THE HUMAN MICROBIOME DRUG METABOLISM DATABASE

THE HUMAN MICROBIOME DRUG METABOLISM DATABASE

By AMOGELANG R. RAPHENYA, B.Eng.

A Thesis Submitted to the School of Graduate Studies in Partial Fulfillment of the Requirements for the Degree of Master of Science

You may hate gravity, but gravity does not care

—Albert Einstein

McMaster University, MASTER OF SCIENCE (2023), Hamilton, Ontario (Health Sciences)

TITLE: The Human Microbiome Drug Metabolism Database

AUTHOR: Amogelang R. Raphenya B.Eng. (McMaster University)

SUPERVISOR: Andrew G. McArthur

NUMBER OF PAGES: xx, 156

Lay Abstract

We use medications in our everyday life to treat infections and manage diseases. Yet, bacteria residing within the human gut can interact with these medications, which can cause undesirable outcomes. Many bacteria in the human gut produce biological catalysts known as enzymes that break down chemicals, including drugs. Medication is measured and given to an individual, called a dose, and the oral route is preferred. Enzymes break down oral and biliary system drugs, reducing the effective dose. As a result, medication becomes ineffective or toxic to the body. As such, we must study how each drug is affected by bacterial enzymes. I built a resource, the Human Microbiome Drug Metabolism (HMDM) database, to catalog all the bacterial genes that code for the enzymes reported in scientific papers to break down oral drugs. We can use the HMDM database to study bacterial enzymes that lead to poor drug efficacy.

Abstract

We rely on oral drugs to treat several diseases and infections. Yet, the gut microbiome modifies oral drugs within the human gut by using enzymes, facilitating efficient chemical reactions. These drug modifications impact effective doses and outcomes for individuals. The gut microbiome can convert drugs destined for excretion back to active drugs, and the converse is also true, the microbiome can inactivate active drugs, and both may lead to toxic effects. There is no resource for cataloging bacterial drug-metabolizing genes within the human gut microbiome with analytical tools to annotate these genes in sequenced gut microbiomes. I created a resource called the Human Microbiome Drug Metabolism (HMDM) database. I analyzed 1,196 unpublished sequenced gut bacterial genomes from 8 healthy adult donors to predict genes that encode enzymes capable of metabolizing drugs using two *in silico* methods I developed, namely MAGIS and AutoPhylo. I reviewed the scientific literature and built an ontology-centric database, the HMDM, to catalog the bacterial drug-metabolizing genes and drugs they modify. I developed DME software to predict bacterial genes capable of metabolizing host-directed drugs using the HMDM data. We experimentally validated four novel AMR gene homologs predicted from the genomes. The HMDM is curated with 50 genes reported to metabolize drugs and 45 gene variants of the β-glucuronidase (*uidA*) gene. MAGIS was used to predict 246 putative bacterial drug-metabolizing genes. I predicted the three novel AMR gene homologs that resemble fosfomycin thiol transferase enzymes using AutoPhylo. The MIC experiment shows that *fosD1*, *fosD2*, and *fosD3* have MIC of 8μg/mL, 8μg/mL, and >512μg/mL, respectively. The genes *fosD1* and *fosD2* are of unknown

function, and FosD3 converts fosfomycin. The HMDM database is limited to bacterial genes. The *in silico* methods are critical for studying bacterial drug metabolism to predict drug fate and patient outcomes.

Acknowledgements

I would like to thank my supervisor Dr. Andrew G. McArthur, for taking me under his wing and taking a considerable risk to help guide a computer guy to be a well-rounded bioinformatician. I thank my wife every day for convincing me to meet with you for an interview in 2015.

I thank my committee members, Dr. Michael G. Surette and Dr. Gerard D. Wright, for their guidance. We did great things together, as Gerry put it, and I'm proud of the work done to date. From the Wright Lab, thanks to Samini H.R Kank, Dr. Adam Schaenzer, Dr. Akosiererem Sokaribo, and Dr. Michael Cook for performing the validation experiments, and Dr. Emily Bordeleau for introduction to PyMol.

From the Andres Lab, Dr. Sara Andres, for helping me with protein structure predictions which didn't make the final cut, but I'm sure we will collaborate in the future as this is much-needed research to uncover protein functions.

Dr. Nicholas Waglechner, for supporting my wild idea of automated phylogenetic tree generation, and I'm glad to say that it works!

I would also like to thank the volunteers for this project, Mugdha Dave, Akash Mehta, Nawal Masood, Michaela Hughes-Butler, and Mahrukh Khan, for taking their precious time to learn about bacterial drug metabolism.

Thank you to the Weston family foundation for funding this project, and I hope you continue to support basic science.

My McArthur lab mates, past and present, thank you for your continued support, and a special thanks to Brian P. Alcock, Dr. Kara K. Tsang, Arman Edalatmand, and Martins Oloni. I still think Joya has the best sushi in Hamilton, and of course, Will Ferrell is the funny man!

To my siblings Ditso, Mary, Wantlha, Mojaboswa, and Moeteledi, thank you for your love and support. Dad will have been proud. Ditso, I still remember you reading me Tom Sawyer and thank you. To my mom, Ntswelepelo, thank you for caring for me and loving me. To Thato, Bakang, Ella, and Lerei, your uncle is very sad that he is far away from seeing you grow, but I'm only a call away.

To my mahal, Claudine Raphenya, thank you for caring for our three kids and me. Thank you to my mother-in-law Virginia Sumadsad and Frederick Poland for their continued support. Henri, Harrison, and Halle, your tata loves you.

Contents

List of Tables

List of Figures

Declaration of Academic Achievement

I, Amogelang R. Raphenya, declare that this thesis titled, *The Human Microbiome Drug Metabolism Database* and the work presented in it are my own. I confirm that:

I did most of the research and the writting, except where indicated in the preface of each chapter.

Chapter 1

Introduction

1.1 Drug Usage

Drugs treat, manage, and prevent a variety of medical conditions (e.g., headache, Parkinson's disease, inflammatory bowel disease, etc.) [1]. Patients use multiple routes to introduce drugs into the body, such as topical (skin), enteral (oral), and parenteral (intravascular) [2]. Oral administration is the preferred method due to ease of use [2]. Oral drugs pass through the human gut epithelial cells into the bloodstream to reach target organs. Orally administered drugs encounter variable conditions in the human gut before reaching their target, which includes high acid content, variable absorption rates, and bacterial enzymes [3,4]. The human gut has trillions of microorganisms that express a variety of enzymes that may alter drugs, leading to poor drug efficacy (i.e., rendering drugs ineffective) and toxicity [4]. Patients receive a specific drug dose (i.e., the amount of drug to use) based on weight [5]. Endogenous substances or enzymes that might interfere with the drug can affect the effective drug dose [5]. For example, patients are administered

Levodopa with an inhibitor that blocks the human tyrosine decarboxylase enzyme from interacting with the drug. The inhibitor ensures that Levodopa is at an effective dose, but the bacterial tyrosine decarboxylase enzyme in the human gut (if present) can still interact with the drug, leading to lower levels of effective drug. As such, a higher dose is needed to overcome this underdose due to the bacterial enzyme [5].

1.2 Drug Metabolism

A chemical change to a drug after administration can lead to changes in effective dose and undesirable outcomes. Drug metabolism is the biotransformation of endogenous and exogenous compounds by making them more polar, which helps their elimination from the body [6]. Drug metabolism can be performed by various systems in the human body (liver, intestines, and others) and there are three phases (i.e., phase I, phase II, and phase III). Phase I involves drug modifications by reduction, oxidation, and hydrolysis processes. Phase II processes, such as glucuronidation and methylation, combine molecules and prepares them for excretion by making them water soluble and not active [7]. Phase III processes primarily prepare drugs for excretion by further modifications, such as using adenosine triphosphate (ATP) to facilitate the pumping of drugs out of cells via efflux pumps. Various enzymes facilitate Phase I, II, and III processes in human cells and within bacteria.

The term microbiota refers to microorganisms which include viruses, fungi, and bacteria within a particular environment [8]. For this thesis, the microbiota I mainly refer to is the collection of bacteria in the human gastrointestinal

(gut) and their genomes (i.e., microbiome) and how they metabolize orally administered drugs. The human gut microbiota can degrade or modify oral pharmaceuticals [9], and the products of these modifications can have harmful or beneficial effects on the human body [8,10]. As a result, there is a need to study drug metabolism performed by the gut microbiota to understand better and predict adverse outcomes. As there has been more research emphasis on drug metabolism by human enzymes than microbial metabolism, the main goal of my thesis is to develop bioinformatic resources that other researchers can use to study gut microbial drug metabolism. The bioinformatic resource will allow for the prediction of drug-metabolizing genes in the human gut microbiome. The resource will give a sense of how prevalent or rare drug-metabolizing genes are and could ultimately lead to personalized medicine approaches to the application of therapeutic drugs.

1.3 Bacterial Drug Metabolism

While much is still unknown about the underlying mechanisms of bacterial drug metabolism, I will highlight studies by various groups to understand where bacterial drug metabolism can lead to unwelcome outcomes [11].

1.3.1 Drug activation and inactivation by diazo-reductase and NAT

Crouwel and colleagues conducted a literature review and found four drugs used to treat inflammatory bowel disease (IBD) are subjected to microbial metabolism: mesalazines, methotrexate, glucocorticoids, and thioguanine. The azo-bonds

(double-bonded nitrogen group) in sulfasalazine, balsalazide, and olsalazine are broken by bacterial azo-reductase enzymes in the human gut, releasing an active moiety 5-aminosalicylic acid (5-ASA or mesalazine). This is an example where effective treatment directly depends upon the microbiota, as gut bacterial enzymes are required to produce activated 5-ASA to effectively treat inflammation [7,8,12]. Yet, 5-ASA was also found to be inactivated via acetylation by bacterial N-acetyltransferase (NAT) enzymes in the human gut as well as by epithelial NATs from the human host [13] (see Figure 1.1). NATs transfer the acetyl group from acetyl coenzyme A (AcCoA) to nitrogen or oxygen atoms, and bacterial NATs have been associated with adverse side effects (e.g., pancreatitis, hepatitis, and renal toxicity) for 5-ASA use [14]. As such, an effective dose of 5-ASA is a balance between azo-reductase and N-acetyltransferase activity. There is an ongoing effort to use bacterial azo-reductase mechanisms as a delivery method for more therapeutic drugs [15–17].

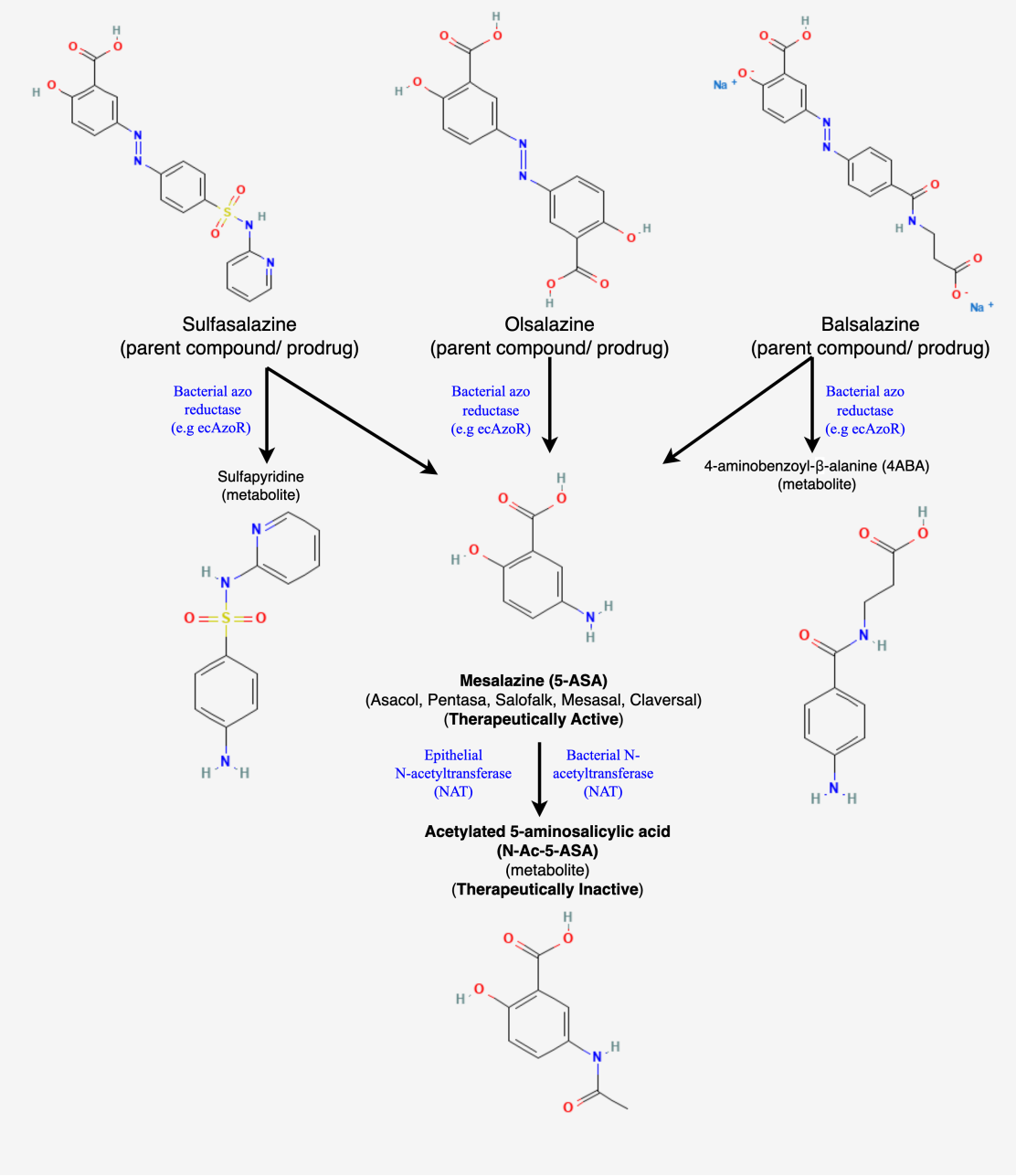


FIGURE 1.1: Drugs metabolized by bacterial azo-reductases releasing 5-ASA and in-activated by NATs enzymes to form N-Ac-5-ASA. (Figure adapted from Sandborn *et al.*[13]).

1.3.2 Bacterial Cytochrome P450s

In human cells, cytochrome P450 (CYP450) enzymes have been shown to metabolize many drugs. In fact, six CYP450 enzymes are estimated to metabolize 90% of all administered drugs [18]. Among patients of different ethnic groups, polymorphisms in CYP450 genes have led to variable drug responses. Since the 1960s, there has also been an increase in studies uncovering a diversity of bacterial CYP450s [19–23]. Overall, there are fewer CYP450s in bacteria than in the human liver, especially in the intestines [24]. To date, it is unclear if gut bacterial P450s are involved in drug metabolism similar to their counterparts in human liver cells.

1.3.3 Tyrosine Decarboxylase

The tyrosine decarboxylase (TDC) enzyme from commensal *Enterococcus* and *Lactobacillus* activates the important Parkinson's disease (PD) drug Levodopa outside the brain [5,25]. Levodopa is needed in the brain, where it is converted to dopamine to restore dopamine levels for PD patients. The hallmark of PD is impaired neurons due to less dopamine leading to uncontrollable movements. Levodopa is susceptible to breakdown by human decarboxylase, so it is administered with a tyrosine decarboxylase inhibitor [5], e.g., benserazide. Yet, the bacterial tyrosine decarboxylase is not inhibited by this inhibitor, and some PD patients need a higher dose of Levodopa due to the abundance of TDC genes in the human gut [5]. We need to understand how common this phenomenon is, i.e., can we predict bacterial enzymes using the same substrates as human enzymes?

1.3.4 Cardiac glycoside reductase

The cardiac glycosides digoxin, digixon, and digoxigen are reduced by the human gut bacterium *Eggerthella lenta*, leading to their inactivation [26,27]. Cardiac glycosides are used to treat heart conditions (e.g., atrial fibrillation), and their mechanism of action is to help increase heart contraction, improving blood flow [28]. There is variability among individuals; one study found that conversion happens in ~10% of patients [27], and others have found conversion in more than 40% of patients [11,29]. The products of a two-gene operon cardiac glycoside reductase (*cgr*) are responsible for this inactivation, namely genes *cgr1* and *cgr2* [11]. Cgr1 is anchored to the cell membrane and helps with electron transfer, while the Cgr2 protein contains an oxygen-sensitive cluster that transfers electrons to the flavin adenine dinucleotide (FAD) molecule producing $FADH^-$ which in turn reduces glycosides [11,29] (see Figure 1.2). Multiple strains of *Eggerthella lenta* (i.e., DSM2243, 11c, DSM11767, CC86D54, AB12n2, AB8n2, 326IFAA, and DSM18163) are capable of metabolizing digoxin. Koppel and colleagues speculated that the *cgr* operon protects the host from plant toxins, as digoxin originates from plants, and no benefits were observed for *cgr+ Eggerthella lenta* strains [11,30].

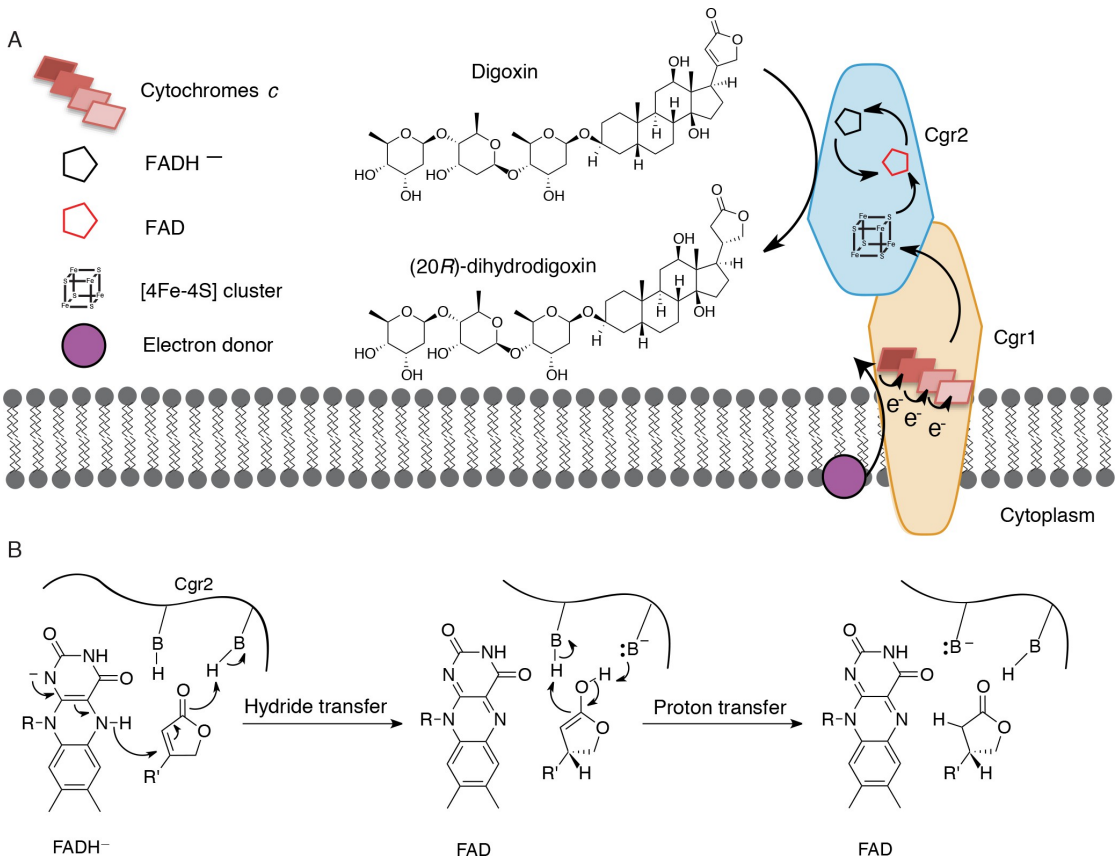


FIGURE 1.2: (A) The interaction between Cgr1 and Cgr2 in the conversion of Digoxin. (B) The reduction mechanism by the Cgr proteins. (Figure reproduced without modification from Koppel *et al.* [30]).

1.3.5 β-glucuronidase

The prodrug CPT-11 (irinotecan) is given intravenously and is converted to an active form, SN-38, by human carboxylases in the liver. SN-38 is then transported via the circulatory system to treat colorectal cancer by inhibiting the human topoisomerase I enzyme, which is responsible for breaking DNA (Deoxyribonucleic acid) strands. Ultimately, the SN-38 molecule is conjugated in the liver by human UDP-glucuronosyltransferase (UGT) enzymes to SN-38G, which is then transported to the intestines for excretion. Commensal bacteria (e.g., *E. coli*) in the intestines can cleave the sugars (glucuronide) from SN-38G to use as a carbon source via bacterial enzymes called β-glucuronidases (GUS). This process releases SN-38 within the gut lumen, which induces dose-limiting diarrhea and changes effective dose of the active compound [31] (see Figure 1.3). Pollet and colleagues created an "Atlas" for GUS enzymes in the human gut microbiome totaling 3,013 unique proteins. They found 279 unique GUS proteins grouped into six structural categories with differing functional capabilities, and more studies are needed to elucidate functions [32].

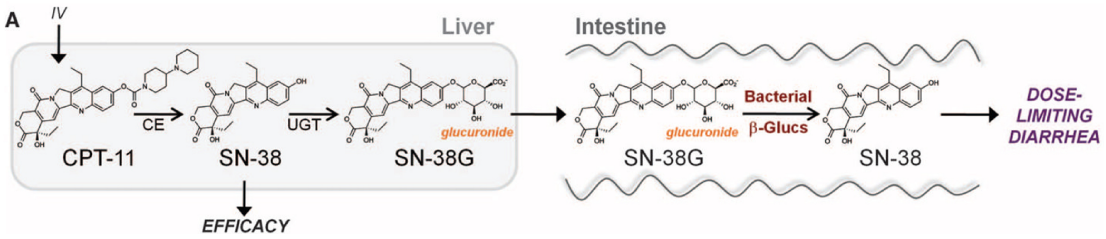


FIGURE 1.3: The drug SN-38, which is given through IV (Intravenously) as a prodrug (CPT-11), is converted back to an active drug in the gut by bacterial β-glucuronidase leading to diarrhea. (Figure reproduced without modification from Wallace *et al.* [31])

1.4 Next Generation Sequencing

The human gut comprises a complex and variable microbial community that performs various functions [33]. Bacteria constitute most microorganisms in the human gut [34]. To understand the various functions performed by bacteria in the gut, we need to be able to obtain the genetic materials from these bacteria, sequence their genomes, and perform predictive analysis. Fortunately, next-generation sequencing (NGS) can help us sequence many microbial genomes within the gut [35–37]. NGS is a technology for sequencing an organism's genome using a massively parallel process [38]. Multiple methods are used to obtain the bacterial DNA and sequence, including culture and culture-independent sequencing. The culture-independent approach (molecular method such as metagenomics) is defined as obtaining all DNA from a sample and sequencing [39]. Culture-dependent requires that the microorganisms be grown using defined conditions before obtaining the DNA for sequencing. Culturing ensures that the DNA obtained for sequencing is from live bacteria compared to the culture-independent method, which can include DNA from dead microorganisms [40,41]. Some of the bacteria in the human gut are hard to culture and require specific conditions to proliferate [42]. Cultures can also limit the organisms identified when using a high threshold, as low abundant organisms can be missed [42]. The opposite is true, and culture is much more sensitive than culture-independent methods see Lau *et al.* [43]. Lau and colleagues compared culture-enriched and culture-independent sequencing using multiple culturing conditions and concluded that culture enrichment allows for the identification of additional isolates [43]. Culture enrichment is different from conventional culture methods as it combines culture-based and molecular

methods, increasing sensitivity [43,44]. Other groups also were able to identify novel bacterial species by using multiple culturing conditions [40,41,45]. To date, most [35] of these data available from the human gut microbiome studies is culture-independent, i.e., metagenomes. The largest data sets are in the Human Microbiome Project [35,35] (HMP), MetaHIT [46], and NCBI databases. The MetaHIT cataloged 3.3 million non-redundant microbial genes from samples obtained from 124 individuals of European descent, totaling 576.7 gigabases of sequence data [46]. The HMP used a population of 242 healthy adults, up to 18 body sites sampled, and obtained 5,177 microbial taxonomy profiles. The total data generated for the HMP is 3.5 terabases of metagenomic sequences. Even with this wealth of sequencing data, for some organisms, we still find it hard to determine the potential functionality [42] or resolve taxonomic rank [47] for the obtained sequences. Nonetheless, these data sets can still help to identify possible bacterial function. Some of the poor functional characterizations can be due to the quality of reference data used for comparison [47], which is one of the issues my thesis attempts to resolve for drug metabolism by gut bacteria.

With sequencing information increasingly available for gut bacteria, it is essential that we build tools to use these bacterial sequenced data. We must build knowledge-based platforms using FAIR (Findable, Accessible, Interoperable, and Re-usable) principles to make the sequenced data reusable, standardized, and easy to use [48]. Good reference data sets are needed to gain new knowledge as more sequencing is performed. For building robust reference databases, ontologies (i.e., controlled vocabularies) can standardize these data by defining subject domains. For example, FoodOn [49] and CARD [50] use ontologies to organize

food and antimicrobial resistance concepts. Oxford Languages defines ontology as defining key concepts, relationships, and properties to describe a subject area, such as antimicrobial drug resistance and food. Building a bacterial drug metabolism resource using an ontology-centric approach ensures an easy expansion to incorporate new knowledge.

1.5 Resources

There are a few solutions developed to predict microbial drug-metabolism, which include the MASI (Microbiota Active Substance Interactions) database [51], PharmacoMicrobiomic [52], Disbiome [53], gutMDisorder [54], MagMD (Metabolic action of gut Microbiota to Drugs) [55], MicrobeFDT [56], and SIMMER (Similarity algorithms that Identify MicrobioMe Enzymatic Reactions) [57], and GutBug [58]. The MASI, PharmacoMicrobiomic, and gutMDisorder databases have summaries of gut microbiota metabolism of drugs based on published literature but provide no genomic or gene level information. The Disbiome database provides standardized microbiota linked to disease data, for example, *Akkermansia* and *Acinetobacter* are associated with Parkinson's Disease but no gene level information. The MASI and gutMDisorder databases are missing microbes implicated in drug metabolism, for example, MASI outlines that certain bacteria impact a particular substance, but no specific microbe is stated. For example, the nitrazepam entry in MASI (identifier PMDBD595) has interpretive information, e.g., "Transform drugs and phytochemicals into toxic metabolites" or "Nitroreductase," "Drug metabolism," or "Decrease Toxicity," but the microbe or specific enzymes are listed as unknown. Some resources are missing complete

annotations reported in the literature. The gutMDisorder is missing some diseases, for example, Parkinson's, as well as drugs used to manage this disease (i.e., Levodopa), and MagMD annotation for digoxin is shown as "unknown," which should be a "reduction" process. MicrobeFDT and MASI are limited to the Kyoto Encyclopedia of Genes and Genomes (KEGG) database enzymes, which does not cover the full diversity of bacterial metabolism. SIMMER is limited by reaction products and cofactors required to acquire high accuracy for the predictions. The GutBug database uses machine learning methods and reference databases to predict possible enzymes that might metabolize a given drug and provide the Enzyme Commission Number (EC numbers) for the predicted enzymes. Most of these resources provide valuable information for studying bacterial drug metabolism within the human gut but lack completeness in connecting the bacterial drug metabolizing genes to the drugs they transform. Frequent updates are needed to improve prediction, and some of these resources are not up to date with the last updates over two years ago.

1.6 Research Goals

The human gut microbiome's biotransformation of host-directed (i.e., non-antibiotics) drugs is poorly understood. Few studies provide evidence of specific drugs being metabolized to a particular metabolite by a specified bacterial enzyme. Most studies report a drug conversion with no enzyme or bacterium identified. Large drug panels and highly sensitive assays are required to fully understand the underlying mechanism of bacterial drug metabolism [59,60]. Comparative genomics could help us identify many bacterial drug-metabolizing genes that might

contribute to drug efficacy. Current resources need key information for studying bacterial drug metabolism in the human gut, such as gene level information, and some are incomplete. There are no genome-based annotation tools currently available. I hypothesize that a well-curated, experimentally verified informatics platform, the Human Microbiome Drug Metabolism Database (HMDM), can be developed. The HMDM can be used in genomic and metagenomic studies to predict bacterial drug metabolism prevalence and can be applied towards personalized approaches to treat various diseases.

In order to build the HMDM resource, my goal was to:

1. Analyze the sequenced genomes of over 1,200 human gut microbiome samples and identify potential drug-metabolizing genes using bioinformatics tools. Using various enzyme characteristics, baseline bioinformatic standards were to be developed to identify suitable drug metabolism candidates from raw genome sequences. Biochemical methods will then be used on the predicted genes to validate their activity. The candidate genes will be curated into the HMDM database after validation.

2. Systematically review the literature on bacterial drug metabolism in the human gut using manual searches and computer-aided software (i.e., text mining) to triage relevant literature. I will then use the knowledge gained from the literature reviews to design an ontology structure to describe current known bacterial drug metabolizing genes, their underlying mechanisms, and the host-directed drugs they transform within the human gut.

3. Develop curation tools in the form of a web interface to allow for the curation

of the literature into the HMDM database. The web tool will be built to be accessible, user-friendly, and easy to use. The latest and well-supported technologies will be used to build the HMDM database. Additionally, prediction tools for drug metabolism will be built, which use these data curated into the HMDM and make predictions on new bacterial genomes from the human gut. The main goal of this thesis was to build a resource, the HMDM, with comprehensive knowledge of bacterial drug metabolism in the human gut and to be a good reference for bacterial drug metabolism studies.

Chapter 2

Using *in silico* methods to predict putative microbial drug-metabolizing genes

2.1 Chapter 2 Preface

The biochemical experiments in this chapter were conducted by our collaborators from the Wright Lab, as outlined below.

Amogelang R. Raphenya[a,b,c] , Akosiererem Sokaribo[a,b,c], Michael Cook[a,b,c], Samini H. R. Kank[a,b,c], Adam J. Schaenzer[a,b,c], Michael G. Surette[a,b,c], Gerard D. Wright[a,b,c], Andrew G. McArthur[a,b,c]

Author contributions: ARR, MGS, GDW, and AGM conceived the project. AS, MC, SHRK, and AJS performed the biochemical experiments. MGS provided

the human gut microbiome genome sequences. ARR designed bioinformatics experiments, performed the analysis, and wrote the chapter.

[a] Michael G. DeGroote Institute for Infectious Disease Research, McMaster University, Hamilton, Ontario, Canada

[b] Department of Biochemistry and Biomedical Sciences, McMaster University, Hamilton, Ontario, Canada

[c] David Braley Centre for Antibiotic Discovery, McMaster University, Hamilton, Ontario, Canada

2.2 Abstract

2.2.1 Objective

The objective of this study was to develop *in silico* methods that can be used to predict novel bacterial drug-metabolizing enzymes or homologs to antimicrobial resistance (AMR) genes that may metabolize non-antibiotic drugs within the human gastrointestinal (gut) microbiome.

2.2.2 Methods

We built a custom snakemake bioinformatics pipeline (MAGIS; in latin which means “more”) to predict genes encoding drug-inactivating enzymes within the human gut microbiome and built a custom tool, AutoPhylo (an Automatic Phylogenetic Pipeline), to compare these putative genes to known genes using

phylogenetic trees. The predicted homolog genes *fosD1*, *fosD2*, *fosD3*, and *rphB2* were tested for function using biochemical assays.

2.2.3 Results

Based on gene discovery and phylogenetic results, we predicted one fosfomycin inactivation gene (*fosD3*), one rifampicin inactivation gene (*rphB2*), and two homologs of unknown function (*fosD1* and *fosD2*). Both FosD1 and FosD2 have minimum inhibitory concentration (MIC) of 8μg/mL against fosfomycin, FosD3 has MIC of >512μg/mL against fosfomycin, and RphB2 has MIC of >4μg/mL against rifampin. The three fosD gene homologs were predicted from the nine genomes of the commensal gut bacterium *Staphylococcus saprophyticus* and *rphB2* was from the genera *Bacillus* and *Peribacillus*.

