Université de Montréal

Hunting for causal variants in microbial genomes

par

Peter E. Chen

Département de Sciences Biologiques
Faculté des Arts et des Sciences

Thèse présenté(e) en vue de l'obtention du grade de *Philosophiae doctor*
en Sciences Biologiques

November 2021

Université de Montréal
Département de Sciences Biologiques, Faculté des Arts et des Sciences

---

Cette thèse intitulée

Hunting for causal variants in microbial genomes

Présenté par
Peter E. Chen

A été évalué par un jury composé des personnes suivantes

Timothée Poisot
Président-rapporteur

Jesse Shapiro
Directeur de recherche

Luis Barreiro
Membre du jury

Timothy Read
Examinateur externe

Résumé

L'un des objectifs centraux de la biologie est de comprendre comment l'ADN, la séquence primaire, donne lieu à des traits observables. À cette fin, nous examinons ici des méthodes pour identifier les composants génétiques qui influencent les traits microbiens. Par « identifier », nous entendons l'élucidation à la fois l'état allélique et de la position physique de chaque variante causale d'un phénotype d'intérêt à la résolution des nucléotides de paires de bases. Nous nous sommes concentrés sur les études d'association génomique (*genome-wide association studies;* GWAS) en tant qu'approche générale d'étudier l'architecture génétique des traits. L'objectif global de cette thèse était d'examiner de manière critique les méthodologies GWAS et de les considérer en pratique dans des populations microbiennes fortement clonales et non-clonales (*i.e.* avec recombinaison fréquent). Le domaine de la GWAS microbienne est relativement nouveau par rapport aux quinze dernières années de la GWAS humaine, et en tant que tel, nous avons commencé par un examen de l'état de la GWAS microbienne. Nous avons posé deux questions principales : 1) Les méthodes GWAS humaines fonctionnent-elles facilement et sans modification pour les populations microbiennes ? 2) Et sinon, quels sont les problèmes méthodologiques centraux et les modifications nécessaires pour la GWAS microbienne? À partir de ces résultats, nous avons ensuite détaillé le déséquilibre de liaison (*linkage disequilibrium;* LD) comme principal obstacle dans la GWAS microbien, et nous avons présenté une nouvelle méthode, POUTINE, pour relever ce défi en exploitant les mutations homoplasiques pour briser implicitement la structure LD. Le reste de la thèse présente à la fois les méthodes traditionnelles GWAS (comptage des allèles) et POUTINE (comptage d'homoplasies) appliquées à une population hautement recombinogène de génomes de vibrions marins. Malgré une taille d'échantillon modeste, nous donnons un premier aperçu de l'architecture génétique de la résistance aux bactériophages dans une population naturelle, tout en montrant que les récepteurs des bactériophages jouent un rôle primordial. Ce résultat est en pleine cohérence avec des expériences en laboratoire de coévolution phage-bactérie. Il est important de noter que cette architecture met en évidence à quel point la sélection positive peut sculpter certains traits microbiens différemment de nombreux traits complexes humains, qui sont généralement soumis à une faible sélection purificatrice. Plus précisément, nous avons identifié des mutations à effet important à haute fréquence qui sont rarement observées dans les phénotypes complexes humains où de nombreuses mutations à faible effet contribuent à l'héritabilité. La thèse se termine par des perspectives sur les voies à suivre pour la GWAS microbienne.

Mots-clés: étude d'association génomique microbienne, déséquilibre de liaison, comptage d'allèles, comptage d'homoplasie, architecture génétique

Abstract

One of the central goals of biology is to understand how DNA, the primary sequence, gives rise to observable traits. To this aim, we herein examine methods to identify the genetic components that influence microbial traits. By "identify" we mean the elucidation of both the allelic state and physical position of each causal variant of a phenotype of interest down to the base-pair nucleotide resolution. Our focus has been on genome-wide association studies (GWAS) as a general approach to dissecting the genetic architecture of traits. The overarching aim of this thesis was to critically examine GWAS methodologies and to consider them in practice in both strongly clonal and highly recombining microbial populations. The field of microbial GWAS is relatively new compared to the over fifteen years of human GWAS, and as such, we began this work with an examination of the state of microbial GWAS. We asked and attempted to answer two main questions: 1) Do human GWAS methods readily work without modification for microbial populations? 2) And if not, what are the central methodological problems and changes that are required for a successful microbial GWAS? Building from these findings, we then detailed linkage disequilibrium (LD) as the primary obstacle in microbial GWAS, and we presented a new method, POUTINE, to address this challenge by harnessing homoplasic mutations to implicitly break LD structure. The remainder of the thesis showcases both traditional GWAS methods (allele counting) and POUTINE applied to a highly recombining population of marine vibrio genomes. Despite a small sample size, we provide a first glimpse into the genetic architecture of bacteriophage resistance in a natural population and show that bacteriophage receptors play a primary role consistent with experimental populations of phage-bacteria coevolution. Importantly, this architecture highlights how strong positive selection can sculpt some microbial traits differently than many human complex traits, which are generally under weak purifying selection. Specifically, we identified common frequency, large-effect mutations that are rarely observed in human complex phenotypes where many low-effect mutations are thought to contribute to the bulk of heritability. The thesis concludes with perspectives on ways forward for microbial GWAS.

Keywords: microbial GWAS, linkage disequilibrium, allele counting, homoplasy counting, genetic architecture

Table of contents

List of tables

List of figures

Acknowledgements

I would like to acknowledge the mentors I have had in lives' past. I stand on their shoulders and all that they have instilled in me. To Aravinda Chakravarti, whose insightfulness and brilliance have left me with ideas that I still ponder about to this day. To David Cutler, who along with Aravinda, taught me much of what I know about population genetics. To Tim Read, who took a chance on a young mind, and showed me the wonderous world of microbes.

To Jesse Shapiro, I thank you for the patience and freedom you have extended to me. You let loose more rope each time I needed it. I sincerely look forward to all the ideas we will be chatting about as we deepen our collaborative spirits.

To my parents, Howard and Susan, in this moment I cannot but think of you both. I think of all the sacrifices that we have made as a family, crossing an ocean to start a new life in a strange world where we did not speak the local language. I remember all those mornings when I was in high school, and mom would wake up first to press fresh-squeezed juice to keep me healthy. I remember being greatly fatigued each morning from getting scant hours of sleep the night before as I toiled away at my homework. But each morning, I had something healthy waiting for me, made by the loving hands a mother has for her son. I could not have arrived at this moment without the sacrifices my parents have made for me. For this, I am deeply grateful. As I continue to nurture the next generation, I will be guided by my memories of your sacrifices.

In a moment of levity, I would also like to thank the fine folks at the Lagavulin, Teeling, Buffalo Trace, Jameson, Redbreast, Green Spot, Glenmorangie, Bushmills, Talisker, Kilkerran, Connemara, and Glendronach distilleries for producing the whiskeys that have been my quiet company while I wrote this thesis on many sleepless nights.

Introduction

Every generation, the genomes of organisms are passed down with great fidelity and give rise to a diversity of phenotypes. The physical entity that carries this heritability was not always believed to be deoxyribonucleic acid (DNA), and for some time it was widely believed that proteins were the carriers of heritable information. It was not until the Avery–MacLeod–McCarty experiment in 1944 [1] and the Hershey-Chase experiment in 1952 [2] that DNA was confirmed to be the material responsible for heredity. In this thesis, we explore and further develop genome-wide association studies (GWAS) as a general approach to elucidate the genetic components that underlie the diversity of heritable phenotypes encoded by genomes. The overarching aim of this thesis was to critically examine GWAS methodologies and to consider them in practice in both strongly clonal and highly recombining microbial populations.

Viewing microbial GWAS within its historical context, two facets stand out: the diversity of life and the nascent history of microbial association studies. Broadly, the diversity of life can be broken down into prokaryotes and eukaryotes. Yet despite their genetic diversity and long evolutionary history, prokaryotes make up only a minority of genotype to phenotype studies. Historically, much effort has gone into deciphering the genome sequence of eukaryotes, primarily *Homo sapiens*. As of October 2021, the Human GWAS Catalog of the National Human Genome Research Institute (NHGRI) of the United States and the European Bioinformatics Institute (EBI) officially reports 5,419 GWAS publications, each investigating a minimum of 100,000 single nucleotide polymorphisms (SNPs). Within these studies, it is reported that greater than 60,000 SNPs are associated with more than 600 phenotypes at genome-wide significance (p-value $\leq 5.0 \times 10^{-8}$) (http://www.genome.gov/gwastudies/) [3]. Moreover, within the realm of eukaryotic studies, there have been prominent GWA studies in *Arabidopsis thaliana*, rice, maize, cattle, sheep, dog, and mice [4–10].

As with human phenotypes, prokaryotes and viruses also possess compelling traits of interest. Importantly, genotype to phenotype studies in microbial populations can

elucidate the genetic basis of pathogenesis in infectious diseases. Highlighting the translational and clinical importance of this area of study, a small sampling of infectious disease traits includes such phenotypes as drug resistance, transmissibility, virulence, biofilm formation, persistence in the host (evasion of the host immune system for prolonged periods), and vaccine attenuation (many vaccines are formulated from attenuated strains with no known knowledge of how the strains were attenuated).

A Brief History of Genetic Mapping Studies

The history of genotype to phenotype studies can be broken down into bottom-up and top-down approaches. The bottom-up methods attempt to modify and perturb DNA to test its effect on phenotype. Changes to the phenotype are thought to be due to the changes made to the DNA. This class of methods includes gene knockouts, over-expression studies, and various mutagenesis techniques. In contrast, top-down studies begin with the phenotype and attempt to "map" the underlying genetic components back to the primary sequence. This class of methods is inherently agnostic in that there is no prior knowledge of the variants that contribute to the phenotype. Top-down approaches also have the significant advantage of studying natural genetic variation as it occurs in the natural environment, as opposed to bottom-up approaches which are limited by the genetic variation that can be constructed in the lab (*e.g.,* gene deletions or targeted mutations). Despite the challenges involved in top-down approaches, they have the potential to shed light on evolutionary processes as they occur in nature.

Genetic mapping dates back to the work of Alfred Henry Sturtevant and Thomas Hunt Morgan at Columbia University. In Sturtevant's 1913 publication, he put forth the logic of genetic mapping still in use today and created the world's first genetic map [11]; this map was of the common fruit fly, *Drosophila melanogaster*. The central idea here is that the genetic components underlying traits not only have regular positions along a chromosome but that their linear position can be deduced by exploiting the frequency

of recombination as a measure of approximate location (genetic distance), with loci farther apart experiencing more frequent recombination events (crossovers) between them. In Sturtevant's 1965 "A History of Genetics" [12], he writes: "In the latter part of 1911, in conversation with Morgan ... I suddenly realized that the variations in strength of linkage, already attributed by Morgan to differences in the spatial separation of the genes, offered the possibility of determining sequences in the linear dimension of a chromosome. I went home and spent most of the night (to the neglect of my undergraduate homework) in producing the first chromosome map, which included the sex-linked genes y, w, v, m, and r, in the order and approximately the relative spacing that they still appear on the standard maps."

The above logic guided an explosion of genetic maps in various model organisms, but it was not until 1980 that the same logic was published for human genomes [13]. David Botstein and colleagues realized that naturally occurring human sequence variation (in this study they exploited RFLPs, restriction fragment length polymorphisms) can act as genetic markers to trace the inheritance patterns across human families in a manner analogous to the controlled crosses first done by Sturtevant with *Drosophila*. These early human linkage studies had their most success in rare diseases where the causal locus is monogenic and behaves in a simple Mendelian manner. However, these linkage studies largely failed to elucidate the genetic components of many traits of medical relevance that departed from Mendelian behavior and operated under a more complex pattern of inheritance (today these traits are known as complex traits).

Addressing this problem and bringing us into the modern era of genotype to phenotype studies, Risch and Merikangas in 1996 proposed association studies as a way to capture higher-frequency, smaller-effect polymorphisms; in contrast, linkage studies were mainly suitable for capturing causal loci of large effect [14]. Briefly, a basic case and control GWAS attempts to statistically associate specific alleles with a phenotype of interest by looking for over-represented alleles in cases relative to controls. On a

practical level, association studies possess an advantage in that it is easier to collect larger numbers of unrelated individuals from the general population versus the families (trios of two parents and one child) needed for linkage studies. The main component missing was a readily available set of human sequence variation to power these association studies. These insights provided the impetus to design and build a haplotype map (HapMap) of the human genome. To date, this map consists of over three million SNPs scattered across the genome in four major world-wide populations consisting of 269 individuals [15]. The first HapMap was published in 2005, making it practical for association studies to move to the scope of whole-genome scans of association between variants and phenotypes [16].

The advent of microbial GWAS

Although the genetic basis of traits was of interest to those studying bacterial genomes, there had not been a concerted effort to bridge the gap between genotype and phenotype akin to that seen in human disease studies and other eukaryotic model organisms. It was not until 2006 that Falush and Bowden broached the subject in a paper entitled “Genome-Wide Association Mapping in Bacteria?” [17]. Drawing inspiration from the recently published HapMap, the authors argued for top-down methods, particularly a case and control association study design, to be employed for bacteria. In brief, the authors claimed that the same principles that have guided human GWA studies are fundamentally unchanged for bacterial populations. Presciently, in their final paragraph, they write: “However, to realize the full potential of these methods, a more detailed knowledge of how variation is distributed and transmitted within bacterial populations must be developed. For example, the effectiveness of SNP typing as a method of resolving genetic relationships depends crucially on the rate of recombination, and association studies will be of limited use in organisms that are completely clonal or recombine infrequently.”

Perhaps one of the first computational efforts to link genotype to phenotype in bacterial genomes was published earlier that year [18]. Notably, the approach was not based upon the association methods that were in vogue at the time in human studies. Rather, the authors presented a method to determine lineage-specific molecular adaptations using comparisons of selection signals from different phenotypic populations. The trait of interest was uropathogenicity in *E. coli*, a human-virulent lifestyle markedly different from the usual commensal state of this common gut bacteria. Tests for positive natural selection at the protein level were performed using PAML (Phylogenetic Analysis Using Maximum Likelihood) [19], and the comparative portion of their approach was simply the presence or absence of positive selection in UPEC (uropathogenic *E. coli*) strains versus non-UPEC strains (i.e., a gene is UPEC-specific because it showed positive selection in UPEC strains but no selection in non-UPEC strains). This study reported 29 genes under positive selection in only the UPEC strains, while remarkably using only seven genomes, of which two were UPEC strains.

Highlighting the paucity of genotype to phenotype studies in bacterial genomes, it is not until 2013 that the first GWAS in bacterial genomes appears in the literature [20]. In this study, the phenotype of interest was host preference and the two host types examined were chicken and cattle. The authors studied a dataset of 192 genomes of *Campylobacter jejuni* and *Campylobacter coli* (these bacteria are common causative agents of food poisoning in humans). The authors provided an interesting variant of the classic GWAS approach. While most GWAS methods rely on SNPs to “tag” nearby causal variants that are in linkage disequilibrium (LD) with the statistically associated SNP, the authors used unique 'words' of 30 nucleotides (30-mers) as the base unit of association. This change allowed their method to capture not only SNPs but also whole gene absence or presence, as well as smaller insertions or deletions (indels). 9,034 of these 30-mers were significantly associated with either cattle or chicken and mapped to 97 genes grouped into 10 genomic regions. Within these 10 regions, the vast majority of the cattle-associated words mapped (minimum of 70% sequence identity and 50% alignment length) to three adjacent genes, *panB*, *panC*, and *panD*.

These genes have been known to be involved in the pantothenic acid (vitamin B5) synthesis pathway. This association between the *panBCD* genes and cattle host preference was further supported by the authors' *in vitro* functional study which consisted of growing *Campylobacter* strains possessing and lacking the *panBCD* region on growth medium with and without vitamin B5. This set of tests showed that cattle-associated isolates do indeed have a greater capacity to grow in a vitamin B5 depleted environment compared to isolates from chicken. The authors postulated that host preference of *Campylobacter* was largely due to host diets where vitamin B_5 is abundant in cereals and grains (main diet of chicken) but is at low concentration in grasses (main diet of cattle), thus suggesting that *Campylobacter* needs to produce the vitamin itself in order to persist in cattle.

Thesis outline

Rarely does one receive the fortune of being in the beginnings of a new field. Thus, it was natural that we started our work with a thorough review of the microbial association studies available at that time, as presented in Chapter 1. We note that this chapter was published in 2015 and that all but one paper reviewed was published within two years prior to this date. Since 2015, a number of microbial GWAS studies have been published, highlighting the growing interest in this nascent field. Chapter 1 is not just a literature review, but also defines the terms 'allele counting' and 'homoplasy counting,' and suggests the latter as a promising GWAS method in strongly clonal populations. We compared these two approaches in a GWAS using a dataset of *Mycobacterium tuberculosis* genomes and antibiotic resistance phenotypes. Briefly, homoplasies are mutations that occur independently in different lineages, thereby breaking the dependence structure of a clonal phylogeny. Homoplasies are discussed in greater detail in Chapter 1 and provide the basis for a novel homoplasy-based GWAS method (POUTINE) developed in Chapter 2. In this next chapter, we examined the effects of LD on association signals from both allele and homoplasy counting methods in a GWAS using 1,330 *M. tuberculosis* genomes and isoniazid drug

resistance as the phenotype. Importantly, we showed how strong and long-range LD can prevent allele-counting methods from distinguishing between a putative causal locus and its linked sites scattered throughout the genome. In contrast, we showed that POUTINE hits were mostly unlinked from other sites, and thus these homoplasy-counting hits could be considered the true driver of the association signal. In Chapter 3, we again compared allele-counting and homoplasy-counting GWAS methods in a dataset of aquatic *Vibrio breoganii*, a much more highly recombining and less clonal population than *M. tuberculosis.* Despite a small sample size, we identified three genome-wide significant mutations of large effect, which highlights how strong positive selection can shape the genetic architecture of microbial traits in contrast to weakly purifying selection likely shaping many human complex phenotypes. Consistent with phage-bacteria coevolution in experimental lab studies, we showed that modifications in bacteriophage receptors, whether directly via a mutation in the receptor gene or indirectly via a mutation in a gene that modifies receptor structure, played a primary role in bacteriophage resistance. The thesis concludes with some perspectives on the field of microbial GWAS going forward.

References

1. Avery OT, Macleod CM, McCarty M. Studies on the chemical nature of the substance inducing transformation of pneumococcal types : Induction of transformation by a desoxyribonucleic acid fraction isolated from Pneumococcus Type iii. J Exp Med. 1944;79: 137–158.

2. Hershey AD, Chase M. Independent functions of viral protein and nucleic acid in growth of bacteriophage. J Gen Physiol. 1952;36: 39–56.

3. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019;47: D1005–D1012.

4. Atwell S, Huang YS, Vilhjálmsson BJ, Willems G, Horton M, Li Y, et al. Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature. 2010;465: 627–631.

5. Kump KL, Bradbury PJ, Wisser RJ, Buckler ES, Belcher AR, Oropeza-Rosas MA, et al. Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. Nat Genet. 2011;43: 163–168.

6. Yano K, Yamamoto E, Aya K, Takeuchi H, Lo P-C, Hu L, et al. Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice. Nat Genet. 2016;48: 927–934.

7. Bouwman AC, Daetwyler HD, Chamberlain AJ, Ponce CH, Sargolzaei M, Schenkel FS, et al. Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals. Nat Genet. 2018;50: 362–367.

8. Li X, Yang J, Shen M, Xie X-L, Liu G-J, Xu Y-X, et al. Whole-genome resequencing of wild and domestic sheep identifies genes associated with morphological and agronomic traits. Nat Commun. 2020;11: 2815.

9. Karlsson EK, Sigurdsson S, Ivansson E, Thomas R, Elvers I, Wright J, et al. Genome-wide analyses implicate 33 loci in heritable dog osteosarcoma, including regulatory variants near CDKN2A/B. Genome Biol. 2013;14: R132.

10. Flint J, Eskin E. Genome-wide association studies in mice. Nature Reviews Genetics. 2012;13: 807–817.

11. Sturtevant AH. The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association. J Exp Zool. 1913;14: 43–59.

12. Sturtevant AH. A history of genetics. Previously published. New York: Harper & Row; 1965.

13. Botstein D, White RL, Skolnick M, Davis RW. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet. 1980;32: 314.

14. Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science. 1996;273: 1516–1517.

15. The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007;449: 851–861.

16. International HapMap Consortium. A haplotype map of the human genome. Nature. 2005;437: 1299–1320.

17. Falush D, Bowden R. Genome-wide association mapping in bacteria? Trends Microbiol. 2006;14: 353–355.

18. Chen SL, Hung C-S, Xu J, Reigstad CS, Magrini V, Sabo A, et al. Identification of genes subject to positive selection in uropathogenic strains of Escherichia coli: a comparative genomics approach. Proc Natl Acad Sci U S A. 2006;103: 5977–5982.

19. Yang Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol Biol Evol. 2007;24: 1586–1591.

20. Sheppard SK, Didelot X, Meric G, Torralbo A, Jolley KA, Kelly DJ, et al. Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in Campylobacter. Proceedings of the National Academy of Sciences. 2013;110: 11923–11927.

Chapter 1: The advent of genome-wide association studies for bacteria

Chen PE, Shapiro BJ. The advent of genome-wide association studies for bacteria. Curr Opin Microbiol. 2015;25: 17–24.

Abstract

Significant advances in sequencing technologies and genome-wide association studies (GWAS) have revealed substantial insight into the genetic architecture of human phenotypes. In recent years, the application of this approach in bacteria has begun to reveal the genetic basis of bacterial host preference, antibiotic resistance, and virulence. Here, we consider relevant differences between bacterial and human genome dynamics, apply GWAS to a global sample of *Mycobacterium tuberculosis* genomes to highlight the impacts of linkage disequilibrium, population stratification, and natural selection, and finally compare the traditional GWAS against phyC, a contrasting method of mapping genotype to phenotype based upon evolutionary convergence. We discuss strengths and weaknesses of both methods, and make suggestions for factors to be considered in future bacterial GWAS.

Introduction

A central goal of biology is to understand how DNA, the primary sequence, gives rise to observable traits. Historically, much effort has gone into deciphering the primary sequence of eukaryotes, primarily *Homo sapiens*. As of August 8, 2014, the National Human Genome Research Institute (NHGRI) reported 1,961 publications of genome-wide association studies (GWAS). Within these studies, a total of 14,014 single nucleotide polymorphisms (SNPs) are associated with over 600 phenotypes. The advent of GWAS in bacteria has mainly occurred in the last two years [1**, 2**, 3**, 4**, 5**, 6**], and provides an unbiased "top-down" framework [7] to dissect the genetic basis of bacterial phenotypes. In principle, any measurable bacterial phenotype (or archaeal phenotype, although here our focus is on bacteria) can be dissected with a GWAS approach. To date, bacterial GWAS have focused on clinically-relevant phenotypes such as virulence and antibiotic resistance, but there is also great potential to investigate environmentally or industrially relevant phenotypes as well.

Bacterial genomes experience strong linkage, strong stratification, and strong selection

Are bacterial genetic mapping studies any different from eukaryotic studies? Although there are many fundamental differences, this review highlights three features that are most germane to GWAS. The impact of the first two differences, in linkage and population stratification, have been recognized before [6**, 7], but we identify the strength of natural selection relative to drift as a third and under-appreciated factor to consider in bacterial GWAS.

First, unlike eukaryotic recombination which occurs predominantly via the crossing-over of two homologous chromosomes during meiosis, bacterial recombination occurs via gene conversion of relatively short stretches of DNA. In bacteria, recombination is

not coupled with reproduction, and can occur multiple times within a cell's lifespan, or not at all. Without any recombination, purely clonal transmission of DNA leaves the entire bacterial chromosome in complete linkage (in strong linkage disequilibrium; LD). As with eukaryotic genomes, bacterial recombination events break this linkage, but the landscape of LD is markedly different from that seen in eukaryotes; gene conversion events leave a "patchwork" of recombined tracts on top of a genomic background of linked regions called a clonal frame [8]. In contrast to eukaryotic LD patterns, all regions of the clonal frame are in complete linkage, and these regions may be quite distant from one another. The clonal frame phenomenon limits the utility of classic genetic mapping methods mainly by obscuring the true causal variant from the rest of the linked sites in the clonal frame. Here, we define a variant as causal if it plays a functional role in the phenotype of interest, as opposed to only being correlated with the phenotype.

Second, as with eukaryotes, bacterial genomic diversity may be shaped by population stratification. Stratification refers to a "situation in which the population of interest includes subgroups of individuals that are on average more related to each other than to other members of the wider population" [9]. These subpopulations give rise to spurious associations when "cases" (with phenotype A) are on average more closely related with each other than with "controls" (without phenotype A); in other words, associations due to genetic relatedness rather than causality for the phenotype of interest. The problem of population stratification is particularly acute in highly clonal (rarely recombining) bacteria, and in those with separate geographic or host-associated subpopulations [6**].

Third, the phenotypes of most interest in bacterial GWAS are largely different from many human disease phenotypes. In particular, bacterial phenotypes tend to be shaped by strong natural selection (e.g. positive directional selection driving drug resistance), while many human disease phenotypes evolve largely by genetic drift owing to historically small effective population sizes (e.g. due to population

bottlenecks); in this scenario, drift overpowers purifying selection and leaves slightly deleterious alleles in the population that underlie disease traits [10, 11]. This is not to say that bacteria do not experience genetic drift (particularly in frequently bottlenecked populations), but simply that many traits of interest (e.g. resistance, virulence, host-association) have evolved recently and under strong positive selection. These bacterial traits might also be controlled by mutations with large effect sizes on the phenotypes of interest. If this is the case, relatively small samples of bacterial genomes should be sufficient to identify causal mutations [11, 12].

Units of genetic and phenotypic variation

The two basic requirements for GWAS are genotypic and phenotypic measurements from a sample of organisms. Phenotypes are usually broken into either discrete units (e.g. resistance/sensitive or high/low virulence) or continuous traits (e.g. human height). Phenotypes must be reproducible, and easy to measure, ideally in high-throughput if hundreds or thousands of samples are being studied. At the genotypic level, a set of bacterial genomes can be broken down into a ‘‘core’’ genome shared among nearly all members and an ‘‘accessory’’ genome composed of elements present in some strains but not others (typically including genes involved in environmental adaptation) [13, 14]. The genetic units of a GWAS may be variants in the core (e.g. single nucleotide polymorphisms (SNPs) or small indels) [2**, 3**, 4**, 5**] or in the flexible genome (e.g. presence/absence of larger pieces of DNA including genes or operons [1**, 15, 16, 17] (Table 1). While most bacterial GWAS to date have studied either SNPs or gene presence/absence, Sheppard et al. [1**] described a method that uses n-mers ("words" of DNA) as the basic unit of association, allowing them to study both the core and flexible genome simultaneously.

Table 1. Examples of bacterial genome-wide associations studies to date. SNP is single nucleotide polymorphism, MAF is minor allele frequency, CGH is comparative genomic hybridization.

Study	Year	Taxa	Relative recombination rate	# genomes	Phenotype	Association method	Addresses accessory genome?	Unit of genetic variation studied	# of variants	Correction for population stratification
Sheppard et al. [1]	2013	*C. jejuni*	moderate	29 (+ validation in 161)	host specificity	allele counting	yes	30-bp DNA sequences (words)	>10,000 words (?)	simulation of word gain/loss along the phylogenetic tree
Farhat et al. [2]	2013	*M. tuberculosis*	low	123	antibiotic resistance	homoplasy counting	no	SNPs	~25,000	implicit in phylogenetic convergence criterion
Chen & Shapiro (This review)	2015	*M. tuberculosis*	low	123	antibiotic resistance	allele counting	no	SNPs	~3,000 MAF > 0.05)	inferred ancestry clusters
Laabei et al. [3]	2014	*S. aureus*	low	90	virulence	allele counting	no	SNPs & small indels	~3000	genomic control
Alam et al. [4]	2014	*S. aureus*	low	75	antibiotic resistance	allele counting and homoplasy counting	no	SNPs	~55,000	inferred ancestry clusters
Chewapreecha et al. [5]	2014	*S. pneumoniae*	high	3085 (+ validation in 616)	antibiotic resistance	allele counting	no	SNPs	~400,000 (MAF > 0.01)	inferred ancestry clusters
Salipante et al. [16]	2014	*E. coli*	low-moderate	312	antibiotic resistance	allele counting	yes	gene presence/absence	~15,000 genes	inferred ancestry clusters
Chaston et al. [17]	2014	41 strains	N/A	41	host development time and triglyceride content	allele counting	yes	gene presence/absence	~12,000 genes	consideration of genes with unique phylogenetic distributions
van Hemert et al. [15]	2010	*L. plantarum*	low	42	host immune response	allele counting	yes	gene presence/absence	? (CGH)	none

Allele counting and homoplasy counting approaches to GWAS

GWAS approaches for bacteria can be broadly broken down into allele counting [1**, 3**, 4**, 5**] and homoplasy counting [2**, 12] methods (Table 1 and Graphical Abstract). The primary association signal for allele counting methods is derived from an over-representation of an allele at the same site in cases relative to controls, which can later be corrected for population stratification. In contrast, homoplasy counting methods (in this case, phyC [2**]) derives its evidence of association by counting

repeated and independently emerged mutations occurring more often on branches of cases relative to controls. Homoplasy, as an indicator of convergent evolution, is a well-known signal of positive selection [28]. Combining this signal of selection with phenotypic associations (e.g. convergent mutations that occur only in cases and not in controls) provides the basis for homoplasy-based association tests.

