UNIVERSITY OF CALIFORNIA, SAN DIEGO

Identifying and Accommodating Context Dependent Effects in Studies of Genetic Variation and Human Disease

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Biomedical Sciences

by

Danjuma Quarless

Committee in charge:

Professor Nicholas Schork, Chair
Professor Richard Kolodner, Co-Chair
Professor John Kelsoe
Professor Victor Nizet
Professor Bing Ren

2017

The dissertation of Danjuma Quarless is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____
Co-Chair

_____
Chair

University of California, San Diego

2017

DEDICATION

This thesis is dedicated most importantly to my family and friends. Beyond contributing my drop in the scientific bucket, my motivation for completing this degree was demonstrate that goals are achievable. After my family, I'd like to address the incredible support from an incredible group of individuals which as guided me this far. Gwen White and the entire community at Bellarmine who uplifted my intellectual abilities. Dr. Whitehouse, Tim Herron, Erin Jones, Dr. Kent Jones, Dr. Pond who first introduced me to genomics and the entire Whitworth community who challenged me in the perfect combination of support and doubt. Coach Toby and Coach Basket who solidified my character and Dr. Brown and Dr. Whitman for investing in me during the summer that changed my life. Travis Styles, Erick Scott, the Schork Lab, and the BMS program at large: they listened, guided, and supported me throughout the entire process. Jessica who's become my anchor, where I started this journey without her, however there's no way I could have finished in the same state. Thank you for dealing with me at all time.

I'd like especially to express my sincere gratitude for the patience, tutelage, and support of Dr. Nicholas Schork. I could write an additional thesis chapter to describe the lessons that I've acquired from his guidance over the years. I thank you for being the scientist, individual, and mentor that you've been, and for shaping my life for the better.

Finally and most importantly, I'd like to dedicate this degree to my

mother who passed away before my completing this program and was unable to participate during that time. I know she was more excited than anyone to see where this journey will lead, and I'm sure that she still is. I press forward in her memory to be an individual that will continue to make her proud.

EPIGRAPH

One good thing about music, when it hits you, you feel no pain.
— Bob Marley

Science is hard.
— Erick Scott

Searching
— Roy Ayers

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Chapter 2, in full is currently being prepared for submission for publication, Liu X.*, Pfeiffer W.*, Quarless D., Lee J., Oliveira G., Diamant J., and Schork N. "Clinical Bacterial Species Determination by Whole-Genome Sequencing: A Proof of Concept Study".

Chapter 3, in full is currently being prepared for submission for publication, Quarless Q., Liu X.*, Pfeiffer W.*, Lee J., Oliveira G., and Schork N. "Implications of Methicillin-Resistant Staphylococcus Aureus (MRSA) Reference Genome Choice for Investigating Clinical Correlates". The dissertation author is the primary researcher and author on this paper.

Chapter 4, in full is currently being prepared for submission for publication, Quarless Q., Mitra I., ADGC GROUP, Schellenberg G., and Schork N. "Comprehensive Gene Expression-Based Mediator-Wide Association Study of Alzhei-mer's Disease". The dissertation author is the primary researcher and author on this paper.

VITA

2010	B.A. in Mathematics and Computational Biology, Whitworth University
2013	San Diego Foundation Match Fellow, University of California, San Diego
2013-2014	Howard Hughes Medical Institute, Med-Into-Grad Fellow, University of California, San Diego
2013-2014	Science Bridge Socrates Fellow, University of California, San Diego
2017	Ph.D. in Biomedical Sciences, University of California, San Diego

Danjuma Quarless, "Epigenomic Diversity in a Global Collection of *Arabidopsis thaliana* Accessions.", *Cell*, 166(2), 2016.

Danjuma Quarless, "Admixture and clinical phenotypic variation", *Human Heredity*, 77(1-4), 2014.

ABSTRACT OF THE DISSERTATION

Identifying and Accommodating Context Dependent Effects in Studies of Genetic Variation and Human Disease

by

Danjuma Quarless

Doctor of Philosophy in Biomedical Sciences

University of California, San Diego, 2017

Professor Nicholas Schork, Chair
Professor Richard Kolodner, Co-Chair

Genetic variants, or changes in DNA sequence, are known to contribute to both complex and Mendelian diseases. The identification of individual and collections of variants, both common and rare, associated with diseases can help elucidate pathogenic mechanisms contributing to those diseases, since it is known that genetic variants can impact gene function and drive pathophysiology. Unfortunately, there is no consensus on the best strategies for identifying genetic associations and effects. In fact, many methods simply involve testing each variant in the genome for association with a trait directly, and ignore the fact that most molecular and physiological systems are quite complex and involve a number of interacting parts. In this light,

the effect of any one variant may be masked by, or interact with, other variants and phenomena (such as environmental factors). This is a likely reason why many attempts to identify genetic variants associated with most diseases have not been able to explain the majority of the heritable component of those diseases. It is therefore important to consider genetic association analysis methods that are sensitive to the fact that genetic variants may exhibit effects that are "context dependent" in that their effects depend on the existence of other variants or environmental factors.

Quantifying the extent to which genetic variants interact with other factors remains a challenge in genetic studies. This is the case despite the fact that there have been numerous historical studies exposing the existence of context dependent genetic effects in very broad settings that should motivate greater concern for context dependency in modern genetic association studies. For example, many model organism studies, highly contrived *in vitro* studies, studies of tumor responsiveness to targeted therapies, and general clinical studies of monogenic diseases have all suggested that the phenotypic impact of certain genetic factors is dependent on other factors. We believe that ignoring the genetic and overall context within which a genetic variant is operating can negatively impact understanding disease pathogenesis and human biology.

In the following, we explore two broad settings in which genetic background and context can have an effect on the interpretation of the impact of genetic variation on a clinically meaningful phenotype. The first setting involves associating genetic variation exhibited by the pathogen Methicillin-Resistant *Staphylococcus Aureus* (MRSA) and the clinical outcomes of patients harboring an infection induced by that pathogen. Essentially, the current manner in which MRSA genetic variants are identified requires the choice of a reference strain genome whose genetic background relative to the strains of interest could influence the characterization, association and interpretation of the impact of those variants. The second setting considers the identification of genetic factors that collectively influence Alzheimer's Disease (AD) in a manner that is dependent on the genetic background of the individuals studied through intermediate phenotypes. We ultimately believe that the approaches and findings in our work should motivate further research and a sensitivity to the numerous contexts in which genetic variants may impact phenotype development.

Chapter 1

Context Dependent Effect Considerations For Genomics Research

1.1 General Introduction

Genetics studies seeking to identify casual factors associated with human disease are often highly problematic, since many confounding factors complicate their implementation and interpretation. These confounding factors are rooted in the biological and genetic complexity of disease pathogenesis but are exacerbated further in studies leveraging DNA sequencing and genotyping protocols. For example, DNA extraction, sample processing, variant identification techniques, etc. as well as statistical analysis methods that can't possibly control for all relevant parameters that might affect a study's results. As a result, substantial interaction effects involving additional factors influencing disease and context dependent biological effects, which in many instances often go unnoticed, have been identified in recent investigations. These elements are important to consider in the elucidation of the contribution of various factors to natural phenotype variation, particularly in the context of whole genome sequencing (WGS) and genome-wide association studies (GWAS) [1,2].

A prime motivation for the work outlined in this thesis is that it addresses the shortcomings of current genome-wide association study (GWAS) methodologies to

explain the heritable portion of most common diseases[3–11]. Genetic variations, single nucleotide variation (SNVs) in particular, are known to have small individual effects on disease susceptibility but often interact in subtle ways. Thus, more appropriate methods are needed to understand the full context in which genetic factors impact disease, above-and-beyond the individual effects of SNVs.

The context dependency of the effects of genetic factors is particularly important in precision medicine (PMed), through investigation of either complex or Mendelian disease, since PMed requires the identification of drug targets that could be specific SNVs and structural mutations. The assumption, however, is that those genetically-mediated targets are not modified by other factors, which, again, may be the critical oversight in GWAS methodologies. In many cases, the biology behind a disease relevant to particular PMED is known to be rather complicated. Take as an example cancer, where common practice often relies on treatments that target specific somatically-acquired tumor-initiating mutations. However, many of these individual mutation-targeting therapies fail due to preexisting mechanisms not accounted for or THAT ARE acquired and create resistance that are unaccounted for at the time of treatment initiation[12–14]. The occurrence of these mutations enables continued tumor proliferation. Although this example is specific to the etiology of cancer, it points out how mutations not accounted for, i.e., context specific mutations or mutations in the genetic background, can be important for the successful treatment of cancer. Thus, cancer mutations that are the targets of specific drugs often operate in a very context-specific manner, given that the existence of other factors could mitigate their success in combatting the tumor.

The context-dependency of targeted therapies and the existence of biological and genetic bypass mechanisms confounding potential therapies are not unique to cancer. For example, it has long been established that anti-microbial treatments targeting specific pathogens often unknowingly assume a particular genetic architecture and genetically-mediated capabilities of the pathogen for most infectious diseases. As a result, many pathogens have eluded successful treatment, in a similar fashion to certain tumors, through an ability to rapidly evolve mechanisms that withstand effective treatments. In addition, many complex neurological diseases, such as Alzheimer's Disease (AD), are known to be influenced by many factors whose complex interac-

tions make it difficult to identify successful treatments that consider only one of those factors[15].

The biological complexity surrounding cancer, infectious disease and common chronic congenital diseases like AD makes it hard enough to elucidate genetic mechanisms contributing to them that might lead to treatments, but the very process of identifying those factors adds even more complexity. This observation motivated our study of the choice of a MRSA reference genome, as a goal of that study was to uncover technical sources of variation that complicate clinically-meaningful interpretations of MRSA genome associations. For example, WGS and genotyping protocols for identifying genetic variants that might be implicated in cancer, infectious disease and common chronic diseases use constructs, like reference genomes, that may not be suitable for the identification of all the relevant genetic factors.

These observations suggest that gene and DNA variations responsible for human diseases do not work in isolation, but rather have effects that are 'shaped' or influenced by the activities of other genes and variants. In addition, current strategies for assembling and characterizing genomes in anticipation of mining them for important, clinically meaningful genetic variants are also impacted by the assumptions they make about the variation they are interrogating. In this light, there is overwhelming evidence in the scientific literature that the 'background,' primarily genetic background, in which a specific variant can be identified and in which it operates biologically influences both the identification and the ultimate effect of the variant, ultimately suggesting that the genetic background associated with a specific variant creates a true 'context specific' manner in which genetic variants can be characterized and have an influence on a human disease. See figure 1.1 for a diagram of the complexity associated with disease gene identification.

1.2 Terminology

The term "genetic background" has numerous definitions in the biomedical literature. However, it historically denotes the phenomenon whereby the phenotypic penetrance or expression of one genetic factor, such as a specific SNV or structural variant, can influence the expression or function, or lack thereof, of a second genetic

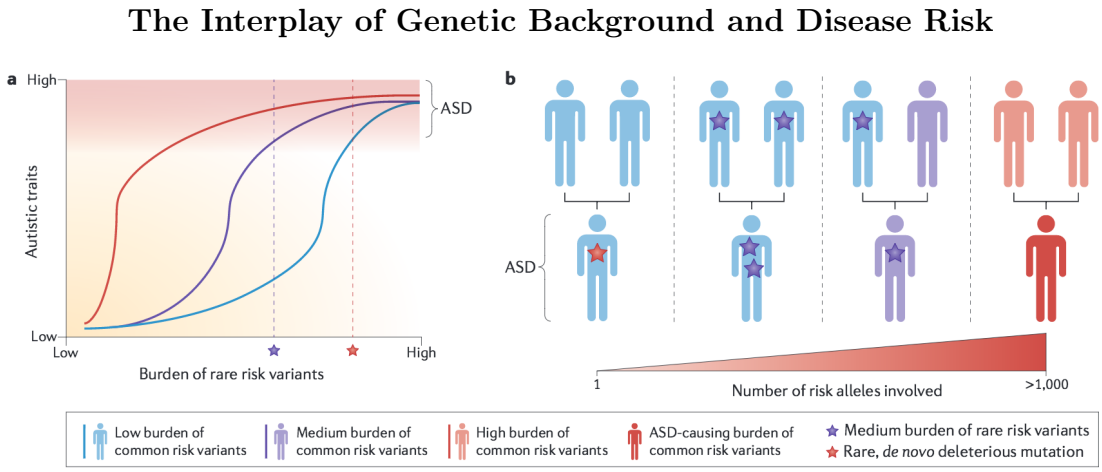

Figure 1.1: Figure from Bourgeron et. al (2015): A) Panel depicting various disease burdens. B) Disease transmission may occur through multiple paths: a *de novo* highly penetrant mutation (far left), a medium burden of rare variants passed on from each parent (left of middle), if one parent has a medium load of common risk variants and one has medium burden of rare risk variants (right of middle), and lastly when children develop disease where both parents have a high dose of common risk variants (far right).

factor. This type of interaction can often affect a phenotype of interest in a non-additive manner[16]. In this context, the regulated factor in question that is influenced by the presence or ability of another factor, is often said to be "modified" by the other factor. Most human phenotypes are known to be under control of extensive collections of genes, where variations in many of these genes contribute to the total phenotype variation and hence modify each other's effects. Ignoring modifiers is known to confound research studies[16–18].