2.2.4 Conclusions

The *in silico* methods MAGIS and AutoPhylo were used to predict genes that encode putative drug-metabolizing enzymes. The biochemical experiments (e.g., MIC experiment) can be complex, so these *in silico* methods are essential in narrowing down the search for putative bacterial drug-metabolizing genes in the human gut microbiome before embarking on biochemical tests. We predicted and validated three homologs to fos genes and one *rphB* homolog. The enzyme encoded by *fosD3* gene metabolizes fosfomycin, but more studies are required to determine the function of two fos homologs (*fosD1* and *fosD2*). The gene *rphB2* encodes an enzyme that metabolizes rifampicin and has the same architecture as seven other

putative *rphs* in the genomes analyzed for this study. The *fosD1* and *fosD2* are the positive results for this study, and *rphB2* and *fosD3* are negative results.

2.3 Introduction

The human microbiota are capable of decomposing or modifying xenobiotics (i.e., externally administered drugs), regulating host gene expression, and modulating xenobiotic absorption [61]. These phenomena have been documented since the 1900s [9]. Yet, since everyone is colonized by different gut microbes, there are heterogeneous responses to therapeutics among individuals [60,62]. Microbial drug metabolism is one driver behind increased rates of adverse reactions in older adults, further compounded by comorbidities [63,64]. Oral drug administration is a widely used and preferred patient method [1]. Yet, drugs taken orally have many limitations, such as the inability to reach their target due to variable absorption rates, variable concentrations, high acid content, and the action of many digestive enzymes. Drug development and clinical trials are costly, so there is a need to understand microbial drug metabolism as it can reduce time and resources during the drug development process by avoiding adverse reactions or treatment failure by way of the gut microbiome. To avoid these shortcomings, *in silico* methods based on genomic data could be used to predict drug metabolism for a given drug before the drug is placed in clinical trials [65].

2.3.1 Study rationale

In order to set up a method to predict bacterial drug metabolizing genes from the human microbiome, we used homologs of antibiotic resistance proteins and

non-AMR open reading frames as proof of principle. Gene homologs can diverge due to relative position from replication origin in the bacterial chromosome, and the further away genes are from the origin, the more non-synonymous mutations accumulate [66]. Furthermore, genome repair can occur at different rates on the chromosome, which can lead to more mutations that can alter genes and enzyme function [66]. Enzymes can have multiple substrates, and mutations in the active site can be selected to act on non-native substrates [67,68]. Other mechanisms include the switching of metal cofactors which can lead to new functions [69,70]. By using homologs to AMR genes, we hoped this strategy could yield enzymes with different substrates, additionally, the human gut is one of the reservoirs for AMR genes [71,72]. To find possible functions of predicted non-AMR genes, we can use the Pfam functional domain database because the sequences have been grouped based on a shared common ancestor [73]. Even though most of these grouped sequences, i.e. homologs, share a common ancestor, they may have different functions [73]. We can use the domains and identify different architectures as a hint of probable alternate functions. In this study, we predicted three homologs to the *fosD* gene, which codes for an enzyme that inactivates the fosfomycin antibiotic. One homolog (*fosD3*) codes for a fosfomycin inactivation enzyme, and two homologs are of unknown function (*fosD1* & *fosD2*). We used a combination of *in silico* and biochemical methods to predict and verify the activity of these enzymes. One of the *in silico* tools we developed uses phylogenetic trees (described below) to tease out the relationships between our predicted proteins to known proteins.

2.3.2 Phylogenetic trees

In biology, phylogenetic trees are used to understand the changes observed in genes over time and determine their evolutionary relationships [74]. Homologous sequences trace back to ancestral sequences because they share a common ancestor. Phylogenetic trees can be used to reveal the relationship between sequences [74] and infer the probable function of new sequences given the information available for known sequences. Phylogenetic trees can be generated using various sequence information, including gene sequences or the whole genome for an organism. There are several methods for constructing phylogenetic trees with different advantages and disadvantages; for this work, we used the maximum likelihood (ML) and neighbor-joining (NJ) algorithms. A NJ tree is built using distances for taxa under investigation by comparing each pair, building a subtree. The next step is reducing the taxon set by minimizing the criterion denoted by formula (1) from Gascuel and Steel [75]. ML is a method of searching for a population parameter that maximizes the probability for a given data set [76]. In order to estimate ML trees, assumptions are made for each data set. These assumptions are described by Whelan and Morrison [74], briefly highlighted as follows: 1) The multiple sequence alignments are assumed to be homologous that are inherited from a common ancestor, for example, via substitution or duplication. This is also true for NJ or any other tree-generation algorithms. 2) Estimating how likely a set of observed data occurred for a given substitution model by applying statistical method function. 3) Sequence mutations that have occurred in a population and are now fixed, which can happen due to random chance, adaptation, or positive selection. 4) Statistical models called substitution models are used to describe sequence evolution. The rate of substitution for both amino acids and nucleotides

is taken into account. 5) A heuristic method is used to search for the best tree. Each tree assessed for topology is scored and the highest-scoring tree is selected. The ML methods for this work use bootstrap values as a confidence value for the trees generated, i.e., how well the branching patterns are supported by the data. Traditionally, researchers have relied on a manual process of phylogenetic tree construction [77], which includes selecting sequences to generate alignment and removing regions with uncertain homology after visual inspection. In this study, we developed an automated phylogenetic tree-building pipeline called AutoPhylo used to predict genes encoding novel enzymes that may be involved in drug metabolism. AutoPhylo assesses how each predicted gene relates to a known drug-inactivating gene. AutoPhylo uses gene sequences found in common bacterial phyla in the human gut.

2.4 Methods

2.4.1 Genomes

We analyzed 1,196 unpublished and assembled human microbiome genomes provided by Dr. Michael G. Surette of McMaster University. This data set contains genotypic information from isolates obtained from eight healthy adults with no recent history of using antibiotics. The isolates are from human stool and respiratory tract and were cultured using conditions described by Lau *et al.* [43] and Sibley *et al.* [44], respectively. The genomes were sequenced using Illumina, and some were sequenced with Pacbio [78,79]. Library construction methods are described in Derakhshani *et al.* [80]. Unicycler (version v0.5.0) [81] was used to assemble all the genomes for both short reads and for hybrid assemblies. PROKKA

(version 1.14.6) [82] and bakta (version 1.5.0) [83] were used for gene prediction and annotations.

2.4.2 Genome quality assessment

The quality of the 1,196 unpublished assembled bacterial genomes was assessed using assembly-stats (version 1.0.1; https://github.com/sanger-pathogens/assembly-stats). Genome completeness was calculated using CheckM (version v1.2.0) [84] and taxonomic classification of sequences was predicted using GTDBTk (version v1.7.0) [85].

2.4.3 Prediction of putative drug-metabolizing genes using MAGIS

A custom snakemake pipeline termed MAGIS (version 1.0.0) (Figure 2.1) was created to analyze the 1,196 bacterial genomes using the Comprehensive Antibiotic Resistance Database's (CARD [86]; version 3.1.1) software tool Resistance Gene Identifier (RGI; version 5.1.1). RGI was used to predict homologs to antimicrobial resistance (AMR) genes but likely to metabolize other small molecules. RGI yields results with three major labels: Perfect, Strict, or Loose annotations [86]. Perfect annotations match reference AMR protein sequences in CARD. Strict annotations fall above a manually curated similarity bitscore cut-off and are generally functional AMR variants. Loose annotations fall below the manually curated bitscore and are either novel AMR genes, distant homologs of known AMR genes, or spurious matches. We first examined 'Loose' annotations predicted by RGI and predicted functional domains within the encoded proteins (described

below). We implemented five bioinformatic filters described below to predict putative homologs of AMR genes that may metabolize non-antibiotic molecules.

RGI uses predicted open reading frames (ORFs) via Prodigal (version v2.6.3) [87] and uses CARD reference data to annotate AMR-associated ORFs in bacterial genomes. Most of the ORFs predicted by Prodigal are not examined by RGI as they are dissimilar to what is curated into CARD, i.e. encode proteins uninvolved in AMR, while all others are annotated under the Perfect, Strict, Loose paradigm outlined above. To predict additional putative drug-metabolizing genes, we secondarily examined ORFs via Prodigal that were unused by RGI (i.e., non-AMR related ORFs). For both RGI Loose annotations and the unused ORFs, functional domains were predicted using Pfam [88] data (version 34.0), pfam_scan (conda version 1.6), InterPro [89] (version 81.0), and Resfams [90] (version v1.2.2). Pfam is a database containing protein functional domains built by profiling multiple sequences using hidden Markov models [91] (HMMs). The Pfam entries are refined to ensure no overlaps between protein families, reducing false positives. InterPro combines protein signatures from 13 databases to make query protein or nucleotide sequence predictions. Resfams predict domains based on HMM profiles built using AMR reference databases, including CARD, Lactamase Engineering Database (LacED; `http://www.laced.uni-stuttgart.de`), and Jacoby and Bush's collection (`http://www.lahey.org/Studies`). These methods allow the assessment of functional domains encoded within each putative gene.

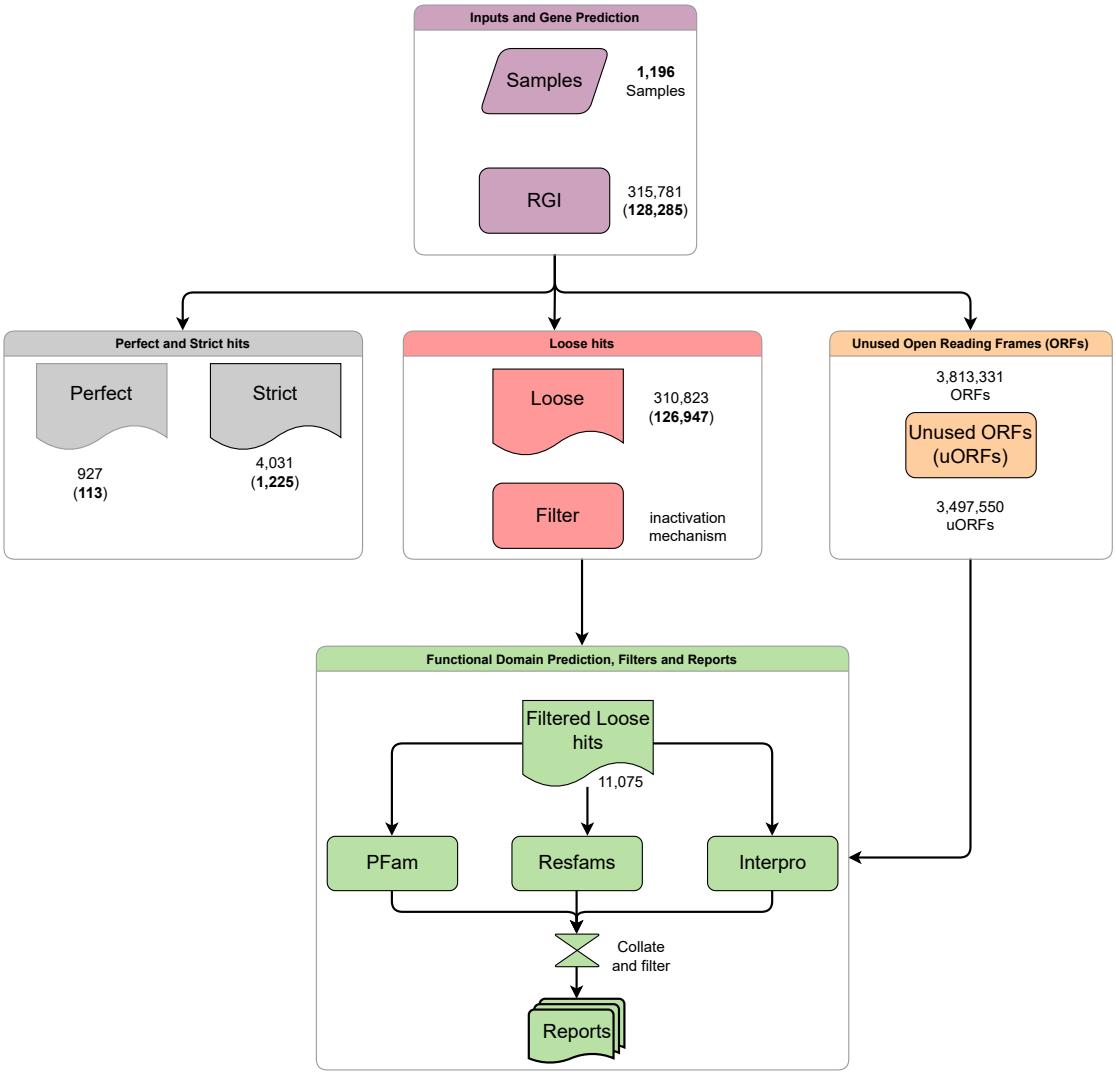


FIGURE 2.1: The workflow diagram for the snakemake pipeline (MAGIS) was used to analyze 1,196 bacterial genomes using the Comprehensive Antibiotic Resistance Database's (CARD) Resistance Gene Identifier (RGI) to predict potential enzymes similar to AMR genes but likely to metabolize other small molecules (red box). MAGIS also uses unused ORFs (orange box) from RGI's Prodigal step to predict drug metabolizing genes (i.e., not AMR homologs). The workflow diagram shows the tools used and the results in each step. The numbers in brackets represent unique protein sequences and others are total protein sequences. The Perfect and Strict annotations (in grey box) were not used as they are dedicated AMR genes. Functional domains are predicted for filtered Loose annotations and unused ORFs using Pfam, Resfams, and Interpro (in green box).

2.4.4 Bioinformatic filters used by MAGIS

We implemented five bioinformatic filters to group related predicted proteins that might have similar functions. A bioinformatic filter is a computer method to sort results, for example, sorting proteins by a "transferases" annotation. The five filters are described below:

Filter 1

The criteria for the first filter are Loose annotations from RGI with antibiotic inactivation mechanism, "transferases", percentage length of reference greater than 80%, and percent identities above 75%.

Filter 2

We used the "unused ORFs" via Prodigal and selected genes annotated as "transferases" by both Pfam and InterPro. Only sample GC1078 was used for this filter as a pilot case.

Filter 3

For this filter, RGI Loose annotations labeled for antibiotic inactivation that also included domain of unknown function (DUF) annotations coupled with other domains were used. DUFs are putative functional domains that have not been characterized or defined in the Pfam database. Pfam groups proteins by family and DUFs contains a group of similar proteins, none of which have been characterized [92]. The idea behind this filter was to find new AMR homologs with additional, novel functional domains that might have activity toward non-antibiotics.

Filter 4

Using the active site predictions (by Pfam) on the RGI Loose annotations as the first-pass filter, we created two separate lists of annotations: those with known active sites and those without known active sites, with the goal of investigating the latter. Our rationale was that the predictions without active sites will yield new AMR homologs with polymorphisms or mutations leading to new function.

Filter 5

For filter 2 we used one sample as a test case. We expanded this to include unused ORFs for all 1,196 samples and predicted functional domains looking for different families associated with drug metabolism, for example, including domains for cytochrome P450 (CYP450).

2.4.5 The Automatic Phylogenetic Pipeline (AutoPhylo)

AutoPhylo (version 1.0.0) is divided into six modular sections with quality checks at each stage (see Figure 2.2). This pipeline has been specialized for bacteria found in the human gut. AutoPhylo uses a user-submitted protein sequence in FASTA format for a putative enzyme to search for homologous sequences from the pre-calculated NCBI [93] nr (National Center for Biotechnology Information non-redundant) database using BLASTP. Sampling is performed from eight commonly found phyla in the human gut, namely *Actinobacteriota*, *Bacteroidota*, *Desulfobacterota*, *Firmicutes*, *Fusobacteria*, *Proteobacteria*, *Synergistota*, and *Verrucomicrobiota*. The sampled protein sequences are aligned using alignment software MUSCLE v5 [94] with default options and automatic mode. Alignment

trimming is performed to remove gaps and non-homologous columns in the alignment using trimAI [95] in two steps. The first step is to trim using the automatic option *-automated1*, which uses the heuristic method to obtain an optimal alignment. As a result, this step removes columns with poor alignment or uncertain homology. The second step is to remove sequences that are overly similar to each other and thus not informative for overall placement of homologs in evolutionary context, for example differing by two or three amino acids. For this step, the pairwise distance calculated by trimAI is used. The resulting alignment file is used thus to generate a NJ tree or a ML tree. Within AutoPhylo, the FastTree [96] software is used for generating a neighbor-joining (NJ) tree, and a publication-grade maximum likelihood (ML) tree is created using RAxML (version 8) [97]. FastTree was used with default parameters to generate an approximate tree and assess automatic tree generations options while building the pipeline. For RAxML, we used the option PROTCATIJTTF which specifies both GAMMA and JTT. The GAMMA option allows for different rates of changes among sites. The JTT amino acid substitution model is best suited for enzymes. The JTT model uses empirical frequencies allowing the substitution model to incorporate unequal frequencies of amino acids in the data.

FIGURE 2.2: The automatic phylogenetic pipeline (AutoPhylo). The AutoPhylo uses user submitted single or multiple proteins (purple box). The submitted user sequences are used to sample the NCBI database and obtain homologous sequences to the query sequences using BLASTp algorithm. Optionally, users can sample sequences using hmmer algorithm to sample proteins based on functional domains. The multiple sequence alignment is performed on the obtained sequences using the MUSCLE algorithm and the alignment is trimmed automatically using TrimAI or Gblocks (red box). The phylogenetic trees are built using Fasttree or RAxML (green box).

2.4.6 Determining genomic context for putative genes

To examine the genomic context of putative genes, we used PROKKA (version 1.14.6) [82] to annotate the genomes in their entirety and plotted the features using DNAPlotter (version v18.2.0) [98].

2.4.7 Testing AMR homologs for activity against antibiotics

RGI Loose annotations could possibly be novel AMR genes. As such, the function of these homologs were tested using the Antibiotic Resistance Platform (ARP) [99]. Briefly, gBlocks (Integrated DNA Technologies: IDT) for the putative AMR homologs were cloned into pGDP2 [99] linearized using restriction enzymes XhoI and NocI, and transformed into *E. coli* TOP10 cells. The clones were identified by colony PCR (polymerase chain reaction), validated by Sanger sequencing, and transformed into the hyperpermeable, efflux-deficient mutant strain *E. coli* BW25113 Δbam Δ*tolC* for the determination of the minimum inhibitory concentration (MIC [100]) for the applicable antibiotic. The MIC experiment was performed using 96-well round bottom plates (Sarstedt) and antibiotic concentrations were varied at 2, 4, 8, 16, 32, 64, 128, 256, and 512 μg/ml (fosfomycin) or 0.125, 0.25, 0.5, 1, 2, and 4 μg/ml (rifampin). The assays were incubated for 16 – 20 hours at 37°C either statically or with shaking at 200 rpm (revolutions per minute), and the wavelength at 600 nm was read using a BioTek synergy A1 microplate reader (BioTek). All assays were performed with at least three biological replicates. The positive controls used were AMR gene (*fosA* and *rox*) reference sequences in CARD, while for the negative control the

strain alone (*E. coli* BW25113 Δbam ΔtolC) was used (i.e., no vector).

2.4.8 Code Available

The code for both the AutoPhylo and MAGIS is available on GitHub at https://github.com/raphenya/autophylo.git and https://github.com/raphenya/magis.git, respectively (please note that the stated repositories are currently private and will be made public after publication of the corresponding manuscript).

2.5 Results

2.5.1 Genomes

We compared genome length with the number of contigs obtained from each of the 1,196 genome assemblies shown in Figure 2.3, and most genomes had less than 500 contigs, with few outliers. Using assembly-stats, the average number of contigs per genome for all 1,196 genomes was 86, with an average N50 of 521,426 bp and an average N50n of nine contigs. The quality of an assembled genome is evaluated using a N50 method which is determined by sorting all contigs from largest to smallest and picking the contig that sits at the 50% mark of the total assembly as a metric (Supplementary Figure 2.10) [101]. N50n is the number of contigs constituting the 50% of the genome underlying the N50 metric. Overall these are high quality assembled genomes and most of the genome taxonomy was predicted with high confidence using GTDBTk. The summary of organisms in the 1,196 genomes is shown in Figure 2.4, summarized by families. There was an average completeness of 99.27% for the 1,196 genomes but 11 genomes were flagged for high

contamination using the CheckM tool (Table 2.1). CheckM uses lineage-specific maker genes identified from reference genomes to estimate genome completeness and contamination. A genome has contamination if different species' marker genes are binned together.

FIGURE 2.3: Number of contigs and genome length for 1,195 examined bacterial genomes. One sample (GC1513) was omitted from the plot it has 6,584 contigs and a genome length of ~18M base pairs.

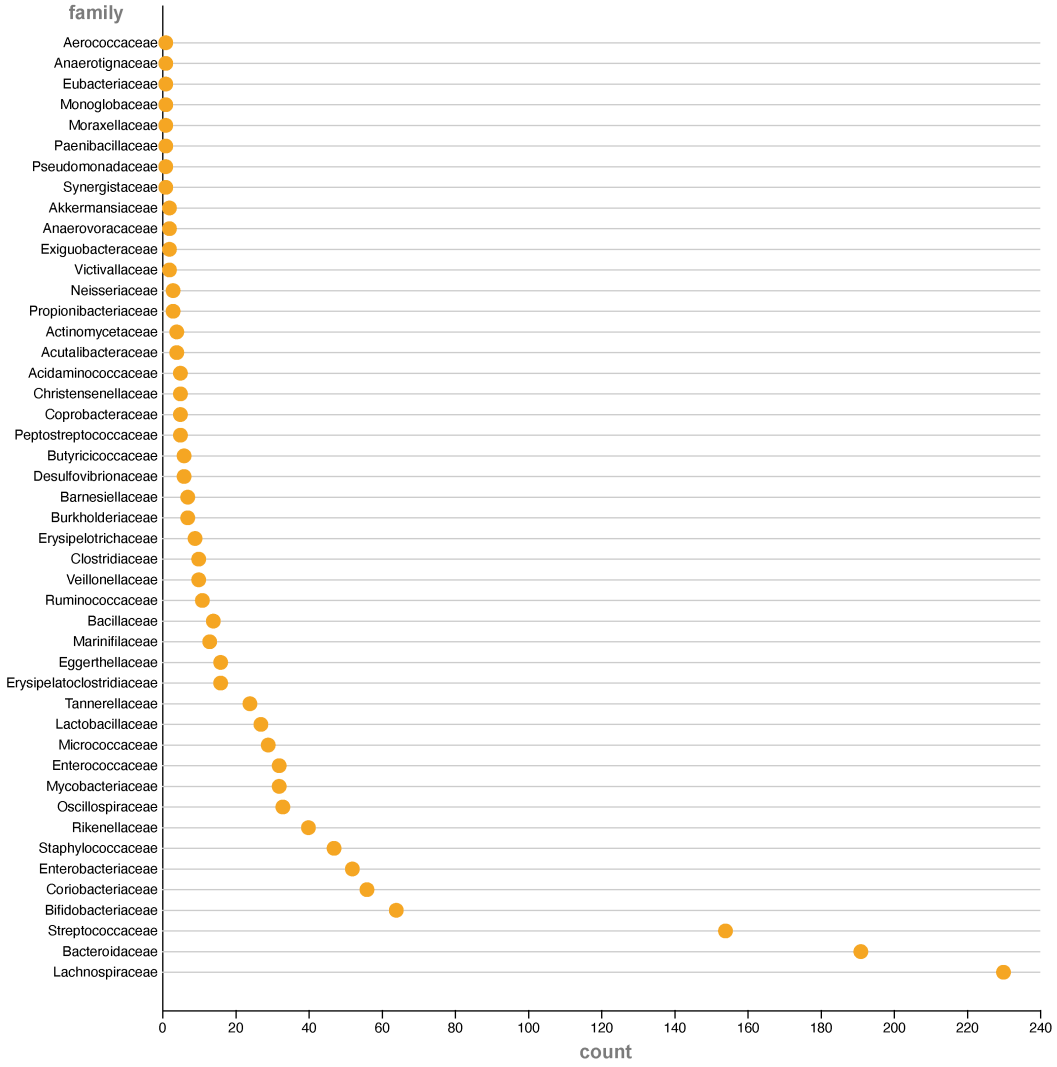


FIGURE 2.4: The summary for 1,187 out of 1,196 genomes grouped by families (9 genomes had no GTDBTk prediction see Table 2.1). The taxonomy was predicted using GTDBTk. The "count" represents the number of genomes in each bacterial family.

TABLE 2.1: CheckM analysis showing genome completeness for 11 samples are predicted to have high contamination (the 8/11 samples were predicted as Archaea by GTDBTk). "Sample" is the genome identifier, "Marker lineage" is the taxonomy call for the genome, "Completeness" is the proportion of maker genes identified per genome, and "Contamination" is the proportions of closely related species' marker genes binned together.

Sample	**Marker Linage**	**Completeness**	**Contamination**	**GTDBTk**
GC1035	Bacteria (UID203)	100.00	97.41	No prediction
GC1146	root (UID1)	100.00	100.00	No prediction
GC1002	root (UID1)	100.00	100.00	No prediction
GC15_hybrid_assembly	Bacteria (UID203)	98.28	92.52	No prediction
GC1188	root (UID1)	100.00	100.00	No prediction
GC1123	root (UID1)	100.00	100.00	No prediction
GC1513	root (UID1)	100.00	359.69	No prediction
GC75	root (UID1)	100.00	100.00	No prediction
GC1066	Bacteria (UID203)	100.00	93.10	No prediction
GC571	Bacteria (UID203)	98.28	82.13	*Staphylococcus warneri*
GC813	Bacteria (UID203)	100.00	52.38	*Phocaeicola vulgatus*

2.5.2 Putative drug-metabolizing genes predicted using MAGIS

MAGIS produced 12 annotations covering three drug classes (see Table 2.2) using filter 1 (i.e., RGI loose annotations). Three were putative fosfomycin-inactivating enzyme (*fosA*) homologs, eight were *rphB* (rifampin phosphotransferases family) homologs, and one a *vatB* (streptogramin *vat* acetyltransferase family) homolog.

MAGIS filter 2 produced 4,043 candidates. Two annotations were examined in detail (Table 2.2). The putative Arylamine N-acetyltransferase-like is similar to *Alicyclobacillus acidiphilus* and *Paenibacillus herberti* arylamine N-acetyltransferase from the NCBI nr database using BLAST and a uncharacterized protein (with accession T0BS57_ALIAG from *Alicyclobacillus acidoterrestris*) from UniProt. The other putative gene including a DUF domain has no characterization to date.

Filter 3 produced 28 annotations shown in Supplementary Table 2.5. There were 18 serine hydrolase homologs, three GCN5-related N-acetyltransferases (GNAT) family N-acetyltransferase homologs, five aminoglycoside nucleotidyltransferase (ANT) homologs, and two ole glycosyltransferase homologs. The putative acetyltransferases could participate in antibiotic inactivation of nucleosides, macrolides, fluoroquinolones, and aminoglycosides antibiotics. The DUF4111 domain is in both ANT-like putative predictions and CARD's ANTs reference sequences, which can be indicative of their possible function. This filter can be helpful to discover new serine hydrolases based on annotation of both the β-lactamase domain and the signal motif for secretion. Given that there are

many serine hydrolase homologs, we generated a phylogenetic tree for all of the serine hydrolases and used CARD hydrolase sequences for comparison. In the ML phylogenetic tree, the 18 serine hydrolyse homologs formed their own clade within the class C β-lactamases and are likely active upon beta-lactams (not shown).

Filter 4 produced 9,090 unique annotations without active site annotations. We focused on UDP glucoronosyl and UDP glucosyltransferase annotations as they have been shown to metabolize drugs, finding 50 unique sequences without active site predictions (Supplementary Table 2.6). These putative genes are annotated as macrolide, ole and rifampin glycosyltransferases family members. Using phylogeny, we predicted that the putative *Bacteroides* glycosyltransferases share a common ancestor with rifampin glycosyltransferases from *Streptomyces* and *Nocardia*. Both *Streptomyces* and *Nocardia* rifampin glycosyltransferases have been characterized by Spanogiannopoulos *et al.* [102], and Yazawa *et al.* [103], respectively. There is no publication on rifamycin glycosyltransferases in *Bacteroides*, which motivates us to follow up with these results in future studies.

For filter 5, we first searched for CYP450 annotations, which produced 154 unique bacterial candidates from 74 samples shown in Supplementary Table 2.7. The next steps were to compare these annotations to known CYPs and build a phylogenetic tree around these sequences. We downloaded annotated CYP sequences from the bacterial CYPs site (`https://drnelson.uthsc.edu/bacteria/`) and examined these alongside all 154 putative CYPs. We clustered all the sequences using UCLUST [104] with a percent identity of at least 90%. We selected CYPs that clustered with known drug metabolism CYPs. We then used the MEME [105] suite (version 5.4.1) to annotate the putative and identify key CYPs

motifs. Overall, we annotated 10 CYPs: one missing the heme motif (putative GC1084_2_119) annotated as CYP107A1, one annotated as CYP107A1, five unknown CYPs similar to Erythromycin 12 hydroxylase, two CYP106A1, and one CYP102A1. CYPs have four main motifs, namely I-Helix(I), K-Helix(K), Meander Coil(M), and Heme Loop(H). The I-helix is used for proton delivery while M and K stabilize the structure. The heme loop interacts with the heme cofactor in the active site, as seen in eukaryotes [106]. The bacterial CYPs use ferredoxin and ferredoxin reductases for electron donors in their active site [24]. All the nine putative CYPs have the four main CYPs motifs, i.e., I-Helix(I), K-Helix(K), Meander Coil(M), and Heme Loop(H), with the same arrangement: MIKH (see Supplementary Figure 2.11 and Supplementary Table 2.8).

After obtaining results from the five filters we performed further analysis on *rphBs* and *fos* gene homologs before selecting some candidates for validation experiments which are described below.

2.5.3 *Fos* homologs

We performed an AutoPhylo analysis on the three *fos* homologs to assess their evolutionary history as a means to gain insight into their possible function. The candidates *fosD1*, *fosD2*, and *fosD3* are in the same clade as CARD references *fosBs* and *fosDs*, which have fosfomycin thiol transferase activity (Figure 2.5). The genes *fosD1* and *fosD2* share common ancestry with 75% bootstrap value and jointly common ancestry with *fosD3* with bootstrap value of 95%, indicative of a recent event for sequence divergence. All three genes are closely related to *fosD* with bootstrap of 75%. We annotated the genomic context of *fosD1*, *fosD2*,

and *fosD3*. The features of one of the genomes (sample GC605) containing *fosD1* are plotted using DNAplotter in Figure 2.6. All the three genes are annotated within the same chromosome with *fosD1* and *fosD2* close together, suggestive of possible evolution of functional diversification. MIC experiments reveal *fosD3* as a functional homolog of *fosD*, while *fosD1* and *fosD2* do not metabolize fosfomycin (Table 2.4).

