Implementation and Application of Genomic Association Methods to *Clostridium difficile* Toxicity and Clinical Infection Outcomes

by

Katie M. Saund

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Microbiology and Immunology)
in The University of Michigan
2020

Doctoral Committee:

Assistant Professor Evan Snitkin, Chair
Professor Philip Hanna
Associate Professor Stephen Smith
Associate Professor Cristen Willer
Professor Vincent Young

Katie M. Saund

katiephd@umich.edu

ORCID: 0000-0002-6214-6713

DEDICATION

To my Bupa

Dugald Cameron

1932 - 2015

ACKNOWLEDGEMENTS

To paraphrase Snitkin Lab alumna Alex Wells, it takes a lab to raise a scientist. I am grateful to the people who have kindly trained and supported me. Thank you all:

The scientists on who taught me how to tame *C. difficile* and bend it to my will: Kim Vendrov, Alex Standke, Aline Penkevich, and Dr. Anna Seekatz.

Dr. Marc Sze who gave me a glimpse at life in the pharmaceutical industry.

Two mentors prior to graduate school who demonstrated through their work the importance of doing science to improve lives: Dr. Indi Trehan and Dr. Courtney Crane.

Several postdocs whose influence was critical both to my scientific growth and navigating the quagmire of graduate school: Dr. Kristen Haberthur, Dr. Josie Libertucci, and Dr. Arianna Miles-Jay.

Evan and the rest of the Snitkin Lab. You Snitfits are a good bunch.

My family: my parents, my in-laws, and particularly my husband for their support. Brad, this thesis brings us one step closer to finally establishing The Saund Institute for Biological Robots.

Contents

List of Figures

List of Tables

ABSTRACT

Clostridium difficile is a major cause of healthcare-associated infections in the United States. A *C. difficile* infection can lead to a range of outcomes including diarrhea, intensive care unit admission, abdominal surgery, or death. Pathogenesis is mediated by the release of toxin from *C. difficile* cells growing in the intestines. Some patients are more vulnerable to infection, including those with previous antibiotic exposure and advanced age. Host factors can affect the likelihood of infection but also the severity of infection. Additionally, infection severity can be influenced by the genome of the infecting strain(s). Host-pathogen interactions are extremely complex and very little is known about the interplay between host factors and *C. difficile* genomic variation with respect to infection likelihood and outcomes. With the recent deluge of whole genome sequencing data, the contribution of bacterial genomic variation to infections can be more comprehensively evaluated than ever before. The work described in this dissertation used two different approaches to test for associations between *C. difficile* genomic variation and clinically relevant phenotypes.

In the first approach we implemented and applied a novel convergence-based bacterial genome-wide association study (bGWAS) algorithm for quantitative traits. We introduce the algorithm using a set of data generated *in silico* to realistically model bacterial genome variation and phenotypes under various evolutionary regimes. When the algorithm was applied to *C. difficile* genomic variants and toxin activity our bGWAS identified known toxin regulatory genes associated with toxin activity, supporting the value of our approach. Besides identifying key cis-regulatory variants in the toxin-producing locus, we observed

several associations that connect toxin activity to a complex network of trans-regulatory genes. Many highly associated variants occur in flagellar genes and indicate coregulation of toxicity and motility. We propose new variants associated with toxin activity for future functional validation. This study focused on a complex phenotype, toxin activity, within a highly controlled *in vitro* system.

We next investigated the impact of bacterial genetic variation on human infections. The increased complexity of this human-pathogen interaction justified a different association approach to better understand the independent contribution of bacterial genomic variation to infection. In a set of clinically derived isolates, we tested for the association between variants in trehalose metabolism operons and infection severity while incorporating and controlling for infection severity-modulating patient characteristics. Trehalose utilization variants were recently proposed to modulate *C. difficile* infections in a mouse model. Interestingly, we observed that this *in vivo* result did not translate to our clinical cohort as we found no evidence of an association between any of the trehalose utilization variants and patient infection outcomes. Taken together, these results demonstrate the utility of applying multiple approaches for identifying genomic variants associated with clinical outcomes that account for either bacterial population structure or host factors.

Chapter 1

Introduction

1.1 Motivation

Many factors contribute to the clinical outcomes of bacterial infections including host, bacterial, and environmental characteristics. This thesis focuses on the contribution of *Clostridium difficile* genomic variants to clinically relevant *in vitro* phenotypes and clinical infection outcomes. *C. difficile* genomic variants that affect infection outcomes may someday be included in diagnostic testing so that patients can be stratified for specialized treatments to improve outcomes. Additionally, these same genomic variants may also provide additional targets for future *C. difficile* treatments and provide insights into the nature of human-pathogen interactions. Discovering these clinically-important variants is possible through multiple approaches, including both phylogenetically aware genome-wide association studies (GWAS) and association methods that control for host factors.

1.2 *C. difficile* infection, pathogenesis, and metabolism

C. difficile is a healthcare-associated bacterial pathogen. *C. difficile* was first recognized to cause antibiotic-associated diarrhea in 1977[1]. Today, *C. difficile* is the most common healthcare-associated infection in the US and costs associated with *C. difficile* hospital-

onset infections reach $1 billion annually[2]. *C. difficile* prevalence is slowly decreasing in hospitals but is not declining in the community[2]. In 2017, *C. difficile* caused 223,900 infections in hospitalized patients and 12,800 deaths[2]. The *C. difficile* species is diverse and contains strains with and without the ability to produce toxin. Only toxin-producing *C. difficile* strains cause clinically important manifestations. Infections with toxin-producing *C. difficile* can be asymptomatic or symptomatic (Figure 1.1)[3]. Clinical manifestations of *C. difficile* infection include diarrhea, fever, abdominal pain, and colitis[3]. In more severe cases *C. difficile* infection may lead to shock, toxic megacolon, intestinal perforation, peritonitis, and death[3]. After infection resolution, some patients may suffer from additional *C. difficile* infections that are either novel infections or a recurrence of the same infecting strain[3].

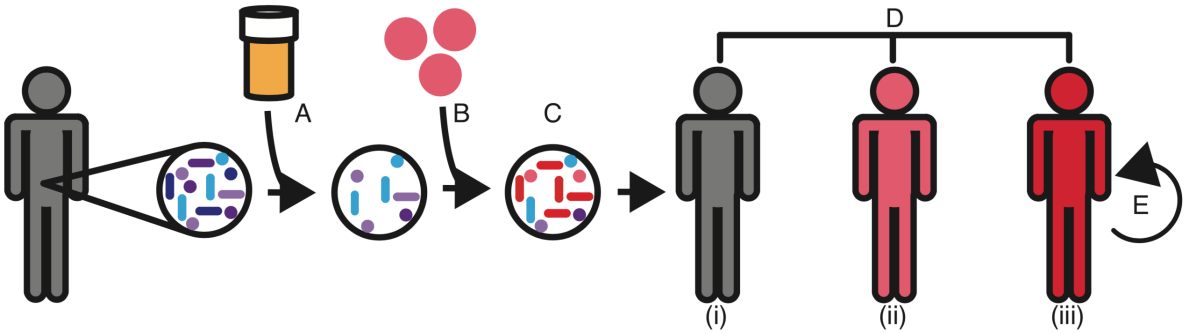

Figure 1.1: Model of a typical *C. difficile* infection

A) Antibiotic treatment disrupts gut microbiota. B) Patient ingests *C. difficile* spores. C) Spores germinate into vegetative cells in the intestines and secrete toxin. D) Patient may have (i) an asymptomatic infection, (ii) a mild infection, or (iii) a severe infection. E) Infection may recur after initial resolution.

C. difficile is a Gram-positive, obligate anaerobic bacterium. It survives aerobic conditions in its spore form (Figure 1.2). Spores are released in the stool of colonized or infected people. Patients swallow *C. difficile* by touching spores on contaminated surfaces or contaminated people. Dormant spores are resistant to the alcohol in hand disinfectants which

contributes to the persistence of *C. difficile* in healthcare settings and even hand washing can fail to remove spores[3]. Transmission to patients in healthcare settings can be reduced by the use of personal protective equipment by medical staff (gowns and gloves), isolating patients in individual rooms, and judicious use of antimicrobials that are associated with increased risk of *C. difficile* infections. Patients with increased risk for *C. difficile* infection may have been recently treated with proton pump inhibitors or certain antibiotics, particularly with clindamycin, fluoroquinolones, or cephalosporins. Additionally, advanced age, prior or lengthy hospitalization, residence in a long-term care facility and certain comorbidities may increase risk for developing a *C. difficile* infection.

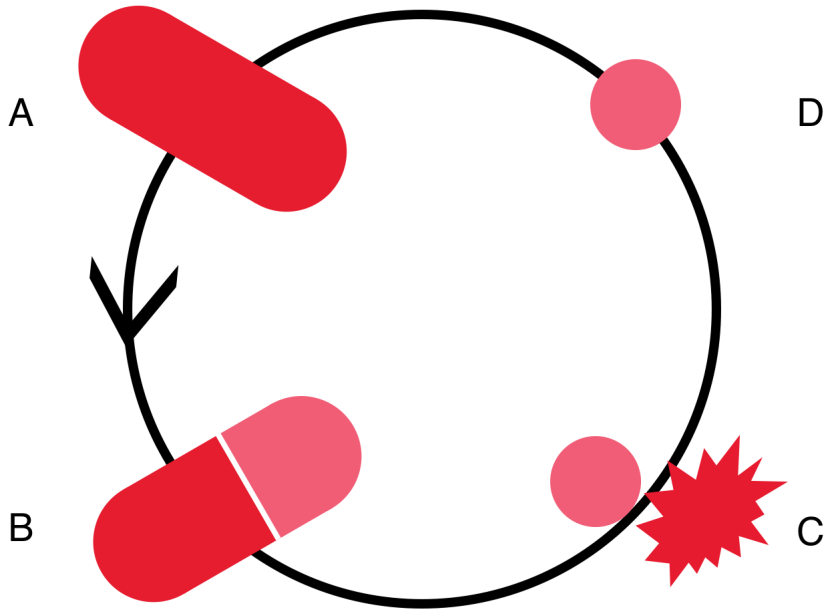

Figure 1.2: *C. difficile* life cycle

A) Vegetative cell living in colon. B) In low nutrient conditions the vegetative cell asymmetrically divides to create a mother cell and a forespore. C) Mother cell lyses to release the spore. D) The spore is resistant to oxygen, heat, and ethanol. Upon ingestion the spore may germinate in the presence of primary bile acids.

The US Centers for Disease Control and Prevention designated *C. difficile* as an urgent antibiotic resistant threat because patients receiving antibiotics are at increased risk for in-

fection[2]. The rapid spread of *C. difficile* in the US has in part been due to the emergence and spread of the epidemic lineage ribotype 027 (RT027) which took on global prominence in the 2000s [4, 5]. RT027 harbors resistance to fluoroquinolones [4, 5]. *C. difficile* treatment includes switching from the currently-prescribed antibiotic to either metronidazole or vancomycin, to which *C. difficile* is commonly susceptible[6].

The pathogenesis of *C. difficile* is mediated by the release of toxin into the gut lumen. Disease-causing *C. difficile* must produce at least one toxin, but may produce up to three toxins. The three toxins are: Toxin A (*tcdA*), Toxin B (*tcdB*), and binary toxin (encoded by two genes *cdtA* and *cdtB*)[7]. Toxin A and Toxin B are secreted by *C. difficile* cells and enter host intestinal cells through receptor-mediated endocytosis. Toxins A and B form a pore in the endosome, secrete a glycosyltransferase that binds to and deactivates Rho family GTPases. Deactivation of these GTPases causes improper functioning of the actin cytoskeleton and cell-cell junctions leading to cell rounding and apoptosis. The damage to the intestinal cells leads to inflammation and immune cell recruitment thus causing common clinical manifestations of *C. difficile* infection including diarrhea and colitis. More severe symptoms are caused by extensive tissue damage which can allow bacteria from the gut to infect the bloodstream and cause sepsis.

Toxins A and B are located in the pathogenicity locus (PaLoc; Figure 1.3). The PaLoc is 19.6kB and contains five genes: *tcdR*, *tcdB*, *tcdE*, *tcdA*, and *tcdC*[8]. *tcdR* encodes a RNA polymerase factor and positively regulates toxin gene expression[8]. *tcdR* may be negatively regulated by *tcdC*[8]. *tcdE* encodes a holin-like protein and is implicated in toxin secretion[8]. Toxins A and B each contain four domains: 1) glucosyltransferase, 2) autoprotease, 3) pore-forming, and 4) C-terminal combined repetitive oligopeptides[9]. Many systems and genes outside of the PaLoc influence toxin regulation including growth phase, access to certain metabolites, sporulation, quorum sensing, and certain flagellar proteins[10]. In particular, toxin production depends on the regulator TcdR which in turn depends on the presence of

specific sugars, amino acids, and fatty acids[11]. *C. difficile* is able to use several different carbon sources including but not limited to arbutin, glucose, glucosamine, leucine, mannitol, mannose, melezilose, salicin, tagatose, and trehalose[12]. Trehalose is of recent interest in the media because of a report suggesting that human consumption of trehalose may have contributed to the spread of two epidemic lineages of *C. difficile*, RT027 and RT078 [12]. We explore this idea further in chapter four.

Figure 1.3: *C. difficile* pathogenicity locus

1.3 *C. difficile* evolutionary history and genomic variation

1.3.1 Evolutionary history

There are 8 recognized monophyletic clades of *C. difficile* (Figure 1.4)[13]. Clade C-III is estimated to have emerged 48 million years ago, while the ancestor to clades 1-5 emerged as recently as 4 million years ago[14]. The PaLoc has had an eventful evolution. A phylogenetic tree built from whole genome sequences suggests frequent gains and losses of the PaLoc and a phylogenetic tree built from just PaLoc sequences suggests that clades 1-5 acquired the PaLoc multiple times[15].

C. difficile strains have variable antibiotic resistance patterns and their resistance depends in part on location and local antibiotic use practices. The majority of clinical *C. difficile* isolates are resistant to multiple antibiotics, commonly cephalosporins, fluoroquinolones, erythromycin, and/or clindamycin[6]. Typically, *C. difficile* infections have been treated with

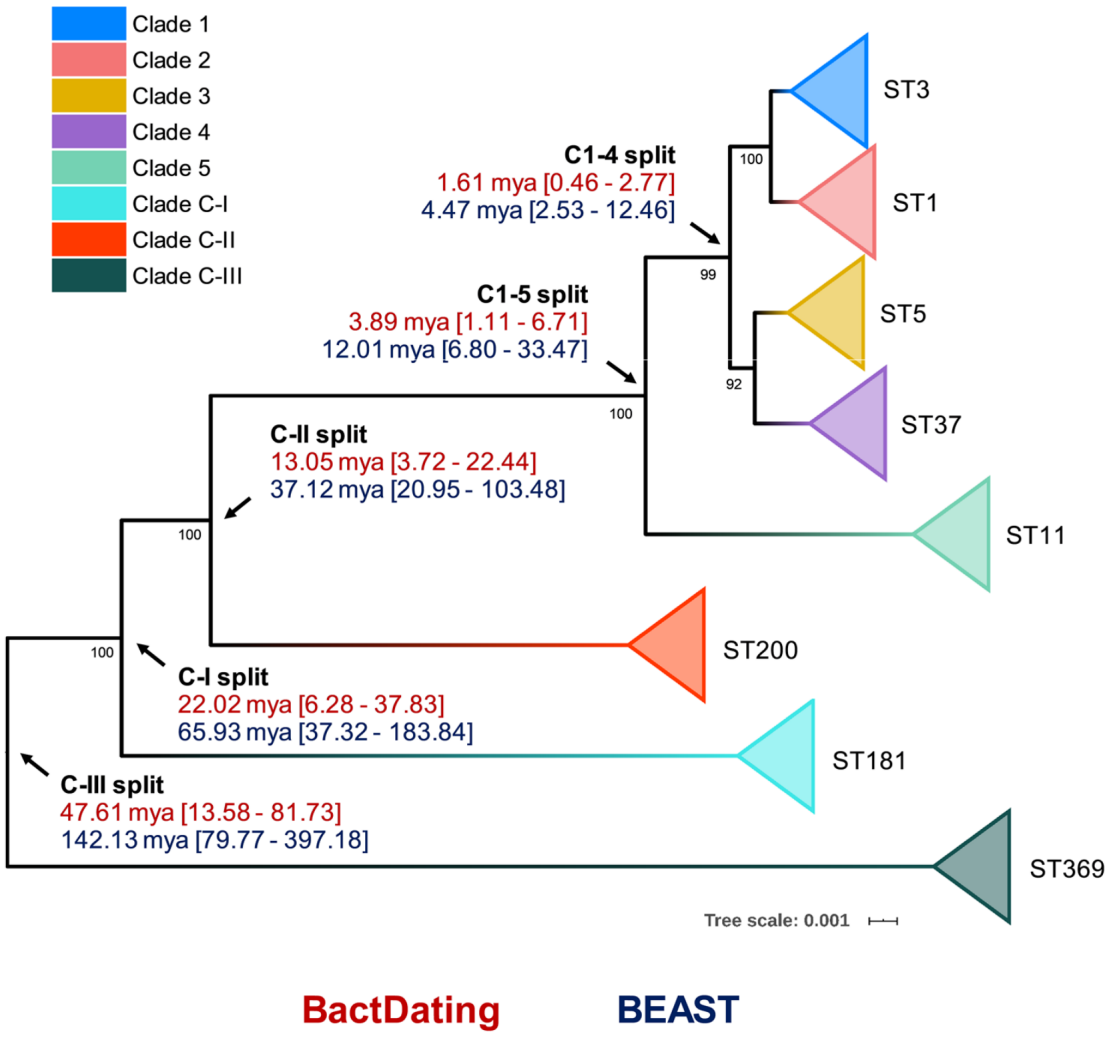

Figure 1.4: *C. difficile* phylogeny

Reproduced from Knight *et al.*[14]. Red dates refer to inferences made by BactDating while blue dates refer to BEAST estimates. The phylogeny was built from concatenated multilocus sequence typing genes.

metronidazole or vancomycin and rates of resistance to these antibiotics remains low[6]. RT027, an epidemic lineage of *C. difficile* in Clade 2, is thought to have risen to prominence in part because it acquired resistance to fluoroquinolones and expanded greatly with their increased clinical use[16]. Analyses in chapters three and four of this thesis focus on genomes from clades 1-5 and include, among other strains, RT027 isolates.

1.3.2 Genomic variation

The variation observed across *C. difficile* genomes is due to multiple causes including but not limited to horizontal gene transfer, a large pangenome, and mutations. Horizontal gene transfer is the movement of genetic material from one organism to another where the recipient is not the offspring of the donor. The *C. difficile* genome is shaped by horizontal gene transfer; for example, the reference genome CD630 is 11% mobile genetic elements[17]. Due to high levels of this lateral genomic transfer it is particularly critical in *C. difficile* studies to mask recombinant sections of the genome from analysis prior to building phylogenetic trees[18]. Genomes include mobile genetic elements such as transposable elements, bacteriophages, and CRISPR-cas elements[13]. A single *C. difficile* genome contains approximately 3,700 genes[14]. Those genes are a subset of the *C. difficile* pangenome. Pangenome estimates vary depending on the bioinformatic tool used, but range between approximately 17,000-33,000 genes[14]. The core genome is approximately 1,300 – 2,200 genes[14]. The low level of genome conservation across *C. difficile* suggests that the accessory genome is a rich source for genomic variation. When focusing on just the core genome, the relatedness of two genomes can be described by the ANI metric, which is a calculation of the average nucleotide identity of all of the orthologous genes shared by the genomes in question. ANI values within *C. difficile* clades is high (98-99%) while between clades ANI sometimes dips as low as 89%, which is well below the species threshold of 96%[14]. Such low ANI values indicate a highly diverse species and have prompted discussions of splitting *C. difficile* into multiple species.