The term 'epistasis' originally described phenomena in which the penetrance of one gene could be suppressed by a modifying gene. However, its current use in biomedical literature refers to other sorts of modifications impacting phenomena beyond gene expression. In addition, other forms of interactions are known to occur, including gene-environment interactions that could themselves complicate characterizations of variant interactions occurring both within the gene, i.e., SNV-SNV interactions, or between variants in different genes[19,20]. In all, terms such as genetic background, modifying factors, epistasis, and additional terms such as synergism, interaction deviation, intragenic complementation and others[21], all convey the fact

that genetic variants often act in a context-dependent or context-specific manner and they will be treated as more or less synonymous and used throughout the remainder of this thesis.

1.3 Examples of Context-Specificity from Biological Studies

Publications in the genetic and biomedical literature going back centuries have noted the non-trivial difficulty in identifying context dependent effects, including the rediscovery of Mendel's laws of genetics[22]. These difficulties mainly derive from the large sample sizes necessary to detect the often-subtle context-specific effects that contribute to many phenotypes. Although there in notable exceptions, mostly in the model organism literature, this fact and others have resulted relatively few specific investigations characterizing context-dependent effects involving very specific genetic variants or factors in the human biomedical literature. Many studies, however, have considered the gross or overall effects of disease modifying factors, as will be discussed later. In addition, there is growing interest in studies investigating human context specific effects that leverage multiple high-throughput genome sequencing technologies and other 'omics' technologies (transcriptomics, proteomics, metabolomics, etc.). In fact, these studies could allow researchers to potentially identify all variants and other genomic factors that interact with a primary genetic factor of interest[23,24] and further could be motivated by and extend historical attempts to characterize instances where genetic context impacts a biological outcome, as described below.

1.3.1 Basic Eukaryotic Model Organism Studies

Multiple historical investigations have exploited inbred mouse strains to assess the transfer of genes and genetic variants from one strain to another[25,26]. These studies demonstrated the importance of genetic context and motivated researchers to consider their implications for human disease. For example, many mouse studies have shown that the transfer of a gene with a lethal embryonic mutation in one mouse strain could actually lead to viable mice when implanted in a mouse strain

with a different genetic background[23,27–29]More recently, it has been estimated that approximately 74% of all variants that modify the effect of a specific mutation, the optomotor-blind gene, in Drosophila are dependent on the genetic background of strains studied.[4,5]. The authors of the study investigating the optomotoer-blind gene ultimately noted that the genetic background effects on the expression of the mutant were likely underestimated in the study, and that the wider consequences of such genetic influences are poorly understood[5,6]. Finally, Kruglyak and colleagues showed that many phenotypically-impactful variants in yeast had effects that were modified by the genetic backgrounds of different yeast strains[3].

1.3.2 Cancer and Treatment Response

As noted in the introduction, it is well-documented that cancer is initiated and sustained by the coordinated activities of multiple oncogenes, often harboring inherited and somatically-acquired mutations, and the activities associated with specific gene and protein networks[30]. For example, Vogelstein et al. recently reviewed the current state of knowledge about tumorigenesis in various cancer types and concluded that each tumor is likely a product of multiple primary driver mutations in addition to the activities and impact of many subtler 'backseat' driver mutations. These secondary mutations often work independently and in tandem, such that if one mutated gene is therapeutically targeted, other mutated genes sustain tumor growth and may initiate metastasis by circumventing any targeted treatments focusing on a single mutated gene. In addition, it has been shown recently that tumors acquire copies of growth sustaining genes, such as housekeeping genes, and lose copies of tumor suppressor genes during the evolution of the tumor. Such gains and losses provide a unique 'permissive' background for tumor growth, allowing the primary driver mutations to thrive and work in an uninhibited environment[31]. In addition, it is becoming increasingly clear that tumorigenesis is not only influenced by *de novo* acquired somatic mutations, but also by variants present in the host genome, as some inherited germline variants may themselves create a more permissive environment for tumorigenesis while increasing cancer susceptibility[32].

1.3.3 Human Ancestry and Disease

It has been shown unequivocally that disease prevalence rates vary by country, community affiliation, and geoethnic origins. For example, genetic differences between ethnic groups are known to modify the association of specific disease-causing variants with disease risk and transmission[33]. This is often attributed to dietary and cultural differences between the population groups, but also to genetic differences between populations[34]. As a result of genetic background differences between individuals that modify specific variant effects, individuals that are admixed between different populations (e.g., African Americans, Hispanics, Brazilians, etc.) often exhibit phenotypes and disease rates that are intermediate between the two relevant parental populations, suggesting that genetic background does indeed influence human phenotypic expression and can modify the effects of individual variants associated with a particular disease of phenotype in the population at large.[35,36]

1.3.4 The Polygenic Model of Human Diseases

Recent studies have sought to determine the degree to which the combined or collective effects of genetic variants, each with a minor or non-substantive phenotypic effect, contribute to a disease state[37–39,39–41]. Such studies suggest that most complex human diseases do indeed have a large polygenic, or genetic background effect, which contributes to their manifestation. Although this may make it difficult to identify each and every variant contributing to a particular disease, given the small effect each individual variant may have on the disease, it does suggest that the cumulative effects of many variants shape phenotypic expression. This further suggests that the genetic 'context' of an individual shapes their phenotypic presentation, i.e. the state of gene expression, protein levels, metabolic profile, physiologic function, and overt clinical profile *in vivo*[7]. Also, there are many examples in the literature where a small number of genes or genetic variants have been shown to influence a phenotype, also suggesting that the genetic background context within which a gene or variant operates is important to consider in the assessment of the contribution of any one gene or variant[3].

1.3.5 Infectious Disease Caused by Bacterial Pathogens

Many antibiotic agents target individual genetic regions of essential genes, which are responsible for bacterial fitness[42]. Resistance to these therapies can arise if mutations, as small as individual nucleotides in conserved genes throughout the genome, arise that counteract the mechanism targeted by the therapy. It is now well accepted that infectious diseases caused by bacterial pathogens can reach epidemic proportions due to evolutionary mechanisms involving the accumulation of mutations that allow them to survive and overcome most antibiotics and interventions[43,44]. This suggests that a host's defense mechanisms cannot always deal with a pathogen and creates the potential for host-pathogen interactions influencing the evolutionary mechanisms that contribute to sustained human infectious disease. In this light, it is well known that host genetic background and a particular pathogen may be better 'matched' and lead to a symbiotic or mutually beneficial relationship[45]. A number of studies on this particular topic have investigated the ability of host and pathogen genetic background to contribute to disease pathogenesis as well as shape features of bacteria[46–49].

1.3.6 General Gene x Environment Interactions

There is a great deal of literature on the modifying influence of gross environmental factors on the impact of a gene or genetic variant on disease susceptibility or general phenotypic expression in humans[50]. For example, lactose intolerance is known to be influenced by genetic variants, but only really manifests in societies with sufficient access to milk[51–53]. Other well-known examples that need further study and validation include obesity and diabetes, many forms of cancer and addiction, all of which are likely influenced by gene-environment interactions[54–56].

1.3.7 General Epistasis the Non-Additive Effects of Genetic Variants

As noted, epistasis involves the interaction between multiple biological intermediates such as proteins, genes, SNVs, and other factors, and is known to contribute

to both disease susceptibility and organismal function. Epistatic effects can be revealed through either direct biological studies or statistical analyses by identifying non-additive effects of combinations of factors. However, methods in to identify biological or statistical epistasis have particular challenges[57–60]. Both *In vivo* and *in vitro* systems and strategies exist to characterize epistasis, although identifying specific genes and/or mutations for potential analyses is not trivial for many reasons, not the least of which has to do with cost of testing thousands or millions of potential interacting factors[61]. Some statistical methods for detecting epistasis that do not just consider individual variant associations seek to statistically *prioritize* collections of SNVs for study in laboratory assays. As such, potential epistatic events identified through statistical analyses could reveal non-additive or multiplicative effects that may explain missing heritability of a disease or reveal novel drug targets that can be further assessed in focused laboratory investigations[62]. A caveat with statistical association-based epistatic screens is that they often suffer from difficulties associated with validation and replication in different data sets, whereas lab based epistatic screens often do not. Despite this, an advantage of sophisticated 'big-data' analytic methodologies for assessing interactions is that they can be used to screen thousands of genetic and phenotypic variables at fractions of the cost and time compared to functional assays and prioritize findings for further study.

Interestingly, in the context of Alzheimer's disease (AD), evidence suggests that APOE variants exhibit epistatic, or context dependent effects. For example, the $\epsilon 4$ allele is known to impact disease in a dosage dependent manner where two copies more than doubles the disease odds ratio. However, having either an $\epsilon 2$ or $\epsilon 3$ allele decreases the odds of having AD even if an $\epsilon 4$ allele is present; and possessing two $\epsilon 2$ alleles can protect against AD development[63]. These complexities suggest that the ApoE is involved in a number of processes controlling AD pathogenesis that may involve interactions of the sort that could be teased out via statistical methods. In addition, as mentioned, ethnicity and ancestral genetic background, which is a measurable genetic phenomenon, can impact differential disease associations. For example, African-American and Hispanics show weak disease association with the $\epsilon 4$ allele, although individuals from a Caucasian and Japanese ethnic background exhibit more pronounced allele dosage dependent effects of $\epsilon 4$ allele[64]. These phenomena are

likely attributable to epistatic or context-specific interactions amenable to statistical analyses, but not likely to be identified through conventional additive disease modeling[65]. One hypothesis we explore is that APOE, specifically the $\epsilon 4$ allele, interacts with groups or collections of other SNVs in a manner that would be missed by simple pairwise interaction-based association testing, or an assessment of the additive main or marginal effects or individual genetic variants.

1.4 The Genomics Era and Context-Specificity

1.4.1 Sequencing and High-Throughput Technologies

The advent of affordable genome-wide DNA sequence interrogation tools, such as genotyping microarray chips and high-throughput sequencing technologies, has enabled a data-driven biological era. Through these technologies, researchers can probe genetic aberrations throughout the genome for their association with diseases using genome-wide association study (GWAS) strategies[66]. Given the massive amounts of data that modern genetic and genomic technologies generate, particularly in the context of GWAS strategies, the challenges that impede their use to elucidate the biological complexity of diseases involve computational efficiency, management and storage of data, and mathematical and statistical analysis of data. This is in distinction to a great deal of biomedical research in the past where the solutions needed for many problems were biological or chemical in nature having to do with the creation of appropriate laboratory assays.

In this light, contemporary genomic studies currently need analytical methods that can: 1) robustly associate the vast number of DNA sequence variations with relevant clinical disease phenotypes via GWAS strategies; 2) Accommodate complexities, such as interactions between genetic variants, environmental factors, and general context dependencies in association studies; and 3) correct for the incredibly large number of statistical calculations needed to test each variant, or combination of variants and other factors, for association with disease. This chapter will address these three issues and describe methods to improve the GWAS analysis and the interpretation of information resulting from GWAS.

1.4.2 GWAS and Statistical Analyses

Since the first draft of the human genome was published in 2003, subsequent genomic studies have sought out to identify genomically-mediated pathophysiologic factors that positively or negatively modulate disease expression, either under the assumption that these factors involve multiple loci (i.e., complex 'diseases'), or one or a few individual loci (i.e., Mendelian 'diseases').[67] These efforts led to the initiation of the GWAS era of genetic studies, in which naturally occurring DNA sequence variants throughout the genome are interrogated or 'genotyped' on a large number of individuals either with or without a phenotype of interest or measured on a specific quantitative phenotype such as weight or cholesterol level, and tested for association with the phenotype using traditional statistical methods, such as linear regression methods for quantitative traits or logistic regression methods for qualitative traits[68,69]. Identifying genetic variants influencing diseases is particularly important because these variants can be used to predict disease risk and aid the development of novel therapeutics by revealing drug targets. To date, the NHGRI GWAS catalogue, which records the results of GWAS and is maintained by the National Human Genome Research Institute, lists $\geq$ 29,000 SNV-disease or SNV-trait associations that pass simple criteria for statistical significance (e.g., P(x) $\leq$ 5.0x10ˆ-8) for hundreds of Mendelian and complex diseases. However, as noted, the majority of these associations fail to explain large portions of the heritable components of more complex, polygenic disease[70].