Architecture of a strong association signal

GWAS signals from allele counting and homoplasy counting methods are not expected to perfectly overlap because each method represents different strengths and weaknesses. However, with a sufficiently large sample size, allele counting methods theoretically can detect all convergent sites (identified by homoplasy counting methods) as well as non-convergent sites. Still, ever-increasing sample size does not directly address the confounding effects of both population stratification and LD on allele counting methods. Homoplasy counting intrinsically accounts for these effects by virtue of its phylogenetic convergence criterion. In contrast, allele counting methods have no such phylogenetic requirement. Thus, a monophyletic group containing many cases with the same over-represented allele at the same site may provide a strong signal for allele counting while providing no signal for homoplasy counting. Conversely, homoplasy counting requires a smaller count of homoplasy events (versus allele counts) in order to reach statistical significance; thus, a relatively small sample size with a strong paraphyletic structure may provide homoplasy counting with a much stronger signal than allele counting.

A genome-wide association study of antibiotic drug resistance in *Mycobacterium tuberculosis*

To examine the impacts of clonal frames (strong LD) and population stratification, we performed a 'traditional' GWAS using PLINK on a population of 123 *M. tuberculosis* (MTB) genomes that had been previously analyzed by phylogenetic convergence (phyC) [2**]. Of the 123 strains, 47 (cases) are resistant to at least one antibiotic and 76 strains are sensitive to all antibiotics (controls). This dataset contains 11 'gold standard' experimentally-verified antibiotic resistance alleles, all of which were identified by phyC, along with 39 new phyC hits in nonsynonymous coding sites and intergenic regions, and 7 hits in synonymous sites. We chose this particular MTB dataset as it allows a comparison of the results from traditional GWAS and phyC, and also because MTB genomes possess extensive LD and strong population structure, making them challenging subjects for traditional GWAS.

Clonal frames and the resolution of GWAS signals

MTB is considered to be a highly clonal pathogen, with very little detectable recombination [18]. Consistent with this, we observe a clonal frame consisting of linked sites across the genome. This clonal frame is evident from the extensive genome-wide linkage (black or red in Figures 1a and b, respectively), interrupted by a few homoplasic sites (small white or black points, respectively) identified by the four-gamete test [19] or the D' measure [20] of linkage (Figure 1a and b). The r^2 measure [21] does not directly measure recombination or homoplasy, but rather how well the allelic state at one site in the genome can predict the allele present at another site. The r^2 analysis confirms that MTB has extensive genome-wide LD, posing a challenge to pinpointing causal variants (Figure 1c and d). Other more highly recombining bacteria, such as *Streptococcus pneumoniae* (Figure 1e) have less long-range LD and more localized, shorter LD blocks (black triangles near the horizontal axis), facilitating GWAS [5**]. Because the extent of

genome-wide linkage is unlikely to be known *a priori*, an important first step before performing a bacterial GWAS is to characterize LD, as illustrated here (Figure 1).

Figure 1. Patterns of linkage disequilibrium (LD) in bacterial genomes assessed by different metrics. The x-axis of heat maps (a-d) represents the physical position along the MTB genome; (e) shows the S. pneumoniae genome. Each square in the heat map represents a pairwise calculation of LD.

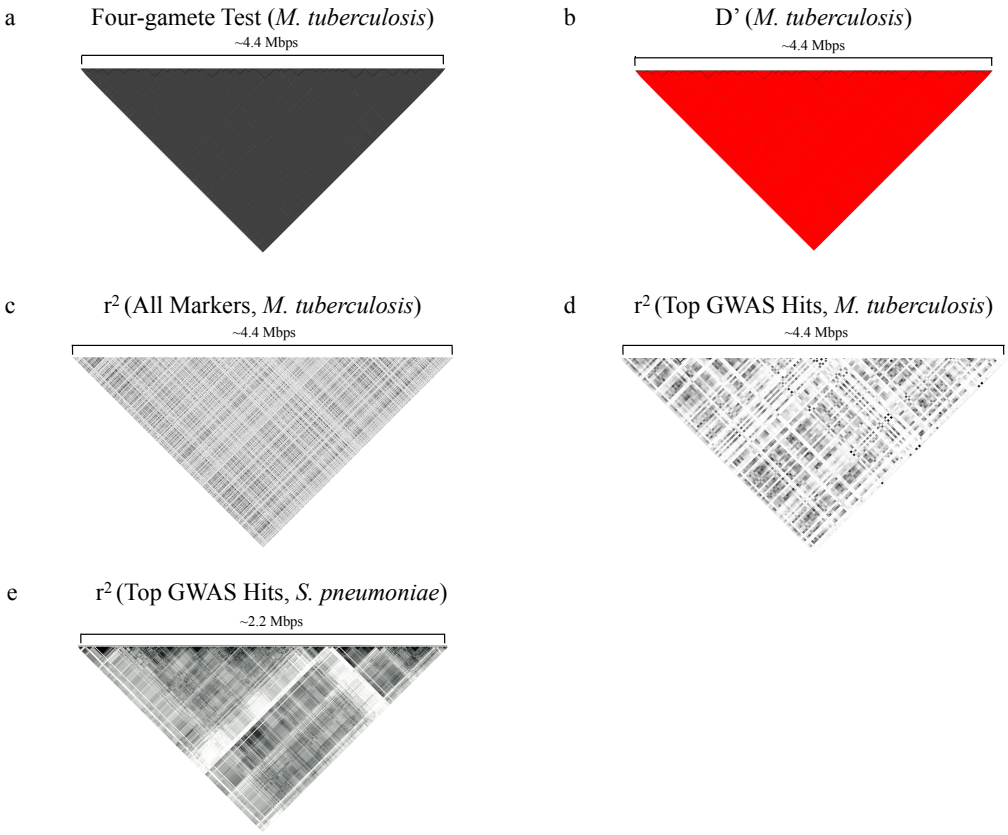
a) Four-gamete test. White squares denote four observed haplotypes indicating recombination may have occurred between the two sites. Black squares denote three or fewer observed haplotypes (strong linkage).

b) Pairwise |D'| measurements (range of |D'| values: $0 \leq |D'| \leq 1$). Red squares denote |D'| = 1 (strong linkage). Black squares denote |D'| < 1.

c) Pairwise r^2 measurements (range of r^2 values: $0 \leq r^2 \leq 1$). Black squares denote r^2 = 1 (strong correlation). The lighter squares denote progressively smaller r^2 values.

d) Pairwise r^2 measurements for the top 133 GWAS hits only. Black squares denote r^2 = 1. The lighter squares denote progressively smaller r^2 values.

e) Pairwise r^2 measurements of beta-lactam resistance associated variants co-detected in two separate *S. pneumoniae* populations [5**]. Black squares denote $r^2 = 1$. The lighter squares denote progressively smaller r^2 values.



Correcting for population stratification

The strong clonal nature of MTB also creates strong population substructures that in turn may lead to false positive associations. Without any population stratification correction we observe a substantial systematic inflation of the association test p-values (Figure 2a), likely due to both causal and non-causal resistance-associated mutations being linked on the same clonal frame. We assessed two classic methods of addressing population stratification. The first method, called genomic control [22], normalizes all p-values by a single inflation factor λ, which is the observed median chi-square divided by the expected median chi-square with 1 degree of freedom. Due to a relatively large observed inflation factor ($\lambda = 12.20$), genomic control seems to over-correct, leaving no statistically significant GWAS hits (Figure 2b). A less conservative

correction for population stratification is to infer ancestry by identifying genetic subpopulations within the overall population, and then testing for association conditional on these subpopulations. Subpopulations can be inferred using a variety of methods (e.g. multi-dimensional scaling in PLINK [23], principal component analysis in EIGENSTRAT[24], and BAPS [25]), then used as covariates in association testing (e.g. with the Cochran-Mantel-Haenszel test). Here, we defined subpopulations based on 14 previously defined MTB epidemiological clusters [2**]. Using these epi-clusters as covariates reduced the inflation factor to 1, suggesting that it effectively controls for population stratification (Figure 2c). Although this procedure clearly changes the Manhattan plot (Figure 2, right panels), producing more clean 'hits' that stand out from the average p-value, we note that none of these hits pass correction for multiple hypothesis testing. Therefore, correcting for population stratification can reduce GWAS power significantly – a problem that could potentially be overcome by using larger sample sizes (e.g. thousands rather than hundreds of genomes; [5**]).

Figure 2. GWAS for antibiotic resistance in MTB.

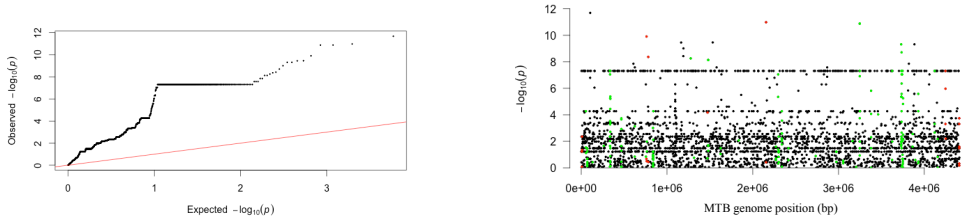
GWAS was performed using Plink version 1.07 [23]. The x=y line (red in QQ plots; left) represents the null hypothesis of no association signal. In Manhattan plots (right), SNPs in 'Gold Standard' resistance genes are shown in red, and SNPs in phyC candidate genes in green (excluding synonymous sites). Different corrections for population stratification were applied:

a) No population stratification correction.

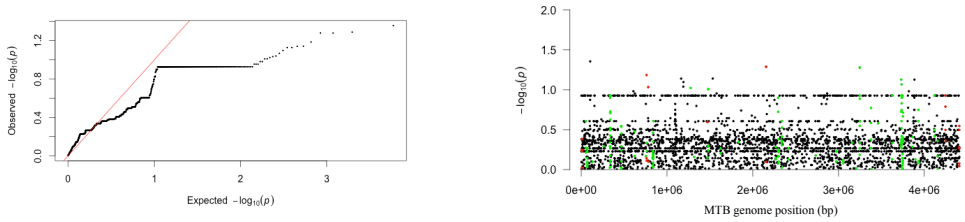
b) Population correction with genomic control.

c) Population correction using "epi-clusters" and Cochran-Mantel-Haenszel 2x2xK test, where K = 14 epi-clusters.

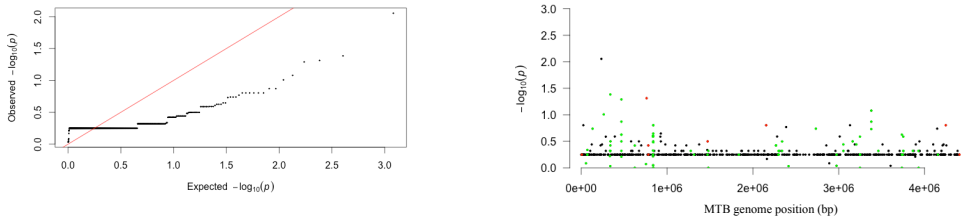
a No Population Stratification Correction



b Genomic Control



c Epi-clusters



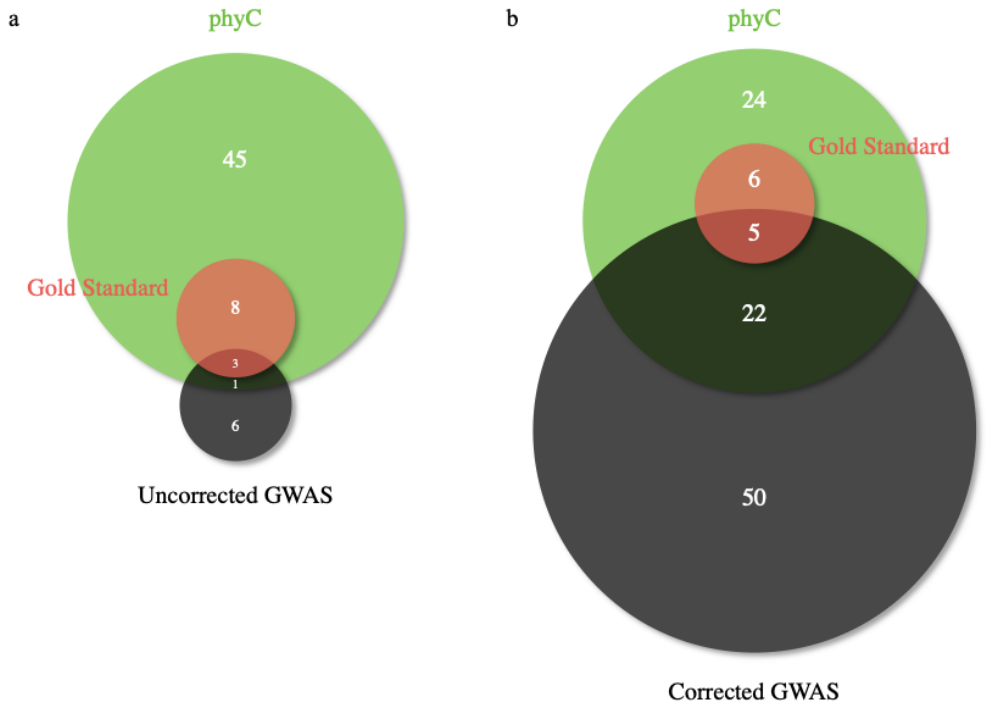
Comparison of GWAS against convergence testing

Despite the lack of significance after multiple testing correction, we identified 133 potential GWAS hits (SNPs) in 77 genes that stood out as outliers from the average genome-wide p-value (Figure 2c), which we will discuss for illustration purposes. These GWAS hits overlapped with 5 of 11 previously known 'gold standard' resistance genes and 22 of 46 additional phyC candidate resistance genes. It is also evident that correcting for population stratification improves the overlap with known resistance genes and phyC hits (Figure 3).

Figure 3. Overlap among GWAS candidates, phyC candidates, and "Gold standard" resistance genes. Numbers in Venn diagrams are in units of genes or intergenic regions.

a) No population stratification correction.

b) Population correction using "epi-clusters" and Cochran-Mantel-Haenszel 2x2xK test, where K = 14 epi-clusters.



Potential new drivers of drug resistance

Of these top 133 GWAS hits, 75 SNPs (in 50 genes) did not overlap with either known resistance genes nor with phyC candidates (Figure 3b). Due to long-range LD, it is not immediately clear without further analysis whether these 75 SNPs represent false

positives due to their correlation with the true drivers of resistance, either 'gold standard' resistance genes or phyC candidates. However, 15 out of these 75 SNPs were relatively uncorrelated ($r^2 < 0.3$) with any of the other 133 top GWAS hits, suggesting they could play causal roles in resistance phenotypes. As an example to illustrate the importance of assessing LD patterns around GWAS hits, the top GWAS hit (a nonsense mutation in an oxidoreductase gene, Rv0197) can be viewed from two different perspectives:

The top GWAS hit may be a false positive because it is in moderate correlation ($0.4 < r^2 < 0.5$) with two phyC candidates (PPE9 and PE_PGRS4 genes) and two other GWAS hits (PE-PGRS30 and PE-PGRS46 genes), and does not represent a true causal variant.

The top GWAS hit may be driving the association. Although it is in moderate correlation with four other phyC or GWAS hits, all four hits reside within the PE/PGRS families of genes, which are highly polymorphic and might represent false positive associations [2**].

Whether this GWAS hit is causal or not can only be firmly established with followup experiments.

Future Directions

We have shown the potential of GWAS for bacterial genomes while highlighting two key obstacles: long-range LD within the clonal frame and extensive bacterial population stratification both reduce our ability to pinpoint causal mutations with confidence. However, a third feature of bacterial genomes – the relative strength of positive selection – provides an opportunity to increase the resolution and confidence of GWAS hits. One could combine positive selection tests and GWAS, as has been done previously for traits shaped by positive selection [26, 27]. This approach may potentially address the problem of clonal frames obscuring true causal variants and

making them indistinguishable from linked non-causal variants. This idea attempts to identify causal variants in two steps:

perform a genome-wide selection scan identifying any genomic regions that are putatively under positive selection
perform a “targeted” association study on each genomic region under positive selection

The rationale here is that each genomic region identified as being under positive selection effectively “unlinks” the putative causal variants from its background clonal frame, provided that the selection test itself can distinguish a positive selection region from the clonal frame upon which it occurred [28]. Since positive selection alone does not provide sufficient evidence that a region is associated with the phenotype of interest, step two targets each of the genomic regions identified in the selection scan and tests each one for association with a phenotype of interest. In phyC, the two steps are done simultaneously, using convergence as the signal of positive selection and the *specificity* of convergence to cases but not controls as the association signal. Future work might 'mix and match' different signals of selection and association.

As this new and growing field develops, we envision a future where multiple genetic mapping approaches – including GWAS, phyC and selection scans – are combined. Each method may harbor its own strengths and weaknesses so that when combined, each method provides distinct information, thus increasing the power to detect true and causal associations.

Acknowledgements

This work was funded by the Natural Sciences and Engineering Research Council, the Canadian Institutes for Health Research and the Canada Research Chairs program. We

would like to thank Luis Barreiro and Jean-Baptiste Leducq for valuable feedback and discussions.

References

1. Sheppard, S. K., Didelot, X., Meric, G., Torralbo, A., Jolley, K. A., Kelly, D. J., et al. (2013). Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in *Campylobacter*. Proceedings of the National Academy of Sciences, 110(29), 11923–11927. doi:10.1073/pnas.1305559110

** This method is unique in that it simultaneously addresses both the core and accessory genomes.

2. Farhat, M. R., Shapiro, B. J., Kieser, K. J., Sultana, R., Jacobson, K. R., Victor, T. C., et al. (2013). Genomic analysis identifies targets of convergent positive selection in drug-resistant *Mycobacterium tuberculosis*. Nature Genetics, 45(10), 1183–1189. doi:10.1038/ng.2747

** Currently, the only GWAS method to explicitly combine evolutionary convergence with association.

3. Laabei, M., Recker, M., Rudkin, J. K., Aldeljawi, M., Gulay, Z., Sloan, T. J., et al. (2014). Predicting the virulence of MRSA from its genome sequence. Genome Research, 24(5), 839–849. doi:10.1101/gr.165415.113

** Despite a relatively low sample size (90 strains) and a low-recombining population, a > 85% predictive accuracy for a toxicity phenotype was achieved, thus highlighting the potential of GWAS.

4. Alam, M. T., Petit, R. A., Crispell, E. K., Thornton, T. A., Conneely, K. N., Jiang, Y., et al. (2014). Dissecting Vancomycin-Intermediate Resistance in *Staphylococcus aureus* Using Genome-Wide Association. Genome Biology and Evolution, 6(5), 1174–1185. doi:10.1093/gbe/evu092

** An example that pinpoints one resistance gene, rpoB, using both allele and homoplasy counting methods.

5. Chewapreecha, C., Marttinen, P., Croucher, N. J., Salter, S. J., Harris, S. R., Mather, A. E., et al. (2014). Comprehensive Identification of Single Nucleotide Polymorphisms Associated with Beta-lactam Resistance within Pneumococcal Mosaic Genes. PLoS Genetics, 10(8), e1004547. doi:10.1371/journal.pgen.1004547.s008

** An example of a high-powered, high-resolution study (3,085 strains) in populations with high levels of recombination.

6. Falush, D., & Bowden, R. (2006). Genome-wide association mapping in bacteria? Trends in Microbiology, 14(8), 353–355. doi:10.1016/j.tim.2006.06.003

** The first paper to broach the potential and challenges of GWAS in bacterial genomes.

7. Read, T. D., & Massey, R. C. (2014). Characterizing the genetic basis of bacterial phenotypes using genome-wide association studies: a new direction for bacteriology. Genome Medicine 6:109 doi:10.1186/s13073-014-0109-z
8. Milkman, R., & Bridges, M. M. K. (1990). Molecular evolution of the *Escherichia coli* chromosome. III. Clonal frames. Genetics, 126(3), 505–517.
9. Balding, D. J. (2006). A tutorial on statistical methods for population association studies. Nature Reviews Genetics, 7(10), 781–791. doi:10.1038/nrg1916
10. Gherman, A., Chen, P. E., Teslovich, T. M., Stankiewicz, P., Withers, M., Kashuk, C. S., et al. (2007). Population bottlenecks as a potential major shaping force of human genome architecture. PLoS Genetics, 3(7), e119. doi:10.1371/journal.pgen.0030119
11. Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., et al. (2009). Finding the missing heritability of complex diseases. Nature, 461(7265), 747–753. doi:10.1038/nature08494

12. Farhat, M. R., Shapiro, B. J., Sheppard, S. K., Colijn, C., & Murray, M. (2014). A phylogeny-based sampling strategy and power calculator informs genome-wide associations study design for microbial pathogens. Genome Medicine 6:101. doi:10.1186/s13073-014-0101-7
13. Lapierre, P., & Gogarten, J. P. (2009). Estimating the size of the bacterial pan-genome. Trends in Genetics : TIG, 25(3), 107–110. doi:10.1016/j.tig.2008.12.004
14. Vernikos, G., Medini, D., Riley, D. R., & Tettelin, H. (2015). Ten years of pan-genome analyses. Current Opinion in Microbiology, 23, 148–154. doi:10.1016/j.mib.2014.11.016
15. van Hemert, S., Meijerink, M., Molenaar, D., Bron, P. A., de Vos, P., Kleerebezem, M., et al. (2010). Identification of Lactobacillus plantarum genesmodulating the cytokine response of humanperipheral blood mononuclear cells. BMC Microbiology, 10(1), 293. doi:10.1186/1471-2180-10-293
16. Salipante, S. J., Roach, D. J., Kitzman, J. O., Snyder, M. W., Stackhouse, B., Butler-Wu, S. M., et al. (2014). Large-scale genomic sequencing of extraintestinal pathogenic *Escherichia coli* strains. Genome Research. doi:10.1101/gr.180190.114
17. Chaston, J. M., Newell, P. D., & Douglas, A. E. (2014). Metagenome-Wide Association of Microbial Determinants of Host Phenotype in Drosophila melanogaster. mBio, 5(5), e01631–14–e01631–14. doi:10.1128/mBio.01631-14
18. Smith NH, Gordon SV, de la Rua-Domenech R, Clifton-Hadley RS, Hewinson RG (2006). Bottlenecks and broomsticks: the molecular evolution of *Mycobacterium bovis*. Nature Reviews Microbiology, 4(9), 670–681. doi:10.1038/nrmicro1472
19. Hudson, R. R., & Kaplan, N. L. (1985). Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics, 111(1), 147–164.
20. Lewontin, R. C. (1964). The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. Genetics, 49(1), 49–67.
21. Hill, W. G. & Robertson, A. Linkage disequilibrium in finite populations. Theor. Appl. Genet. 38, 226–231 (1968).

22. Devlin, B., & Roeder, K. (1999). Genomic control for association studies. Biometrics, 55(4), 997–1004.
23. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. The American Journal of Human Genetics, 81(3), 559–575. doi:10.1086/519795
24. Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics, 38(8), 904–909. doi:10.1038/ng1847
25. Corander, J., Marttinen, P., Sirén, J., & Tang, J. (2008). Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. BMC Bioinformatics, 9(1), 539. doi:10.1186/1471-2105-9-539
26. Cheeseman, I. H., Miller, B. A., Nair, S., Nkhoma, S., Tan, A., Tan, J. C., et al. (2012). A Major Genome Region Underlying Artemisinin Resistance in Malaria. Science, 336(6077), 79–82. doi:10.1126/science.1215966
27. Karlsson, E. K., Harris, J. B., Tabrizi, S., Rahman, A., Shlyakhter, I., Patterson, N., et al. (2013). Natural selection in a bangladeshi population from the cholera-endemic ganges river delta. Science Translational Medicine, 5(192), 192ra86. doi:10.1126/scitranslmed.3006338
28. Shapiro, B. J., David, L. A., Friedman, J., & Alm, E. J. (2009). Looking for Darwin's footprints in the microbial world. Trends in Microbiology, 17(5), 196–204. doi:10.1016/j.tim.2009.02.002

Chapter 2: Classic genome-wide association methods are unlikely to identify causal variants in strongly clonal microbial populations

Chen PE, Jesse Shapiro B. Classic genome-wide association methods are unlikely to identify causal variants in strongly clonal microbial populations. bioRxiv. 2021. p. 2021.06.30.450606. doi:10.1101/2021.06.30.450606

Abstract

Since the advent of genome-wide association studies (GWAS) in human genomes, an increasing sophistication of methods has been developed for more robust association detection. Currently, the backbone of human GWAS approaches is allele-counting-based methods where the signal of association is derived from alleles that are identical-by-state. Borrowing this approach from human GWAS, allele-counting-based methods have been popularized in microbial GWAS, notably the generalized linear model using either dimension reduction for fixed covariates and/or a genetic relationship matrix as a random effect in a mixed model to control for population stratification. In this work, we show how the effects of linkage disequilibrium (LD) can potentially obscure true-positive genotype-phenotype associations (i.e., genetic variants causally associated with the phenotype of interest) and also lead to unacceptably high rates of false-positive associations when applying these classical approaches to GWAS in weakly recombining microbial genomes. We developed a GWAS method called POUTINE (https://github.com/Peter-Two-Point-O/POUTINE), which relies on homoplastic mutation to both clarify the source of putative causal variants and reduce likely false-positive associations compared to traditional allele counting methods. Using datasets of *M. tuberculosis* genomes and antibiotic-resistance phenotypes, we show that LD can in fact render all association signals from allele counting methods to be fully indistinguishable from hundreds to thousands of

sites scattered across an entire genome. These classic GWAS methods thus fail to pinpoint likely causal genotype-phenotype associations and separate them from background noise, even after applying methods to correct for population structure. We therefore urge caution when utilizing classical approaches, particularly in populations that are strongly clonal.

Introduction

To date, human genome-wide association studies (GWAS) have revealed meaningful evidence both about genetic architectures and the underlying pathways involved in human disease and other heritable traits. Microbial GWAS, while gaining popularity, is still in its infancy compared to human GWAS. Borrowing knowledge and methodology from human studies, microbial association studies are prone to similar pitfalls but also present new challenges due to the distinct and diverse population genetics of microbes [1]. Notably, population stratification and linkage disequilibrium (LD) present substantial impediments to identifying potential causal genotype-phenotype associations in strongly clonal populations. The first obstacle is a similar confounder seen in human GWAS, though microbial populations often exhibit a much higher magnitude of stratification owing to clonal descent. The second obstacle is the focus of this paper, and perhaps the greater obstacle of the two. Crucially, microbial populations exhibit both strong and long-range LD due to recombination mostly via relatively short gene conversion events rather than the process of crossing over, leaving large and potentially distant regions of the genome linked in a clonal frame [2].

