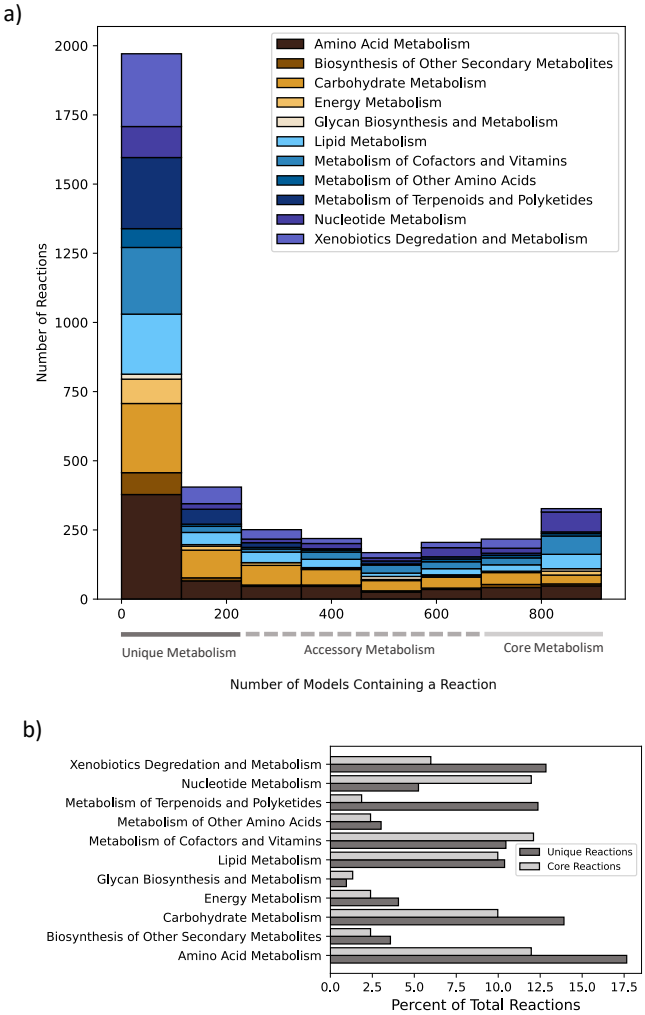


Figure 2 | Core and unique metabolic reaction subsystems across pathobionts. (a) Histogram of annotated reactions across models display prevalent reaction classes used in core metabolism (>75% models have a given reaction) and unique metabolism (<25% models have a given reaction). Notably, the reaction classes xenobiotic degradation/metabolism and metabolism of terpenoids/polyketides are much more prevalent in unique reactions than core reactions. PATHGENN is largest database of GENREs to date (914 GENRES representing 345 species), and the first to include all bacterial pathobionts. (b) Different metabolic subsystems are enriched in core and unique reactions. Amino acid, Xenobiotics, and Terpenoid/Polyketide metabolism is noticeably enriched in unique reactions, while Nucleotide metabolism is noticeably enriched in core reactions.