Given the abundance of genomic data generated from genomic sequencing and high-throughput genotyping technologies, GWAS analyses have become rooted in the "Small N, Large P" statistical and mathematical problems, which have increasingly become commonplace in many research challenges. Thus, genomic analyses typically involving relatively small samples size (N) that have been leveraged to collect a large number of variables (P), such as phenotypes or genetic information.[71] The problem with these types of analyses is often low statistical power to draw compelling inferences, given the small sample size and the number of statistical tests which require a multiple hypothesis test correction. In the context of the AD studies that we have investigated and will discuss below, N=14,000+ Alzheimer's disease patients and controls have been evaluated for P=3,000,000+ DNA variants. It is expected

in such studies that the majority of variants will exhibit little or no association with disease[72]; however, a small subset of these variants will demonstrate an association with disease. In this light, the problem of most pressing concern is trying to distinguish the appropriate true association 'signals' from the background 'noise' induced from all the other non-associated variants. In addition, for most diseases, like AD, the belief is that they have many genetic determinants with weak effects that are likely to be context specific, resulting in genetic studies that ultimately explain a far lower proportion of the heritability of the disease than originally anticipated since the identification of variants with weak effects in genome-wide studies requires very large sample sizes. Ultimately, then, deciphering the root genetic cause(s) of the complex and multigenic nature of most chronic conditions, like AD, can be considered largely a data analysis and large-scale study design endeavor.[73,74]

1.4.3 GWAS and Context-Specific Genetic Effects

As noted, a major problem inherent to GWAS methodologies, and genetic association studies in general, is that the majority of variants in the human genome that could be interrogated are simply not associated with the disease. In addition, if an association can be detected, it is typically not very strong for a variety of reasons, including not accounting for interactions involving the relevant gene or variant[7,75,76]. Despite implicating thousands of disease variants, the traditional or canonical single-variant-focused GWAS framework used to date will likely fail to explain the relationship between genetic variants and disease since it ignores complexities, like genetic interactions. One can argue that to overcome this issue, one could increase statistical power to detect the effects of individual loci through the analysis of larger samples. However, this may not work when certain types of interaction are at play.[16] This is not to devalue the utility of current association analysis efforts or the entire paradigm of GWAS, however, but to suggest that improving the power of variant detection via association analysis could involve statistical analysis models that consider more than just the marginal effects of genetic variants and robustly account for and explore interactions.

The challenges associated with identifying pairwise or interacting variants are

apparent when considering more complex phenotypes, e.g., diseases associated with the genomically complex antimicrobial resistance or neuropsychiatric disorders, that are known to be among the most heterogeneous in terms of phenotypic presentation and causal etiology. Complex pathologies and phenotypes are thought to arise from perturbations in multifaceted gene networks, the interplay between multiple genetic-factors which influence biochemical and biophysical processes, or environmental factors that exacerbate genetic effects.

As discussed in the previous sections, the analysis of epistatic interactions and general context-specificity of disease-mediating factors is typically not pursued because of the additional statistical and computational burden involved, i.e. "Small N, Large P". For example, without knowing which variants may be interacting, one may test all possible combinations of variants for a potential interaction effect. With an initial dataset of millions of genetic variant, this would create an astronomical number of statistical tests. Thus, although *it is recognized that gene-gene, gene-environment, SNV-SNV, and other interactions exist and influence disease*, and may potentially produce genetic effects on par with the main effects of individual genetic variant, they are hard to tease out with standard statistical analysis methodologies. Of the methods proposed to assess gene-gene or SNV-SNV interaction effects, for example, the majority of them exhaustively test all gene or SNVs for association in a pairwise fashion, creating an enormous statistical problem given the need to correct for multiple hypotheses.[77–85] Alternative statistical methods could elucidate combinatorial effects missed in traditional single locus or SNV association studies, however. For example, in the context of our analysis of AD, we propose a two-step, nested logistic regression approach to testing the hypothesis that a particular variant known to be associated with a disease influences or interacts with the collective or combined effects of many other variants. We apply this approach to the study of Alzheimer's disease (AD). In particular, we are interested in characterizing the ability of the $\epsilon 4$ allele in the Apolipoprotein E (ApoE) gene to interact with and/or modulate groups or collections of other SNVs, possibly throughout the genome. This approach obviates the need for directly testing all possible individual SNV-SNV interactions that could number in the tens of millions, if not more. We apply this two-step approach to a set of SNVS thought to be associated with AD obtained from the NHGRI GWAS catalogue.

Ultimately we find some evidence that the e4 allele modulates the influence of a group of other susceptibility SNVs on AD, but only in the context of the cohorts that have been studied, with the ADGC data. We also consider analyses that explore the identification of variants impacting infectious disease and show that the current techniques for identifying variants in a pathogen and then relating them to a clinical outcome in the host is very context-specific, and ultimately requires care and an attention to detail that precedes testing specific interactions involving those variants. We also show that despite this issue, DNA sequencing of pathogens does often result in more confident clinical classification of pathogens than standard culture-based methods.

1.5 Overview and Organization of Dissertation

The examples and issues discussed in the previous sections demonstrate both the interest in, and inherent difficulties associated with, characterizing the context-specific effects of genetic factors influencing disease susceptibility and general phenotypic expression. Genetic background and general context-specificity may be subtle and impact even the bioinformatic and assay workflow used to identify and characterize genetic factors and their relationship to a phenotype. This suggests that a variety of bioinformatics and statistical genetic analysis methods are needed. Ultimately, we hypothesize that failure to control for genetic background and general context-specificity, e.g., multiple genetic loci interacting to influence disease, may lead to: 1. confounding effects and mischaracterization of the contribution of genetic factors to disease; 2. variability across studies when associating disease outcomes with genetic factors, especially in clinical contexts; and 3. a need for sensitivity to the design and implementation of studies, limitations of data analysis methods, and clinical genetic assay interpretation. As an example of this last point, consider researchers developing genetic assays for mutation detection in clinical diagnostic settings that focus on the identification of single mutations where the result could mislead prognoses if the mutation of interest has an effect that is modified by another gene or genetic background as a whole. We have embarked on three different studies designed to characterize the context-specificity and the effect of genetic background on human

diseases. First, in chapter 2, we explore the utility of DNA sequencing relative to traditional culture-based methods in the context of identifying bacterial pathogens in the clinical setting. Second, in chapter 3, we explore the context-specificity of genetic variant identification protocols involving the MRSA pathogen and show how they can impact clinical interpretations about the severity and outcomes of the infection in humans. Third, in chapter 4, we consider the influence of a single genetic factor, the much-studied APOE4 locus on the impact of a number of other AD susceptibility variants and show that these variants have differential effects depending on whether an individual does or does not have the APOE4 variant and in what set of individuals the study is pursued in. Fourth, in chapter 5, we consider the identification genetic variants that influence AD by considering whether they influence gene expression levels that are ultimately associated with AD via mediator-wide association study (mWAS) methodology. We conclude in Chapter 6 with a general discussion of the results, limitations of the studies and areas for future research.

Chapter 2

Clinical Bacterial Species Determination by Whole-Genome Sequencing: A Proof of Concept Study

2.1 Abstract

Whole-genome sequencing (WGS) has drastically improved bacterial pathogen identification, which is crucial for diagnosing and managing infectious disease. WGS technologies present potential advantages over current complex, labor-intensive, and often slow clinical laboratory identification practices. However, few studies have addressed the clinical validity of WGS to accurately determine bacterial species. We compared the performance of WGS and traditional laboratory-based methods to identify species from 354 bacterial cultures collected from routine clinical microbiology laboratory practices in the Scripps Hospital system in Southern California. To determine pathogen identity, contiguous DNA sequences from cultures were assembled and aligned against the NCBI genome database, where species identification was defined as ≥97% assembly similarity with a database species across ≥35% of that speciesâĂŹ genome. Laboratory identification methods were defined by Clinical Laboratory Standards Institute (CLSI) protocols established at the Scripps Microbiolog-

ical Laboratory. Our results showed the WGS approach adequately determined 35 distinct species for 339 samples (95.8 of the total), which closely matched the 340 samples with species determinations by the clinical laboratory methods. Discordance between routine clinical and WGS protocols occurred in 29 of the 339 samples (8.6), in addition to 17 samples where WGS detected additional bacterial species that met the WGS match criteria. We conclude that WGS identification methods utilized in conjunction with the NCBI database provided more resolute species determinations, at a lower cost, and shorter turnaround time. Nonetheless, a more comprehensive genome database and improved species match thresholds may be necessary for WGS adoption in routine clinical pathogen identification strategies.

2.2 Introduction

Bacterial infections cause roughly 16% of annual deaths, which significantly impacts global rates of morbidity and mortality. Management of antimicrobial interventions and therapies for infectious disease critically rely on efficient pathogen identification. However, current clinical pathogen identification practices are often labor-intensive, costly, and slow[86]. Emerging whole-genome sequencing (WGS) technologies routinely characterize research grade bacterial genomes and could potentially improve clinical microbial pathogen identification strategies. However, clinical implementation of WGS identification is not common-place, despite the first complete bacterial, *Haemophilus influzenzae*, having been sequenced in 1995[87]. In addition, the current National Center for Biotechnology Information (NCBI) genomic database contains complete genomes for over 5,000 bacterial species, and numerous bioinformatic tools exist to accuratley compare the genomes of clinically isolated bacteria which could be exploited in clinical settings[88].

Many large-scale epidemiological research projects have leveraged WGS in ways that could be of clinical relevance. For example, retrospecitve WGS studies have elucidated transmission dynamics for outbreaks associated with *Escherichia coli*, *Vibrio cholerae*, *Klebsiella pneumoniae*, and *Mycobacteria*[89]. Real-time high-throughput WGS methods have also uncovered primary outbreak sources of hospital infections caused by methicillin-resistant strains *Staphylococcus aureus* (MRSA)[90,91]

vancomycin-resistant *Enteroccocus facecium*, and carbapenem-resistant *Enterobacter cloacae*[92]. Despite these examples, additional validity and feasibility studies are still required are motivate the routine implementation of WGS technologies for clinical pathogen identification. Thus, to investigate the performance of WGS-based bacterial pathogen identification, we compared a WGS-based identification strategy to a traditional Clinical Laboratory Standards Institute (CLSI) identification protocol. The two methods were applied to 354 clinical isolate cultures collected from a standard routine clinical microbiology workflow not screened for specific pathogenic features and thus reflect the incidental pathogenic variation arising in routine clinical care.

2.3 Methods

2.3.1 Sample Collection

We obtained 354 bacterial specimens from the Scripps Heath system-associated Sorrento Mesa Microbiological Laboratory in Southern California. Potentially pathogenic samples were collected from various body sites and cultured overnight. Resulting colonies were then inoculated on two duplicate plates: one for clinical laboratory testing and one for WGS sequencing analysis. The clinical laboratory bacterial species identification strategy adhered to the following CLSI guidelines: direct microscopic examination, gram-staining, elective media culture, and biochemical assays. The BD Phoenix^{TM} Automated Microbiology System further analyzed all colonies to confirm bacterial species determinations and antimicrobial susceptibility.

2.3.2 DNA Processing and Whole-Genome Sequencing

DNA was extracted from 354 bacterial isolate plate colonies using ThermoFisherâĂŹs ChargeSwitch technology, and prepared for sequencing using the Illumina Nextera XT library preparation kit. Libraries were rapidly sequenced on an Illumina HiSeq 2500 sequencer. Single-ended, 50-bp reads were generated in batches of up to 48 samples. Read coverage ranged from 15 -247 apart from three low-coverage outliers. Sequence reads were trimmed for quality using the trimmotmatic tool and were subjected to reference-free de novo read-based assembly (Velvet assembler V1.2.10,

k-mer seed = 25)[93]. All computational analyses were performed on the Triton Shared Computer Cluster at the San Diego Super Computer Center (8 or 16 cores per node).