Variation between *C. difficile* isolates can be described at various levels of resolution. For example, ANI comparisons require whole genome sequencing of core genes, while the tree shown in Figure 1.4 only required sequencing a set of seven genes used in multilocus sequence typing. Another common method to categorize and compare *C. difficile* isolates is ribotyping. Ribotyping is a simple, PCR-based typing method that amplifies the spacer region between

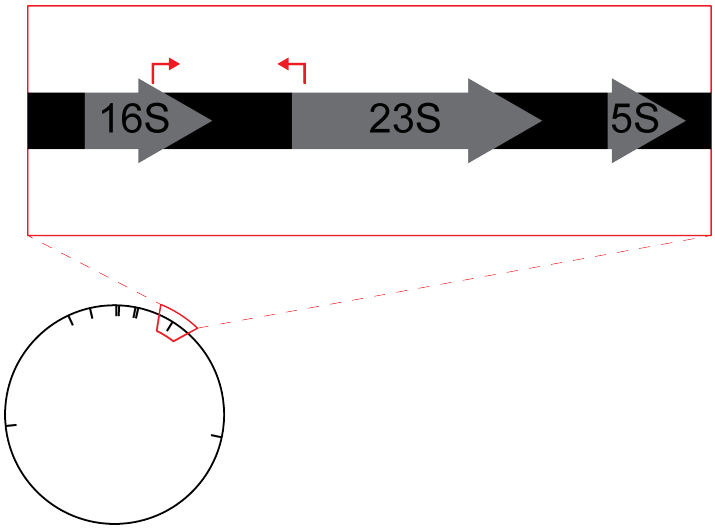

Figure 1.5: *C. difficile* ribotyping scheme

Ribotyping is a PCR-based assay wherein the region between the 16S and 23S genes in the rRNA operon is amplified and the banding pattern produced determines the isolate's identity. The circle represents the CD630 reference genome, black lines indicate genomic locations for rRNA operons. In the callout red arrows indicate the binding locations of ribotyping PCR primers.

the 16S and 23S rRNA genes (Figure 1.5)[19]. *C. difficile* has 9-12 genomic copies of the rRNA operon, and the size of the 16S-23S spacer region can vary operon to operon [20]. A ribotype is a group of *C. difficile* isolates with identical PCR banding patterns. Ribotyping, and other methods, can differentiate strains with different clinical properties (toxin presence, antibiotic resistance, etc..). The identities of ribotypes highly prevalent in clinical settings depends on location, but ribotypes that are frequently high prevalence include RT027 (Clade 2), RT014-020 (Clade 1), RT078 (Clade 5), and RT106 (Clade 2)[21]. Different regions have different relative proportions of ribotypes in the environment, community, and healthcare setting and the prevalence of individual ribotypes changes over time[22]. Ribotypes have been shown to transmit globally and, for example, there is evidence of multiple introductions of RT027 from North America to other continents[16].

Multiple studies suggest that RT027 infections lead to more severe infection outcomes in

clinical settings compared to other ribotypes[23]. RT027 quickly rose to epidemic prominence in North America and Europe in the early 2000s[16]. This ribotype is resistant to fluoroquinolones, encodes binary toxin, and secretes high levels of toxins A and B possibly due to a frameshift deletion in tcdC [23, 24]. RT014-020 can produce Toxin A and Toxin B but is binary toxin negative. RT078 has similar disease severity as RT027[25]. It can produce Toxin A, Toxin B, and binary toxin[22]. RT078 is more commonly associated with community acquired infections and infections occur in younger populations than in RT027[26].

1.4 Associating genomic variation with phenotypic variation

This thesis investigates the contribution of *C. difficile* genomic variants to clinically relevant *in vitro* phenotypes, such as toxin activity, and clinical infection outcomes via multiple approaches. Previously, researchers have used ribotype as a factor in models predicting *C. difficile* infection risk or severity with some success. Our work expands on these studies by using more nuanced whole genome sequence data from collections of clinical *C. difficile* samples to define genomic variation. There exist many methods, including genome-wide association studies (GWAS), to find statistical associations between genomic variation and phenotypic variation. We can capitalize on diverse collections of bacterial isolates by applying such methods to the existing genomic variation and identify genotype-phenotype associations. Using the naturally occurring variation in bacterial collections is an important approach because it has certain advantages over traditional microbiological methods. Bacterial GWAS (bGWAS) are particularly attractive for use in bacterial species that are hard to genetically modify. *C. difficile* is a fairly intractable bacterial species and while knock-outs and knock-ins are possible, these experiments are performed in laboratory adapted strains such as CD630, are laborious and difficult to implement[27, 28]. Additionally, bGWAS allows

researchers to examine clinical phenotypes such as *C. difficile* infection outcome rather than just those phenotypes that can be measured *in vitro* or *in vivo*. While *in vitro* and *in vivo* experiments are essential to expanding our understanding of bacterial mechanisms, findings derived from laboratory research do not always translate into clinically relevant applications.

1.5 Thesis outline

This thesis investigates the interplay of *C. difficile* genomic variation with *in vitro* toxin activity and clinical infection outcomes. In chapter two I introduce three methods to characterize the association between bacterial genomic variation with phenotypic variation. In chapter three I present preliminary results from a bGWAS examining the relationship of *C. difficile* genomic variation and *in vitro* toxin activity. In chapter four I demonstrate a lack of association between the presence of trehalose utilization variants and severe *C. difficile* infection outcomes in a clinical patient cohort. Finally, in chapter five, I discuss implications of these results and sketch out future directions for related work.

Chapter 2

Hogwash: Three Methods for Genome-Wide Association Studies in Bacteria

2.1 Preamble

This chapter provides motivation for and describes the implementation of three bacterial genome-wide association methods (bGWAS). It applies these methods to sets of synthetically generated data that model specific aspects of bacterial evolution relevant to modifying bGWAS performance. The next chapter applies this method to a set of clinical *C. difficile* isolates and their *in vitro* phenotypes.

2.2 Impact statement

We introduce hogwash, an R package with three methods for bGWAS. There are two methods for handling binary phenotypes, including an implementation of PhyC[29], as well as one method for handling continuous phenotypes. We formulate novel indices quantifying the relationship between phenotype convergence and genotype convergence on a phylogenetic tree. These indices shape an intuitive understanding for the ability of hogwash to detect significant intersections of phenotype convergence and genotype convergence and how to interpret hogwash outputs.

2.3 Introduction

2.3.1 Bacterial genome-wide association studies

bGWAS infer statistical associations between genotypes and phenotypes. Seminal bGWAS papers identified novel variants associated with antibiotic resistance in *Mycobacterium tuberculosis* and host specificity in *Campylobacter*[29, 30]. Since then, there have been numerous applications of bGWAS that have further highlighted the potential of this approach to identify genetic pathways underlying phenotypic variation and provide insights into the evolution of phenotypes of interest. Association studies can use various genetic data types including single nucleotide polymorphisms (SNPs), k-mers, copy number variants, accessory genes, insertions, and deletions. To improve the power and interpretability of bGWAS inclusion criteria or weighting can be applied to these variants based on predicted functional impact, membership in pathways of interest, or other user preferences[31, 32]. Differences between human and bacterial GWAS have been reviewed extensively by Power et al.[33]. Of note, clonality and horizontal gene transfer complicate the application of human GWAS methodology to bacteria. However, bGWAS approaches can leverage unique features of bacterial evolution, including frequent phenotypic convergence and genotypic convergence, to identify phenotype-genotype correlations.

2.3.2 bGWAS software

Several different variations of bGWAS approaches have been applied, including methods for SNPs, accessory genes (Scory)[34], or k-mers (pyseer)[35], methods using regression (pyseer)[35, 36] or phylogenetic convergence (PhyC, treeWAS)[29, 37], and methods designed for humans (PLINK)[38] or specifically for bacteria[35, 37]. Differences between standard and convergence-based bGWAS were expertly reviewed by Chen and Shapiro[39]. Convergence-based methods identify events where a genomic mutation arises independently on differ-

ent edges of a phylogeny more often in the presence of the phenotype of interest than expected by chance (Figure 2.1C). Convergence-based methods can yield higher significance with a smaller sample size, but may fail to identify some statistical associations that traditional GWAS approaches would identify when the population is clonal[39]. Additionally, convergence-based methods are limited to smaller data sets because of their large memory requirements and computational time relative to traditional methods[40], but can surmount issues of clonality.

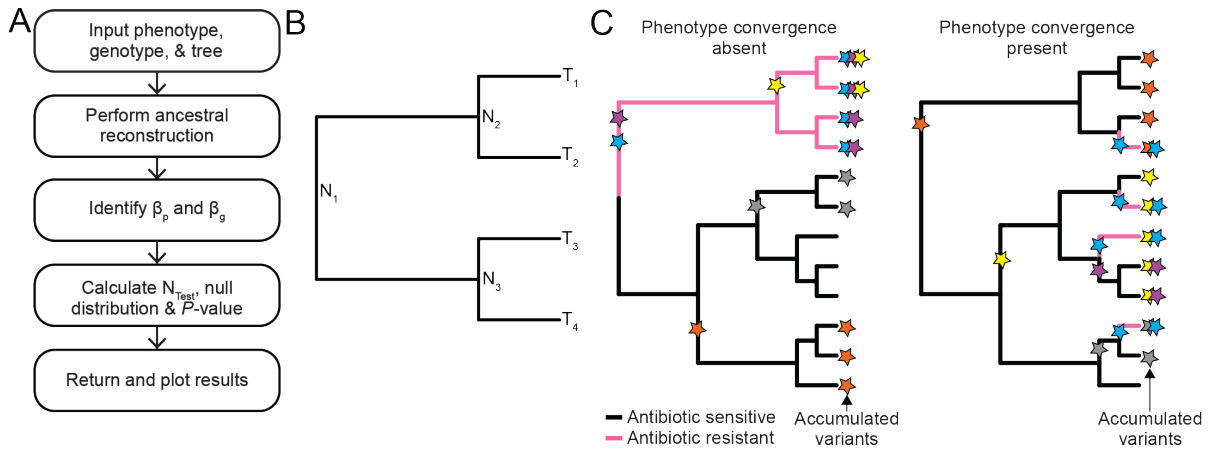

Figure 2.1: Hogwash workflow, tree nomenclature, and convergence example

A) Software workflow. B) In this example phylogenetic tree $N1$ is the root. Tree nodes are labeled $N1$ – $N3$. Tree tips are labeled $T1$ – $T4$. $N1$ is a parent node to $N2$ and $N3$. $N2$ is a child of $N1$ and a parent to $T1$ and $T2$. Edges are lines connecting a parent node to a child node or a parent node to a tip. C) A conceptual example of a phylogenetic tree with a phenotype that has arisen under two different scenarios. In the left tree antibiotic resistance, encoded by pink edges, arises once and therefore does not converge on the tree. In the right tree antibiotic resistance arises four times and therefore converges. Each colored star represents a unique genomic variant, such as single nucleotide polymorphism (SNP), that arises. Stars on edges indicate the time at which the SNP is inferred to have arisen. The stars at each tip indicate the accumulated variants found in each sample. In both trees the blue variant occurs in 4/4 antibiotic resistant isolates and 0/8 antibiotic sensitive isolates. Convergence-based association methods could only ascertain the relationship between the blue variant and antibiotic resistance in the right tree.

2.3.3 Objective

As the popularity of bGWAS increases there is a need for more widely available software that addresses specific aspects of bacterial evolution and is appropriate for various kinds of data sets. This work introduces two novel methods for convergence-based bGWAS with these needs in mind: the Synchronous Test and the Continuous Test. Users can implement these methods using hogwash, a new R package available on GitHub. Hogwash also contains an implementation of PhyC which is a bGWAS algorithm introduced by Farhat et al.[29]. The Synchronous Test is a stringent variation of PhyC, requiring a tighter relationship between the genotype and phenotype. We describe the algorithms and evaluate them on a set of simulated data. The hogwash wiki contains further explanation of bGWAS, a more conceptual introduction to these three algorithms and specific user instructions for hogwash on a set of data provided with the software package https://github.com/katiesaund/hogwash/wiki.

2.3.4 Grouped genotype analysis

Some phenotypes are not well correlated with commonly occurring genomic variants. In these cases, rare variants may provide some additional explanation for trait variability. There are multiple approaches to studying rare variants including various burden testing methods which can group loci into meaningful groups, such as mapping SNPs to genes[41, 42]. Analyzing aggregated loci can improve both the interpretability of GWAS results and improve power to detect associations[41, 42, 43, 44]. Hogwash implements two such grouping approaches to improve convergence detection for related but weakly penetrant genotypes.

2.3.5 Data simulation

We evaluate hogwash results on simulated data generated to capture aspects of bacterial evolution pertinent to these bGWAS approaches. We simulated data with a range of phylogenetic signals and convergence distributions to highlight the critical impact of these features on bGWAS results. The simulated data are publicly available and could be used to compare the impact of convergence patterns within phenotypes, genotypes, and their intersection when bench-marking various convergence-based bGWAS methods.

2.4 Package description

We developed hogwash to allow users to perform three bGWAS methods, including an open source implementation of the previously described PhyC algorithm[29], and aggregate genotypes by user-defined groups of mutations. The hogwash function minimally requires a phenotype, a phylogenetic tree, and a set of genotypes. An optional argument may be supplied to facilitate grouping genotypes. The genotypes and tree can be prepared from a multiVCF file by the variant pre-processing tool prewas[45]. Hogwash assumes that the genotype is encoded such that 0 refers to wild type and 1 refers to a mutation and that binary phenotypes are encoded such that 0 refers to absence and 1 refers to presence.

In brief, the hogwash workflow (Figure 2.1A) begins with the user supplying a phenotype, a set of genotypes, and a tree. Hogwash performs ancestral state reconstruction for the phenotype and genotypes to assign phenotype and genotype values to each tree edge (Figure 2.1B). The interaction of the phenotype with the genotypes is uniquely defined for each of the three association tests. To establish the significance of the interaction the genotypes are permuted and their intersection with the phenotype is recorded as a null distribution. Finally, we introduce an additional metric, ε, to capture the interaction between the convergence of the phenotype and genotypes.

2.4.1 Definitions

To describe the association algorithms, we introduce terms to characterize phenotypes, genotypes, and their interactions. We evaluate node values in a phylogenetic tree through ancestral state reconstruction. β is vector where each element corresponds to an edge in this tree.

- $\dot{\beta}_p$ is a binary vector indicating phenotype presence, with a value of 1 for exactly the edges with a child node with value 1 and otherwise 0.
- $\overleftrightarrow{\beta_p}$ is a binary vector indicating phenotype transitions, with a value of 1 for exactly the edges where the parent differs from the child and otherwise 0.
- $\hat{\beta}_p$ is a continuous vector that has value $\Delta_{edge} = |phenotype_{parent\ node} - phenotype_{child\ node}|$ for each edge, where Δ_{edge} values are normalized from 0 to 1.
- $\vec{\beta}_g^i$ is a binary vector indicating a genotype arising on the tree. It has a value of 1 for exactly the edges where the parent node has value 0 and the child node has value 1, for each genotype i in the set of all genotypes.
- $\overleftrightarrow{\beta_g^i}$ is a binary vector indicating genotype transitions, with a value of 1 for exactly the edges where the parent differs from the child and otherwise 0, for each genotype i in the set of all genotypes.
- We define the elementwise sum of β as $\sum \beta$.

Our three methods use different combinations of β_p and β_g. PhyC is concerned with presence and appearance ($\dot{\beta}_p$, $\vec{\beta}_g^i$). The Synchronous Test is concerned with transitions ($\overleftrightarrow{\beta_p}$, $\overleftrightarrow{\beta_g^i}$). The Continuous Test is concerned with deltas and transitions ($\hat{\beta}_p$, $\overleftrightarrow{\beta_g^i}$).

The interaction of the phenotype and genotypes are summarized as N for each method.

- We define the number of edges where both a genotype arises and the phenotype is present as $N^i_{PhyC} = \sum \vec{\beta}^i_g \wedge \dot{\beta}_p$, for each genotype i in the set of all genotypes.

- We define the number of edges where both a genotype changes and the phenotype changes as $N^i_{Synchronous} = \sum \overleftrightarrow{\beta^i_g} \wedge \overleftrightarrow{\beta_p}$, for each genotype i in the set of all genotypes.

- We define the sum of the absolute value of phenotype change on only genotype transitions edges as $N^i_{Continuous} = \overleftrightarrow{\beta^i_g}^{\top} \hat{\beta}_p$, for each genotype i in the set of all genotypes.

2.4.2 PhyC

PhyC is a convergence-based bGWAS method introduced by Farhat et al.[29] that identified novel antibiotic resistance-conferring mutations in *M. tuberculosis*. To our knowledge, the original PhyC code is not publicly available, but the algorithm is well described in the original paper. The algorithm addresses the following question: Does the genotype transition from wild type, 0, to mutant, 1, occur more often than expected by chance on tree edges where the phenotype is present, 1, than where the phenotype is absent, 0? By requiring the overlap of the phenotype with the genotype transition, instead of genotype presence, associations are not inflated by clonal sampling and thus this approach controls for population structure. We implement the PhyC algorithm as described in Farhat *et al.*[29].

For permutation tests to determine the significance of associations genotype transitions are randomized on the tree with probability proportional to the branch length. The number of edges where the permuted genotype mutation intersects with phenotype presence edges is recorded for each permutation; these permuted N^i_{PhyC} values create a null distribution. An empirical P-value is calculated based on the observed N^i_{PhyC} as compared to the null distribution.

Our PhyC implementation (Figure 2.2) has several important differences from the original paper. First, multiple test correction in hogwash is performed with False Discovery

Rate instead of the more stringent Bonferroni correction. Second, hogwash reduces the multiple testing burden by testing only those genotype-phenotype pairs for which convergence is detectable; genotypes with $\sum \vec{\beta}_g^i < 2$ are excluded and genotype-phenotype pairs with $N_{PhyC}^i < 2$ are assigned a P-value of 1. Third, ancestral state reconstruction for genotypes and phenotypes is performed using only maximum likelihood. Finally, users sacrifice some robustness in exchange for ease of use by supplying one phylogenetic tree instead of three.

2.4.3 Synchronous Test

This test (Figure 2.2) is an extension of PhyC but requires more stringent association between the genotype and phenotype. The Synchronous Test addresses the question: Do genotype transitions occur more often than expected by chance on phenotype transition edges than on phenotype non-transition edges? As in PhyC, the Synchronous Test is only appropriate for binary phenotypes.

Genotypes with $\sum \overleftrightarrow{\beta}_g^i < 2$ are removed, genotype-phenotype pairs with $N_{Synchronous}^i < 2$ are assigned a P-value of 1, and the remaining genotypes are permuted and a null distribution of $N_{Synchronous}^i$ is calculated to determine the significance of each genotype.

This test is similar to the Simultaneous Score in treeWAS[37]. The Simultaneous Score is derived from the number of edges on the tree where the genotype and phenotype transition in the same direction (both have a parent node of 0 and a child node of 1 or both have a parent node of 1 and child node of 0). In contrast, our newly developed Synchronous Test allows for the phenotype and genotype transition directions to mismatch, thus allowing for a genotype to have opposing effects on a phenotype. Such opposing effects of a genotype on a phenotype could arise when grouping mutations in the same gene that differentially impact gene function, or even for an individual mutation whose phenotypic impact may be dependent on genetic background.