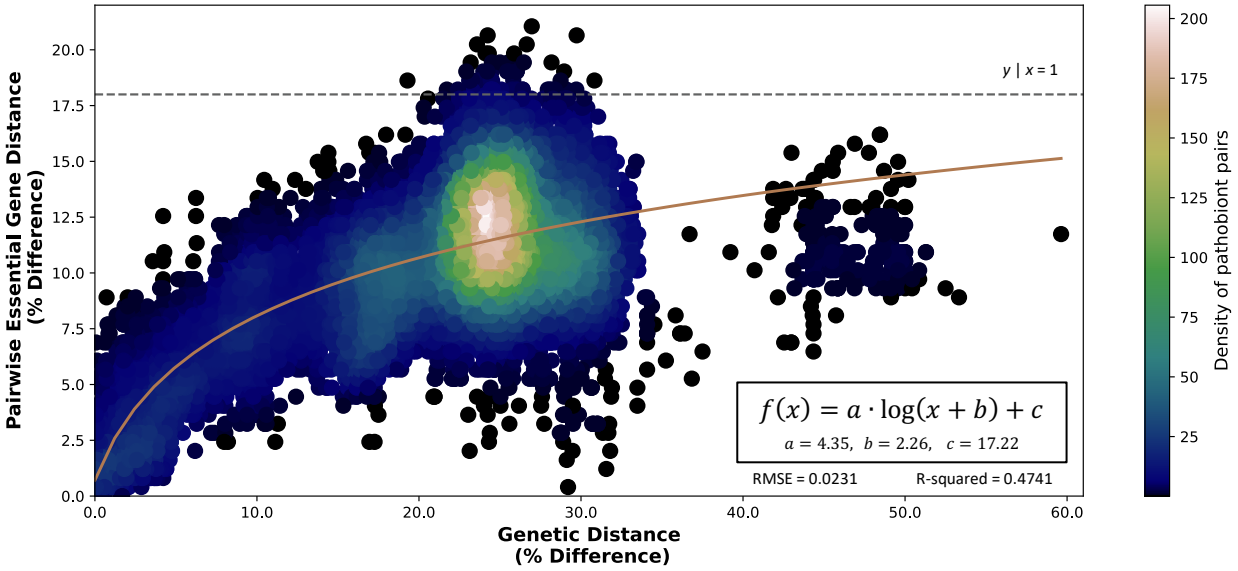


Figure 3 | Differences in metabolic function of pathobiont pairs are related to their genetic distance. The relationship between pairwise essential gene profile distance and genetic distance of 245 pathobionts suggests adaptive pressure for closely related pairs of organisms to evolve to occupy their own distinct metabolic niche. This result further suggests that metabolic composition of environment is a major governing principle of evolution of functional metabolism.