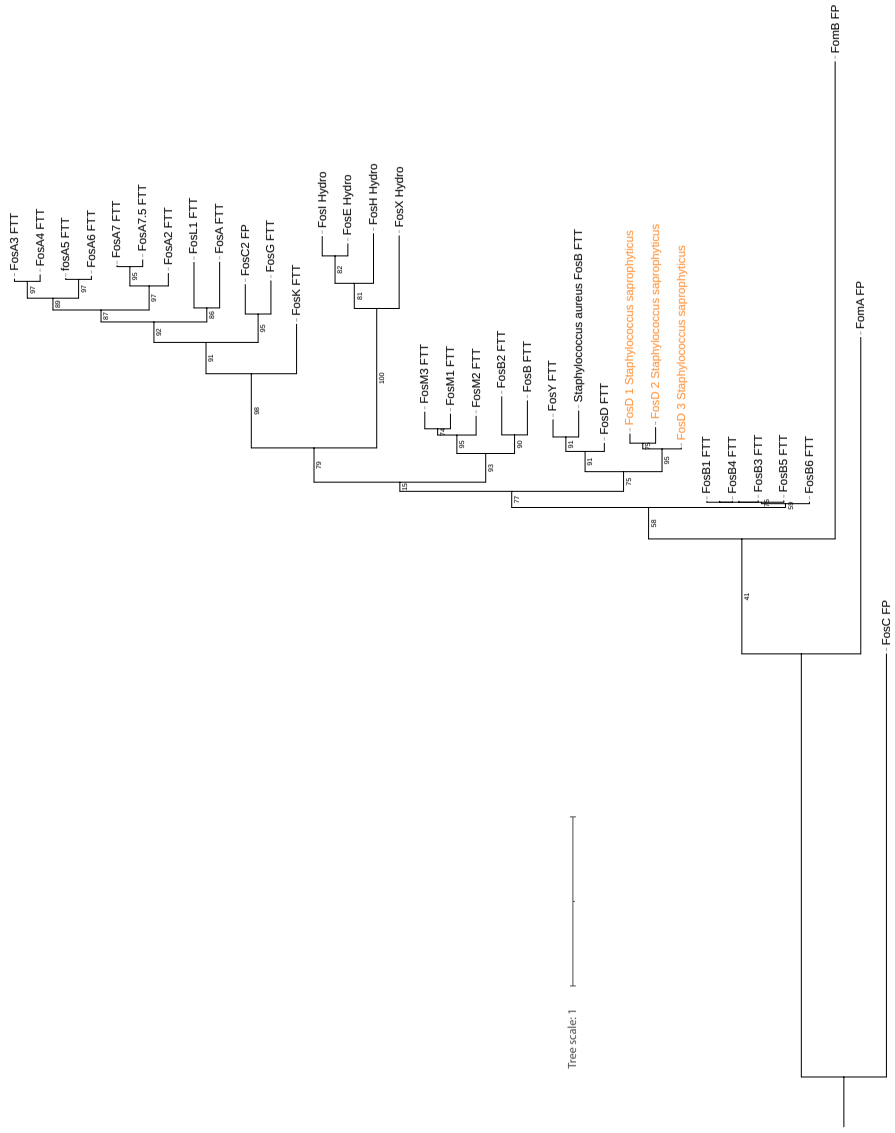


FIGURE 2.5: A phylogenetic tree generated using AutoPhylo for three *fosDs* homologs comparing all known fosfomycin resistance enzymes. (FTT = fosfomycin thiol transferase; FP = fosfomycin phosphotransferase; Hydro = fosfomycin resistance hydrolase). The homologs *fosD1*, *fosD2*, and *fosD3* are colored in orange.

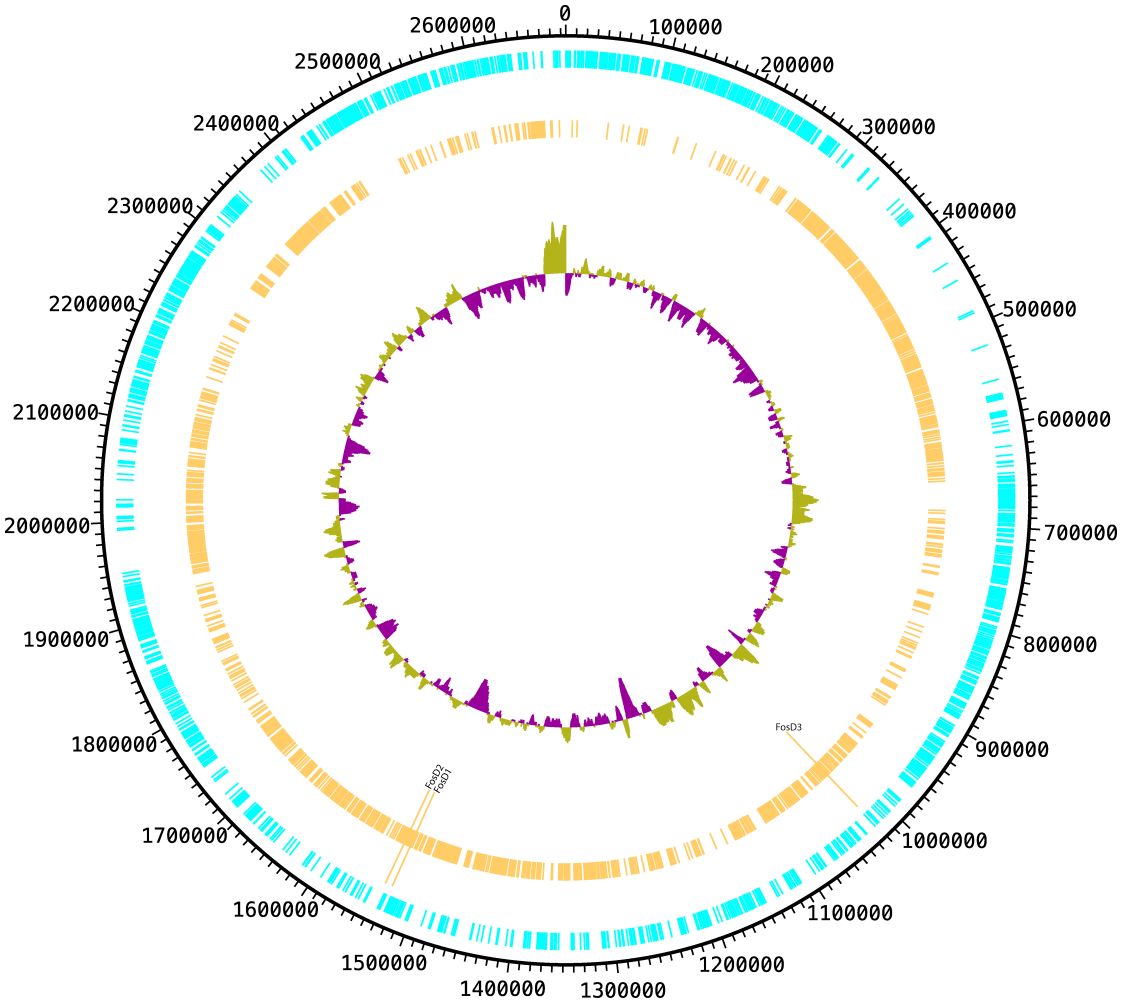


FIGURE 2.6: A plot showing *fosD1*, *fosD2*, and *fosD3* positions in genome GC605. All three putative are coded in the negative strand (gold) and have below-average GC content (purple). The *fosD1* and *fosD2* are closer to each other with coordinates 1528684..1529103 and 1536209..1536628, respectively. The *fosD3* has coordinates 1020747..1021166. The protein-coding genes are in blue color on the positive strand and gold on the negative strand.

TABLE 2.2: The MAGIS results for filters 1 (black) and 2 (red). The "Sample ID" is the identifier for the genome, "ID" is a unique identifier for each predicted protein, "PFAM Family" are annotations from Pfam, and "NCBI BLASTp [nr database]" are top scoring alignments from NCBI.

Samples	**ID**	**PFAM FAMILY**	**NCBI BLASTp NR**	**Percent Identity**
GC1078	vatB1	galactoside acetyltransferase-like	Vat family streptogramin A O-acetyltransferase [*Bacillus vini*]	91.43
GC1084	rphB1	Pyruvate phosphate dikinase, AMP/ATP-binding domain; PEP-utilising enzyme, mobile domain	phosphoenolpyruvate synthase [*Bacillus subtilis*]	100.00
GC1086	rphB2	Pyruvate phosphate dikinase, AMP/ATP-binding domain; PEP-utilising enzyme, mobile domain	MULTISPECIES: phosphoenolpyruvate synthase [*Bacillus*]	100.00
GC1160; GC76	rphB3	Pyruvate phosphate dikinase, AMP/ATP-binding domain; PEP-utilising enzyme, mobile domain	MULTISPECIES: phosphoenolpyruvate synthase [*Bacillus*]	100.00
GC380 hybrid assembly; GC381 hybrid assembly; GC709	rphB4	Pyruvate phosphate dikinase, AMP/ATP-binding domain; PEP-utilising enzyme, mobile domain	MULTISPECIES: phosphoenolpyruvate synthase [*Bacillus*]	100.00

GC602	rphB5	Pyruvate phosphate dikinase, AMP/ATP-binding domain; PEP-utilising enzyme, mobile domain	MULTISPECIES: phosphoenolpyruvate synthase [*Bacillus*]	99.65
GC793	rphB6	Pyruvate phosphate dikinase, AMP/ATP-binding domain; PEP-utilising enzyme, mobile domain	phosphoenolpyruvate synthase [*Bacillus licheniformis*]	100.00
GC873	rphB7	Pyruvate phosphate dikinase, AMP/ATP-binding domain; PEP-utilising enzyme, mobile domain	phosphoenolpyruvate synthase [*Paenibacillus maysiensis*]	88.29
WAC8344 hybrid assembly	rphB8	Pyruvate phosphate dikinase, AMP/ATP-binding domain; PEP-utilising enzyme, mobile domain	MULTISPECIES: phosphoenolpyruvate synthase [*Bacillaceae*]	100.00
GC1123; GC605; GC624; GC631; GC708; GC711; GC719; GC812; GC835	fosD1	Glyoxalase/Bleomycin resistance protein/Dioxygenase superfamily	MULTISPECIES: FosB/FosD family fosfomycin resistance bacillithiol transferase [*Staphylococcus*]	100.00
GC1123; GC605; GC624; GC631; GC708; GC711; GC719; GC812; GC835	fosD2	Glyoxalase/Bleomycin resistance protein/Dioxygenase superfamily	FosB/FosD family fosfomycin resistance bacillithiol transferase [*Staphylococcus saprophyticus*]	100.00

GC1123; GC605; GC624; GC631; GC708; GC711; GC719; GC812; GC835	fosD3	Glyoxalase/Bleomycin resistance protein/Dioxygenase superfamily	FosB/FosD family fosfomycin resistance bacillithiol transferase [*Staphylococcus saprophyticus*]	100.00
GC1078_6_97	N/A	Arylamine N-acetyltransferase	arylamine N-acetyltransferase [*Neobacillus cucumis*]	78.00
GC1078_19_22	N/A	Protein of unknown function DUF436 (PFam) ; Aminoglycoside 3-N-acetyltransferase-like (SUPERFAMILY); Aminoglycoside 3-N-acetyltransferase-like (Gene3D)	TIGR01440 family protein [*Bacillus yapensis*]	88.95

2.5.4 *rphBs* homologs

As an alternative to obtaining sequences for building phylogenetic trees, we built a custom HMM model using the eight putative *rphBs* and used the model to search for sequences in the UniProt database. We combined these with putative *rphB1* to generate an ML phylogenetic tree using AutoPhylo (Figure 2.7). The putative *rphB1* was in the same clade as a reviewed entry with accession PPS_BACSU on UniProt which describes a gene *pps* encoding a putative phosphoenolpyruvate synthase from *Bacillus subtilis* (strain 168). The putative *rphB1* is from *Bacillus subtilis* using annotation by the GTDBTk tool. The next annotation closely related to *rphB1* was gene *yvkC* which encodes uncharacterized phosphotransferase YvkC (accession YVKC_BACSU) from *Bacillus subtilis*. Using the UniProt ID mapping summary page for all 85 accessions revealed that the enzyme class was pyruvate kinase (18 results), histidine kinase (1 result), phosphotransferase with a nitrogenous group as acceptor (31 results), and phosphotransferase with paired acceptors (30 results). The taxonomic classification for all of these proteins was *Bacillus subtilis*. We performed sequence comparison for the eight putative *rphB* homologs and identified similar motifs and architecture described by Stogios *et al.* [107] and Spanogiannopoulos *et al.* [108] for rifampin phosphotransferases. All eight had the same architecture (ATP-RIF-HIS), with catalytic His in the C-terminus which phosphorylates rifamycin. MIC experiments revealed RphB2's ability to inactivate rifampicin (Table 2.3).

TABLE 2.3: Rifampin antibiotic susceptibility test results for RphB2.

Enzyme	**MIC (μg/mL)**	**Information**
RphB2	>4	Resistant
BLANK	0.5	Negative Control
Rox	>4	Positive Control

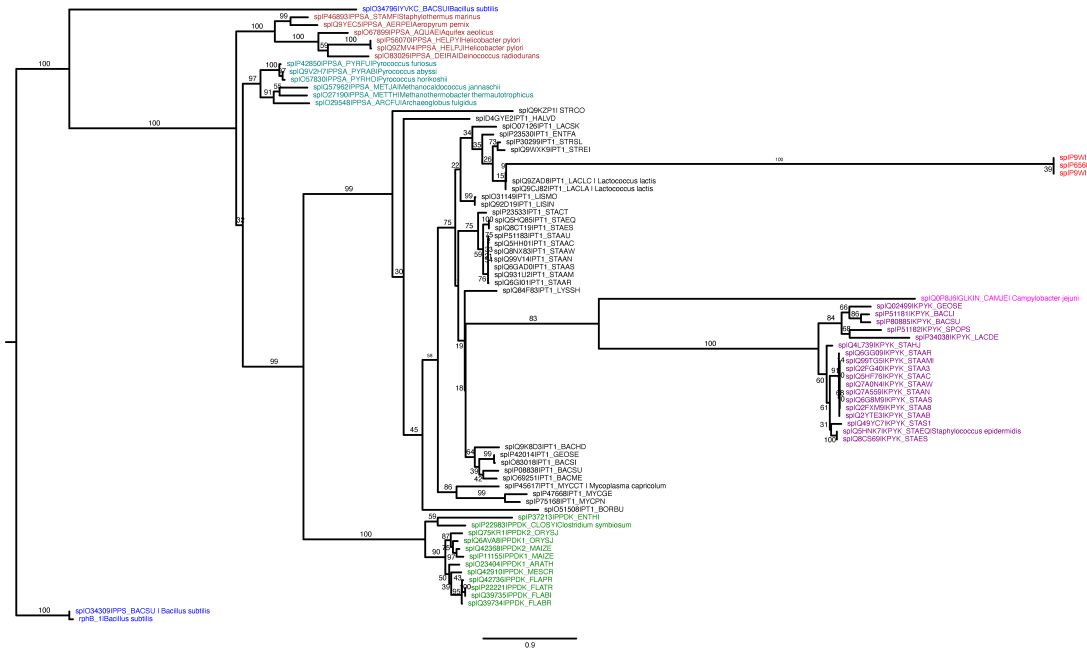


FIGURE 2.7: The maximum likelihood tree for *rphB1* (labelled as rphB_1|Bacillus subtilis) after sampling UniProt using HMM model. The putative *rphB1* (in blue) is in the same group with putative phosphoenolpyruvate synthase (PPS) from *Bacillus subtilis*. In brown, phosphoenolpyruvate synthase, one *Aeropyrum pernix* (Archaea). In teal, other putative phosphoenolpyruvate synthases (PPS) and one from *Archaeoglobus fulgidus* (Archaea). The PPS enzymes are involved in converting pyruvate into phosphoenolpyruvate (PEP). In green pyruvate, phosphate dikinase (PPDK). PPDK helps to convert pyruvate and Adenosine triphosphate (ATP) to Adenosine monophosphate (AMP) and PEP. In black, phosphoenolpyruvate-protein phosphotransferase (PT1), which transfers the phosphoryl group from PEP to a phosphoryl carrier protein. In purple, pyruvate kinase (KPYK). In pink, L-glutamine kinase, involved in pathway capsule polysaccharide biosynthesis. In red, uncharacterized proteins from *Mycobacterium tuberculosis* and *M. bovis* with transferase activity. GO annotations indicated transferring phosphorus-containing groups.

TABLE 2.4: Fosfomycin antibiotic susceptibility test results from experiment number 3 for FosD1, FosD2, and FosD3 (repeated twice).

Enzyme	**MIC (μg/mL)**	**Information**
FosD1	8	Susceptible
FosD2	8	Susceptible
FosD3	>512	Resistant
BLANK	8	Negative Control
FosA	>512	Positive Control

2.6 Discussion

We analyzed high-quality (average N50n of 9.4945 contigs) sequenced genomes from the human gut microbiome. The samples were from volunteers with no history of using antibiotics. Despite this, MAGIS predicted 113 unique Perfect annotations and 1,225 unique Strict annotations to AMR genes via RGI. These results show that commensal bacteria are capable of causing antibiotic drug resistance by expressing the genes annotated with Perfect and Strict annotations. FosD3 also shows that commensals have AMR genes not predicted by RGI Perfect & Strict, hence RGI has false Loose rate. The manually curated bitscore could be contributing to these missed annotations. The phylogenetic approaches, such as AutoPhylo, could be used to group homologous sequences and pick an appropriate bit-score to use as a cutoff. For this work, we were interested in uncovering AMR homologs and non-AMR related genes that are capable of metabolizing non-antibiotic drugs. As a result, we built MAGIS for this task. MAGIS uses five bioinformatic filters to predict putative drug-metabolizing genes from bacterial genomes. MAGIS allows filters based on protein domains or motifs to predict putative drug-metabolizing genes from the predicted proteins. After predicting the putative genes, we wanted a way to predict their function, and for that, we implemented a phylogenetic based software, AutoPhylo. AutoPhylo automatically builds a phylogenetic tree from a user submitted protein sequence. The predicted proteins obtained from MAGIS are used in AutoPhylo to infer possible functions for each protein based on the phylogenetic trees constructed. We predicted 246 putative genes using MAGIS (see Supplementary Table 2.5, Supplementary Table 2.6, and Supplementary Table 2.7).

There are two scenarios for the AutoPhylo predictions, the first is that the putative will be in the same clade or group as genes of known function and the second is that the putative will be an outgroup to such genes. The outgroup genes may have novel function. AutoPhylo uses eight defined phyla, using these specific phyla ensures that the clades we find would have already been established in the human gut instead of transient phyla from the environment.

To reduce the number of very overly distant homologous sequences when sampling, a user-defined percent positive scoring (i.e., percent identity adjusted for conservative amino acid substitutions) is used; by default, this value is set to 50%. Sampled sequences are also filtered by length to remove very long and very short sequences relative to the query sequence. Overall, AutoPhylo produces phylogenetic trees that can be used to infer functions of putative genes. From the predicted putatives, we focused on *fos* and *rphs* genes and validated three *fos* genes and one *rph* gene as proof of principle.

2.6.1 Fosfomycin background

Fosfomycin is an antibiotic containing an epoxide ring and carbon-phosphorus bonds [109] discovered in 1969 [109] from *Streptomyces fradiae* [110]. It is used to treat urinary tract infections and is also effective against methicillin-resistant *Staphylococcus aureus* (MRSA) infections [111–113]. The mechanism of action for fosfomycin is to inhibit the MurA enzyme, hence blocking the incorporation of UDP-GlcNAc (UDP-N-acetylglucosamine) and PEP (phosphoenolpyruvate) into new cell wall during peptidoglycan synthesis [110]. Fosfomycin can be inactivated by three types of bacterial metalloenzymes (coded by genes *fosA*, *fosB*, & *fosX*) and

two bacterial kinases (coded by genes *fomA* and *fomB*). The *fosB* genes are found in low GC monoderm [114] or gram-positive bacteria which use bacillithiol instead of glutathione; examples include *Staphylococcus saprophyticus*, *Staphylococcus aureus*, *Bacillus subtilis*, *Bacillus anthracis*, and *Staphylococcus epidermidis* [115]. The metalloenzymes use magnesium ion (Mg^{2+}) as a cofactor, and there is evidence that zinc ion (Zn^{2+}) inhibits these enzymes [116,117]. FosB enzymes use Mg^{2+} while FosX and FosA use manganese ion (Mn^{2+}) [110,118]. Thompson and colleagues demonstrated that deletion of the *fosB* gene or bacillithiol synthetic machinery from bacteria leads to a susceptible phenotype towards fosfomycin [116]. Roberts and colleagues came to a similar conclusion [115]. The *Streptomyces wedmorensis* species, producers of fosfomycin, use the kinases *fomA* and *fomB* to protect themselves from the antibiotic [119]. Another kinase is *fosC*, an ortholog of *fomA*, from fosfomycin producer *Pseudomonas syringae* [120]. All the kinases transfer phosphates from the adenosine triphosphate (ATP) cofactor and use Mg^{2+} to inactivate fosfomycin [121].

2.6.2 *Fos* homologs

We used AutoPhylo to compare the three putative *fosDs* genes (from filter 1) with known genes, and the putative *fosDs* resemble fosfomycin thiol transferases. From the preliminary biochemical test, the enzyme encoded by *fosD3* gene has a high MIC towards fosfomycin compared to putative enzymes encoded by *fosD2* and *fosD1* genes. After obtaining the antibiotic test preliminary results for FosD1 and FosD2, we performed further sequence analysis and compared residues in all three putative genes (Figure 2.8). The sequence analysis suggests that the putative FosDs are similar to enzymes that inactivate fosfomycin and are more closely

related to metalloenzyme FosB. The sequence analysis and motif identification suggest that FosD1 is a possible fosfomycin resistance element with subtlety with either metal ion or thiol preferred. We speculated that the three genes use different thiols, so we investigated the genetic context of the genes. The three genes are found within the same organism (*Staphylococcus saprophyticus*), there could be more than one thiol donor in an organism, but we were limited by the expression system in *E. coli*, which is not the native host. We found that the genes *fosD1* and *fosD2* are within an incomplete prophage with accession phage (gi77020174) using PHASTER (`https://phaster.ca`) (Supplementary Figure 2.12 & Supplementary Figure 2.13). This prediction shows that they originate from *Bacillus* gamma phages. The NCBI complete *Bacillus* gamma phage genomes show that they usually carry one copy of *fosB* gene. We used multiple rounds of biochemical tests to determine the activity of the three FosDs putative enzymes, and this highlights that biochemical methods are more complex, even for testing highly similar homologs. Multiple tests were required due to some mislabeled samples and inconsistencies with the negative controls having random growth amounts. The MIC results are inconsistent for FosD1 and FosD2 but consistent for FosD3. It is fair to say another round of tests are required to be certain of FosD1 and FosD2 susceptibility towards fosfomycin. It is not feasible to test all putative enzymes biochemically but using multiple *in silico* methods on the predicted genes can help elucidate function (i.e., domain predictions/MAGIS & AutoPhylo).

2.6.3 rphs homologs

The rifamycin antibiotics are important drugs for treating diseases such as tuberculosis [123]. These drugs target the bacterial RNA polymerase enzyme

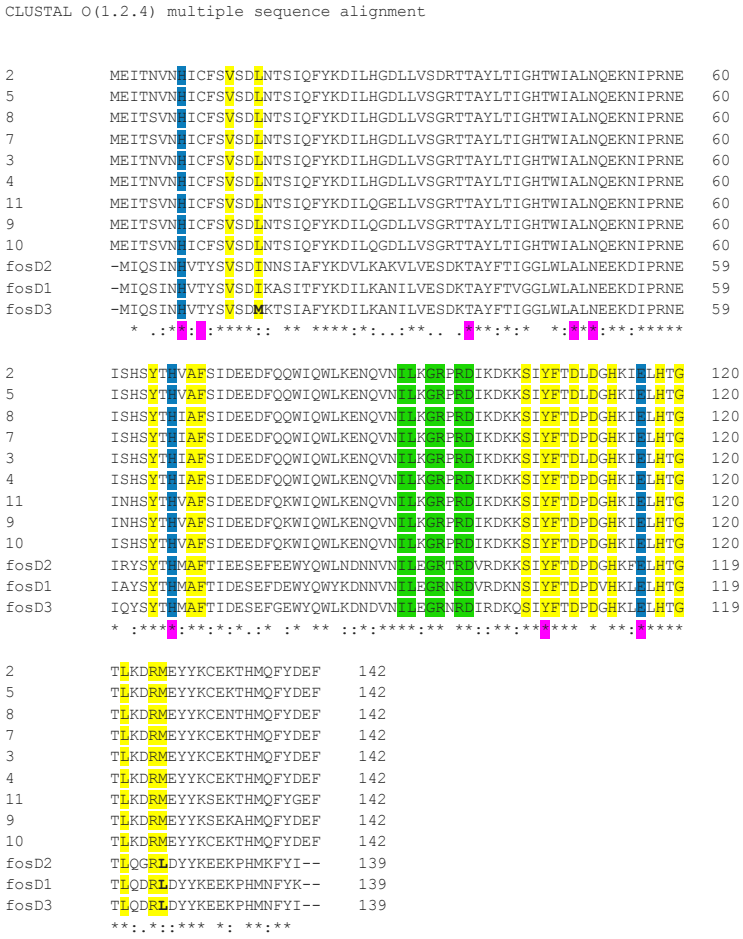


FIGURE 2.8: The figure shows sequence alignment for the 12 putative fosfomycin resistance genes predicted from the 1,196 bacterial genomes. The phosphonate binding loop is shown in green color, the metal binding sites in blue, and the active site residues in pink. In yellow are conserved residues for all sequences from Travis *et al.* [122], and other conserved residues are shown by the asterisks at the bottom of the alignment.

which is essential in protein transcription [124]. As a result, bacteria employ various methods to protect themselves from drugs that interfere with this important enzyme [108,124]. The mechanisms include efflux, mutation to the RNA polymerase, and dedicated genes to directly modify the drugs that target the RNA polymerase. In this study we uncovered nine dedicated rifamycin antibiotic inactivating gene homologs, including four *arr-1*, 76 *Bifidobacterium adolescentis rpoB* mutants conferring resistance to rifampicin, 330 *iri*, one *LAP-2*, 134 *RbpA*, 39 *rgt1438*, 45 *rphA*, 1,061 *rphB*, and 194 *rpoB2*. MAGIS filter 1 was able to capture eight of the 1,061 *rphB* homologs and this means this filter is highly stringent. Multiple filters can be used in MAGIS to obtain targeted results. All the filters implemented in this study are dedicated to domains and inactivation mechanisms. We anticipate using an antibiotic class, i.e drug class, or specific drug (such rifampicin antibiotic) can be used to filter out annotations that are not worthwhile to pursue. For this study, the *rphB* (i.e., *rphB2*) homolog is considered a negative result because the objective was to predict homologs to AMR genes that might metabolize non-antibiotic drugs. Spanogiannopoulos and colleagues highlighted that rif metabolism is found in *Streptomyces* genus sourced from the soil [102], and Pawlowski and colleagues identified *rphB* from a cave bacterium *Paenibacillus sp.* LC231 [125], but in this study we predicted several rif inactivating homologs which suggests that the gut microbiome can be another source for these rif inactivating elements.

2.6.4 Limitations of Phylogenetic Trees

The limitation of this study is that phylogenetic trees (e.g., from AutoPhylo) can be large and hard to interpret, but we can circumvent this by using sequence

clustering tools, for example, Enzyme Function Initiative - Enzyme Similarity Tool (EFT-EST [126]; https://efi.igb.illinois.edu/efi-est/). The EFT-EST Tool clusters proteins by function based on similarity cutoff values. We used the EFT-EST Tool for the FosDs which shows that they cluster with other fosfomycin inactivating genes (Figure 2.9).

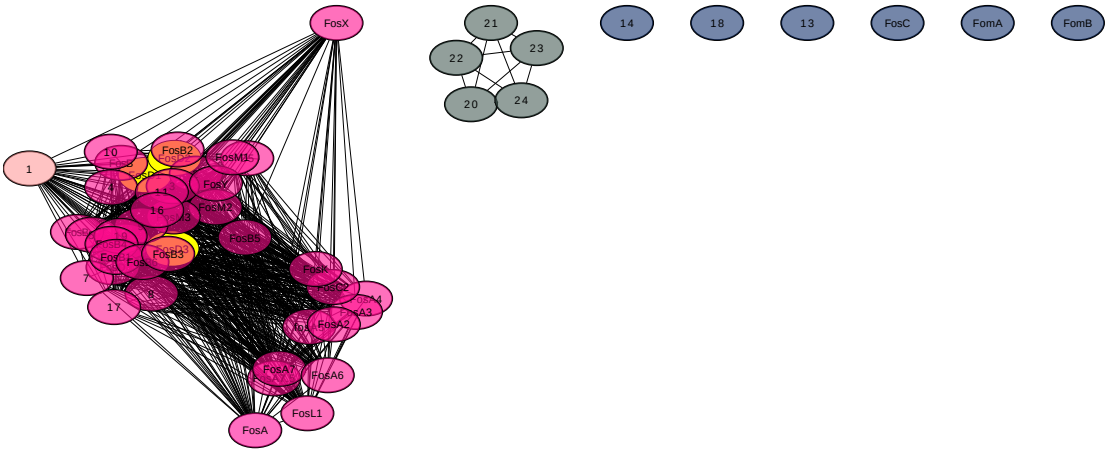


FIGURE 2.9: A similarity sequence network (SSN) was generated using the Enzyme Function Initiative - Enzyme Similarity Tool (EFT-EST [126]; https://efi.igb.illinois.edu/efi-est, Uniprot Version 2022_03, InterPro Version 90) and visualized using Cytoscape version 3.9.1. The SSN groups gene sequences with similar function, each gene is represented by a bubble or circle with a connecting line or edge to show relationship between sequences. The nodes in grey are closely related but not to those in pink and blue. The nodes in blue are singletons dissimilar to all the sequences examined. The nodes in pink have similar functions and *fosD1*, *fosD2*, and *fosD3* are coloured in yellow. The alignment score cutoff of 40, and sequence identity of 90%. The reference sequences were obtained from CARD version 3.2.5.

2.7 Conclusion

The *in silico* methods MAGIS and AutoPhylo help to predict and identify bacterial drug-metabolizing genes within the human gut. We predicted and verified one fosfomycin inactivating homolog to *fosD* (*fosD3*) and one rifampicin inactivating homolog to *rphB* (*rphB2*), while finding two *fosD* homologs with possible new function (*fosD1* and *fosD2*). The FosD1 and FosD2 findings are positive results in this study as they do not metabolize antibiotic tested (i.e fosfomycin) despite their high similarity to the fosfomycin inactivating enzyme FosD3. As a result, both methods are valuable for drug development and clinical trial evaluations of potential bacterial drug-metabolizing genes. RGI within the MAGIS was able to predict antibiotic inactivation enzymes and others are of unknown functions, which means that the bitscore set for the *fosDs* genes needs to be adjusted to be able to at least capture the *fosD3*. We also used flanking sequence around the *fosD1* gene, which led to identifying similar genes (data not shown) from the bioproject PRJNA636387 (samples sourced from human urinary tract). This result indicates that the *fosD1* and *fosD2* genes are complete and are found in multiple body sites i.e., human urinary tract and human gut. The MIC experiments were not enough to elucidate the possible function of the *fosD1* and *fosD2* genes, as a result, more studies are required to identify the function of these homologs. Given that there is evidence of prophages (i.e., genes transferred from other sources), perhaps the two encoded enzymes require specific co-factors to function or fosfomycin is not the natural substrate.