The human haplotype map [3–5] exploited the block-like LD created by crossing over (i.e., recombination during meiosis in which homologous chromosomes exchange segments) to provide a shortcut where a relatively small subset of genome-wide markers within haplotype blocks could ‘tag’ other markers in LD as a proxy. The non-block-like structure of clonal frames seen in many microbial populations presents a situation that is the converse of that seen in the human haplotype map; where blocks allowed a shortcut to genome-wide coverage, clonal frames resemble one large haplotype block covering the length of entire microbial genomes, thus obscuring potential causal sites that are not distinguishable from the rest of the frame. This scenario is perhaps analogous to the fine-mapping problem where an attempt is made to clarify the source of the putative causal signal within an LD region [6], except in highly clonal populations the region to fine-map is the entire clonal frame. Population

genetic theory informs us that, in genomes exhibiting strong LD, only sites under convergent selection, which experience homoplasic mutations in independent lineages and thus not in LD with other sites, are likely to be distinguishable from other sites [1]. Yet, much of the recent literature on microbial GWAS tends to report results without explicit mention of the LD profiles of the top hits, making it difficult to assess which mutations or genes are more likely to be driving the association and how many others are likely associated by LD.

Currently, GWAS methods can be broadly broken down into two general approaches based upon the source of their primary association signal: allele counting and homoplasy counting [1]. Regardless of the association model used, all allele counting methods derive their signal from alleles that are identical-by-state, whereas homoplasy counting methods derive their signal strictly from alleles that are identical-by-state but not identical-by-descent, often called homoplasic, convergent, or parallel mutations that arise repeatedly and independently on different genetic backgrounds. To date, most microbial GWAS in the literature have primarily relied on the classical allele counting methods invented for human studies, notably generalized linear regression using either dimension reduction for fixed covariates and/or a genetic relationship matrix as a random effect in a mixed model to control for population stratification [7–9].

Here we build off of our earlier homoplasy counting association method [10] and describe a next-generation GWAS method, which we call POUTINE. We apply POUTINE to two datasets of *M. tuberculosis* genomes and antibiotic-resistance phenotypes and find that it identifies known resistance mutations while minimizing likely false-positive associations. Utilizing our new tool, we further explore a major question concerning the state of classic allele counting and its use for strongly clonal microbial populations: Do allele counting methods identify signals outside of convergent sites, and are these likely true or false-positive associations?

Results

A new homoplasy counter

We begin by describing key aspects of POUTINE, a GWAS method based on homoplasy counting, with specific details elaborated in the Methods. Homoplasies at each nucleotide position (site) are identified by finding all identical alleles that do not share a most recent common ancestor. Using only homoplasic mutations offers solutions to two of the most substantial obstacles in microbial GWAS. First, homoplasic mutations having arisen on independent genetic backgrounds are not linked to other sites in the genome. This feature allows homoplasies to naturally bypass the LD problem because convergent sites are by definition unlinked from the clonal frame and thus provide truly independent association signals. Second, there is no need to further correct for the confounding effect of population stratification, which arises due to genetic ancestry. By definition, homoplasic mutations are not identical-by-descent, and thus do not contribute to spurious associations caused by subpopulations where cases are on average more genetically related with each other than controls. Avoiding population stratification correction preserves statistical power that is otherwise potentially lost due to this correction.

Our goal with POUTINE was to develop a homoplasy-based GWAS method that is both robust and user-friendly. Currently, POUTINE input phenotypes are strictly discrete and binary. Different from the original phyC method [10], ancestrally reconstructed phenotypic states are avoided due to the noise they may potentially introduce to the association signal. With the growing scale of genome sequencing, larger sample sizes can compensate for excluding ancestral genotypes and phenotypes from the homoplasy counts. Genotypes are strictly from the core genome and only biallelic single nucleotide variants (SNVs) are currently tested for associations with a discrete phenotype using a binomial test. The background mutation rate across the genome can potentially vary, thus allowing some sites to have a higher expected level of homoplasies. The binomial test incorporates the total number of observed

homoplasies at each site, thus accounting for any varying background mutation rates. Critical to homoplasy counting is the robustness of the input tree topology as this directly determines which mutations are called homoplasic. The input ancestral phylogenetic reconstruction is optional; users may choose to precompute an ancestral reconstruction or opt for POUTINE to compute one based upon the topology of the input tree.

Homoplasic mutations are often taken as hallmarks of positive selection. The intuition is that adaptive mutations under positive selection will appear repeatedly in independent lineages experiencing the same selective pressure [11,12]. However, some baseline level of selectively neutral homoplasy is expected, and in the context of GWAS, some of these homoplasies will be spuriously associated with the phenotype under study. In POUTINE, we therefore establish an empirical null distribution (with no association between phenotype and genotype) by permuting the phenotypes and leaving the genotypic structures completely intact. We assess sites likely to be under convergent selection using Westfall and Young's max(T) resampling scheme [13]. This marks an additional improvement from the earlier phyC method. Principally, it produces a familywise error rate (FWER) that is far less conservative than methods that treat each hypothesis as being independent. This feature is particularly appropriate in strongly clonal populations where many regions are in complete LD and thus all sites in the region present as identical hypotheses. In addition, in the context of strongly clonal populations, the max(T) method is also more favorable than classic false discovery rate (FDR) based approaches. Methods to strongly control the FDR, such as Benjamini and Hochberg's step-up procedure [14] and Storey's q-value [15], were originally proposed under the assumption of independent tests. As such, these methods are unproven to strongly control the FDR in the face of pervasive dependence structures between sites. We note that there have been recent developments in relaxing this assumption to allow for increasingly arbitrary dependence structures [16–20].

As a further improvement upon phyC, POUTINE has been designed to handle much larger datasets. To scale to larger sample sizes in the tens of thousands of genomes, our implementation of POUTINE uses several optimizations including parallelizing the most time-consuming step, which is resampling. To further decrease runtime, the user may set a minimum homoplasy count to ignore sites with low counts that are unlikely to be statistically significant. We find this option to be helpful as many sites potentially have only one or two homoplasic mutations.

Benchmarking POUTINE on a test dataset

To empirically test the sensitivity and specificity of POUTINE, we reanalyzed a 'test' dataset of 123 *M. tuberculosis* genomes previously analyzed with phyC to identify convergent mutations associated with a broad resistance phenotype (defined as resistance to any anti-TB drug by conventional drug susceptibility testing) [10], and compared our new results to this reference set. Table S1 shows the six genome-wide significant hits from the reanalysis. All four genes previously identified in the literature as causal genes for drug resistance (*rpoB, embB, rpsL, rrs*) were re-identified as top hits, except for *rpsL.* The lack of signal for *rpsL* could be due to the relatively small sample size of 123 genomes and only four homoplasic mutations at this site, not including mutations in ancestrally reconstructed internal branches. By including internal branches, the original phyC may have gained power to identify associations in *rpsL*, but at the likely risk of additional false-positive associations. In particular, phyC identified associations in 16 PE/PPE genes in this dataset, which were reasoned to be false-positive associations [10], while the POUTINE reanalysis did not identify a single association in a PE/PPE gene. The family of PE/PPE genes in *M. tuberculosis* is known to be problematic for both sequencing and alignment due to their similarity and repetitive nature; thus they are prone to GWAS false positives [10]. Three additional genes were identified by POUTINE that were not seen in the previous analysis (Rv0853c, Rv0587, Rv1639c). An examination of the literature reveals that each of these three genes has been implicated in drug resistance in *M. tuberculosis* [21–23]. Overall, this real-world example suggests that POUTINE is similarly sensitive compared

to phyC (detecting likely true positives), and also more specific (reducing likely false positives in PE/PPE genes).

An LD perspective on allele and homoplasy counting signals

To compare the LD profiles of convergent vs. non-convergent GWAS hits, we analyzed a second 'discovery' dataset of 1330 *M. tuberculosis* genomes and an isoniazid drug-resistance phenotype [24]. To identify non-convergent sites, we used a standard allele-counting GWAS approach, implemented in PLINK using logistic regression with principal components as covariates to control for population stratification (Methods).

Using POUTINE, we identified three SNVs significantly associated with isoniazid resistance after correcting for multiple hypothesis testing, and another three 'secondary' hits which did not survive multiple test correction (max(T)-corrected P = 0.058; Table 1). For illustrative purposes, we consider these secondary hits because all three sites are one homoplasic case count away from genome-wide statistical significance, suggesting that a larger sample size or combining more drug-resistant phenotypes into one broad resistant phenotype would show these sites to be significant. In addition, two of the three secondary hits (sites 1674048 and 1674481) are in the same region as the primary hit at site 1673425, suggesting that a set-based test (i.e., combining mutation counts in the same gene or region) would likely identify these two secondary hits as genome-wide significant.

Table 1. Discovery set top hits (POUTINE + PLINK-only hits).

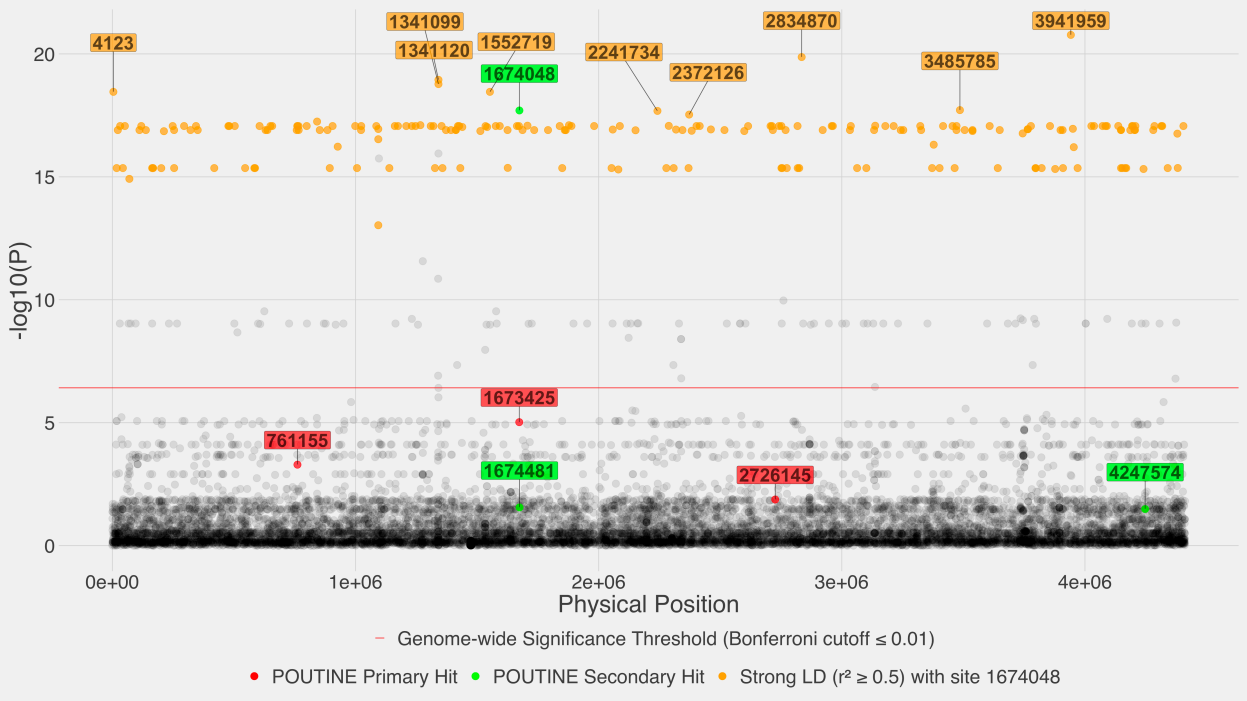
Physical Position	Gene Annotation	Homoplasy Counts [cases, controls]	Allele Counts [cases, controls]	max(T) FWER	Firth P-value	Odds Ratio (95% CI)
POUTINE Hits						
761155*	Rv0667 (*rpoB*)	[10, 0]	[17, 0]	1.00e-05	5.10e-04	170 (9-3073)
1673425*	intergenic: Rv1482c (hypothetical protein) - Rv1483 (*fabG1/mabA*)	[32, 0]	[42, 0]	1.00e-05	9.52e-06	586 (35-9841)
2726145*	intergenic: Rv2427A (*oxyR*) - Rv2428 (*ahpC*)	[5, 0]	[5, 0]	8.01e-03	1.33e-02	58 (2-1425)
1674048**✦	Rv1483 (*fabG1/mabA*)	[4, 0]	[50, 29]	5.80e-02	2.00e-18	89 (33-245)
1674481**	Rv1484 (*inhA*)	[4, 0]	[4, 0]	5.80e-02	2.79e-02	40 (1-1062)
4247574**	Rv3795 (*embB*)	[4, 0]	[4, 0]	5.80e-02	3.25e-02	35 (1-935)
PLINK-only Hits						
3941959	intergenic: Rv3511 (PE_PGRS55) - Rv3512 (PE_PGRS56)	[0, 0]	[46, 35]	1	1.69e-21	39 (18-84)
2834870	Rv2518c (*ldtB*)	[0, 0]	[42, 21]	1	1.35e-20	59 (25-140)
1341120	Rv1198 (*esxL*)	[2, 5]	[50, 47]	1	1.13e-19	23 (12-45)
1341099	Rv1198 (*esxL*)	[3, 5]	[52, 52]	1	1.69e-19	18 (10-33)
4123	Rv0003 (*recF*)	[0, 0]	[45, 25]	1	3.49e-19	80 (31-210)
1552719	Rv1379 (*pyrR*)	[0, 0]	[45, 25]	1	3.49e-19	80 (31-210)
3485785	Rv3120 (hypothetical protein)	[0, 0]	[46, 31]	1	1.93e-18	67 (26-173)
2241734	Rv1997 (*ctpF*)	[0, 0]	[45, 28]	1	2.08e-18	70 (27-183)
2372126	Rv2112c (*dop*)	[0, 1]	[46, 30]	1	2.94e-18	66 (26-171)

* Denotes primary POUTINE hit. ** Denotes secondary POUTINE hit. ✦ Denotes the only site that is both a POUTINE and PLINK hit. Cases refer to resistant phenotypes and controls to sensitive.

In contrast to POUTINE, it proved difficult to distinguish GWAS hits from background noise using PLINK, even with standard corrections for population stratification. Due to pervasive LD, a genome-wide significance threshold cannot be relied upon to identify a subset of plausible hypotheses because of the overwhelming numbers of false positives that LD drags below this threshold. Consider a Bonferroni-corrected *P*-value cutoff of 0.01 (red line in Figure 1). Even using such a conservative threshold would still include too many false-positive associations and would leave the investigator with 288 genome-wide significant PLINK hits – a daunting number to consider for experimental follow-up. Manhattan plots of strongly clonal populations often feature groupings of sites in strong LD that we refer to as 'LD frames', visible as horizontal lines of sites with near-identical *P*-values (Figure 1). We refer to these plots as Montreal plots to reflect the low, horizontal skyline in contrast to the vertical skyscrapers of Manhattan.

For our purpose of examining LD profiles of allele counting hits, we arbitrarily chose the 10 sites visually distinguishable above the top LD frame as the top PLINK hits (Figure 1). Of these 10, only site 1674048 overlaps with a secondary POUTINE hit (Table 1). The careful reader will note that the top PLINK hit by raw allele counts (42 cases to 0 controls in Table 1) should be site 1673425 but this is not the case in Figure 1. Due to the phenomenon of complete statistical separation at this site (caused by allele counts showing all cases and zero controls) as well as a small sample size at the minor allele, even the Firth correction used does not sufficiently improve the accuracy of the significance estimate (note in Table 1 the wide confidence intervals around the estimated effect size). However, this problem is avoided when using an exact test such as Fisher's exact test, for which we do see site 1673425 as the top hit (Figure S2). For the purposes of the further analyses below, this detail does not affect our general conclusions.

Figure 1. Montreal plot of the discovery set with correction for population stratification. *P*-values along the y-axis were calculated using PLINK's logistic regression with the Firth correction. Population stratification was corrected using the first four principal components. The x-axis shows the nucleotide positions along the *M. tuberculosis* genome.

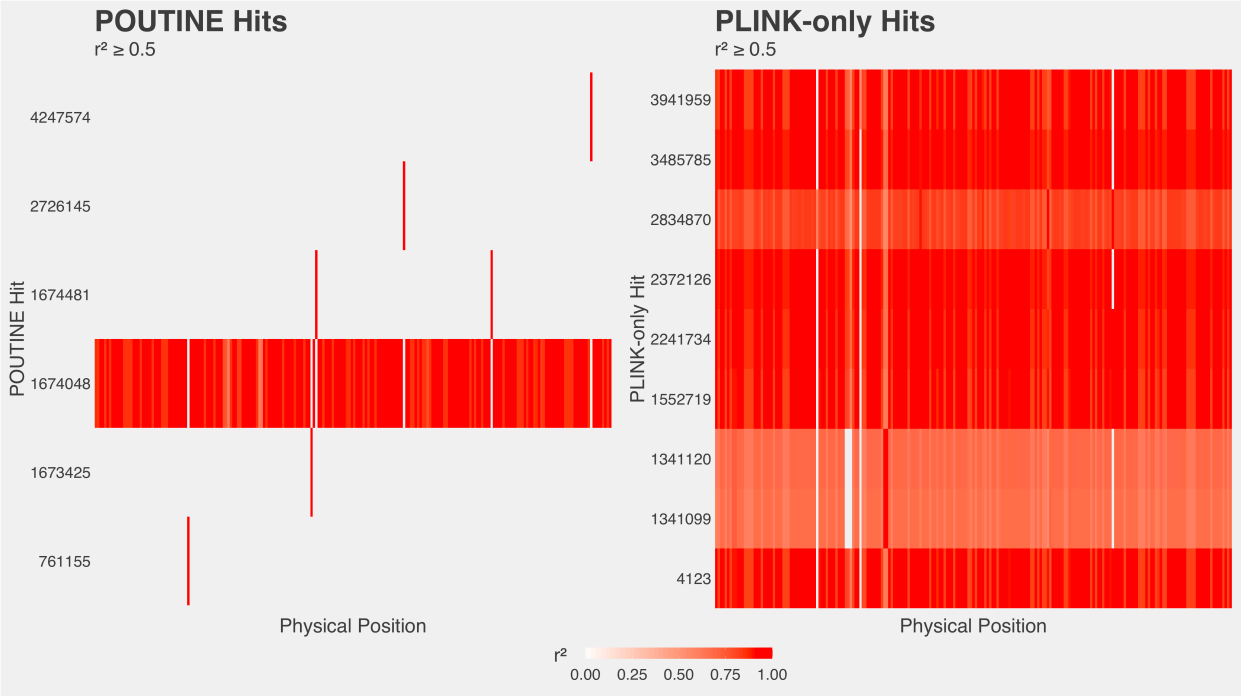


To further investigate the effects of linkage on GWAS hits, we compared how PLINK-only hits were linked to other sites in the genome to the six POUTINE hits. To focus on relatively strong LD, we plot sites across the genome that are linked to a top GWAS hit with a threshold of r^2 >= 0.5 (Figure 2). The complete distribution of r^2 values is shown in Supplementary Figure S1. Contrasting the two sets of LD profiles shows that the homoplasy-counting-based POUTINE signals are predominately free of strong linkage from each other and from the rest of the genome. Conversely, the allele-counting hits show strong linkage to sites throughout the genome. The only POUTINE hit that shows a similar LD pattern to the set of PLINK-only hits is position 1674048. This is the only POUTINE hit where the mutations are predominantly non-homoplasic; there are 79 minor alleles at this site, only four of which are homoplasies (Table 1). The other

five POUTINE hits are at sites that consist of entirely or predominantly homoplasies. These five hits suggest that sites composed of predominantly homoplasic mutations can be considered the sole source of an association signal, i.e., there are no other sites in strong LD that can be driving this signal, or hitchhiking along with a causal association. In contrast, the associations at site 1674048 and the set of nine PLINK-only hits cannot be disentangled from linked mutations across the genome (Figure 2). Because these hits show strong linkage to each other and to many sites across the entire genome, it is unclear how many and which of these sites are driving the association signal.

To determine what might be driving the POUTINE hit at site 1674048, we identified all the sites to which it is strongly linked ($r^2 >= 0.5$). Strikingly, all sites in strong LD with site 1674048 (orange points in Figure 1) include the entire top LD frame and all nine PLINK-only hits. Among these 10 potential hits, site 1674048 is the only one with a known causal mutation (a silent mutation that confers isoniazid resistance) in the literature [25]. It, therefore, seems likely that this site is driving the association signal, with the other sites being associated due to linkage.

Figure 2. Genome-wide LD (r^2 >= 0.5) of top hits (POUTINE + PLINK-only hits).

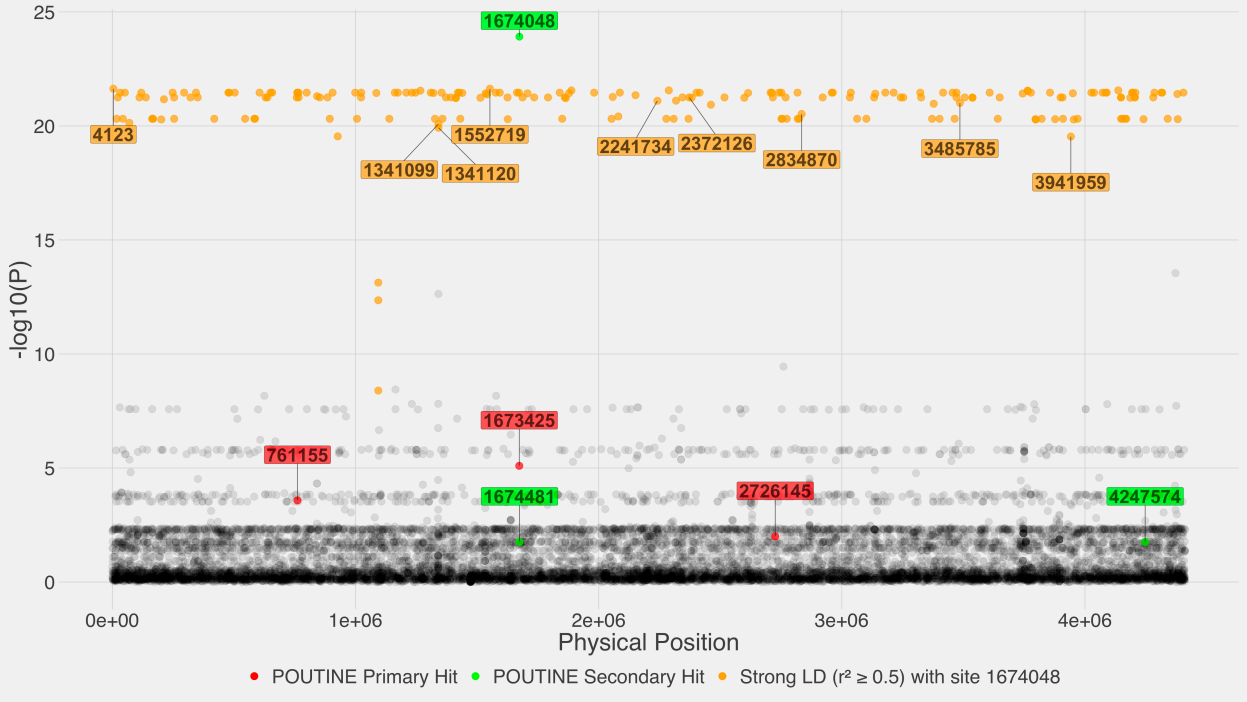


What would the association results look like if one relied only on classical allele counting methods (i.e., all homoplasy information is removed from the above analyses)? First, one would mistakenly identify 9 of the 10 top hits as plausible candidates of association to isoniazid resistance, with site 1674048 being the only known causal mutation. Even upon further examination of LD for all 10 sites, one is left with no meaningful signal because each top hit is strongly linked to sites across the entire genome; the likely causal site 1674048 does not stand out in any identifiable way. Second, one would miss all five other POUTINE hits (sites 761115, 1673425, 1674481, 2726145, 4247574). Since these five sites are predominantly homoplasic, it is possible that with increased sample sizes allele counting would identify them, provided that the increase in sample size also increases the minor allele frequencies at these sites. Only then could further examination of LD reveal that these five sites are not strongly linked to one another nor to the rest of the genome, and thus be considered as plausible candidates.

Population stratification corrections do not break up LD

As an alternative to homoplasy counting, allele-counting GWAS methods are typically corrected in an attempt to remove the confounding effect of population stratification. Regardless of the population stratification control used, these corrections effectively reweight statistical significance at each affected site and thus only shifts the site up or down along the y-axis of a Manhattan plot. Crucially, these corrections say nothing about the correlational structures between sites. Consider how the Manhattan plot changes when one removes the stratification control; here we simply removed all principal components used in the regression model (Figure 3). The nine PLINK-only hits identified with stratification correction (Figure 1; Table 1) are no longer discernible from the top LD frame, leaving only site 1674048 as the lone top hit which rises above the top LD frame in the absence of correction (Figure 3). Therefore, stratification correction of an allele counting GWAS identifies eight additional hits compared to an uncorrected approach, none of which overlap with homoplasy counting hits. Without LD information, one is susceptible to being misled into thinking that these allele counting hits are distinguishable from the top LD frame, when in fact their LD profiles say otherwise (Figure 2). In summary, stratification correction alone cannot bypass the confounding effects of LD.

Figure 3. Montreal plot of the discovery set without correction for population stratification. The analysis is identical to Figure 1, except no population stratification correction was done.



Discussion

The state of allele counting methods for strongly clonal populations

To demonstrate the impediment LD presents to classical GWAS methods, we chose *M. tuberculosis*, a clonal species that features strong and long-range LD across the entire genome. We sought to answer a critical question regarding the utility of homoplasy- and allele-counting GWAS approaches: Do allele counting methods meaningfully identify signals outside of sites evolving under convergent evolution? The answer to this question is likely no. Although allele counting methods may produce strong GWAS signals for causal variants that are not convergent, those signals are indistinguishable from other similarly strong signals in linked sites. Crucially, the vast majority of these

linked sites – which can easily number in the hundreds or thousands – are likely false positives. The LD frames that represent groupings of strongly linked sites feature prominently in Manhattan plots of strongly clonal microbial populations. In such populations, Manhattan plots are more appropriately called Montreal plots because they more closely resemble the relatively flat skyline of a city like Montreal (where regulations prevent buildings taller than its namesake Mount Royal, with an elevation of only 233 m) rather than the skyscrapers punctuating the Manhattan skyline.

Is LD friend or foe?

The implications of these findings can be extended to other clonal populations, and likely to many other populations across the gamut of microbial recombination rates. It is unclear if there exist microbial populations with a sufficiently high rate of recombination to exhibit similar block-like LD structures seen in eukaryotic genomes. If so, these populations would be more amenable to allele counting methods. We note that there do exist recombination hotspots in bacterial species, such as *E. coli* [26], *C. jejuni* [27], and others [28]. These distinct regions are relatively unlinked from the rest of the genome, and as such can potentially provide a cleaner signal of association. In one such notable example, Chewapreecha *et al.* analyzed a highly recombining *S. pneumoniae* population to identify six common recombination hotspots [29]. A follow-up GWAS applying an allele-counting approach (specifically, the Cochran-Mantel-Haenszel test using population clusters identified with BAPS [30]) on beta-lactam antibiotic resistance in this population identified plausible association signals in genes encoding penicillin-binding proteins and involved in peptidoglycan synthesis – which tended to reside within one of the six common recombination hotspots [31]. It remains to be seen if allele-counting GWAS approaches can identify hits outside of such hotspots.

A viral population of current substantial interest is SARS-CoV-2. This population offers a timely example of the ramifications of our findings. This population is strongly clonal with evidence showing that there has been little realized recombination, although the

potential for recombination is present [32]. Intriguingly, many of the variants of concern (VOCs) identified thus far harbor mutations (e.g., E484K in the spike gene) that are thought to be under convergent selection [33]. This highlights that even potentially recombining populations may be effectively clonal at the early stages of an outbreak (or pandemic), at which time homoplasy-counting methods are likely to be much more effective than allele-counting to identify genotype-phenotype associations.