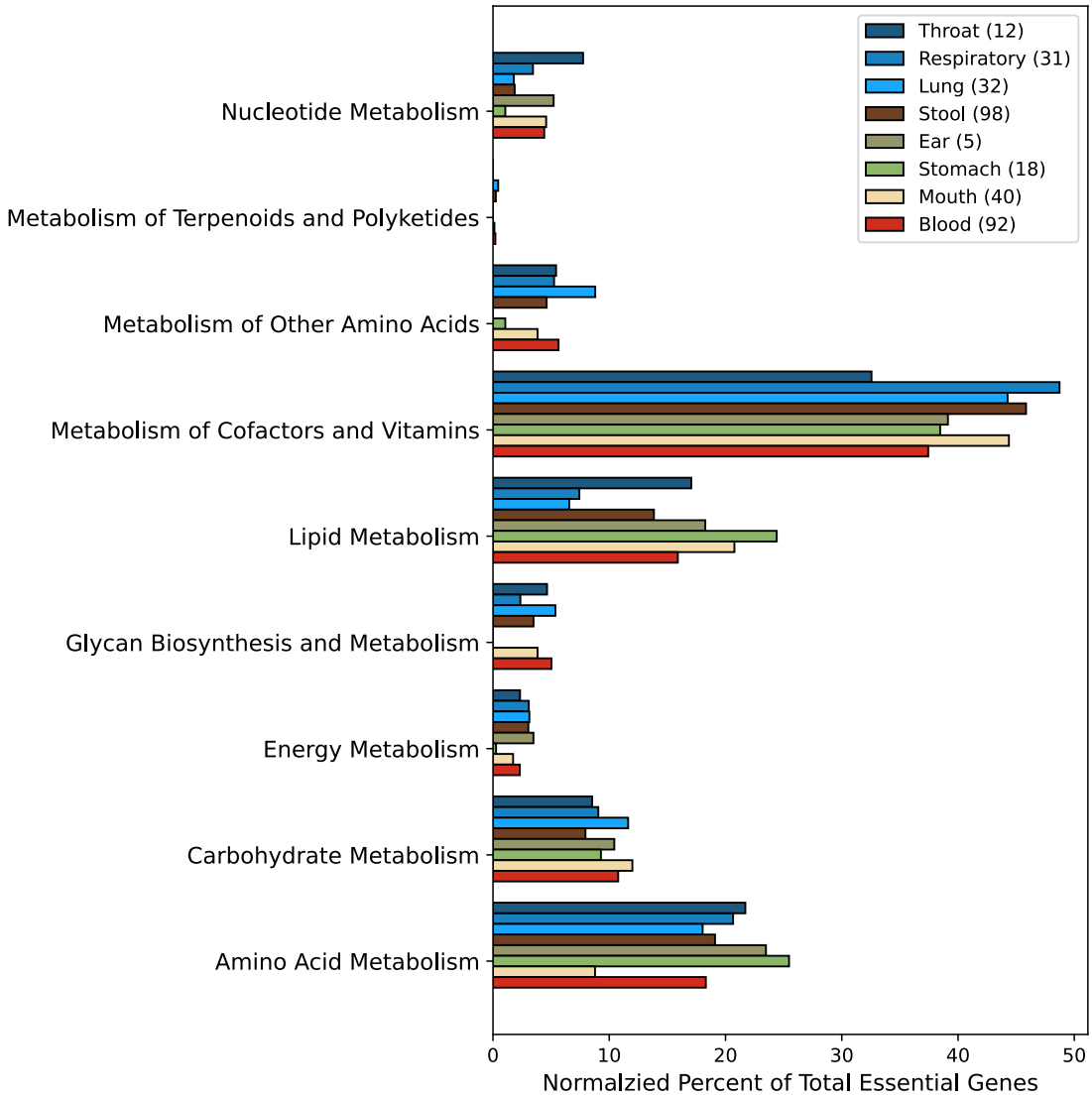


Figure 4 | Essential gene subsystems vary by isolate location. Enrichment of amino acid and lipid metabolism in stomach isolates is evident, along with an absence of essential genes used in glycan biosynthesis and energy metabolism. Each subsystem indicates differential metabolic subsystem utilization by isolate location (ANOVA $p < 0.05$).