2.8 Supplementary material

2.8.1 Supplementary Figures

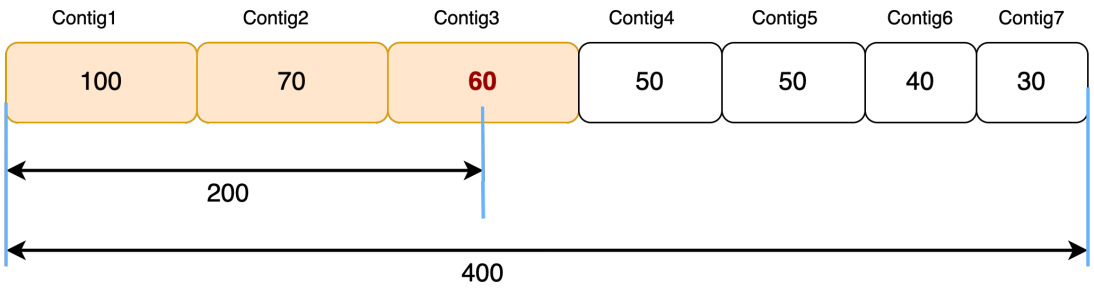


FIGURE 2.10: Illustration for N50 calculation, which requires sorting the contigs by length starting with the largest and picking the contig that lies in the middle to the whole assembly. In this example the N50 is 60bp with the assumption that a single unit is a base pair (i.e., 1bp). (The figure was adapted from https://www.molecularecologist.com)

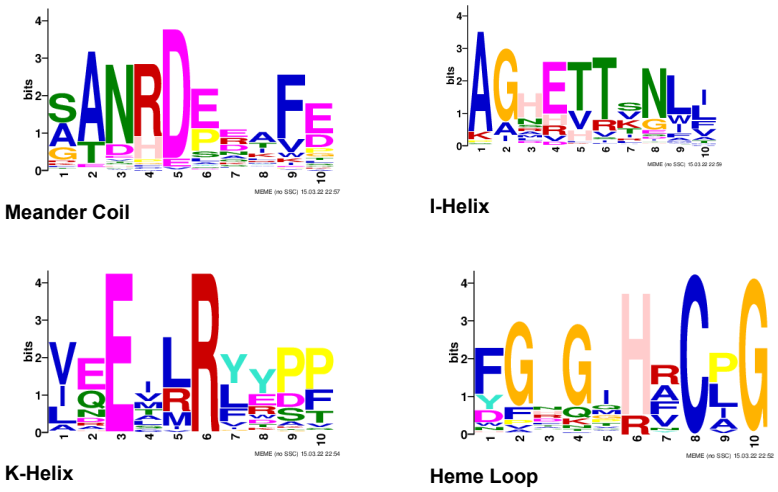
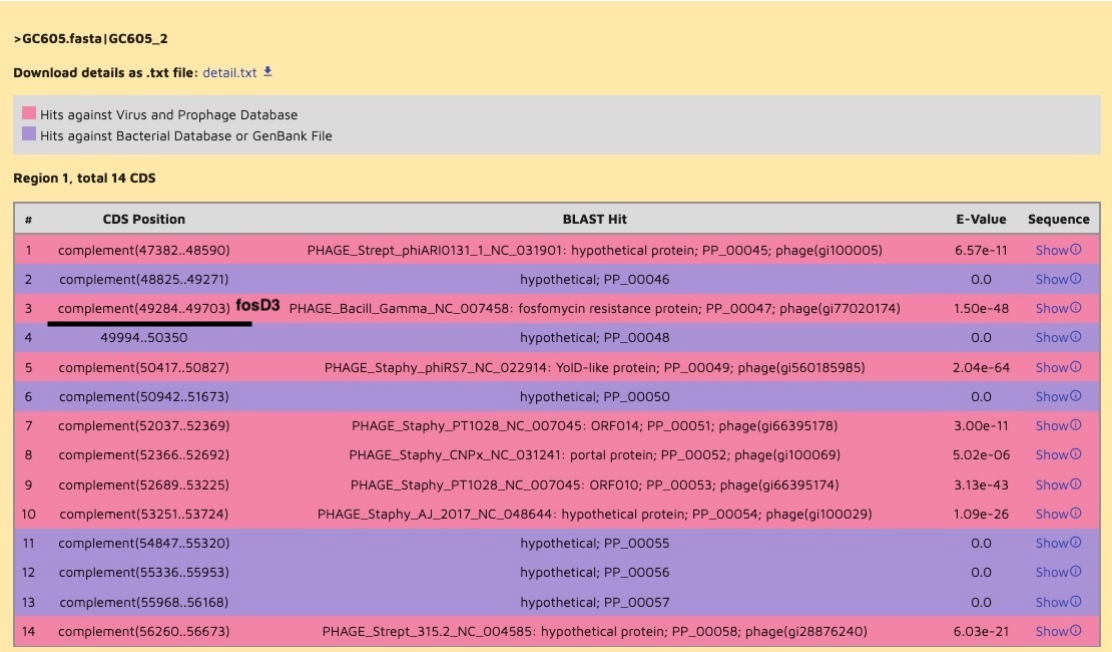


FIGURE 2.11: CYPs motifs identified using the MEME suite for all 10 putative bacterial CYPs.



>GC605.fasta|GC605_2

Download details as .txt file: detail.txt

Hits against Virus and Prophage Database
Hits against Bacterial Database or GenBank File

Region 1, total 14 CDS

#	CDS Position	BLAST Hit	E-Value	Sequence
1	complement(47382..48590)	PHAGE_Strept_phiARI0131_1_NC_031901: hypothetical protein; PP_00045; phage(gi100005)	6.57e-11	Show
2	complement(48825..49271)	hypothetical; PP_00046	0.0	Show
3	complement(49284..49703) fosD3	PHAGE_Bacill_Gamma_NC_007458: fosfomycin resistance protein; PP_00047; phage(gi77020174)	1.50e-48	Show
4	49994..50350	hypothetical; PP_00048	0.0	Show
5	complement(50417..50827)	PHAGE_Staphy_phiRS7_NC_022914: YolD-like protein; PP_00049; phage(gi560185985)	2.04e-64	Show
6	complement(50942..51673)	hypothetical; PP_00050	0.0	Show
7	complement(52037..52369)	PHAGE_Staphy_PT1028_NC_007045: ORF014; PP_00051; phage(gi66395178)	3.00e-11	Show
8	complement(52366..52692)	PHAGE_Staphy_CNPx_NC_031241: portal protein; PP_00052; phage(gi100069)	5.02e-06	Show
9	complement(52689..53225)	PHAGE_Staphy_PT1028_NC_007045: ORF010; PP_00053; phage(gi66395174)	3.13e-43	Show
10	complement(53251..53724)	PHAGE_Staphy_AJ_2017_NC_048644: hypothetical protein; PP_00054; phage(gi100029)	1.09e-26	Show
11	complement(54847..55320)	hypothetical; PP_00055	0.0	Show
12	complement(55336..55953)	hypothetical; PP_00056	0.0	Show
13	complement(55968..56168)	hypothetical; PP_00057	0.0	Show
14	complement(56260..56673)	PHAGE_Strept_315.2_NC_004585: hypothetical protein; PP_00058; phage(gi28876240)	6.03e-21	Show

FIGURE 2.12: The figure shows *fosD3* annotation using PHASTER from sample GC605. The *fosD3* CDS position is underlined in black. In pink are annotations predictions against prophage reference databases are highlighted and annotations in purple are against Bacteria reference databases.

>GC605.fasta|GC605_3

Download details as .txt file: detail.txt

Hits against Virus and Prophage Database
Hits against Bacterial Database or GenBank File

Region 1, total 21 CDS

#	CDS Position	BLAST Hit	E-Value	Sequence
1	239559..239570	attL	0.0	Show
2	complement(239921..240241)	PHAGE_Staphy_phiRS7_NC_022914: YolD-like protein; PP_00239; phage(gi560185985)	8.80e-53	Show
3	complement(240251..240442)	PHAGE_Staphy_phiRS7_NC_022914: hypothetical protein; PP_00240; phage(gi560185984)	1.75e-39	Show
4	240562..240741	hypothetical; PP_00241	0.0	Show
5	complement(240857..241381)	hypothetical; PP_00242	0.0	Show
6	241874..242275	hypothetical; PP_00243	0.0	Show
7	complement(242525..242644)	hypothetical; PP_00244	0.0	Show
8	complement(242641..242814)	hypothetical; PP_00245	0.0	Show
9	243068..244036	PHAGE_Bacill_phIS3501_NC_019502: phage integrase; PP_00246; phage(gi422934298)	1.81e-20	Show
10	complement(244400..244642)	hypothetical; PP_00247	0.0	Show
11	complement(245103..245423)	PHAGE_Paenib_Harrison_NC_028746: head-tail adaptor protein; PP_00248; phage(gi971482318)	2.16e-15	Show
12	complement(245853..246299)	hypothetical; PP_00249	0.0	Show
13	complement(246312..246731) **fosD2**	PHAGE_Bacill_Gamma_NC_007458: fosfomycin resistance protein; PP_00250; phage(gi77020174)	1.08e-49	Show
14	complement(246760..247230)	PHAGE_Erwini_vB_EamM_Caitlin_NC_031120: putative SSB protein; PP_00251; phage(gi100096)	9.08e-10	Show
15	complement(247359..247769)	hypothetical; PP_00252	0.0	Show
16	complement(247780..248193)	hypothetical; PP_00253	0.0	Show
17	complement(248203..248607)	hypothetical; PP_00254	0.0	Show
18	248641..248652	attR	0.0	Show
19	complement(248694..249029)	PHAGE_Staphy_PT1028_NC_007045: ORF014; PP_00255; phage(gi66395178)	2.23e-13	Show
20	complement(249026..249352)	PHAGE_Staphy_CNPx_NC_031241: portal protein; PP_00256; phage(gi100069)	9.40e-06	Show
21	complement(249349..249885)	PHAGE_Staphy_PT1028_NC_007045: ORF010; PP_00257; phage(gi66395174)	3.90e-42	Show

FIGURE 2.13: The figure shows *fosD2* annotation using PHASTER from sample GC605 (PHASTER failed to annotate FosD1 with CDS complement(238787..239206)). The *fosD2* CDS position is underlined in black. In pink are annotations predictions against prophage reference databases are highlighted and annotations in purple are against Bacteria reference databases.

2.8.2 Supplementary Tables

TABLE 2.5: The MAGIS results for Filter 3 produced 28 candidates. There are 18 serine hydrolases, 3 GCN5-related N-acetyltransferases (GNAT) family N-acetyltransferase, 5 aminoglycoside nucleotidyltransferases (ANT) putative, and 2 ole glycosyltransferases. The "SAMPLES" column is the identifier for the genome, "ID" is a unique identifier for each predicted protein, "PROTEIN_FAMILY[PFAM]" are annotations from Pfam, and "NCBI BLASTp [nr database]" are annotations from NCBI.

Samples	**ID**	**PFAM FAMILY**	**NCBI BLASTp NR**
GC80b_2_15; GC80_hybrid_0_881; GC643_2_15	SatA_1	Acetyltransferase (GNAT) domain; Domain of unknown function (DUF5613)	MULTISPECIES: GNAT family N-acetyltransferase [*Streptococcus*]; 100% identity
GC453_2_109	AAC(6′)Ibcr_1	Acetyltransferase (GNAT) family; Domain of unknown function (DUF4081)	GNAT family N-acetyltransferase [*Cutibacterium acnes*]; 100% identity
GC1045_3_102; GC1117_2_105	AAC(6′)Ibcr_2	Acetyltransferase (GNAT) family; Domain of unknown function (DUF4081)	GNAT family N-acetyltransferase [*Cutibacterium acnes*]; 100% identity
GC832_0_313	Escherichia coli ampC1 beta-lactamase_1	Domain of unknown function DUF302; Beta-lactamase	serine hydrolase [*Akkermansia muciniphila*]; 97% query coverage; 100% identity
GC392_hybrid_0_1188	Escherichia coli ampC1 beta-lactamase_2	Domain of unknown function DUF302; Beta-lactamase	serine hydrolase [*Alkalihalobacillus clausii*]; 91.76% identity
GC1160_0_1964; GC76_0_1974	Escherichia coli ampC beta-lactamase_1	Domain of unknown function (DUF3471); Beta-lactamase	MULTISPECIES: serine hydrolase [*Bacillus*]; 100% identity

GC382_1_39; GC553_4_39; GC591_1_39	Yrc-1_1	Domain of unknown function DUF302; Beta-lactamase	serine hydrolase [*Blautia coccoides*]; 99.8% identity
GC417_hybrid_0_565	Yrc-1_2	Domain of unknown function DUF302; Beta-lactamase	serine hydrolase [*Anaerotruncus massiliensis*]; 99.61% identity
GC305a_1_384; GC626_33_20; GC630_99_3; GC658_2_306;	Yrc-1_3	Domain of unknown function DUF302; Beta-lactamase	serine hydrolase [*Hungatella hathewayi*]; 100% identity
GC272_2_394; GC483_12_73; GC73_3_223	SRT-1_1	Domain of unknown function DUF302; Beta-lactamase	serine hydrolase [*Bacteroides fragilis*]; 100% identity
GC238_1_352	SRT-1_2	Domain of unknown function DUF302; Beta-lactamase	serine hydrolase [*Bacteroides fragilis*]; 100% identity
GC1010_1_325; GC1011_1_325; GC1041_18_41	ACC-3_1	Domain of unknown function (DUF3471); Beta-lactamase	serine hydrolase [*Dorea sp.* OM02-2LB]; 99.57% identity
GC1069_9_70	ACC-3_2	Domain of unknown function (DUF3471); Beta-lactamase	serine hydrolase [*Enterocloster asparagiformis*]; 100% identity
GC1160_2_137; GC76_2_266	ACT-37_1	Domain of unknown function (DUF3471); Beta-lactamase	MULTISPECIES: serine hydrolase [*Bacillus*]; 100% identity
GC793_4_146	ACT-37_2	Domain of unknown function (DUF3471); Beta-lactamase	MULTISPECIES: serine hydrolase [*Bacillus*]; 100% identity

GC1031_25_14	CMY-119_1	Domain of unknown function (DUF3471); Beta-lactamase	serine hydrolase [*Clostridium beijerinckii*]; 98.13% identity
GC1066_4_42; GC1146_12_42	CMY-104_1	Domain of unknown function (DUF3471); Beta-lactamase	serine hydrolase [*Peptococcus niger*]; 98% query coverage; 48.13% identity
GC471_25_28	CMY-98_1	Beta-lactamase; Domain of unknown function DUF302	serine hydrolase [*Cloacibacillus porcorum*]; 100% identity
GC1084_0_689	Rhodobacter sphaeroides ampC beta-lactamase_1	Domain of unknown function (DUF3471); Beta-lactamase	MULTISPECIES: serine hydrolase [*Bacillus*]; 100% identity
GC380_hybrid_0_747; GC381_hybrid_0_1934; GC709_1_346	Rhodobacter sphaeroides ampC beta-lactamase_2	Domain of unknown function (DUF3471); Beta-lactamase	MULTISPECIES: serine hydrolase [*Bacillus*]; 100% identity
GC458_hybrid_0_1113; GC589_12_121; GC590_8_44	PDC-82_1	Domain of unknown function DUF302; Beta-lactamase	serine hydrolase [*Blautia marasmi*]; 99% query coverage; 82.41% identity
GC432_28_3; GC764_28_73	ANT(9)-Ia_1	Domain of unknown function (DUF4111)	MULTISPECIES: DUF4111 domain-containing protein [*Bacteroidales*]; 100% identity

GC1069_40_5	ANT(9)-Ia_2	Domain of unknown function (DUF4111)	nucleotidyltransferase domain protein [*Clostridium asparagiforme* DSM 15981; *Clostridium asparagiforme* DSM 15981]; 99.61% identity
GC353_2_230	ANT(3")-Ib_1	Domain of unknown function (DUF4111)	MULTISPECIES: DUF4111 domain-containing protein [*Bacillus*]; 100% identity
GC747_12_14	ANT(3")-Ib_2	Domain of unknown function (DUF4111)	MULTISPECIES: DUF4111 domain-containing protein [*Oscillospiraceae*]; 100% identity
GC1211_10_63	ANT(3")-Ib_3	Domain of unknown function (DUF4111)	MULTISPECIES: DUF4111 domain-containing protein [*Oscillospiraceae*]; 99.27% identity
GC162b_25_14; GC162_hybrid_0_2848; GC253_25_14; GC276_21_10; GC278_22_10; GC707_23_31; GC60_hybrid_0_400	oleI_1	Protein of unknown function (DUF1205)	IroB [*Escherichia coli*]; 100% identity

GC61_hybrid_1_649	oleI_2	Protein of unknown function (DUF1205)	MULTISPECIES: salmochelin biosynthesis C-glycosyltransferase IroB [*Enterobacterales*]; 100% identity

TABLE 2.6: The MAGIS results for Filter 4, showing annotations with UDP glucoronosyl and UDP glucosyltransferase domains. The list was selected from annotations without active site prediction totaling 50 unique sequences. The "SAMPLES" is the identifier for the genome, "ID" is a unique identifier for each predicted protein, "AMR_GENE_FAMILY" are annotations from RGI, and "NCBI BLASTp [NR] database]" are annotations from NCBI. The "PROTEIN_FAMILY[PFAM]" annotations from Pfam are "UDP-glucoronosyl and UDP-glucosyltransferase" for all proteins. The proteins with "ole glycosyltransferase" are annotated with macrolide antibiotic drug class from RGI and "rifampin glycosyltransferase" has rifamycin antibiotic.

SAMPLES	**ID**	**AMR_GENE_FAMILY**	**NCBI BLASTp [NR]**
GC667_6_94_na	gimA_1	gimA family macrolide glycosyltransferase	MULTISPECIES: glycosyl transferase [*Staphylococcus*]; 100% identity
GC383_hybrid_0_1872_na	gimA_2	gimA family macrolide glycosyltransferase	UDP-glucosyltransferase [*Priestia megaterium*]; 93.165% identity
GC106_hybrid_0_1281_na	gimA_3	gimA family macrolide glycosyltransferase	glycosyltransferase [*Streptococcus mutans* UA159-FR]; 99.472% identity
GC893_3_27_na	oleD_1	ole glycosyltransferase	glycosyltransferase family 28 protein [*Actinomyces sp.* oral taxon 170 str. F0386]; 84.197% identity
GC260_13_48_na	oleD_2	ole glycosyltransferase	hypothetical protein [*Coprobacillus cateniformis*]; 100% identity

GC793_6_49_na	oleD_3	ole glycosyltransferase	Glycosyl Transferase Family 1 [*Bacillus licheniformis* DSM 13 = ATCC 14580]; 99.433% identity
GC222_33_34_na	oleI_1	ole glycosyltransferase	MGT family glycosyltransferase [*Lachnospiraceae bacterium* 3-1]; 97.744% identity
GC1052_36_21_na	oleI_2	ole glycosyltransferase	glycosyltransferase family 1 protein [*Faecalibacillus intestinalis*]; 100% identity
GC52_14_43_na	oleI_3	ole glycosyltransferase	glycosyl transferase family protein [*Flavonifractor plautii*]; 97.733% identity
GC1031_12_90_na	oleI_4	ole glycosyltransferase	glycosyl transferase [*Clostridium beijerinckii*]; 97.543% identity
GC1084_1_993_na	oleI_5	ole glycosyltransferase	YojK [*Bacillus subtilis subsp. subtilis str.* 168]; 99.303% identity
GC1086_6_72_na	oleI_6	ole glycosyltransferase	UDP-glycosyltransferase GT-1, partial [*Bacillus subtilis subsp. spizizenii* ATCC 6633]; 62.972% identity

GC1123_2_92_na	oleI_7	ole glycosyltransferase	glycosyl transferase [*Staphylococcus equorum*]; 61.809% identity
GC383_hybrid_0_198_na	oleI_8	ole glycosyltransferase	glycosyl transferase family 1 [*Priestia megaterium*]; 96% identity
GC602_12_71_na	oleI_9	ole glycosyltransferase	UDP-glycosyltransferase GT-1, partial [*Bacillus subtilis subsp. spizizenii* ATCC 6633]; 62.72% identity
GC1160_4_40_na	oleI_10	ole glycosyltransferase	putative UDP-glucosyltransferase YdhE [*Bacillus licheniformis* DSM 13 = ATCC 14580]; 99.495% identity
GC790_5_95_na	oleI_11	ole glycosyltransferase	glycosyl transferase family protein [*Flavonifractor plautii*]; 96% identity
WAC8344_hybrid_0_224_na	oleI_12	ole glycosyltransferase	glycosyl transferase family 1 (plasmid) [*Brevibacillus laterosporus*]; 81% identity
GC793_5_64_na	oleI_13	ole glycosyltransferase	putative UDP-glucosyltransferase YdhE [*Bacillus licheniformis* DSM 13 = ATCC 14580]; 99.492% identity

GC102_hybrid_18_24_na	oleI_14	ole glycosyltransferase	glycosyl transferase family protein [*Flavonifractor plautii*]; 97.481% identity
GC1086_4_19_na	oleI_15	ole glycosyltransferase	putative UDP-glucosyltransferase YdhE [*Bacillus licheniformis* DSM 13 = ATCC 14580]; 69.797% identity
GC1103_14_60_na	oleI_16	ole glycosyltransferase	glycosyl transferase family protein [*Flavonifractor plautii*]; 97.229% identity
GC602_2_390_na	oleI_17	ole glycosyltransferase	putative UDP-glucosyltransferase YdhE [*Bacillus licheniformis* DSM 13 = ATCC 14580]; 69.543% identity
GC704_22_31_na	oleI_18	ole glycosyltransferase	glycosyl transferase family protein [*Flavonifractor plautii*]; 96.474% identity
GC789_8_50_na	oleI_19	ole glycosyltransferase	glycosyl transferase family protein [*Flavonifractor plautii*]; 100% identity
GC980_20_26_na	oleI_20	ole glycosyltransferase	TPA: glucosyltransferase [*Oscillibacter sp.*]; 62.113% identity

GC1160_0_1326_na	oleI_21	ole glycosyltransferase	glycosyltransferase [*Bacillus licheniformis*]; 94.697% identity
GC1048_40_16_na	oleI_22	ole glycosyltransferase	glucosyltransferase [*Clostridium phoceensis*]; 99.747% identity
GC1084_5_18_na	oleI_23	ole glycosyltransferase	putative UDP-glucosyltransferase YdhE [*Bacillus licheniformis* DSM 13 = ATCC 14580]; 69.211% identity
GC1160_2_336_na	oleI_24	ole glycosyltransferase	Glycosyl Transferase Family 1 [*Bacillus licheniformis* DSM 13 = ATCC 14580]; 99.746% identity
GC1509_286_1_na	oleI_25	ole glycosyltransferase	glucosyltransferase [*Clostridioides difficile*]; 75.192% identity
GC380_hybrid_0_1996_na	oleI_26	ole glycosyltransferase	putative UDP-glucosyltransferase YdhE [*Bacillus licheniformis* DSM 13 = ATCC 14580]; 69.466% identity
GC415_11_73_na	oleI_27	ole glycosyltransferase	glucosyltransferase [*Clostridium phoceensis*]; 100% identity

GC472_15_15_na	oleI_28	ole glycosyltransferase	glucosyltransferase [*Clostridium phoceensis*]; 98.987% identity
GC730_0_1898_na	oleI_29	ole glycosyltransferase	glucosyltransferase [*Clostridium phoceensis*]; 98.985% identity
GC778_3_340_na	oleI_30	ole glycosyltransferase	glucosyltransferase [*Clostridium phoceensis*]; 99.494% identity
GC1086_0_652_na	oleI_31	ole glycosyltransferase	hypothetical protein U471_12310 [*Bacillus amyloliquefaciens* CC178]; 98.731% identity
GC602_0_291_na	oleI_32	ole glycosyltransferase	hypothetical protein U471_12310 [*Bacillus amyloliquefaciens* CC178]; 98.477% identity
GC792_26_51_na	oleI_33	ole glycosyltransferase	glycosyl transferase [*Flavonifractor plautii*]; 97.297% identity
GC383_hybrid_0_317_na	oleI_34	ole glycosyltransferase	MULTISPECIES: glycosyl transferase [*Priestia*]; 80.208% identity
GC383_hybrid_0_2009_na	rgt1438_1	rifampin glycosyltransferase	glycosyl transferase, family 28 [*Bacillus megaterium* QM B1551]; 98.826% identity

PES_hybrid_0_784_na	rgt1438_2	rifampin glycosyltransferase	MULTISPECIES: glycosyltransferase [*Pseudomonas*]; 100% identity
GC1031_52_20_na	rgt1438_3	rifampin glycosyltransferase	sterol 3-beta-glucosyltransferase [*Priestia megaterium*]; 64.578% identity
GC190_6_107_na	rgt1438_4	rifampin glycosyltransferase	glycosyltransferase family 1 protein, partial [*Bacteroides ovatus*]; 99.659% identity
GC760_12_28_na	rgt1438_5	rifampin glycosyltransferase	glycosyltransferase family 1 protein, partial [*Bacteroides ovatus*]; 100% identity
GC189_1_124_na	rgt1438_6	rifampin glycosyltransferase	glycosyltransferase family 1 protein, partial [*Bacteroides ovatus*]; 99.317% identity
GC401_hybrid_0_2634_na	rgt1438_7	rifampin glycosyltransferase	glycosyltransferase family 1 protein, partial [*Bacteroides ovatus*]; 99.659% identity
GC1062_0_105_na	rgt1438_8	rifampin glycosyltransferase	glycosyltransferase [*Bacteroides ovatus*]; 91.81% identity

GC1032_0_109_na	rgt1438_9	rifampin glycosyltransferase	Glycosyl transferases%2C related to UDP-glucuronosyltransferase [*Bacteroides faecis*]; 100% identity
GC228_43_18_na	rgt1438_10	rifampin glycosyltransferase	uncharacterized protein BN607_02294 [*Bacteroides faecis* CAG:32]; 100% identity

TABLE 2.7: The MAGIS results for Filter 5 showing 154 unique putative bacterial CYP450s predicted from 74 samples. The "SAMPLE_ORF" is identifier for the predicted protein, "START" and "END" shows sections which aligns with the Pfam domain predicted i.e p450.

SAMPLE ORF	START	END	DOMAIN (Pfam)
GC1035_15_18	274	364	p450
GC1035_60_15	210	374	p450
GC1084_0_625	16	372	p450
GC1084_1_389	17	450	p450
GC1084_1_433	24	400	p450
GC1084_1_81	14	365	p450
GC1084_2_119	66	269	p450
GC1084_2_120	2	62	p450
GC1084_3_52	49	373	p450
GC1084_4_134	12	446	p450
GC1084_6_138	277	393	p450
GC1086_0_148	85	374	p450
GC1086_0_20	21	368	p450
GC1086_0_653	46	386	p450
GC1086_0_778	280	380	p450
GC1086_1_360	17	449	p450
GC1086_1_601	163	353	p450
GC1086_5_26	12	443	p450
GC1095_26_2	268	327	p450
GC1095_7_63	285	405	p450

GC1160_0_1103	53	399	p450
GC1160_0_447	16	449	p450
GC1160_0_471	83	400	p450
GC1160_1_294	16	373	p450
GC1160_3_158	19	368	p450
GC1160_3_96	44	373	p450
GC144_3_211	272	367	p450
GC144_6_66	271	370	p450
GC1500_0_1111	277	397	p450
GC1503_0_1111	277	397	p450
GC1504_0_1111	277	397	p450
GC1505_2_273	277	397	p450
GC156_hybrid_0_1131	271	370	p450
GC156_hybrid_0_593	272	367	p450
GC157_2_211	272	367	p450
GC157_5_66	271	370	p450
GC210_1_33	279	399	p450
GC230_4_34	279	399	p450
GC23_hybrid_3_17	279	399	p450
GC24_hybrid_0_193	279	399	p450
GC258_13_35	279	399	p450
GC25_hybrid_3_110	279	399	p450
GC26_hybrid_0_17	279	399	p450
GC292_1_41	287	402	p450
GC293_0_693	287	402	p450

GC298_3_29	272	367	p450
GC298_5_66	271	370	p450
GC299_3_211	272	367	p450
GC299_5_66	271	370	p450
GC29_hybrid_2_63	279	399	p450
GC300_3_211	272	367	p450
GC300_5_66	271	370	p450
GC301_3_211	272	367	p450
GC301_7_37	271	370	p450
GC30_hybrid_3_63	279	399	p450
GC31_hybrid_0_84	279	399	p450
GC31_hybrid_9_34	279	399	p450
GC321_6_4	162	373	p450
GC32_19_30	22	399	p450
GC322_hybrid_0_1688	162	373	p450
GC323_5_31	163	376	p450
GC324_4_111	163	375	p450
GC325_2_53	161	368	p450
GC326_4_66	163	375	p450
GC33_hybrid_9_31	22	399	p450
GC34_19_14	22	399	p450
GC353_31_32	278	389	p450
GC35_hybrid_1_93	22	399	p450
GC380_hybrid_0_13	17	450	p450
GC380_hybrid_0_1641	277	392	p450

GC380_hybrid_0_2162	12	446	p450
GC380_hybrid_0_2659	49	373	p450
GC380_hybrid_0_3179	66	373	p450
GC380_hybrid_0_320	14	365	p450
GC380_hybrid_0_810	17	372	p450
GC381_hybrid_0_1200	17	450	p450
GC381_hybrid_0_1507	14	365	p450
GC381_hybrid_0_1997	17	372	p450
GC381_hybrid_0_2828	277	392	p450
GC381_hybrid_0_3349	12	446	p450
GC381_hybrid_0_352	66	373	p450
GC381_hybrid_0_3846	49	373	p450
GC383_hybrid_0_1993	6	443	p450
GC383_hybrid_0_2399	53	372	p450
GC383_hybrid_0_2591	17	373	p450
GC383_hybrid_0_633	13	389	p450
GC383_hybrid_0_786	8	366	p450
GC383_hybrid_0_827	53	399	p450
GC383_hybrid_4_6	143	261	p450
GC392_hybrid_0_2903	277	395	p450
GC392_hybrid_0_817	197	370	p450
GC396_4_50	161	370	p450
GC46_0_188	218	370	p450
GC51_3_63	279	399	p450
GC518_6_125	275	388	p450

GC533_2_70	161	369	p450
GC534_6_34	161	370	p450
GC535_2_150	161	370	p450
GC536_3_60	164	369	p450
GC55_3_2	268	369	p450
GC602_0_157	280	380	p450
GC602_0_290	46	386	p450
GC602_1_257	86	374	p450
GC602_1_392	21	368	p450
GC602_15_30	12	443	p450
GC602_3_237	17	449	p450
GC602_7_211	163	354	p450
GC65_hybrid_0_515	183	381	p450
GC667_2_136	98	186	p450
GC67_3_2	268	369	p450
GC68_5_98	283	387	p450
GC69_hybrid_0_1484	265	369	p450
GC70_3_50	22	471	p450
GC709_0_1033	14	365	p450
GC709_0_726	17	450	p450
GC709_1_409	17	372	p450
GC709_2_354	49	373	p450
GC709_2_874	66	373	p450
GC709_3_142	12	446	p450
GC709_5_140	277	392	p450

GC76_0_1037	53	399	p450
GC76_0_377	16	449	p450
GC76_0_401	83	400	p450
GC76_1_810	16	373	p450
GC76_3_30	19	368	p450
GC76_3_92	44	373	p450
GC77_3_108	254	353	p450
GC793_0_248	16	449	p450
GC793_0_272	83	400	p450
GC793_0_915	53	399	p450
GC793_2_232	16	373	p450
GC793_5_131	19	368	p450
GC793_5_203	44	373	p450
GC868_7_42	163	375	p450
GC869_2_147	161	370	p450
GC871_1_60	161	369	p450
GC873_1_114	8	133	p450
GC873_17_2	223	286	p450
GC873_17_2	291	389	p450
GC873_31_15	276	394	p450
GC878_33_12	278	400	p450
GC892_1_114	197	409	p450
GC892_4_44	272	363	p450
GC981_3_39	163	372	p450
GC981_4_113	272	358	p450

GC992_4_150	163	372	p450
GC992_5_113	272	358	p450
GC997_12_55	275	387	p450
PES_hybrid_0_2292	194	450	p450
PES_hybrid_0_5263	309	406	p450
PES_hybrid_0_634	157	368	p450
PES_hybrid_0_986	228	398	p450
WAC8344_hybrid_0_3716	19	443	p450
WAC8344_hybrid_0_377	16	445	p450
WAC8344_hybrid_0_660	35	378	p450

TABLE 2.8: Bacterial CYPs predicted and annotated using MAGIS. The MOTIFs were annotated using MEME software. The protein GC1084_2_119 is missing a heme loop. All the putative are arranged MIKH (M = Meander Coil, I = I-Helix, K = K-helix, H = Heme Loop). In red is CYPs annotated with "CYPs_unknown".