2.3.3 WGS Species Determination Based on NCBI Database Matching

BLAST+ V2.2.29 was used to align assembled contiguous sequences (contigs) for each bacterial isolate against the NCBI genomic database, which at the time of analysis (November 10, 2014) contained over 28,000 whole bacterial genomes[93,94]. Quality matches for common species were based on the nucleotide (nt) collection of finished assemblies. However, uncommon species samples required querying the whole-genome shotgun (wgs) database of unfinished assemblies. Custom Perl scripts processed BLAST outputs, identified optimal contig matches, and derived match metrics, i.e. matched contigs frequency, DNA species match length (nucleotides), and species length above a specified identity threshold. Adequate species determination for individual samples was defined as $\geq$97% identity with a species in the database across $\geq$35% of the species' genome.

2.3.4 Partial Genomic Matches

Custom Perl scripts assessed samples exhibiting partial contig matches to multiple closely related database species and discrepant identifications between the WGS and the clinical laboratory methodologies. The average nucleotide identity and match length between the assembled genomes and the additional database matches were then calculated.

2.4 Results

2.4.1 Bactria Sample Distribution by Culture Sources and Species

178 of 354 (50%) presumed pathogen samples were from wounds; 79 samples (21%) from blood and sterile sites; and 43 samples (12%) and 42 samples (12%) from

sputum and urine, respectively (Table 1). This represented a typical communal distribution of common bacterial pathogens. Using the previously described thresholds, WGS adequately determined bacterial species in 339 of 354 samples (95.8%), which was comparable to the 340 clinical laboratory species determinations. WGS methods revealed 15 inadequate samples (Table 2.1). Figure 2.1 (complete list in Table S5 of the Supplementary Appendix) describes the frequencies of the 35 distinct bacterial species adequately identified by the WGS species determination method. The most common pathogens were *Staphylococcus aureus* (35%), *Escherichia coli* (15%), *Enterococcus spp.* (13%), *Coagulase-negative Staphylococcus spp.* (8%), and *Pseudomonas aeruginosa* (6%) (complete list in Table S5). Didelot, et al., reported a similar common bacterial pathogen species distribution[86].

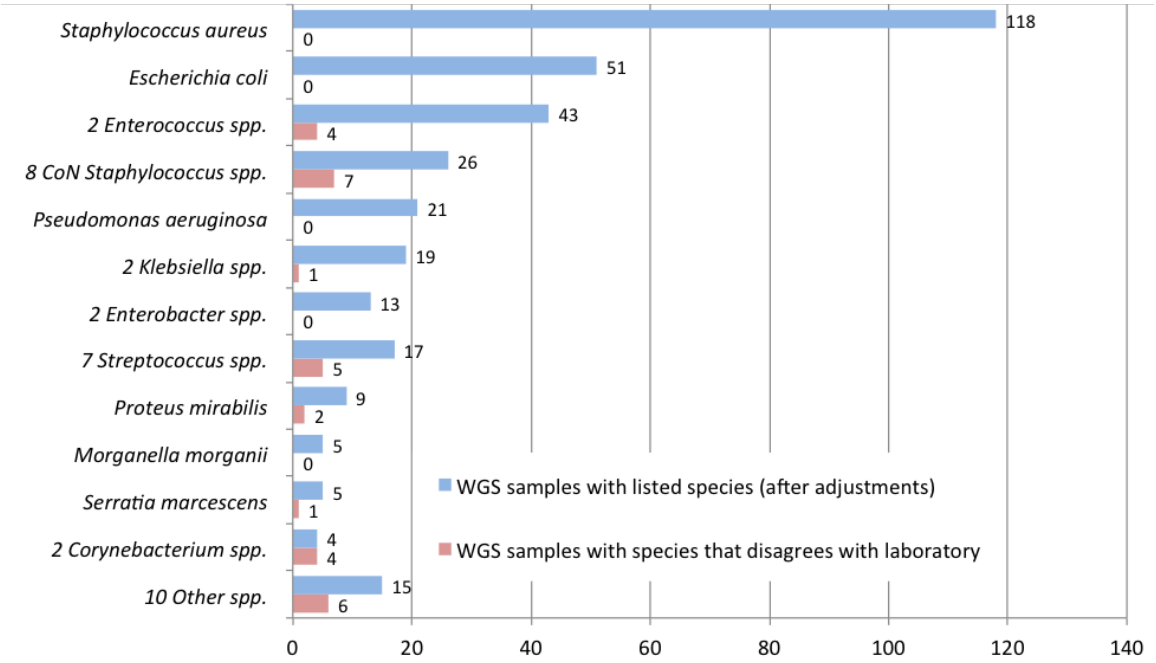

Figure 2.1: Species Distribution for 346 samples with a single species adequately determined by WGS. Histogram depicts the occurrence frequency of WGS based sample identification where a subset of samples disagree with laboratory identification methods.

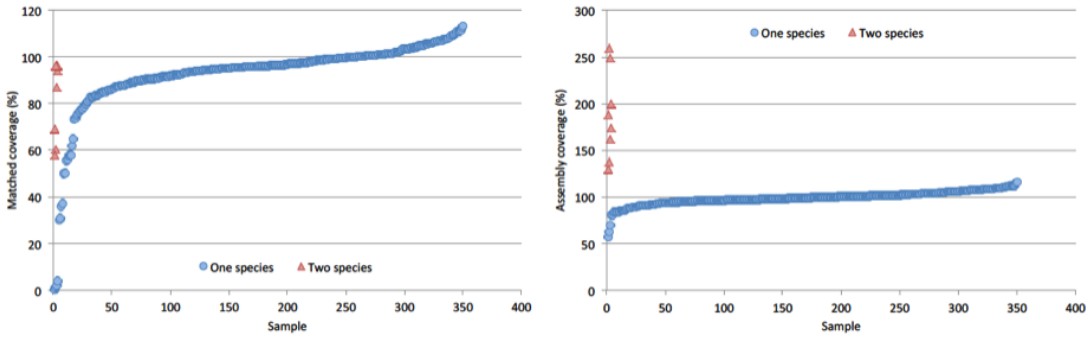

Figure 2.2: Left Panel: Depicts the distribution of matched coverage, i.e., the matched blast length divided by length of genome of determined species. Right Panel: Distribution of assembly coverage, i.e., assembly length divided by length of genome of determined species.

Table 2.1: Distribution of bacterial samples by culture source. Additional information includes the result species determination by WGS and the clinical laboratory.

Culture source	Samples (%)		Determination result	WGS samples		Clinical laboratory samples	
Wound	178	(50%)		*Initial*	*Adjusted*	*Initial*	*After repeats*
Blood	67	(19%)	Adequate determination of single species	317	346	337	340
Sputum	43	(12%)	Adequate determination of multiple species	33	4	0	0
Urine	42	(12%)	Inadequate determination	4	4	17	14
Sterile sites	12	(3%)					
Unspecified	12	(3%)					
Total	354			354	354	354	354

2.4.2 Clinical Laboratory Species Determination

337 of 354 bacterial samples initially had complete clinical laboratory species determinations. 17 samples exhibited incomplete results, 13 determinations with only genus specifications and 4 characterized either "mixed flora" or "skin contamination", which are common when microscopic or biochemical examinations suggest a non-pathogenic species. WGS provided adequate species determinations for 13 of the 17 unidentifiable samples. Beyond incomplete species determinations, 21 other samples returned discordant species determinations between the laboratory and WGS methods. This raised human intervention error concerns for the laboratory method, i.e. multiple handlings or poor sensitivity. To assess the reproducibility of the laboratory determinations and also which method, laboratory or WGS, might be problematic, the

laboratory determination process was repeated on the 23 discrepant samples. Species determination results were updated in 12 of these 23 samples (Table S7). After repeat testing, 6 originally discordant samples became concordant between the two methods tested, while the remaining 6 samples remained discordant (2 with incomplete species information).

The clinical laboratory initially determined 37 bacterial species in 337 samples; the remaining 17 samples had incomplete species information, with 13 determined only to the genus and four characterized as "mixed flora" or "skin contamination". The latter characterizations are common when routine microscopic examination and biochemical assays suggest the species are not pathogenic. WGS adequately determined the species in 16 of the preceding 17 samples. In addition to the samples with incomplete species determination, 24 others had an initial laboratory species that did not fully agree with the WGS species. This suggested possible handling errors and poor sensitivity for the laboratory species determination. To assess its reproducibility, 23 samples (20 with discordant determinations) were returned to the laboratory for repeat testing, and the species information changed in 12 of these samples (Table S7 of the Supplementary Appendix). Whereas none of these 12 samples had agreement between the WGS and laboratory species initially, six agreed after repeat sequencing. The number of samples with incomplete species determination decreased from 17 to 14.

2.4.3 Whole Genome Sequencing Species Determination

Contigs from 339 samples with adequate WGS species determination matched almost exclusively to a single species (Table 2.4). However, 29 samples had sizable matches (longer than 300kb) to multiple species (Table 2.6). For 25 of 29 samples, investigations revealed the total lengths of additional matched assemblies were comparable to, or less than, the lengths of the original matched genome. This implied that 25 samples contained a single species; however the remaining 4 samples inadvertently contained two species, as denoted by the additional database matches. The resulting adjustment to single and multi-species matched coverage samples is reflected in the "Adjusted" column of Table 2.1.

To support our adjustments, we noted that all 29 discordant samples exhibited nearly identical DNA content based on additional BLAST optimal strain matches (Table S4). Indeed, matched species results for 22 of these samples had Latin names, while other NCBI matches returned only numbered ID identifiers. Numbered IDs indicate these NCBI genomes may not characterize fully complete species as that designation typically corresponds to named species. Also, 2 of the 24 samples matched to two clearly distinct NCBI database species, which potentially indicates contaminated entries. To determine the species of each adjusted sample, either the longest named species match or the correct species in the case of a database error was chosen. One single-species sample matched to two species with numbered IDs that were not close enough for adjustment which suggests inadequate species determinations.

For the 354 samples that underwent WGS analysis, Figure depicts the matched coverage sample distribution, i.e., matched contig length $\geq$97% identity for the best species divided by a typical genome length for that species. For the 24 samples with adjustments, we added the matched lengths of the species that were nearly identical at the DNA sequence level to calculate the matched coverage. The matched coverage threshold was set at 35%, due to 339 samples (95.8%) falling above 40%. Figure shows the corresponding assembly coverage sample distribution, i.e., the assembly length divided by the best matched species typical genome length, which is greater than 55% in all cases. The triangles in the figure correspond to the four samples with multiple WGS species results, which had assembly coverage values between 129% and 249%.

15 samples failed the NCBI database match criteria and were deemed inadequate for WGS species determination. 13 of these samples had long assemblies; however, they exhibited drastically low read coverage which is calculated by total read length multiplied by total sequencing reads, divided by genome length (Table 2.9). When the identity threshold was relaxed from 97 to 95%, seven of the nine sample coverage values were increased substantially. We hypothesize our WGS procedure likely produced correct species identifications for these samples, apart from three with only numbered IDs. Four samples had leq2% matched coverage at geq97% identity, thus have no proximal NCBI database genome matches.

2.4.4 Wgs and Clinical Laboratory Discrepant Species Determinations

Of the 339 samples with adequate WGS species determination, only 29 were discrepant with the laboratory determination, even after repeat clinical testing. Upon separating discordant samples into four categories (Table 2.2)), Category 1 contained 11 samples where the clinical laboratory species determination failed. Eight of these samples were determined only to the genus, and three samples were deemed identified incomplete by laboratory testing. This highlights two potential limitations of current laboratory procedures: 1) subjectivity by laboratory technicians; and 2) poor biochemical assays sensitivity, although this limitation may be specific to less commonly annotated bacterial species and not inherent to the laboratory method. In Category 2, 12 samples had complete species determination by the clinical laboratory and agreement with WGS to genus, but not species. For 10 of these samples, the NCBI database contained whole genomes of the laboratory species designations, however WGS species method matched more robustly than the laboratory-determined species. This suggests the laboratory determinations were less accurate. The two remaining samples in Category 2 were not in the NCBI database and thus whether a concordant species determination exists remains unclear.

In Category 3, 2 samples had discrepant genus species determination. Category 4 had four samples that contained two WGS species, but only one laboratory species result (WGS not fully determining both species in one of these samples). Nonetheless, these events suggest WGS methods may potentially improve multiple pathogen identification events compared to the more stringent laboratory practices.

Category 3 contains two samples that did not agree even to the genus. Category 4 contains four samples that had two species according to WGS, but only one according to the laboratory, which indicates that WGS is better able to identify multiple pathogens in a single sample. Even though our WGS procedure was unable to determine the species of four samples, the laboratory species for three of those were inconsistent with the WGS data (Table S6 of the Supplementary Appendix). Since the laboratory species are in the NCBI database, they would have been found by our procedure if they were correct.