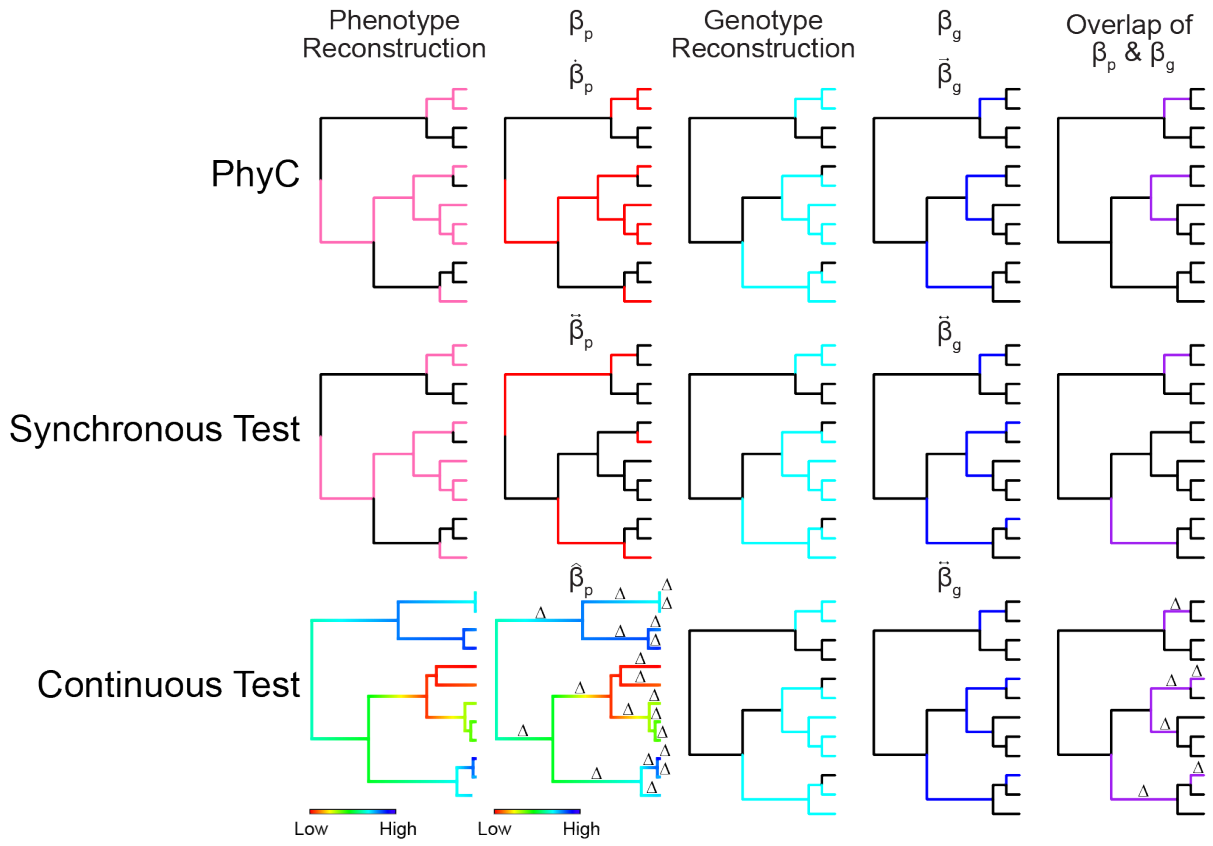

Figure 2.2: Schematic of PhyC, Synchronous, and Continuous Tests

For all binary trees black indicates 0 and a solid color indicates 1. The Phenotype Reconstruction indicates the ancestral state reconstruction for a simulated phenotype; either binary for PhyC and Synchronous Test or a range of values for the Continuous Test. The β_p indicates the test-specific β_p value taken on each tree edge; 0 or 1 for PhyC and Synchronous Test or the normalized Δ_{edge} for the Continuous Test. The Genotype Reconstruction column indicates the ancestral state reconstruction for a simulated genotype; the values are 0 or 1 in all algorithms. The β_g indicates the test-specific β_g value taken on each tree edge; the values are 0 or 1 in all algorithms. The Overlap of β_p and β_g represents the components of N_{test} The variables β_g, β_p, and N_{test} are described in the Definitions section.

2.4.4 Continuous Test

The Continuous Test (Figure 2.2) is a novel application of a convergence-based GWAS method to continuous phenotypes. The Continuous Test addresses the question: Does the phenotype change more than expected by chance on genotype transition edges than on geno-

type non-transition edges?

As above, the genotypes with $\sum \overleftrightarrow{\beta_g^i} < 2$ are removed; the remaining genotypes are permuted and a null distribution of the $N^i_{Continuous}$ is calculated to determine the significance of each genotype.

2.4.5 User inputs

The user must provide a phylogenetic tree, a set of genotypes, and a phenotype. The user may optionally provide a key that maps individual genomic loci into groups in order to use hogwash's grouping feature. For a detailed description of the user inputs please see the Supplementary Package Description.

2.4.6 Hogwash outputs

The package produces two files per test: data (.rda) and plots (.pdf). The data file contains many pieces of information, including P-values for each tested genotype. The plots are described in the Results section.

2.4.7 Grouping feature

To identify an association between a genomic variant and a phenotype hogwash requires that a variant occur in multiple different lineages. Hogwash may classify some causal variants as independent of a phenotype if they are weakly penetrant. To surmount this issue, related genomic variants may be aggregated to capture larger trends at the grouped level. For example, a user may apply this method to group only nonsynonymous SNPs by gene to use hogwash to detect associations between the mutated gene and the phenotype. Grouping related variants can improve power through a reduction in the multiple testing correction penalty. However, the power benefits are dependent on grouping variants with similar effect

directions.

By default, hogwash implements the grouping features by first performing ancestral state reconstruction for each individual locus (Figure 2.3). Then those loci are joined as indicated in the user supplied key. Grouped loci with $\sum \beta_g^i < 2$ are excluded from analysis. After this point hogwash runs as previously described for non-grouped genotypes. Alternatively, users may group together related genomic variants prior to ancestral reconstruction (Supplementary Methods). The two grouping approaches are compared in Figure S1.

2.5 Methods

2.5.1 Data simulation

Trees

We simulated eight random coalescent phylogenetic trees with 100 tips each; four trees were used for the Continuous Test and four trees were used for the binary tests.

Tree edge filtering Low confidence edges are defined as those edges with low bootstrap support (default $< 70\%$), those that are more than 10% of the total tree length, or those with low genotype or phenotype ancestral reconstruction support (maximum likelihood <0.875). Low confidence edges are ignored during permutation testing.

Phenotypes

Motivation for simulating phenotypes under two evolutionary models For each tree we simulated phenotypes under different evolutionary models: either Brownian motion or white noise. A phenotype modeled well by Brownian motion follows a random walk along the tree. A phenotype modeled well with white noise appears to be independent of tree structure and may suggest a role for horizontal gene transfer, gene loss, or convergent

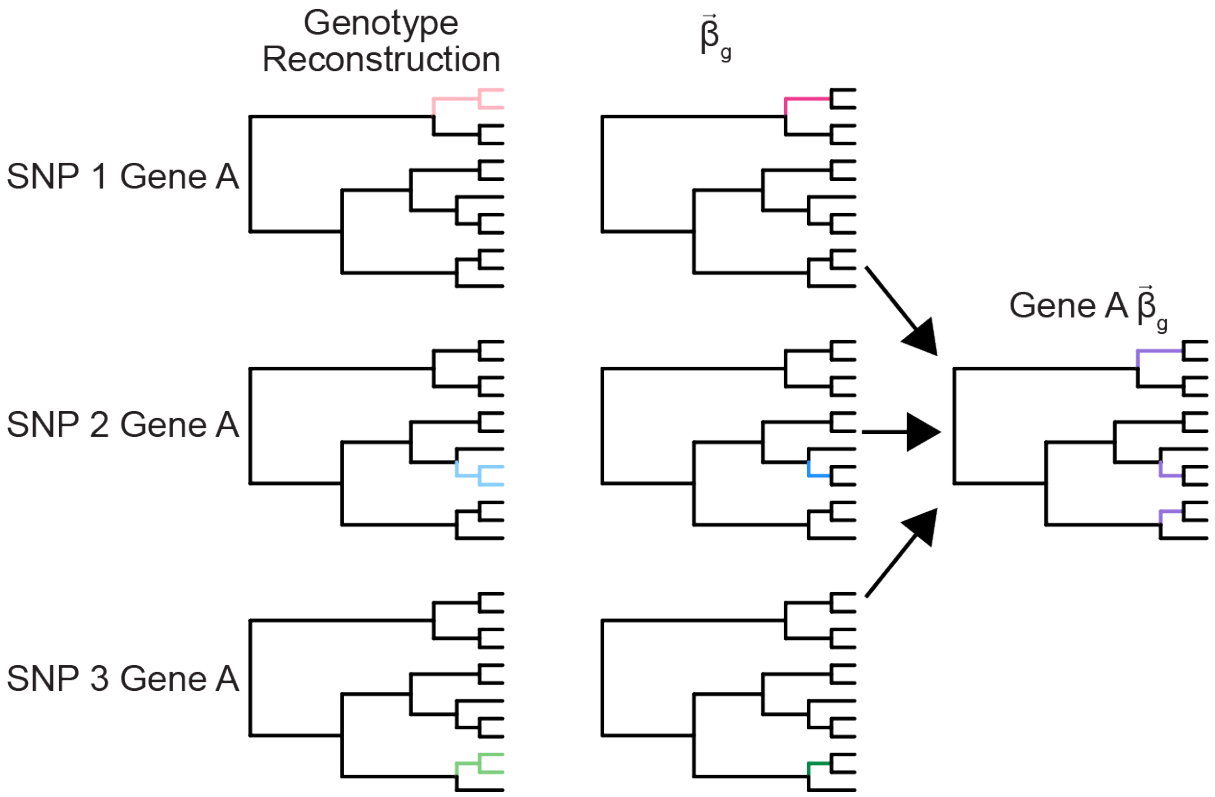

Figure 2.3: Example of hogwash grouping feature on simulated data

In this case, three SNPs are found in the same gene (Gene A). No individual SNP is convergent on the tree. Hogwash performs ancestral state reconstruction on each SNP. The edges where SNP presence is inferred are colored. Next, hogwash identifies the transitions for each SNP (colored edges). Finally, hogwash combines the three SNPs transitions together to create the Gene A transitions (purple edges). When the SNPs are grouped into Gene A the genotype converges on the tree. In this example, pre-ancestral reconstruction and post-ancestral reconstruction grouping results are identical. See Figure S1 for scenarios illustrating differences in the two grouping approaches.

evolution[46]. A white noise phenotype may be better suited to the hogwash algorithms than a phenotype modeled by Brownian motion given the requirement for phylogenetic convergence.

Calculation of phylogenetic signal Phylogenetic signal is a metric that captures the tendency for closely related samples on a tree to be more similar than random samples.

Phylogenetic signal is calculated by different metrics for continuous and binary traits; continuous traits are measured by λ while binary traits are measured by D (Figure S2). A continuous phenotype that is modeled well by Brownian motion has a λ near 1 while a white noise phenotype has a λ near 0[47]. In contrast, a binary phenotype that is modeled well by Brownian motion has a D near 0 while a white noise phenotype has a D near 1[48].

Simulation of phenotypes on trees For each tree we simulated four phenotypes fitting a Brownian motion model and four phenotypes fitting a white noise model. For phenotypes modeling Brownian motion, binary phenotypes were restricted to $-0.05 < D < 0.05$ and continuous phenotypes to $0.95 < \lambda < 1.05$. For phenotypes modeling white noise, binary phenotypes were restricted to $0.95 < D < 1.05$ and continuous phenotypes to $-0.05 < \lambda < 0.05$.

Genotypes

For each simulated tree a set of unique binary genotypes were generated. We generated genotypes that span a range of phylogenetic signals, degree of similarity to the phenotype, and prevalence.

Genotypes to be used in PhyC and the Synchronous Test First, 25,000 binary genotypes were generated using ape::rTraitDisc; these genotypes have a range of phylogenetic signals[49]. Second, these genotypes were duplicated and randomized with the following approach to reduce their phylogenetic signal: one quarter had 10% of tips changed, one quarter had 25% of tips changed, one quarter had 40% of tips changed, and one quarter were entirely redistributed. Third, we removed any simulated genotypes present in 0, 1, $N-1$, or N samples. Fourth, we subset the genotypes to keep only unique presence/absence patterns. Fifth, we subset genotypes to only those within a range of $-1.5 < D < 1.5$. These filtering steps reduced the data set size (range 2,214-2,334).

Genotypes to be in used in the Continuous Test In addition to the five steps above we added two more data generation steps. First, we made all possible genotypes based on the rank of the continuous phenotype. Second, we made genotypes based on which edges of the tree had high Δ_{edge}. The filtering steps reduced the data set size (range 1,234-1,310).

2.5.2 Hogwash on simulated data

We ran hogwash for each of the tree-phenotype-genotype sets. In addition to generating P-values for each tested genotype, hogwash also reports convergence information. We ran hogwash with the following settings for single locus analysis: permutations = 50,000; false discovery rate = 0.0005 (binary), 0.05 (continuous); bootstrap value = 0.70; no genotype grouping key was provided. For grouped analyses the settings were identical except that a grouping key was generated, and hogwash was run with both grouping methods (pre- and post-ancestral reconstruction). For the grouped analyses only PhyC was run on simulated Brownian motion phenotype 1, simulated genotype 1, and simulated tree 1. The grouping key assigned each approximately 3 unique simulated variants to each created "gene;" resulting in approximately one-third as many input genotypes when compared to the single locus analysis.

Calculation of ε

We introduce ε to quantify the degree of shared phenotype convergence and genotype convergence. Low values of ε indicate a lack of overlap in the edges where the phenotype and genotype converge. High values of ε indicate many instances of overlap in the edges where the phenotype and genotype converge. By reducing these patterns of convergence into a simple number, ε, we can more easily contextualize convergence-based bGWAS results. We define an ε for each algorithm.

- $\varepsilon^i_{PhyC} = \frac{2 \times N^i_{PhyC}}{\sum \vec{\beta}^i_g + \sum \hat{\beta}_p}$, for each genotype i in the set of all genotypes.
- $\varepsilon^i_{Synchronous} = \frac{2 \times N^i_{Synchronous}}{\sum \overleftrightarrow{\beta}^i_g + \sum \overleftrightarrow{\beta}_p}$, for each genotype i in the set of all genotypes.
- $\varepsilon^i_{Continuous} = \frac{N^i_{Continuous}}{\sum \overleftrightarrow{\beta}^i_g + \sum \hat{\beta}_p - N^i_{Continuous}}$, for each genotype i in the set of all genotypes.
- For each ε, $0 \leq \varepsilon \leq 1$.

2.5.3 Data analysis

Statistical analyses were conducted in R v3.6.2[50]. The R packages used can be found in the simulate_data.yaml file on GitHub[49, 51, 52, 53, 54, 55] and can be installed using miniconda[56].

2.6 Results

2.6.1 Motivation for evaluating hogwash on simulated data

Given our lack of comprehensive knowledge of the genetic variation contributing to any phenotype, it is not feasible to quantify sensitivity/specificity on real data. We therefore generated data that simulates genotype and phenotype distributions covering a spectrum of realistic evolutionary scenarios (spanning Brownian motion to white noise). Our goal is not to validate the premise of using convergence-based approaches, as these have been previously shown to provide useful biological insights, but rather to understand how our approach detects convergence for phenotypes with different evolutionary regimes. The following analyses can guide users in the appropriate use cases and the applicability of this method to their data. Particularly, these results provide context for interpreting the strength of the observed associations.

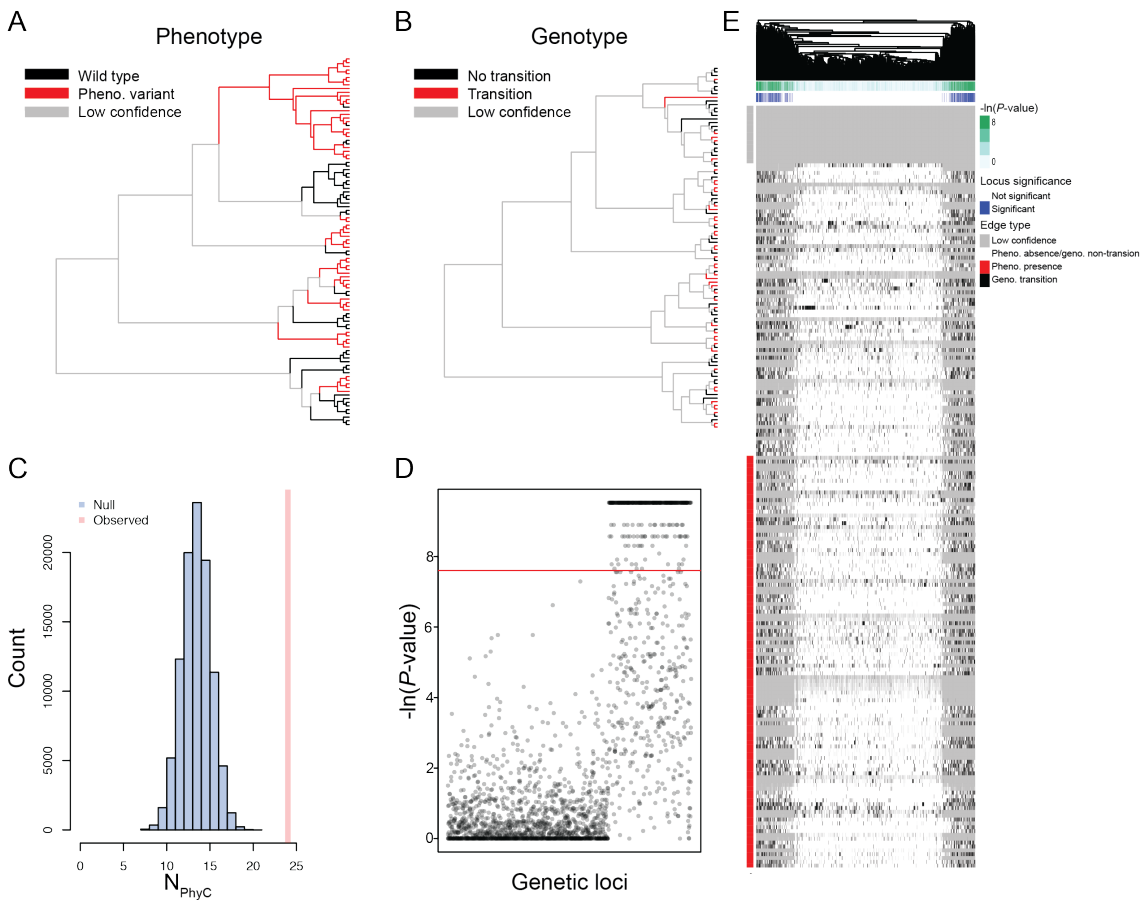

Figure 2.4: Example output from hogwash PhyC results from simulated data

A) Phenotype reconstruction. Edges with: phenotype presence in red; phenotype absent in black; low confidence in tree or low confidence phenotype ancestral state reconstruction in gray. B) Genotype transitions. Edges with: genotype mutations that arose in red; genotype mutation did not arise in black; low confidence in tree or low confidence genotype ancestral state reconstruction in gray. C) Null distribution of N_{PhyC} D) Manhattan plot. The genetic loci were simulated to achieve a range of phylogenetic signals. The left most two-thirds of genetic loci were simulated under Brownian motion models (mean $D = 0.16$) while the remaining third were modeled by white noise (mean $D = 0.99$). E) Heatmap with tree edges in the rows and genotypes in the columns. The genotypes are hierarchically clustered. The genotypes are classified as being a transition edge in black or non-transition edge in white. The column annotations pertain to loci significance; green indicates the P-value while blue indicates that the P-value is more significant than the user-defined threshold. The row annotation classifies the phenotype at each edge; red indicates phenotype presence and white indicates phenotype absence. Gray indicates a low confidence tree edge; low confidence can be due to low phenotype ancestral state reconstruction likelihood, low genotype ancestral state reconstruction likelihood, low tree bootstrap value, or long edge length.

2.6.2 Hogwash output for simulated data

Hogwash outputs two sets of results: a data file and a PDF file with plots. Each run of PhyC produces at least three plots: the phenotype reconstruction (Figure 2.4A), a Manhattan plot (Figure 2.4D), and a heatmap of genotypes (Figure 2.4E). The phenotype presence is highlighted on the tree (Figure 2.4A). The Manhattan plot shows the distribution of P-values from the hogwash run (Figure 2.4D). The heatmap shows the genotype reconstruction and phenotype reconstruction for each tree edge (rows) and genotype (columns) (Figure 2.4E). The genotypes are clustered by the presence/absence pattern. Two additional plots are produced for each genotype that is significantly associated with the phenotype: a phylogenetic tree showing the genotype transition edges (Figure 2.4B) and the null distribution of N_{PhyC} (Figure 2.4C). As expected, the two grouping approaches identified different associations as compared to non-grouped PhyC results (Figure S3).

The Synchronous Test and Continuous Test output plots that reflect their test-specific β and N definitions (Figure S4, S5). Running hogwash on 100 samples required < 3 hours and < 2 GB of memory for binary data and < 5 hours and < 2 GB of memory for continuous data (Figure S6).

2.6.3 Hogwash evaluation on simulated data

To help users identify optimal use cases and also interpret hogwash results we describe the behavior of hogwash on simulated data. We note that this assessment is not meant to convey performance in the sense of calculating sensitivity and specificity, but rather evaluate whether hogwash can robustly detect the association between phenotypic and genotypic convergence. To guide our assessment, we compared the relationship between the P-value and ε values produced by hogwash on sets of simulated data constructed using different evolutionary models (Figure 2.5). ε is a quantification of the relationship between phenotype

	Phenotype	
	Brownian motion	White noise
PhyC	0.91	0.93
Synchronous Test	0.60	0.94
Continuous Test	NA	0.08

Table 2.1: Mean Spearman's rank correlation coefficient for -ln(P-value) versus ε from hogwash run on simulated data.

The ρ could not be calculated for the results from the Continuous Test on the Brownian motion phenotypes because, after multiple testing correction, all P-values are identical.

convergence and genotype convergence. Low ε values indicate little to no intersection of phenotype convergence and genotype convergence, while higher ε values indicate their increased intersection. The ε value is always a fraction between 0 and 1 and therefore obscures information about the sample size; to account for the number of samples in the tree we recommend always interpreting ε value for any locus with its P-value.