Limitations of POUTINE and future directions

Homoplasy-based methods such as POUTINE provide a promising lifeline for tackling association studies in microbial populations. Its major limitation is that if the causal variants to discover are not convergent mutations, then there simply will be no signal to discover. It is unclear at this time what proportion of causal variants are sculpted by convergent evolution. In addition, it is also unclear how much genetic heterogeneity underlies the genetic architecture of many microbial traits. Both locus and allelic heterogeneity can dilute the association signal, requiring higher sample sizes to recapitulate any signal. To address allelic heterogeneity, we plan on adding a set-based test to aggregate individual variants into localized regions to boost the signal. Currently, our initial implementation assays biallelic SNVs inside the core genome. As such, POUTINE excludes sources of variation including tri/quad-allelic sites, small indels, and the accessory genome. In the future, it would be straightforward to include tri/quad-allelic sites using a multinomial test instead of the binomial. We note that it is currently possible to recode loci in the accessory genome as present or absent among the population and run POUTINE as if these recoded loci were core SNVs. When taking this approach, one should proceed with caution because the recombination dynamics of the accessory genome may differ from the core. Lastly, a further limitation of homoplasy-based methods can be their inability to identify hemiplasies from homoplasies. A hemiplasy is a form of incomplete lineage sorting that can mask as a homoplasy [34,35]. If hemiplasies were mistakenly identified as homoplasies, it would no longer be appropriate to consider a homoplasy-based hit to be free from linkage to other sites. As such, sites composed of hemiplasies would present the same problem

to homoplasy counting methods as that seen in non-convergent sites for allele counting methods.

This work highlights the necessity to both examine and report the LD profiles of top association signals. However, the importance of examining LD should not steer the reader away from utilizing prior knowledge of their phenotype of interest, as such knowledge can play a clarifying role in narrowing down hypotheses. However, lesser understood phenotypes serve as prime targets for the agnostic view of GWAS to reveal underlying mutations in genes and other loci we know nothing or little about.

Because many causal variants may be hiding in non-convergent sites, it is critical that we understand if allele counting methods can provide meaningful association signals in the face of pervasive LD observed across microbial populations. For strongly clonal populations that are not amenable to allele counting approaches, we must improve upon these classical methods, and if these methods prove intractable to LD, we must open a new line of inquiry perhaps beyond homoplasy-based solutions in hopes of capturing non-convergent causal variants. Until such time, we urge caution when using classical GWAS methods to tackle microbial populations, particularly those with little measurable recombination.

Methods

Sample collection, genotyping, and phenotyping

The reference set of 123 *Mycobacterium tuberculosis* genomes comprises 14 major phylogenetic clusters from different micro-epidemics and 23 geographically diverse drug-sensitive isolates. Of the 123 isolates, 47 were resistant to one or more antibiotics. Isolate selection, sequencing, variant calling, and phenotyping are described in detail in [10].

The discovery set of 1330 *M. tuberculosis* genomes (GenBank BioProject accession: PRJNA413593) includes isolates collected by the British Columbia Public Health Laboratory of the British Columbia Centre for Disease Control. Isolate collection and phenotypic drug susceptibility testing are detailed in [24]. Samples were sequenced on the Illumina HiSeq2500 platform at the Michael Smith British Columbia Genome Sciences Centre using 125-bp paired-end reads [36]. All reads were quality checked using FastQC [37], trimmed with Trimmomatic [38], and mapped against the H37Rv reference genome (GenBank Reference Sequence accession: NC000962.2) using BWA-MEM [39]. All variants were called using GATK [40] requiring a Phred quality score > 20 and read depth > 5. Variants were further filtered out if a site had a missing call rate > 10% of samples, and only biallelic SNVs were kept.

Allele counting

Both PLINK 1.9 and 2.0 [41] were used for both preprocessing and association testing described below. All PLINK analyses were done with PLINK version 1.90b6.21 64-bit (19 Oct 2020) except for firth regression which used PLINK version 2.00a3 AVX2 (28 Mar 2021). For all runs, --chr-set -1 was used to designate a single haploid genome.

Population substructures

To capture population substructures, LD pruning was done in two ways: 1) whole-genome pruning using one window including all markers, and using the PLINK option -

-indep-pairwise with varying r^2 thresholds of 0.50, 0.90, 0.99; 2) local pruning using non-overlapping windows of 1000 markers at a time, and using --indep-pairwise with an r^2 threshold of 0.99. Principal component analyses were run on the above LD-pruned sets and once without any LD pruning using --pca 20 header var-wts (Figure S3).

Associating testing

Fisher's exact testing was run with a mid-p correction and 10^7 permutations using PLINK options --assoc fisher-midp mperm=10000000, and filtering out sites with a minor allele count < 3, sites with a missing call rate > 0.10, and samples with a missing call rate > 0.10 using --mac 3, --geno 0.1, --mind 0.1, respectively.

Logistic regression with a Firth penalty was conducted using --glm firth cols=+a1count,+totallele,+a1countcc,+a1countcc,+totallelecc,+a1freq,+a1freqcc. Firth correction serves as a penalty during maximum likelihood estimation to avoid non-convergence issues due to statistical separation [42].

To control for population stratification, --covar was used with the LD-pruned set, described above, which was calculated with non-overlapping windows of 1000 markers. The first four principal components were selected as fixed covariates using --covar-name PC1-PC4 (Figure S4). Sites were removed from Firth regression using --mac 4, --geno 0.1, and --mind 0.1. Multiple-testing correction reports were generated using --adjust. 95% confidence intervals around each odds ratio were reported using --ci 0.95.

Linkage disequilibrium

r^2 was calculated using PLINK options --r2, --ld-snps followed by sample ID names, --ld-window 69722 and --ld-window-kb 5000 to remove default settings to allow all

genome-wide pairs of markers conditional on the specified sites in --ld-snps, and --ld-window-r2 0 to report all r^2 values.

Homoplasy counting

All steps except for the building of an input tree are implemented in POUTINE, the method introduced here.

Input tree

Phylogenies were inferred using both FastTree [43] version 2.1.11 double precision (No SSE3) and raxml-ng [44] version 1.0.2. FastTree was run with -nt -gtr settings, and raxml-ng was run with --model GTR+G settings. To improve the inference of the tree topology, likely homoplasic regions (39 genes previously associated with drug resistance from [45]), as well as repetitive regions (e.g., PE/PPE and PGRS genes), were filtered out (273 genes; 10% of the genome; genes listed in [46]). Both tree topologies were effectively identical and did not affect the final homoplasy-based hits (Figure S5). Newick format parsing was done using the Coevolution library [47].

Homoplasy identification

Ancestral genotypic reconstructions were done using TreeTime [48] version 0.7.6. The ancestral subcommand was used with the --gtr infer settings and used the default joint maximum likelihood method. To identify homoplasic mutations, ancestral mutations were mapped onto the input tree, and a homoplasic event was called if at least two identical mutations/alleles at the same segregating site occurred on independent genetic backgrounds, i.e., the two mutations do not share a most recent common ancestor.

Association model

Associations of homoplasic mutations to the phenotype of interest is conducted using a binomial test using the binomial distribution defined as binomial(n, p, x), where n is

the total number of homoplasic mutations observed at each site, p is the probability of success of each Bernoulli trial and is equal to the case to control ratio for all sites, and x is the count of homoplasic mutations in cases at each site. We assume that the mutation rate at a particular site is a proxy for the background rate of expected homoplasic mutations at that site. Considering n at each site helps address any possible variation in the background mutation rate across the genome, and also satisfies the exchangeability principle between sites to allow for resampling of permutations. The implementation of the binomial test is from the Apache Math Commons 3.6.1 library (https://commons.apache.org/proper/commons-math/).

Significance assessment

P-values are derived using a resampling-based multiple-testing method by Westfall & Young [13]. This approach is sometimes referred to as max(T) for the maximum test statistic. max(T) is a single-step adjustment method:

$$\tilde{p}_i = \Pr\left(\max_{1 \le j \le k} |T_j| \ge |t_i| \,\middle|\, H_0\right)$$

$\tilde{p}_i$ is the probability that the largest test statistic in the resampled data set, T_j, is larger than the observed test statistic, t_i, given that all null hypotheses, H_0, are true. k is the number of tests.

Specifically, for each site a familywise error rate (FWER) is calculated as follows:

1) pointwise estimate: for each replicate, only phenotypes are permuted while genotypes and their dependency structures are left unmodified and completely intact. For each site, a resampled null distribution of no association is constructed from m replicates, where m is defined as the user-specified total number of replicates/permutations. The pointwise estimate is defined as $(r_{point} + 1)/(m + 1)$ [49], where r_{point} is the number of replicates equal to or more extreme than the observed binomial test statistic.

2) familywise estimate: for each replicate, the maximum test statistic is saved, and a second null distribution is constructed using all m maximum test statistics across all replicates. The familywise estimate is defined as $(r_{family} + 1)/(m + 1)$, where r_{family} is the total number of maximum test statistics equal to or more extreme than the observed binomial test statistic.

Availability of POUTINE

Download and installation instructions for POUTINE are available at GitHub: https://github.com/Peter-Two-Point-O/POUTINE.

Acknowledgments

We would like to thank Jennifer Gardy and Ben Sobkowiak for their assistance with the discovery set (GenBank BioProject accession: PRJNA413593). We further thank James Tanner, Len Taing, Arnaud N'Guessan, and Gavin Douglas for alpha testing our software and also for helpful discussions and valuable feedback.

References

1. Chen PE, Shapiro BJ. The advent of genome-wide association studies for bacteria. Curr Opin Microbiol. 2015;25: 17–24.

2. Milkman R, Bridges MMK. Molecular evolution of the Escherichia coli chromosome. III. Clonal frames. Genetics. 1990;126: 505–517.

3. International HapMap Consortium. The International HapMap Project. Nature. 2003;426: 789–796.

4. International HapMap Consortium. A haplotype map of the human genome. Nature. 2005;437: 1299–1320.

5. International HapMap Consortium, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, et al. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007;449: 851–861.

6. Schaid DJ, Chen W, Larson NB. From genome-wide associations to candidate causal variants by statistical fine-mapping. Nat Rev Genet. 2018;19: 491–504.

7. Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet. 2005;38: 203–208.

8. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38: 904–909.

9. Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, et al. Efficient Control of Population Structure in Model Organism Association Mapping. Genetics. 2008;178: 1709–1723.

10. Farhat MR, Shapiro BJ, Kieser KJ, Sultana R, Jacobson KR, Victor TC, et al. Genomic analysis identifies targets of convergent positive selection in drug-resistant Mycobacterium tuberculosis. Nature Publishing Group. 2013;45: 1183–1189.

11. Shapiro BJ, David LA, Friedman J, Alm EJ. Looking for Darwin's footprints in the microbial world. Trends Microbiol. 2009;17: 196–204.

12. Lieberman TDTD, Michel J-BJB, Aingaran MM, Potter-Bynoe GG, Roux DD, Davis MRMR, et al. Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes. Nat Genet. 2011;43: 1275–1280.

13. Westfall PH, Stanley Young S. Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment. John Wiley & Sons; 1993.

14. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society: Series B (Methodological). 1995. pp. 289–300. doi:10.1111/j.2517-6161.1995.tb02031.x

15. Storey JD. A direct approach to false discovery rates. J R Stat Soc Series B Stat Methodol. 2002;64: 479–498.

16. Benjamini Y, Yekutieli D. The Control of the False Discovery Rate in Multiple Testing under Dependency. Ann Stat. 2001;29: 1165–1188.

17. Ge Y, Sealfon SC, Speed TP. SOME STEP-DOWN PROCEDURES CONTROLLING THE FALSE DISCOVERY RATE UNDER DEPENDENCE. Stat Sin. 2008;18: 881–904.

18. Brzyski D, Peterson CB, Sobczyk P, Candès EJ, Bogdan M, Sabatti C. Controlling the Rate of GWAS False Discoveries. Genetics. 2017;205: 61–75.

19. Korthauer K, Kimes PK, Duvallet C, Reyes A, Subramanian A, Teng M, et al. A practical guide to methods controlling false discoveries in computational biology. doi:10.1101/458786

20. Fithian W, Lei L. Conditional calibration for false discovery rate control under dependence. arXiv [stat.ME]. 2020. Available: http://arxiv.org/abs/2007.10438

21. Raman K, Yeturu K, Chandra N. targetTB: a target identification pipeline for Mycobacterium tuberculosis through an interactome, reactome and genome-scale structural analysis. BMC Syst Biol. 2008;2: 109.

22. Danelishvili L, Wu M, Young LS, Bermudez LE. Genomic approach to identifying the putative target of and mechanisms of resistance to mefloquine in mycobacteria. Antimicrob Agents Chemother. 2005;49: 3707–3714.

23. Hang NTL, Le Hang NT, Hijikata M, Maeda S, Thuong PH, Ohashi J, et al. Whole genome sequencing, analyses of drug resistance-conferring mutations, and correlation with transmission of Mycobacterium tuberculosis carrying katG-S315T in Hanoi, Vietnam. Scientific Reports. 2019. doi:10.1038/s41598-019-51812-7

24. Guthrie JL, Kong C, Roth D, Jorgensen D, Rodrigues M, Hoang L, et al. Molecular Epidemiology of Tuberculosis in British Columbia, Canada: A 10-Year Retrospective Study. Clin Infect Dis. 2018;66: 849–856.

25. Ando H, Miyoshi-Akiyama T, Watanabe S, Kirikae T. A silent mutation in mabA confers isoniazid resistance on Mycobacterium tuberculosis. Mol Microbiol. 2014;91: 538–547.

26. Didelot X, Meric G, Falush D, Darling AE. Impact of homologous and non-homologous recombination in the genomic evolution of Escherichia coli. BMC

Genomics. 2012;13: 256.

27. Yahara K, Didelot X, Ansari MA, Sheppard SK, Falush D. Efficient Inference of Recombination Hot Regions in Bacterial Genomes. Mol Biol Evol. 2014. doi:10.1093/molbev/msu082

28. Yahara K, Didelot X, Jolley KA, Kobayashi I, Maiden MCJ, Sheppard SK, et al. The Landscape of Realized Homologous Recombination in Pathogenic Bacteria. Mol Biol Evol. 2016;33: 456–471.

29. Chewapreecha C, Harris SR, Croucher NJ, Turner C, Marttinen P, Cheng L, et al. Dense genomic sampling identifies highways of pneumococcal recombination. Nat Genet. 2014;46: 305–309.

30. Corander J, Marttinen P, Sirén J, Tang J. Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. BMC Bioinformatics. 2008;9: 539.

31. Chewapreecha C, Marttinen P, Croucher NJ, Salter SJ, Harris SR, Mather AE, et al. Comprehensive Identification of Single Nucleotide Polymorphisms Associated with Beta-lactam Resistance within Pneumococcal Mosaic Genes. PLoS Genet. 2014;10: e1004547.

32. VanInsberghe D, Neish AS, Lowen AC, Koelle K. Recombinant SARS-CoV-2 Genomes Circulated at Low Levels Over The First Year of The Pandemic. Virus Evolution. 2021. doi:10.1093/ve/veab059

33. Martin DP, Weaver S, Tegally H, San EJ, Shank SD, Wilkinson E, et al. The emergence and ongoing convergent evolution of the N501Y lineages coincides with a major global shift in the SARS-CoV-2 selective landscape. doi:10.1101/2021.02.23.21252268

34. Mendes FK, Livera AP, Hahn MW. The perils of intralocus recombination for inferences of molecular convergence. Philos Trans R Soc Lond B Biol Sci. 2019;374: 20180244.

35. Lee KM, Coop G. Population genomics perspectives on convergent adaptation. Philos Trans R Soc Lond B Biol Sci. 2019;374: 20180236.

36. Romanowski K, Sobkowiak B, Guthrie JL, Cook VJ, Gardy JL, Johnston JC. Using Whole Genome Sequencing to Determine the Timing of Secondary Tuberculosis in British Columbia, Canada. Clin Infect Dis. 2020. doi:10.1093/cid/ciaa1224

37. Andrews S. FastQC: a quality control tool for high throughput sequence data. Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom; 2010. Available: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

38. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina

sequence data. Bioinformatics. 2014;30: 2114–2120.

39. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v2. 2013. Available: http://arxiv.org/abs/1303.3997

40. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20: 1297–1303.

41. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4: 559.

42. Firth D. Bias Reduction of Maximum Likelihood Estimates. Biometrika. 1993;80: 27–38.

43. Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One. 2010;5: e9490.

44. Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics. 2019;35: 4453–4455.

45. Sandgren A, Strong M, Muthukrishnan P, Weiner BK, Church GM, Murray MB. Tuberculosis drug resistance mutation database. PLoS Med. 2009;6: e2.

46. Comas I, Chakravartti J, Small PM, Galagan J, Niemann S, Kremer K, et al. Human T cell epitopes of Mycobacterium tuberculosis are evolutionarily hyperconserved. Nat Genet. 2010;42: 498–503.

47. Yip KY, Patel P, Kim PM, Engelman DM, McDermott D, Gerstein M. An integrated system for studying residue coevolution in proteins. Bioinformatics. 2008;24: 290–292.

48. Sagulenko P, Puller V, Neher RA. TreeTime: Maximum-likelihood phylodynamic analysis. Virus Evol. 2018;4: vex042.

49. Belinda Phipson GKS. Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. arXiv:1603.05766v1 [stat.AP]. 2016. Available: https://arxiv.org/abs/1603.05766

Supplementary material

Table S1. All six genome-wide significant POUTINE hits in the reference set of 123 *M. tuberculosis* genomes.

Physical Position	Gene Annotation	Homoplasy Counts [cases, controls]	max(T) FWER
1473246*	Rvnr01 (rrs)	[9,0]	8.89e-04
761155*	Rv0667 (rpoB DNA-directed RNA polymerase subunit beta)	[8,0]	6.30e-03
949535**	Rv0853c (pdc alpha-keto-acid decarboxylase)	[8,0]	6.30e-03
4247429*	Rv3795 (embB arabinosyltransferase B)	[7,0]	2.90e-02
685461**	Rv0587 (yrbE2A hypothetical protein: ABC-type transporter Mla maintaining outer membrane lipid asymmetry, permease component MlaE [Cell wall/membrane/envelope biogenesis])	[7,0]	2.90e-02
1847919**	Rv1639c (hypothetical protein: Enterochelin esterase or related enzyme [Inorganic ion transport and metabolism])	[8,1]	4.23e-02

* Denotes an overlapping hit with phyC. ** Denotes a new POUTINE hit not identified by phyC.

Figure S1. Distribution of genome-wide r^2 values for the set of POUTINE hits vs the set of PLINK-only hits.

Figure S2. Montreal plot of the discovery set. *P*-values along the y-axis were calculated using PLINK's Fisher's exact test with the mid-p correction.

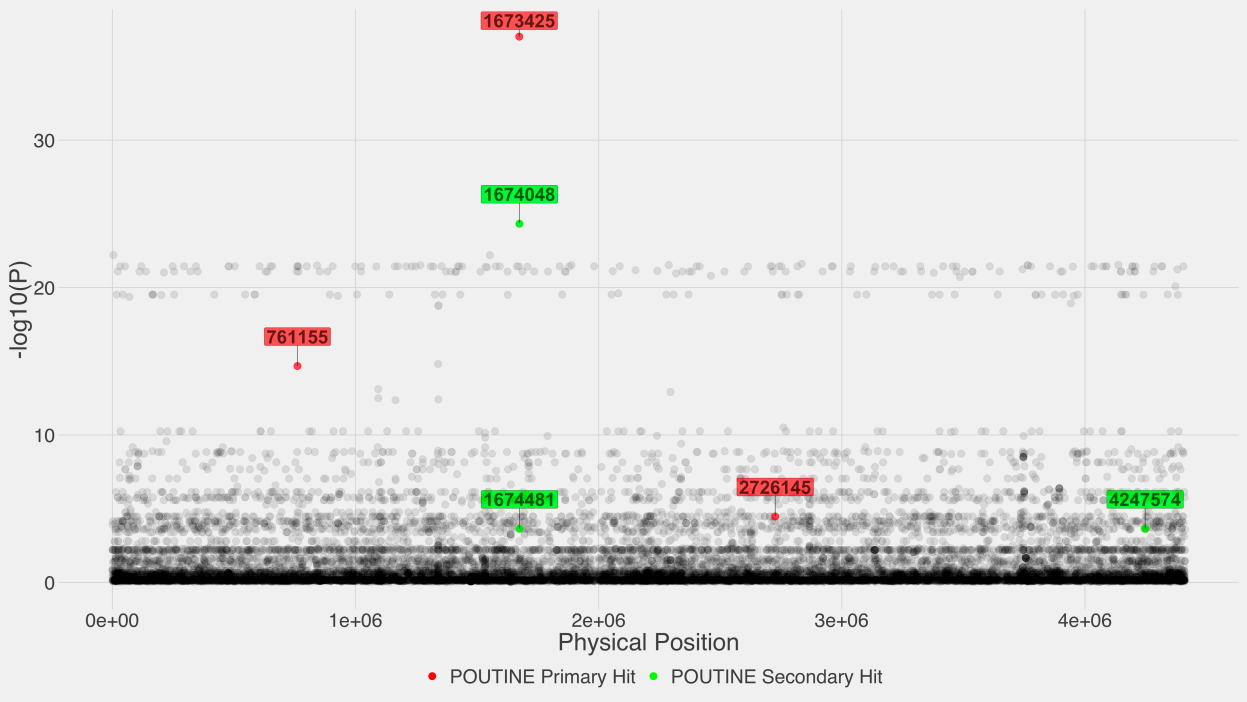


Figure S3. Score plots of the first four principal components derived from the LD pruned set using $r^2 > 0.99$ and non-overlapping windows of 1000 sites. The three colors (red, green, and blue) highlight three broad subpopulations as inferred from the phylogeny of the discovery set (Figure S5).

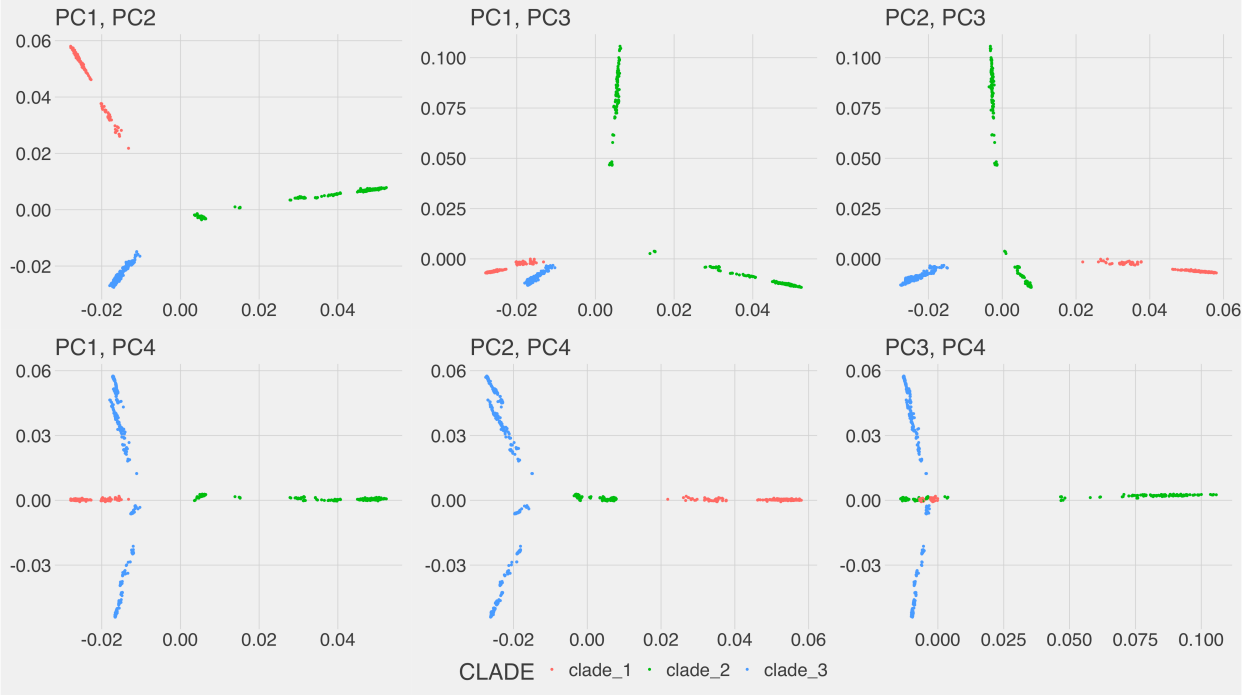


Figure S4. Scree plot of the first 20 principal components derived from the LD pruned set using $r^2 > 0.99$ and non-overlapping windows of 1000 sites.

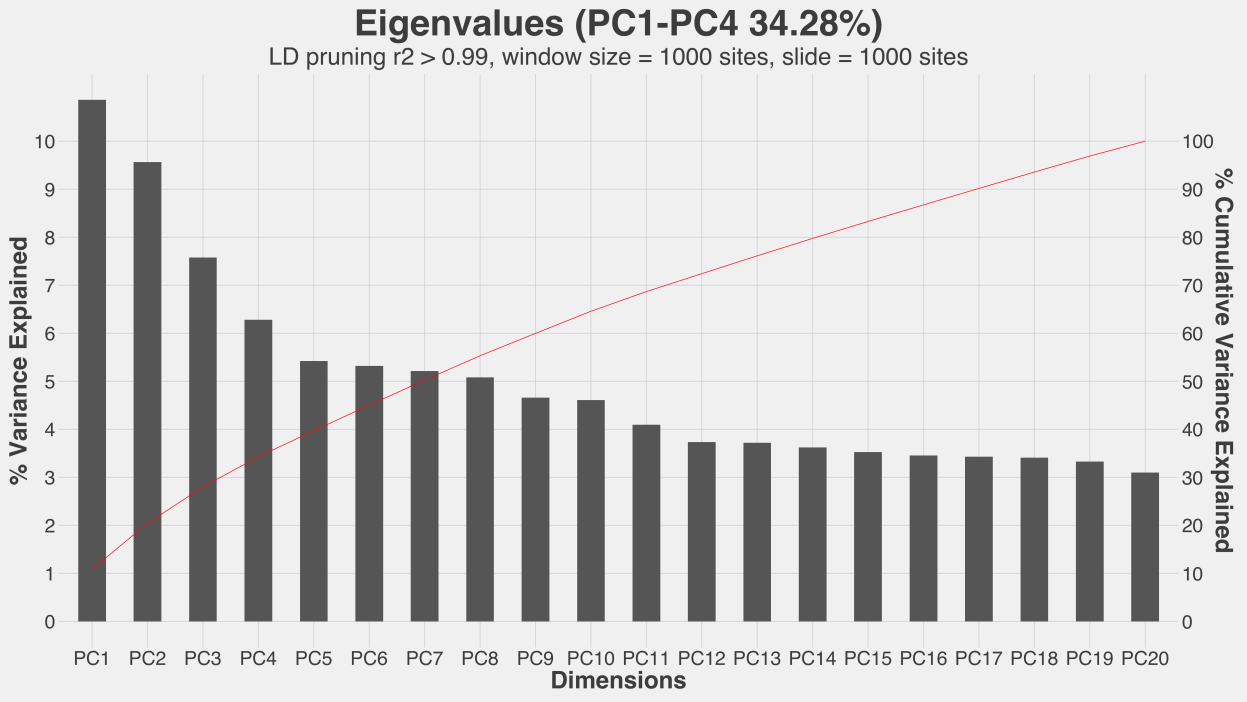
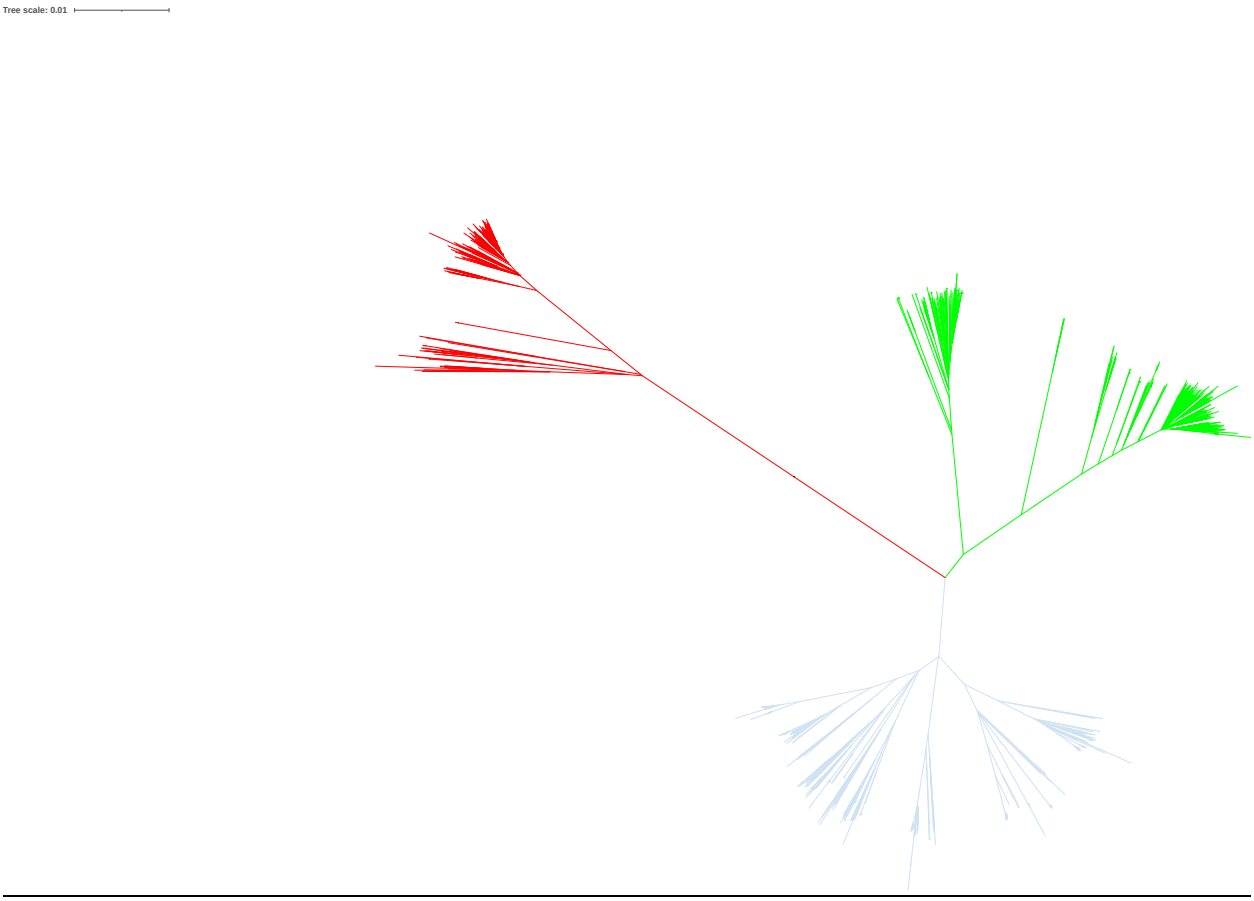


Figure S5. Phylogeny inferred using FastTree (double precision version) of the discovery set of 1330 *M. tuberculosis* genomes. The three colors (red, green, and blue) highlight three broad subpopulations.