CYP	Protein ID	Length	I	K	M	H
CYP107A1	GC709_2_874	405	AGLETT	ELLR	FTPRT	FGFGIHFCLG
CYP107A1	GC1084_2_119	311	AGLETT	ELLR	FTPRT	-
similar to Erythromycin 12 hydroxylase	GC325_2_53	403	GELSTV	EIMR	FTPER	YGDGIHVCPG
similar to Erythromycin 12 hydroxylase	GC869_2_147	403	GELSTV	EIMR	FTPER	YGDGIHVCPG
similar to Erythromycin 12 hydroxylase	GC992_4_150	398	GELSTV	EIMR	FTPER	YGDGIHVCPG
similar to Erythromycin 12 hydroxylase	GC324_4_111	398	GELSTV	EIMR	FTPER	YGDGIHVCPG
similar to Erythromycin 12 hydroxylase	GC321_6_4	398	GELSTV	EIMR	FTPER	YGDGIHVCPG
CYP106A1	WAC8344_hybrid_0_660	410	AGVETT	EMLR	FDREK	FGNGPHFCLG
CYP106A1	GC383_hybrid_0_633	410	AGVETT	EMLR	YDQER	FGNGPHFCLG
CYP102A1	GC1160_0_1103	404	AGNETT	EMLR	RDELK	FGFGIHFCLG

Chapter 3

The Human Microbiome Drug Metabolism (HMDM) Database

3.1 Chapter 3 Preface

Some of the literature reviews were conducted by volunteers for this chapter.

Amogelang R. Raphenya[a,b,c], Akash Mehta[d], Nawal Masood[d], Michaela Hughes-Butler[d], Mugdha Dave[a,b,c], Mahrukh Khan[d], Michael G. Surette[a,b,c], Gerard D. Wright[a,b,c], Andrew G. McArthur[a,b,c]

Author contributions: ARR, MGS, GDW, and AGM conceived the project. AM, NM, MHB, MD, and MK performed literature reviews. AM and MD performed some of the curations in the database. ARR designed and developed the database as a whole (i.e., the ontology, schema, software, and website), performed curation into the database, performed the analysis, and wrote this chapter.

[a] Michael G. DeGroote Institute for Infectious Disease Research, McMaster University, Hamilton, Ontario, Canada

[b] Department of Biochemistry and Biomedical Sciences, McMaster University, Hamilton, Ontario, Canada

[c] David Braley Centre for Antibiotic Discovery, McMaster University, Hamilton, Ontario, Canada

[d] Undergraduate program, McMaster University, Hamilton, Ontario, Canada

3.2 Abstract

3.2.1 Objective

It is increasingly evident that bacterial drug metabolism by the human microbiota plays a critical role in drug efficacy. Yet, there are no resources with well-curated gene annotations that can be used to study human gastrointestinal (gut) microbiome drug metabolism. The aim of this chapter is to build a resource to annotate microbiome genomic data for potential bacterial drug metabolism genes.

3.2.2 Methods

We used ontologies to build an ontology-centric database and systematically catalog all reported bacterial drug-metabolizing genes and their encoding gene sequences that contribute to bacterial drug metabolism.

3.2.3 Results

We built a resource, the HMDM database, to catalog all known bacterial drug metabolism genes from the human gut microbiome that code for enzymes that modify or degrade drugs. We also built annotation software, the DME, to use the reference data in the HMDM to predict potential new bacterial drug metabolism genes. We additionally developed HMDM*Shark software to triage relevant microbial drug metabolism literature for monthly curation.

3.2.4 Conclusions

The HMDM database and the accompanying software DME are vital resources to help predict bacterial drug-metabolizing genes from genomic sequences from the human gut microbiome. These resources could be used to interrogate human gut microbiomes for their ability to transform drugs.

3.3 Introduction

The human gastrointestinal tract (gut) is host to trillions of microbes that affect drug effectiveness by degrading or modifying oral therapeutics [62,127]. The gut microbiota modifies drugs by releasing enzymes into the gut lumen, facilitating efficient biochemical reactions [4]. These drug modifications impact effective doses and how individuals respond to their medication. Yet, due to variable gut microbiota composition, individual drug responses can differ [60–62]. On the other hand, xenobiotics can affect microbiome viability. Maier and colleagues looked at 1,000 drugs and found that 24% slow down bacterial growth [128]. Bacterial drug metabolism is not limited to orally administered drugs, the microbiota also

converts metabolites destined for excretion via the gut, including drug conjugates from the liver [129,130]. To date, no resources systematically catalog all reported bacterial genes and enzymes involved in gut microbiome drug metabolism [131]. Drug development and clinical trials are costly, so it is essential that bacterial drug-metabolizing genes be identified before the drugs are further developed [65].

There are methods previously developed by other groups to predict microbial drug metabolism showing promising results, namely, MicrobeFDT [56], SIMMER (Similarity algorithms that Identify MicrobioMe Enzymatic Reactions) [57], MASI (Microbiota Active Substance Interactions) database [51], PharmacoMicrobiomic [52], Disbiome [53], gutMDisorder [54], and MagMD (Metabolic action of gut Microbiota to Drugs) [55]. Most of these tools lack gene-level information and connections between bacterial drug-metabolizing genes and drugs, and some of the annotations are incomplete. As a result, we built an ontology-centric database to systematically catalog all bacterial drug-metabolizing genes and their encoding gene sequences that contribute to (if present) microbial drug metabolism, coined the Human Microbiome Drug Metabolism Database (HMDM; `https://hmdm.mcmaster.ca`). The HMDM is a manually curated database with supporting literature describing the bacterial drug-metabolizing genes involved in bacterial drug metabolism of mainly host-directed drugs. The HMDM database will help researchers by providing a reference resource to annotate microbiome data and inform potential bacterial drug metabolism sources. Additionally, the HMDM can guide the discovery of new bacterial drug metabolism genes within human microbiome samples.

3.4 Methods

3.4.1 Hardware and Setup

The HMDM has a development and a production branch. Currently, both are available behind the McMaster University firewall. The development branch is used to curate and quality check data before each release to the public-facing production branch (public release pending). Both instances are built using Ubuntu version 20.04 with 60GB of memory, 8 cores, and 1TB of storage.

3.4.2 Website and Schema Design

The HMDM website uses Laravel version 9 (https://laravel.com/docs/9.x/releases), a PHP-based framework. The Postgresql (Version 13.1) object-relational database software stores these data. The web server running the HMDM is NGINX version 1.18. The SQL schema combined a suite of *de novo* approaches and pre-existing modules from CARD [86], some of which have origins from the Generic Model Organism Database schema [132] (GMOD, http://gmod.org/wiki/Main_Page).

3.4.3 The HMDM ontology

The HMDM ontology describes bacterial drug metabolism or degradation by bacterial enzymes in the human gut. An extensive literature review was performed using a manual search of PubMed, Semantic Scholar, and using a custom-built automated text mining algorithm, HMDM*Shark. To keep the HMDM database comprehensive with reported drug-metabolizing genes, HMDM*Shark is used to

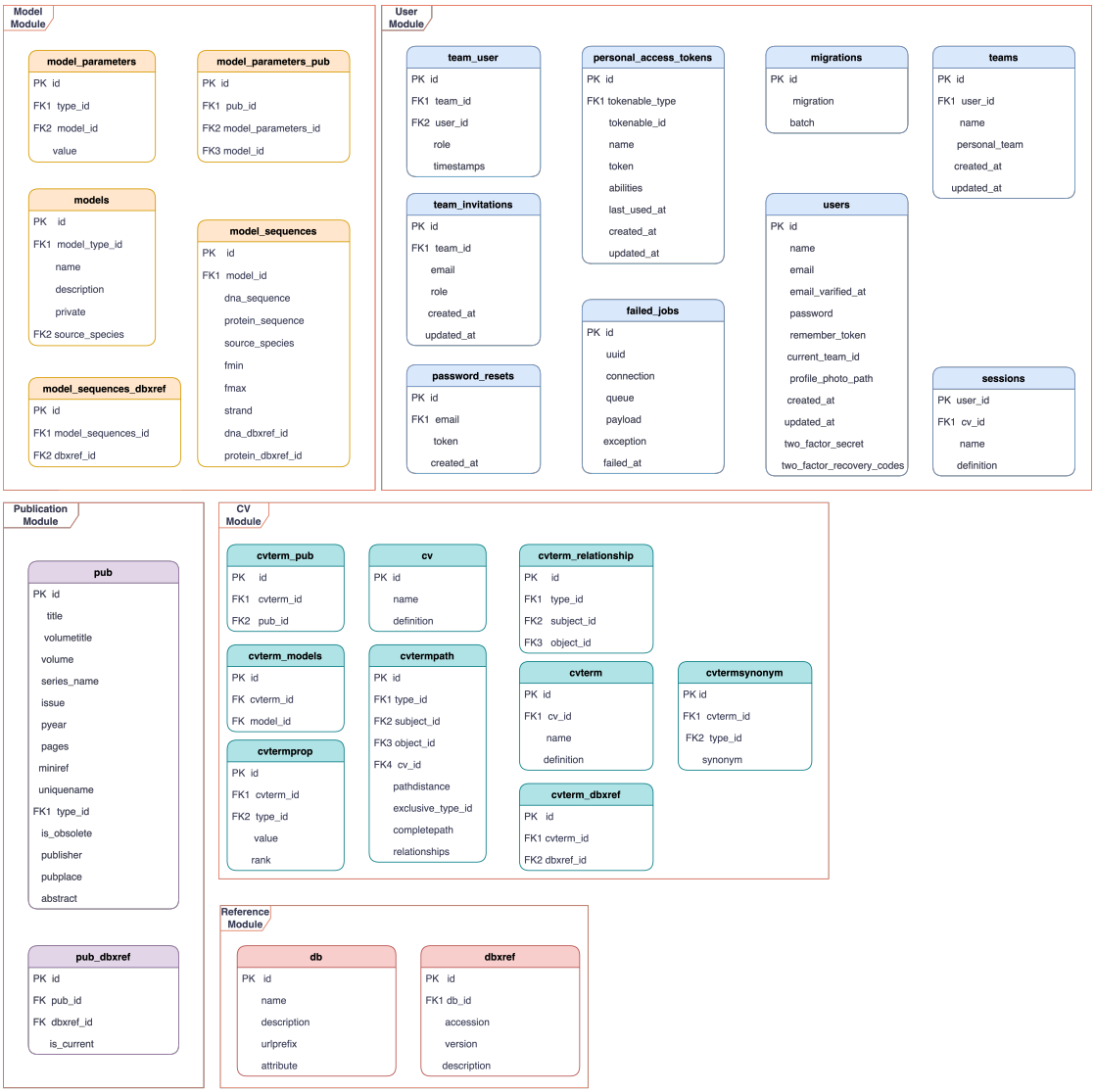


FIGURE 3.1: HMDM database schema showing all five modules. The user module (in blue) captures the user's information while using the database. The model module (in orange) stores molecular information for the bioinformatic models for each drug metabolizing gene in the HMDM database. The publication module (in purple) stores identifiers for the literature describing drug metabolism in the human gut, which includes abstracts and titles. The controlled vocabulary (CV) module (in light green) stores terms used to describe drug-metabolism concepts, e.g., the name of the gene as well as the drug it modifies. The reference module (in red) stores reference databases, e.g., accession numbers for protein and nucleotide sequences for each gene.

TABLE 3.1: The HMDM database contains six modules described in the table.

Module	**Number of Tables**	**Description**
Publication	2	This module stores information on articles or literature used in the database. This includes peer-reviewed papers describing the subject matter of the HMDM database.
User	9	All users and their activity within the database are stored in this module.
Reference	2	This module stores accession numbers from external resources, for example, accession numbers for gene ontology (GO) terms imported to be used together with HMDM database terms.
Controlled Vocabulary (CV)	9	This module stores terms used to describe concepts of HMDM ontology
Model	5	Gene and molecular information is stored in this module
Prevalence	7	Records the gene frequencies of HMDM genes in the sampled bacterial genomes.

search for literature relevant to the HMDM database at the start of every month. The literature review allowed for developing the ontology terms that describe bacterial drug metabolism in the human gut. HMDM*Shark's software was based on CARD*Shark version 2 [133] and a brief description of HMDM*Shark is provided below.

3.4.4 Predicting microbial drug metabolizing genes using novel Drug Metabolizing Enzyme (DME) software

A software termed Drug Metabolizing Enzyme (DME), which we built using Python programming language to use these data in the HMDM database, predicts bacterial drug-metabolizing genes in the draft and complete genomes. Details on this software and its annotation criteria are provided below.

3.4.5 Selecting taxa for the HMDM*Prevalence module

We used three approaches to determine an appropriate list of bacterial species or strains to analyze for bacterial drug-metabolizing gene prevalence. First, we used strains or species from which the genes annotated in HMDM were discovered. For this, a species list was selected to download bacterial genomes from NCBI Datasets (`https://www.ncbi.nlm.nih.gov/datasets`), analyze the genomes using DME, and calculate the prevalence or frequency of each gene in the HMDM database, thus producing the HMDM*Prevalence data set. Secondly, we analyzed the proteome data available at the UniProt [134] database using DME to determine taxa associated with bacterial drug-metabolizing proteins. Third, using DME, we analyzed the in-house human gut microbiome genomes and culture-enriched

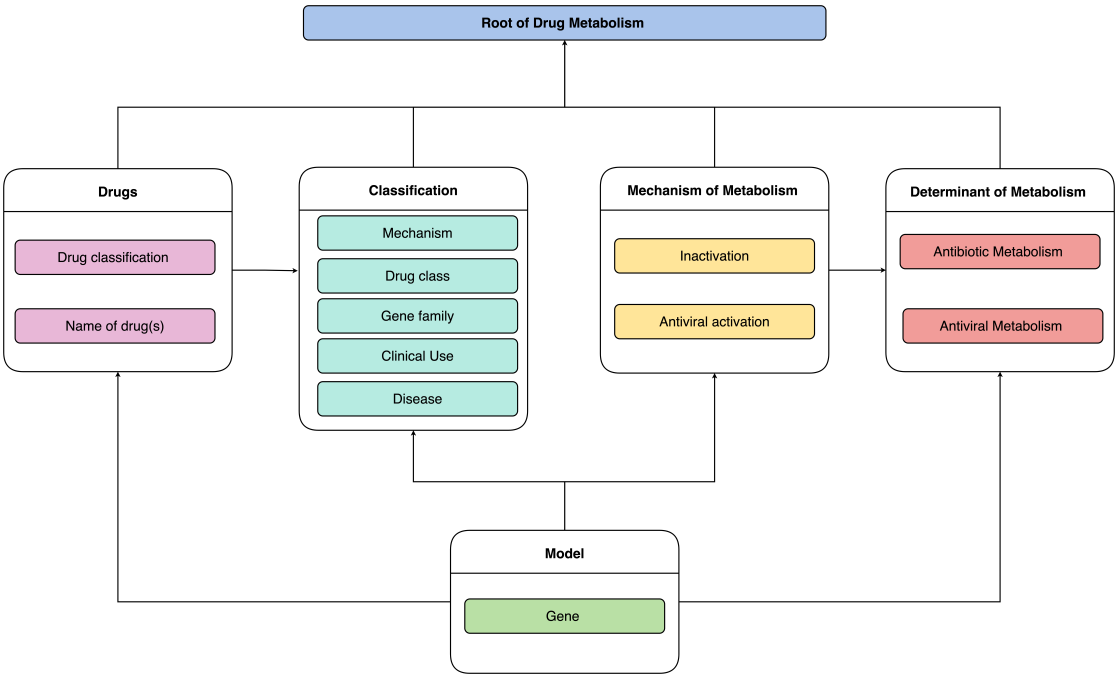


FIGURE 3.2: The overall HMDM topology. The HMDM database contains five main branches: Drugs, Classification, Mechanism of Drug Metabolism, and Determinant of Drug Metabolism, and all are connected at the bottom with a model. The Drugs branch captures drug names, descriptions, and classifications. The classification module tag each gene connected with "Drug class", "Gene family", and "mechanism" at minimum. "Clinical Use" and "Disease" tags are used to allow for an expanded search. The determinant branch describes how the encoded enzymes modify the drugs e.g. "Drug inactivation". The mechanism branch adds general high-level terms for the mechanism e.g. "Antiviral metabolism".

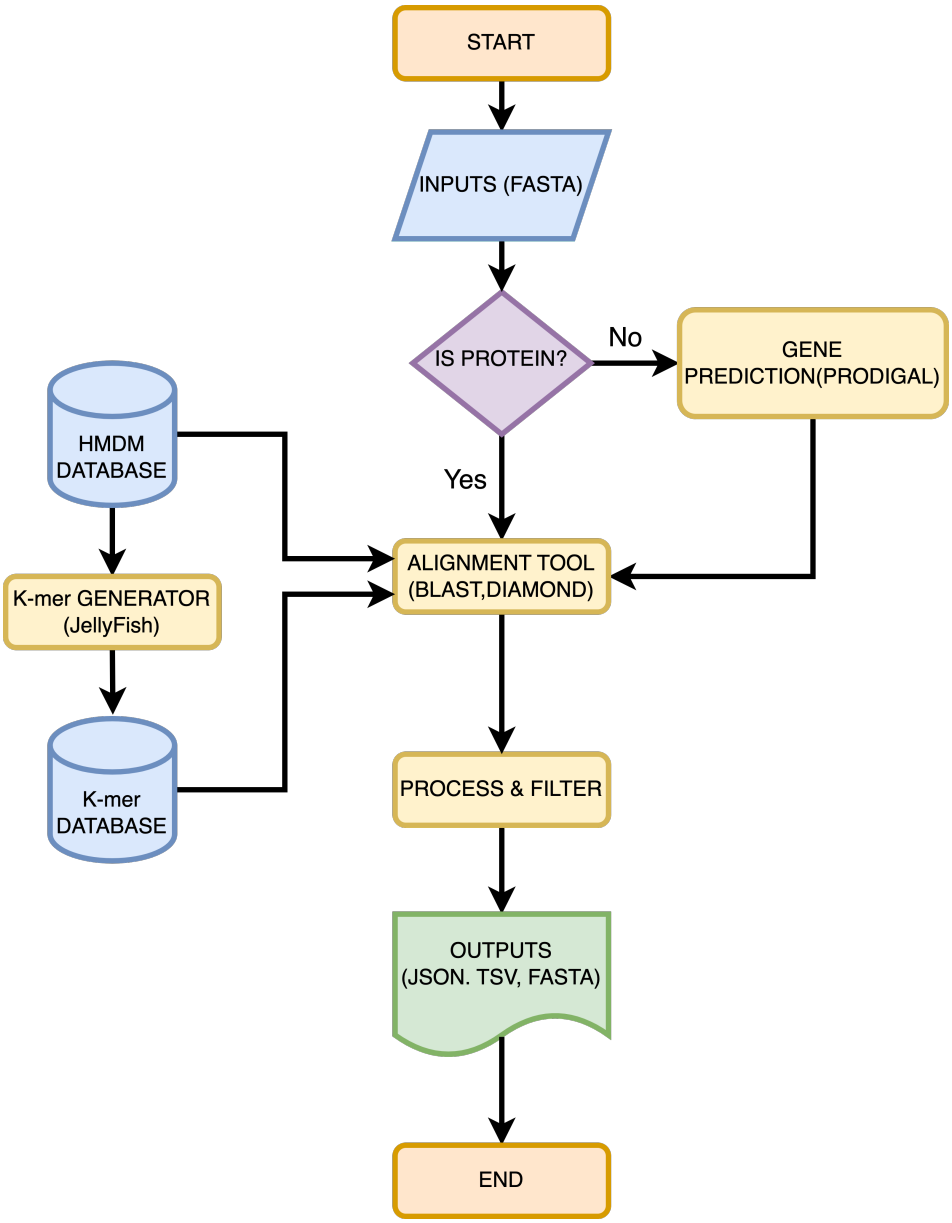


FIGURE 3.3: The Drug Metabolizing Enzyme (DME) software flowchart. The user inputs sequences in FASTA format, if a nucleotide sequence is submitted, Progidal is used to call the genes. The protein sequences are used to align to reference gene sequences in the HMDM database using BLAST or DIAMOND. The results are sorted and annotated with the Perfect-Strict-Loose paradigm depending on the query sequence bitscore obtained relative to the manually curated bitscore for each reference gene. The DMEgt algorithm is used for each predicted gene to predict possible strains carrying the gene. A summary file is provided in tabular format, and a detailed results file is provided in JSON format.

Table 3.2: There are 17 Drug Classes defined in the HMDM.

Accession	Drug Class	Description
HMDM:0000824	anti-diabetic agent	The anti-diabetic agent are used to prevent blood glucose from rising in the treatment of diabetes mellitus.
HMDM:0000791	anti-diarrheal agent	An anti-diarrheal agent is a drug used to slow or stop diarrhea.
HMDM:0000839	anti-inflammatory agent	An anti-inflammatory agent is a drug that helps reduce swelling or inflammation.
HMDM:0000076	anti-neoplastic agent	Anti-neoplastic agents are medications used to treat cancer. They are the most common type of systemic drug therapy to treat cancer. These drugs interfere with cancer cells' ability to grow and spread in a variety of ways. Antineoplastic drugs are also referred to as anticancer, chemotherapy, chemo, cytotoxic, or hazardous drugs.
HMDM:0000261	anti-parkinson agent	The inhibitor used in the therapy of Parkinson's disease.
HMDM:0000005	antiviral	An agent that kills a virus or that suppresses its ability to replicate and, hence, inhibits its capability to multiply and reproduce. For example, amantadine (Symmetrel) is a synthetic antiviral. It acts by inhibiting the multiplication of the influenza A virus.
HMDM:0000490	calcium channel blockers	Calcium ions are essential for the chain of events that leads to myocardial contraction. Its role in the cardiac cycle has been studied extensively for years. Calcium is thought to be effective in slow channels.
CHEBI:83970	cardiac glycoside	Steroid lactones containing sugar residues that act on the contractile force of the cardiac muscles. Cardiac glycosides are a class of organic compounds that increase the output force of the heart and increase its rate of contractions by acting on the cellular sodium-potassium ATPase pump. Their beneficial medical uses are as treatments for congestive heart failure and cardiac arrhythmias; however, their relative toxicity prevents them from being widely used.
HMDM:0000025	cardiac therapeutic	A group of therapeutics targeting the heart.
HMDM:0000021	central nervous system agent	The central nervous system agents are medicines that affect the central nervous system (CNS). There are many different types of drugs that work on the CNS, including anesthetics, anticonvulsants, antiemetics, antiparkinson agents, CNS stimulants, muscle relaxants, narcotic analgesics (pain relievers), nonnarcotic analgesics (such as acetaminophen and NSAIDs), and sedatives.
HMDM:0000074	herbal medication and supplementation	Products made from botanicals, or plants, that are used to treat diseases or to maintain health are called herbal products, botanical products, or phytomedicines. A product made from plants and used solely for internal use is called an herbal supplement.
HMDM:0000281	hormones or endobiotics	Hormones are largely responsible for the integrated communication network responsible for modulating cellular signaling for protein synthesis. All aspects from production, release, transportation, and tissue uptake to intracellular signaling affect the cell signaling and communication that govern the basic activities of cells and coordinate all cellular actions.
CHEBI:50503	laxative	An agent that produces a soft formed stool, and relaxes and loosens the bowels, typically used over a protracted period, to relieve constipation. Compare with cathartic, which is a substance that accelerates defecation. A substances can be both a laxative and a cathartic.
HMDM:0000784	Nitroimidazoles	Nitroimidazoles are a class of antimicrobial drugs that have remarkable broad spectrum activity against parasites, mycobacteria, and anaerobic Gram-positive and Gram-negative bacteria. While nitroimidazoles were discovered in the 1950s, there has been renewed interest in their therapeutic potential, particularly for the treatment of parasitic infections and tuberculosis.
HMDM:0000009	nucleoside antiviral	Nucleoside analogs represent the largest class of small molecule-based antivirals, which currently form the backbone of chemotherapy of chronic infections caused by HIV, hepatitis B or C viruses, and herpes viruses.
HMDM:0000789	Phenothiazine	Phenothiazines belong to the oldest synthetic antipsychotic drugs, which do not have their precursor in the world of natural compounds. Apart from their fundamental neuroleptic action connected with the dopaminergic receptors blockade, phenothiazine derivatives also exert diverse biological activities, which account for their cancer chemopreventive-effect, such as calmodulin- and protein kinase C inhibitory actions, anti-proliferative effect, inhibition of P-glycoprotein transport function and reversion of multidrug resistance.
HMDM:0000708	statins	Statins, inhibitors of the hydroxymethylglutaryl-CoA (HMG-CoA) reductase enzyme, are molecules of fungal origin. Statins are powerful cholesterol lowering medications and have provided outstanding contributions to the prevention of cardiovascular disease.

metagenomes provided by Dr. Michael G. Surette of McMaster University (see Chapter 2).

3.4.6 Sample SHCM1

Similar to the 1,196 genomes (also provided by Dr. Michael G. Surette), the sample SHCM1 (unpublished) was obtained from a healthy adult donor with no history of using antibiotics in the last six months. The sample SHCM1 was from human stool and was prepared and cultured using conditions described by Lau *et al.* [43]. After DNA extraction, 16S rRNA gene sequencing was performed, and the 16S profiles (obtained using Plate Coverage Algorithm (PLCA)) were used to determine representatives from the culture-enriched plates which better represent the bacteria in the sample [43]. The representatives (called cultured-enriched) were then sequenced using shotgun metagenomics. The DNA extraction and library preparation was performed using methods described by Derakhshani *et al.* [80]. Sequencing was done on an Illumina HISeq with a depth of 15-30M paired-end reads (2x150) per sample. The obtained sequencing reads were processed using Fastp [135] (version: v0.23.0) to remove low-quality reads and sequencing primers. We used KneadData pipeline (version: v0.7.2; `https://github.com/biobakery/kneaddata`), which uses bowtie2 [136] (version: v2.4.3) to trim and remove any reads containing human DNA. The remaining reads were assembled using metaSPADE [137] (version: 3.15.1) and binned using Metabat2 [138] (version: V2021). CheckM [84] (version: v1.1.5) was used to identify high-quality metagenomic assembled genomes (MAGs) by identifying single-copy core genes in each bin, and taxonomic classification was performed using GTDB-tk [85] (version: 2.1.0).

3.4.7 Data and Code Available

These data for the HMDM database are available for browsing and download at the https://hmdm.mcmaster.ca website. Each data set is under the download section and is available with accompanying DOIs links at Zenodo (https://zenodo.org). The HMDM ontology is available at https://github.com/raphenya/hmdm-ontology. The DME software is available at https://github.com/raphenya/dme. Please note that the Zenodo links, HMDM ontology, and DME software will be made available after publication of the manuscript describing the database (prior access is available upon request).

3.5 Results

3.5.1 The HMDM Schema

The SQL tables and modules are described in Figure 3.1 and Table 3.1. The HMDM is curated with 120249 records, 349 HMDM terms, eight RO terms, seven MOP terms, 135 ChEBI terms, one MI term, and one GO term. There are 92 unique reference protein sequences curated in the database.

3.5.2 The HMDM Ontology

The HMDM ontology has four main branches (Figure 3.2) from the root node: drugs, classifications, determinants of drug metabolism, and mechanism of drug metabolism. The 'drugs' branch includes drug names and classifications based on other ontologies, such as Chemical Entities of Biological Interest [139] (ChEBI). Some drugs are used in multiple settings; in that case, we used the anatomical

therapeutic chemical classification (ATC; https://www.whocc.no/atc_ddd_index/?code=S01AD02) developed by the World Health Organisation (WHO) to capture these types of drugs. The classification branch defines gene families, drug classes, and mechanisms involved for each gene in the HMDM ontology. The drug metabolism determinant branch describes metabolism requirements for each major drug group, for example, antiviral determinants connected to the gene terms. The mechanism of drug metabolism captures terms explaining how the drugs are being changed, for example, drug activation or inactivation. An additional model ontology (MO) branch connects with all the components from the HMDM at the bottom of the ontology structure. The MO defines genes and their molecular sequences as described in CARD [86].

There are 50 bacterial drug metabolizing genes with complete classification (i.e., drug class, gene family, and mechanism). We have added 100 drugs commonly used in Canada to the ontology (https://studywithclpna.com/medicationadministration/docs/top100meds.pdf), improving the prediction for HMDM*Shark, which triages newly published papers by drug class (outlined below). We have added the clinical use of drugs and diseases associated with the drugs in the HMDM ontology for a few terms. The enzyme *BT_4096* was used for this pilot (see Figure 3.4). This addition will allow for a much more flexible search by disease and clinical use of drugs rather than by only using gene names. Curation of this level of information for all genes is ongoing. We added 30 reported bacterial drug metabolizing genes from Zimmermann and colleagues [127], seven hypothetical proteins, one putative sialic acid-specific acetylesterase, one putative general stress protein (BT_1429), one putative xylanase (BT_1192), and one

putative nitroreductase (BT_1006). Most of these bacterial drug-metabolism-encoded enzymes metabolize several drugs tested by Zimmermann and colleagues. Overall, the ontology includes 349 terms, covering 50 genes and 85 published papers.

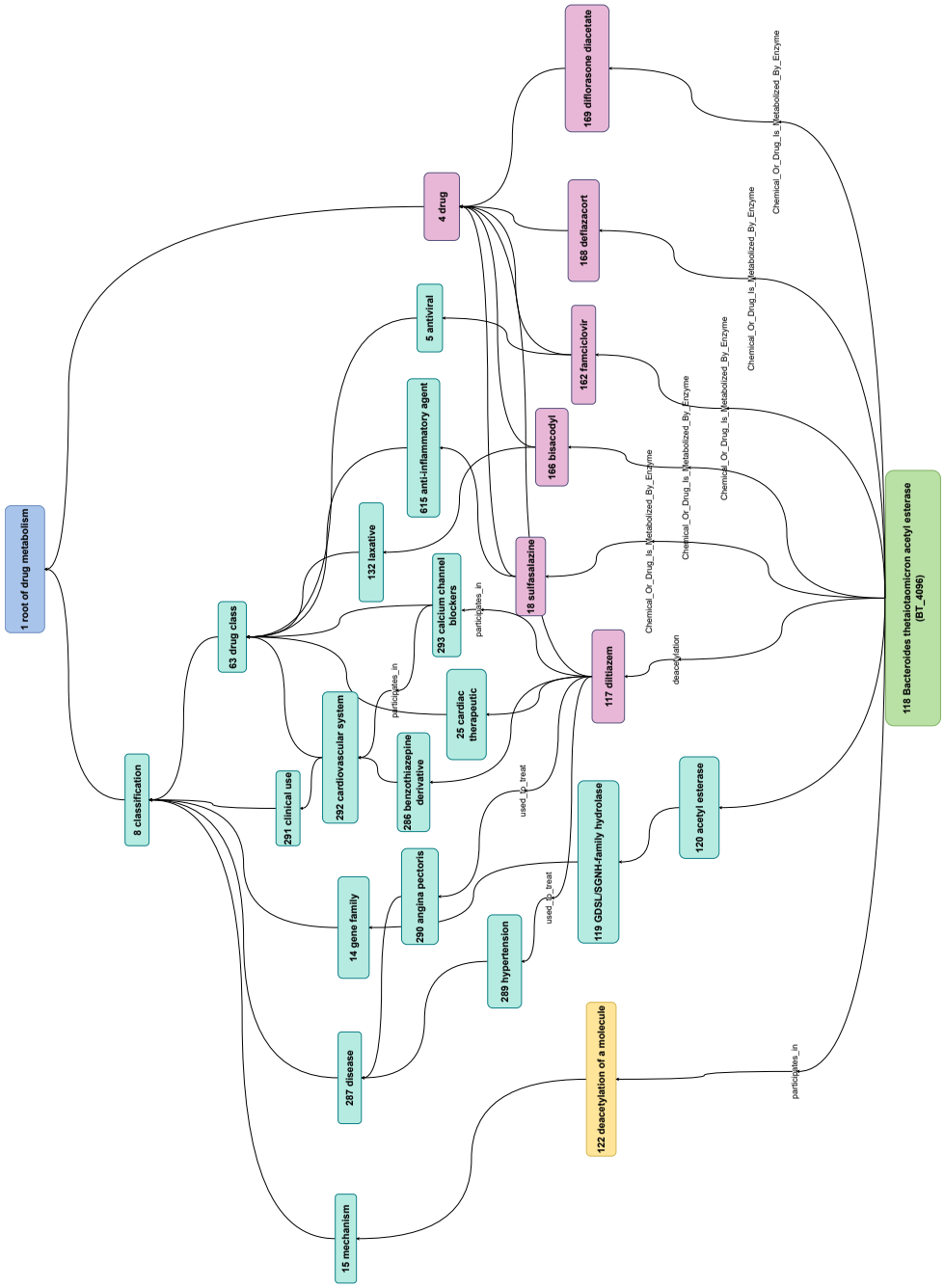


FIGURE 3.4: The figure shows a use case with branches for "clinical use" and "disease" tags added for gene *BT_4096*. The terms with yellow backgrounds have classification tags, and green terms have a model attached.