For binary phenotypes, we observe an overall strong positive association between -ln(P-value) and ε, demonstrating that as the intersection of phenotype convergence and genotype convergence increase hogwash predicts that it is less likely that they intersect due to chance (Table 2.1). In other words, below a certain ε_{binary} threshold (ε_{binary} is ε_{PhyC} or $\varepsilon_{Synchronous}$), hogwash attributes the association between the genotype convergence and phenotype convergence to chance; from Figure 2.5 the user can get a sense for the range of this ε_{binary} threshold under different evolutionary regimes.

For the simulated continuous data an $\varepsilon_{Continuous}$ threshold that separates meaningful genotype-phenotype associations from associations by chance is less apparent. Higher ε, low significance values demonstrate that some overlap of β_g and β_p is likely by chance given the data. Low ε, high significance genotype-phenotype pairs demonstrate that sometimes small amounts of β_g and β_p overlap are unlikely, however that does not necessarily suggest that these hits are the best candidates for *in vitro* follow up. We suspect that these associations are largely driven by poor exploration of the sampling space, despite running many permutations,

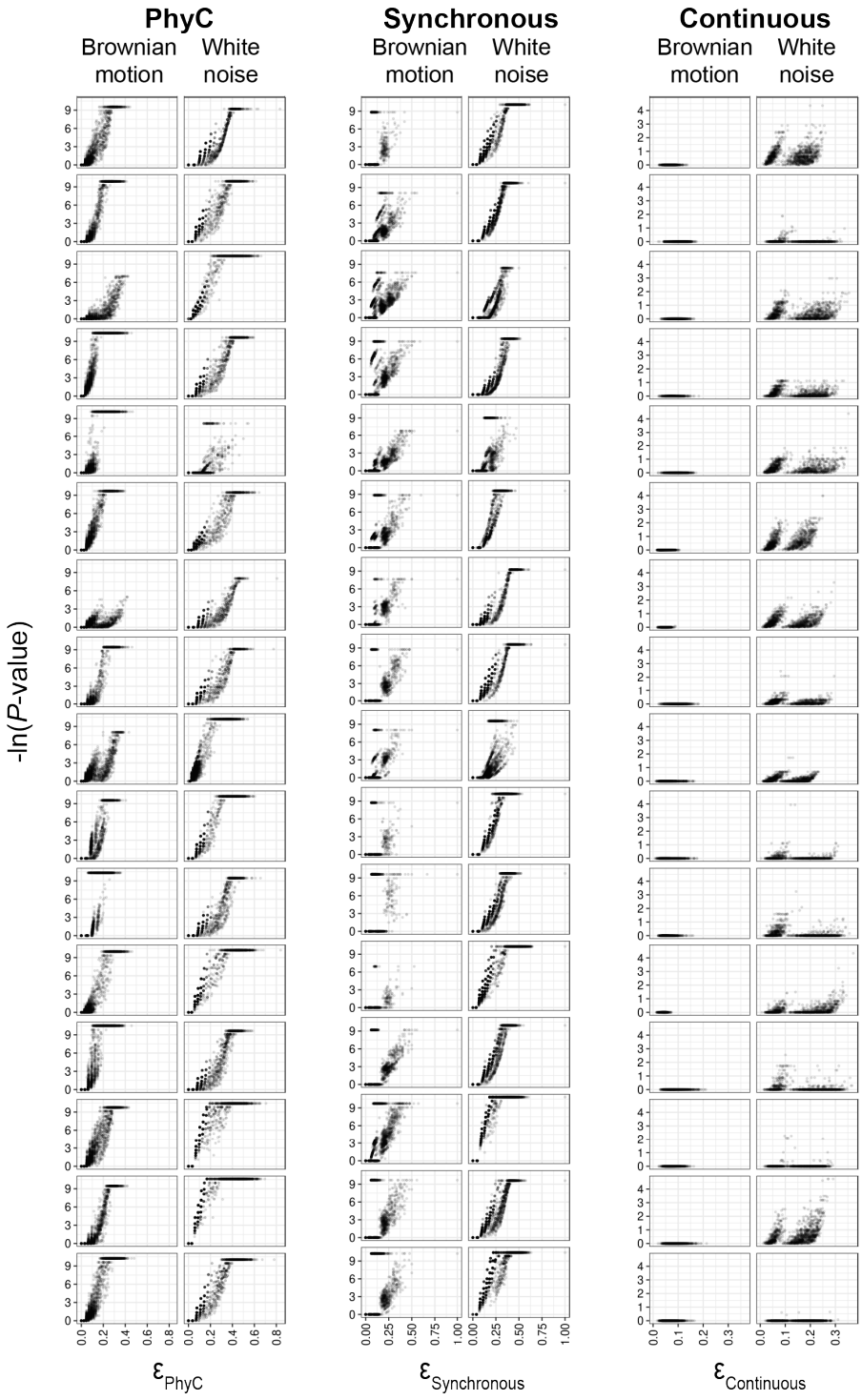

Figure 2.5: High ε values correlate with increased significance

Each plot is a tree-phenotype pair. Each point represents one genotype-phenotype pair. Brownian motion and white noise refer to the evolutionary regime modeled by the phenotype. The genotypes span a range of phylogenetic signals.

because of the edge-length based sampling probability of the permutation method. Therefore, it is essential P-values be interpreted within the context of ε. Notably, the Continuous Test was only able to detect significant genotype-phenotype associations for phenotypes modeled by white noise, suggesting this method is particularly sensitive to the phenotype's evolutionary model. We observe for both the binary and continuous tests that ε is more tightly correlated with -ln(P-value) for phenotypes characterized by white noise than by Brownian motion (Table 2.1), indicating that hogwash performs better under a white noise model. Therefore, we suggest using the report_phylogenetic_signal function on the phenotype prior to running hogwash to ascertain the appropriateness of these algorithms for the dataset.

2.7 Discussion

We have developed two algorithms for convergence-based bGWAS that are particularly well suited for phenotypes modeled by white noise. Hogwash, is straightforward to install in R, accepts easy-to-format data inputs (described in detail on the wiki), and provides publication ready plots of the GWAS results. Hogwash also implements grouping features to aggregate related genomic variants to increase detection of convergence for weakly penetrant genotypes. Hogwash is best used for datasets comprising binary or continuous phenotypes, phenotypes fitting white noise models, situations where convergence may occur at the level of genes or pathways and with datasets whose size can be accommodated given the time and memory constraints of convergence methods.

The results of running hogwash on simulated data suggest that after a certain ε threshold, it unlikely that the intersection between phenotype convergence and genotype convergence occurs by chance, particularly for white noise phenotypes. Given the variability in results within each method, as shown in Figure 2.5, users may want to contextualize the statistical significance of the tested genetic loci with the amount of convergence possible for any one

particular data set; to facilitate this the hogwash output includes both P-values and ε.

The simulated data set presented here is published to serve as a resource or template for future work focused on benchmarking convergence-based bGWAS software as such a dataset has not yet, as far as we are aware, been made available[57]. The simulated data set is available on GitHub and includes convergence information for each phenotype, genotype, and their intersection.

2.8 Acknowledgements

We thank Brad Saund for his help formalizing the continuous algorithm ε definition.

2.9 Supplement

2.9.1 Extended package description

User provided inputs

Tree The provided phylogenetic tree may be rooted; unrooted trees will be midpoint rooted. The tree must be fully bifurcating. We recommend building the phylogenetic tree with an outgroup using a maximum likelihood framework, root to the outgroup, and then remove the outgroup prior to running hogwash. We also recommend that trees be built from recombination-filtered genomic data[18].

Genotype The required structure of the genotype data object is a matrix. The rows correspond to samples and should be ordered to match the tips of the phylogenetic tree. The row names must exactly match the tree's tip labels. The columns correspond to individual genotypes. The matrix should have both row names and column names. Genotypes can be SNPs (core genome), genes (accessory genome) or other types (indels, pathways,

etc...). Genotypes must be encoded in binary (0/1). We assume that the user has chosen a meaningful reference and recoded the reference allele as 0 and variants as 1.

Phenotype The required structure of the phenotype data object is a matrix. The rows correspond to samples and should be ordered to match the tips of the phylogenetic tree. There should only be one column, which contains the phenotype data. The matrix should have both row names and a column name. The row names must exactly match the tree's tip labels. The phenotype can either be binary (0/1) or continuous. We assume that the user has chosen a meaningful reference for binary phenotypes.

Optional: Genotype grouping key To perform grouping, for example of SNPs into genes, the user must provide a key that maps the individual loci into groups. This matrix has two columns, where the first column has names of the genotypes included in the provided genotype matrix. The second column has names for the group to which the item in the first column belongs. If the individual locus should be mapped to more than one group, the user should add additional rows for each additional group. Row names are not required. The column names are used in output plots and therefore must be included.

Filters and checks performed on user-provided inputs

Genotype Hogwash filters will remove genotypes whose variants are present in only 0, 1, N or $N-1$ genomes because convergence is not detectable. This removal step occurs prior to ancestral state reconstruction for the single locus (non-grouped) test and in the pre-ancestral reconstruction grouping method. This removal step occurs after ancestral state reconstruction for the post-ancestral reconstruction grouping method. Additionally, variants with missing information are removed as are variants with fewer than two high confidence transition edges because otherwise no convergence is detectable ($\sum \beta_g < 2$). Any edges

with low genotype ancestral state reconstruction support (maximum likelihood < 0.875) are excluded from the analysis.

Phenotype Hogwash filters will exclude any edges with low phenotype ancestral state reconstruction support (maximum likelihood < 0.875) from the analysis. Hogwash reports the phenotype's phylogenetic signal to the user. The user can check the phylogenetic signal prior to running hogwash with the function report_phylogenetic_signal.

Tree Hogwash hard filters exclude the following tree edges from analysis: A) low bootstrap support (default: $< 70\%$) and B) overly long ($> 10\%$ of total tree edge length). Overly long edges are excluded because ancestral state reconstruction accuracy decreases with edge length[58]. The tree must be fully bifurcating. Unrooted trees are midpoint rooted.

Grouping The user can choose between two grouping methods: either pre- or post-ancestral reconstruction grouping (Figure S1). Pre-ancestral reconstruction grouping occurs before the genotype ancestral states are inferred; genotypes are grouped, and then ancestral reconstruction is performed. Pre-ancestral reconstruction grouping is fast but is not as sensitive as post-ancestral reconstruction grouping. Post-ancestral reconstruction grouping occurs after the ancestral states and genotype transitions are determined for each individual (un-grouped) genotype.

2.9.2 Supplementary figures

For supplementary figures visit https://www.biorxiv.org/content/10.1101/2020.04.19.048421v2.

Chapter 3

Genomic Variants Associated with Toxin Activity in *Clostridium difficile*

3.1 Preamble

This chapter applies hogwash's Continuous Test to *in vitro* toxin activity in a set of *C. difficile* samples. We explore these hogwash results. First, we examine the presence of variants in genes known to modulate toxin production in the set of variants found to significantly associate with toxin activity. Second, we speculate on the potential role of variants and genes found to associate with toxin activity that have not been previously described and suggest potential experiments for functional validation.

3.2 Introduction

Clostridium difficile is a toxin-producing, healthcare-associated pathogen. Infections with *C. difficile* infection may manifest as diarrhea, fever, abdominal pain, colitis, shock, intestinal perforation, or death[3]. Disease is caused by secreted toxins, primarily Toxins A (*tcdA*) and B (*tcdB*), which damage the cytoskeletons of intestinal cells causing cell death and leading to gut inflammation. These two toxins are large proteins with four domains: glucosyltransferase, autoprotease, pore-forming, and C-terminal combined repetitive oligopeptides

(CROPs)[9]. Toxins A and B are colocated within the pathogenicity locus (PaLoc) with three other genes: *tcdR*, *tcdE*, and *tcdC*. *tcdR* is a positive regulator of *tcdA* and *tcdB* and encodes an RNA polymerase factor[8]. Mixed evidence suggests that *tcdC* is a possible negative regulator of *tcdR*[8]. *tcdE* encodes a holin-like protein and may contribute to toxin secretion[59]. Many factors and systems are implicated in PaLoc regulation including growth phase, access to specific metabolites, sporulation, quorum sensing, and some flagellar proteins[10].

Approaches for uncovering regulation and evolution of phenotypes, such as toxin activity, include *in vitro* assays, comparative genomics, and bacterial genome-wide association studies (bGWAS). Knock-out and complementation *in vitro* assays are frequently informative, but they only resolve biological questions in the context of a specific strain and *in vitro* condition. The contextual specificity of such experiments can lead to conflicting reports, such as the controversial findings about the role TcdE in toxin secretion in *C. difficile*. Some experiments point to a plausible involvement of TcdE in toxin export while others suggest completely TcdE-independent release of toxin from cells[59, 60]. Comparative genomics approaches are increasingly common in *C. difficile* studies as thousands of genomes are now available in public repositories. These comparative reports offer insight into general trends and evolutionary events such as horizontal gene transfer, which has certainly shaped the history of the PaLoc[61, 62]. Comparative genomics approaches tend to focus on trends at the level of strain types and largely ignore individual genomic variants[61]. bGWAS harnesses the natural variation within a bacterial population and thus surmounts the strain-dependence of *in vitro* experimentation and highlights specific variants frequently overlooked in comparative genomics approaches. bGWAS have been applied to *C. difficile* populations to uncover accessory genes associated with specific clades[14] or disease severity in a mouse model of infection[63]. Some studies have identified putative variants associated with strains producing higher levels of toxin, such as an adenosine deletion at position 117 in *tcdC* in ribotype 027 (RT027); these studies have not taken a genome-wide approach and seem to

be confounded by population structure[64]. Notably, Lewis *et al.* used only 33 samples and both C. difficle bGWAS studies used only accessory genes as genomic variants. To uncover genomic variants within *C. difficile* strains that may mediate changes in toxin activity we performed a genome-wide association study (GWAS). In our approach we include single nucleotide polymorphisms (SNPs), insertions and deletions (indels), and accessory genes to capture more of the genomic variation within the study population.

3.3 Methods

3.3.1 Study population and *in vitro* toxin activity

The University of Michigan Institutional Review Board approved all sample and clinical data collection protocols used in this study (HUM00034766). Where applicable, written, informed consent was received from all patients prior to inclusion in this study. Stool samples were collected from a cohort of 106 Michigan Medicine patients with CDI from 2010-2011, which included all severe cases during the study period[65]. A clonal spore stock from each patient was used for ribotyping and *in vitro* studies. Experiments characterized the toxin activity of these isolates as well as two laboratory *C. difficile* strains VPI and CD630[65]. Vero cells were incubated with *C. difficile* supernatant; toxin activity was measured as a function of Vero cell viability compared to viability in the presence of known quantities of purified Toxin B in ng/ml[65].

3.3.2 Genomic analysis

The spore stocks were grown in an anaerobic chamber overnight on taurocholate-coition-cycloserine-fructose agar plates. The next day a single colony of each sample was picked and grown in Brain Heart Infusion medium with yeast extract liquid culture media overnight. The

vegetative *C. difficile* cells were pelleted by centrifugation, washed, and then total genomic DNA was extracted. Genomic DNA extracted with the MoBio PowerMag Microbial DNA Isolation Kit (Qiagen) from *C. difficile* isolates (N=108) was prepared for sequencing using the Illumina Nextera DNA Flex Library Preparation Kit. Sequencing was performed on either an Illumina HiSeq 4000 System at the University of Michigan Advanced Genomics Core or on an Illumina MiSeq System at the University of Michigan Microbial Systems Molecular Biology Laboratories. Quality of reads was assessed with FastQC[66] Adapter sequences and low-quality bases were removed with Trimmomatic[67]. Details on sequenced strains are available in Table S1. Sequence data are available under Bioproject PRJNA594943. Pangenome analysis was performed with roary[68]. Accessory genes annotations were assigned by prokka[69]. SNP and indel annotations were extracted from the reference genome CD630 (AM180355.1).

3.3.3 Genome-wide association study

GWAS was performed with hogwash v1.2.4[70]. Toxin activity data were log transformed. Individual locus analysis hogwash settings: bootstrap threshold=0.95, and permutations=10,000. Individual loci for analysis included SNPs, indels, and accessory genes. Grouped locus analysis hogwash settings: bootstrap threshold=0.95, and permutations=10,000, grouping=post-ancestral reconstruction. For the grouped locus analysis only SNPs and indels were considered; coding SNPs and coding indels were grouped into the appropriate gene while non-coding SNPs and non-coding indels were grouped into the intergenic region between the two nearest genes.

3.3.4 Variant calling

Variants were identified by mapping filtered reads to the CD630 reference genome (GenBank accession number AM180355.1) using bwa[71], removing polymerase chain reaction duplicates with Picard[72], and calling variants with SAMtools and bcftools[73]. Variants were filtered from raw results using GATK's VariantFiltration (QUAL, >100; MQ, >50; >=10 reads supporting variant; and FQ, <0.025)[74]. Variants were referenced to the ancestral allele using prewas[45].

3.3.5 Phylogenetic analysis

Consensus files generated during variant calling were recombination filtered using Gubbins[18]. The alleles at each position that passed filtering were concatenated to generate a non-core variant alignment relative to the CD630 reference genome. Alleles that did not pass filtering were considered unknown (denoted as N in the alignment). Variant positions in the alignment were used to reconstruct a maximum likelihood phylogeny with IQ-TREE v1.5.5 using ultrafast bootstrap with 1,000 replicates[75, 76]. ModelFinder limited to ascertainment bias-corrected models was used to identify the best nucleotide substitution model[77]. The tree was midpoint rooted.

3.3.6 Data analysis

Data analysis in R v3.6.2[50] was performed with following packages: ape v5.3[49], phytools v0.6-99[52], tidyverse v1.3.0[53], ggtree v2.0.4[78], aplot v0.0.6[79], data.table v1.12.8[80]. pheatmap v1.0.12[81].

Permutation testing

The empirical P-value for enrichment of toxin variants in the significant individual GWAS results and flagellar genes and intergenic regions in the significant grouped GWAS results were generated by permutation testing. The null distribution was generated from random sampling without replacement repeated in 10,000 trials.

3.4 Results

3.4.1 Individual locus GWAS identifies variants associated with toxin activity

To identify all loci associated with toxin activity we performed a GWAS. *In vitro* toxin activity was assessed in a set of 106 *C. difficile* strains isolated from stool samples of infected patients[65]. We used a convergence-based bacterial GWAS approach[70] to control for population structure. We included SNPs, indels, and accessory genes as sources of genomic variation. A False Discovery Rate of 3% was applied to the resulting P-values (Figure 3.1).