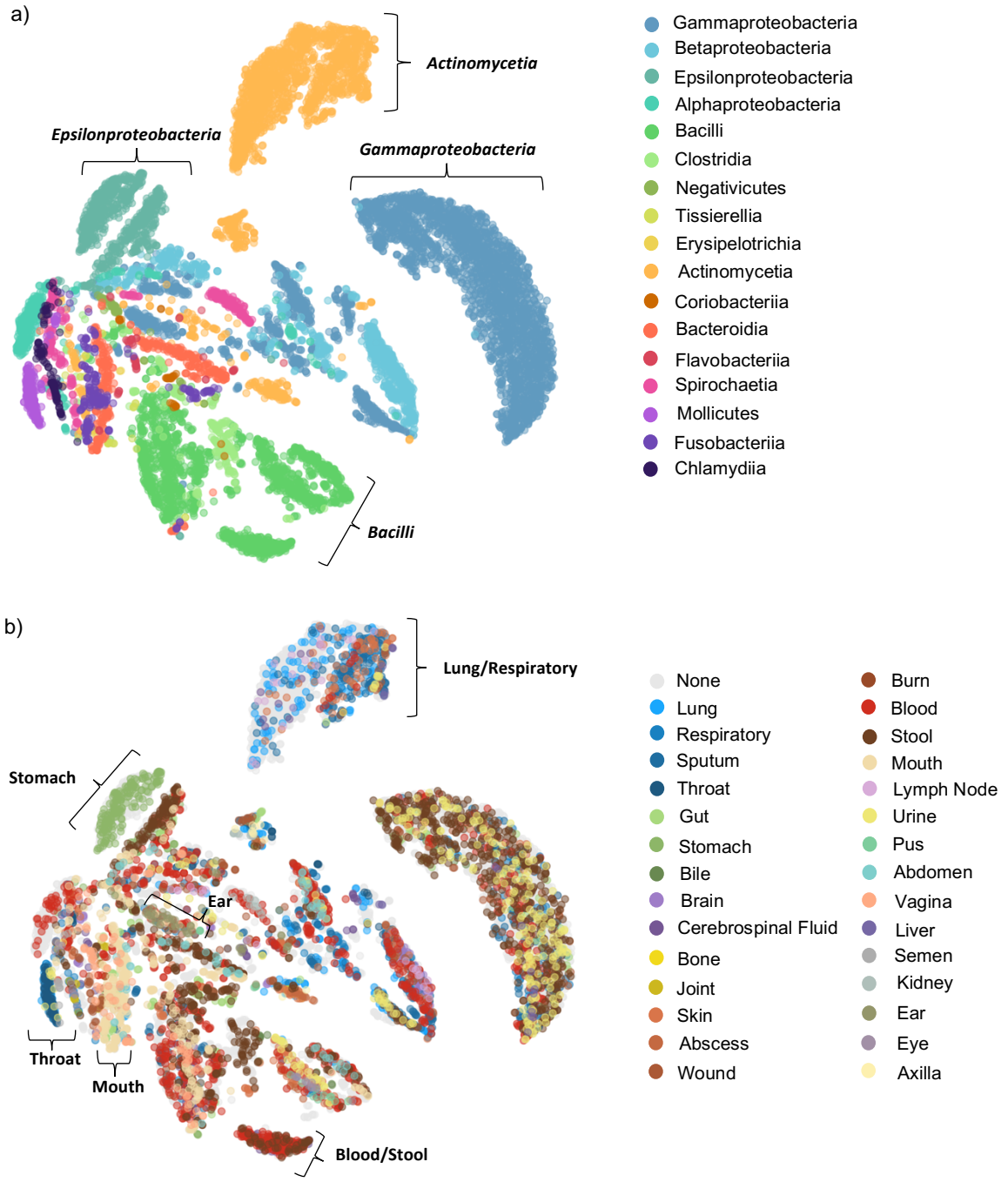


Figure 5 | tSNE of Flux Samples Clustering on Taxonomic Class and Isolation Site. 10 flux samples across all 914 GENREs were plotted using tSNE, and points were colored on taxonomic class (a) and isolation site (b).