Chapter 3: The genetic architecture of bacteriophage resistance in a natural population of *Vibrio breoganii*

Peter E. Chen[a], Fatima Aysha Hussain[b], Martin F. Polz[b,c], B. Jesse Shapiro[a,d]

[a] Département de sciences biologiques, Université de Montréal, Montréal, QC H3C 3J7, Canada.

[b] Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA

[c] Division of Microbial Ecology, Department of Microbiology and Ecosystem Science, Center for Microbiology and Environmental Systems, University of Vienna, Vienna, Austria

[d] Department of Microbiology & Immunology, McGill Genome Centre, Faculty of Medicine, McGill University, Montreal, Quebec, Canada.

Note: This is an ongoing study. Currently, we are waiting for an updated dataset from our collaborators, which will likely expand our sample size to over 100 host genomes. In addition, we will be including a new phenotyping assay that tests for the lack of phage replication post adsorption.

Abstract

Phages are key regulators of bacterial abundance and diversity across Earth's ecosystems, and phage therapy is a promising approach to combat antimicrobial resistant infections. Phages typically bind a specific receptor on the surface of bacterial hosts, and laboratory studies have shown how mutations in receptor genes can confer phage resistance, triggering cycles of co-evolution in which phage evolve to bind the mutated receptors. In addition to surface receptor mutations, bacteria encode various anti-phage systems, including restriction modification systems and CRISPR-Cas. The relative importance of these different phage resistance genes or mutations has not been studied extensively in natural bacterial and phage populations. Here, we apply a genome-wide association study (GWAS) to identify genes and mutations associated with phage resistance in a natural population of *Vibrio breoganii.* Both *V. breoganii* and sympatric phages were isolated from the same seawater sampled off the coast of Massachusetts, USA. Despite a limited sample size of 32 *V. breoganii* genomes, we identify three genome-wide significant associations between phage resistance and point mutations in *lamB,* a known phage receptor in *E. coli*, a sugar transferase that modifies the host cell surface, and a hypothetical protein. Secondary GWAS hits (not genome-wide significant) involving both point mutations and gene gain/loss also point to *lamB* homologs and other surface proteins as the major determinants of resistance. Our results contrast with a recent study of phage resistance in another much more closely-related *Vibrio* population, which found mobile phage defense elements to be the major determinants of resistance, not surface receptors. Together, this suggests that surface receptor variation can explain phage resistance on longer evolutionary time scales, while more recently evolved resistance is mostly due to mobile defense elements.

Introduction

Since the discovery of bacteriophages (also known as phages) in the early 1900s, they have come to be understood as the most abundant biological entities on the planet, estimated to have a population size of 10^{31} - 10^{32} phage particles [1,2]. They have been isolated from a diversity of niches from the human gut to the marine environment, and it is believed that there is at least one type of phage to infect every single bacterial cell [3]. Many studies have detailed the coevolution of these bacterial-viral systems, but we have yet to develop a complete picture of the genetic architecture of phage resistance, a potentially complex set of mechanisms where resistance to phage infection can occur at any stage of the viral life cycle, from the initial stage of attachment to the bacterial host surface to the lysis of the cell [4]. With the emergence of multi-drug antibiotic resistance, phage therapy has once again gained popularity and further highlights our need to better understand how bacterial resistance to phages can evolve during treatment [5].

Although there have been many experimental studies examining phage resistance, particularly using experimental evolution and transposon insertion sequencing [6–10], there have been few studies examining natural populations. The sequences of laboratory-evolved strains do not fully reflect the natural history of a population and depending on the experimental design may only reflect laboratory-scale time of evolution between host and a single phage. A natural population in its niche will have experienced a long history of coevolution with a multitude of phages along with other external pressures both biotic and abiotic. Thus, the genome sequences of natural phage-resistant and -sensitive bacterial populations capture a more complete picture of the genetic architecture underlying phage-bacteria interactions. A recent and notable study used 259 diverse *Staphylococcus aureus* strains challenged against eight phages belonging to all three morphological categories (*Siphoviridae*, *Myoviridae*, and *Podoviridae*) that infect the host species. The authors performed a genome-wide association study (GWAS) to discover underlying genetic components of phage

resistance, and then followed up newly reported putative causal loci not found in the literature with wet-lab verification using both transposon mutants and complemented strains [11].

In this study, we conducted a GWAS in a natural population of 32 *Vibrio breoganii* to examine the underlying genetic elements that influence phage resistance against 22 marine phages (these bacteria and phages were collected as part of a much wider study on a new lineage of non-tailed dsDNA Vibrio phages [12]). *V. breoganii* is a rarely characterized non-motile marine bacterium that specializes in algal carbohydrates [13,14]. While little is known about this species, and our sample size is quite small, we demonstrated that there is sufficient statistical power to detect high-effect, common mutations of known mechanisms to phage resistance.

Results

Underlying all our association analyses is a set of 32 *V. breoganii* genomes and a panel of 22 phages (Table S1) isolated sympatrically off the coast of Massachusetts. Although plaque assays were performed against 22 phages, 18 plaque assays showed an insufficient number of controls (i.e., phage susceptibility) to be of further use for association analysis. In addition, upon further examination of the full panel of plaque assays, nine of the *V. breoganii* genomes showed resistance to all 22 phages. We thus introduced a new phenotype labeled, “all-resistant”. In total, our association analyses were conducted using five phenotypes: phage resistance against four individual phages (labeled as 1.034.O, 2.117.0, 1.139.A, 1.117.0) and the all-resistant phenotype.

Although *Vibrio* genomes tend to exhibit relatively more recombination [15] than typical clonal populations, such as *Mycobacterium tuberculosis* [16], bacterial genomes in general lack the block-like linkage disequilibrium (LD) structures necessary for classic "allele-counting" GWAS approaches to distinguish between the true drivers of

association and hitchhiking background passengers that are in strong LD [17]. Thus, we applied both an allele-counting approach and a homoplasy-counting approach to both the core and accessory genomes. For allele counting, we used PLINK [18], and for homoplasy counting, we used our recently developed method, POUTINE [17]. A summary of all the primary (genome-wide significant, after correction for multiple hypothesis tests) and secondary (uncorrected p-values < 10^{-4}) hits is shown in Table 1.

Table 1. Summary of all hits in both the core and accessory genomes using both PLINK and POUTINE across all five phage phenotypes. Red represents primary hits while all other hits are secondary.

1.034.O Phage Phenotype	All-resistant Phenotype	1.117.O Phenotype	2.117.O Phenotype	1.139.A Phenotype
Core Genome Associations				
lamB homologue maltoporin phage receptor	Sugar transferase	Hypothetical protein	Tyrosine kinase/ phosphatase	Permease
	lamB homologue maltoporin phage receptor	Tyrosine kinase/ phosphatase		
	Permease			
	malM (part of the lamB operon)			
	Efflux pump			
Accessory Genome Associations				
lamB homologue maltoporin phage receptor	lamB homologue maltoporin phage receptor			
	ompF			

Core genome association

Variant calling was conducted using the 1C10 *V. breoganii* genome from the National Center for Biotechnology Information (NCBI) database because it was longer (4.15 Mbp) than all other assemblies in our study population. Using 1C10 as the reference better ensures that portions of the core genome are not left out of the core genome multiple sequence alignment as the assemblies in our study population varied in total genome length (Figure S1). After variant calling, 265,936 biallelic nucleotide sites were observed across the 32 *V. breoganii* genomes. To increase the power of the study given the small sample size, we reduced the number of markers to 84,607 biallelic sites by filtering out all sites with a minor allele frequency < 0.20. This cutoff was chosen as a reasonable detectable limit of association for logistic regression and a relatively small sample size of 32. We further reduced the number of markers tested by filtering out all sites that are in complete linkage disequilibrium (LD) with an $r^2 = 1$, while keeping one representative site for each subset of sites in complete LD. Thus, the total number of markers used in the core genome association analysis was 42,072 sites.

Using these 42,072 biallelic sites, we examined the magnitude of population stratification in our population. Across all five phenotypes, we observed relatively low levels of stratification (Figure S2) compared to the higher levels typically observed in microbial populations [19]. We note that this lower level of stratification aids in the detection of significant associations, in that any stratification correction will be less severe and thus lead to a smaller loss of power. The first principal component (Figure S3) by itself captures the five V. *breoganii* genomes (222.51.E5, 261.52.F8, 286.51.B5, FF-50, 261.52.C1) that are most phylogenetically distant from the rest of the isolates (Figure S4). Thus, using the first two principal components, which explained 27% of the total variance, reasonably captured the population structures that contributed to stratification.

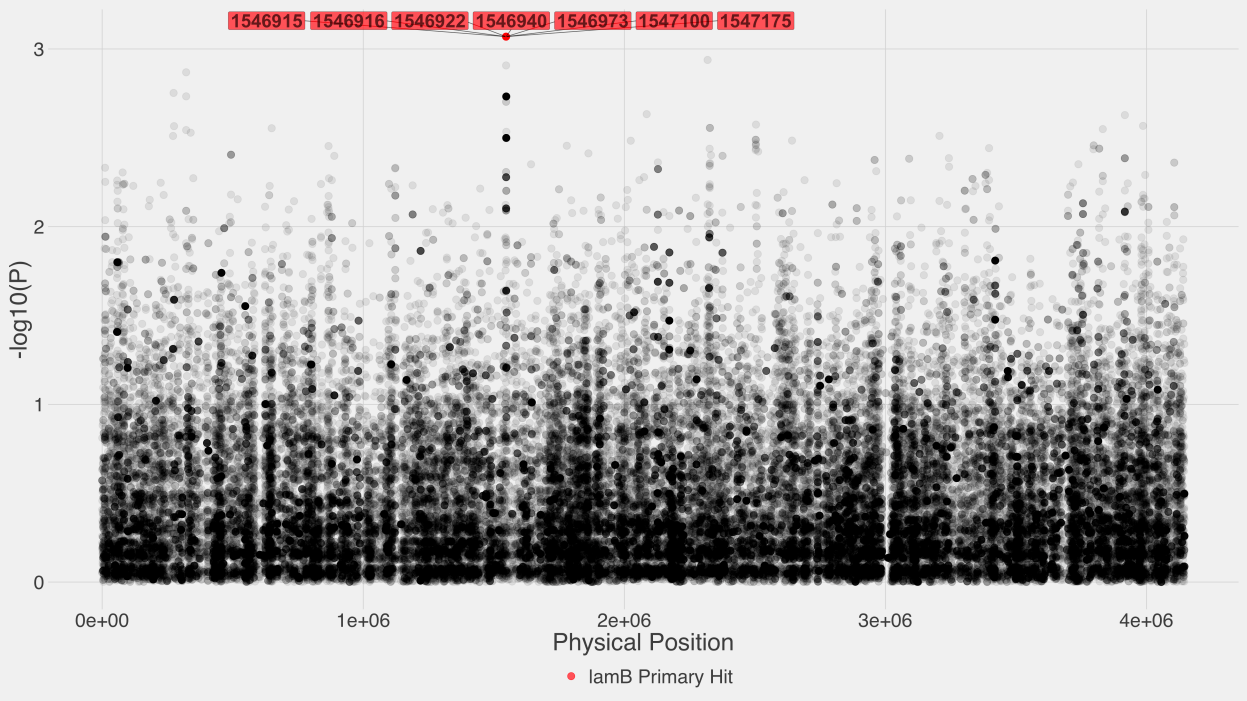
Primary GWAS hits

Although the sample size of 32 genomes is relatively small, we observed three genome-wide significant GWAS hits in the core genome with a familywise p-value ≤ 0.05. Two of the three hits are PLINK-only hits while the third hit is a POUTINE-only hit.

PLINK hit in a *lamB* gene

In the 1.034.O phage phenotype, PLINK revealed a cluster of seven nearby nucleotide sites which all showed a familywise p-value = 0.02 and reside inside a *lamB* gene (Figure 1). Five of the sites are nonsynonymous single nucleotide variants (SNVs) and the other two are synonymous (Table S2). The *lamB* gene codes for an outer membrane protein, and is known to serve two broad functions: (1) as a beta barrel that transports maltose (a disaccharide comprised of two glucose molecules) across the outer membrane [20,21], and (2) as a phage receptor for various phages including the classic λ phage that infects *E. coli* [22–24]. Thus, this maltoporin has binding sites for both maltosaccharide and the tip attachment J protein of the tail of phages [25]. In fact, resistance to phage λ has been associated with partial or complete loss of the ability to grow on maltose [26]. It is unsurprising that a genome-wide significant hit was found inside a classic phage receptor, but what is striking is the near-complete penetrance of the putative causal alleles of these seven sites (Table S2). Only two host genomes (10N.261.51.F2, 10N.286.51.A6) that possess the putative causal variant are not resistant. It is this high level of penetrance that amplifies the statistical signal to allow for genome-wide significance to occur with such a small sample size. The effect size of each of these seven sites is extremely large (Odds Ratio of 189). To place this massive effect size into context, a typical genome-wide significant association hit in common traits in humans shows an OR of 1.5 or less [27], and even these estimates are likely to be inflated [28]. As hypothesized in our previous work [19], the genetic architecture of prokaryotes can differ from multicellular eukaryotes in that common mutations of large effect size may be prevalent, whereas among human traits common, large-effect mutations are mainly observed in monogenic and Mendelian traits [29].

Figure 1. Montreal plot for the 1.034.O phage phenotype, showing the primary hit in *lamB*. Seven nearby sites (all seven sites overlap onto one single red dot due to the scale of the plot) are genome-wide significant inside the *lamB* gene. Y-axis shows unadjusted p-values from logistic regression using PLINK. Red labels show the physical positions of each of the seven sites.

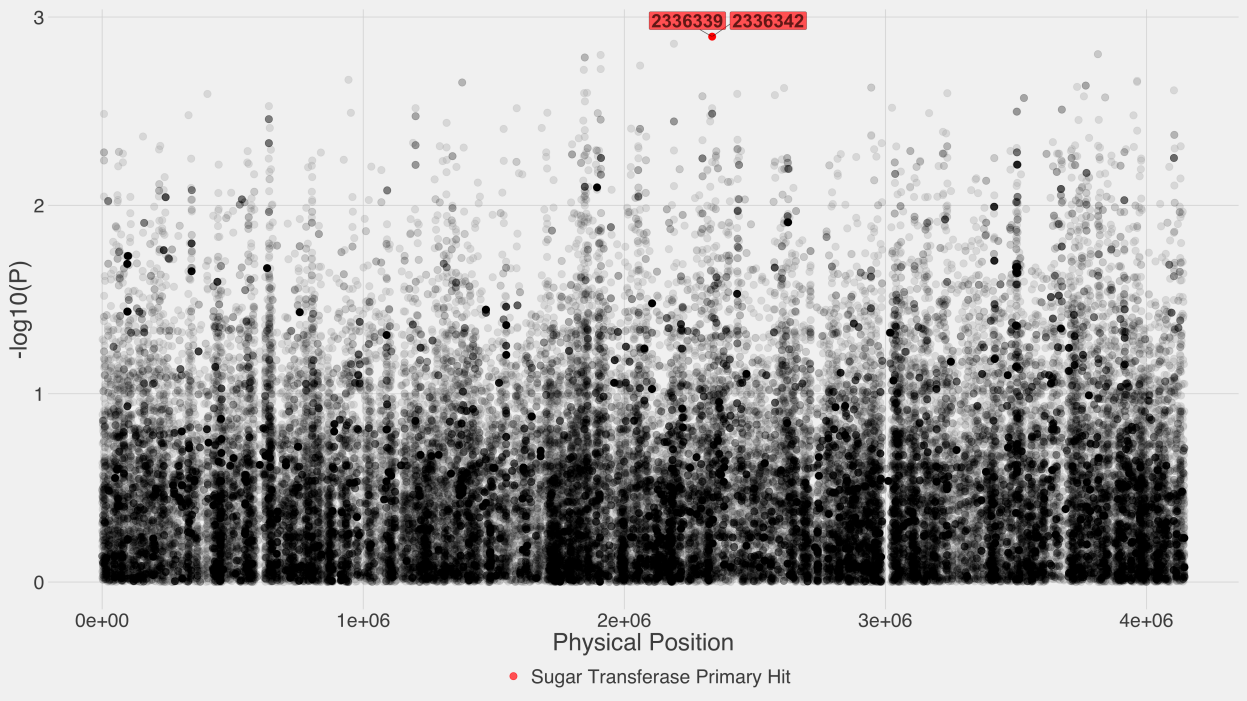


PLINK hit in a sugar transferase gene

In the all-resistant phage phenotype, PLINK identifies two nearby sites as genome-wide significant with a familywise p-value = 0.05 (Figure 2). These two sites reside in a sugar transferase gene, and both mutations are nonsynonymous. Based upon RAST and Prokka annotations, as well as manual searches in the NCBI non-redundant database and UniProt database, there are many homologs of this gene. These homologs include *sypR*, *tuaA*, *rfa*/*rfb*, and *tag* genes like *tagO*. Despite this apparent diversity, they all function as sugar transferases. The closest sugar transferase homolog to this *V. breoganii* gene is the undecaprenyl-phosphate N-acetylgalactosaminyl 1-phosphate transferase gene. This gene and close homologs

have been shown to cause phage resistance in both gram-negative [8,30] and gram-positive bacteria [31,32] and also cyanobacteria [33]. Although there are variations to the underlying mechanism of phage resistance by this gene, all mechanisms involve the disruption of the sugar transferase in lipopolysaccharide (LPS) synthesis. LPS structures on the outer membrane of gram-negative bacteria are commonly used as phage receptors [34]. In all but a few cases, phage adsorption involves either constituents of the cell wall or structures protruding from the cell wall. Specifically, this sugar transferase is known to modify the O-antigen portion of the LPS, thus disrupting phage adsorption; the O-antigen portion being a common interaction point between bacteria and phage [35]. For this all-resistant phage phenotype, the implication of an LPS sugar transferase suggests that it is very likely that there exists at least one LPS phage receptor among these nine host genomes that are entirely resistant to all 22 phages assayed. The minor allele frequency of this putative causal mutation is 0.28 and its effect size is an OR of 38. Again, we observe a potential phage resistance mutation that is common in frequency with a relatively large effect size.

Figure 2. Montreal plot of the sugar transferase gene primary hit in the all-resistant phenotype. Two nearby sites are genome-wide significant inside the sugar transferase gene. Y-axis shows unadjusted p-values from logistic regression using PLINK. Red labels show the physical positions of the two sites.

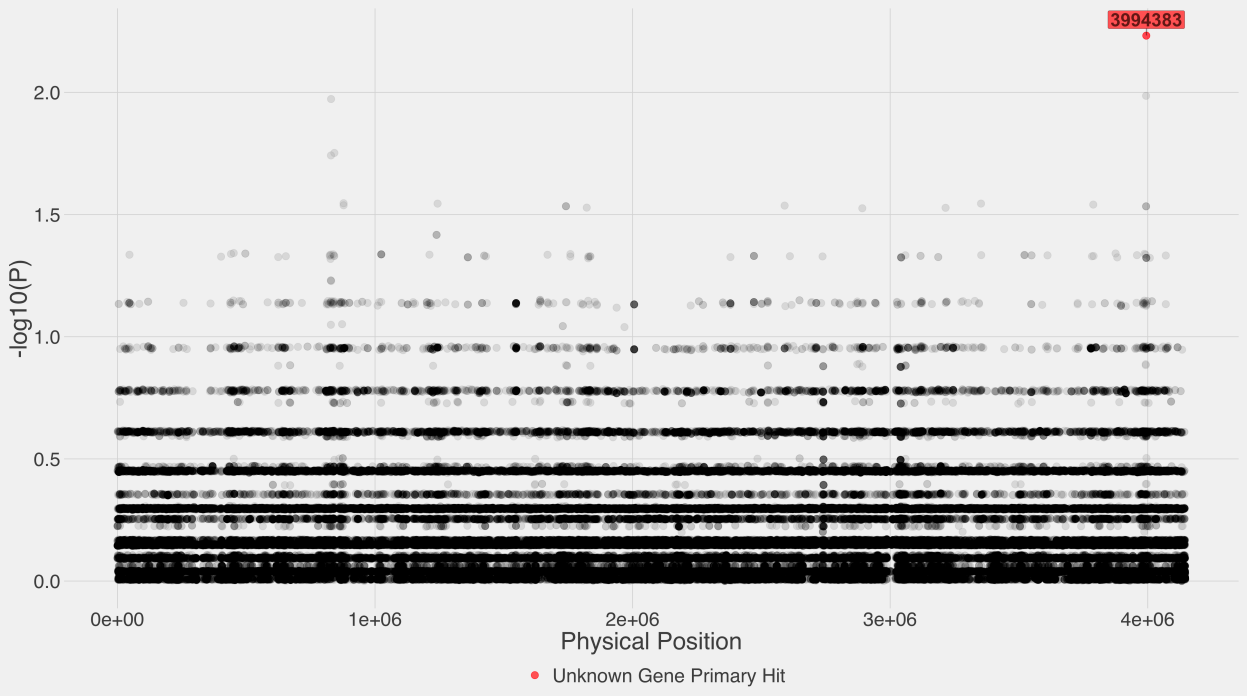


POUTINE hit in a hypothetical protein

In the 1.117.O phage phenotype, POUTINE identified one genome-wide significant site with a familywise p-value = 0.04 (Figure 3). This putative causal variant resides in a hypothetical protein and is a nonsynonymous mutation. The minor allele frequency at this site is 0.375 which corresponds to 12 minor alleles. POUTINE shows that all 12 alleles are resistant to phage 1.117.O, and again we observe a common mutation with a large effect size, where the minor allele has complete penetrance (i.e., all genomes that harbor this minor allele are resistant to phage 1.117.O). This site is also a "hybrid site"; 12 of the alleles are homoplasic while the remaining 20 alleles are identity-by-descent alleles. Thus, it is not always the case that convergent sites consist of only homoplasic mutations. Other such hybrid sites have been observed, for example in a causal mutation in isoniazid drug resistance in an *M. tuberculosis* population [17]. The

function of this hypothetical protein is currently unknown. Both RAST and Prokka annotations show this gene to be a hypothetical protein. A search for sequence homology in the NCBI database reveals 35% identity over 91% coverage to the VP1478 hypothetical protein in Vibrio *parahaemolyticus* RIMD 2210633 and 95.4% identity across the entire length of a hypothetical protein in the FF-50 V. *breoganii* genome. Thus, this gene has no homology to any available sequence to date that has an annotated function. This finding is consistent with the fact that *V. breoganii* is less characterized relative to other vibrios such as *V. cholerae*. The only *V. breoganii* genomes submitted to any database have been provided by the Polz laboratory.

Figure 3. Montreal plot of the unknown gene primary hit in the 1.117.O phage phenotype. One site is genome-wide significant inside this hypothetical protein. Y-axis p-values are pointwise estimates using POUTINE.



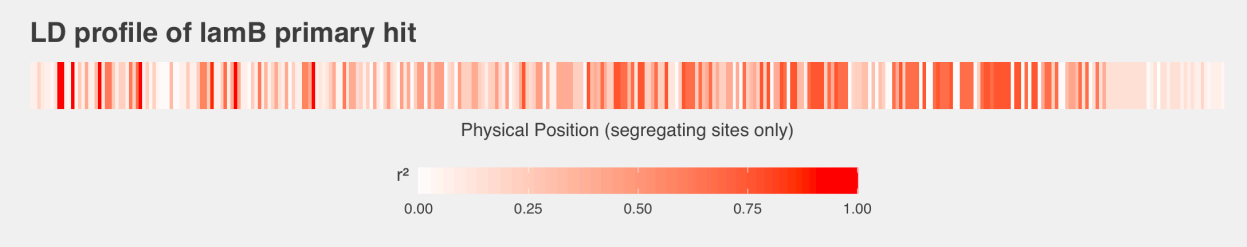
LD profiles of primary hits

Because bacterial genomes exhibit strong and long-range LD even in relatively highly recombining species like the vibrios, it is crucial to examine all other loci that are in strong LD with these hits to determine the likely driver of the association signal [17].

lamB LD profile

The *lamB* primary hit contains seven nucleotide sites that are in complete LD ($r^2 = 1$). We surveyed the entire genome for sites in relatively strong LD ($r^2 > 0.5$) with this cluster of genome-wide significant hits, and this revealed only two loci. The first locus is a 1074 bp region located completely within the *lamB* gene and surrounds the cluster of seven sites. This region spans almost the entirety of the *lamB* gene, with the rest of the gene in relatively weak LD to this region (Figure 4). The second locus is a single nonsynonymous site ($r^2 = 0.62$) located in a nearby gene, *malR*, approximately 1 Kbp from *lamB*. The *malR* gene in some gram-positive bacteria is known to be part of the maltose regulon and functions as a repressor of the regulon in the absence of maltose [36]. It is unclear at this time what the full complement of genes is that participate in the maltose regulon for *V. breoganii* species, however, the proximity of *malR* to *lamB* is consistent with how operons are spatially organized in clusters within the genome. Even if *malR* does function as a repressor for the operon that contains *lamB*, it is unclear how an altered *malR* function, potentially caused by this nonsynonymous mutation, would affect *lamB* expression. Since no other loci within the genome are in strong LD with the *lamB* primary hit, it is reasonable to conclude that the *lamB* gene, and potentially also the linked *malR* gene, are causal loci for resistance against phage 1.034.O.

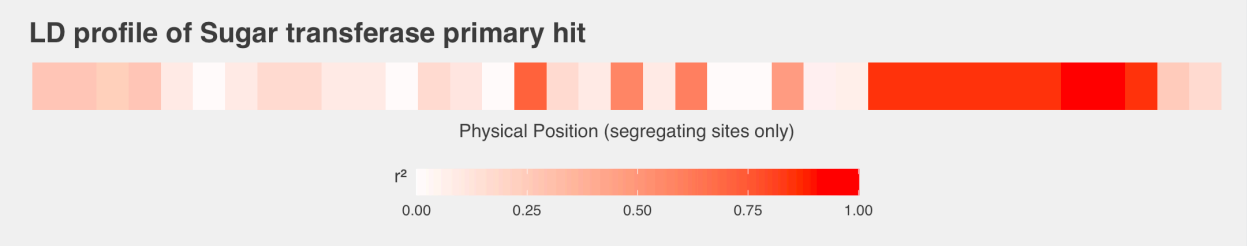
Figure 4. Heatmap of linkage disequilibrium within the *lamB* gene primary PLINK hit. LD of the seven genome-wide significant sites against all other segregating sites within the *lamB* gene (1302 bp).