3.5.3 The HMDM Ontology Curation Rules

The genes curated into the HMDM database, at a minimum, should have a drug to gene relationship, *is_a* relationship to enzyme family, and the biotransformation mechanism performed by the bacterial gene product (i.e., enzyme) towards the drug. A publication describing the experiment must be added, with accompanying molecular information (i.e., protein sequence, nucleotide sequence, etc.), and a bit-score cutoff for its bioinformatics annotation model. Currently, the HMDM database exclusively uses the Comprehensive Antibiotic Resistance Databases' Protein Homolog Model (PHM) annotation model type [86] to annotate a protein sequence based on its similarity to a curated reference sequence in the HMDM database.

3.5.4 Predicting bacterial drug metabolizing genes using novel Drug Metabolizing Enzyme (DME) software

The DME accepts protein sequences as input which are aligned to protein reference sequences from the HMDM database using NCBI's BLAST [140] (version 2.9.0) or DIAMOND (version 0.8.36) [141,142] BLASTp algorithm. Prodigal (version v2.6.3) [87] predicts genes when the user submits a nucleotide genome or assembly contig. The predicted protein sequences from Prodigal are then used as described above. Each gene in the HMDM is manually annotated with a similarity score called bit-score to distinguish it from other genes that likely metabolize other small molecules. The DME software uses similar concepts from the Resistance Gene Identifier (RGI; `https://github.com/arpcard/rgi`) software such that the DME gene prediction annotation is assigned using CARD's Perfect, Strict, and

Loose paradigm. A Perfect annotation means the predicted protein completely matches the reference protein sequence in the HMDM. A Strict annotation passes the manually curated bit-score and is likely a functional variant, and the Loose annotation falls below the set bit-score, hence a distant homolog or spurious match (Supplementary Figure 3.11). The DME software flowchart is depicted in Figure 3.3.

We analyzed three genomes with known drug-metabolizing genes using DME and other tools (GutBug [58], MASI [51]) to measure DME performance based on predicted gene absence or presence [143]. True positives (TP) were defined as genes that were correctly predicted to be present, false negatives (FN) were genes that were incorrectly predicted to be absent, and false positives (FP) were genes that were incorrectly predicted to be present. Recall, precision, and F1 scores were calculated using the following formulae:

$$Precision = \frac{TP}{TP + FN}$$

$$Recall = \frac{TP}{TP + FP}$$

$$F1 = 2 * \frac{(Precision * Recall)}{(Precision + Recall)}$$

Two genomes were sourced from NCBI: *Enterococcus faecalis* V583 (accession: ASWP01000005.1) and *Eggerthella lenta* DSM 2243 (accession: CP001726.1), both with papers outlining functional characterization of their drug metabolism genes (`https://hmdm.mcmaster.ca/cvterms/24` and `https://hmdm.mcmaster.ca/cvterms/36`). One genome was from Dr. Surette's 1,196 genomes (sample GC796,

Bacteroides thetaiotaomicron). As such, only DME Perfect annotations were used from sample GC796 to ensure that the genes predicted are truly functional (assuming expression). Overall, the DME performed better than GutBug and MASI by predicting all drug-metabolizing genes in all three genomes (Table 3.3). Notably, these two other tools are only available via manual web interfaces and cannot be used for bulk analyses like DME at the command line.

TABLE 3.3: DME performance in prediction of drug-metabolizing genes vs. MASI and GutBug.

Species	**Metric**	**DME**	**MASI**	**GutBug**
Bacteroides thetaiotaomicron	Precision	1.0	1.0	N/A
Enterococcus faecalis V583	Recall	1.0	0.2	0
Eggerthella lenta DSM 2243	F1	1.0	0.33	N/A
Bacteroides thetaiotaomicron	Precision	1.0	N/A	N/A
Enterococcus faecalis V583	Recall	1.0	0	0
Eggerthella lenta DSM 2243	F1	1.0	N/A	N/A
Bacteroides thetaiotaomicron	Precision	1.0	1.0	N/A
Enterococcus faecalis V583	Recall	1.0	1.0	0
Eggerthella lenta DSM 2243	F1	1.0	1.0	N/A

3.5.5 Predicting bacterial strains using the DME's gene triad

To determine specific bacterial strains for the predicted drug-metabolizing genes using DME, we implemented a Gene-Triad (DMEgt) algorithm. The DMEgt involves using bacterial genes curated in the HMDM database in the following four steps;

1. Identifying two flanking genes for each gene curated from a genome annotation file (i.e., GFF, General Feature Format formatted file)

2. Calculating k-mers (using k size 63) for the 3 genes (HMDM entry + 2 flanking genes) and tying them to a specific bacterial strain

3. Create a custom BLAST database using the nucleotide k-mers

4. Use BLASTn for strain prediction via shared k-mers

The DMEgt performed well in predicting strains encoding the drug-metabolizing genes as part of the DME prediction. We were able to accurately predict the source species for the sample GC1188 as *Escherichia coli* ATCC 8739 using DMEgt, which had no GTDBTk predictions (Chapter 2). The GTDBTk software predicts genome taxonomy classification using reference trees based on bacterial and archaeal genomes [85]. GC813 was predicted as from *Phocaeicola vulgatus* via GTDBTk, but DMEgt predicted possible strains for GC813 as *Collinsella aerofaciens* ATCC 25986 and *Phocaeicola dorei* DSM 17855. The species *Phocaeicola dorei* has been misidentified as *Phocaeicola vulgatus* by a previous study as they are closely related

[144]. As a result, the sample GC813 would require further analysis or resequencing to determine specific strain and species.

3.5.6 Text mining literature using HMDM*Shark

HMDM*Shark uses two sets of abstracts, one from the HMDM database searched using drug classes defined by the Human Microbiome Drug Metabolism Ontology (or simply HMDM*Ontology) and a second set from PubMed (`https://pubmed.ncbi.nlm.nih.gov`) using the drug class names from HMDM. The set from HMDM is called target and from PubMed is called global. The two sets create a scoring matrix from which to rank each new paper searched in PubMed. The title and abstract are used to identify the frequency of words appearing in both sets. If a word appears more in the global set, it is assigned a low score, otherwise, a high score if it's more frequent in the target set. Papers are retrieved for the last 30 days from PubMed using Biopython's Entrez module. The drug classes used by HMDM*Shark are defined by the HMDM*Ontology, e.g., antiviral, central nervous system agent, antineoplastic agent, cardiac therapeutic, and calcium channel blockers (see Table 3.2 for the complete list). HMDM*Shark is modeled after the CARD*Shark 2 algorithm used to search for literature for the CARD database [133]. The HMDM*Shark has been run four times and found 14 papers, most of which are curated into the HMDM database, but other genes have not been curated because the drug is not FDA-approved (such as albiflorin, with four esterases that metabolize this drug).

3.5.7 Determining bacterial drug-metabolizing gene frequencies using the HMDM*Prevalence Module

Currently, there is no easy way to understand the prevalence of bacterial drug-metabolizing genes in the human gut microbiome. Knowing the frequency of bacterial drug-metabolizing genes, we can better prioritize the encoded enzymes contributing to poor drug efficacy during drug development. A prevalence module for the HMDM database was used to assess the frequency of the bacterial drug-metabolizing genes from the phyla commonly found in the human gut microbiome. This module uses genomic data from the National Center for Biotechnology Information (NCBI) Datasets (https://www.ncbi.nlm.nih.gov/datasets). For the pilot of this module, only complete chromosomes (6,855) and plasmid assemblies (11,428) were downloaded, totaling 18,283 FASTA files (downloaded on 27 March 2022). These sequences were analyzed using the DME software to predict potential bacterial drug-metabolizing genes based on curated genes in the HMDM database. We built this module as a custom snakemake automated pipeline that cleans and updates seven SQL tables (Figure 3.5) with new data on each prevalence run. The HMDM*Prevalence module is run every month to obtain the frequency of bacterial drug-metabolizing genes. All three approaches outlined below were used to curate the comprehensive list of bacterial strains for the prevalence module in the HMDM database. This list will be used moving forward and will be updated as more genes are reported or curated into the HMDM database.

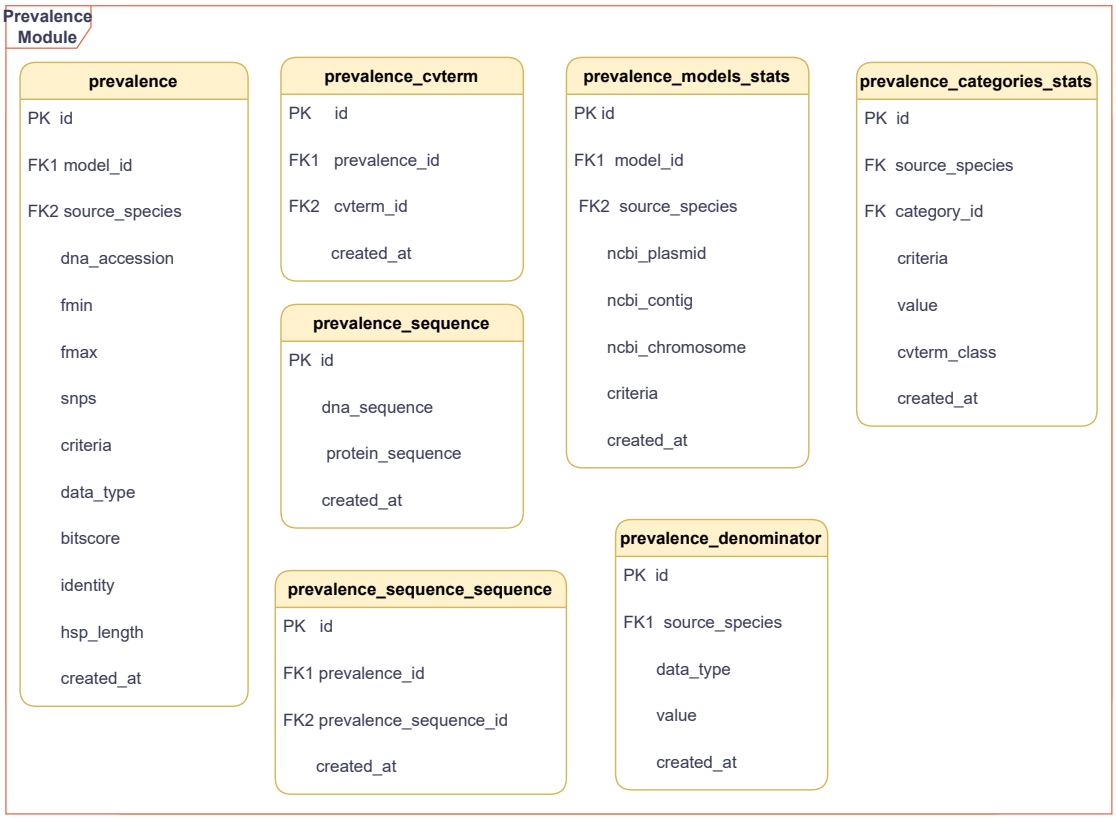


FIGURE 3.5: The figure shows seven SQL tables that are used for the prevalence module. The main table is "prevalence," which records the prevalence data, which includes accessions and coordinates to the gene annotation in the analyzed genome. The "prevalence_cvterm" records ontology terms related to the annotated gene. The sequence information is recorded using "prevalence_sequence" and "prevalence_prevalence_sequence". The statistics are stored using "prevalence_categories_stats", "prevalence_models_stats", and "prevalence_denominator".

3.5.8 Results for selecting taxa to use for the HMDM*Prevalence Module

Approach 1 - Using strains tied to drug-metabolizing genes

The first analysis was performed on 27 March 2022 after completing the HMDM*Prevalence pipeline. Currently, this module is biased toward 215 bacterial species found in the human gut. There were 7,164 Perfect and Strict annotations amongst the 18,283 analyzed genomes (Figure 3.6). Twenty-seven unique genes had Perfect annotations and 36 unique genes had Strict annotations to genes curated in the HMDM database. All the Perfect annotations were from chromosomes only. The highest frequency gene was from *Escherichia coli* O139:H28 str. E24377A β-glucuronidase (*uidA*) at 10.77% followed by *Escherichia coli* uridine phosphorylase (*udp*) at 9.98%. Third was *Enterococcus faecalis* tyrosine decarboxylase (TDC) at 6.21%, and fourth was *Escherichia coli* thymidine phosphorylase (*deoA*) at 3.37% for Perfect annotations. For Strict annotations, *udp* and *deoA* are at 37.21% and 37.2%, respectively. The *deoA* gene is 0.01% or ~115 of the analyzed plasmids. The β-glucuronidase genes are more common in these data sets, suggesting they are more prevalent.

Approach 2 - UniProt Proteome

We downloaded reference proteomes totaling 8821 proteins (6.9GB) curated by UniProt on 13-December-2022 for UniProt release 2022_05. The proteomes were analyzed using DME and 119 of 8821 files produced annotations to the HMDM genes. We identified 20 bacterial families associated with the genes in the HMDM database (Figure 3.7). The β-glucuronidase genes are also encoded in multiple

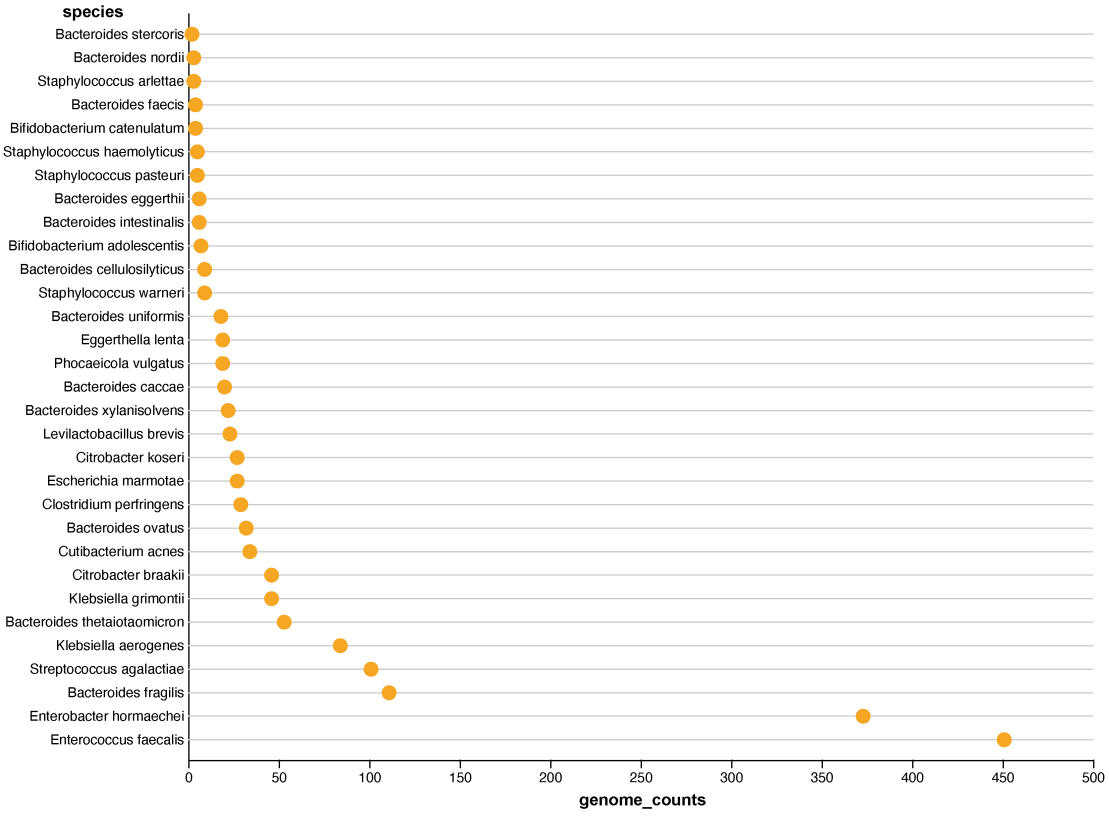


FIGURE 3.6: The species that produced annotations for HMDM*Prevalence run on 27 March 2022 from NCBI datasets genomes. *Escherichia coli* (not shown in the figure) had 7,913 genomes with DME annotations.

bacterial families.

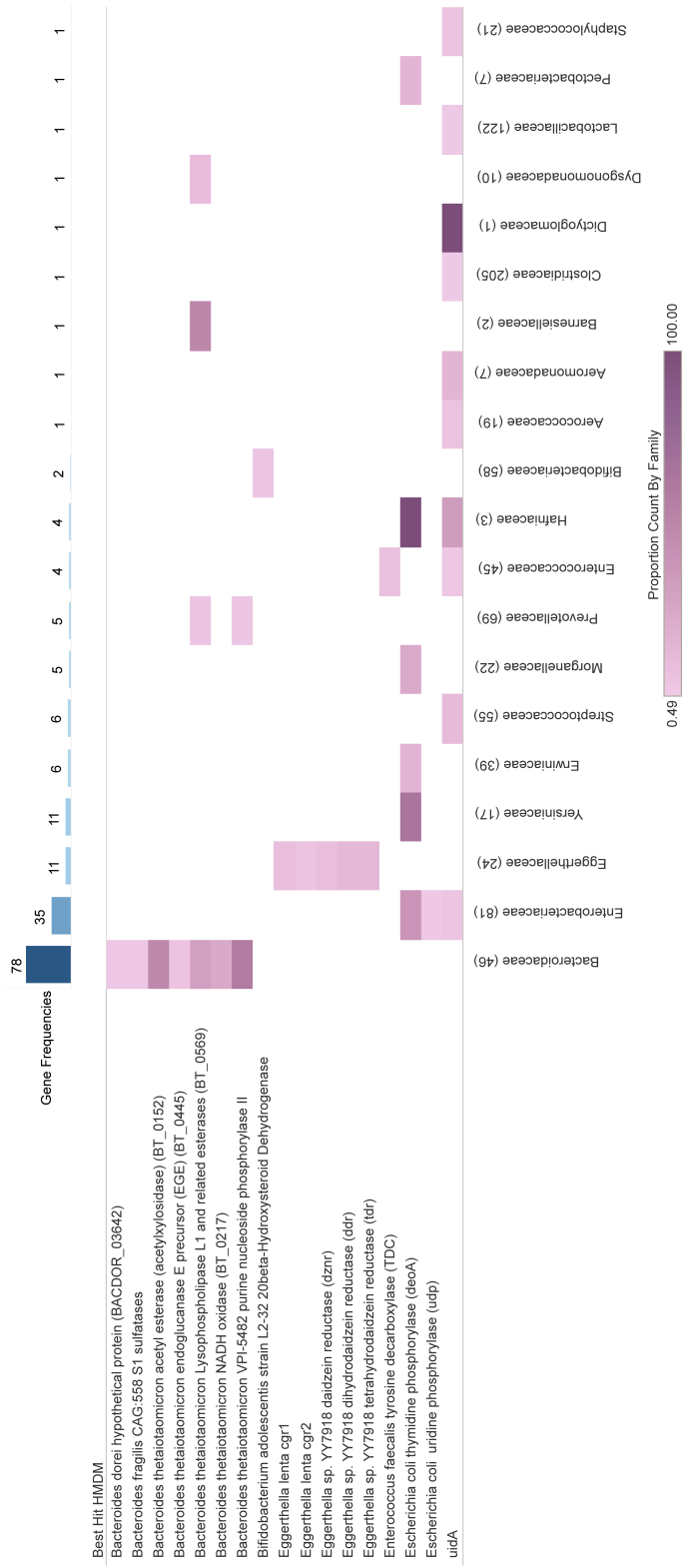


FIGURE 3.7: The figure shows HMDM genes predicted from the UniProt Reference Proteomes plotted as a heatmap for the percentage of genes predicted within each family and gene frequencies. The gene *uidA* encoded in different species was collapsed to *uidA*. The number in brackets is proteomes sampled from each bacterial family.

Approach 3 - In-house gut microbiome and culture-enriched metagenomes

Using DME, we analyzed 1,196 in-house genomes and 1 (sample SHCM1) culture-enriched metagenome assembly. 18 gene families were identified from the 1,196 genomes (Figure 3.8) and β-glucuronidases were also in multiple bacterial families. Antivirals were the most metabolized by the genes predicted in the 1,196 genomes (see Supplementary Figure 3.14). Sample SHCM1 produced annotations to 1 class, 2 families, 2 genera, 3 orders, and 9 species. Taxonomic assignment was performed using kraken2 [145] on each of the 35 contigs in sample SHCM1 (Figure 3.9).

FIGURE 3.8: The frequency of the HMDM genes for the 1,196 in-house genomes (bar plot in blue). The genes are on the y-axis, and bacterial gene families are on the x-axis. The heatmap shows the percentage of genes predicted within each family. The gene *uidA* encoded in different species was collapsed to *uidA*. The number in brackets is genomes sampled from each bacterial family.

3.6 Discussion

We rely on drugs to prevent and manage diseases [1]. Still, we have limited knowledge of how these drugs can affect our bodies, and most of the research has been on the host drug metabolism, neglecting the microbes' ability to metabolize drugs [11]. Our knowledge is not easily accessible or available in a portable format that can be re-used to answer research questions, such as which bacterial family carries the most drug-metabolizing genes. To leverage the wealth of information available via next-generation sequencing, we need a gold standard data set to parse genomic data and make sense of it [35–37]. There are a few resources (MASI [51], PharmacoMicrobiomic [52], Disbiome [53], gutMDisorder [54], MagMD [55], MicrobeFDT [56], SIMMER [57], and GutBug [58]) developed for predicting microbial drug metabolism but they lack completeness, genomic information, and are not accessible for bulk analysis.

We developed the HMDM database to systematically catalog all known drug-metabolizing genes in the human gut microbiome. The HMDM uses an ontology centric approach to catalog determinants of bacterial drug metabolism, i.e., genes that encode drug-metabolizing enzymes. The ontology allows for a concise description of drug metabolism concepts and, at minimum, the genes are connected to at least four classification terms, namely, "Drug Class", "Gene Family", "Mechanism", and "Drug" (see Figure 3.10, Supplementary Figure 3.12, and Supplementary Figure 3.13). Each gene in the HMDM database is connected to the drug it metabolizes by using the enzyme it encodes, i.e., the essential drug-to-gene relationship. These classifications are used to annotate results when using our annotation software (DME). By annotating results with classification

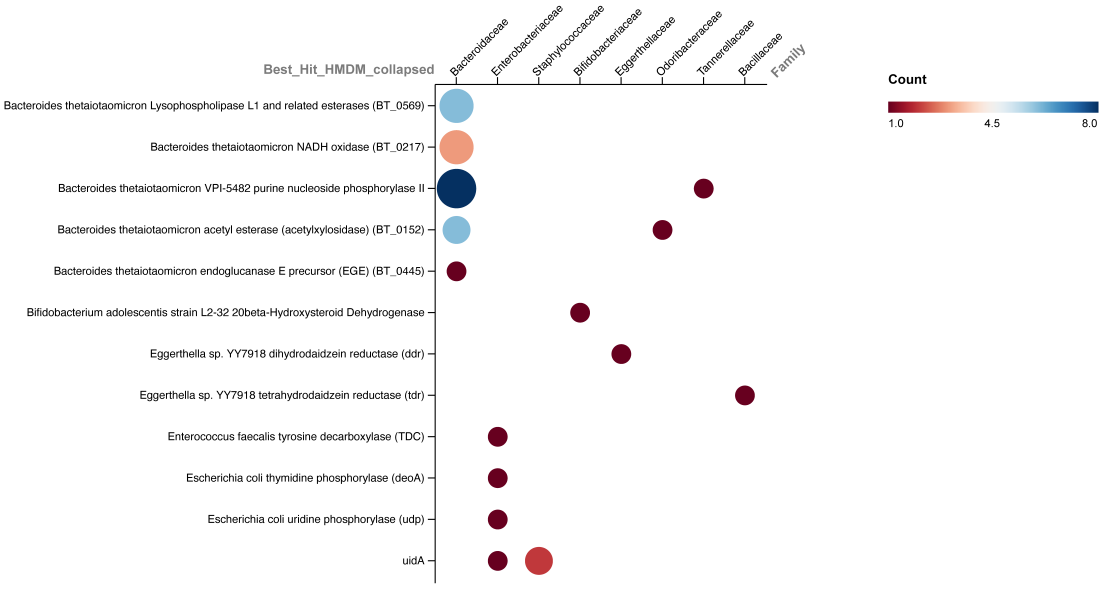


FIGURE 3.9: The HMDM*Prevalence culture-enriched metagenome sample SHCM1. There are 12 drug metabolizing genes in this sample covering eight bacterial families with most genes in *Bacteroidaceae* family. The gene *uidA* encoded in different families was collapsed to *uidA*.

terms, it makes it easier to compare and summarize results from different samples. When new drug-metabolizing genes are reported in the literature, they can be easily curated into the HMDM database based on these classifications and specific branches in the HMDM ontology. For example, if another gene encoding a lyase enzyme is identified, it will be connected to the "lyases family" (Figure 3.10), with connections to "Drug Class", "Mechanism", and the drug(s) metabolized. Like for CARD, these data in the HMDM database can be used to develop machine learning algorithms since the knowledge is structured, portable, and reusable [146].

The HMDM database is also accompanied by prediction software, the DME, which uses gene sequence data and ontological structure to predict known drug-metabolizing genes and their variants in bacterial genomes with a higher level of accuracy than existing tools (Table 3.3). The associated DMEgt algorithm can help predict known drug-metabolizing genes in different gene neighborhoods. For example, if a gene is predicted to be a Perfect annotation, and the k-mers annotations do not have 100% identity, it is an indication of new gene arrangements or mutation. The DMEgt was implemented as a targeted algorithm and uses a smaller database as compared to algorithms like Kraken2.

One key piece of information in drug development is the prevalence of genes or mutations that could lead to unexpected outcomes [147,148]. For this, the HMDM provides a separate data set (HMDM*Prevalence) from the "canonical" or known genes by mining the NCBI Datasets or in-house data sets to determine frequencies for each gene in the HMDM database relative to obtained genomes for each data release. As a result, this will help prioritize bacterial genes that might metabolize oral drugs being developed. We can essentially use the

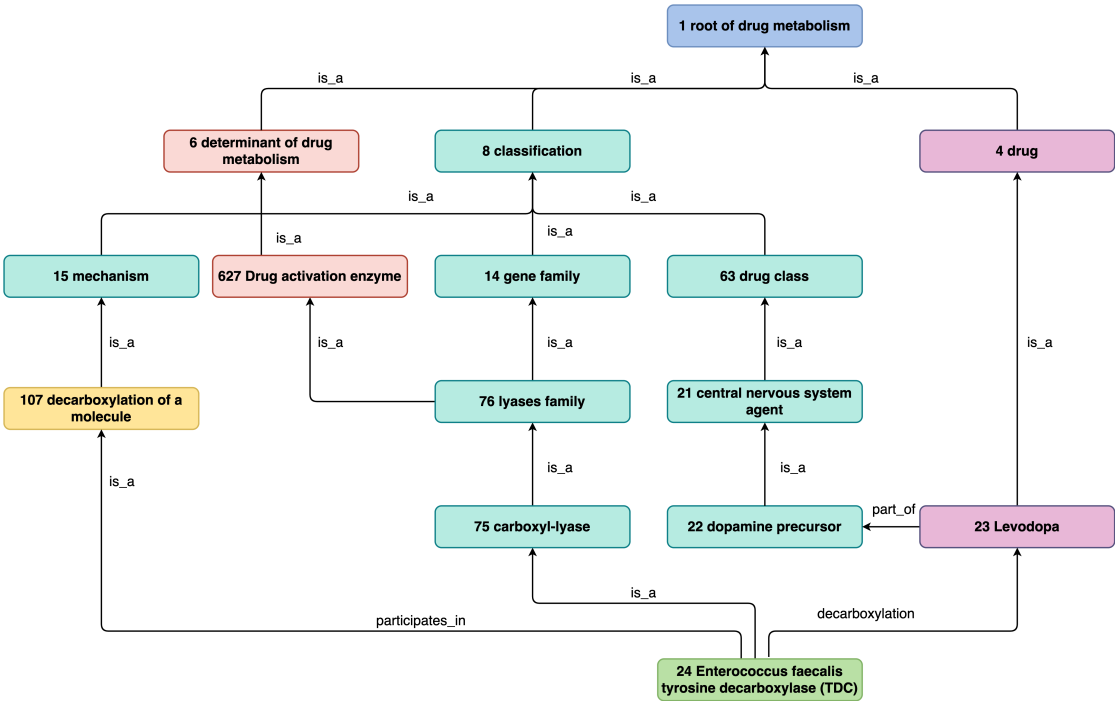


FIGURE 3.10: The *Enterococcus faecalis* tyrosine decarboxylase (TDC) gene (in green) which encodes an enzyme that decarboxylates (the mechanism, in blue) a Parkinson's disease drug Levodopa (purple branch). The TDC belongs to a lyases enzyme family (blue branch), and Levodopa is an agent that targets the central nervous system. The numbers indicate the HMDM cvterm ids which are used to locate the terms (e.g. https://hmdm.mcmaster.ca/cvterms/24 for the TDC page)

HMDM*Prevalence data set to profile genomes from specific data sets and we used three approaches to assess HMDM gene frequencies. The first approach revealed that drug-metabolizing genes are predominantly located in bacterial chromosomes, not mobile plasmids. In addition, the β-glucuronidases-glucuronidases were most common among bacterial genomes, which was also supported by the other two approaches and previous studies [149–151]. The β-glucuronidases are primarily used to metabolize carbohydrates, deconjugate glucuronides, and hydrolyze 0- or S-glycosidic moieties from glycosides, releasing aglycones [149]. The second and third approaches showed that most drug-metabolizing genes are carried by the *Bacteroidaceae* family. The *Bacteroides* species are dominant from birth to adulthood and have been shown to be both protective flora and opportunistic pathogens [152,153]. Diet has been associated with *Bacteroides* species by Ferrocino and colleagues [154]. They analyzed microbiomes from 153 healthy volunteers in three diet groups (i.e., vegans, vegetarians, and omnivores), with 51 individuals in each group. The results showed that *Bacteroides fragilis* was more abundant in the omnivores and less in the other groups. Even though some *Bacteroides* species can be tied to a particular diet, there is variability which can be due to factors such as host genetics [155]. Since our results show that the HMDM drug-metabolizing genes are located in the bacterial chromosomes, we could use these data for profiling specific diets as a predisposition to certain bacterial drug-metabolizing genes.

The HMDM database is built to be expandable as new drug-metabolizing genes are discovered, as we can update the database and thus improve DME software predictions via curator interfaces. A combined prediction of bacterial and host

drug metabolism enzymes will provide a better view of drug transformation and efficacy. The HMDM database provides a platform for bacterial drug metabolism prediction.

3.6.1 Limitations for HMDM*Shark

We used the CARD*Shark 2 approach for HMDM*Shark to triage papers, but this method underperforms and highly ranks unrelated papers even for a small database such as HMDM with under 100 papers curated [133]. We reasoned that since the corpus of bacterial drug metabolism of drugs has fewer publications, HMDM*Shark 2 was sufficient for our purpose of triaging papers despite its high rate of false predictions, i.e. the total number of papers to review is low. CARD*Shark 2 has high recall, i.e. rarely misses valuable papers, and we anticipate the same for HMDM*Shark [133]. In addition, we curated relevant terms specific to bacterial drug metabolism of drugs, which included words such as "microbiome-encoded enzymes", "gut microbiome drug metabolized", "drug deglycosylation by human gut microbiome", and "drug metabolism by *Eggerthella*" to improve the predictions. We also filtered more stopwords (e.g., "in", "the" etc.) using the Natural Language Toolkit (NLTK) library [156], which surpassed CARD*Shark 2's stopwords, thus improving the related words used for scoring the papers.