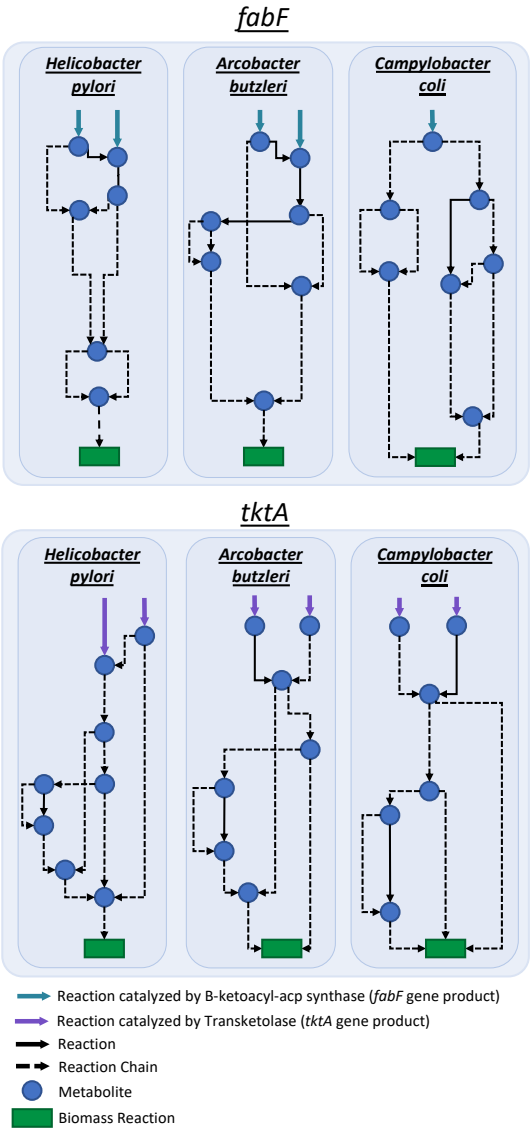
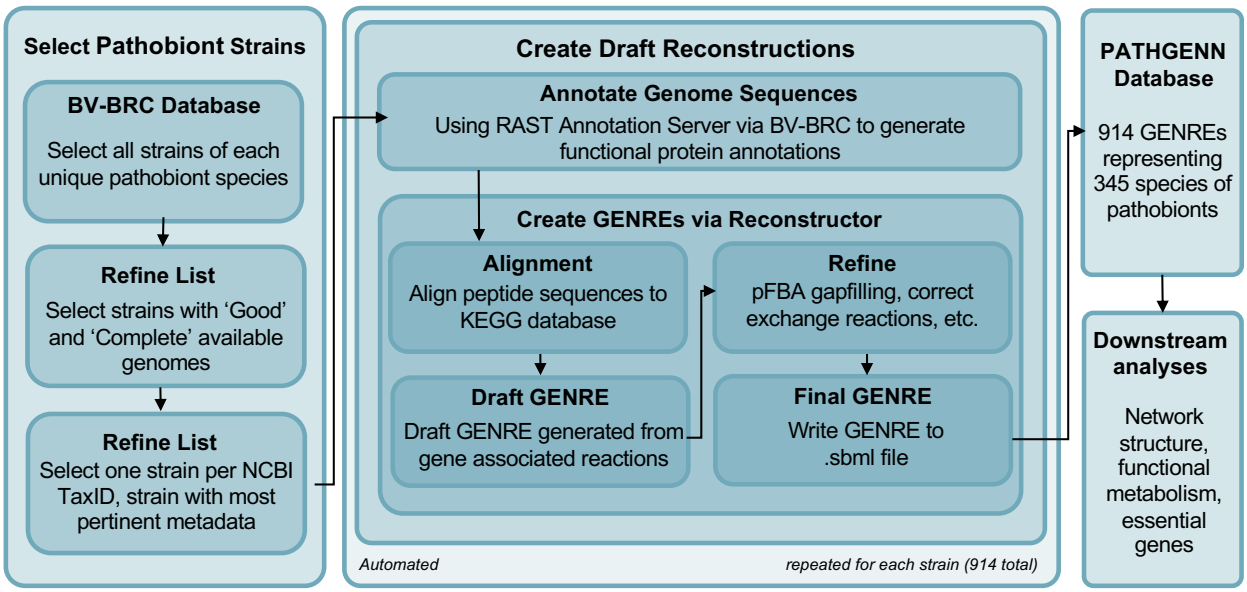
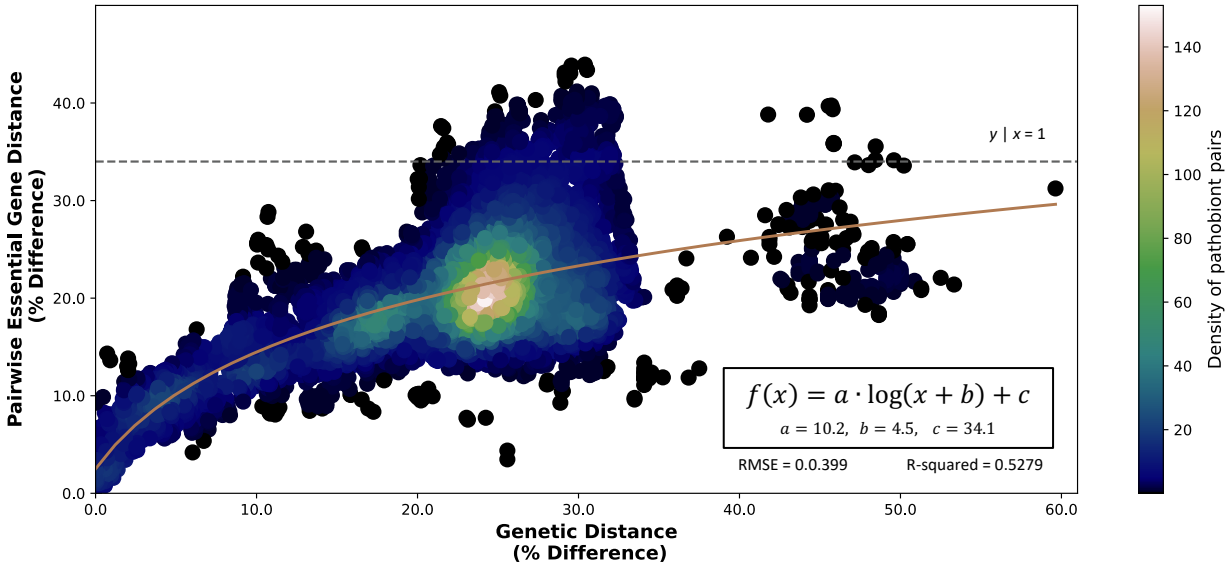