Sugar transferase LD profile

The sugar transferase primary hit consists of two sites in complete LD. Similar to the *lamB* LD profile, we observed a 228 bp region of strong LD surrounding the two sites and located completely within the transferase gene. This region spans close to half the length of the transferase gene, with the rest of the gene in relatively weak LD (Figure 5). In addition, we also observed a distant site ~474 Kbp away in strong LD (r^2 = 0.57) and located inside a gene annotated as an N-carbamoyl-L-amino acid hydrolase. A literature search revealed no known connection with phage resistance and therefore it is reasonable to conclude that the sugar transferase alone is the likely causal gene, although any causal effect from the hydrolase cannot currently be ruled out.

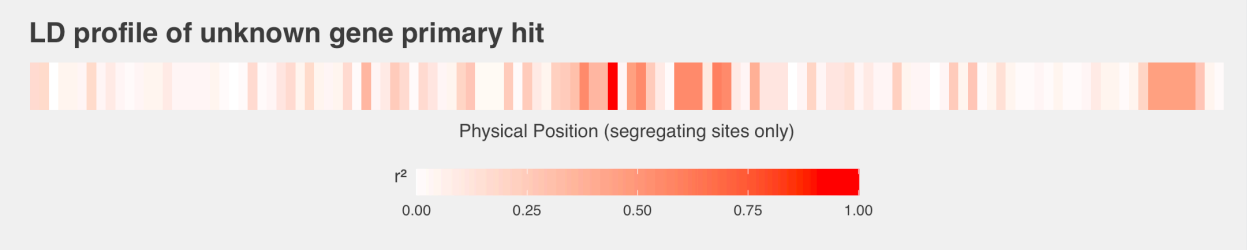
Figure 5. Heatmap of linkage disequilibrium within the sugar transferase gene primary PLINK hit. LD of the two genome-wide significant sites against all other segregating sites within the sugar transferase gene primary hit (639 bp).



Hypothetical protein LD profile

This primary hit consists of only one site located inside an unknown protein. Consistent with the fact that this site is a target of convergent mutation, we observed no other sites within the genome that is in complete LD with this POUTINE hit. Similar to the other two primary hits, this convergent site is only in strong LD with seven other sites that are spread across a relatively small 72 bp region that spans ~10% of the length of the unknown gene (Figure 6). In addition, this primary hit is in strong LD (r^2 = 0.54) with one synonymous site in the *malT* gene that is located ~3.37 Mbps away. The canonical *E. coli malT* gene is a transcriptional activator of the operon containing *lamB* [37], and mutations disrupting *malT* are known to lower the expression of *lamB* and thus diminish phage adsorption [10]. The linkage between the unknown protein and *malT* is difficult to interpret partly owing to the lack of any functional annotation for the primary hit. Moreover, without a higher sample size, we cannot conclude that *malT* is not a causal gene for phage 1.117.O, albeit the current statistical evidence does not show this site to be genome-wide significant. One reasonable interpretation is that the linkage between the two genes is not functional but rather a side-effect of the pattern of recombination where even in moderate to high recombining microbial populations, there will still exist long-range LD [15,19]; we note that an $r^2 > 0.5$ is an arbitrary threshold for this analysis and does not signify a precise biological meaning. Finally, it is unclear if *malT* is interacting with the unknown protein, perhaps as an uncharacterized member of the maltose regulon, or that the unknown protein is acting alone in a non-*lamB*-related mechanism.

Figure 6. Heatmap of linkage disequilibrium within the unknown gene primary POUTINE hit. LD of the one genome-wide significant site against all other segregating sites within the unknown gene primary hit (1020 bp).



The three primary hits reside in recombination hotspots

All three LD profiles of the primary hits (Figures 4-6) showed the same pattern of LD where each set of genome-wide significant sites is surrounded by a region of strong LD relatively unlinked to the rest of the genome. No other regions in the genome (except for the three exceptions detailed above) showed strong LD ($r^2 > 0.5$) with these three highly localized regions. This pattern is consistent with recombination hotspots. This genomic feature has allowed for a cleaner dissection of which sites are likely the true positive drivers underlying the association signals for the *lamB* and sugar transferase allele counting hits. Curiously, we did not observe homoplasy counting hits in either putative recombination hotspot. By definition, recombinant tracts are homoplasic, so the expectation is that if a site is genome-wide significant inside a hotspot, then it will likely be identified as also convergent. A likely explanation is that homoplasic mutations are not being accurately identified. Two main reasons may prevent the identification of homoplasies: 1) Our small sample size provides little sample diversity and therefore there is insufficient polyphyletic structure in the inferred phylogeny to observe all homoplasic mutations. 2) The branches of the phylogeny are not well-resolved due to the level of historical recombination present, and thus the inferred phylogeny does not accurately reflect clonal relationships but rather the population's recombination history [38]. We examined the level of homoplasy within and surrounding the hotspots and we did not observe an elevated level of homoplasies inside the two putative recombination hotspots relative to their surrounding regions. If

recombination is ongoing, the expectation would be that the hotspots harbor elevated levels of homoplasic mutations relative to the rest of the genome. The lack of homoplasies in the two putative hotspots supports the theory that these mutations are not being identified with sufficient accuracy.

Secondary hits

Despite the small sample size and low statistical power in our study, both common frequency and large effect size mutations allowed the identification of three putative causal genes for phage resistance at genome-wide significance. To further probe for other potential candidates of phage resistance, we examined a relatively small subset of sites enriched at a pointwise p-value (resampling-derived point estimates from permutation testing) of less than 10^{-4}, and that also showed an annotated function and underlying mechanism that has been reported in the literature to be involved in phage resistance. We consider these hits as 'secondary hits' in contrast to the three 'primary hits'. Here, we highlight the following promising candidates and note that if they are causal genes, they can potentially reach genome-wide significance with higher sample size.

PLINK hit located 5' of a permease gene

In the 1.139.A phage phenotype, the top PLINK site with the smallest pointwise p-value estimate contains a nonsynonymous mutation that resides in the non-coding region that is 5' relative to the start of a permease gene. Permease proteins in general catalyze the transport of specific classes of molecules across cell membranes and are found in virtually all cells ranging from multicellular eukaryotes to archaea [39]. Gram-negative bacteria such as vibrios also contain many permease genes for both signal transduction and energy transport; one of the classic permeases, *lacY*, is a transmembrane protein that is part of the lac operon in gram-negative *E. coli* and facilitates the active transport of lactose [40]. Interestingly, in *E. coli* it has been shown

that the mannose permease, encoded by the *manXYZ* locus of three genes, is not only necessary for the transport of mannose and maltose but also the penetration of λ phage DNA across the inner membrane [41–43]. This secondary hit in a non-coding region 5' of a *V. breoganii* permease could function as a cis-acting regulatory element that modifies the expression of its corresponding permease gene, which then acts similarly as the mannose permease for λ phage.

PLINK hit inside a tyrosine kinase/phosphatase system

Examining the most significant pointwise estimates revealed that PLINK identified nearly the same set of sites inside a tyrosine kinase and tyrosine phosphatase gene in both the 2.117.O and 1.117.O phage phenotypes. For the tyrosine phosphatase locus, the same four sites were found between the two phenotypes. In the tyrosine kinase locus, each phenotype revealed a different site only 30 bps apart. A relevant study in *L. monocytogenes* [44] revealed a mechanism for the tyrosine phosphorylation system to cause resistance to *Listeria* phage A511 and P35. In this study, the authors created a deletion mutant that lacks all four of the highly conserved tyrosine phosphatase genes seen in the strains studied. The deletion mutant was observed to lack *N*-acetylglucosamine in its wall teichoic acid, a structure that protrudes beyond the cell wall and capsule of gram-positive bacteria analogous to LPS structures and outer membrane proteins in gram-negative bacteria, thus capable of binding with phages. Wall teichoic acid is also a primary phage receptor for *Staphylococcus* species [32]. As further evidence that the tyrosine phosphatase genes are necessary for phage resistance, the authors created a set of complement strains harboring varying sets of the four genes and showed that phage susceptibility was recapitulated to varying degrees depending on the complement strain used. Phage adsorption assays further showed that the mechanism underlying resistance was likely at the attachment phase of the phage to the host cell wall. Although the example above is in a gram-positive bacteria and wall teichoic acids are not found in gram-negative bacteria, there are reported examples in gram-negative bacteria of how tyrosine kinase/phosphatase systems can modify various cell wall entities. In enteropathogenic *E. coli* species, the

Etk/Etp tyrosine kinase/phosphatase genes are required for capsule formation [45], a structure that resides outside of the outer membrane and cell wall. In the gram-negative plant pathogen, *Erwinia amylovora*, an *Etk* homolog was shown to play a role in exopolysaccharide formation. The authors created a complement strain using a phage-resistant mutant and showed the restoration of phage susceptibility to *E. amylovora* phage Ea1h [46]. Broadly, the tyrosine phosphatase system has been shown to play a role in the regulation of cell wall integrity [47] and biofilm formation [48], thus disruption of either the kinase or phosphatase component may lead to a modification of various cell wall structures and downstream modification of the magnitude of phage adsorption.

Multiple PLINK secondary hits in the all-resistant phenotype

Examining the most significant sites in the all-resistant phenotype, PLINK revealed multiple candidates, some of whose functions have already been detailed above. First, among the top secondary hits were four sites in complete LD located inside a *lamB* maltoporin phage receptor homolog that is different than the *lamB* porin identified as a primary hit in the 1.034.O phage phenotype. Of the four sites, one is nonsynonymous and is likely the true driver of the association signal. Because these nine host strains are resistant to all 22 phages tested, it is not unlikely that more than one phage receptor has been disrupted. Second, another permease gene was identified, this time with one site located inside the gene harboring a nonsynonymous minor allele. Third, two sites in complete LD were identified inside the *malM* gene, which is part of the maltose regulon that contains the *lamB* gene. Importantly, a third *lamB* homolog is located next to this *malM* gene. Little is known about the function of *malM* since its first reporting in the literature [49]. However, *malM* is not just a part of the maltose regulon, it is also part of the operon containing *lamB*, which hints at a possible mechanism that affects lamB function as the actual phage receptor. Curiously, in an *E. coli* and λ phage study [50], the authors identified various mutations that confer phage resistance in both the *lamB* gene and *malT*, the transcriptional activator of the *lamB* operon. In their study, the *malM* gene that is a secondary hit in our study was not

sequenced and analyzed. As such, it remains to be seen if these two sites in *malM* underlie a mechanism to disrupt phage binding to the *lamB* receptor like its sister *malT* gene. Lastly, two nonsynonymous sites were identified inside an efflux pump. These structures are protein complexes that span the outer membrane and extend beyond the cell wall. Thus, like LPS, wall teichoic acid, and porin channels such as *lamB*, efflux pumps can provide a feasible structural target for phage binding. In one study using *E. coli* and phage LTS [51], the authors showed that the *tolC* protein, which can act as an efflux pump, can harbor missense mutations mostly in two hypervariable regions of the gene that cause resistance to phage LTS. In an intriguing study using *E. coli* and phage U136B [30], the authors examined the evolutionary trade-offs between multi-drug efflux pumps and phage resistance. Their study showed that there exists an antagonistic pleiotropy where an increase in phage resistance resulted in a decreased sensitivity to antibiotics, owing to both phage and antibiotics sharing the same efflux pumps. Relevant to our study is that these authors demonstrated that phages can be highly dependent on efflux pumps as phage receptors.

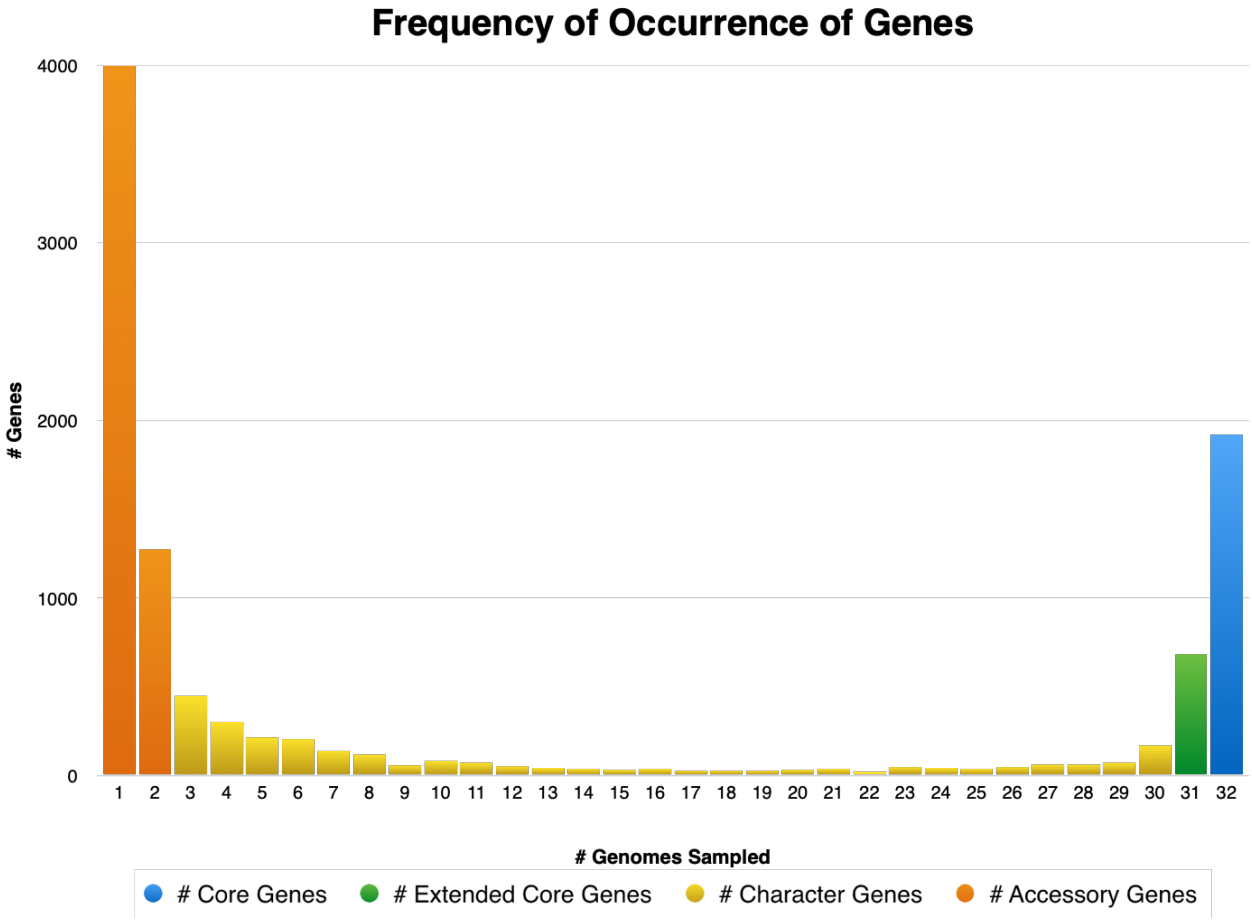
Accessory genome association

The concept of the pangenome, the total set of genes among a group of organisms (typically a species), was introduced in 2005 [52] using only eight S. *agalactiae* genomes. In an excellent review [53], Lapierre and Gogarten defined the genes in the pangenome into four gene classes: core, extended core, character, and accessory. The core genes are defined as genes present among all members of a group, the extended core genes are genes present in at least 99% of members, the character genes are present in some but not all members, and the accessory genes are absent from at least 95% of members, with many accessory genes only observed in one member (i.e., ORFans). Importantly, the authors calculated the pangenome of 573 diverse bacterial genomes and showed that the core plus extended core genome of an average bacterial genome consists of ~250 gene families spread across only 8% of the length of the genome. The small genomic territory covered by the core genome

highlights the need to examine the non-core portion of the pangenome when conducting association studies. A major difference between human and bacterial genomes is that the latter possesses markedly larger pangenomes. It is not to say that there is no human pangenome [54] but that they have not been included in human GWAS, partly because few novel sequences, on the order of ~30 Mbps equal to ~1% of the human reference sequence, have been identified in human pangenome studies [55].

We first characterized the pangenome of our *V. breoganii* population by frequency of gene occurrence, and we showed the characteristic 'U' shape [53] from the dominant frequencies of accessory (5271 genes) and core genes (1919 genes) out of a total of 10403 genes (Figure 7). By definition, each host genome contributes to the core genome the same number of genes, while in the other gene classes individual contributions vary particularly in the accessory genes, with sample 10N.261.45.E9 contributing 22x more accessory genes than sample 10N.261.52.F5 (Figure S1).

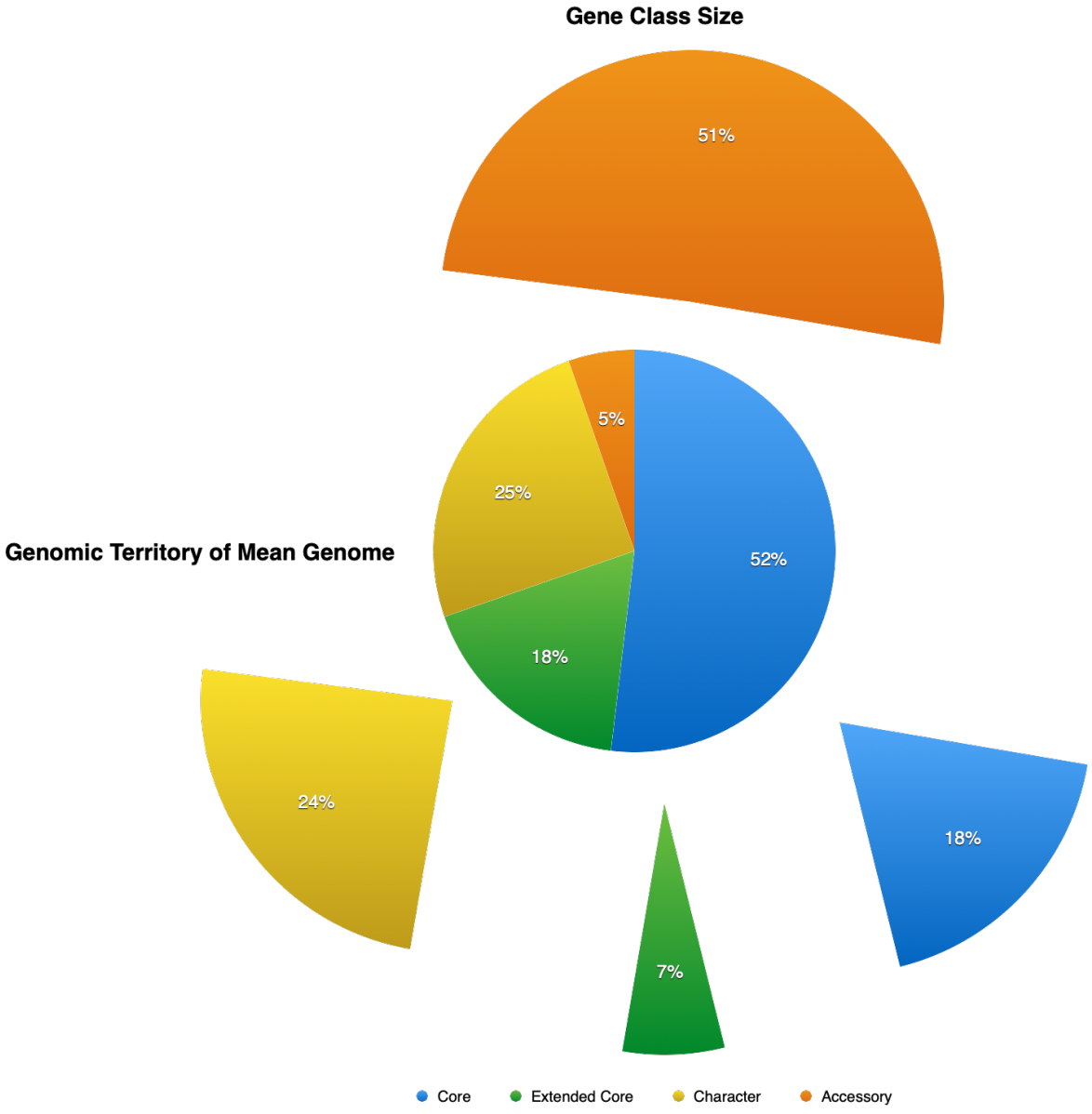
Figure 7. All genes in the population of the 32 V. *breoganii* genomes broken down by their frequency of occurrence.



Also as expected in our study host population, we observed that the pangenome is 'open' (Figure S5); as more genomes are sampled, the number of core genes plateau while the number of non-core genes continues to rise, demonstrating the large genic diversity not sampled if one were to constrain their association study to only the core genome. Concretely in our study, if one considered the genomic territory occupied by each gene class for the "mean genome", we see that the core genome only accounts for 52% of this genome (Figure 8). Furthermore, the core genome only accounts for 18% of the total number of genes in the pangenome, leaving 82% of the total genes unconsidered in a strictly core genome association (Figure 8). In other words, without

examining the pangenome, one effectively is excluding approximately half of each genome as potentially playing a role in the phenotype.

Figure 8. The four gene class sizes. The inner pie chart represents the proportion of genomic territory occupied in the "mean genome" by each gene class; the mean genome is simply the mean of the proportions for each class across all samples. The outer pie chart represents the proportion of total genes in each gene class.



To expand our study to the pangenome, we recoded each gene as if it were a biallelic SNV. Specifically, for each gene, the absence of the gene from a particular genome was coded as an "A" allele and a presence of the gene was coded as a "G" allele. For purposes of simplification and consistent with the pangenome literature, we refer to all genes outside of the core genome as the accessory genome.

Primary and secondary accessory genome hits

Using both PLINK for allele counting and POUTINE for homoplasy counting, we did not observe any genome-wide significant genes in the accessory genome. In the same fashion as our core genome analysis, we examined our most significant pointwise estimates for each phenotype to probe for potential causal genes that may become genome-wide significant with higher sample size. Here, we report two such promising candidates with a known phage resistance mechanism. Consistent with the literature, many genes in the *V. breoganii* pangenome are hypothetical proteins; 42% of the genes in this pangenome have no known function, with 93% of these hypothetical proteins residing outside of the core genome (Figure S6). Genes outside of the core genome are less conserved, with accessory genes often specific to strains and serotypes, thus their functions are mostly unknown. Considering this, we note that many of the secondary hits identified using pointwise estimates are hypothetical proteins and would require a higher sample size to potentially provide stronger statistical evidence for further consideration.

PLINK hit in a *lamB* homolog phage receptor

The first secondary candidate is a PLINK hit to another *lamB* phage receptor homolog. In contrast to the two core *lamB* genes identified in our core genome association analysis, this accessory *lamB* gene is completely absent in 16 of the 32 genomes, with 14 out of 21 cases missing this gene and 9 out of 11 controls with this gene present. The interpretation for this accessory genome hit is more straightforward (*i.e.,* presence or absence) than core genes in which it is unclear how a silent or missense mutation affects the gene, and perhaps even more unclear are the downstream effects of a

mutation in a non-coding region. Here, we observe the *lamB* gene is mostly absent from cases (phage resistant) but not controls, which is consistent with the abolishing of this *lamB* porin as a potential receptor for phage 1.034.O. We further note a large effect size for this gene of OR = 17. As further evidence this accessory *lamB* gene may be causal, we observed a secondary hit for this same gene in the all-resistant phenotype. We again observed a large effect size of OR = 18, and 8 out of the 9 all-resistant host strains are missing this *lamB* gene in its accessory genome.

PLINK hit in the ompF gene

In the all-resistant phenotype, another outer membrane protein, *ompF*, was identified as a secondary hit. Similar to the *lamB* porin, the *ompF* protein folds into beta barrels and has been reported to function in a trimeric complex as a general porin [56]. The *ompF* porin is known to allow the transport of a range of molecules including sugars [57], bacteriocins [58], and antibiotics [59,60]. Although the amino acid sequences of *lamB* and *ompF* are quite different, they both show similar crystal structures [61–63]. Crucially, *ompF* is known to bind to various phages including *E. coli* phage K20 [64,65]. In our study population, the *ompF* gene is absent in 8 out of 9 cases and present in 15 out of 23 controls. Again, we observe a large effect size of OR = 23.

In a 2012 experimental evolution study [7], the authors report that λ phage evolved to use the *E. coli ompF* porin as an alternative receptor in lieu of its primary receptor, the *lamB* porin. The authors observed an interesting historical contingency of mutations that involved the coevolution of both the *lamB* receptor and the J protein of the λ phage tail that binds to this receptor. In total, four mutations in the phage J protein are required for phage utilization of the *ompF* receptor in what the authors describe as an "all-or-none form of epistasis". However, the order of mutational events in both host and phage are critical in allowing an adaptive path to this novel receptor phenotype, as such not all λ populations in their study evolved this new capability. First, the host bacteria evolved resistance to λ phage infection via mutations in the *malT* positive,

transcriptional regulator of *lamB*. Second, despite the malT disruption, spontaneous inductions of the lamB gene generated a subpopulation of phenotypically sensitive cells that sustained a small population of phages that continued to utilize the *lamB* receptor. Third, perhaps due to the diminished availability of *lamB* receptors, λ phage evolved mutations in the J protein that improved its performance on the *lamB* receptor. Fourth, it is these mutations that are required for the fourth and final mutation to have occurred, thus enabling the phage to target the new receptor, *ompF*. Interestingly, this historical contingency demonstrates that the final adaptive selection for binding to *ompF* was not directly responsible for the rise of the three prior mutations. It is unclear in our study if the *ompF* gene is lost in the accessory genome only after the disruption of the production of the *lamB* protein, nevertheless the loss of an alternative phage receptor is consistent with resistance to all 22 *Vibrio* phages.

Discussion

The role of low-hanging fruits in the genetic architecture

In this work, we demonstrate that using a small sample size paired with phenotypes likely operating under strong selection can yield genome-wide significant associations. Although higher sample size is expected to reveal causal sites of lower penetrance and thus of lower effect, in this study we have captured some of the “low-hanging fruits” of very large effect. Specifically, we identified three putative causal hits with common frequency in the *V. breoganii* population. Two of the three genes, the *lamB* phage receptor and the sugar transferase that alters the O-antigen binding site of various LPS receptors, are known phage resistance mechanisms in gram-negative bacteria. The third putative causal gene in a hypothetical protein is perhaps a common mechanism whose role is yet to be revealed for phage resistance in *V. breoganii*. An examination of the primary and secondary hits shown in Table 1 highlights the role that they play in the phage adsorption phase of infection. Aside from the unknown gene identified and

the two permeases, all hits either play a direct role in binding as receptors (*lamB*, *ompF*, efflux pump) or indirectly as disruptors of binding by modifying the receptor (sugar transferase, *malM*, tyrosine kinase/phosphatase). Moreover, all the above mechanisms plus the two permease structures highlight how critical the integrity of the bacterial cell wall, both inner and outer membranes, is to viral infection.

Although there are still other causal loci to discover, the discovery of low-hanging fruits in our small sample study begins to reveal a genetic architecture of these phage phenotypes that is reminiscent of the omnigenic model of complex traits [66]. In this model, a small number of "core genes" that can have the strongest single effects on the traits, are followed by a larger number of "peripheral genes" of smaller effect that together contribute a larger proportion of heritability; we note that this latter point on heritability remains unclear in the genetic architecture of phage resistance. In addition, this genetic architecture has also been observed in the cancer genome landscape for common forms of human cancer where a small set of "mountains", genes altered in a high percentage of tumors, are followed by a much larger number of "hills", genes altered infrequently [67]. Moving beyond low-hanging fruits by applying larger sample sizes, we expect to observe higher levels of allelic and locus heterogeneity in causal loci due to the bewildering diversity of mechanisms of resistance to phage infection.