PaLoc variants We expected that variants located within the PaLoc would be significantly associated with toxin activity. Consistent with this, we observed PaLoc variants in the pool of significant results. Of the 134 individual loci significantly associated with toxin activity, 15 occur in *tcdB* and 1 occurs in the *tcdR-tcdB* intergenic region (Figure 3.2), which is a significant enrichment compared to the number of variants within or flanking *tcdB* that are expected by chance (P-value=0.0001; median=1; range=0-8). *tcdB* variants were found in all four protein domains, but those 15 significantly associated with toxin activity are mostly found within the autoprotease domain (Figure 3.2). Certain significant missense variants within *tcdB* have plausible functional impacts on Toxin B such as the adenosine

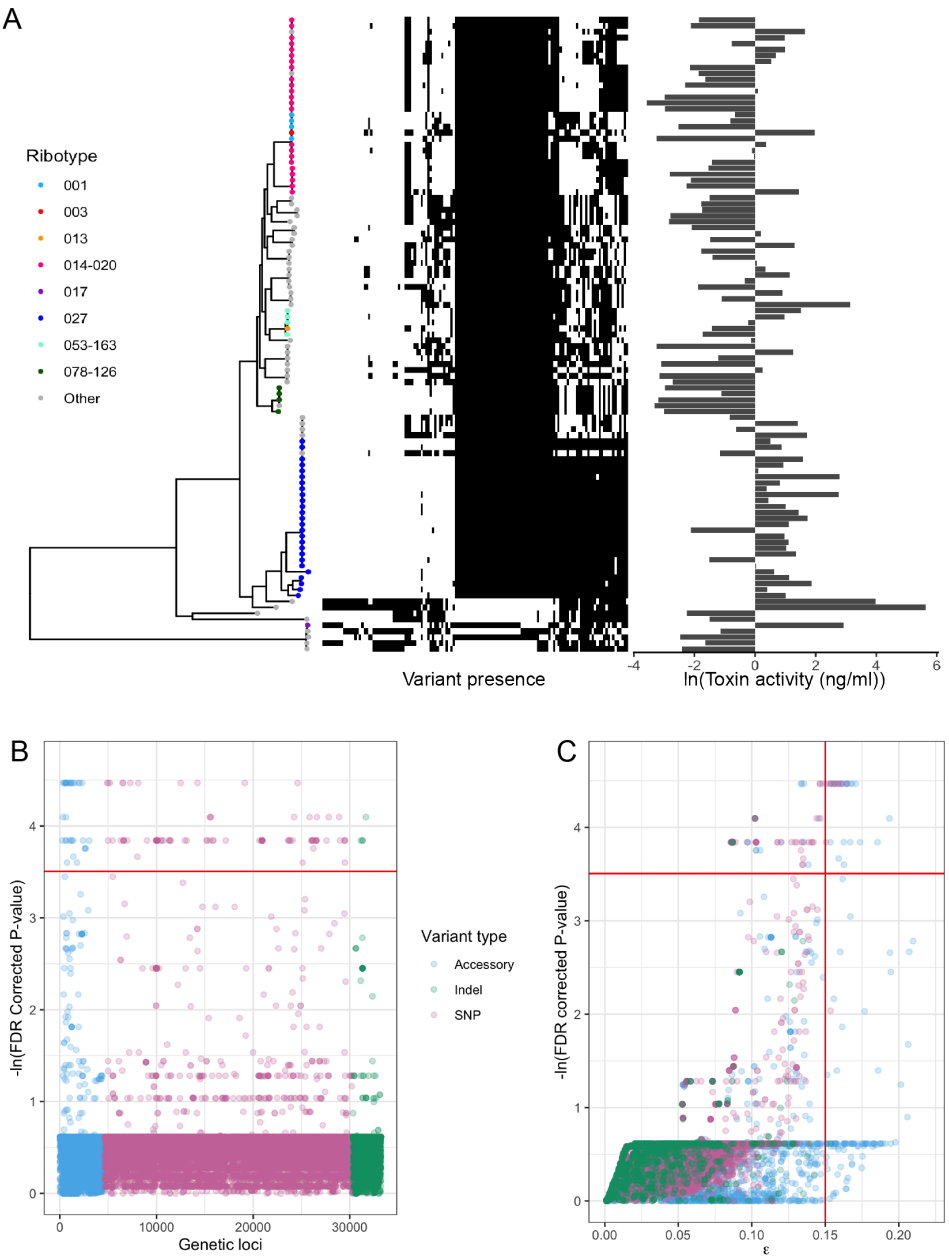

Figure 3.1: Individual locus GWAS identifies significant associations between various types of genomic variants and toxin activity

A) Left: Phylogenetic tree labeled by ribotype. Center: Heatmap indicating the presence of individual loci strongly associated with toxin activity (those variants located in the top right quadrant in C. Right: Toxin activity. B) The Manhattan plot of all individual loci assigned a P-value in hogwash. Loci are either accessory genes (blue; N=4,671), SNPs (pink; N=25,668), or indels (green; N=2,895). C) The loci from B are visualized in a dot plot of P-value vs ε, a convergence metric. The red horizontal line indicates a False Discovery Rate of 3%. The red vertical line separates low vs high convergence.

to cytoseine transversion at position 1967 which changes an aspartic acid to alanine; this mutation occurs near the zinc binding site and could theoretically affect the protein's ability to autoprocess within the host cell. Most of the 15 variants occur on the protein's surface, but an alanine to glycine mutation at residue 1231 occurs in the middle of a beta-strand in the pore-forming domain and likely destabilizes the protein. Of the 15 tested variants that occur within the *tcdR-tcdB* intergenic region, only 1 was significantly associated with toxin activity. This variant (C $\rightarrow$ T) occurs within a *tcdB* promoter suggesting a potential role in modulating sigma factor binding and therefore altering *tcdB* transcription. A notable lack of association was observed between an adenosine deletion at nucleotide 117 in *tcdC* that has been suggested to cause increased toxin production in RT027[64]. This deletion was found in 26/26 RT027 samples as well as in 3 additional samples ("Other" ribotype) but did not reach significance (-ln FDR-corrected P-value=0.56).

Novel variants

This GWAS was performed to generate hypotheses about as yet unknown associations between genomic variants and toxin activity. The variants that were significant and had high ε, a metric of shared convergence, are summarized in Tables 3.1 and 3.2. A single ε value captures the number of tree edges where a genotype mutated and the toxin activity value had a large fluctuation. ε values close to zero suggest that the genotype mutates on very few edges where the toxin activity changes drastically.

Novel SNPs There are several variants that may be worth pursuing in follow up functional studies, to be performed in more genetically tractable bacterial species. Genes of top priority may include CD630_10170 and CD630_04840 which could potentially contribute to Toxin A and B secretion. The export process of Toxins A and B is not well understood. *tcdE* has been proposed to encode for a potential export protein but these results have been disputed

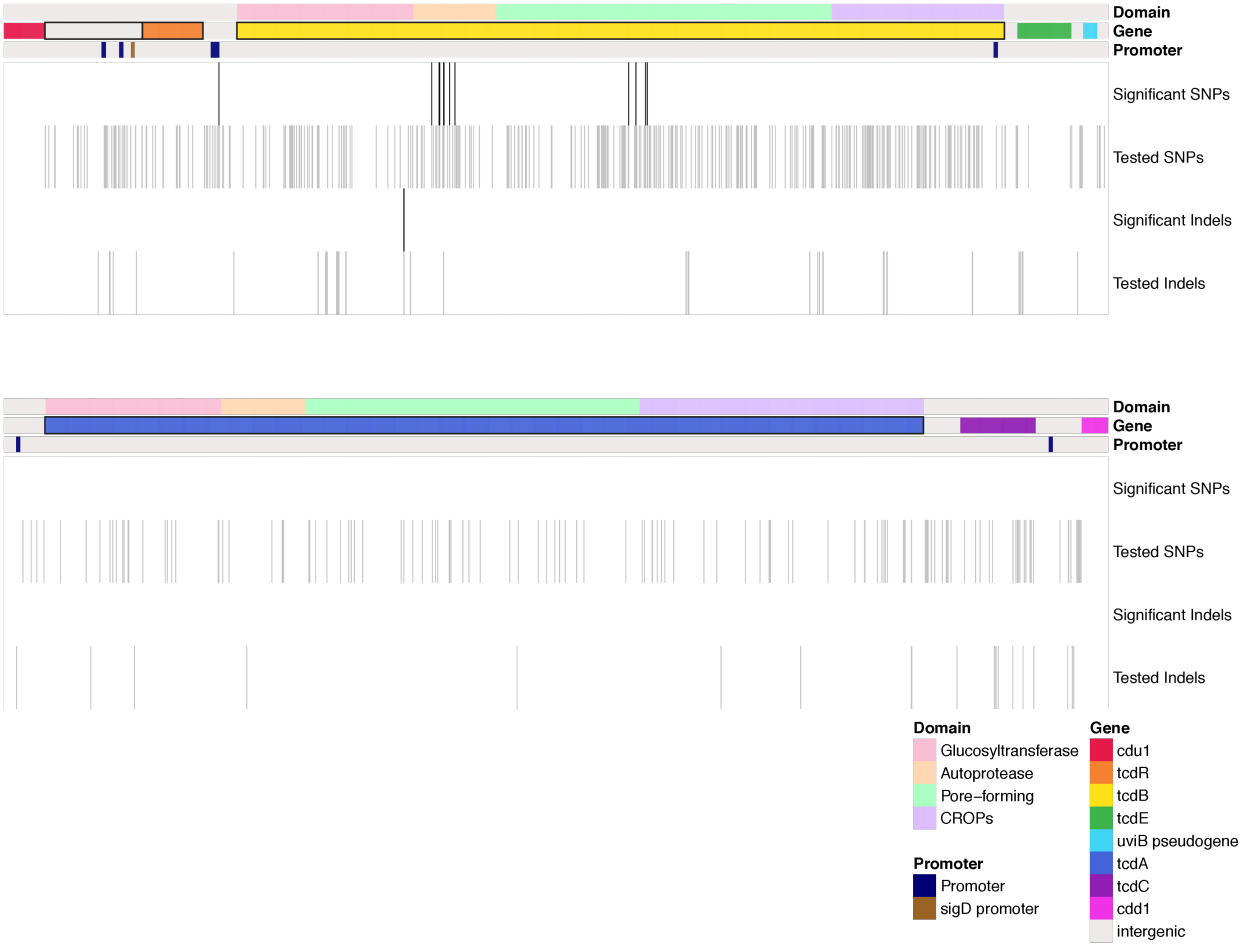

Figure 3.2: Genomic locations of PaLoc variants tested in individual locus and grouped locus GWAS

All PaLoc variants tested by GWAS are plotted in the second and fourth rows of the heatmap (SNP N=563, Indel N=70). Variants significantly associated in the individual locus test are plotted in the first and third rows (SNP N=14, Indel N=2). All tested variants were grouped into the appropriate gene or intergenic region for the grouped locus test. Loci significantly associated in the grouped locus GWAS are outlined with black boxes. Schematics of the PaLoc plus adjacent genes *cdu1* and *cdd1* are shown in the annotations. Top annotation: protein domains for Toxins A and B[9]. Center annotation: gene. Bottom annotation: promoter locations[82, 83, 8]

in another *C. difficile* strain[59, 60]. CD630_10170 and CD630_04840 are both ATP-binding cassette (ABC) transport system proteins. ABC transport systems are ubiquitous molecular importers and exporters. ABC transporters have been known to export toxins such as hemolysin[84] and therefore could potentially contribute to toxin secretion in *C. difficile*. An important caveat to proposed investigations of novel variant is that some of the highly associated variants are inherited together, having identical presence/absence patterns among isolates, and therefore perhaps only one in each inherited group, if any, cause changes in toxin activity.

Novel accessory genes The top set of accessory genes generally lack informative annotation and therefore future work will need to check these putative proteins for homology to protein domains that could plausibly affect steps in toxin production. Toxin production relies on sensors, transcriptional regulators, and secretory proteins; accessory genes might affect any of these three steps in toxin production.

SNP locus	*P*-value	ε	Reference	Variant	Variant type	Nucleotide position	Annotation
CD630_02731	4.469	0.164	G	A	missense	54	Null or hypothetical protein
CD630_21090-CD630_21100	4.469	0.162	A	G	intergenic	NA	Putative anion:Na+ symporter - putative transaldolase C-like
CD630_00340	4.469	0.159	T	C	missense	994	Putative xylose transporter, sodium:galactoside symporter family
CD630_00470	4.469	0.159	A	G	missense	283	ispD: 2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase
CD630_01090-CD630_r0130	4.469	0.157	T	C	intergenic	NA	nrdG: Anaerobic ribonucleoside-triphosphate reductase-activating protein - 16s rRNA
CD630_29780	4.469	0.156	A	C	missense	1269	Putative CRISPR-associated Cas3 family helicase
CD630_10170	4.469	0.154	C	A	missense	1358	ABC-type transport system, multidrug-family ATP-binding protein/permease
CD630_03540	4.469	0.153	C	T	intron	NA	Pseudogene: Fragment of putative membrane protein
CD630_04840	4.469	0.151	C	T	missense	356	ABC-type transport system, ATP-binding protein
CD630_07810	3.842	0.150	C	A	missense	819	Putative penicillin-binding protein

Table 3.1: Individual locus GWAS results: SNPs

There are 10 SNPs associated with toxin activity. *P*-values are reported as -ln(FDR corrected *P*-value). Nucleotide position refers to position within the gene where applicable. Gene annotations were taken from the CD630 reference genome.

Accessory gene	P-value	ε	Annotation
group_134	4.469	0.171	Hypothetical protein
sttH_1	4.469	0.168	NA
group_2326	4.469	0.167	Hypothetical protein
group_1064	4.469	0.165	Helix-turn-helix domain protein
group_688	4.469	0.165	Transposase from transposon Tn916
group_988	4.469	0.165	Excisionase from transposon Tn916
group_1846	4.469	0.161	Hypothetical protein
group_1566	4.469	0.161	Hypothetical protein
group_438	4.469	0.161	Hypothetical protein
mucD	4.469	0.159	Putative periplasmic serine endoprotease DegP-like precursor
group_1142	4.469	0.157	LytTr DNA-binding domain protein
group_1989	4.469	0.156	Putative periplasmic serine endoprotease DegP-like precursor
group_1376	4.469	0.154	Putative transposase DNA-binding domain protein
group_1666	4.469	0.154	Putative licABCH operon regulator
group_1663	4.097	0.194	Putative isomerase YddE
group_4126	3.842	0.185	Hypothetical protein
group_1647	3.842	0.174	Penicillinase repressor
group_1803	3.842	0.167	Hypothetical protein
group_823	3.842	0.163	Hypothetical protein
group_2327	3.842	0.154	Hypothetical protein
group_2712	3.667	0.159	Hypothetical protein
group_1696	3.602	0.163	Helix-turn-helix domain protein

Table 3.2: Individual locus GWAS results: Accessory genes

There are 22 accessory genes associated with toxin activity. Two accessory genes, *mucD* and *sttH*, were assigned known gene names while the remaining accessory genes were given temporary names and assigned annotations by prokka. P-values are reported as -ln(FDR corrected P-value).

3.4.2 Grouped locus GWAS identifies variants associated with toxin activity

We repeated the GWAS but grouped loci together by gene or intergenic region using the post-ancestral reconstruction method in hogwash[70]. This second GWAS was performed to generate hypotheses about as yet unknown associations between genes and intergenic regions with toxin activity. The grouping approach, a type of burden testing, increases power by reducing multiple testing penalties and can find patterns of association lost when penetrance is weak among individual loci. We included only SNPs and indels as sources of genomic variation. A False Discovery Rate of 0.1% was applied to the resulting P-values (Figure 3.3). The top 10 associations are shown in Table 3.3.

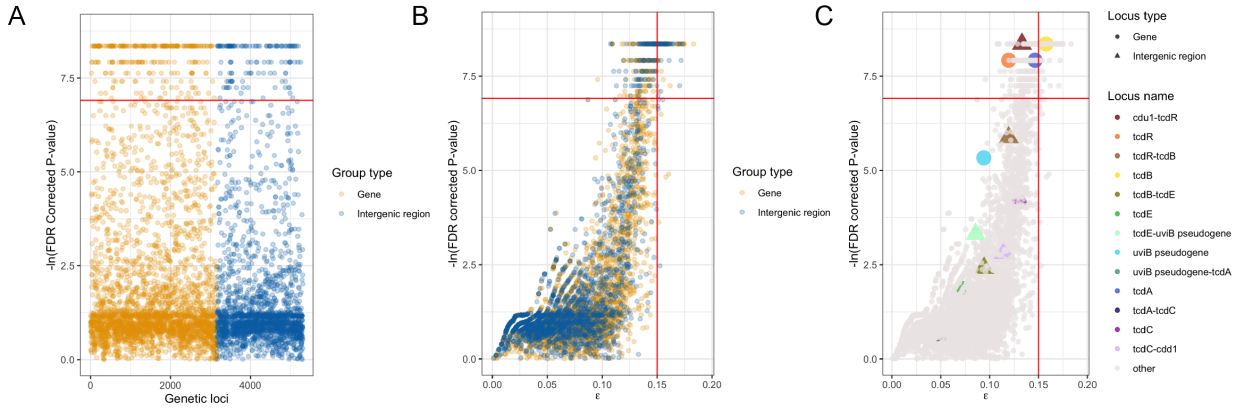

Figure 3.3: Grouped locus GWAS identifies significant associations between both genes and intergenic regions and toxin activity

A) The Manhattan plot of all grouped loci assigned a P-value in hogwash. Loci are either genes (maize; N=3,160) or intergenic regions (blue; N=2,149). B) The loci from A are visualized in a dot plot of P-value vs ε. The red horizontal line indicates a False Discovery rate of 0.1%. The red vertical line separates low vs high convergence. C) The data from B are plotted again, but the PaLoc results are highlighted.

PaLoc variants We expected that PaLoc genes and intergenic regions would be significantly associated with toxin activity and we observed both in the pool of significant results.

Locus tag	P-value	ε	Annotation
CD630_25210	8.352	0.183	leuS: leucyl-tRNA synthetase
CD630_02620	8.352	0.175	flhB: bifunctional flagellar biosynthesis protein FliR/FlhB
CD630_17690	8.352	0.175	Putative CoA enzyme activase
CD630_10160	8.352	0.174	Transcriptional regulator, MerR family
CD630_35570	8.352	0.173	Putative peptidase
CD630_16580	8.352	0.173	gcvPB: glycine decarboxylase
CD630_01820-CD630_01830	8.352	0.173	Hypothetical protein - putative cell wall hydrolase
CD630_35840	8.352	0.173	ABC-type transport system, permease
CD630_17010	8.352	0.173	recQ: ATP-dependent DNA helicase
CD630_14170-CD630_14180	8.352	0.171	Putative ATP-binding protein - putative drug/sodium antiporter, MATE family

Table 3.3: Grouped locus GWAS results: Top 10 results

There are 112 grouped loci with both significant P-value and high ε values. This table shows the top 10 loci when ranked by P-value and then ε.

Four of the 396 grouped loci significantly associated with toxin activity occur in the PaLoc: *tcdA*, *tcdB*, *tcdR*, and the *cdu1-tcdR* intergenic region (Figures 3.2 and 3.3C). To investigate the plausibility particularly of the *cdu1-tcdR* association we checked for the presence of variants within the two *tcdR* promoters. We observed that of the 35 variants that were included in the *cdu1-tcdR* intergenic region 5 SNPs fall into the *tcdR* promoters which suggests a possible functional role of these loci in modulating *tcdR* transcription and therefore affecting *tcdA* and *tcdB* regulation (Figure 3.2).

Flagellar variants While no individual locus within a flagellar gene associated with toxin activity, in the grouped locus GWAS flagellar genes were frequently associated. Fourteen of the 396 grouped loci significantly associated with toxin activity were flagellar genes or intergenic regions adjacent to flagellar genes. This is an approximately threefold enrichment compared to the number of flagellar regions that are be expected by chance (P=0.0001; median=4; range=0-13). The second most strongly associated grouped locus was *flhB/fliR*; this finding is consistent with a report that a *flhB/fliR* knockout mutant in CD630 had significantly reduced toxin activity in an assay of cell rounding (Table 3.3)[85]. Both the enrichment of flagellar genes in the significant results and the particular relationship to *flhB/fliR* are striking, but perhaps not surprising given known co-regulation that occurs

between flagellar and toxin systems in *C. difficile* that is mediated in part by SigD, a sigma factor that binds to a *tcdR* promoter region and positively regulates *tcdR*[86, 87]. Note that no tested variants occurred within the SigD promoter sequence upstream of *tcdR*.

Novel associations Within the highly ranked grouped results is an ABC transporter, CD630_35840, which could potentially contribute to toxin secretion. Future work will evaluate the 112 strongly associated grouped loci for their plausibility in contributing to toxin transcriptional initiation, transcriptional regulation or secretion.

3.5 Discussion

These preliminary findings demonstrate the ability of bGWAS to track key genetic events that modulate toxin activity. In addition to cis-regulatory variants within the PaLoc, there were numerous strong associations in trans that indicate a complex regulatory network mediating toxin activity. This regulatory network includes SigD, Agr, and Hfq which all positively regulate both flagella and toxin genes while Spo0A, SigH, and RstA each negatively regulate both gene types[87]. The grouped GWAS results are consistent with previous reports of the co-regulation of toxin flagellar genes, either through the effect of one group of genes on the other or through third-party mediators. Flagella and toxin genes may be under strong co-regulation to allow the bacterium to quickly evade the host immune system as products of both activate the host immune system; flagellin is recognized by Toll-like receptor 5 in humans and toxins can activate inflammasomes.

3.5.1 Future directions

We have access to additional *in vitro* phenotype data for the same set of *C. difficile* isolates and can perform similar GWAS using these data. The phenotypes include germination effi-

ciency, maximal growth rate, absolute spore production and spore viability[65, 88]. Studying additional *in vitro* phenotypes with this approach could generate lists of novel genomic loci for investigation. Given a known inverse relationships between sporulation and toxin genes, particularly evidenced by spo0A which is a regulator that controls both toxicity and sporulation, it may be fruitful to look for overlap between toxin and sporulation GWAS results to prioritize loci for further characterization in functional validation assays.

While many regulators of toxin activity are well described, mechanisms by which individual variants affect toxin activity could be better understood through experimentation. In particular, there are various assays to investigate the functions of each of the four toxin domains. We propose that the significant variants identified by this GWAS in *tcdB* be tested for their impact on their respective protein domains. Our genomic investigation into the *tcdB* variants revealed that there are many frameshift mutations, of which one was significantly associated with toxin activity. We found that this mutation, at residue 516, is a compensatory frameshift that restores *tcdB* back into reading frame one after a previous deletion causes a frameshift at residue 337. Additionally, we found that most of the RT027 samples have a series of frameshift mutations that move the C-terminal region of the protein into reading frame two. We are unaware of any previous work investigating the functional impact of this RT027 frameshift on toxin structure or activity.