3.7 Conclusion

The bioinformatics models curated into the HMDM from the literature can help annotate genomes for bacterial drug-metabolizing genes and help predict novel homologs. The limitation of this study (i.e., DME predictions based on the

HMDM) is the assumption that predicted genes will lead to the expression, resulting in drug metabolism. The HMDM database is a vital resource to help identify bacterial drug-metabolizing genes and can be used to assess unexpected outcomes for candidate drugs in drug discovery pipelines. The HMDM database provides a gold standard data set that can be used in machine learning algorithms to understand bacterial drug metabolism in the human gut. This resource will help reduce costs and time by identifying potential bacterial drug metabolism early in drug development and could lead to personalized medicine.

3.8 Supplementary material

3.8.1 Supplementary Figures

FIGURE 3.11: DME paradigm at a glance. Perfect annotation matches the "Reference Gene", Strict annotation passes manually similarity score curated the "Bitscore Cutoff" (set at 400) and Loose annotation falls below the cutoff.

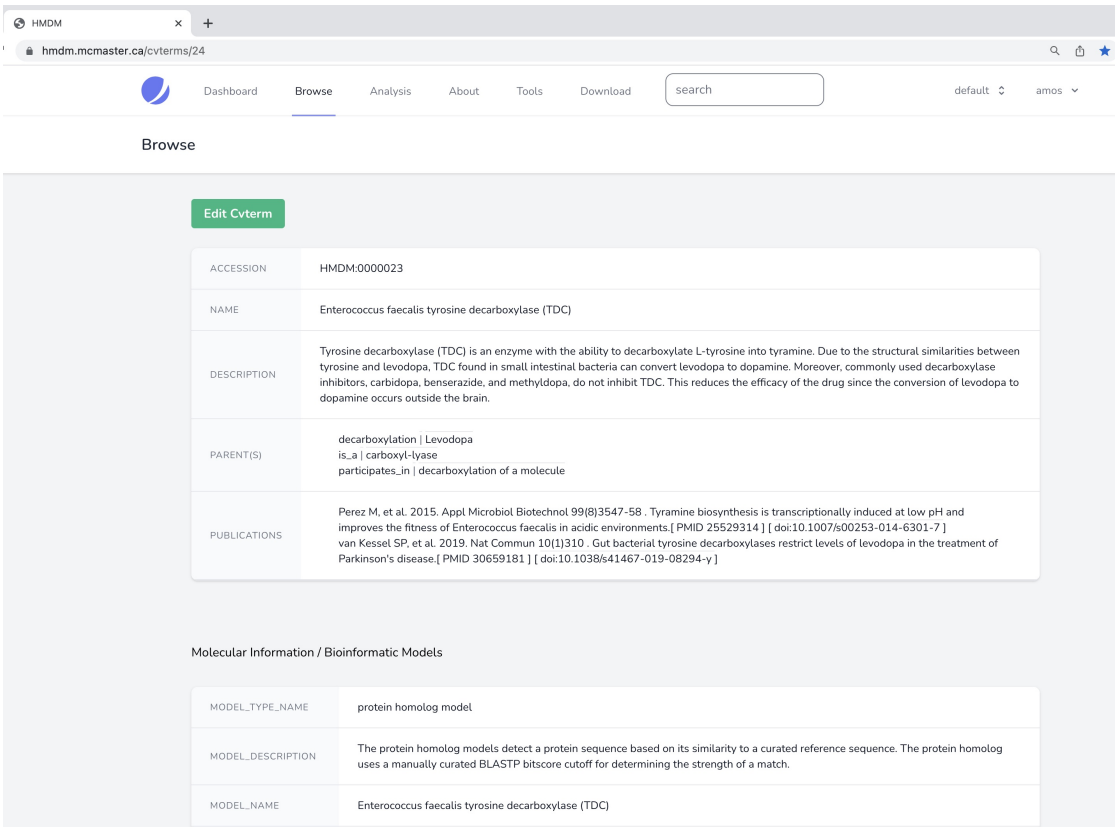
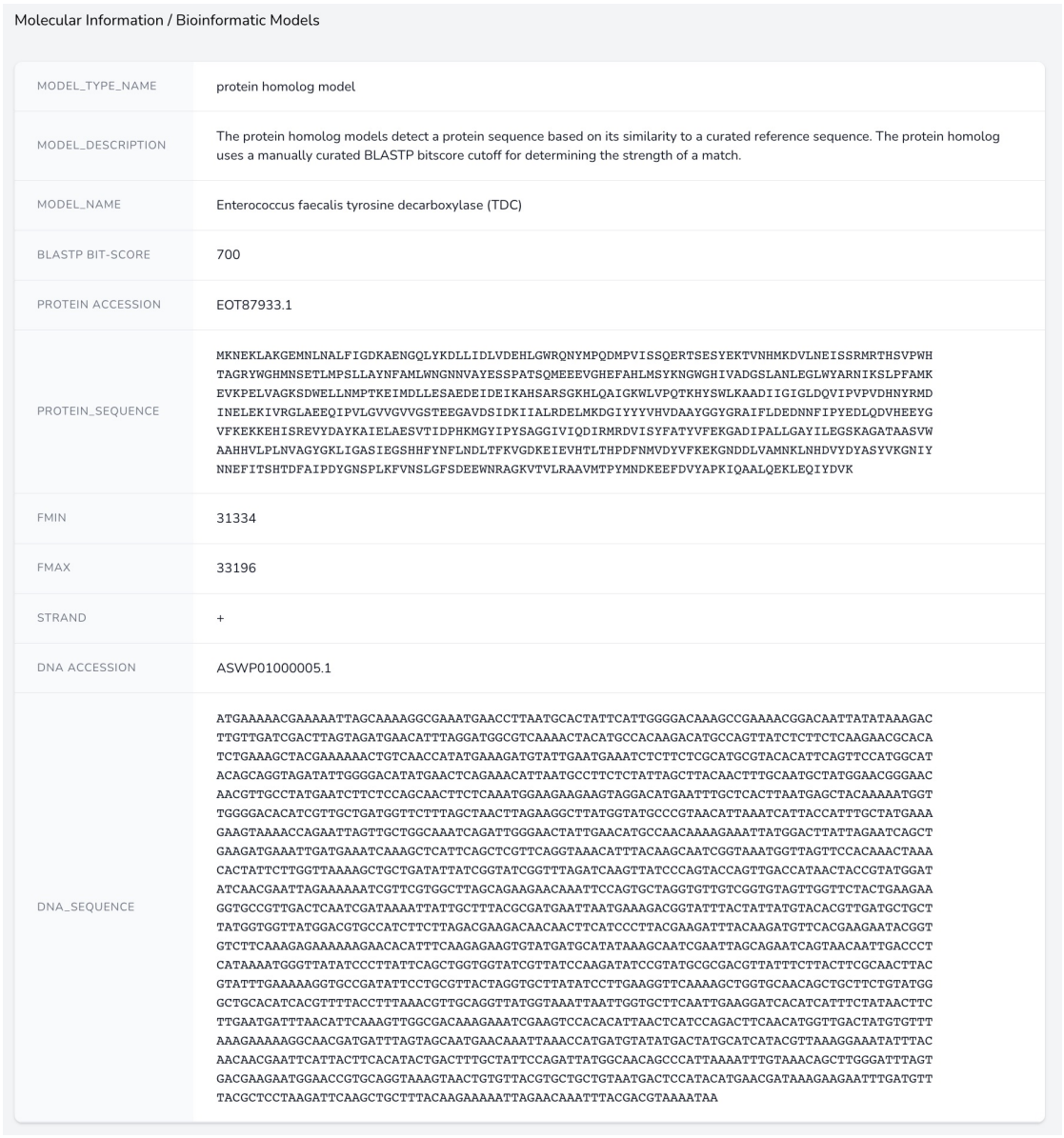


FIGURE 3.12: Page view for the gene TDC in the HMDM database showing gene name, description, publications, and immediately connected cvterms.



Molecular Information / Bioinformatic Models

MODEL_TYPE_NAME	protein homolog model
MODEL_DESCRIPTION	The protein homolog models detect a protein sequence based on its similarity to a curated reference sequence. The protein homolog uses a manually curated BLASTP bitscore cutoff for determining the strength of a match.
MODEL_NAME	Enterococcus faecalis tyrosine decarboxylase (TDC)
BLASTP BIT-SCORE	700
PROTEIN ACCESSION	EOT87933.1
PROTEIN_SEQUENCE	MKNEKLAKGEMNLNALFIGDKAENGQLYKDLLIDLVDEHLGWRQNYMPQDMPVISSQERTSESYEKTVNHMKDVLNEISSRMRTHSVPWH TAGRYWGHMNSETLMPSLLAYNFAMLWNGNNVAYESSPATSQMEEEVGHEFAHLMSYKNGWGHIVADGSLANLEGLWYARNIKSLPFAMK EVKPELVAGKSDWELLNMPTKEIMDLLESAEDEIDEIKAHSARSGKHLQAIGKWLVPQTKHYSWLKAADIIGIGLDQVIPVPVDHNYRMD INELEKIVRGLAEEQIPVLGVVGVVGSTEEGAVDSIDKIIALRDELMKDGIYYYVHVDAAYGGYGRAIFLDEDNNFIPYEDLQDVHEEYG VFKEKKEHISREVYDAYKAIELAESVTIDPHKMGYIPYSAGGIVIQDIRMRDVISYFATYVFEKGADIPALLGAYILEGSKAGATAASVW AAHHVLPLNVAGYGKLIGASIEGSHHFYNFLNDLTFKVGDKEIEVHTLTHPDFNMVDYVFKEKGNDDLVAMNKLNHDVYDYASYVKGNIY NNEFITSHTDFAIPDYGNSPLKFVNSLGFSDEEWNRAGKVTVLRAAVMTPYMNDKEEFDVYAPKIQAALQEKLEQIYDVK
FMIN	31334
FMAX	33196
STRAND	+
DNA ACCESSION	ASWP01000005.1
DNA_SEQUENCE	ATGAAAAACGAAAAATTAGCAAAAGGCGAAATGAACCTTAATGCACTATTCATTGGGGACAAAGCCGAAAACGGACAATTATATAAAGAC TTGTTGATCGACTTAGTAGATGAACATTTAGGATGGCGTCAAAACTACATGCCACAAGACATGCCAGTTATCTCTTCTCAAGAACGCACA TCTGAAAGCTACGAAAAAACTGTCAACCATATGAAAGATGTATTGAATGAAATCTCTTCTCGCATGCGTACACATTCAGTTCCATGGCAT ACAGCAGGTAGATATTGGGGACATATGAACTCAGAAACATTAATGCCTTCTCTATTAGCTTACAACTTTGCAATGCTATGGAACGGGAAC AACGTTGCCTATGAATCTTCTCCAGCAACTTCTCAAATGGAAGAAGAAGTAGGACATGAATTTGCTCACTTAATGAGCTACAAAAATGGT TGGGGACACATCGTTGCTGATGGTTCTTTAGCTAACTTAGAAGGCTTATGGTATGCCCGTAACATTAAATCATTACCATTTGCTATGAAA GAAGTAAAACCAGAATTAGTTGCTGGCAAATCAGATTGGGAACTATTGAACATGCCAACAAAAGAAATTATGGACTTATTAGAATCAGCT GAAGATGAAATTGATGAAATCAAAGCTCATTCAGCTCGTTCAGGTAAACATTTACAAGCAATCGGTAAATGGTTAGTTCCACAAACTAAA CACTATTCTTGGTTAAAAGCTGCTGATATTATCGGTATCGGTTTAGATCAAGTTATCCCAGTACCAGTTGACCATAACTACCGTATGGAT ATCAACGAATTAGAAAAAATCGTTCGTGGCTTAGCAGAAGAACAAATTCCAGTGCTAGGTGTTGTCGGTGTAGTTGGTTCTACTGAAGAA GGTGCCGTTGACTCAATCGATAAAATTATTGCTTTACGCGATGAATTAATGAAAGACGGTATTTACTATTATGTACACGTTGATGCTGCT TATGGTGGTTATGGACGTGCCATCTTCTTAGACGAAGACAACAACTTCATCCCTTACGAAGATTTACAAGATGTTCACGAAGAATACGGT GTCTTCAAAGAGAAAAAAGAACACATTTCAAGAGAAGTGTATGATGCATATAAAGCAATCGAATTAGCAGAATCAGTAACAATTGACCCT CATAAAATGGGTTATATCCCTTATTCAGCTGGTGGTATCGTTATCCAAGATATCCGTATGCGCGACGTTATTTCTTACTTCGCAACTTAC GTATTTGAAAAAGGTGCCGATATTCCTGCGTTACTAGGTGCTTATATCCTTGAAGGTTCAAAAGCTGGTGCAACAGCTGCTTCTGTATGG GCTGCACATCACGTTTTACCTTTAAACGTTGCAGGTTATGGTAAATTAATTGGTGCTTCAATTGAAGGATCACATCATTTCTATAACTTC TTGAATGATTTAACATTCAAAGTTGGCGACAAAGAAATCGAAGTCCACACATTAACTCATCCAGACTTCAACATGGTTGACTATGTGTTT AAAGAAAAAGGCAACGATGATTTAGTAGCAATGAACAAATTAAACCATGATGTATATGACTATGCATCATACGTTAAAGGAAATATTTAC AACAACGAATTCATTACTTCACATACTGACTTTGCTATTCCAGATTATGGCAACAGCCCATTAAAATTTGTAAACAGCTTGGGATTTAGT GACGAAGAATGGAACCGTGCAGGTAAAGTAACTGTGTTACGTGCTGCTGTAATGACTCCATACATGAACGATAAAGAAGAATTTGATGTT TACGCTCCTAAGATTCAAGCTGCTTTACAAGAAAAATTAGAACAAATTTACGACGTAAAATAA

FIGURE 3.13: Page view for the gene TDC in the HMDM database showing bioinformatic model (with sequence annotations).

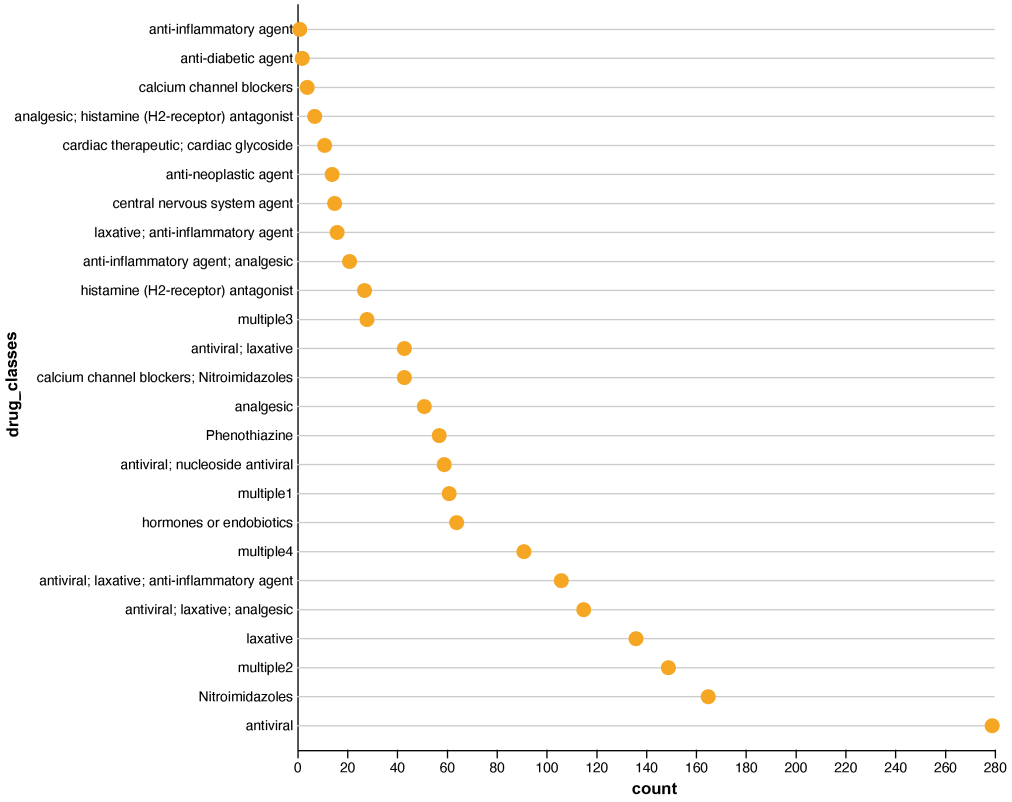


FIGURE 3.14: The drug classes affected by the predicted genes in the 1,196 genomes. Antivirals are the most metabolized, and there are genes that metabolize drugs several drugs. (multiple1 = anti-parkinson agent; calcium channel blockers; Nitroimidazoles; anti-inflammatory agent; catechol-O-methyltransferase (COMT) inhibitor), (multiple2 = anti-parkinson agent; calcium channel blockers; Nitroimidazoles; catechol-O-methyltransferase (COMT) inhibitor), (multiple3 = antiviral; cardiac therapeutic; laxative; calcium channel blockers; anti-inflammatory agent), (multiple4 = antiviral; laxative; anti-inflammatory agent; analgesic)

Chapter 4

Discussion and future directions

4.1 Discussion

Drugs are used every day to battle diseases, and we need to understand bacterial drug metabolism as this can lead to undesirable outcomes. Bacterial drug metabolism is not limited to the human gut, but can occur anywhere in the human body where there are bacteria. As outlined in this thesis, understanding and predicting the genes involved in bacterial drug metabolism holds promise for efficient drug development and adoption of personalized medicine. In this thesis, I analyzed genomes and metagenomes from healthy individuals and identified more drug-metabolizing genes in the *Bacteroidaceae* family (Chapter 3). It is unclear how big of a problem this is in other data sets, for example, patients with a particular disease. I anticipate analyzing more data sets would help to identify trends in bacterial drug metabolism.

4.1.1 Importance of *in silico* methods

There is a wealth of microbial sequencing data provided by multiple projects, which includes the National Center for Biotechnology Information (NCBI; https://pubmed.ncbi.nlm.nih.gov) and the Human Microbiome Project [157]. We need to devise new *in silico* methods (i.e., computer-aided methods) to mine these data and identify new microbial functions. Multiple *in silico* tools are required to complement each other and achieve the same goal. For example, I needed to determine the gene context for the 3 fosfomycin homologs, I used PHASTER and PROKKA for genome annotation. This thesis used multiple *in silico* tools to identify putative bacterial drug-metabolizing genes from sequenced genomes (Chapter 2). It is essential that these *in silico* tools be built using a modular approach, as this will ensure an easy swapping of modules if a better method is conceived. I used the modular method to build MAGIS and AutoPhylo. There are a lot of homologous genes to known bacterial drug-metabolizing genes, for example, I identified 28 homologs to fosfomycin metabolizing genes in the 1,196 gut microbial genomes. As such, it's not feasible to test all the identified homologs, and *in silico* methods can help triage annotations and select representatives to characterize using biochemical methods (i.e., MIC experiment, etc.). Even testing only the representative or unique ones can be expensive and laborious. For example, in this thesis (Chapter 2), I selected the three homologs to fosfomycin metabolizing genes to test in the lab, which involves gene cloning (takes 1 day), transforming plasmid with the gene into an E. coli strain (1 day), checking correct clones (1 day), sequencing clones for confirmation (1 day), and analysis of the data (~2-3 days) for one gene. This timeline can be variable, and the cost of gene cloning can vary too, depending on gene size. The results for the three fosfomycin homologs

shows that two (FosD1 and FosD2) clearly do not metabolize fosfomycin despite their high similarity to fosfomycin inactivating enzymes. As such, less expensive methods to elucidate the function of these two enzymes are to use more *in silico* methods and guide biochemical tests.

Another method I tried during my research was to predict the possible function of a putative enzyme by using molecular docking. This method involved predicting a protein structure, docking multiple drugs to the predicted structure, and assessing possible activity. The current methods need to be more reliable to produce good results so this work was not included in my thesis. For example, for a given novel protein, it's not easy to predict the oligomer state suitable for the downstream analysis, i.e., docking.

4.1.2 Application of the HMDM database

Resources like the HMDM database are essential to support precision medicine. A good example of precision medicine is targeting bacterial strains which inactivate drugs using bacteriophages [158]. Duan and colleagues successfully removed a strain of *Enterococcus faecalis*, which produces a metabolite, cytolysins, that is toxic to human cells [159]. The HMDM resource can help identify such bacterial strains that can inactivate drugs leading to poor drug efficacy or toxicity [160]. Probiotics are used to promote health by providing specific microbiome metabolic outputs, e.g., modifying immune response [161]. These probiotic strains need to be able to colonize an unoccupied niche in the human gut [162]. The full functional capabilities of these probiotic strains require applications such as the HMDM database to assess potential drug-metabolizing genes that might affect

individual drug responsiveness. The machine learning method, GutBug [58], uses reference databases to make predictions, and the HMDM database can improve the quality of the predictions as it contains an experimentally validated data set. I tested GutBug to predict genes curated in the HMDM for drugs like Acarbose, and it failed to predict the *acbk* gene, which encodes the AcbK enzyme that metabolizes this drug. This test can be expanded to test all the drugs curated in the HMDM database to get the complete assessment for the GutBug software, albeit GutBug is not designed for high-throughput analysis.

4.2 Future Directions

4.2.1 The HMDM database sustainability

To ensure sustainability for the HMDM database, community engagement is required by publishing the resource and providing training materials on how to use its tools. Providing mailing lists for users to send questions will help with communications and keeping the HMDM database correct and up to date. Periodic updates are required, preferably monthly, that will ensure that there is not a lot of overwhelming information when updating, as generally, publications describing bacterial drug-metabolizing gene mechanisms can be sporadic. The HMDM database involves manual curation by human experts in the field of bacterial drug-metabolizing enzymes to ensure that the data is correct. As a result, human capital is required to keep the HMDM database up to date, and we need a long-term way to fund the project. There is interest in the HMDM database for evaluating clinical trial drugs. I was approached by a vendor at IIDR trainee day 2022 to use the DME software to predict drug-metabolizing genes

in routinely collected microbiome data. Given these potential collaborations, I recommend that the HMDM database be released on a subscription model, as this will ensure its sustainability. The DME software needs to be open source to encourage peer review and community evaluation, as there is little public software to predict bacterial drug metabolism from human microbiome data.

4.2.2 The HMDM database improvements

Currently, there is a basic search function added to the web application. I will be implementing a detailed search function, adding a download page for the data on the HMDM website, and will continue to add enzymes reported from the literature into the HMDM database. The website will be made public upon submission of the corresponding manuscript. I will be scheduling a run of prevalence data sets every month with the data releases for the database. After implementing batching for the inputs, the prevalence data set will be expanded to include scaffolds, contigs, and draft genomes from NCBI. I would also like to add any data sets Dr. Surette has to periodically annotate these genomes as more genes are added to the HMDM database. The documentation for the HMDM database will be updated, operating procedures will be documented, and HMDM will be added to the McArthur Lab biocuration operations under my supervision.

4.2.3 The HMDM database validation

I have added 45 β-glucuronidase [163] into the HMDM database as GUS enzymes metabolize glucuronidated drug conjugates [164]. Currently, these are added to the HMDM without connections to glucuronidated drugs. Elmassry and colleagues

used computational methods to link 100 drugs (see drug list in their Supplementary Data S1) to the β-glucuronidases, such as acetaminophen, but stated that the enzymes have not been experimentally validated [163], which makes it difficult to curate this information into the HMDM. The immunosuppressant prodrug mycophenolate mofetil (MMF) has the active form mycophenolic acid (MPA), which is conjugated in the liver to mycophenolate glucuronide (MPAG) [165]. MMF treats autoimmune diseases (such as Crohn's disease) and limits organ rejection during transplants. The GUS enzymes can be highly specific, MPAG was found to be only reactivated to MPA by flavin mononucleotide (FMN) binding GUS enzymes [166]. Simpson and colleagues used a combination of metagenomics data and biochemical methods to identify the FMN-binding GUS [166], but we need biochemical validation experiments to test all 45 β-glucuronidases before we can connect each to a glucuronidated drug in the HMDM. As a first step, I would like to group the 45 β-glucuronidase into the six groups proposed by Pollet *et al.* [32] and come up with representatives to test experimentally first. I would also group the 100 drugs identified by Elmassry and colleagues using drug classes in the HMDM ontology so that we can provide better interpretation. We can then compare the drug classes with structural similarity analysis performed by Elmassry and colleagues (see their Fig 4) [163].

The DME is a good tool for predicting drug-metabolizing genes from human gut microbiome data, as shown in Chapter 3, but to solidify these results, biochemical tests are required. I also used a small sample size to test for performance, and a large sample size would be ideal. There can also be miss-annotation, as annotation can slip into the Loose category due to the curated bit-score set being too high. I

did not explicitly examine for the possibility of false Loose annotations, but it will be valuable to assess these using both biochemical tests and *in silico* methods and adjust the bit-scores appropriately. These validation experiments will build trust in using the HMDM resources.

4.2.4 Improving *in silico* methods

With high-quality data curated in the HMDM database, we can expand our efforts to move beyond using alignments and use other methods to enrich the HMDM database. The SIMMER [167] tool was able to achieve high accuracy by using known chemical reactions, chemical structures, and protein similarity to predict drug-metabolizing enzymes in the human gut. To archive these results, Bustion and colleagues highlighted a full reaction description is required. SIMMER uses the MetaCyc reaction database [168] because there are validations for the curated enzymes. For a given predicted homolog, we can cluster with sequences predicted by SIMMER and predict the possible function of the homologs. Most drugs are metabolized in the human liver, and the drug conversions are known [18], we can use these data to build *in silico* methods, i.e., machine learning, perhaps, to classify homologs from the bacteria. Combining these methods with enzyme structures around the active sites, I anticipate, would yield drug metabolizing candidates.

References

1. Kaur G, Arora M, Kumar MNVR. Oral drug delivery technologies—a decade of developments. *Journal of Pharmacology and Experimental Therapeutics*. 2019;370: 529–543. doi:10.1124/jpet.118.255828
2. Mignani S, Kazzouli SE, Bousmina M, Majoral J-P. Expand classical drug administration ways by emerging routes using dendrimer drug delivery systems: A concise overview. *Advanced Drug Delivery Reviews*. 2013;65: 1316–1330. doi:10.1016/j.addr.2013.01.001
3. Al-Hilal TA, Alam F, Byun Y. Oral drug delivery systems using chemical conjugates or physical complexes. *Advanced Drug Delivery Reviews*. 2013;65: 845–864. doi:10.1016/j.addr.2012.11.002
4. Jourova L, Anzenbacher P, Anzenbacherova E. Human gut microbiota plays a role in the metabolism of drugs. *Biomedical Papers*. 2016;160: 317–326. doi:10.5507/bp.2016.039
5. Kessel SP van, Frye AK, El-Gendy AO, Castejon M, Keshavarzian A, *et al.* Gut bacterial tyrosine decarboxylases restrict levels of levodopa in the treatment of parkinson's disease. *Nature Communications*. 2019;10. doi:10.1038/s41467-019-08294-y
6. Almazroo OA, Miah MK, Venkataramanan R. Drug metabolism in the liver. *Clinics in Liver Disease*. 2017;21: 1–20. doi:10.1016/j.cld.2016.08.001

7. Döring B, Petzinger E. Phase 0 and phase III transport in various organs: Combined concept of phases in xenobiotic transport and metabolism. *Drug Metabolism Reviews*. 2014;46: 261–282. doi:10.3109/03602532.2014.882353
8. Crouwel F, Buiter HJC, Boer NK de. Gut microbiota-driven drug metabolism in inflammatory bowel disease. *Journal of Crohn's and Colitis*. 2020;15: 307–315. doi:10.1093/ecco-jcc/jjaa143
9. Sun C, Chen L, Shen Z. Mechanisms of gastrointestinal microflora on drug metabolism in clinical practice. *Saudi Pharmaceutical Journal*. 2019;27: 1146–1156. doi:10.1016/j.jsps.2019.09.011
10. Perez M, Calles-Enriquez M, Nes I, Martin MC, Fernandez M, *et al.* Tyramine biosynthesis is transcriptionally induced at low pH and improves the fitness of enterococcus faecalis in acidic environments. *Applied Microbiology and Biotechnology*. 2014;99: 3547–3558. doi:10.1007/s00253-014-6301-7
11. Haiser HJ, Gootenberg DB, Chatman K, Sirasani G, Balskus EP, Turnbaugh PJ. Predicting and manipulating cardiac drug inactivation by the human gut bacterium eggerthella lenta. *Science*. 2013;341: 295–298. doi:10.1126/science.1235872
12. Simmonds, Millar, Blake &R. Antioxidant effects of aminosalicylates and potential new drugs for inflammatory bowel disease: Assessment in cell-free systems and inflamed human colorectal biopsies. *Alimentary Pharmacology and Therapeutics*. 1999;13: 363–372. doi:10.1046/j.1365-2036.1999.00484.x
13. Sandborn WJ, Hanauer SB. The pharmacokinetic profiles of oral mesalazine formulations and mesalazine pro-drugs used in the management of ulcerative colitis. *Alimentary Pharmacology and Therapeutics*. 2002;17: 29–42. doi:10.1046/j.1365-2036.2003.01408.x

14. Grisham MB, Ware K, Marshall S, Yamada T, Sandhu IS. Prooxidant properties of 5-aminosalicylic acid. *Digestive Diseases and Sciences*. 1992;37: 1383–1389. doi:10.1007/bf01296008

15. Naeem M, Kim W, Cao J, Jung Y, Yoo J-W. Enzyme/pH dual sensitive polymeric nanoparticles for targeted drug delivery to the inflamed colon. *Colloids and Surfaces B: Biointerfaces*. 2014;123: 271–278. doi:10.1016/j.colsurfb.2014.09.026

16. Ryan A. Azoreductases in drug metabolism. *British Journal of Pharmacology*. 2016;174: 2161–2173. doi:10.1111/bph.13571

17. Li X, Li J, Gao Y, Kuang Y, Shi J, Xu B. Molecular nanofibers of olsalazine form supramolecular hydrogels for reductive release of an anti-inflammatory agent. *Journal of the American Chemical Society*. 2010;132: 17707–17709. doi:10.1021/ja109269v

18. Lynch T, Price A. The effect of cytochrome P450 metabolism on drug response, interactions, and adverse effects. *Am Fam Physician*. 2007;76: 391–396.