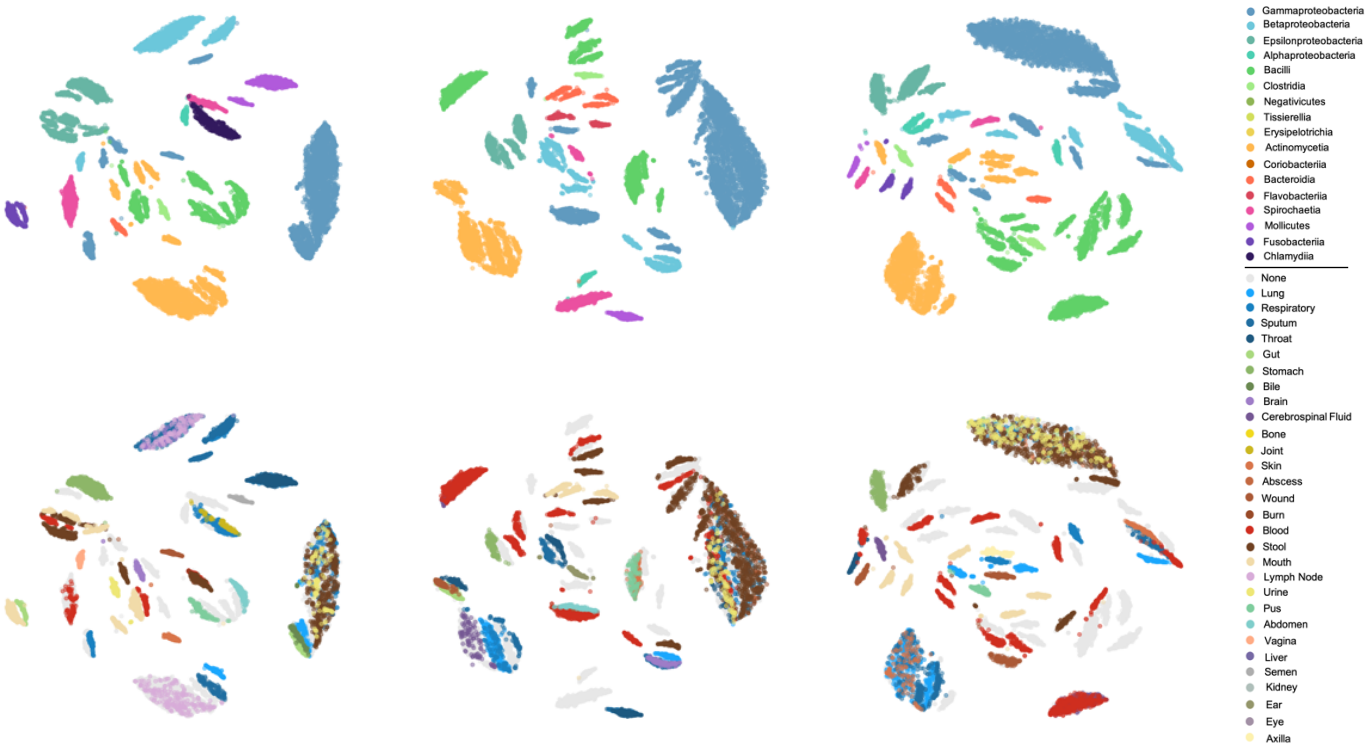
Figure 6 | *fabF* (a) and *tktA* (b) metabolic pathways in three stomach pathobionts: *Helicobacter pylori*, *Arcobacter butzleri*, and *Campylobacter coli*. There are differences in pathway structures in both *fabF* and *tktA* pathways across three stomach pathobionts. This figure was adapted from pathways generated with fluxer[30].



S1 | PATHGENN Development Pipeline. The BV-BRC database[45] was used to select pathobiont genome strains that satisfied quality criteria. These genome strains were then annotated using the RAST annotation toolbox[38,39] to generate the amino acid FASTA file that was then used in Reconstructor[13] to generate the 914 GENREs of PATHGENN.



S2 | Reaction differences in pathobiont pairs are related to their genetic distance. The relationship between pairwise reaction presence profile distance and genetic distance of 245 pathobionts can be approximated with log functions.



S3 | t-SNE plot of 100 flux samples for 100 GENREs. The clustering relationships seen in Figure 4 with 10 flux samples for each of 914 models are consistent with the clusters seen here with three randomly selected subsets of 100 GENREs with 100 flux samples each.