3.6 Acknowledgements

We thank Ali Pirani for bioinformatics support. We thank Phil Hanna and Vince Young for their helpful suggestions. We thank Borden Lacy for her insights into the structural implications of the *tcdB* variants.

Chapter 4

Genetic Determinants of Trehalose Utilization Are Not Associated With Severe *Clostridium difficile* Infection Outcome

4.1 Preamble

The previous two chapters have introduced and highlighted the utility of bGWAS in identifying associations in synthetic and *in vitro* bacterial datasets. In this chapter we take a different approach to association testing because the phenotype in question, clinical infection outcome, is deeply influenced by host factors.

4.2 Introduction

C. difficile infection (CDI) is a health care–associated infection that can result in a range of patient outcomes. Of greatest concern is the development of severe disease outcome, defined as intensive care unit admission, intra-abdominal surgery (such as colectomy), and/or death attributable to the infection[89]. Specific patient factors such as age, antibiotic use, and female gender have been associated with severe infection outcome. The genetic background of the infecting *C. difficile* isolate may also influence clinical outcome. Prior studies have reported an increased risk of severe infection outcome for ribotypes RT027 and RT078[90, 91]. One recently proposed mechanism for increased virulence of specific *C. difficile* lineages

is the differential capacity to utilize the dietary disaccharide trehalose[12]. These authors observed enhanced trehalose utilization in RT027 strains, identified an additional trehalose operon in RT078, and demonstrated increased virulence of an RT027 strain in a mouse model of infection when physiologically relevant quantities of trehalose were given. This report also noted the coincidence between the introduction of trehalose as a food additive in 2000 and the global emergence of RT027 and RT078 shortly thereafter.

Subsequently, Eyre *et al.* examined clinical *C. difficile* isolates for the ability to use trehalose, noting that variants conferring improved trehalose utilization were not confined to successful epidemic lineages. Moreover, Eyre *et al.* found no evidence of association between a trehalose utilization variant and 30-day all-cause mortality in RT015, a ribotype in which enhanced trehalose utilization is variably present[92]. Here, we set out to more comprehensively evaluate the potential contribution of trehalose utilization to clinical outcomes by quantifying the independent contribution of trehalose utilization variants to infection severity across ribotypes, when controlling for all clinical factors independently associated with risk for severe infection outcome.

4.3 Methods

4.3.1 Study population

All subjects signed written consent to participate in this study. This study was approved by the University of Michigan Institutional Review Board (Study HUM33286). We considered a cohort of 1144 cases of CDI from hospitalized adults diagnosed with CDI between October 2010 and January 2013 at the University of Michigan Health System[90]. The following predictors of severe outcome were noted in the window 24–48 hours post–CDI diagnosis: age, gender, metastatic cancer, concurrent antibiotic use, systolic blood pressure, creatinine, bilirubin, and white blood cell count. Of the 981 unique patients, 898 had complete clinical

information. CDI was classified as severe outcome if any of the following outcomes attributable to CDI occurred within 30 days of diagnosis: admission to an intensive care unit, intra-abdominal surgery, or death[89]. Ribotyping was performed[93], and 137 ribotypes were identified (Supplementary Table 1). Cases were diagnosed as health care–associated CDI if symptoms developed >48 hours after admission[90].

4.3.2 Data analysis

Summary statistics, matching, modeling, and imputation were conducted in R, version 3.5.0[50]. The R code is available at: github.com/katiesaund/clinical_cdifficile_trehalose_variants.

4.3.3 Severe outcome risk score matching

Unique subjects with complete clinical information (n=898) were assigned a severe outcome risk score based on a logistic regression model with severe CDI outcome as the response variable and the following predictors: age (years), female gender, metastatic cancer, concurrent antibiotic use, systolic blood pressure (mm Hg), creatinine (>1.5 mg/dL), bilirubin (>1.2 mg/dL), and white blood cell count (cells/μL)[90]. Where possible, isolates were sorted into strata, each with exactly 1 case (severe CDI outcome) and at least 1 and up to 4 controls (nonsevere CDI outcome), with control scores within $\pm$0.10 of the case score. The matching process identified 323 CDI case isolates, of which only 247 *C. difficile* isolates were successfully grown, isolated, and sequenced. Due to growth or sequencing failures, only 235 isolates were from 59 complete strata (strata with at least 1 control and 1 case). Each stratum contained 1 case (n=59, severe CDI outcome) and up to 4 controls (n=176, nonsevere CDI outcome; total n=235). The median number of controls per case (range) was 3 (1–4). There were 61 RT027 isolates (19/59 cases; 46/176 controls), 5 RT078 isolates (0/59 cases; 5/176

controls), and only 1 RT015 (1/59 cases; 0/176 controls) in this matched data set.

4.3.4 Conditional logistic regression model for matched samples

Logistic regressions were modeled with severe CDI outcome as the response variable, conditioned on strata, with a trehalose variant as the predictor. Bonferroni-corrected P-values were reported.

4.3.5 Genomic analysis

Genomic DNA extracted with the MoBio PowerMag Microbial DNA Isolation Kit (Qiagen) from *C. difficile* isolates (n=247) was prepared for sequencing using the Illumina Nextera DNA Flex Library Preparation Kit. Sequencing was performed on either an Illumina HiSeq 4000 System at the University of Michigan Advanced Genomics Core or on an Illumina MiSeq System at the University of Michigan Microbial Systems Molecular Biology Laboratories. Quality of reads was assessed with FastQC[66]. Adapter sequences and low-quality bases were removed with Trimmomatic[67]. Details on sequenced strains are available in Supplementary Table 1. Sequence data are available under Bioproject PRJNA561087. For variant calling details, see the Supplementary Methods. Pangenome analysis was performed with roary[68], and the insertion putatively conferring enhanced trehalose utilization was inferred based on the presence of 4 genes: *treA2*, *ptsT*, *treX*, and *treR2* (Supplementary Data).

4.4 Results

Given the limited clinical data regarding the potential contribution of trehalose utilization variants to CDI severity, we set out to test for an association between trehalose utilization across all strains of *C. difficile*–causing infection while comprehensively controlling for patient factors associated with severe outcome. Patient factors associated with CDI severity in our

clinical cohort (n=1,144) were age, female gender, metastatic cancer, concurrent antibiotic use, systolic blood pressure, creatinine, bilirubin, and white blood cell count[90]. To quantify the independent contribution of trehalose utilization variants to severe patient outcomes, we implemented a retrospective, matched case–control study to control for these patient factors. Each *C. difficile* episode was assigned a severe outcome risk score utilizing variables available around the time of diagnosis, which is the patient's predicted probability of having severe CDI outcome, based on a logistic regression model of CDI built from the 8 patient factors. Unique patients were grouped into strata based on their risk score.

We identified the presence of reported trehalose utilization variants: 2 amino acid substitutions in the transcriptional repressor, *treR* (TreR C171S and TreR L172I), and a set of 4 accessory genes, called the 4-gene trehalose insertion, that contain an additional phosphotrehalase enzyme and transcriptional regulator[12, 94]. Consistent with previous reports, we found TreR C171S in all 4 RT017 isolates and 3 closely related ribotypes (Figure 4.1). Similarly, we found TreR L172I in all 61 RT027 isolates and in 7 closely related ribotypes. However, in contrast to Collins *et al.*, who found the 4-gene trehalose insertion only in RT078, we identified the insertion in 25 isolates that were broadly distributed across the phylogeny[12, 92].

To investigate if trehalose utilization variants were individually associated with severe CDI outcome, we performed unadjusted logistic regressions conditioned on severe outcome risk strata (Supplementary Table 2). The presence of the TreR L172I trehalose utilization variant was associated with an increased, but not statistically significant, odds of severe CDI outcome (OR, 1.60; 95% CI, 0.84–3.05; P=0.52). The presence of the 4-gene trehalose insertion was associated with decreased, but not statistically significant, odds of a severe CDI outcome (OR, 0.35; 95% CI, 0.09–1.32; P=0.52). None of the other 4 amino acid substitutions observed in TreR were significantly associated with severe CDI outcome (Supplementary Table 2). Due to low prevalence (n=7), we did not test for an association between

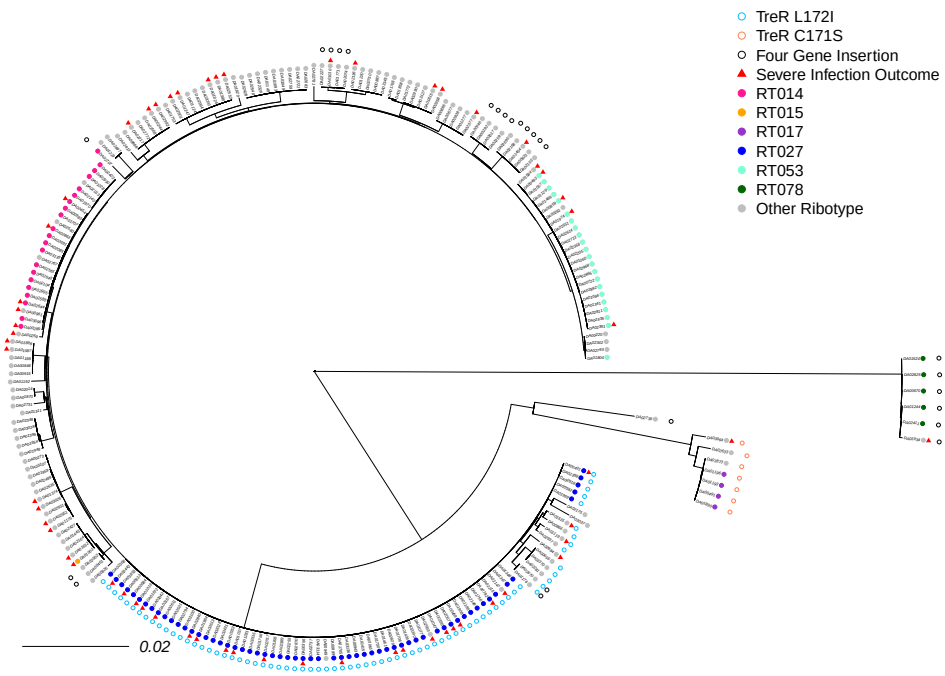

Figure 4.1: Comparative analysis of trehalose genetic variants in *C. difficile*.

A maximum-likelihood phylogeny was constructed (n=235) using variants identified in the genome (scale in mutations per site in the whole genome). Trehalose variants are either point mutations in *treR* (C171S empty orange circle, L172I empty blue circle) or the presence of the 4-gene trehalose insertion (empty black circle). A severe infection outcome is indicated by a red triangle. Ribotype is indicated by the filled circles.

the presence of the TreR C171S trehalose utilization variant and severe CDI outcome.

To evaluate the overall effect of the 3 previously described trehalose utilization variants on CDI outcome, we combined the 3 variants into a single predictor of severe CDI outcome and performed a logistic regression conditioned on severe outcome risk strata. When combined, the presence of any of the 3 trehalose utilization variants does not significantly affect the odds of severe CDI outcome (OR, 1.10; 95% CI, 0.61–1.99; P=0.86). To evaluate whether these results translated to the original unmatched cohort, we also tested for associations between trehalose utilization variants and severe infection outcome using severe outcome risk score

as a covariate rather than a matching criterion. To infer trehalose utilization genotypes in the larger cohort for which genomic data were unavailable, we performed an imputation where the presence of trehalose utilization variants (TreR C171S, TreR L172I, and the 4-gene trehalose insertion) was assumed to have the same distribution across ribotypes as was observed in the matched cohort (Supplementary Methods). We then performed a logistic regression modeling severe outcome with presence of any of the 3 trehalose utilization variants as a predictor and severe outcome risk score as a covariate. When combined, the presence of any of the 3 trehalose utilization variants did not significantly affect the odds of severe CDI outcome (median OR [range], 0.99 [0.87–0.99]; median P [range] =0.97 [0.64–0.99]; n=586). Note that using this logistic regression framework we were powered to recapitulate the original finding in Rao *et al.*, that RT027 is significantly associated with severe outcome when not conditioning on risk score (OR, 2.22; 95% CI, 1.20–4.04; P=0.01).

4.5 Discussion

The ability to predict the clinical outcome of *C. difficile* infection based on patient and pathogen characteristics could help guide therapy for this important nosocomial infection. An enhanced ability to utilize trehalose was shown to be associated with increased virulence in mice, prompting us to evaluate this relationship using a case–control study performed with a large clinical data set. Our controlled analysis, in patients from a diverse set of ribotypes, failed to detect a significant association between any of the previously described trehalose utilization variants and severe CDI outcome. Although it is possible that an effect may be observed in a larger cohort, this is not supported by the odds ratio being close to 1 both in our matched analysis and when imputing to the larger cohort. We also note that while the TreR L172I variant alone has an odds ratio >1, the tight association between this variant and RT027 leave unclear its independent contribution to worse outcomes in patients

with RT027 infection. Our results are consistent with Eyre *et al.*'s study of patients infected with RT015, which observed a lack of association between the 4-gene trehalose insertion and 30-day all-cause mortality. We also confirmed Eyre *et al.*'s observation that the 4-gene trehalose insertion is not exclusive to RT078, but rather is found throughout the *C. difficile* phylogeny[92].

The lack of association between trehalose utilization variants and severe CDI outcome in hospitalized patients emphasizes the need to incorporate clinical results earlier into the genetic variant discovery pipeline. Indeed, a more relevant hypothesis-generating process could begin by identifying genetic loci of interest first through comparative genomic analysis of clinical isolates, followed by validation in lab strains with molecular and animal studies. A critical component of this discovery work is to control for both patient factors and strain background, as both may confound analysis, which we did above in separate analyses using 2 distinct modeling strategies. Given the association between ribotype, recent acquisition, and CDI outcome, controlling for patient factors is critical for identifying the genetic variants associated with severe CDI outcome[95].

Although these clinical data cannot rule out a potential role of trehalose utilization variants in the success of hypervirulent *C. difficile* lineages, they do emphasize the difference between human infection and murine models. More broadly, generating hypotheses in controlled laboratory systems and having them generalize to genetically and clinically heterogeneous human populations will limit insights to only the most penetrant phenotypes. With increasing availability of electronic health record and microbial genomic data, it is now becoming feasible to flip the script and generate hypotheses through analysis of human data, which can subsequently be tested in appropriate animal and in vitro systems with a priori knowledge of relevance to human disease.

4.6 Potential conflicts of interest

K.S. reports grants from National Institutes of Health during the conduct of the study; K.R. reports personal fees from Bio-K+, Inc., outside the submitted work; V.B.Y. reports grants from National Institutes of Health, during the conduct of the study; non-financial support from Vedanta Biosciences, non-financial support from Bio-K+ International, non-financial support from Pantheryx, outside the submitted work; E.S.S. reports grants from National Institutes of Health, during the conduct of the study.

4.7 Supplement

4.7.1 Methods

Variant calling

Variants were identified by mapping filtered reads to the CD630 reference genome (GenBank accession number AM180355.1) using bwa[71], removing polymerase chain reaction duplicates with Picard[72], and calling variants with SAMtools and bcftools[73]. Variants were filtered from raw results using GATK's VariantFiltration (QUAL, >100; MQ, >50; >=10 reads supporting variant; and FQ, <0.025)[74]. Of the single nucleotide polymorphisms identified in *treR*, all were biallelic and only nonsynonymous variants were analyzed.

Phylogenetic analysis

Consensus files generated during variant calling were recombination filtered using Gubbins[18]. The alleles at each position that passed filtering were concatenated to generate a non-core variant alignment relative to the CD630 reference genome. Alleles that did not pass filtering were considered unknown (denoted as N in the alignment). Variant positions in the alignment were used to reconstruct a maximum likelihood phylogeny with IQ-TREE v1.5.5

using ultrafast bootstrap with 1,000 replicates[75, 76]. ModelFinder limited to ascertainment bias-corrected models was used to identify the best nucleotide substitution model[77]. The tree was rooted on a cryptic clade *C. difficile* isolate as the outgroup (Run: ERR232398, Sample: C00007672, Bioproject: PRJEB1483).

Imputation of trehalose utilization variants for non-sequenced isolates

To assign trehalose utilization variants to isolates that were either not sequenced or not matched the set of 1,144 isolates was subset to include only unique individuals and to exclude (1) isolates with a ribotype classified as unique, other or missing and (2) isolates whose ribotype was not found in the sequenced, case-control study. Imputation of trehalose utilization variants was performed on N=351 (6 cases; 345 controls) non-sequenced isolates based on N=165 (43 cases; 122 controls) sequenced isolates. In the imputation steps described below, the process was applied to cases and controls separately (sequenced cases were used to impute variant presence in only non-sequenced cases and sequenced controls were used to impute variants presence in only non-sequenced controls). For each ribotype found in the sequenced, case-control isolates the percentage of isolates that had any trehalose utilization variant (C171S, L172I, and/or four gene trehalose insertion) was calculated. Each non-sequenced isolate was assigned imputed trehalose utilization variants as a function of the proportion of trehalose utilization variants within the sequenced members of the ribotype. Assignment of variants was performed with 1,000 replicates.

Statistical analysis on imputed isolates combined with sequenced isolates

Statistical analyses were performed on the matched isolates (N=235) combined with the imputed isolates (N=351; total N=586). Logistic regressions were modeled with severe CDI outcome as the response variable, severe outcome risk score as a continuous covariate, and a trehalose variant as the predictor. Each regression was performed once per imputation

(1,000 total tests) and summary statistics were reported.

Data analysis

Data analysis in R was performed with following packages: ape 5.3[49], seqinr 3.4-5[96], survival 2.44-1.1[97], and tidyverse 1.2.1[53].

Files and tables

Supplementary files and tables may be accessed at https://academic.oup.com/ofid/article/7/1/ofz548/5696590.

Chapter 5

Discussion

5.1 Major thesis contributions

The previous chapters focus on the detection of genotype-phenotype associations with data derived from *in silico*, *in vitro*, and clinical settings. Chapter one described the implementation of three bGWAS methods in hogwash, a software package, and demonstrated the package's utility on *in silico* data. This chapter introduced two metrics, N and ε, to quantify the relationship between phenotype convergence and genotype convergence on a given phylogenetic tree. Finally, it generated an *in silico* dataset of phenotypes and genotypes with a range of convergence and phylogenetic signals to be used for future convergence-based GWAS benchmarking studies. Chapter two validated hogwash's continuous algorithm on *in vitro* toxin activity data from a clinical cohort of *C. difficile* isolates. The GWAS results found significant associations between toxin activity and variants in both flagellar genes and the PaLoc. The strong associations observed between *tcdB* variants and toxin activity particularly suggest that hogwash is a robust method for genome-wide association studies. Chapter three applied an alternative genotype-phenotype association method that controlled for patient factors rather than tree structure. This method found no statistically significant association between three trehalose utilization variants proposed to contribute to poor *C. difficile* infection outcomes.

5.1.1 Implementation of existing and novel convergence-based bGWAS methods

As whole genome bacterial sequences are increasingly available due to plummeting costs and improved bioinformatic pipelines there is a demand for more bacteria-specific software. In response, we developed a bGWAS software package that leverages bacterial whole genome sequence data and incorporates aspects of bacterial evolution. Hogwash implements three bGWAS methods, two burden testing approaches, and describes a novel index, ε, that quantifies the relationship between phenotype convergence and genotype convergence. The burden testing methods provide two choices for combining genomic variation into biologically meaningful groups that may potentially surmount the stringent requirements for convergence in the individual locus test. When users combine the ε metric with the P-value they are more likely to identify associations of highest interest as the two values together contextualize the penetrance and non-randomness of genotype-phenotype associations. The investigation of genotype-phenotype associations in the *in silico* dataset demonstrated several features of hogwash. When comparing the individual locus test to a grouped locus test, the grouped results revealed associations that were masked in the individual locus test. The novel findings of the grouped test seem to be due to lower penetrance of genotype convergence at the individual variant level that is overcome by aggregation of multiple genotypes together. Additionally, the *in silico* dataset reveals that hogwash is better able to identify associations between genotypes and phenotypes when the phenotypes are modeled better by white noise than by Brownian motion; this performance difference is likely because of the increased convergence in white noise modeled phenotypes. The *in silico* modeled phenotypes, genotype, and phylogenetic tree dataset may serve as a resource for future work benchmarking various convergence-based bGWAS methods. In particular, this dataset contains phenotypes and genotypes with a range of convergence and phylogenetic signals that can highlight the

conditions wherein each method has the highest ability to discriminate genotype-phenotype associations.