Table 2.2: Clinical isolates with adequate WGS species determination status which partially of fully failed to agree with the clinical laboratory determinations (N=34).

Sample ID	WGS determination	Laboratory determination	Culture source
Category 1: Incomplete species determination by laboratory (n=13)			
2014-060	*Achromobacter xylosoxidans*	*Achromobacter species*	bronchial washing
2014-066	*Staphylococcus epidermidis*	*Staphylococcus, coagulase negative (changed after repeat)*	blood
2014-092	*Bacillus licheniformis*	*Bacillus species, not anthracis (changed after repeat)*	blood
2014-147	*Enterococcus faecalis*	*Mixed gram positive flora*	urine
2014-149	*Enterococcus faecalis*	*Mixed flora*	urine
2014-176	*Streptococcus parasanguinis*	*Prevotella species*	blood
2014-205	*Corynebacterium striatum*	*Many Corynebacterium*	sternum
2014-208	*Staphylococcus simulans*	*Many mixed cutaneous flora*	left hallux
2014-255	*Corynebacterium striatum*	*Many Corynebacterium species, not JK*	right leg
2014-330	*Corynebacterium amycolatum*	*Mod Corynebacterium species, not JK*	right foot
2014-340	*Streptococcus anginosus*	*Alpha Streptococcus, not Enterococcus*	urine
2014-392	*Achromobacter xylosoxidans*	*Achromobacter species*	right hand
2014-393	*Staphylococcus epidermidis*	*Staphylococcus, coagulase negative*	urine
Category 2: Agreement in genus, but disagreement in species (n=15)			
2A: Whole genome of laboratory species *is* in the NCBI database (n=10)			
2014-047	*Staphylococcus caprae*	*Staphylococcus capitis (did not change after repeat)*	right femoral
2014-048	*Enterococcus faecium*	*Enterococcus faecalis (changed after repeat)*	urine
2014-062	*Klebsiella pneumoniae*	*Klebsiella oxytoca*	abdomen
2014-067	*Staphylococcus epidermidis*	*Staphylococcus hominis (changed after repeat)*	blood
2014-114	*Staphylococcus hominis*	*Staphylococcus epidermidis (did not change after repeat)*	blood
2014-148	*Enterococcus faecalis*	*Enterococcus faecium*	urine
2014-237	*Staphylococcus lugdunensis*	*Staphylococcus chromogenes (changed after repeat)*	scalp ulcer
2014-258	*Proteus mirabilis*	*Proteus vulgaris (did not change after repeat)*	peg tube
2014-331	*Serratia marcescens*	*Many Serratia plymuthica*	sputum
2014-363	*Proteus mirabilis*	*Proteus vulgaris (did not change after repeat)*	peg tube
2B: Whole genome of laboratory species *is not* in the NCBI database (n=5)			
2014-182	Streptococcus parasanguinis	Streptococcus viridans	blood
2014-272	*Aeromonas hydrophila (adjusted)*	*Aeromonas sobria*	lower back
2014-276	*Staphylococcus HGB0015*	*Mod Staphylococcus schleiferi*	right breast
2014-286	*Staphylococcus HGB0015*	*Mod Staphylococcus schleiferi*	right breast
2014-333	*Citrobacter freundii (adjusted)*	*Citrobacter braakii*	left breast
Category 3: Disagreement in genus (n=2)			
2014-170	*Corynebacterium amycolatum*	*Klebsiella pneumoniae (changed after repeat)*	right knee
2014-352	*Sphingomonas paucimobilis (adjusted)*	*Elizabethkingia meningoseptica*	urine
Category 4: Second species found by WGS, but not by laboratory (n=4)			
2014-081	*Enterococcus faecalis, Staphylococcus aureus*	*Enterococcus faecalis (did not change after repeat)*	blood
2014-112	*Pseudomonas aeruginosa, Proteus mirabilis*	*Pseudomonas aeruginosa*	Groshong catheter swab
2014-228	*Streptococcus intermedius, Staphylococcus aureus*	*Streptococcus intermedius*	abdomen
2014-268	*Proteus mirabilis, Streptococcus parasanguinis*	*Rare Streptococcus parasanguinis*	left wrist

The culture source distribution of the 29 discordant samples was similar to the distribution of the collection at large. However, a heterogeneous species distribution was observed among discordant samples (Figure 2.1). No disagreement was seen in *Staphylococcus aureus*, *Escherichia coli*, and *Pseudomonas aeruginosa*; however *coagulase-negative Staphylococcus spp.* and other clinically uncommon species caused a high-percentage disagreement events. These results indicate that WGS can produce higher resolution and accurate bacterial species determinations for at least the more common species.

2.4.5 WGS Cost Efficiency in Real-Time Clinical Applications

Cost and turnaround time is crucial for real-time clinical implementation. The costs associated with our WGS method were approximately $50 per bacterial genome,

which includes the costs of bacterial cell culture, DNA extraction, library preparation, and sequencing(Table 2.3). Sequencing technologies are rapidly advancing which has significantly decreased associated costs over the past decade (estimated WGS cost in 2012 was $150 per isolate) [7]. Didelot, et al., estimated a cost of $25 per Mb of assembled sequence, and bacterial genomes typically range from 2-6 Mb[86].

Approaches	Cost	Turnaround time	Major limitations
WGS	Reagent: $50 per sample	Culture for colony (overnight) Sample processing for WGS (12~18 h) Bioinformatics analysis (<1 h)	Whole genomes of some species not available in databases Inadequate curation of species with only numbered IDs in databases Occasional errors in databases Lack of validated bioinformatics thresholds for species determination
Clinical laboratory	Reagent: variable	Culture for colony (overnight) Standard procedures# (12 h) Confirmation on automated system& (12 h)	Difficulty in identifying uncommon species Difficulty in distinguishing closely related species Failure to identify multiple species Incomplete determination of species for samples thought to be nonpathogenic Poor reproducibility

#: including direct microscopic examination, gram stain, culture on elective media, and biochemical assays; &: BD PhoenixTM automated Microbiology system

Table 2.3: The cost, turnaround time, and major limitations of WGS and clinical laboratory determinations of bacterial species.

To replicate the typical community hospital microbiology lab workload, 48 samples were processed per WGS analysis batch. Genomic sequences were obtained approximately 18 hours after bacterial colonies were available for a 48-sample pool (3 hours for DNA extraction, 3 hours for library preparation, and 12 hours for sequencing). A similar WGS turnaround time was reported by Hasman et al.[95], which roughly compares to routine clinical laboratory turnaround time. Bioinformatic species determination analyses required <1 hour per sample and all samples can be computationally processed in parallel.

2.5 Discussion

Currently, clinical laboratories follow a variety CLSI guided methods for pathogen identification which rely on organism culture and phenotypic characterization, i.e. gram staining and biochemical properties. These processes are complex, time-consuming, and often species-specific with variable sensitivity and specificity. WGS, in conjunction with comprehensive reference genome databases and highly-accurate bioinformatic workflows, can ultimately resolve and infectious microorganism identities, at potentially greater resolutions. Our study demonstrates that WGS-based bacterial

species identification is at least as accurate as laboratory methods, which highlights the growing validity of WGS as a scalable clinical microbiology diagnostic tool with low-cost and short turnaround time.

Over a third of the 29 discordant species determinations between the two methods had incomplete laboratory results; i.e. they returned as mixed flora or *coagulase negative Staphylococcus* (CoNS). Incomplete laboratory determinations were more probable when pathogenicity was âĂIJruled outâĂİ during the initial CLSI examination. However, informing clinicians of these types of pathogen identification discrepancies is crucial, particularly because previously recognized non-pathogenic strains have emerged as pathogenic. CoNS bacteria commonly colonize skin and mucosa and had long been considered non-pathogenic infections. However, as intravascular device use increases, CoNS infections have become a major cause of nosocomial bacteremia[96]. A blood culture sample in our study was initially determined as Staphylococcus epidermidis by WGS and MRSA by clinical laboratory testing respectively. Upon repeat laboratory testing, the sample identity was determined as CoNS, thus becoming concordant with the WGS determination. Interestingly, the mecA gene that confers β-lactam resistance, i.e. methicillin, was found in the DNA sequences of this sample. MRSA and CoNS are are closely related, thus a mischaracterization event based on the phenotypic contribution of methicillin resistance is possible. Mischaracterizations events such as this example can introduce variability in disease management by causing heterogeneous clinical treatment and outcomes.

Certain bacterial species demonstrated more frequent clinical laboratory mischaracterization, including *Enterococcus faecalis* and *faecium*, CoNS, and *Klebsiella spp.* Exemplifying *Enterococci*, the most reliable laboratory testing schemes for differentiating *Enterococcus faecalis* and *faecium* from other *Enterococcus* species includes 8 procedures: acid production from sorbose, sucrose, ribose and l-arabinose, utilization of pyruvate, deamination of arginine, motility, and pigment production on tellurite[97]. Efficient rapid and cost-effective strategies to identify the *Enterococcus* genus relies on antibiogram analysis, which requires continuous surveillance and is region-specific[98]. The complexity of existing laboratory testing schemes, with poor sensitivity and specificity, complicate optimal pathogen characterization approaches. Moreover, the application of our WGS methodology led to the identification of additional pathogens

in one blood culture sample, where the antibiotics necessary for the two distinct infections often varies amongst physicians, e.g., WGS identified *Enterococcus faecalis* and *Staphylococcus aureus* while the clinical laboratory only identified *Enterococcus faecalis*. Antibiotics necessary to combat these two distinct infections often vary amongst physicians. Thus, misdetection of additional pathogens through standard laboratory methods could confound patient care and significantly increase the occurrence of adverse events, especially in patients presenting bacteremia. Antibiotics necessary to combat these two distinct infections often vary amongst physicians. Thus, misdetection of additional pathogens through standard laboratory methods could confound patient care and significantly increase the occurrence of adverse events, especially in patients presenting bacteremia.

An unintended, yet important, finding was poor reproducibility of the clinical laboratory. Twenty-three discordant samples were sent for repeat testing and species results differed for more than half. Six initially discordant samples agreed completely with WGS after repeat. This error rate was unexpectedly high and thus raised concerns regarding human handling errors in current gold-standard identification methods. Exemplifying one blood culture sample, WGS returned Enterococcus faecalis while clinical laboratory initially returned Streptococcus parasanguinis and then Enterococcus faecalis after repeat. Clearly, initial mischaracterizations could mislead clinicians and produce differential diagnostic decisions.

2.5.1 Method Limitations

Elements of our proof-of-concept study suggest two major limitations of WGS determination methods to routinely identify bactieral species: (1) genome databases lack comprehensive bacterial representation, which will likely improve as additional bacterial genomes are deposited into the NCBI database; and (2) appropriate computational sequence matching thresholds are uncertain. Our threshold of geq97% identity and geq35% coverage is based upon similar studies[99], which conducted a study that leveraged 28 bacterial genomes in six phylogenetically distinct groups and determined the 95% threshold sufficient to delineate species. A more extensive study[100] involving 536 pairwise genome comparisons from the NCBI database sug-

gested an average nucleotide identity (ANI) threshold of 95-96%, which is closer to our study and the more stringent ANI threshold (Table 2.9). We leveraged a highly stringent threshold to ensure confident species determinations and the highest degree of clinically applicability. Nonetheless, bacterial species identification can be ambiguous because genomic variation in each species is non-uniform. As additional WGS studies are completed, the uncovered distribution of genomic variation in bacterial species, within and between bacterial species, will reveal more reliable DNA based differentiators as opposed to continued use of phenotypic differentiators.

Two extensions could futher improve clinical WGS approaches. First, drug resistance profiling during species determination would improve pathogen surveiliance efforts. ARG-ANNOT is a recently developed drug-resistance gene database with potential[101], however comprehensively investigating its clinical effectiveness is necessary. Second, cultureless species determination of patient samples from WGS could greatly reduce human errors compared to laboratory methods, and has been demonstrated previously[95]. However, the necessary metagenomic analyses are more computationally demanding, and will likely complicate processing samples from various sources.

2.6 Conclusion

WGS combined with the NCBI database is able to determine the species of most bacteria typically found in a clinical setting with high resolution, low cost, and short turnaround. Nonetheless, a more comprehensive reference genome database and validated species identity thresholds may be necessary for clinical implementation of WGS.

2.7 Support

The authors acknowledge partial support from the National Institutes of Health (NIH) grant UL1TR001442 and grant KL2 TR001112. Our study protocols were approved by Scripps Institutional Review Board (IRB-10-5522). We thank Scripps Sorrento Mesa Microbiology Laboratory for their services.