Limitations of our study and future directions

Aside from the obvious need to increase the number of host samples, this study does not address variation outside of either biallelic SNVs in the core genome or whole gene presence or absence in the accessory genome. Small indels within the core genome were not included in the association analysis. These indels may include spacer elements as part of the broader requisite CRISPR machinery of the host to defend against viral attack [68]. In addition, allelic diversity in the accessory genome was not examined. It is possible that aside from whole gene absence, accessory genes can contain SNVs that allow them to play an altered role in the phenotype. In our study, we lacked the host sample diversity to observe more homoplasic mutations in portions of

the tree. These portions likely include the three recombination hotspots identified surrounding our three primary hits. Furthermore, the accuracy of the topology of the inferred tree is critical to identifying homoplasies [17]. Because *Vibrio* genomes have undergone substantial historical recombination, a future study could attempt to identify these recombination tracts and remove them from the set of variants used for tree inference to better capture clonal ancestry. For a future study, increasing the number of host samples can potentially bring the rest of the unused phages in the 22-phage panel for use as individual phenotypes. For instance, for phage 1.206.O no plaques were observed (i.e., all host samples were resistant to this phage) and as such there currently are no host sample controls (Table S1). Another avenue to explore may be to look for correlations between multiple phages to identify any pleiotropic causal loci. In this study, we show two potential examples. First, the sugar transferase primary hit identified in the all-resistant phenotype may cause modifications to multiple LPS receptors that disrupt adsorption to different phages. Second, the two secondary hits in the tyrosine kinase/phosphatase system may play a role in altering the cell wall structure which impedes binding in both phages 2.117.O and 1.117.O. A more complete picture of the genetic architecture in V. *breoganii* may reveal other phages that share similar mechanisms underlying resistance to viral infection.

A recent study using 19 *V. lentus* genomes showed that these near-clonal host genomes were differentially infected by a panel of 22 lytic siphovirus phages [69]. Specifically, the phage broke down into two groups, with each infecting a different set of hosts. The authors reported discovering phage defense genes clustered together in large and highly diverse mobile genetic elements in the accessory genome (called phage defense elements or PDEs). Crucially, three PDEs were specific to one group of hosts while two other PDEs were specific to the remaining hosts. In *V. lentus,* these PDEs were primarily responsible for the differing phage resistance phenotypes between the two host groups, while phage receptors played a secondary role due to the near-clonality of the host genomes. In contrast, we showed that part of the genetic architecture of phage resistance in *V. breoganii* involved constituents of the cell wall

including potential phage receptors. This finding is similar to the results from some studies using experimental populations that showed receptors to play a primary role. We caution that any conclusions regarding the genetic architecture revealed thus far in our study must be interpreted in the context of a small sample size and that a larger study could potentially reveal more diverse mechanisms of resistance, perhaps even in support of similar phage defense elements identified in mobile elements of the accessory genome. Alternatively, it is possible that there are species-specific differences in genetic architecture, particularly when considering less clonal populations, like the *V. breoganii* population presented in this study, compared to the near-clonal 19 *V. lentus* genomes. Moreover, examining populations from vastly differing evolutionary time scales or differing recombination capabilities may also potentially show varying architectures to phage resistance not yet revealed in experimental evolution studies or the few studies in natural populations.

Methods

Sample collection, genotyping, phenotyping, and plaque assays

Both bacterial and viral samples were collected from the littoral marine zone at Canoe Cove, Nahant, Massachusetts, USA. The details of the sample collection are in [12]. Briefly, the bacterial samples were collected using size-fractionation and selective-medium cultivation-based methods described in [70]. The protocol for the preparation of bacterial genomic libraries and sequencing using Illumina HiSeq is described in [71]. Viral samples were collected using a previously described iron flocculation approach described in [72]. Sequencing of viral samples, using both Illumina MiSeq and HiSeq, and plaque assays are described in [12].

Accessory Genome

Genome assemblies were done using CLC Genomics Workbench v6.5.1 and v8.5.1 and CLC assembly cell v4.4.2.133896. Annotations were done using both Prokka [73] and Rast [74] programs on default settings. The accessory genome was built using default settings in Roary [75].

Allele-counting

All allele-counting associations were conducted using PLINK 1.9 [18]. For all runs, --chr-set -1 was used to designate a single haploid genome.

Principal component analysis was done using –pca 20 header. The first two principal components used for population stratification correction were incorporated as covariates in logistic regression using the following settings:

--chr-set -1
--allow-no-sex
--no-fid
--no-parents
--pheno
--all-pheno
--logistic mperm=100000
--covar
--covar-name PC1-PC2
--maf 0.20

All final p-values were calculated using the max(T) resampling scheme of Westfall and Young [76].

Homoplasy counting

The phylogenetic inference for the input tree to POUTINE was done using raxml-ng using --model GTR+G settings [77]. POUTINE was run using default settings [17].

LD profiles

All r2 values were calculated using PLINK and the following options:

--allow-no-sex

--no-fid

--no-parents

--chr-set -1

--r2

--ld-snp sample ID names for each primary hit

--ld-window 265936

--ld-window-kb 5000

--ld-window-r2 0

References

1. Keen EC. A century of phage research: bacteriophages and the shaping of modern biology. Bioessays. 2015;37: 6–9.

2. Wittebole X, De Roock S, Opal SM. A historical overview of bacteriophage therapy as an alternative to antibiotics for the treatment of bacterial pathogens. Virulence. 2014;5: 226–235.

3. Clokie MR, Millard AD, Letarov AV, Heaphy S. Phages in nature. Bacteriophage. 2011;1: 31–45.

4. Labrie SJ, Samson JE, Moineau S. Bacteriophage resistance mechanisms. Nat Rev Microbiol. 2010;8: 317–327.

5. Luong T, Salabarria A-C, Roach DR. Phage therapy in the resistance era: Where do we stand and where are we going? Clin Ther. 2020;42: 1659–1680.

6. Bishop-Lilly KA, Plaut RD, Chen PE, Akmal A, Willner KM, Butani A, et al. Whole genome sequencing of phage resistant Bacillus anthracis mutants reveals an essential role for cell surface anchoring protein CsaB in phage AP50c adsorption. Virol J. 2012;9: 246.

7. Meyer JR, Dobias DT, Weitz JS, Barrick JE, Quick RT, Lenski RE. Repeatability and contingency in the evolution of a key innovation in phage lambda. Science. 2012;335: 428–432.

8. Kulikov EE, Golomidova AK, Prokhorov NS, Ivanov PA, Letarov AV. High-throughput LPS profiling as a tool for revealing of bacteriophage infection strategies. Sci Rep. 2019;9: 2958.

9. Kortright KE, Chan BK, Turner PE. High-throughput discovery of phage receptors using transposon insertion sequencing of bacteria. Proc Natl Acad Sci U S A. 2020;117: 18670–18679.

10. Mutalik VK, Adler BA, Rishi HS, Piya D, Zhong C, Koskella B, et al. High-throughput mapping of the phage resistance landscape in E. coli. PLoS Biol. 2020;18: e3000877.

11. Moller AG, Winston K, Ji S, Wang J, Hargita Davis MN, Solís-Lemus CR, et al. Genes Influencing Phage Host Range in Staphylococcus aureus on a Species-Wide Scale. mSphere. 2021;6. doi:10.1128/mSphere.01263-20

12. Kauffman KM, Hussain FA, Yang J, Arevalo P, Brown JM, Chang WK, et al. A major lineage of non-tailed dsDNA viruses as unrecognized killers of marine bacteria. Nature. 2018;554: 118–122.

13. Beaz Hidalgo R, Cleenwerck I, Balboa S, Prado S, De Vos P, Romalde JL. Vibrio breoganii sp. nov., a non-motile, alginolytic, marine bacterium within the Vibrio halioticoli clade. Int J Syst Evol Microbiol. 2009;59: 1589–1594.

14. Corzett CH, Elsherbini J, Chien DM, Hehemann J-H, Henschel A, Preheim SP, et al. Evolution of a vegetarian Vibrio: Metabolic specialization of Vibrio breoganii to macroalgal substrates. J Bacteriol. 2018;200. doi:10.1128/JB.00020-18

15. Keymer DP, Boehm AB. Recombination Shapes the Structure of an Environmental Vibrio cholerae Population. Appl Environ Microbiol. 2011;77: 537–544.

16. Smith NH, Gordon SV, de la Rua-Domenech R, Clifton-Hadley RS, Hewinson RG. Bottlenecks and broomsticks: the molecular evolution of Mycobacterium bovis. Nat Rev Microbiol. 2006;4: 670–681.

17. Chen PE, Jesse Shapiro B. Classic genome-wide association methods are unlikely to identify causal variants in strongly clonal microbial populations. bioRxiv. 2021. p. 2021.06.30.450606. doi:10.1101/2021.06.30.450606

18. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4: 559.

19. Chen PE, Shapiro BJ. The advent of genome-wide association studies for bacteria. Curr Opin Microbiol. 2015;25: 17–24.

20. Klebba PE. Mechanism of maltodextrin transport through LamB. Res Microbiol. 2002;153: 417–424.

21. Thoma J, Ritzmann N, Wolf D, Mulvihill E, Hiller S, Müller DJ. Maltoporin LamB unfolds β hairpins along mechanical stress-dependent unfolding pathways. Structure. 2017;25: 1139-1144.e2.

22. Randall-Hazelbauer L, Schwartz M. Isolation of the bacteriophage lambda receptor from Escherichia coli. J Bacteriol. 1973;116: 1436–1446.

23. Clément JM, Hofnung M. Gene sequence of the λ receptor, an outer membrane protein of E. coli K12. Cell. 1981;27: 507–514.

24. Charbit A, Clement JM, Hofnung M. Further sequence analysis of the phage lambda receptor site. Possible implications for the organization of the lamB protein in Escherichia coli K12. J Mol Biol. 1984;175: 395–401.

25. Wang J, Michel V, Hofnung M, Charbit A. Cloning of the J gene of bacteriophage lambda, expression and solubilization of the J protein: first in vitro studies on the interactions between J and LamB, its cell surface receptor. Res Microbiol. 1998;149: 611–624.

26. Burmeister AR, Sullivan RM, Lenski RE. Fitness costs and benefits of resistance to phage lambda in experimentally evolved Escherichia coli. Evolution in Action: Past, Present and Future. Cham: Springer International Publishing; 2020. pp. 123–143.

27. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007;447: 661–678.

28. Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nat Genet. 2003;33: 177–182.

29. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461: 747–753.

30. Burmeister AR, Fortier A, Roush C, Lessing AJ, Bender RG, Barahman R, et al. Pleiotropy complicates a trade-off between phage resistance and antibiotic resistance. Proc Natl Acad Sci U S A. 2020;117: 11207–11216.

31. Azam AH, Hoshiga F, Takeuchi I, Miyanaga K, Tanji Y. Analysis of phage resistance in Staphylococcus aureus SA003 reveals different binding mechanisms for the closely related Twort-like phages ϕSA012 and ϕSA039. Appl Microbiol Biotechnol. 2018;102: 8963–8977.

32. Moller AG, Lindsay JA, Read TD. Determinants of Phage Host Range in Staphylococcus Species. Appl Environ Microbiol. 2019;85. doi:10.1128/AEM.00209-19

33. Xiong Z, Wang Y, Dong Y, Zhang Q, Xu X. Cyanophage A-1(L) adsorbs to lipopolysaccharides of Anabaena sp. Strain PCC 7120 via the tail protein lipopolysaccharide-interacting protein (ORF36). J Bacteriol. 2019;201. doi:10.1128/JB.00516-18

34. Bertozzi Silva J, Storms Z, Sauvageau D. Host receptors for bacteriophage adsorption. FEMS Microbiol Lett. 2016;363: fnw002.

35. Rakhuba DV, Kolomiets EI, Dey ES, Novik GI. Bacteriophage receptors, mechanisms of phage adsorption and penetration into host cell. Pol J Microbiol. 2010;59: 145–155.

36. Afzal M, Shafeeq S, Manzoor I, Kuipers OP. Maltose-Dependent Transcriptional Regulation of the mal Regulon by MalR in Streptococcus pneumoniae. PLoS One. 2015;10: e0127579.

37. Thirion JP, Hofnung M. On some genetic aspects of phage lambda resistance in E. coli K12. Genetics. 1972;71: 207–216.

38. Sakoparnig T, Field C, van Nimwegen E. Whole genome phylogenies reflect the distributions of recombination rates for many bacterial species. Elife. 2021;10. doi:10.7554/eLife.65366

39. Saier MH Jr. Molecular phylogeny as a basis for the classification of transport proteins from bacteria, archaea and eukarya. Adv Microb Physiol. 1998;40: 81–136.

40. Guan L, Kaback HR. Lessons from lactose permease. Annu Rev Biophys Biomol Struct. 2006;35: 67–91.

41. Erni B, Zanolari B, Kocher HP. The mannose permease of Escherichia coli consists of three different proteins. Amino acid sequence and function in sugar transport, sugar phosphorylation, and penetration of phage lambda DNA. J Biol Chem. 1987;262: 5238–5247.

42. Esquinas-Rychen M, Erni B. Facilitation of bacteriophage lambda DNA injection by inner membrane proteins of the bacterial phosphoenol-pyruvate: carbohydrate phosphotransferase system (PTS). J Mol Microbiol Biotechnol. 2001;3: 361–370.

43. Saris P. The ptsL, pel/ptsM (manXYZ) locus consists of three genes involved in mannose uptake in Escherichia coli K12. FEMS Microbiol Lett. 1987;44: 371–376.

44. Nir-Paz R, Eugster MR, Zeiman E, Loessner MJ, Calendar R. Listeria monocytogenes tyrosine phosphatases affect wall teichoic acid composition and phage resistance. FEMS Microbiol Lett. 2012;326: 151–160.

45. Peleg A, Shifrin Y, Ilan O, Nadler-Yona C, Nov S, Koby S, et al. Identification of an Escherichia coli operon required for formation of the O-antigen capsule. J Bacteriol. 2005;187: 5259–5266.

46. Ilan O, Bloch Y, Frankel G, Ullrich H, Geider K, Rosenshine I. Protein tyrosine kinases in bacterial pathogens are associated with virulence and production of exopolysaccharide. EMBO J. 1999;18: 3241–3248.

47. Obadia B, Lacour S, Doublet P, Baubichon-Cortay H, Cozzone AJ, Grangeasse C. Influence of tyrosine-kinase Wzc activity on colanic acid production in Escherichia coli K12 cells. J Mol Biol. 2007;367: 42–53.

48. Danese PN, Pratt LA, Kolter R. Exopolysaccharide production is required for development of Escherichia coli K-12 biofilm architecture. J Bacteriol. 2000;182: 3593–3596.

49. Rousset JP, Gilson E, Hofnung M. malM, a new gene of the maltose regulon in Escherichia coli K12. II. Mutations affecting the signal peptide of the MalM protein. J Mol Biol. 1986;191: 313–320.

50. Chaudhry WN, Pleška M, Shah NN, Weiss H, McCall IC, Meyer JR, et al. Leaky resistance and the conditions for the existence of lytic bacteriophage. PLoS Biol. 2018;16: e2005971.

51. German GJ, Misra R. The TolC protein of Escherichia coli serves as a cell-surface receptor for the newly characterized TLS bacteriophage. J Mol Biol. 2001;308: 579–585.

52. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, et al. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome." Proc Natl Acad Sci U S A. 2005;102: 13950–13955.

53. Lapierre P, Gogarten JP. Estimating the size of the bacterial pan-genome. Trends Genet. 2009;25: 107–110.

54. Li R, Li Y, Zheng H, Luo R, Zhu H, Li Q, et al. Building the sequence map of the human pan-genome. Nat Biotechnol. 2010;28: 57–63.

55. Sherman RM, Salzberg SL. Pan-genomics in the human genome era. Nat Rev Genet. 2020;21: 243–254.

56. Stenberg F, Chovanec P, Maslen SL, Robinson CV, Ilag LL, von Heijne G, et al. Protein complexes of the Escherichia coli cell envelope. J Biol Chem. 2005;280: 34409–34419.

57. Alva A, Sabido-Ramos A, Escalante A, Bolívar F. New insights into transport capability of sugars and its impact on growth from novel mutants of Escherichia coli. Appl Microbiol Biotechnol. 2020;104: 1463–1479.

58. Chai T, Wu V, Foulds J. Colicin A receptor: role of two Escherichia coli outer membrane proteins (OmpF protein and btuB gene product) and lipopolysaccharide. J Bacteriol. 1982;151: 983–988.

59. Harder KJ, Nikaido H, Matsuhashi M. Mutants of Escherichia coli that are resistant to certain beta-lactam compounds lack the ompF porin. Antimicrob Agents Chemother. 1981;20: 549–552.

60. Sawai T, Yamaguchi A, Saiki A, Hoshino K. OmpF channel permeability of quinolones and their comparison with beta-lactams. FEMS Microbiol Lett. 1992;74: 105–108.

61. Schirmer T, Keller TA, Wang YF, Rosenbusch JP. Structural basis for sugar translocation through maltoporin channels at 3.1 A resolution. Science (New York, N.Y.). American Association for the Advancement of Science (AAAS); 1995. pp. 512–514.

62. Kefala G, Ahn C, Krupa M, Esquivies L, Maslennikov I, Kwiatkowski W, et al. Structures of the OmpF porin crystallized in the presence of foscholine-12. Protein Sci. 2010;19: 1117–1125.

63. Koebnik R, Locher KP, Van Gelder P. Structure and function of bacterial outer membrane proteins: barrels in a nutshell. Mol Microbiol. 2000;37: 239–253.

64. Silverman JA, Benson SA. Bacteriophage K20 requires both the OmpF porin and lipopolysaccharide for receptor function. J Bacteriol. 1987;169: 4830–4833.

65. Traurig M, Misra R. Identification of bacteriophage K20 binding regions of OmpF and lipopolysaccharide in Escherichia coli K-12. FEMS Microbiol Lett. 1999;181: 101–108.

66. Boyle EA, Li YI, Pritchard JK. An expanded view of complex traits: From polygenic to omnigenic. Cell. 2017;169: 1177–1186.

67. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. Cancer Genome Landscapes. Science. 2013;339: 1546–1558.

68. Barrangou R, Horvath P. CRISPR: new horizons in phage resistance and strain identification. Annu Rev Food Sci Technol. 2012;3: 143–162.

69. Hussain FA, Dubert J, Elsherbini J, Murphy M, VanInsberghe D, Arevalo P, et al. Rapid evolutionary turnover of mobile genetic elements drives bacterial resistance to phages. Science. 2021;374: 488–492.

70. Hunt DE, David LA, Gevers D, Preheim SP, Alm EJ, Polz MF. Resource partitioning and sympatric differentiation among closely related bacterioplankton. Science. 2008;320: 1081–1085.

71. Baym M, Kryazhimskiy S, Lieberman TD, Chung H, Desai MM, Kishony R. Inexpensive multiplexed library preparation for megabase-sized genomes. PLoS One. 2015;10: e0128036.

72. John SG, Mendez CB, Deng L, Poulos B, Kauffman AKM, Kern S, et al. A simple and efficient method for concentration of ocean viruses by chemical flocculation. Environ Microbiol Rep. 2011;3: 195–202.

73. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30: 2068–2069.

74. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, et al. The RAST Server: rapid annotations using subsystems technology. BMC Genomics. 2008;9: 75.

75. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, et al. Roary: Rapid large-scale prokaryote pan genome analysis. 2015. Available: http://biorxiv.org/lookup/doi/10.1101/019315

76. Westfall PH, Stanley Young S. Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment. John Wiley & Sons; 1993.

77. Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics. 2019;35: 4453–4455.

Supplementary Material

Table S1. Panel of plaque assays using 32 V. *breoganii* bacterial genomes against 22 phages. The host bacterial genomes are by row and the phages by column. Plaque formation (*i.e.,* phage sensitivity) is indicated with '1' and empty cells indicate phage resistance.

Strain_ID	1.034.O	2.117.O.	1.139.A.	1.117.O.	1.207.B.	1.209.O	1.261.O	1.188.A.	1.224.A.	1.169.O	1.006.Z.	1.011.O.	1.008.O	1.116.O.	1.157.O	1.167.O	1.172.O	1.176.O	1.182.O	1.206.O	1.228.O	1.012.O	Total Plaques by Host Strain
222.51.E5		1		1	1																		3
261.45.B7	1	1		1	1		1								1						1		7
261.45.E9		1		1	1																		3
261.46.B7	1	1				1		1															4
261.46.C3	1		1						1														3
261.48.B1									1		1	1	1									1	5
261.48.C6	1		1			1																	3
261.48.E3		1	1	1		1					1	1	1									1	8
261.48.E5																							0
261.49.C1	1	1		1											1						1		5
261.49.E4																							0
261.49.F3	1		1			1																	3
261.51.E6	1		1													1							3
261.51.F2	1	1	1	1												1							5
261.52.A10																							0
261.52.B1			1					1		1													3
261.52.C1																							0
261.52.F10																							0
261.52.F5														1			1	1					3
261.52.F6																							0
261.52.F8																							0
261.54.B2		1		1	1																		3
261.54.C7		1																					1
261.55.F5																		1					1
286.46.E1		1		1	1					1									1				5
286.51.A6	1							1		1									1				4
286.51.A9		1		1	1		1																4
286.51.B5	1		1																				2
286.52.C12							1																1
286.52.F9																							0
286.54.A10																							0
FF-50	1		1			1																	3
Total plaques by phage strain	11	11	9	9	6	5	3	3	2	3	2	2	2	1	2	2	1	2	2	0	2	2	

Figure S1. Individual host genome contributions to the four pangenome gene classes.

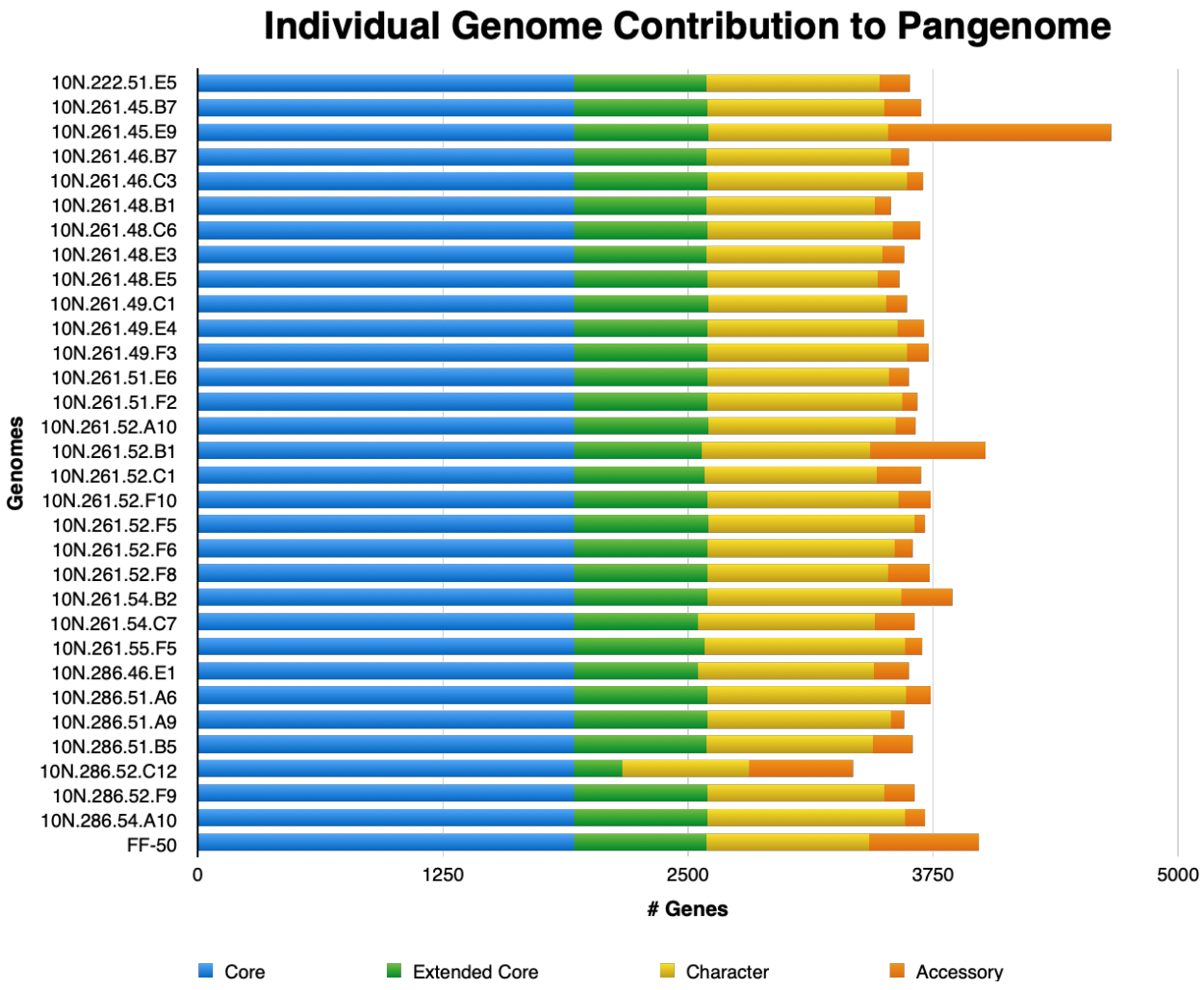


Figure S2. Example quantile-quantile plot (phenotype phage 1.034.O), showing relatively insignificant population stratification. The red line represents the x=y line.

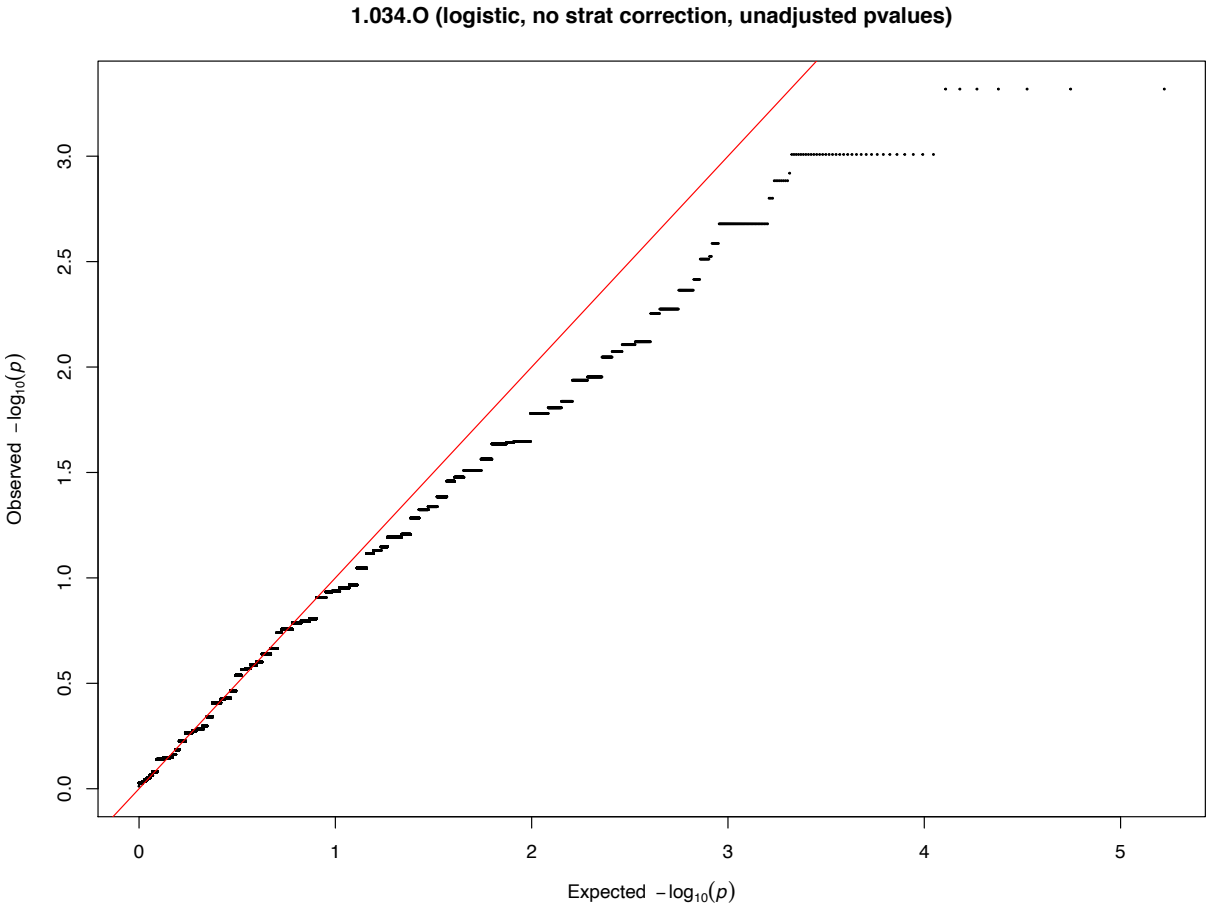


Figure S3. Population structure of 32 V. *breoganii* genomes. Score plot of the first two principal components.