19. Lamb DC, Skaug T, Song H-L, Jackson CJ, Podust LM, *et al.* The cytochrome P450 complement (CYPome) of *Streptomyces coelicolor* A3(2). *Journal of Biological Chemistry*. 2002;277: 24000–24005. doi:10.1074/jbc.m111109200

20. Lamb DC, Ikeda H, Nelson DR, Ishikawa J, Skaug T, *et al.* Cytochrome P450 complement (CYPome) of the avermectin-producer *Streptomyces avermitilis* and comparison to that of *Streptomyces coelicolor* A3(2). *Biochemical and Biophysical Research Communications*. 2003;307: 610–619. doi:10.1016/s0006-291x(03)01231-2

21. Jackson CJ, Lamb DC, Kelly DE, Kelly SL. Bactericidal and inhibitory effects of azole antifungal compounds on *Mycobacterium smegmatis*. *FEMS Microbiology Letters*. 2000;192: 159–162. doi:10.1111/j.1574-6968.2000.tb09375.x

22. Ahmad Z, Sharma S, Khuller GK. Azole antifungals as novel chemotherapeutic agents against murine tuberculosis. *FEMS Microbiology Letters*. 2006;261: 181–186. doi:10.1111/j.1574-6968.2006.00350.x

23. Guardiola-Diaz HM, Foster L-A, Mushrush D, Vaz ADN. Azole-antifungal binding to a novel cytochrome P450 from *Mycobacterium tuberculosis*: Implications for treatment of tuberculosis. *Biochemical Pharmacology*. 2001;61: 1463–1470. doi:10.1016/s0006-2952(01)00571-8

24. Kelly SL, Kelly DE. Microbial cytochromes P450: Biodiversity and biotechnology. Where do cytochromes P450 come from, what do they do and what can they do for us? *Philosophical Transactions of the Royal Society B: Biological Sciences*. 2013;368: 20120476. doi:10.1098/rstb.2012.0476

25. Rekdal VM, Bess EN, Bisanz JE, Turnbaugh PJ, Balskus EP. Discovery and inhibition of an interspecies gut bacterial pathway for levodopa metabolism. *Science*. 2019;364. doi:10.1126/science.aau6323

26. Dobkin JF, Saha JR, Butler VP, Neu HC, Lindenbaum J. Digoxin-inactivating bacteria: Identification in human gut flora. *Science*. 1983;220: 325–327. doi:10.1126/science.6836275

27. Lindenbaum J, Rund DG, Butler VP, Tse-Eng D, Saha JR. Inactivation of digoxin by the gut flora: Reversal by antibiotic therapy. *New England Journal of Medicine*. 1981;305: 789–794. doi:10.1056/nejm198110013051403

28. Patocka J, Nepovimova E, Wu W, Kuca K. Digoxin: Pharmacology and toxicology—a review. *Environmental Toxicology and Pharmacology*. 2020;79: 103400. doi:10.1016/j.etap.2020.103400

29. Koppel N, Bisanz JE, Pandelia M-E, Turnbaugh PJ, Balskus EP. Discovery and characterization of a prevalent human gut bacterial enzyme sufficient for the inactivation of a family of plant toxins. *eLife*. 2018;7. doi:10.7554/elife.33953

30. Koppel N, Rekdal VM, Balskus EP. Chemical transformation of xenobiotics by the human gut microbiota. *Science*. 2017;356. doi:10.1126/science.aag2770

31. Wallace BD, Wang H, Lane KT, Scott JE, Orans J, *et al.* Alleviating cancer drug toxicity by inhibiting a bacterial enzyme. *Science*. 2010;330: 831–835. doi:10.1126/science.1191175

32. Pollet RM, DAgostino EH, Walton WG, Xu Y, Little MS, *et al.* An atlas of β-glucuronidases in the human intestinal microbiome. *Structure*. 2017;25: 967–977.e5. doi:10.1016/j.str.2017.05.003

33. Weinstock GM. Genomic approaches to studying the human microbiota. *Nature*. 2012;489: 250–256. doi:10.1038/nature11553

34. Wang W-L, Xu S-Y, Ren Z-G, Tao L, Jiang J-W, Zheng S-S. Application of metagenomics in the human gut microbiome. *World Journal of Gastroenterology*. 2015;21: 803. doi:10.3748/wjg.v21.i3.803

35. Structure, function and diversity of the healthy human microbiome. *Nature*. 2012;486: 207–214. doi:10.1038/nature11234

36. Suau A, Bonnet R, Sutren M, Godon J-J, Gibson GR, *et al.* Direct analysis of genes encoding 16S rRNA from complex communities reveals many novel molecular species within the human gut. *Applied and Environmental Microbiology.* 1999;65: 4799–4807. doi:10.1128/aem.65.11.4799-4807.1999

37. Tannock GW. Molecular assessment of intestinal microflora. *The American Journal of Clinical Nutrition.* 2001;73: 410s–414s. doi:10.1093/ajcn/73.2.410s

38. Behjati S, Tarpey PS. What is next generation sequencing? *Arch Dis Child Educ Pract Ed.* 2013;98: 236–238. doi:10.1136/archdischild-2013-304340

39. Thomas T, Gilbert J, Meyer F. Metagenomics - a guide from sampling to data analysis. *Microbial Informatics and Experimentation.* 2012;2. doi:10.1186/2042-5783-2-3

40. Lagier J-C, Armougom F, Million M, Hugon P, Pagnier I, *et al.* Microbial culturomics: Paradigm shift in the human gut microbiome study. *Clinical Microbiology and Infection.* 2012;18: 1185–1193. doi:10.1111/1469-0691.12023

41. Rettedal EA, Gumpert H, Sommer MOA. Cultivation-based multiplex phenotyping of human gut microbiota allows targeted recovery of previously uncultured bacteria. *Nature Communications.* 2014;5. doi:10.1038/ncomms5714

42. Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, *et al.* Diversity of the human intestinal microbial flora. *Science.* 2005;308: 1635–1638. doi:10.1126/science.1110591

43. Lau JT, Whelan FJ, Herath I, Lee CH, Collins SM, *et al.* Capturing the diversity of the human gut microbiota through culture-enriched molecular profiling. *Genome Medicine.* 2016;8. doi:10.1186/s13073-016-0327-7

44. Sibley CD, Grinwis ME, Field TR, Eshaghurshan CS, Faria MM, *et al.* Culture enriched molecular profiling of the cystic fibrosis airway microbiome. Planet PJ, editor. *PLoS ONE*. 2011;6: e22702. doi:10.1371/journal.pone.0022702

45. Goodman AL, Kallstrom G, Faith JJ, Reyes A, Moore A, *et al.* Extensive personal human gut microbiota culture collections characterized and manipulated in gnotobiotic mice. *Proceedings of the National Academy of Sciences*. 2011;108: 6252–6257. doi:10.1073/pnas.1102938108

46. Qin J, Ruiqiang Li and, Raes J, Arumugam M, Burgdorf KS, *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*. 2010;464: 59–65. doi:10.1038/nature08821

47. Sankar SA, Lagier J-C, Pontarotti P, Raoult D, Fournier P-E. The human gut microbiome, a taxonomic conundrum. *Systematic and Applied Microbiology*. 2015;38: 276–286. doi:10.1016/j.syapm.2015.03.004

48. Neumann J. FAIR data infrastructure. Smart biolabs of the future. Springer International Publishing; 2022. pp. 195–207. doi:10.1007/10_2021_193

49. Dooley DM, Griffiths EJ, Gosal GS, Buttigieg PL, Hoehndorf R, *et al.* FoodOn: A harmonized food ontology to increase global food traceability, quality control and data integration. *npj Science of Food*. 2018;2. doi:10.1038/s41538-018-0032-6

50. Alcock BP, Huynh W, Chalil R, Smith KW, Raphenya AR, *et al.* CARD 2023: Expanded curation, support for machine learning, and resistome prediction at the comprehensive antibiotic resistance database. *Nucleic Acids Research*. 2022;51: D690–D699. doi:10.1093/nar/gkac920

51. Zeng X, Yang X, Fan J, Tan Y, Ju L, *et al.* MASI: Microbiota—active substance interactions database. *Nucleic Acids Research.* 2020;49: D776–D782. doi:10.1093/nar/gkaa924

52. Aziz RK, Saad R, Rizkallah MR. PharmacoMicrobiomics or how bugs modulate drugs: An educational initiative to explore the effects of human microbiome on drugs. *BMC Bioinformatics.* 2011;12. doi:10.1186/1471-2105-12-s7-a10

53. Janssens Y, Nielandt J, Bronselaer A, Debunne N, Verbeke F, *et al.* Disbiome database: Linking the microbiome to disease. *BMC Microbiology.* 2018;18. doi:10.1186/s12866-018-1197-5

54. Cheng L, Qi C, Zhuang H, Fu T, Zhang X. gutMDisorder: A comprehensive database for dysbiosis of the gut microbiota in disorders and interventions. *Nucleic Acids Research.* 2019;48: D554–D560. doi:10.1093/nar/gkz843

55. Zhou J, Ouyang J, Gao Z, Qin H, Jun W, Shi T. MagMD: Database summarizing the metabolic action of gut microbiota to drugs. *Computational and Structural Biotechnology Journal.* 2022;20: 6427–6430. doi:10.1016/j.csbj.2022.11.021

56. Guthrie L, Wolfson S, Kelly L. The human gut chemical landscape predicts microbe-mediated biotransformation of foods and drugs. *eLife.* 2019;8. doi:10.7554/elife.42866

57. Bustion A, Agrawal A, Turnbaugh PJ, Pollard KS. A novel *in silico* method employs chemical and protein similarity algorithms to accurately identify chemical transformations in the human gut microbiome. 2022. doi:10.1101/2022.08.02.502504

58. Malwe AS, Srivastava GN, Sharma VK. GutBug: A tool for prediction of human gut bacteria mediated biotransformation of biotic and xenobiotic molecules using machine learning. *Journal of Molecular Biology*. 2023; 168056. doi:10.1016/j.jmb.2023.168056

59. Mallory EK, Acharya A, Rensi SE, Turnbaugh PJ, Bright RA, Altman RB. Chemical reaction vector embeddings: Towards predicting drug metabolism in the human gut microbiome. *Pac Symp Biocomput*. 2018;23: 56–67.

60. Lam KN, Alexander M, Turnbaugh PJ. Precision medicine goes microscopic: Engineering the microbiome to improve drug outcomes. *Cell Host and Microbe*. 2019;26: 22–34. doi:10.1016/j.chom.2019.06.011

61. Modis Y, Wierenga R. Two crystal structures of n-acetyltransferases reveal a new fold for CoA-dependent enzymes. *Structure*. 1998;6: 1345–1350. doi:10.1016/s0969-2126(98)00134-8

62. Chae S, Kim DJ, Cho J-Y. Complex influences of gut microbiome metabolism on various drug responses. *Translational and Clinical Pharmacology*. 2020;28: 7. doi:10.12793/tcp.2020.28.e3

63. McLachlan A, Hilmer S, Couteur DL. Variability in response to medicines in older people: Phenotypic and genotypic factors. *Clinical Pharmacology and Therapeutics*. 2009;85: 431–433. doi:10.1038/clpt.2009.1

64. Tulner LR, Frankfort SV, Gijsen GJPT, Campen JPCM van, Koks CHW, Beijnen JH. Drug-drug interactions in a geriatric outpatient cohort. *Drugs and Aging*. 2008;25: 343–355. doi:10.2165/00002512-200825040-00007

65. Kirchmair J, Göller AH, Lang D, Kunze J, Testa B, *et al.* Predicting drug metabolism: Experiment and/or computation? *Nature Reviews Drug Discovery*. 2015;14: 387–404. doi:10.1038/nrd4581

66. Mira A, Ochman H. Gene location and bacterial sequence divergence. *Molecular Biology and Evolution*. 2002;19: 1350–1358. doi:10.1093/oxfordjournals.molbev.a004196

67. Jacob F. Evolution and tinkering. *Science*. 1977;196: 1161–1166. doi:10.1126/science.860134

68. Copley SD. The physical basis and practical consequences of biological promiscuity. *Physical Biology*. 2020;17: 051001. doi:10.1088/1478-3975/ab8697

69. Baier F, Chen J, Solomonson M, Strynadka NCJ, Tokuriki N. Distinct metal isoforms underlie promiscuous activity profiles of metalloenzymes. *ACS Chemical Biology*. 2015;10: 1684–1693. doi:10.1021/acschembio.5b00068

70. Vipond IB, Moon B-J, Halford SE. An isoleucine to leucine mutation that switches the cofactor requirement of the EcoRV restriction endonuclease from magnesium to manganese. *Biochemistry*. 1996;35: 1712–1721. doi:10.1021/bi9523926

71. SALYERS A, GUPTA A, WANG Y. Human intestinal bacteria as reservoirs for antibiotic resistance genes. *Trends in Microbiology*. 2004;12: 412–416. doi:10.1016/j.tim.2004.07.004

72. Lu N, Hu Y, Zhu L, Yang X, Yin Y, *et al.* DNA microarray analysis reveals that antibiotic resistance-gene diversity in human gut microbiota is age related. *Scientific Reports*. 2014;4. doi:10.1038/srep04302

73. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, *et al.* The pfam protein families database in 2019. *Nucleic Acids Research*. 2018;47: D427–D432. doi:10.1093/nar/gky995

74. Whelan S, Morrison DA. Inferring trees. *Methods in molecular biology*. Springer New York; 2016. pp. 349–377. doi:10.1007/978-1-4939-6622-6_14

75. Gascuel O. Neighbor-joining revealed. *Molecular Biology and Evolution*. 2006;23: 1997–2000. doi:10.1093/molbev/msl072

76. Yang S, Angelis DD. Maximum likelihood. *Methods in molecular biology*. Humana Press; 2012. pp. 581–595. doi:10.1007/978-1-62703-059-5_24

77. Gaudet P, Livstone MS, Lewis SE, Thomas PD. Phylogenetic-based propagation of functional annotations within the gene ontology consortium. *Briefings in Bioinformatics*. 2011;12: 449–462. doi:10.1093/bib/bbr042

78. Rhoads A, Au KF. PacBio sequencing and its applications. *Genomics Proteomics Bioinformatics*. 2015;13: 278–289. doi:10.1016/j.gpb.2015.08.002

79. Eid J, Fehr A, Gray J, Luong K, Lyle J, *et al.* Real-time DNA sequencing from single polymerase molecules. *Science*. 2009;323: 133–138. doi:10.1126/science.1162986

80. Derakhshani H, Bernier SP, Marko VA, Surette MG. Completion of draft bacterial genomes by long-read sequencing of synthetic genomic pools. *BMC Genomics*. 2020;21. doi:10.1186/s12864-020-06910-6

81. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. Phillippy AM, editor. *PLOS Computational Biology*. 2017;13: e1005595. doi:10.1371/journal.pcbi.1005595

82. Seemann T. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics*. 2014;30: 2068–2069. doi:10.1093/bioinformatics/btu153

83. Schwengers O, Jelonek L, Dieckmann MA, Beyvers S, Blom J, Goesmann A. Bakta: Rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. *Microbial Genomics*. 2021;7. doi:10.1099/mgen.0.000685

84. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*. 2015;25: 1043–1055. doi:10.1101/gr.186072.114

85. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. GTDB-tk: A toolkit to classify genomes with the genome taxonomy database. Hancock J, editor. *Bioinformatics*. 2019. doi:10.1093/bioinformatics/btz848

86. Alcock BP, Raphenya AR, Lau TTY, Tsang KK, Bouchard M, *et al.* CARD 2020: Antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Research*. 2019. doi:10.1093/nar/gkz935

87. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*. 2010;11. doi:10.1186/1471-2105-11-119

88. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Research*. 2020;49: D412–D419. doi:10.1093/nar/gkaa913

89. Blum M, Chang H-Y, Chuguransky S, Grego T, Kandasaamy S, *et al.* The InterPro protein families and domains database: 20 years on. *Nucleic Acids Research*. 2020;49: D344–D354. doi:10.1093/nar/gkaa977

90. Gibson MK, Forsberg KJ, Dantas G. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *The ISME Journal*. 2014;9: 207–216. doi:10.1038/ismej.2014.106

91. Krogh A, Brown M, Mian IS, Sjölander K, Haussler D. Hidden markov models in computational biology. *Journal of Molecular Biology*. 1994;235: 1501–1531. doi:10.1006/jmbi.1994.1104

92. Bateman A, Coggill P, Finn RD. DUFs: Families in search of function. *Acta Crystallographica Section F Structural Biology and Crystallization Communications*. 2010;66: 1148–1152. doi:10.1107/s1744309110001685

93. Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, *et al.* Database resources of the national center for biotechnology information. *Nucleic Acids Research*. 2021;50: D20–D26. doi:10.1093/nar/gkab1112

94. Edgar RC. High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. 2021. doi:10.1101/2021.06.20.449169

95. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*. 2009;25: 1972–1973. doi:10.1093/bioinformatics/btp348

96. Price MN, Dehal PS, Arkin AP. FastTree 2 – approximately maximum-likelihood trees for large alignments. Poon AFY, editor. *PLoS ONE*. 2010;5: e9490. doi:10.1371/journal.pone.0009490

97. Stamatakis A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. 2014;30: 1312–1313. doi:10.1093/bioinformatics/btu033

98. Carver T, Thomson N, Bleasby A, Berriman M, Parkhill J. DNAPlotter: Circular and linear interactive genome visualization. *Bioinformatics*. 2008;25: 119–120. doi:10.1093/bioinformatics/btn578

99. Cox G, Sieron A, King AM, Pascale GD, Pawlowski AC, *et al.* A common platform for antibiotic dereplication and adjuvant discovery. *Cell Chemical Biology*. 2017;24: 98–109. doi:10.1016/j.chembiol.2016.11.011

100. Kowalska-Krochmal B, Dudek-Wicher R. The minimum inhibitory concentration of antibiotics: Methods, interpretation, clinical relevance. *Pathogens*. 2021;10: 165. doi:10.3390/pathogens10020165

101. Alhakami H, Mirebrahim H, Lonardi S. A comparative evaluation of genome assembly reconciliation tools. *Genome Biology*. 2017;18. doi:10.1186/s13059-017-1213-3

102. Spanogiannopoulos P, Thaker M, Koteva K, Waglechner N, Wright GD. Characterization of a rifampin-inactivating glycosyltransferase from a screen of environmental actinomycetes. *Antimicrobial Agents and Chemotherapy*. 2012;56: 5061–5069. doi:10.1128/aac.01166-12

103. Yazawa K, Mikami Y, Maeda A, Akao M, Morisaki N, Iwasaki S. Inactivation of rifampin by *nocardia brasiliensis*. *Antimicrobial Agents and Chemotherapy*. 1993;37: 1313–1317. doi:10.1128/aac.37.6.1313

104. Chan CX, Mahbob M, Ragan MA. Clustering evolving proteins into homologous families. *BMC Bioinformatics*. 2013;14. doi:10.1186/1471-2105-14-120

105. Bailey TL, Johnson J, Grant CE, Noble WS. The MEME suite. *Nucleic Acids Research*. 2015;43: W39–W49. doi:10.1093/nar/gkv416

106. Pankov KV, McArthur AG, Gold DA, Nelson DR, Goldstone JV, Wilson JY. The cytochrome P450 (CYP) superfamily in cnidarians. *Scientific Reports*. 2021;11. doi:10.1038/s41598-021-88700-y

107. Stogios PJ, Cox G, Spanogiannopoulos P, Pillon MC, Waglechner N, *et al.* Rifampin phosphotransferase is an unusual antibiotic resistance kinase. *Nature Communications*. 2016;7. doi:10.1038/ncomms11343

108. Spanogiannopoulos P, Waglechner N, Koteva K, Wright GD. A rifamycin inactivating phosphotransferase family shared by environmental and pathogenic bacteria. *Proceedings of the National Academy of Sciences*. 2014;111: 7102–7107. doi:10.1073/pnas.1402358111

109. Kahan FM, Kahan JS, Cassidy PJ, Kropp H. The mechanism of action of fosfomycin (phosphonomycin). *Annals of the New York Academy of Sciences*. 1974;235: 364–386. doi:10.1111/j.1749-6632.1974.tb43277.x

110. Silver LL. Fosfomycin: Mechanism and resistance. *Cold Spring Harbor Perspectives in Medicine*. 2017;7: a025262. doi:10.1101/cshperspect.a025262

111. Wu D, Chen Y, Sun L, Qu T, Wang H, Yu Y. Prevalence of fosfomycin resistance in methicillin-resistant *Staphylococcus aureus* isolated from patients in a university hospital in china from 2013 to 2015. *Japanese Journal of Infectious Diseases*. 2018;71: 312–314. doi:10.7883/yoken.jjid.2018.013

112. Manara S, Pasolli E, Dolce D, Ravenni N, Campana S, *et al.* Whole-genome epidemiology, characterisation, and phylogenetic reconstruction of *Staphylococcus aureus* strains in a paediatric hospital. *Genome Medicine*. 2018;10. doi:10.1186/s13073-018-0593-7

113. Yang X, Zhang J, Yu S, Wu Q, Guo W, *et al.* Prevalence of *Staphylococcus aureus* and methicillin-resistant *Staphylococcus aureus* in retail ready-to-eat foods in china. *Frontiers in Microbiology*. 2016;7. doi:10.3389/fmicb.2016.00816

114. Forster BM, Marquis H. Protein transport across the cell wall of monoderm gram-positive bacteria. *Molecular Microbiology*. 2012;84: 405–413. doi:10.1111/j.1365-2958.2012.08040.x

115. Roberts AA, Sharma SV, Strankman AW, Duran SR, Rawat M, Hamilton CJ. Mechanistic studies of FosB: A divalent-metal-dependent bacillithiol-s-transferase that mediates fosfomycin resistance in *Staphylococcus aureus*. *Biochemical Journal*. 2013;451: 69–79. doi:10.1042/bj20121541

116. Thompson MK, Keithly ME, Goodman MC, Hammer ND, Cook PD, *et al.* Structure and function of the genomically encoded fosfomycin resistance enzyme, FosB, from *Staphylococcus aureus*. *Biochemistry*. 2014;53: 755–765. doi:10.1021/bi4015852

117. Thompson MK, Keithly ME, Harp J, Cook PD, Jagessar KL, *et al.* Structural and chemical aspects of resistance to the antibiotic fosfomycin conferred by FosB from *Bacillus cereus*. *Biochemistry*. 2013;52: 7350–7362. doi:10.1021/bi4009648

118. Castañeda-García A, Blázquez J, Rodríguez-Rojas A. Molecular mechanisms and clinical impact of acquired and intrinsic fosfomycin resistance. *Antibiotics*. 2013;2: 217–236. doi:10.3390/antibiotics2020217

119. Kuzuyama T, Kobayashi S, OHara K, Hidaka T, Seto H. Fosfomycin monophosphate and fosfomycin diphosphate, two inactivated fosfomycin derivatives formed by gene products of fomA and fomB from a fosfomycin producing organism *Streptomyces wedmorensis*. *The Journal of Antibiotics*. 1996;49: 502–504. doi:10.7164/antibiotics.49.502

120. García P, Arca P, Suárez JE. Product of fosC, a gene from pseudomonas syringae, mediates fosfomycin resistance by using ATP as cosubstrate. *Antimicrobial Agents and Chemotherapy*. 1995;39: 1569–1573. doi:10.1128/aac.39.7.1569

121. Thompson MK, Keithly ME, Sulikowski GA, Armstrong RN. Diversity in fosfomycin resistance proteins. *Perspectives in Science*. 2015;4: 17–23. doi:10.1016/j.pisc.2014.12.004

122. Travis S, Shay MR, Manabe S, Gilbert NC, Frantom PA, Thompson MK. Characterization of the genomically encoded fosfomycin resistance enzyme from *Mycobacterium abscessus*. *MedChemComm*. 2019;10: 1948–1957. doi:10.1039/c9md00372j

123. Aristoff PA, Garcia GA, Kirchhoff PD, Showalter HDH. Rifamycins – obstacles and opportunities. *Tuberculosis*. 2010;90: 94–118. doi:10.1016/j.tube.2010.02.001

124. Adams RA, Leon G, Miller NM, Reyes SP, Thantrong CH, *et al.* Rifamycin antibiotics and the mechanisms of their failure. *The Journal of Antibiotics*. 2021;74: 786–798. doi:10.1038/s41429-021-00462-x

125. Pawlowski AC, Wang W, Koteva K, Barton HA, McArthur AG, Wright GD. A diverse intrinsic antibiotic resistome from a cave bacterium. *Nature Communications*. 2016;7. doi:10.1038/ncomms13803

126. Gerlt JA, Bouvier JT, Davidson DB, Imker HJ, Sadkhin B, *et al.* Enzyme function initiative-enzyme similarity tool (EFI-EST): A web tool for generating protein sequence similarity networks. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*. 2015;1854: 1019–1037. doi:10.1016/j.bbapap.2015.04.015

127. Zimmermann M, Zimmermann-Kogadeeva M, Wegmann R, Goodman AL. Mapping human microbiome drug metabolism by gut bacteria and their genes. *Nature*. 2019;570: 462–467. doi:10.1038/s41586-019-1291-3

128. Maier L, Pruteanu M, Kuhn M, Zeller G, Telzerow A, *et al.* Extensive impact of non-antibiotic drugs on human gut bacteria. *Nature*. 2018;555: 623–628. doi:10.1038/nature25979

129. Little MS, Pellock SJ, Walton WG, Tripathy A, Redinbo MR. Structural basis for the regulation of β-glucuronidase expression by human gut *Enterobacteriaceae*. *Proceedings of the National Academy of Sciences*. 2017;115. doi:10.1073/pnas.1716241115

130. Ervin SM, Simpson JB, Gibbs ME, Creekmore BC, Lim L, *et al.* Structural insights into endobiotic reactivation by human gut microbiome-encoded sulfatases. *Biochemistry*. 2020;59: 3939–3950. doi:10.1021/acs.biochem.0c00711

131. Lazarević S, anic M, Al-Salami H, Mooranian A, Mikov M. Gut microbiota metabolism of azathioprine: A new hallmark for personalized drug-targeted therapy of chronic inflammatory bowel disease. *Frontiers in Pharmacology*. 2022;13. doi:10.3389/fphar.2022.879170

132. Mungall CJ, and DBE. A chado case study: An ontology-based modular schema for representing genome-associated biological information. *Bioinformatics*. 2007;23: i337–i346. doi:10.1093/bioinformatics/btm189

133. Edalatmand A, McArthur AG. CARDshark: Automated prioritization of literature curation for the comprehensive antibiotic resistance database. *Database*. 2023;2023. doi:10.1093/database/baad023

134. Alex Bateman and, Martin M-J, Orchard S, Magrane M, Ahmad S, *et al.* UniProt: The universal protein knowledgebase in 2023. *Nucleic Acids Research*. 2022;51: D523–D531. doi:10.1093/nar/gkac1052

135. Chen S, Zhou Y, Chen Y, Gu J. Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*. 2018;34: i884–i890. doi:10.1093/bioinformatics/bty560

136. Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. *Nature Methods*. 2012;9: 357–359. doi:10.1038/nmeth.1923

137. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: A new versatile metagenomic assembler. *Genome Research*. 2017;27: 824–834. doi:10.1101/gr.213959.116

138. Kang DD, Li F, Kirton E, Thomas A, Egan R, *et al.* MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*. 2019;7: e7359. doi:10.7717/peerj.7359

139. Hastings J, Owen G, Dekker A, Ennis M, Kale N, *et al.* ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Research*. 2015;44: D1214–D1219. doi:10.1093/nar/gkv1031

140. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, *et al.* BLAST+: Architecture and applications. *BMC Bioinformatics*. 2009;10. doi:10.1186/1471-2105-10-421

141. Buchfink B, Reuter K, Drost H-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods*. 2021;18: 366–368. doi:10.1038/s41592-021-01101-x

142. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nature Methods*. 2014;12: 59–60. doi:10.1038/nmeth.3176

143. Gabrielaite M, Marvig RL. GenAPI: A tool for gene absence-presence identification in fragmented bacterial genome sequences. *BMC Bioinformatics*. 2020;21. doi:10.1186/s12859-020-03657-5

144. Cobo F, Perez-Carrasco V, Rodriguez-Guerrero E, Sampedro A, Rodriguez-Granger J, *et al.* Misidentification of *Phocaeicola* (*Bacteroides*) *dorei* in two patients with bacteremia. *Anaerobe*. 2022;75: 102544. doi:10.1016/j.anaerobe.2022.102544

145. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with kraken 2. *Genome Biology*. 2019;20. doi:10.1186/s13059-019-1891-0

146. Tsang KK, Maguire F, Zubyk HL, Chou S, Edalatmand A, *et al.* Identifying novel β-lactamase substrate activity through in silico prediction of antimicrobial resistance. *Microbial Genomics.* 2021;7. doi:10.1099/mgen.0.000500

147. Kurokawa K, Itoh T, Kuwahara T, Oshima K, Toh H, *et al.* Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Research.* 2007;14: 169–181. doi:10.1093/dnares/dsm018

148. Ellrott K, Jaroszewski L, Li W, Wooley JC, Godzik A. Expansion of the protein repertoire in newly explored environments: Human gut microbiome specific protein families. Jones DT, editor. *PLoS Computational Biology.* 2010;6: e1000798. doi:10.1371/journal.pcbi.1000798

149. Awolade P, Cele N, Kerru N, Gummidi L, Oluwakemi E, Singh P. Therapeutic significance of β-glucuronidase activity and its inhibitors: A review. *European Journal of Medicinal Chemistry.* 2020;187: 111921. doi:10.1016/j.ejmech.2019.111921

150. Gloux K, Berteau O, oumami HE, Béguet F, Leclerc M, Doré J. A metagenomic β-glucuronidase uncovers a core adaptive function of the human intestinal microbiome. *Proceedings of the National Academy of Sciences.* 2010;108: 4539–4546. doi:10.1073/pnas.1000066107

151. The intestinal microbiome and estrogen receptor–positive female breast cancer. *JNCI: Journal of the National Cancer Institute.* 2016. doi:10.1093/jnci/djw029

152. Murphy EC, Mörgelin M, Cooney JC, Frick I-M. Interaction of *Bacteroides fragilis* and *Bacteroides thetaiotaomicron* with the kallikrein–kinin system. *Microbiology.* 2011;157: 2094–2105. doi:10.1099/mic.0.046862-0

153. Zafar H, Saier MH. Gut *Bacteroides* species in health and disease. *Gut Microbes*. 2021;13. doi:10.1080/19490976.2020.1848158

154. Ferrocino I, Cagno RD, Angelis MD, Turroni S, Vannini L, *et al.* Fecal microbiota in healthy subjects following omnivore, vegetarian and vegan diets: Culturable populations and rRNA DGGE profiling. Heimesaat MM, editor. *PLOS ONE*. 2015;10: e0128669. doi:10.1371/journal.pone.0128669

155. Voreades N, Kozil A, Weir TL. Diet and the development of the human intestinal microbiome. *Frontiers in Microbiology*. 2014;5. doi:10.3389/fmicb.2014.00494

156. Bird S, Klein E, Loper E. Natural language processing with python: Analyzing text with the natural language toolkit. O'Reilly Media, Inc.; 2009.

157. Gevers D, Knight R, Petrosino JF, Huang K, McGuire AL, *et al.* The human microbiome project: A community resource for the healthy human microbiome. *PLoS Biology*. 2012;10: e1001377. doi:10.1371/journal.pbio.1001377

158. Aziz RK, Rizkallah MR, Saad R, ElRakaiby MT. Translating pharmacomicrobiomics: Three actionable challenges/prospects in 2020. *OMICS: A Journal of Integrative Biology*. 2020;24: 60–61. doi:10.1089/omi.2019.0205

159. Duan Y, Llorente C, Lang S, Brandl K, Chu H, *et al.* Bacteriophage targeting of gut bacterium attenuates alcoholic liver disease. *Nature*. 2019;575: 505–511. doi:10.1038/s41586-019-1742-x

160. Yan A, Culp E, Perry J, Lau JT, MacNeil LT, *et al.* Transformation of the anticancer drug doxorubicin in the human gut microbiome. *ACS Infectious Diseases*. 2017;4: 68–76. doi:10.1021/acsinfecdis.7b00166

161. Kumar R, Sood U, Kaur J, Anand S, Gupta V, *et al.* The rising dominance of microbiology: What to expect in the next 15 years? *Microbial Biotechnology.* 2021;15: 110–128. doi:10.1111/1751-7915.13953

162. Cunningham M, Azcarate-Peril MA, Barnard A, Benoit V, Grimaldi R, *et al.* Shaping the future of probiotics and prebiotics. *Trends in Microbiology.* 2021;29: 667–685. doi:10.1016/j.tim.2021.01.003

163. Elmassry MM, Kim S, Busby B. Predicting drug-metagenome interactions: Variation in the microbial β-glucuronidase level in the human gut metagenomes. Loor JJ, editor. *PLOS ONE.* 2021;16: e0244876. doi:10.1371/journal.pone.0244876

164. Biernat KA, Pellock SJ, Bhatt AP, Bivins MM, Walton WG, *et al.* Structure, function, and inhibition of drug reactivating human gut microbial β-glucuronidases. *Scientific Reports.* 2019;9. doi:10.1038/s41598-018-36069-w

165. Park H. The emergence of mycophenolate mofetilin dermatology: From its roots in the world of organ transplantation to its versatile role in the dermatology treatment room. *J Clin Aesthet Dermatol.* 2011;4: 18–27.

166. Simpson JB, Sekela JJ, Graboski AL, Borlandelli VB, Bivins MM, *et al.* Metagenomics combined with activity-based proteomics point to gut bacterial enzymes that reactivate mycophenolate. *Gut Microbes.* 2022;14. doi:10.1080/19490976.2022.2107289

167. Bustion AE, Nayak RR, Agrawal A, Turnbaugh PJ, Pollard KS. SIMMER employs similarity algorithms to accurately identify human gut microbiome species and enzymes capable of known chemical transformations. *eLife.* 2023;12. doi:10.7554/elife.82401

168. Caspi R, Billington R, Keseler IM, Kothari A, Krummenacker M, *et al.* The MetaCyc database of metabolic pathways and enzymes - a 2019 update. *Nucleic Acids Research.* 2019;48: D445–D453. doi:10.1093/nar/gkz862