2.8 Acknowledgements

Chapter 2, in full is currently being prepared for submission for publication, Liu X.*, Pfeiffer W.*, Quarless D., Lee J., Oliveira G., Diamant J., and Schork N. "Clinical Bacterial Species Determination by Whole-Genome Sequencing: A Proof of Concept Study". The dissertation author is the primary researcher and author on this paper.

2.9 Supplemental Appendix

2.9.1 Species Delineation By Average Nucleotide Identity

To compute the average nucleotide identity (ANI) between a sample genome and a genome in the NCBI database, we used a variant of the algorithm adopted by Goris, et al.[102]. Instead of splitting the sample genome into 1-kb fragments for the alignments with BLAST[103], we just used the variable-length contigs. Often each database genome also consisted of multiple contigs, rather than a single chromosome, so this worked as well when comparing two database genomes. In all cases we used the default blastn settings. By contrast, Goris, et al., used alternate settings to obtain better sensitivity when comparing two distantly related genomes, but we were most interested in closely related genomes. By comparing ANI and DNA-DNA hybridization values for 28 bacterial genomes in six phylogenetically distinct groups, Goris, et al., concluded that two genomes with an ANI value above 95% that covers more than 69% of the genomes are within the same species. They also showed that the ANI value and its coverage are highly correlated, so it is often sufficient to just report the former. In a much more extensive study involving 536 pairwise comparisons of genomes in 85 groups from the NCBI database, a species delineation threshold of 95-96% ANI was observed[104]. They further noted that two genomes with an ANI value between 94% and 96% are in a “transition zone" within which it is less certain whether they are the same species or not.

2.9.2 Illustrative Output for WGS Species and Strain Determination

To illustrate the information obtained from our WGS analysis, we provide results here for a typical sample: 2014-043. Its Velvet assembly generated 538 contigs with a total length of 2,704,594 b. The longest contig had a length of 96,326 b, and the N50 value was 34,602 b. The assembly took 2 min 27 s running on 8 cores of an Intel Sandy Bridge processor. The subsequent BLAST analysis with blastn, using -max_target_seqs 1 and default settings otherwise, found the best match for each contig against all of the genomes in the NCBI database as of November 10, 2014. For Sample 2014-043 this took 28 min 11s on 16 cores of two Intel Sandy Bridge processors.

Two Perl scripts that we wrote processed the BLAST output. These binned the best matches of the contigs by species and by strain above our 95% identity threshold and then output the corresponding number of matching contigs and the total length of the matches. These scripts took less than a second to run. Outputs for Sample 2014-043 are shown in Tables 2.4 and 2.4. Table 2.4 shows that practically all contigs of the assembly match to *Staphylococcus aureus* genomes in the database. The few matches to genomes of other species are presumably due to contamination or incorporation of short segments of DNA from the other species in the genome of our sample. Table 2.4 shows that the matches are too many different strains of *S. aureus*. The strain with the longest match length is *S. aureus* A8117.

Evolution of the NCBI Database

At the outset of our study we downloaded the NCBI database as of January 23, 2014 to SDSC, and this was used for the initial BLAST alignments of all of our samples. We found several genomes that had been assigned incorrect species and reported them to NCBI. By the time that our study was nearing completion, many more genomes had been added to the database, so we downloaded a newer version as of November 10, 2014. We used this to reanalyze all samples with inadequate species determinations by WGS or ones for which the WGS and laboratory determinations disagreed. We found several samples with much better matches using the newer database, and these improved results are those reported here. We also found that

two errors we reported had been corrected, though another one affecting our results had been added.

As the NCBI database continues to grow, the BLAST run times will likely increase. If this is deemed a problem, the local version of the database could be restricted to a smaller subset of relevant genomes.

2.9.3 Long Match/Multiple Species WGS Samples

Most samples had long matches to a single named species as shown in Table 2.4. However, some samples had long matches to multiple species, and these required further investigation. Table 2.6 contains information on the 33 samples that initially appeared to contain multiple species, since they had relatively long matched lengths of >300 kb to two to four species. However, examination of the assembly coverage for these samples suggested that all but four contained only a single species. To characterize these samples, we separated them into four categories in the table.

Category 1 of Table 2.6 contains 25 samples with long matches to multiple species, at least one of which has a numbered ID. According to NCBI, such unnamed isolates "are clearly distinct from currently recognized species [and] are tentatively designated at the species level". These unnamed isolates have not yet been characterized by traditional methods, or the species name has not yet been validly published." Such species are unknown to a clinical laboratory and cannot agree with the laboratory-determined species, which are named. This led us to ask how close genomically the multiple matching species are to each other and, especially, whether the species with numbered IDs are really distinct from named species. To quantify our answers, we computed the ANI between various combinations of sample-strain and strain-strain pairs. Some relevant results are shown in Table 2.7, and four are discussed here.

Sample 2014-038 has long matches to *Enterobacter cloacae UCICRE 5, Enterobacter hormaechei YT3 Enterobacter MGH 1*, and *Enterobacter MGH 33*. As shown in Table 2.7, all four strains have ANIs ≥98.7% relative to the sample. Thus they seem to be the same species, which we took to be *E. cloacae* because of its longest match. It also agrees with the species determined by the laboratory, but calls into question the validity of the *E. hormaechei YT3* strain being within a named species

distinct from E. cloacae in the NCBI database. Sample 2014-272 has long matches to *Aeromonas hydrophila 173*, *Aeromonas MDS58*, and *Aeromonas dhakensis AAK1*. All three strains have ANIs ≥96.9% relative to the sample and between each other. Thus they seem to be the same species, which we took to be *A. hydrophila* based upon its much longer match to the sample. This also calls into question the validity of *A. dhakensis AAK1* being a strain within a named species distinct from *A. hydrophila* in the NCBI database. The laboratory-determined species was *Aeromonas sobria*, which is not in the NCBI database. Sample 2014-330 has long matches to *Corynebacterium ATCC 6931* and *Corynebacterium HFH0028*. Relative to the sample, the first strain has an ANI of 96.4, while the second strain has an ANI of 95.4. These two strains are close genomically, with an ANI between them of 94.9, which is just below the 95 species identity threshold of Goris, et al 2.7. Neither strain has a Latin species name, however, whereas *Corynebacterium amycolatum SK46* does and has an ANI relative to the sample that is only slightly lower at 94.5. We took the latter, named species to be that of the sample and adjusted the matched length to be the sum of the lengths for *C. ATCC 6931* and *C. HFH0028*. Note that the laboratory also had difficulty identifying the species of this sample and reported it as "Mod *Corynebacterium* species, not JK". Sample 2014-352 has long matches to *Sphingomonas S17* and *Sphingomonas paucimobilis NBRC 13935*, and the two strains have ANIs ≥99.3 relative to the sample. Thus they are clearly the same species, which we took to be *S. paucimobilis*, even though its match was shorter. By comparison, the laboratory-determined species was *Elizabethkingia meningoseptica*. The ANI value between the sample and *E. meningoseptica 502*, a typical species, is 79.0 (and only over a very short portion of the genome), so the laboratory determination is incorrect.

2.9.4 Multiple Close Species in Different Genera Samples Matches

Category 2 of Table 2.6 contains two samples that had long matches to both *Stenotrophomonas maltophilia RR-10* and *Pseudomonas geniculata N1*. These strains assigned to different genera have an ANI between them of 96.2, so they are actually the same species, namely *Stenotrophomonas maltophilia*, which has been noted before [S7].

2.9.5 Two Distant Species Long Matches Due to NCBI Database Error

The two samples in Category 3 of Table 2.6 had long matches to NCBI genomes of very different species because of errors in the database. Sample 2014-356 had long matches to *Morganella morganii F675* and *Strongyloides ratti*. The latter is a nematode, not a bacterium, so it was likely contaminated with *M. morganii* when the genome was sequenced. Similarly, Sample 2014-370 had long matches to *Staphylococcus capitis QN1* and *Balansia obtecta B249*. The latter is a fungus, which again was presumably contaminated with *S. capitis* during sequencing. Samples with high assembly coverage that really did contain two species Category 4 of Table 2.6 contains four samples that had high assembly coverage and so really did contain whole genomes from two different species as well as from different genera.

2.9.6 Adequate Coverage Matches To Numbered ID Species

Five samples had longest matches of adequate coverage to a species with a numbered ID. Two of those, 2014-330 and 2014-352, each had two long matches as discussed previously and were determined to be *Corynebacterium amycolatum* and *Sphingomonas paucimobilis*, respectively. For a third sample, 2014-170, our WGS procedure gave only a single long match to *Corynebacterium HFH0082* with an ANI of 98.1% as shown in Table S4. However, this sample also matches *Corynebacterium amycolatum SK46* with an ANI of 95.2%. Thus we took the latter species with a Latin name to be the one determined by WGS. By contrast, the laboratory initially determined the species to be *Enterobacter cloacae* and changed that to *Klebsiella pneumoniae* after repeat testing. The ANI value between the sample and *K. pneumoniae MGH 18*, a typical species, is 79.9% (and only over a very short portion of the genome), so the laboratory determination is clearly incorrect. For two other samples, 2014-276 and 2014-286, our WGS procedure gave only a single long match to Staphylococcus HGB0015 with an ANI of 95.2%. Since the NCBI database had no species with Latin names that gave close matches, we took the preceding species with a numbered ID to be that determined by WGS. These two samples were the only ones with adequate matched coverage for which we could not find a genomically

equivalent species with a Latin name. The clinical laboratory determined the species for these samples to be *Staphylococcus schleiferi*, which is not in the NCBI database. Thus it is possible that the WGS and laboratory species are, in fact, the same.

2.9.7 Distribution by species of samples with a single species adequately determinated by WGS

Table 2.8 shows the distribution by species of the 346 samples for which our WGS procedure adequately determined a single species after the preceding adjustments were made. There are 39 distinct species.

2.9.8 Samples with inadequate species determination by WGS and their associated ANI values

Using our threshold of $\geq$95% identity across $\geq$25% of the genome, WGS was unable to determine the species for only four samples. These are in the lower left corner of Figure 2A and are the first four listed in Category 1 of Table 2.10, where the samples are ordered by matched coverage. The remaining 13 samples in Category 2 of the table have the next lowest matched coverage values of $\geq$25% but $<$70%, so these all have adequate species determinations. Also listed in the table are the ANI values between each sample and the WGS species as well as between each sample and the laboratory species when it was determined and in the NCBI database. All of the samples in Category 1 have ANI values well below the 95% species identity threshold of Goris, et al. 2.7, and all of the samples in Category 2 have ANI values above the 95% threshold, except for the two *Stenotrophomonas maltophilia* samples with a slightly lower ANI of 93.1%. Three other samples in Category 2 have ANI coverage values below 69% because of low read coverage and so might not strictly pass the species identity requirement of Goris, et al. On balance, though, our species identity threshold and that of Goris, et al., seem nearly equivalent. The first three samples of Category 1 have laboratory species with very low ANI values. This indicates that the laboratory species determination was not correct for these. Table S7 contains results for those samples with repeat species determination by the clinical laboratory.

Table 2.4: Output of BLAST matches binned by species for Sample 2014-043

Sample ID	Matched contigs	Matched length (b)	Species for contigs with ≥95% identity
2014-043	508	2,704,534	Staphylococcus aureus
2014-043	7	4,155	Staphylococcus epidermidis
2014-043	7	605	Staphylococcus massiliensis
2014-043	3	308	Staphylococcus arlettae
2014-043	2	222	Staphylococcus OJ82
2014-043	1	157	Staphylococcus lentus
2014-043	1	78	Staphylococcus AL1
2014-043	1	76	Enterococcus GMD2E

Table 2.5: Complete output of BLAST matches binned by strain for Sample 2014-043.