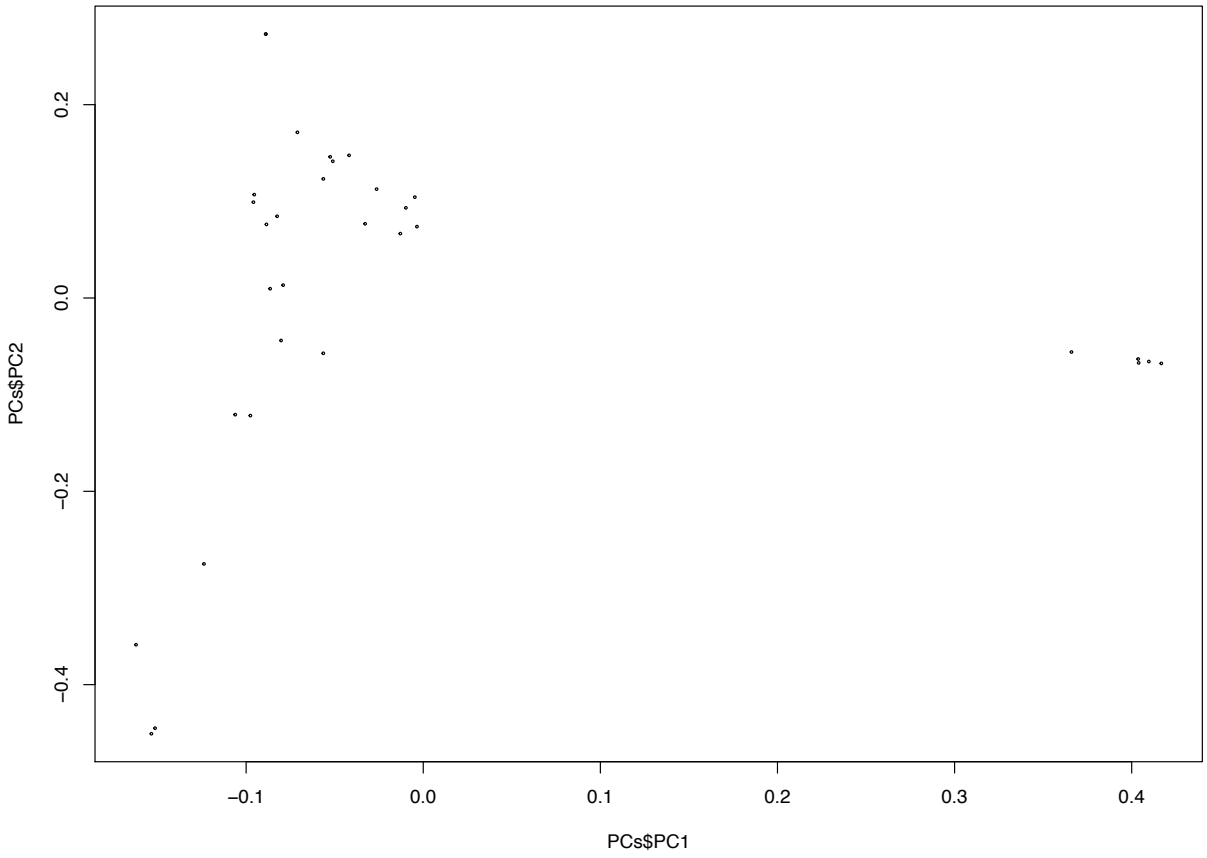


Figure S4. Phylogeny of the 32 V. *breoganii* genomes inferred using raxml-ng.

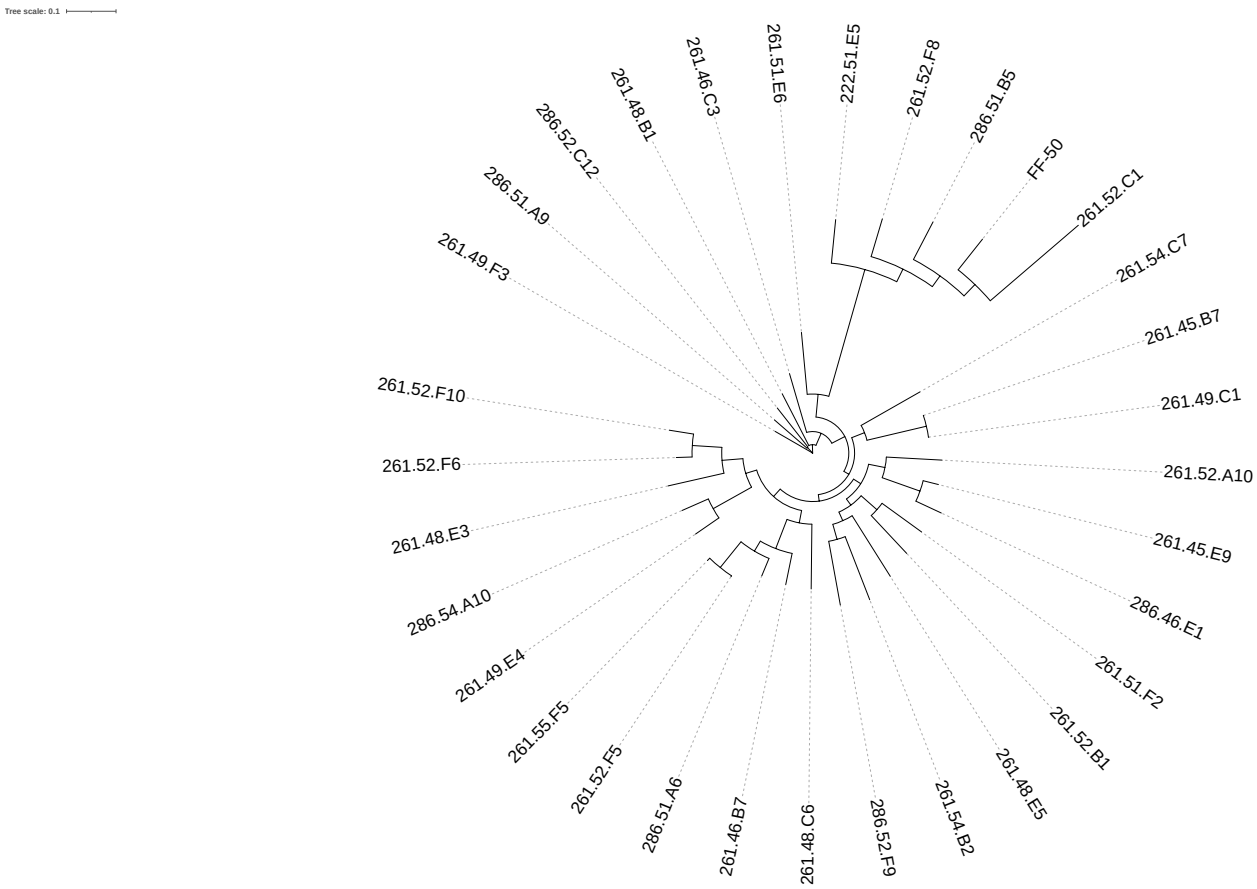


Table S2. Genome-wide significant nucleotide sites in the *lamB* core gene for phenotype 1.034.O. Alleles broken down by cases (red) and controls (blue). Host bacterial genomes are shown in the "Strains" column, with each subsequent column showing the physical position of each site relative to the 1C10 reference sequence.

Strains	1C10_1546915	1C10_1546916	1C10_1546922	1C10_1546940	1C10_1546973	1C10_1547100	1C10_1547175
	nonsyn (T->E, G->A)	nonsyn (C->E, G->Q)	nonsyn (G->R, T->S)	nonsyn (T->N, C->D)	nonsyn (T->N, C->D)	syn (G->Y, A->Y)	syn (G->R, A->R)
10N.286.54.A10	G	G	T	C	C	A	A
10N.286.52.C12	T	C	G	T	T	G	G
10N.261.45.E9	T	C	G	T	T	G	G
10N.261.55.F5	T	C	G	T	T	G	G
10N.261.52.A10	T	C	G	T	T	G	G
10N.261.54.B2	T	C	G	T	T	G	G
10N.261.52.C1	T	C	G	T	T	G	G
10N.261.48.E3	T	C	G	T	T	G	G
10N.261.49.E4	T	C	G	T	T	G	G
10N.261.52.F10	T	C	G	T	T	G	G
10N.261.52.F8	T	C	G	T	T	G	G
10N.261.48.B1	T	C	G	T	T	G	G
10N.261.48.E5	T	C	G	T	T	G	G
10N.261.52.F6	T	C	G	T	T	G	G
10N.286.52.F9	T	C	G	T	T	G	G
10N.222.51.E5	T	C	G	T	T	G	G
10N.261.52.B1	T	C	G	T	T	G	G
10N.261.52.F5	T	C	G	T	T	G	G
10N.286.46.E1	T	C	G	T	T	G	G
10N.261.54.C7	T	C	G	T	T	G	G
10N.286.51.A9	T	C	G	T	T	G	G
10N.261.51.E6	G	G	T	C	C	A	A
FF-50	G	G	T	C	C	A	A
10N.261.45.B7	G	G	T	C	C	A	A
10N.261.48.C6	G	G	T	C	C	A	A
10N.261.46.B7	G	G	T	C	C	A	A
10N.261.46.C3	G	G	T	C	C	A	A
10N.286.51.B5	G	G	T	C	C	A	A
10N.261.51.F2	T	C	G	T	T	G	G
10N.261.49.F3	G	G	T	C	C	A	A
10N.286.51.A6	T	C	G	T	T	G	G
10N.261.49.C1	G	G	T	C	C	A	A

Figure S5. Number of genes observed as a function of the number of genomes sampled. The 'total genes' trend shows that the *V. breoganii* pangenome is 'open'.

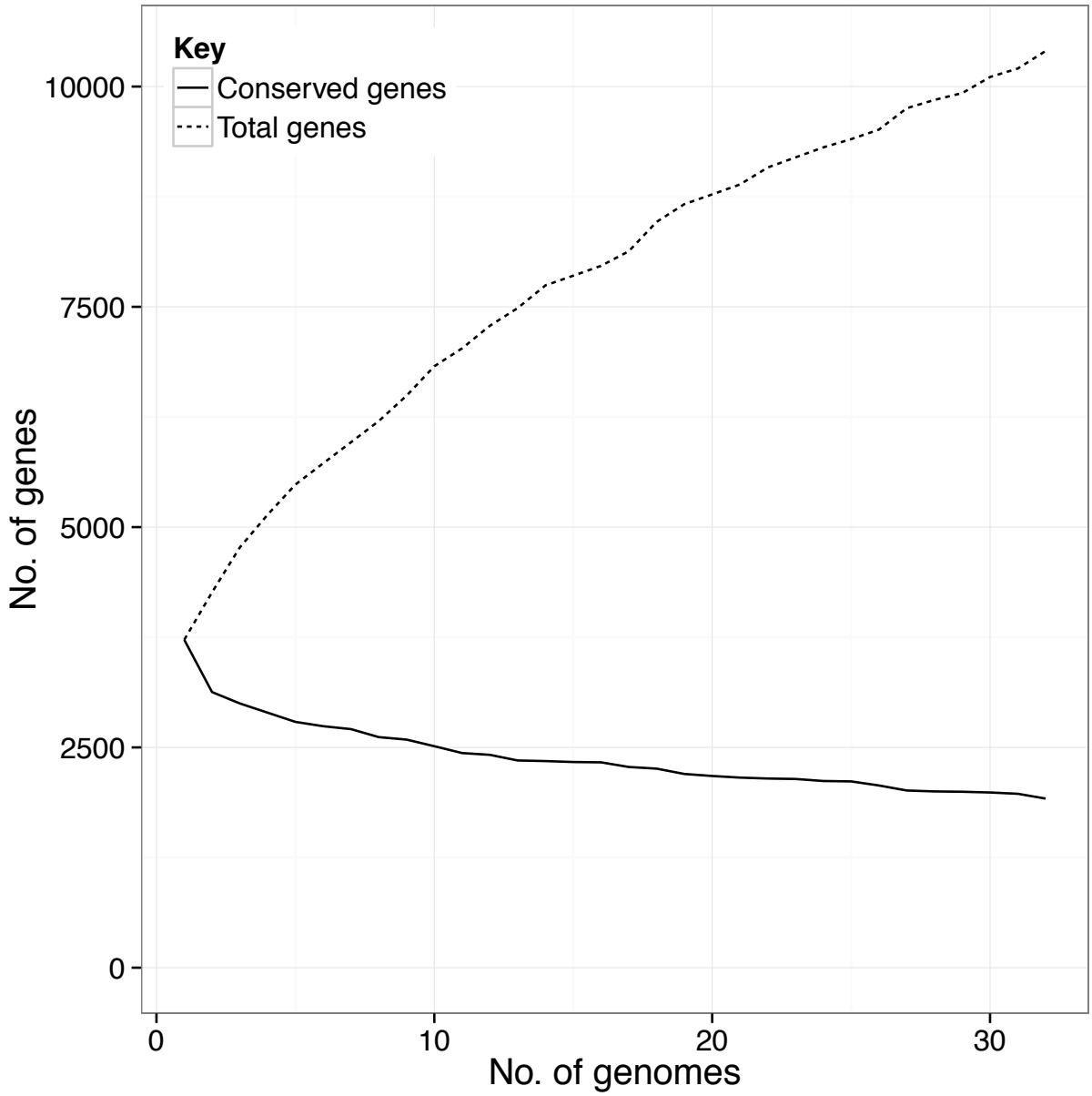
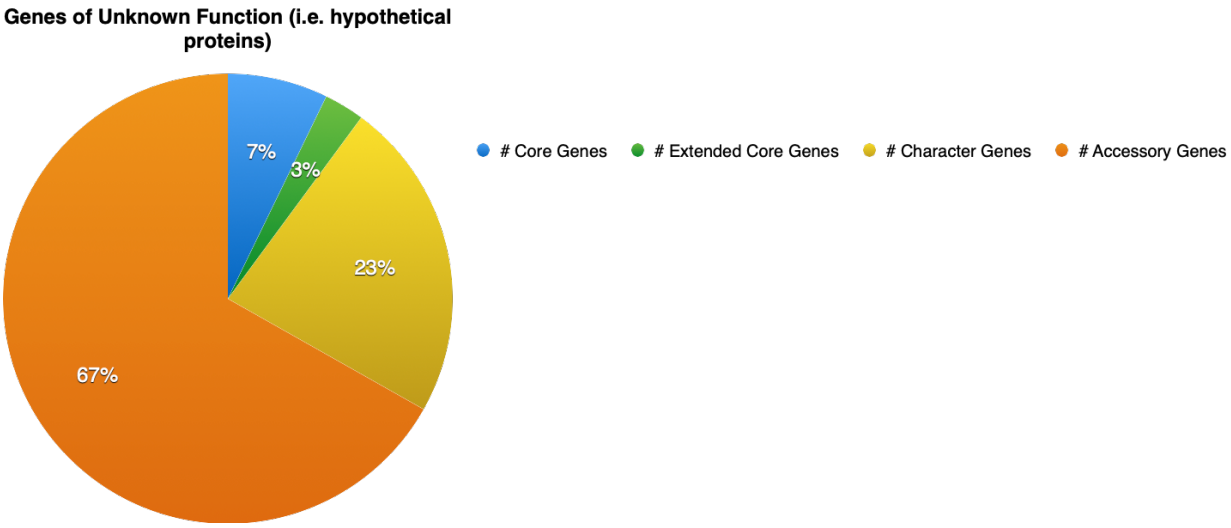


Figure S6. Proportion of genes of no known function (i.e., hypothetical protein) categorized by gene class.



Conclusions and future directions

Our contributions to the nascent field of microbial GWAS

In this thesis, we critically examined GWAS approaches and applied both allele and homoplasy counting methods to strongly clonal and highly recombining populations. In Chapter 1, we presented the origins of microbial GWAS and then took stock of how far the nascent field has progressed. Principally, we asked the question: "Do human GWAS methods readily work for microbial populations, and if not, what are the central problems and changes that are required for a more successful microbial GWAS? In answering this question, we highlighted the major differences between eukaryotic and prokaryotic GWAS; specifically, that prokaryotic genomes tend to experience strong LD, strong population stratification, and strong selection. The first difference, strong positive selection, is mostly an advantage for microbial GWAS in that smaller sample sizes will be required to achieve sufficient statistical power to identify genome-wide significant loci. Unlike human complex traits, many microbial traits under strong positive selection will likely be underlain by variants that are of common frequency and of large effect. The second difference, strong population stratification, can be reasonably dealt with using current methods developed for human GWAS, principally the capturing of population structures either using dimensionality reduction in a fixed model regression framework, or a genetic relationship matrix in a random model regression framework. No major advancements for dealing with population stratification have thus been made for microbial GWAS. We note that there is room for improvement in this area, specifically a quantitative assessment of how well dimensionality reduction and genetic relationship matrices capture fine-scale microbial population substructures. The third of the major differences, strong LD, has presented the greatest challenge. We readily saw this in the association study in a strongly clonal population in Chapter 1. In our GWAS using *M. tuberculosis*, we showed that the problem is not that an association signal cannot be detected (the strong selective

pressure from antibiotic resistance allowed association signals to be observed) but rather that the strong signal is often indistinguishable from linked sites that have hitchhiked from across the genome. In this manner, it is unclear which of the many loci were the true causal drivers of association.

From our findings in Chapter 1, it had become clear that for microbial GWAS to be successful, we needed to identify the degree to which LD limits the detection of true positive causal drivers of association. A number of microbial GWAS studies were published since our publication of Chapter 1, and many of the GWAS hits identified were often reported without any analysis of LD. In our work in Chapter 2, we critically examined allele counting methods in a strongly clonal population, and we showed that, in fact, the vast majority of genome-wide significant hits were likely false positives. Thus, our work suggested that perhaps a substantial proportion of previously reported GWAS hits may be false positives. A corollary to this is that the few reported allele counting hits that are relatively unlinked from the rest of the genome have likely undergone convergent evolution. We showed an example of this in Chapter 2 using a newly developed homoplasy counting method, POUTINE, that directly addressed the challenge of LD. Specifically, in our GWAS we observed one site (site 1674048 in the *M. tuberculosis* genome) that was both a genome-wide significant allele and homoplasy counting hit.

Building from our work in Chapter 2, we used POUTINE and PLINK, as a representative allele counting method, and conducted a GWAS in a relatively highly recombining population of marine vibrios. As we theorized in Chapter 1, microbial populations may experience strong positive selection and thus show a much different genetic architecture than the negative purifying selection thought to be acting on many human complex traits. In our work in Chapter 3, we highlighted this difference by identifying three genome-wide significant putative causal loci, all of which are mutations of common frequency and high effect. These hits represent a departure from the genetic architecture observed in human complex traits where many of the reported hits are

mostly mutations of common frequency and low effect, or sometimes mutations of rare frequency but high effect [1–5]. We note that there likely exist microbial phenotypes that on the spectrum of genetic architectures may lie closer to what has been observed in human complex traits. For these phenotypes, larger sample sizes will be required to capture either mutations of low effect or rare in frequency. However, no sample size will overcome the problem presented by LD.

Future directions

A recurring theme in this thesis is the important role that strong LD from microbial recombination dynamics played in various aspects of microbial GWAS. Namely, the role of LD in obscuring the association signal, and the role of LD in multiple hypothesis testing, where correlated tests can exacerbate the already stringent criterion of controlling for the familywise error rate. The following three projects represent the continuation of the work presented in this thesis, and all three projects highlight either the challenges or even the benefits of strong microbial LD.

Genome-wide LD landscapes across the gamut of microbial recombination rates

Perhaps our most important finding is how LD presents a limit on which microbial populations may be suitable for allele counting methods. There is a pressing need to examine the LD landscape and recombination patterns in more highly recombining populations. Specifically, how prevalent are block-like LD structures that are localized and unlinked from the rest of the genome? If the answer to this question is that in more highly recombining populations there do exist block-like LD structures throughout the genome (like those found in human genomes), then we can similarly deploy allele counting methods as seen in human GWAS. If the answer is that there exists a sparsity of LD blocks, perhaps all isolated to only a few recombination hotspots, then even for highly recombining populations we must develop new

methodological extensions to allele counting in order to discover non-convergent causal loci. Although homoplasy counting does bypass LD, it is completely reliant on convergent evolution for any association signal, and it is unclear how often this form of evolution occurs. We can answer this question by examining empirical datasets of varying recombination rates [6,7]. Because there is a plethora of microbial populations yet to be studied, we can probe the limits to which recombination can shape block-like LD structures using simulations of populations evolving under varying recombination parameters that stretch the gamut of values beyond what has been reported in the literature. Such parameters should likely include the rate of recombination, the length of recombination tracts, and the spatial localization of recombination tracts.

LD-based tests of convergence

Based upon our work in Chapter 2, we have begun formulating a new method to identify convergent sites. The crux of this method is that sites undergoing convergent evolution may often be partially or fully unlinked from the rest of the genome. We observed this feature in the analysis of LD presented in Chapter 2. Thus, a future method may directly use measures of LD to identify convergent sites. Such an LD-based method has a principal advantage over POUTINE in that a phylogeny is not required for the identification of homoplasic mutations. This feature is most appealing in populations where tree inference is problematic either due to rampant recombination interfering with the signal from clonal ancestry, or due to a lack of a sufficient number of mutations for accurate resolving of branching. This latter point has been observed in tree inferences for SARS-CoV-2 causing debate over whether many homoplasies identified are false positives [8–10]. One intriguing possibility is to use such an LD-based method for SARS-CoV-2 to identify variants of concern that have so far been mostly convergent. This phenotype-free use of the method is appealing because a GWAS using SARS-CoV-2 can be problematic if the phenotype cannot be easily measured. A timely example of such a phenotype is the transmissibility of the virus, where quantification of transmission levels of new variants has proven difficult due to the potential level of confounding caused by human behavior and public health

measures being unequally enacted across geography [11,12]. Another intriguing possibility is to combine such an LD-based convergence test with the GWAS framework we developed for POUTINE. For traits that can be feasibly measured, it is appealing to rely on two disparate sources of convergent sites, where concordance between tree-based convergence and LD-based convergence suggests a more robust hit, while discordance between the two sources has the potential for revealing convergent sites that would otherwise have gone undetected by only one method. Concretely, a nucleotide site that is completely unlinked from the rest of the genome is easily identified by the LD-based method, while a tree-based method may potentially miss this convergent site if the topology of the tree is inaccurate and homoplasic mutations at this site are missed. Conversely, a hybrid site harboring both homoplasies and non-homoplasies (one such example is site 1674048 as presented in Chapter 2) can potentially show strong linkage to many other sites and be missed by the LD-based method, while a tree-based method with an accurate topology may have identified homoplasic mutations at this site.

Set-based testing in POUTINE

And finally, our next planned enhancement of POUTINE is to develop a set-based test that can capture allelic heterogeneity. A "set" could be any set of variants within a meaningful genetic region (e.g., a gene), and the test would assess the cumulative effects of the variants in this region. This addition will boost our ability to detect causal loci when multiple convergent sites underlie a single locus, particularly when the sample size is insufficient to capture any individual convergent site. This strategy is akin to various set-testing approaches used in GWAS focused on rare variants [13]. One strategy to explore is integrating a rare variant testing program as an addition to the max(T) resampling scheme used in POUTINE. One such option is the popular SKAT-O program [14] which unifies both burden and kernel-based classes of rare variant testing thus optimizing for a distribution of causal variants that is unknown ahead of time [15]. A second strategy to accomplish set-testing for POUTINE is to explore integrating the harmonic mean p-value test [16]. This method belongs to the

family of approaches best exemplified by the well-known Fisher's combined probability test [17], where the method aggregates multiple hypothesis tests into one composite test. While Fisher's test assumed independence between tests, the harmonic mean p-value is designed to consider dependent tests. A third strategy forward, inspired by the two approaches above, is to consider expanding our max(T) resampling framework to include all homoplasies in a particular region. Variants could be collapsed into one burden variable prior to hypothesis testing (similar to burden testing) or variants could be aggregated after their pointwise p-value estimates have been determined during resampling (similar to kernel-based testing and the harmonic mean p-value).

A note on ethics

As we continue to advance our capability to dissect microbial phenotypes, this progress must go hand-in-hand with our sense of ethics surrounding the proper use of such knowledge. Once Nature's functions have been reverse-engineered, it is entirely possible for humans to bootstrap off this knowledge and construct designs of their own. In this age of genome editing, it is easy to imagine a multitude of nefarious uses one can design if the blueprints for various phenotypes were readily available. To read is to write. As with many potent technologies, they can serve us or harm us. I close this thesis by unequivocally stating that ethics must never be a distant second to this endeavor and must be placed at the tip of the research spear. Humans should not always build what they can, and instead, build what they should.

References

1. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461: 747–753.

2. Park J-H, Gail MH, Weinberg CR, Carroll RJ, Chung CC, Wang Z, et al. Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. Proc Natl Acad Sci U S A. 2011;108: 18026–18031.

3. Simons YB, Bullaughey K, Hudson RR, Sella G. A population genetic interpretation of GWAS findings for human quantitative traits. PLoS Biol. 2018;16: e2002985.

4. Gazal S, Loh P-R, Finucane HK, Ganna A, Schoech A, Sunyaev S, et al. Functional architecture of low-frequency variants highlights strength of negative selection across coding and non-coding annotations. Nat Genet. 2018;50: 1600–1607.

5. O'Connor LJ, Schoech AP, Hormozdiari F, Gazal S, Patterson N, Price AL. Extreme polygenicity of complex traits is explained by negative selection. Am J Hum Genet. 2019;105: 456–476.

6. Vos M, Didelot X. A comparison of homologous recombination rates in bacteria and archaea. ISME J. 2009;3: 199–208.

7. Feil EJ, Holmes EC, Bessen DE, Chan MS, Day NP, Enright MC, et al. Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. Proc Natl Acad Sci U S A. 2001;98: 182–187.

8. Morel B, Barbera P, Czech L, Bettisworth B, Hübner L, Lutteropp S, et al. Phylogenetic analysis of SARS-CoV-2 data is difficult. Mol Biol Evol. 2021;38: 1777–1791.

9. Jo Y-S, Tamuri AU, Towers GJ, Goldstein RA. SARS-CoV-2 convergent evolution cannot be reliably inferred from phylogenetic analyses. bioRxiv. bioRxiv; 2021. doi:10.1101/2021.05.15.444301

10. Turakhia Y, De Maio N, Thornlow B, Gozashti L, Lanfear R, Walker CR, et al. Stability of SARS-CoV-2 phylogenies. PLoS Genet. 2020;16: e1009175.

11. van Dorp L, Richard D, Tan CCS, Shaw LP, Acman M, Balloux F. No evidence for increased transmissibility from recurrent mutations in SARS-CoV-2. Nat Commun. 2020;11: 5986.

12. Obermeyer F, Schaffner SF, Jankowiak M, Barkas N, Pyle JD, Park DJ, et al. Analysis of 2.1 million SARS-CoV-2 genomes identifies mutations associated with transmissibility. bioRxiv. 2021. doi:10.1101/2021.09.07.21263228

13. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. Am J Hum Genet. 2014;95: 5–23.

14. Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. Am J Hum Genet. 2012;91: 224–237.

15. Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, et al. Testing for an unusual distribution of rare variants. PLoS Genet. 2011;7: e1001322.

16. Wilson DJ. The harmonic mean p-value for combining dependent tests. Proc Natl Acad Sci U S A. 2019;116: 1195–1200.

17. Fisher RA. Statistical methods for research workers. 5th ed. Oliver and Boyd: Edinburgh; 1934.