5.1.2 bGWAS on *in vitro* toxin activity underscores utility of hogwash method and suggests co-regulation of toxin and flagellar proteins

We applied hogwash to *in vitro* toxin activity, a clinically relevant phenotype, in a set of clinical *C. difficile* isolates. Toxins are encoded by *tcdA* and *tcdB* and their activity is known to be controlled by the regulator *tcdR*. We observed significant associations between variants in the PaLoc and toxin activity. Individual variants within *tcdB* were significantly associated with toxin activity as were variants grouped into *tcdR*, *tcdA*, *tcdB*, and in the *cdu1-tcdR* intergenic region. These associations suggest the plausibility of hogwash as a reliable bGWAS method. Additionally, flagellar proteins were overrepresented in the group-based hogwash results suggesting a potential evolutionary relationship between toxin and flagellar genes as well as their coregulation. Finally, the bGWAS results point to specific variants and genomic loci to investigate in studies to uncover novel mechanisms of toxin regulation in *C. difficile*.

5.1.3 A case-control study found no statistically significant association between trehalose utilization variants present in *C. difficile* strains and the development of severe infection outcome

Recent reports suggested that trehalose utilization variants reported to be present in the epidemic strains RT027 and RT078 may contribute to these strains' success and virulence.

We evaluated the contribution of three trehalose utilization variants (TreR L172I, TreR C171S, and a four-gene insertion) independently and in aggregate on severe infection outcome in a cohort of clinical patients. This analysis was performed while controlling for eight patient factors independently associated with increased risk of severe infection outcome in order to tease out the contribution specifically of the trehalose utilization variant. The conditional logistic regression suggests no significant association between the presence of the trehalose utilization variants, either independently or in aggregate, and severe infection outcome. We observed that the four-gene insertion was not limited to the epidemic RT078 strain but found in a diverse set of isolates. This investigation was prompted by an *in vitro* screen of carbon sources on *in vitro C. difficile* growth. While these *in vitro* studies have uncovered a previously unknown relationship between trehalose utilization in certain *C. difficile* strains and murine infection severity, they have yet to translate to human clinical outcomes. Perhaps a more relevant pipeline for identifying genomic variants in *C. difficile* that could better stratify and predict infection outcomes would be to identify those genomic variants commonly found in only patients with severe infection outcomes and then follow up with *in vitro* and *in vivo* experiments to determine if these variants are plausibly causal of such outcomes and, if so, reveal the mechanisms of increased virulence. In particular, this approach could be highly specialized to account for patient factors and therefore identify just those genomic factors most likely to be independently contributing to severe infection outcome.

5.2 The future of bGWAS

This thesis developed and tested a bGWAS software tool on toxin activity and applied a conditional logistic regression to test the independent contribution of trehalose utilization variants on clinical infection outcomes. We hope that future work will expand upon both

approaches by incorporating methods to identify causal variants from bGWAS results, contrasting available convergence-based bGWAS software, integrating rich patient data into regression approaches, and including more complex genotypic variation in association testing. Such potential research directions could improve the utility of association results for clinical applications.

5.2.1 Identifying plausible causal variants

bGWAS studies have repeatedly identified genomic variants significantly associated with phenotypes and some have demonstrated causal relationships through functional validation experiments. To our knowledge, no bGWASs have used methods common in human GWASs to refine lists of associated genomic variants associated to those most likely to be causal, such as fine-mapping and colocalization studies. While certain aspects of bacterial genomes may limit the use of these approaches in all cases, we propose the adoption in certain scenarios.

Fine-mapping Fine-mapping is a family of methods that attempt to identify which of the significant SNPs from a GWAS causes a peak in association at a specific locus and is applied when many variants are significantly associated with the phenotype of interest[98]. Fine-mapping is a challenging process because variants are not inherited independently and there are an unknown number of truly associated variants. Current approaches often define a credible set which is a set of significantly associated variants that contain a causal variant with a certain probability[98]. The application of fine-mapping in bGWAS could prioritize the order of and ideally reduce the number of *in vitro* experiments required to validate necessity and sufficiency of a variant's effect on phenotype by winnowing the pool of potential variants to just those within a sufficiently small credible set. Fine-mapping has not caught on in bGWAS, perhaps because of the focus of bGWAS on those bacterial species with high linkage disequilibrium. Fine-mapping is best applied to bacterial species with low linkage

disequilibrium such as *Streptococcus pneumoniae*[36].

Colocalization A genetic locus may be associated with two or more phenotypes, such as a clinical outcome, gene expression, and/or protein expression. Genetic colocalization between a clinical outcome and an intermediate phenotype can help narrow down the genes or pathways involved in mediating the clinical outcome and reduce the pool of variants included in downstream functional testing. Summary statistics of GWAS can be combined with gene expression data to detect associations between genes and phenotypes[99, 100, 101] or with protein level data to detect association between proteins and phenotypes[102, 103, 104]. Transcriptome-wide association studies and proteome-wide association studies have yet to be applied to bacteria, but these approaches could help move the field past associations with simple phenotypes such as antibiotic resistance into more complex, quantitative traits. In particular, transcriptome and proteome data may overcome the barrier imposed by linkage disequilibrium to interpretation of bGWAS results in those species with particularly high levels of linkage disequilibrium.

Given the particular characteristics of the dataset in question and the availability of transcriptomic or proteomic data, either fine-mapping and/or colocalization may be helpful tools for refining lists of variants associated with complex bacterial phenotypes to just those most likely to be causal. Smaller sets of variants for functional testing may reduce the burden posed by these follow on studies and increase rates of experimental successes.

5.2.2 Benchmarking convergence-based bGWAS methods

There are several convergence-based bGWAS methods available (treeWAS[37], hogwash) but their relative performance has yet to be compared. Regression based bGWAS methods have been benchmarked for their performance on datasets with a range of linkage disequilibrium values[57]. A thorough comparison would evaluate F1 score, recall, precision, speed, and

memory usage on various simulated datasets that include a) genomes with a range of linkage disequilibrium values, b) genotypes and phenotypes that span a large range of phylogenetic signals, c) a range of sample sizes. A major obstacle to overcome for such an undertaking is the large computational time required for even moderately sized datasets to be analyzed by both treeWAS and hogwash. An objective comparison of current convergence-based bGWAS tools will aid researchers as they select a tool most appropriate for their dataset and analysis goals.

5.2.3 Integration of host, environmental, and bacterial factors into infection risk scores

The analysis in chapter four examined the potential for associations between a set of genomic variants and severe *C. difficile* infection outcome. This approach controlled for only eight patient factors known to be associated with severe infection outcome. We propose that integrating more and richer data sources could lead to more accurate and actionable infection risk scores for important bacterial pathogens. While such infection risk scores have already been implemented, only a small number consider a wide range of data sources or big data.

Studies have proposed risk scores that predict the likelihood, severity, or recurrence of *C. difficile* infection. Model predictors may include host and environmental factors derived from electronic health records[105, 106], *C. difficile* ribotype information[90], or pre-infection microbiome features[107]. Rarely, models are designed to be used in real-time to aid in clinical decision making[106]. We propose moving towards better models of *C. difficile* infection likelihood, severity, and recurrence through the integration of electronic health records, microbiome data prior to *C. difficile* infection onset, and the genomic features of *C. difficile* strains isolated from diagnosed patients. Patients identified to be at high risk for *C. difficile* infection by such models can be targeted for infection prevention efforts such as antibiotic

use review and educational efforts[108]. More comprehensive approaches may lead to more accurate models and therefore direct healthcare resources efficiently.

5.2.4 Innovation in genotypic measurements for bGWAS

Genomic variants used in bGWAS are typically SNPs, indels, genes, pathways, k-mers, or unitigs. Many studies simplify the genomic information input to the association test by removing more nuanced genetic information such as plasmids, large structural variants, or copy number variants. Copy number variants can be used to weight genomic variants in regression based bGWAS approaches. The increasing accessibility and application of long read sequencing technology from companies such as Oxford Nanopore Technologies and Pacific Biosciences to bacterial species may increase the ability of the field to examine more nuanced genomic information. For example, structural variants can be more accurately resolved from long sequencing reads than from short sequencing reads [109].

The genomic loci found to associate significantly with a phenotype in a bGWAS can often be hard to interpret as loci may fall within intergenic regions or unannotated, hypothetical proteins. While model organisms such as *Escherichia coli* are highly annotated, even clinically relevant species such as *C. difficile* have only $\sim 50\%$ of the proteins in the main reference genome annotated with known functions[110]. More comprehensive annotation projects are necessary to interpret bGWAS hits and to prioritize genomic loci for functional validation studies.

Genome sequences for bGWAS are collected through a labor intensive process of isolating individual clonal strains of bacteria and then growing each strain to sufficient quantity for genomic DNA extraction. In the clinical studies used in chapters three and four a single strain of *C. difficile* was isolated from each infected patient, but this is a simplified representation of the infection as *C. difficile* patients are known to be infected with two or more ribotypes in as many as 16% of infections [111]. The choice to isolate only a single colony was in

part due to the logistical and budgetary burden of isolating multiple strains from a single patient sample. A population sequencing approach, wherein bacteria of a certain species are enriched through culturing and then sequenced as a mixed population, could capture all of the genomic variation occurring in a multi-strain infection. Additionally, a population sequencing approach would remove the colony picking step in protocols, thus saving time and labor costs. Population sequencing not only results in genomic information for multiple strains, whether or not a patient has a multi-strain infection itself is information that can be input to risk score algorithms as multi-strain infections have been shown to be correlated with recurrent infection[112]. Population sequencing can retain some of the nuanced genomic information previously discussed if tools such as chromosome conformation capture (3C) and long read sequencing are used in place of or in addition to short read sequencing. 3C is a method to capture the interactions of DNA strands in 3-D space, and if used in cells prior to lysis can link plasmids with the specific genomes in which they co-occur, thus surmounting a potential loss of information from population sequencing [113].

Increasing the annotation of bacterial genomes and capturing more genomic variation in the form of copy number variants, structural variants, plasmids, and bacterial populations will improve the ability of bGWAS to find the most robust and biologically interpretable associations to phenotypes.

5.3 Conclusion

This thesis work was motivated by a desire to improve outcomes for patients infected with *C. difficile* and focused on the potential for *C. difficile* genomic variation to affect infections. We hope that one day patients infected with *C. difficile* strains will be tested for the still-to-be-identified risky genomic variants during infection. These test results, when combined with patient information, could be used to stratify patients into different treatment regimes

in order to yield the best possible clinical outcomes. Many host, pathogen, and environmental factors may contribute to infection outcomes. The field can build to such a nuanced understanding of *C. difficile* infection and the interplay of host and pathogen by developing increasingly comprehensive models of the infection process.

In the previous chapters, hogwash was introduced and then its validity explored using *C. difficile* toxin activity data. Additionally, a specific set of variants proposed to affect infection outcome were not found to be significantly associated with severe infection outcome. We hope that by using phylogenetically aware bGWAS methods or approaches that control for patient factors we can move on to studies that attempt to identify *C. difficile* genomic variants that strongly associate with severe infection outcomes; these variants may eventually play a role in tailoring patient interventions to improve clinical outcomes.

Bibliography

[1] J G Bartlett et al. "Clindamycin-associated colitis due to a toxin-producing species of Clostridium in hamsters". In: *Journal of Infectious Diseases* 136.5 (Nov. 1977), pp. 701–705. ISSN: 00221899. DOI: 10.1093/infdis/136.5.701. URL: http://www.ncbi.nlm.nih.gov/pubmed/915343.

[2] CDC. *Antibiotic resistance threats in the United States*. Tech. rep. Atlanta, GA: U.S. Department of Health and Human Services, CDC, 2019, pp. 1–139. URL: https://www.cdc.gov/drugresistance/biggest%7B%5C_%7Dthreats.html.

[3] Dale N. Gerding and Vincent B. Young. "Clostridium difficile Infection". In: *Mandell, Douglas, and Bennett's Principles and Practice of Infectious Diseases*. Ed. by John E. Bennett, Raphael Dolin, and Martin J. Blaser. Eighth Edi. Elsevier Inc., 2015. Chap. 245, pp. 2744–2756. ISBN: 978-1-4557-4801-3. DOI: https://doi.org/10.1016/C2012-1-00075-6. URL: https://doi.org/10.1016/C2012-1-00075-6.

[4] L. Clifford McDonald et al. "An epidemic, toxin gene-variant strain of Clostridium difficile". In: *New England Journal of Medicine* 353.23 (2005), pp. 2433–2441. ISSN: 00284793. DOI: 10.1056/NEJMoa051590. URL: http://www.nejm.org/doi/pdf/10.1056/NEJMoa051590.

[5] Vivian G Loo et al. "A Predominantly Clonal Multi-Institutional Outbreak of Clostridium difficile –Associated Diarrhea with High Morbidity and Mortality". In: *New Eng-*

land Journal of Medicine 353.23 (2005), pp. 2442–9. URL: http://www.nejm.org/doi/pdf/10.1056/NEJMoa051639.

[6] Patrizia Spigaglia, Paola Mastrantonio, and Fabrizio Barbanti. "Antibiotic Resistances of Clostridium difficile". In: *Updates on Clostridium difficile in Europe*. Ed. by Paola Mastrantonio and Maja Rupnik. 8th ed. Springer, 2018. Chap. 9, pp. 137–159. DOI: 10.1007/978-3-319-72799-8_15. URL: http://www.springer.com/series/13513.

[7] Dale N Gerding et al. "Clostridium difficile binary toxin CDT". In: *Gut Microbes* 5.1 (2014), pp. 15–27. ISSN: 1949-0976. DOI: 10.4161/gmic.26854. URL: http://www.tandfonline.com/doi/abs/10.4161/gmic.26854.

[8] Marc Monot et al. "Clostridium difficile: New Insights into the Evolution of the Pathogenicity Locus". In: *Scientific Reports* 5 (2015). ISSN: 20452322. DOI: 10.1038/srep15023. URL: www.nature.com/scientificreports/.

[9] Rory N. Pruitt and D. Borden Lacy. "Toward a structural understanding of Clostridium difficile toxins A and B." In: *Frontiers in cellular and infection microbiology* 2.March (2012), p. 28. ISSN: 22352988. DOI: 10.3389/fcimb.2012.00028.

[10] Isabelle Martin-Verstraete, Johann Peltier, and Bruno Dupuy. *The regulatory networks that control Clostridium difficile toxin synthesis*. May 2016. DOI: 10.3390/toxins8050153. URL: http://www.mdpi.com/2072-6651/8/5/153.

[11] Laurent Bouillaut et al. "Integration of metabolism and virulence in Clostridium difficile". In: *Research in Microbiology* 166.4 (2015), pp. 375–383. ISSN: 17697123. DOI: 10.1016/j.resmic.2014.10.002.

[12] J Collins et al. "Dietary trehalose enhances virulence of epidemic Clostridium difficile". In: *Nature* 553.7688 (2018), pp. 291–294.

[13] Daniel R. Knight et al. "Diversity and evolution in the genome of Clostridium difficile". In: *Clinical Microbiology Reviews* 28.3 (July 2015), pp. 721–741. ISSN: 10986618. DOI: 10.1128/CMR.00127-14. URL: http://cmr.asm.org/lookup/doi/10.1128/CMR.00127-14.

[14] Daniel R Knight et al. "The Clostridioides difficile species problem: global phylogenomic analysis uncovers three ancient, toxigenic, genomospecies". In: *bioRxiv* (2020). DOI: 10.1101/2020.09.21.307223. URL: https://doi.org/10.1101/2020.09.21.307223.

[15] Kate E Dingle et al. "Evolutionary history of the Clostridium difficile pathogenicity locus". In: *Genome Biology and Evolution* 6.1 (2014), pp. 36–52. ISSN: 17596653. DOI: 10.1093/gbe/evt204.

[16] Miao He et al. "Supplement: Emergence and global spread of epidemic healthcare-associated Clostridium difficile." In: *Nature genetics* 45.1 (2013), pp. 109–13. ISSN: 1546-1718. DOI: 10.1038/ng.2478. URL: http://dx.doi.org/10.1038/ng.2478.

[17] Mohammed Sebaihia et al. "The multidrug-resistant human pathogen Clostridium difficile has a highly mobile, mosaic genome." In: *Nature genetics* 38.7 (2006), pp. 779–786. ISSN: 1061-4036. DOI: 10.1038/ng1830.

[18] Nicholas J Croucher et al. "Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins". In: *Nucleic acids research* 43.3 (2015), e15–e15.

[19] Valérie Bouchet, Heather Huot, and Richard Goldstein. "Molecular genetic basis of ribotyping." In: *Clinical Microbiology Reviews* 21.2 (Apr. 2008), pp. 262–73. DOI: 10.1128/CMR.00026-07. URL: http://www.ncbi.nlm.nih.gov/pubmed/18400796.

[20] Sandra Janezic and Maja Rupnik. "Genomic diversity of Clostridium difficile strains". In: *Research in Microbiology* 166.4 (2015), pp. 353–360. ISSN: 17697123. DOI: 10.1016/j.resmic.2015.02.002.

[21] Jane Freeman et al. "Five-year Pan-European, longitudinal surveillance of Clostridium difficile ribotype prevalence and antimicrobial resistance: the extended ClosER study". In: *European Journal of Clinical Microbiology and Infectious Diseases* 39.1 (2020), pp. 169–177. ISSN: 14354373. DOI: 10.1007/s10096-019-03708-7.

[22] J Couturier et al. "Ribotypes and New Virulent Strains Across Europe". In: *Updates on Clostridium difficile in Europe*. Ed. by Paola Mastrantonio and Maja Rupnik. 8th ed. Vol. 8. Springer, 2018. Chap. 4, pp. 45–58.

[23] Jennifer R. O'Connor, Stuart Johnson, and Dale N. Gerding. "Clostridium difficile Infection Caused by the Epidemic BI/NAP1/027 Strain". In: *Gastroenterology* 136.6 (2009), pp. 1913–1924. ISSN: 00165085. DOI: 10.1053/j.gastro.2009.02.073. URL: http://dx.doi.org/10.1053/j.gastro.2009.02.073.

[24] Susana Matamouros, Patrick England, and Bruno Dupuy. "Clostridium difficile toxin expression is inhibited by the novel regulator TcdC". In: *Molecular Microbiology* 64.5 (2007), pp. 1274–1288. ISSN: 0950382X. DOI: 10.1111/j.1365-2958.2007.05739.x.

[25] Abraham Goorhuis et al. "Emergence of Clostridium difficile infection due to a new hypervirulent strain, polymerase chain reaction ribotype 078". In: *Clinical Infectious Diseases* 47.9 (2008), pp. 1162–1170. ISSN: 10584838. DOI: 10.1086/592257.

[26] W. N. Fawley et al. "Enhanced surveillance of Clostridium difficile infection occurring outside hospital, England, 2011 to 2013". In: *Eurosurveillance* 21.29 (2016), pp. 1–10. ISSN: 15607917. DOI: 10.2807/1560-7917.ES.2016.21.29.30295.

[27] John T. Heap et al. "The ClosTron: Mutagenesis in Clostridium refined and streamlined". In: *Journal of Microbiological Methods* 80.1 (Jan. 2010), pp. 49–55. ISSN: 01677012. DOI: 10.1016/j.mimet.2009.10.018. URL: https://www.sciencedirect.com/science/article/pii/S0167701209003509?via%7B%5C%%7D3Dihub.

[28] Nigel P. Minton et al. “A roadmap for gene system development in Clostridium”. In: *Anaerobe* 41 (2016), pp. 104–112. ISSN: 10958274. DOI: 10.1016/j.anaerobe.2016.05.011. URL: http://dx.doi.org/10.1016/j.anaerobe.2016.05.011.

[29] Maha R Farhat et al. “Genomic analysis identifies targets of convergent positive selection in drug-resistant Mycobacterium tuberculosis.” In: *Nature genetics* 45.10 (2013), pp. 1183–9. ISSN: 1546-1718. DOI: 10.1038/ng.2747. arXiv: arXiv:1011.1669v3.

[30] Samuel K Sheppard et al. “Genome-wide association study identifies vitamin B 5 biosynthesis as a host specificity factor in Campylobacter”. In: *PNAS* 110.29 (2013), pp. 11923–11927.

[31] Gardar Sveinbjornsson et al. “Weighting sequence variants based on their annotation increases power of whole-genome association studies”. In: *Nature Genetics* 48.3 (Mar. 2016), pp. 314–317. ISSN: 15461718. DOI: 10.1038/ng.3507.