Sample ID	Matched contigs	Matched length (b)	Strain for contigs with ≥95% identity
2014-043	16	461,381	Staphylococcus aureus T59618
2014-043	10	393,857	Staphylococcus aureus A8117
2014-043	13	156,053	Staphylococcus aureus subsp. aureus VRS10
2014-043	243	154,167	Staphylococcus aureus subsp. aureus CM05
2014-043	95	138,680	Staphylococcus aureus NN54 DNA,
2014-043	3	137,946	Staphylococcus aureus DAR104
2014-043	1	85,412	Staphylococcus aureus DAR3176
2014-043	35	83,319	Staphylococcus aureus subsp. aureus VRS11b
2014-043	4	81,164	Staphylococcus aureus subsp. aureus CIGC340D
2014-043	2	76,662	Staphylococcus aureus DAR3198
2014-043	1	62,620	Staphylococcus aureus T15889
2014-043	1	59,024	Staphylococcus aureus strain 502A,
2014-043	5	58,886	Staphylococcus aureus subsp. aureus CIG1150
2014-043	1	58,335	Staphylococcus aureus subsp. aureus N315 DNA,
2014-043	1	48,955	Staphylococcus aureus subsp. aureus 21201
2014-043	2	48,445	Staphylococcus aureus subsp. aureus VRS2
2014-043	1	42,446	Staphylococcus aureus DAR3170
2014-043	1	41,303	Staphylococcus aureus DAR3157
2014-043	1	40,537	Staphylococcus aureus M1291
2014-043	2	36,173	Staphylococcus aureus subsp. aureus 21272
2014-043	1	36,127	Staphylococcus aureus subsp. aureus CIG1750
2014-043	1	35,262	Staphylococcus aureus SMMC6038
2014-043	4	35,140	Staphylococcus aureus subsp. aureus VRS3a
2014-043	1	33,963	Staphylococcus aureus DAR1814
2014-043	1	32,430	Staphylococcus aureus DAR22
2014-043	1	32,055	Staphylococcus aureus DAR26
2014-043	1	30,311	Staphylococcus aureus A9299
2014-043	1	28,939	Staphylococcus aureus DAR1941
2014-043	1	24,395	Staphylococcus aureus DAR3183
2014-043	1	21,276	Staphylococcus aureus F84732
2014-043	2	19,982	Staphylococcus aureus subsp. aureus IS-122
2014-043	3	19,841	Staphylococcus aureus subsp. aureus VRS6
2014-043	1	19,102	Staphylococcus aureus M39274
2014-043	1	15,111	Staphylococcus aureus subsp. aureus 21318
2014-043	1	10,648	Staphylococcus aureus F52326
2014-043	1	10,355	Staphylococcus aureus subsp. aureus VRS4
2014-043	1	7,594	Staphylococcus aureus H61003
2014-043	1	5,182	Staphylococcus aureus subsp. aureus MR1
2014-043	2	3,086	Staphylococcus aureus subsp. aureus VRS7
2014-043	3	2,918	Staphylococcus epidermidis NIHLM020
2014-043	1	2,750	Staphylococcus aureus DAR3166
2014-043	2	2,273	Staphylococcus aureus subsp. aureus CIG1213
2014-043	1	1,954	Staphylococcus aureus M0455
2014-043	3	1,679	Staphylococcus aureus subsp. aureus VRS11a
2014-043	13	1,520	Staphylococcus aureus subsp. aureus GR1
2014-043	1	1,355	Staphylococcus aureus subsp. aureus VRS5
2014-043	1	1,271	Staphylococcus aureus H12893
2014-043	3	726	Staphylococcus epidermidis VCU065
2014-043	1	673	Staphylococcus aureus DAR1174
2014-043	1	643	Staphylococcus aureus M1112
2014-043	7	605	Staphylococcus massiliensis CCUG 55927
2014-043	1	554	Staphylococcus aureus M0460
2014-043	1	520	Staphylococcus aureus M0602
2014-043	1	511	Staphylococcus epidermidis NIHLM053
2014-043	1	495	Staphylococcus aureus OCMM6110
2014-043	1	489	Staphylococcus aureus M0125
2014-043	2	397	Staphylococcus aureus PM1 DNA,
2014-043	1	381	Staphylococcus aureus M1140
2014-043	5	321	Staphylococcus aureus subsp. aureus Newbould 305
2014-043	1	306	Staphylococcus aureus DAR3162
2014-043	2	258	Staphylococcus aureus H20322
2014-043	2	253	Staphylococcus arlettae CVD059 SARL_c56,
2014-043	1	206	Staphylococcus aureus W82239
2014-043	1	157	Staphylococcus lentus F1142
2014-043	1	150	Staphylococcus OJ82 155.SOJ.1_6,
2014-043	1	116	Staphylococcus aureus subsp. aureus VRS9
2014-043	1	101	Staphylococcus aureus subsp. aureus VRS1
2014-043	1	78	Staphylococcus AL1
2014-043	1	76	Enterococcus GMD2E
2014-043	1	72	Staphylococcus OJ82 155.SOJ.1_4,
2014-043	1	58	Staphylococcus aureus W91963
2014-043	1	55	Staphylococcus arlettae CVD059 SARL_c156,
2014-043	1	50	Staphylococcus aureus subsp. aureus CF-Marseille

Table 2.6: Samples that appeared to contain multiple species upon initial analysis by WGS (n=31).

Sample ID	Assembly length (b)	Matched contigs	Matched length (b)	Initial WGS determination	Laboratory determination	Typical genome length (b)	Assembly coverage	Matched coverage	Final WGS determination
Category 1: Samples with long match to at least one close species with a numbered ID in the same genus (n=25)									
2014-038	4,941,638	547	2,464,997	*Enterobacter cloacae*	*Enterobacter cloacae*	5,310,000	93%	46%	One species
2014-038		125	902,347	*Enterobacter hormaechei*		4,810,000	103%	19%	
2014-038		79	354,049	*Enterobacter MGH 1*		5,133,520	96%	7%	
2014-038		59	324,489	*Enterobacter MGH 33*		4,953,280	100%	7%	
2014-075	4,833,596	649	3,013,841	*Enterobacter cloacae*	*Enterobacter cloacae*	5,310,000	91%	57%	One species
2014-075		99	579,348	*Enterobacter MGH 33*		5,133,520	94%	11%	
2014-075		87	420,018	*Enterobacter hormaechei*		4,810,000	100%	9%	
2014-104	5,437,034	1,237	3,258,194	*Serratia marcescens*	*Serratia marcescens*	5,122,570	106%	64%	One species
2014-104		744	1,671,898	*Serratia UC1SER*		5,027,440	108%	33%	
2014-119	4,792,966	257	2,559,969	*Enterobacter BIDMC 27*	*Enterobacter cloacae*	4,838,730	99%	53%	One species
2014-119		637	2,092,613	*Enterobacter cloacae*		5,310,000	90%	39%	
2014-157	3,701,570	426	2,434,534	*Morganella morganii*	*Morganella morganii*	3,799,540	97%	64%	One species
2014-157		97	1,010,670	*Morganella EGD-HP17*		3,947,790	94%	26%	
2014-179	4,714,069	700	2,368,793	*Enterobacter cloacae*	*Enterobacter cloacae*	5,310,000	89%	45%	One species
2014-179		226	937,240	*Enterobacter MGH 37*		5,088,330	93%	18%	
2014-179		179	861,006	*Enterobacter MGH 22*		4,684,380	101%	18%	
2014-232	4,706,619	1,889	2,727,778	*Enterobacter cloacae*	*Enterobacter cloacae*	5,310,000	89%	51%	One species
2014-232		264	508,037	*Enterobacter hormaechei*		4,810,000	98%	11%	
2014-232		362	487,464	*Enterobacter MGH 1*		4,953,280	95%	10%	
2014-232		227	422,731	*Enterobacter MGH 33*		5,133,520	92%	8%	
2014-247	3,963,648	590	2,579,088	*Morganella morganii*	*Morganella morganii*	3,799,540	104%	68%	One species
2014-247		143	1,035,418	*Morganella EGD-HP17*		3,947,790	100%	26%	
2014-272	5,045,699	1,077	3,708,908	*Aeromonas hydrophila*	*Aeromonas sobria*	4,740,000	106%	78%	One species
2014-272		160	377,676	*Aeromonas dhakensis*		4,841,750	104%	8%	
2014-272		106	367,929	*Aeromonas MDS8*		4,760,000	106%	8%	
2014-274	3,962,897	585	2,520,231	*Morganella morganii*	*Morganella morganii*	3,799,540	104%	66%	One species
2014-274		118	1,112,657	*Morganella EGD-HP17*		3,947,790	100%	28%	
2014-280	2,424,386	152	1,756,802	*Staphylococcus capitis*	*Staphylococcus capitis*	2,460,000	99%	71%	One species
2014-280		23	451,642	*Staphylococcus TE8*		2,516,640	96%	18%	
2014-291	4,533,192	634	2,842,750	*Enterobacter cloacae*	*Enterobacter cloacae*	5,310,000	85%	54%	One species
2014-291		117	457,056	*Enterobacter MGH 1*		4,953,280	92%	9%	
2014-291		78	341,915	*Enterobacter MGH 33*		5,133,520	88%	7%	
2014-291		78	328,458	*Enterobacter hormaechei*		4,810,000	94%	7%	
2014-307	3,962,602	625	2,484,770	*Morganella morganii*	*Morganella morganii*	3,799,540	104%	65%	One species
2014-307		157	1,110,228	*Morganella EGD-HP17*		3,947,790	100%	28%	
2014-309	5,162,851	1,608	3,278,678	*Serratia marcescens*	*Many Serratia marcescens*	5,122,570	101%	64%	One species
2014-309		898	1,663,664	*Serratia UC1SER*		5,027,440	103%	33%	
2014-311	4,876,408	5,454	3,536,072	*Serratia marcescens*	*Mod Serratia marcescens*	5,122,570	95%	69%	One species
2014-311		2,378	1,075,738	*Serratia UC1SER*		5,027,440	97%	21%	
2014-312	6,145,600	1,768	2,665,503	*Klebsiella oxytoca*	*Klebsiella oxytoca*	5,970,000	103%	45%	One species
2014-312		2,236	2,543,676	*Klebsiella OBRC7*		6,343,310	97%	40%	
2014-330	2,562,036	154	1,238,945	*Corynebacterium ATCC 6931*	*Mod Corynebacterium species, not JK*	2,471,920	104%	50%	One species
2014-330		111	602,904	*Corynebacterium HFH0082*		2,493,220	103%	24%	
2014-331	5,000,662	1,048	4,166,412	*Serratia marcescens*	*Serratia plymuthica*	5,122,570	98%	81%	One species
2014-331		253	681,697	*Serratia UC1SER*		5,027,440	99%	14%	
2014-332	3,884,032	277	2,188,437	*Acinetobacter baumannii*	*Acinetobacter baumannii*	3,980,000	98%	55%	One species
2014-332		55	602,095	*Acinetobacter NIPH 1847*		3,992,490	97%	15%	
2014-332		55	537,444	*Acinetobacter NIPH 3623*		3,954,090	98%	14%	
2014-333	4,930,626	549	2,158,258	*Citrobacter KTE32*	*Citrobacter braakii*	4,933,940	100%	44%	One species
2014-333		527	1,626,680	*Citrobacter freundii*		4,904,640	101%	33%	
2014-333		243	648,021	*Citrobacter KTE151*		5,266,560	94%	12%	
2014-336	4,290,758	8,624	2,758,203	*Escherichia coli*	*Escherichia coli*	4,720,000	91%	58%	One species
2014-336		3,244	905,591	*Escherichia 1_1_43*		4,708,150	91%	19%	
2014-352	4,396,637	1,704	3,368,475	*Sphingomonas S17*	*Elizabethkingia meningoseptica*	4,268,410	103%	79%	One species
2014-352		142	385,237	*Sphingomonas paucimobilis*		4,327,400	102%	9%	
2014-365	6,407,083	609	3,554,505	*Klebsiella oxytoca*	*Klebsiella oxytoca*	5,970,000	107%	60%	One species
2014-365		466	1,831,240	*Klebsiella OBRC7*		6,343,310	101%	29%	
2014-374	5,232,860	1,467	4,085,154	*Serratia marcescens*	*Serratia marcescens*	5,220,000	100%	78%	One species
2014-374		265	608,142	*Serratia SCBI*		5,101,900	103%	12%	
2014-380	2,446,054	172	1,854,385	*Staphylococcus capitis*	*Staphylococcus capitis*	2,460,000	99%	75%	One species
2014-380		25	358,914	*Staphylococcus TE8*		2,516,640	97%	14%	
Category 2: Samples with long matches to close species assigned to different genera (n=2)									
2014-101	4,621,863	1,112	1,443,191	*Stenotrophomonas maltophilia*	*Stenotrophomonas malthophilia*	4,850,000	95%	30%	One species
2014-101		254	366,338	*Pseudomonas geniculata*		4,511,340	102%	8%	
2014-262	4,615,432	1,046	1,397,972	*Stenotrophomonas maltophilia*	*Mod Stenotrophomonas maltophilia*	4,850,000	95%	29%	One species
2014-262		241	342,958	*Pseudomonas geniculata*		4,511,340	102%	8%	
Category 3: Samples with long matches to two very different species, one of which is the result of an error in the NCBI database (n=2)									
2014-356	4,079,479	2,217	3,188,499	*Morganella morganii*	*Morganella morganii*	3,799,540	107%	84%	One species; *S. ratti* includes
2014-356		1,047	640,076	*Strongyloides ratti*		52,640,000	8%	1%	*M. morganii* in database
2014-370	2,446,261	198	1,885,923	*Staphylococcus capitis*	*Few Staphylococcus capitis*	2,460,000	99%	77%	One species; *B. obtecta* includes
2014-370		14	463,560	*Balansia obtecta*		30,150,000	8%	2%	*S. capitis* in database
Category 4: Samples with high assembly coverage that really did contain two species (n=4)									
2014-081	5,605,844	630	2,797,698	*Enterococcus faecalis*	*Enterococcus faecalis*	3,220,000	174%	87%	Two species
2014-081		612	2,708,217	*Staphylococcus aureus*		2,810,000	199%	96%	
2014-112	10,128,092	7,760	5,893,106	*Pseudomonas aeruginosa*	*Pseudomonas aeruginosa*	6,260,000	162%	94%	Two species
2014-112		1,386	3,899,133	*Proteus mirabilis*		4,060,000	249%	96%	
2014-228	3,634,830	268	1,857,673	*Streptococcus intermedius*	*Streptococcus intermedius*	1,930,000	188%	96%	Two species
2014-228		9,767	1,702,400	*Staphylococcus aureus*		2,810,000	129%	61%	
2014-268	5,576,209	10,983	2,815,320	*Proteus mirabilis*	*Rare Streptococcus parasanguinis*	4,060,000	137%	69%	Two species
2014-268		441	1,244,865	*Streptococcus parasanguinis*		2,150,000	259%	58%	

Table 2.7: Distribution of samples by species adequately determined by WGS.