[32] Audrey E. Hendricks et al. “Rare Variant Analysis of Human and Rodent Obesity Genes in Individuals with Severe Childhood Obesity”. In: *Scientific Reports* 7.1 (Dec. 2017), pp. 1–14. ISSN: 20452322. DOI: 10.1038/s41598-017-03054-8.

[33] Robert A. Power, Julian Parkhill, and Tulio de Oliveira. “Microbial genome-wide association studies: lessons from human GWAS”. In: *Nature Reviews Genetics* (2016). ISSN: 1471-0056. DOI: 10.1038/nrg.2016.132.

[34] Ola Brynildsrud et al. “Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary”. In: *Genome Biology* 17 (2016). ISSN: 1474760X. DOI: 10.1186/s13059-016-1108-8.

[35] John A Lees et al. “pyseer: a comprehensive tool for microbial pangenome-wide association studies”. In: *Bioinformatics* (2018). ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty539. URL: http://pyseer.readthedocs.io.%20https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/bty539/5047751.

[36] Sarah G. Earle et al. "Identifying lineage effects when controlling for population structure improves power in bacterial association studies". In: *Nature Microbiology* 1.5 (2016), pp. 1–8. ISSN: 20585276. DOI: 10.1038/nmicrobiol.2016.41. arXiv: 1510.06863.

[37] Caitlin Collins and Xavier Didelot. "A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination". In: *PLoS Comput Biol* (2018). DOI: 10.1371/journal.pcbi.1005958.

[38] Shaun Purcell et al. "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses". In: *The American Journal of Human Genetics* 81.3 (2007), pp. 559–575. ISSN: 00029297. DOI: 10.1086/519795. arXiv: arXiv:1011.1669v3. URL: www.ajhg.org%20http://linkinghub.elsevier.com/retrieve/pii/S0002929707613524.

[39] Peter E. Chen and B. Jesse Shapiro. *The advent of genome-wide association studies for bacteria*. 2015. DOI: 10.1016/j.mib.2015.03.002.

[40] Jukka Corander et al. "Bacterial Population Genomics". In: *Handbook of Statistical Genomics*. Wiley, July 2019, pp. 997–1020. DOI: 10.1002/9781119487845.ch36.

[41] Seunggeung Lee et al. "Rare-Variant Association Analysis: Study Designs and Statistical Tests". In: *The American Journal of Human Genetics* 95 (2014), pp. 5–23. DOI: 10.1016/j.ajhg.2014.06.009. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4085641/pdf/main.pdf.

[42] Bingshan Li and Suzanne M Leal. "Methods for Detecting Associations with Rare Variants for Common Diseases: Application to Analysis of Sequence Data". In: (). DOI: 10.1016/j.ajhg.2008.06.024.

[43] Michael A. Mooney and Beth Wilmot. "Gene set analysis: A step-by-step guide". In: *American Journal of Medical Genetics, Part B: Neuropsychiatric Genetics* 168.7 (Oct. 2015), pp. 517–527. ISSN: 1552485X. DOI: 10.1002/ajmg.b.32328.

[44] Marquitta J. White et al. “Strategies for Pathway Analysis Using GWAS and WGS Data”. In: *Current Protocols in Human Genetics* 100.1 (Jan. 2019), e79. ISSN: 19348258. DOI: 10.1002/cphg.79.

[45] Katie Saund et al. “Prewas: Data pre-processing for more informative bacterial gwas”. In: *Microbial Genomics* 6.5 (Dec. 2020), pp. 1–8. ISSN: 20575858. DOI: 10.1099/mgen.0.000368.

[46] Ado van Assche et al. “Phylogenetic signal in phenotypic traits related to carbon source assimilation and chemical sensitivity in Acinetobacter species”. In: *Applied Microbiology and Biotechnology* 101 (2016), pp. 367–379. ISSN: 14320614. DOI: 10.1007/s00253-016-7866-0.

[47] Mark Pagel. “Inferring the historical patterns of biological evolution.” In: *Nature* 401.6756 (1999), pp. 877–884. ISSN: 0028-0836. DOI: 10.1038/44766.

[48] Susanne A. Fritz and Andy Purvis. “Selectivity in mammalian extinction risk and threat types: A new measure of phylogenetic signal strength in binary traits”. In: *Conservation Biology* 24.4 (2010), pp. 1042–1051. ISSN: 08888892. DOI: 10.1111/j.1523-1739.2010.01455.x.

[49] Emmanuel Paradis and Klaus Schliep. “Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R”. In: *Bioinformatics* 35.3 (2019), pp. 526–528. ISSN: 14602059. DOI: 10.1093/bioinformatics/bty633.

[50] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2019. URL: https://www.R-project.org/.

[51] David Orme. “The caper package : comparative analysis of phylogenetics and evolution in R”. In: *R package version 0.5, 2* (2013), pp. 1–36. ISSN: 11755326. DOI: 1. arXiv: arXiv:1011.1669v3.

[52] Liam J. Revell. "phytools: An R package for phylogenetic comparative biology (and other things)". In: *Methods in Ecology and Evolution* 3.2 (2012), pp. 217–223. ISSN: 2041210X. DOI: 10.1111/j.2041-210X.2011.00169.x.

[53] Hadley Wickham et al. "Tidyverse: Easily install and load the 'tidyverse'". In: *R package version* 1.1 (2017), p. 2017.

[54] Hadley Wickham and Dana Seidel. *scales: Scale Functions for Visualization*. 2019.

[55] Baptiste Auguie. *gridExtra: Miscellaneous Functions for "Grid" Graphics*. 2017. URL: https://cran.r-project.org/package=gridExtra.

[56] *Anaconda*. URL: https://www.anaconda.com/ (visited on 02/21/2020).

[57] Morteza M Saber and B Jesse Shapiro. "Benchmarking bacterial genome-wide association study methods using simulated genomes and phenotypes". In: *Microbial genomics* 6.3 (2020). ISSN: 20575858. DOI: 10.1099/mgen.0.000337.

[58] Jianzhi Zhang and Masatoshi Nei. *Accuracies of Ancestral Amino Acid Sequences Inferred by the Parsimony, Likelihood, and Distance Methods*. Tech. rep. 1997, pp. 139–146.

[59] Revathi Govind and Bruno Dupuy. "Secretion of Clostridium difficile Toxins A and B Requires the Holin-like Protein TcdE". In: *PLoS Pathogens* 8.6 (2012), e1002727. ISSN: 1932-6203. DOI: 10.1371/. arXiv: arXiv:1208.5792v1. URL: http://www.nsfc.gov.cn/publish/portal1/.

[60] Alexandra Olling et al. "Release of TcdA and TcdB from Clostridium difficile cdi 630 is not affected by functional inactivation of the tcdE gene". In: *Microbial Pathogenesis* 52.1 (2012), pp. 92–100. ISSN: 08824010. DOI: 10.1016/j.micpath.2011.10.009. URL: http://dx.doi.org/10.1016/j.micpath.2011.10.009.

[61] Sandra Janezic et al. “Comparative genomics of Clostridioides difficile toxinotypes identifies module-based toxin gene evolution”. In: *Microbial Genomics* (2020). DOI: 10.1099/mgen.0.000449.

[62] Michael J. Mansfield et al. “Phylogenomics of 8,839 Clostridioides difficile genomes reveals recombination-driven evolution and diversification of toxin A and B”. In: *Biorvix* (2020). URL: https://www.biorxiv.org/content/10.1101/2020.07.09.194449v1.abstract.

[63] Brittany B Lewis et al. “Pathogenicity Locus, Core Genome, and Accessory Gene Contributions to Clostridium difficile Virulence.” In: *mBio* 8.4 (2017), pp. 1–15. ISSN: 2150-7511. DOI: 10.1128/mBio.00885-17. URL: http://mbio.asm.org/content/8/4/e00885-17.full.pdf%20http://www.ncbi.nlm.nih.gov/pubmed/28790208.

[64] Michel Warny et al. “Toxin production by an emerging strain of Clostridium difficile associated with outbreaks of severe disease in North America and Europe”. In: *Lancet* 366.9491 (2005), pp. 1079–1084. ISSN: 01406736. DOI: 10.1016/S0140-6736(05)67420-X. URL: https://ac.els-cdn.com/S014067360567420X/1-s2.0-S014067360567420X-main.pdf?%7B%5C_%7Dtid=fd2c84cd-1334-4fd2-801a-4d11c2f6b80b%7B%5C&%7Dacdnat=1519935230%7B%5C_%7Dc992af2082d9b57565b170cfe836f1b2.

[65] Paul E Carlson et al. “The relationship between phenotype, ribotype, and clinical disease in human Clostridium difficile isolates”. In: *Anaerobe* 24 (2013), pp. 109–116. ISSN: 10759964. DOI: 10.1016/j.anaerobe.2013.04.003. arXiv: NIHMS150003.

[66] Simon Andrews. *FastQC: a quality control tool for high throughput sequence data*. R Foundation for Statistical Computing. Vienna, Austria, 2010. URL: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

[67] Anthony M Bolger, Marc Lohse, and Bjoern Usadel. “Trimmomatic: a flexible trimmer for Illumina sequence data”. In: *Bioinformatics* 30.15 (2014), pp. 2114–2120.

[68] Andrew J Page et al. "Roary: rapid large-scale prokaryote pan genome analysis". In: *Bioinformatics* 31.22 (2015), pp. 3691–3693.

[69] Torsten Seemann. "Prokka: Rapid prokaryotic genome annotation". In: *Bioinformatics* 30.14 (2014), pp. 2068–2069. ISSN: 14602059. DOI: 10.1093/bioinformatics/btu153.

[70] Katie Saund and Evan S Snitkin. "hogwash: Three Methods for Genome-Wide Association Studies in Bacteria". In: *bioRxiv* (2020). DOI: 10.1101/2020.04.19.048421. URL: https://www.biorxiv.org/content/early/2020/08/12/2020.04.19.048421.

[71] Heng Li and Richard Durbin. "Fast and accurate short read alignment with Burrows–Wheeler transform". In: *Bioinformatics* 25.14 (May 2009), pp. 1754–1760. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btp324. eprint: https://academic.oup.com/bioinformatics/article-pdf/25/14/1754/605544/btp324.pdf. URL: https://doi.org/10.1093/bioinformatics/btp324.

[72] *Picard toolkit*. http://broadinstitute.github.io/picard/. 2019.

[73] Heng Li et al. "The Sequence Alignment/Map format and SAMtools". In: *Bioinformatics* 25.16 (June 2009), pp. 2078–2079. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btp352. eprint: https://academic.oup.com/bioinformatics/article-pdf/25/16/2078/531810/btp352.pdf. URL: https://doi.org/10.1093/bioinformatics/btp352.

[74] Aaron McKenna et al. "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data". In: *Genome research* 20.9 (2010), pp. 1297–1303.

[75] Lam-Tung Nguyen et al. "IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies". In: *Molecular biology and evolution* 32.1 (2015), pp. 268–274.

[76] Diep Thi Hoang et al. "UFBoot2: improving the ultrafast bootstrap approximation". In: *Molecular biology and evolution* 35.2 (2018), pp. 518–522.

[77] Subha Kalyaanamoorthy et al. "ModelFinder: fast model selection for accurate phylogenetic estimates". In: *Nature methods* 14.6 (2017), p. 587.

[78] Guangchuang Yu et al. "Ggtree: an R Package for Visualization and Annotation of Phylogenetic Trees With Their Covariates and Other Associated Data". In: *Methods in Ecology and Evolution* 8.1 (2017), pp. 28–36. ISSN: 2041210X. DOI: 10.1111/2041-210X.12628.

[79] Guangchuang Yu. *aplot: Decorate a 'ggplot' with Associated Information*. R package version 0.0.6. 2020. URL: https://CRAN.R-project.org/package=aplot.

[80] Matt Dowle and Arun Srinivasan. *data.table: Extension of 'data.frame'*. R package version 1.12.8. 2019. URL: https://CRAN.R-project.org/package=data.table.

[81] Raivo Kolde. *pheatmap: Pretty Heatmaps*. R package version 1.0.12. 2019. URL: https://CRAN.R-project.org/package=pheatmap.

[82] Thomas Hundsberger et al. "Transcription analysis of the genes tcdA-E of the pathogenicity locus of Clostridium difficile". In: *European Journal of Biochemistry* 244.3 (1997), pp. 735–742. ISSN: 00142956. DOI: 10.1111/j.1432-1033.1997.t01-1-00735.x.

[83] Imane El Meouche et al. "Characterization of the SigD regulon of C. difficile and its positive control of toxin production through the regulation of tcdR". In: *PLoS ONE* 8.12 (2013), pp. 1–17. ISSN: 19326203. DOI: 10.1371/journal.pone.0083748.

[84] Amy L. Davidson et al. "Structure, Function, and Evolution of Bacterial ATP-Binding Cassette Systems". In: *Microbiology and Molecular Biology Reviews* 72.2 (2008), pp. 317–364. ISSN: 1092-2172. DOI: 10.1128/mmbr.00031-07.

[85] Annie Aubry et al. "Modulation of toxin production by the flagellar regulon in Clostridium difficile". In: *Infection and Immunity* 80.10 (2012), pp. 3521–3532. ISSN: 00199567. DOI: 10.1128/IAI.00224-12.

[86] Robert W. McKee et al. "The second messenger cyclic Di-GMP regulates Clostridium difficile toxin production by controlling expression of sigD". In: *Journal of Bacteriology* 195.22 (2013), pp. 5174–5185. ISSN: 00219193. DOI: 10.1128/JB.00501-13.

[87] Brandon R. Anjuwon-Foster and Rita Tamayo. "A genetic switch controls the production of flagella and toxins in Clostridium difficile". In: *PLoS Genetics* 13.3 (2017), pp. 1–33. ISSN: 15537404. DOI: 10.1371/journal.pgen.1006701.

[88] Paul E. Carlson et al. "Variation in germination of Clostridium difficile clinical isolates correlates to disease severity". In: *Anaerobe* 33 (2015), pp. 64–70. ISSN: 10958274. DOI: 10.1016/j.anaerobe.2015.02.003. URL: http://dx.doi.org/10.1016/j.anaerobe.2015.02.003.

[89] L. Clifford McDonald et al. "Recommendations for Surveillance of Clostridium difficile–Associated Disease". In: *Infection Control amp; Hospital Epidemiology* 28.2 (2007), pp. 140–145. DOI: 10.1086/511798.

[90] Krishna Rao et al. "Clostridium difficile ribotype 027: relationship to age, detectability of toxins A or B in stool with rapid testing, severe infection, and mortality". In: *Clinical Infectious Diseases* 61.2 (2015), pp. 233–241.

[91] A Sarah Walker et al. "Relationship between bacterial strain type, host biomarkers, and mortality in Clostridium difficile infection". In: *Clinical infectious diseases* 56.11 (2013), pp. 1589–1600.

[92] David W. Eyre et al. "Clostridium difficile trehalose metabolism variants are common and not associated with adverse patient outcomes when variably present in the same lineage". In: *EBioMedicine* 43 (2019), pp. 347–355. ISSN: 2352-3964. DOI: https://doi.org/10.1016/j.ebiom.2019.04.038. URL: http://www.sciencedirect.com/science/article/pii/S2352396419302774.

[93] Jonathan NV Martinson et al. "Evaluation of portability and cost of a fluorescent PCR ribotyping protocol for Clostridium difficile epidemiology". In: *Journal of clinical microbiology* 53.4 (2015), pp. 1192–1197.

[94] James Collins, Heather Danhof, and Robert A Britton. "The role of trehalose in the global spread of epidemic Clostridium difficile". In: *Gut microbes* 10.2 (2019), pp. 204–209.

[95] Jessica SH Martin et al. "Patient and strain characteristics associated with Clostridium difficile transmission and adverse outcomes". In: *Clinical Infectious Diseases* 67.9 (2018), pp. 1379–1387.

[96] Delphine Charif and Jean R Lobry. "SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis". In: *Structural approaches to sequence evolution*. Springer, 2007, pp. 207–232.

[97] Terry M Therneau. "A Package for Survival Analysis in S; 2015. Version 2.38". In: *URL: https://CRAN. R-project. org/package= survival* (2015).

[98] Daniel J. Schaid, Wenan Chen, and Nicholas B. Larson. "From genome-wide associations to candidate causal variants by statistical fine-mapping". In: *Nature Reviews Genetics* 19.8 (2018), pp. 491–504. ISSN: 14710064. DOI: 10.1038/s41576-018-0016-z. URL: http://dx.doi.org/10.1038/s41576-018-0016-z.

[99] John Lonsdale et al. "The Genotype-Tissue Expression (GTEx) project". In: *Nature Genetics* 45.6 (2013), pp. 580–585. ISSN: 10614036. DOI: 10.1038/ng.2653.

[100] Richard Barfield et al. "Transcriptome-wide association studies accounting for colocalization using Egger regression". In: *Genetic Epidemiology* 42.5 (2018), pp. 418–433. ISSN: 10982272. DOI: 10.1002/gepi.22131.

[101] Nicholas Mancuso et al. “Probabilistic fine-mapping of transcriptome-wide association studies”. In: *Nature Genetics* 51.4 (2019), pp. 675–682. ISSN: 15461718. DOI: 10.1038/s41588-019-0367-1. URL: http://dx.doi.org/10.1038/s41588-019-0367-1.

[102] David Melzer et al. “A genome-wide association study identifies protein quantitative trait loci (pQTLs)”. In: *PLoS Genetics* 4.5 (2008). ISSN: 15537390. DOI: 10.1371/journal.pgen.1000072.

[103] Karsten Suhre et al. “Connecting genetic risk to disease end points through the human blood plasma proteome”. In: *Nature Communications* 8 (2017). ISSN: 20411723. DOI: 10.1038/ncomms14357.

[104] Valur Emilsson et al. “Co-regulatory networks of human serum proteins link genetics to disease”. In: *Science* 361.6404 (2018), pp. 769–773. ISSN: 10959203. DOI: 10.1126/science.aaq1327.

[105] Clare Marley et al. “Evaluation of a risk score to predict future Clostridium difficile disease using UK primary care and hospital data in Clinical Practice Research Datalink”. In: *Human Vaccines and Immunotherapeutics* 15.10 (2019), pp. 2475–2481. ISSN: 2164554X. DOI: 10.1080/21645515.2019.1589288. URL: https://doi.org/10.1080/21645515.2019.1589288.

[106] Jeeheh Oh et al. “A Generalizable, Data-Driven Approach to Predict Daily Risk of Clostridium difficile Infection at Two Large Academic Health Centers”. In: *Infection Control and Hospital Epidemiology* 39.4 (2018), pp. 425–433. ISSN: 15596834. DOI: 10.1017/ice.2018.16.

[107] Sepideh Pakpour et al. “Identifying predictive features of Clostridium difficile infection recurrence before, during, and after primary antibiotic treatment”. In: *Microbiome* 5.1 (2017), p. 148. ISSN: 20492618. DOI: 10.1186/s40168-017-0368-1.

[108] Jennifer L. Kuntz et al. “Predicting the Risk of Clostridium difficile Infection upon Admission: A Score to Identify Patients for Antimicrobial Stewardship Efforts”. In: *The Permanente journal* 20.1 (2016), pp. 20–25. ISSN: 15525775. DOI: 10.7812/TPP/15-049.

[109] Shanika L. Amarasinghe et al. “Opportunities and challenges in long-read sequencing data analysis”. In: *Genome Biology* 21.1 (2020), pp. 1–16. ISSN: 1474760X. DOI: 10.1186/s13059-020-1935-5.

[110] Marc Monot et al. “Reannotation of the genome sequence of Clostridium difficile strain 630”. In: *Journal of Medical Microbiology* 60.8 (2011), pp. 1193–1199. ISSN: 00222615. DOI: 10.1099/jmm.0.030452-0.

[111] Adam A. Behroozian et al. “ Detection of Mixed Populations of Clostridium difficile from Symptomatic Patients Using Capillary-Based Polymerase Chain Reaction Ribotyping ”. In: *Infection Control & Hospital Epidemiology* 34.9 (2013), pp. 961–966. ISSN: 0899-823X. DOI: 10.1086/671728.

[112] Anna M. Seekatz et al. “Presence of multiple Clostridium difficile strains at primary infection is associated with development of recurrent disease”. In: *Anaerobe* 53 (Oct. 2018), pp. 74–81. ISSN: 10958274. DOI: 10.1016/j.anaerobe.2018.05.017. URL: https://linkinghub.elsevier.com/retrieve/pii/S1075996418301021.

[113] Thibault Stalder et al. “Linking the resistome and plasmidome to the microbiome”. In: *ISME Journal* 13.10 (2019), pp. 2437–2446. ISSN: 17517370. DOI: 10.1038/s41396-019-0446-4. URL: http://dx.doi.org/10.1038/s41396-019-0446-4.