Sample ID	Determination approach	Species	BLAST pair of genomes	ANI (%)
2014-038	WGS	*Enterobacter cloacae* +	2014-038 vs. *E. cloacae UCICRE 5* (top 4)	98.8
		Enterobacter hormachei +	2014-038 vs. *E. hormaechei YT3* (top 4)	98.8
		Enterobacter MGH 1 +	2014-038 vs. E. MGH 1 (top 4)	98.8
		Enterobacter MGH 33	2014-038 vs. E. MGH 33 (top 4)	98.7
	Laboratory	*Enterobacter cloacae*	*E. cloacae UCICRE 5* vs. *E. hormaechei YT3*	98.2
2014-101	WGS	*Stenotrophomonas maltophilia* +	2014-101 vs. *S. maltophilia RR-10* (top 2)	93.1
		Pseudomonas geniculata	2014-101 vs. *P. geniculata N1* (top 2)	93.1
	Laboratory	*Stenotrophomonas maltophilia*	*S. maltophilia RR-10* vs. *P. geniculata N1*	96.2
2014-104	WGS	*Serratia marcescens* +	2014-104 vs. *S. marcescens NGS-ED-1015* (top 2)	98.9
		Serratia UC1SER	2014-104 vs. *S. UC1SER* (top 2)	98.9
	Laboratory	*Serratia marcescens*	*S. marcescens NGS-ED-1015* vs. *S. UC1SER*	98.6
2014-157	WGS	*Morganella morganii KT* +	2014-157 vs. *M. morganii KT* (top 2)	98.7
		Morganella EGD-HP17	2014-157 vs. *M. EGD-HP17* (top 2)	98.6
	Laboratory	*Morganella morganii KT*	*M. morganii KT* vs. *M. EGD-HP17*	98.3
2014-170	WGS	*Corynebacterium HFH0082*	2014-170 vs. *C. HFH0082* (best)	98.1
	Laboratory initially	*Enterobacter cloacae*	2014-170 vs. *C. amycolatum SK46* (alternate)	95.2
	Laboratory after repeat	*Klebsiella pneumoniae*	2014-170 vs. *K. pneumoniae MGH 18* (typical)	79.9
			C. HFH0082 vs. *C. amycolatum SK46*	95.3
2014-272	WGS	*Aeromonas hydrophila* +	2014-272 vs. *A. hydrophila 173* (best)	97.2
		Aeromonas MDS8 +	2014-272 vs. *A. MDS8* (2nd best)	97.1
		Aeromonas dhakensis	2014-272 vs. *A. dhakensis AAK1* (3rd best)	97.1
	Laboratory	*Aeromonas sobria*	*A. hydrophila 173* vs. *A. MDS8*	96.9
			A. hydrophila 173 vs. *A. dhakensis AAK1*	97.0
2014-280	WGS	*Staphylococcus capitis* +	2014-280 vs. *S. capitis SK14* (top 2)	98.9
		Staphylococcus TE8	2014-280 vs. *S. TE8* (top 2)	99.1
	Laboratory	*Staphylococcus capitis*	*S. capitis SK14* vs. *S. TE8*	98.5
2014-312	WGS	*Klebsiella oxytoca* +	2014-312 vs. *K. oxytoca 10-5242* (top 2)	99.2
		Klebsiella OBRC7	2014-312 vs. *K. OBRC7* (top 2)	99.0
	Laboratory	*Klebsiella oxytoca*	*K. oxytoca 10-5242* vs. *K. OBRC7*	98.6
2014-330	WGS	*Corynebacterium ATCC 6931* +	2014-330 vs. *C. ATCC 6931* (best)	96.4
		Corynebacterium HFH0082	2014-330 vs. *C. HFH0082* (2nd best)	95.4
	Laboratory	*Mod Corynebacterium species, not JK*	2014-330 vs. *C. amycolatum SK46* (alternate)	94.5
			C. ATCC 6931 vs. *C. HFH0082*	94.9
			C. ATCC 6931 vs. *C. amycolatum SK46*	94.4
2014-332	WGS	*Acinetobacter baumannii* +	2014-332 vs. *A. baumanii 146457* (best)	96.8
		Acinetobacter NIPH 3623 +	2014-332 vs. A. NIPH 3623 (2nd best)	96.5
		Acinetobacter NIPH 1847	2014-332 vs. A. NIPH 1847 (3rd best)	96.5
	Laboratory	*Acinetobacter baumannii*	*A. baumanii 146457* vs. *A. NIPH 3623*	96.2
			A. baumanii 146457 vs. *A. NIPH 1847*	96.2
2014-333	WGS	*Citrobacter KTE32* +	2014-333 vs. *C. KTE32* (best)	99.0
		Citrobacter freundii +	2014-333 vs. *C. freundii ballerup 7851/39* (2nd best)	99.0
		Citrobacter KTE151	2014-333 vs. *C. KTE151* (3rd best)	98.8
	Laboratory	*Citrobacter braakii*	*C. freundii ballerup 7851/39* vs. *C. KTE32*	98.7
			C. freundii ballerup 7851/39 vs. *C. KTE151*	98.5
2014-336	WGS	*Escherichia coli* +	2014-336 vs. *E. coli K-12 MG1655* (best)	99.6
		Escherichia 1_1_43	2014-336 vs. *E. 1_1_43* (2nd best)	99.3
	Laboratory	*Escherichia coli*	*E. coli K-12 MG1655* vs. *E. 1_1_43*	98.1
2014-352	WGS	*Sphingomonas S17* +	2014-352 vs. *S. S17* (best)	99.4
		Sphingomonas paucimobilis	2014-352 vs. *S. paucimobilis NBRC 13935* (2nd best)	99.3
	Laboratory	*Elizabethkingia meningoseptica*	2014-352 vs. *E. meningoseptica 502* (typical)	79.0
			S. S17 vs. *S. paucimobilis NBRC 13935*	98.6
2014-356	WGS	*Morganella morganii* +	2014-356 vs. *M. morganii F675* (best)	99.7
		Strongyloides ratti	2014-356 vs. *S. ratti* (2nd best)	98.2
	Laboratory	*Morganella morganii*	*M. morganii F675* vs. *S. ratti*	97.0
2014-370	WGS	*Staphylococcus capitis* +	2014-370 vs. *S. capitis QN1*	99.6
		Balansia obtecta	2014-370 vs. *B. obtecta B249*	99.7
	Laboratory	*Staphylococcus capitis*	*S. capitis QN1* vs. *B. obtecta B249*	99.4
2014-374	WGS	*Serratia marcescens* +	2014-374 vs. *S. marcescens BIDMC 50* (best)	98.8
		Serratia SCBI	2014-374 vs. *S. SCBI* (2nd best)	98.6
	Laboratory	*Serratia marcescens*	*S. marcescens BIDMC 50* vs. *S. SCBI*	98.3

Table 2.8: Distribution of samples by species adequately determined by WGS.

	WGS samples with listed species (after adjustments)	WGS samples with species that disagrees with laboratory
Single species		
Staphylococcus aureus	118	0
Escherichia coli	51	0
Enterococcus faecalis	39	3
Enterococcus faecium	4	1
2 Enterococcus spp.	43	4
Staphylococcus capitis	4	0
Staphylococcus caprae	1	1
Staphylococcus epidermidis	12	3
Staphylococcus haemolyticus	1	0
Staphylococcus hominis	1	1
Staphylococcus lugdunensis	5	1
Staphylococcus saprophyticus	1	0
Staphylococcus simulans	1	1
8 CoN Staphylococcus spp.	26	7
Pseudomonas aeruginosa	21	0
Klebsiella oxytoca	5	0
Klebsiella pneumoniae	14	1
2 Klebsiella spp.	19	1
Enterobacter aerogenes	5	0
Enterobacter cloacae	8	0
2 Enterobacter spp.	13	0
Streptococcus agalactiae	1	0
Streptococcus anginosus	3	1
Streptococcus intermedius	1	0
Streptococcus oralis	2	0
Streptococcus parasanguinis	4	2
Streptococcus pneumoniae	4	0
Streptococcus HGB0015	2	2
7 Streptococcus spp.	17	5
Proteus mirabilis	9	2
Morganella morganii	5	0
Serratia marcescens	5	1
Corynebacterium amycolatum	2	2
Corynebacterium striatum	2	2
2 Corynebacterium spp.	4	4
Achromobacter xylosoxidans	2	2
Acinetobacter baumannii	3	0
Aeromonas hydrophila	1	1
Bacillus licheniformis	1	1
Citrobacter freundii	1	1
Citrobacter koseri	1	0
Myroides odoratimimus	2	0
Providencia stuartii	1	0
Sphingomonas paucimobilis	1	1
Stenotrophomonas maltophilia	2	0
10 Other spp.	15	6
Total	346	30
Two WGS species		
Enterococcus faecalis	1	0
Proteus mirabilis	2	2
Pseudomonas aeruginosa	1	0
Staphylococcus aureus	2	2
Streptococcus intermedius	1	0
Streptococcus parasanguinis	1	0
Total	8	4

Table 2.9: Samples with <70% matched coverage at ≥95% identity (n=19).

Sample ID	WGS determination	Laboratory determination	Assembly coverage (%) based on WGS species	Matched coverage (%) at ≥95% identity	ANI (%) for WGS species	ANI (%) for lab species	ANI coverage (%) for WGS species
Category 1: Inadequate species determination by WGS with <25% matched coverage for ≥95% identity threshold (n=4)							
2014-318	*Pantoea At-9b*	*Rare Pantoae agglomerans*	70	0.2	80.2	80.2	26.4
2014-386	*Sphingomonas S17*	*Staphylococcus epidermidis*	95	1.2	86.5	79.2	63.0
2014-376	*Sphingobium yanoikuyae*	*Pseudomonas fluorescens*	81	1.8	86.4	78.1	52.2
2014-382	*Paenibacillus MAEPY1*	*Bacillus species, not anthracis or cereus*	91	4.0	91.3	NA	74.4
Category 2: Adequate species determination by WGS with ≥25% but <70% matched coverage for ≥95% identity threshold (n=13)							
2014-063 *	*Streptococcus oralis*	*Streptococcus oralis*	80	30.4	96.1	<- same	62.5
2014-064 *	*Streptococcus oralis*	*Streptococcus oralis*	63	30.9	97.2	<- same	50.1
2014-262	*Stenotrophomonas maltophilia*	*Mod Stenotrophomonas maltophilia*	95	35.9	93.1	<- same	83.0
2014-101	*Stenotrophomonas maltophilia*	*Stenotrophomonas malthophilia*	95	37.3	93.1	<- same	83.3
2014-231	*Streptococcus parasanguinis*	*Steptococcus parasanguinis*	101	49.8	95.3	<- same	85.6
2014-060 *	*Achromobacter xylosoxidans*	*Achromobacter species*	57	50.4	99.0	NA	48.3
2014-276	*Staphylococcus HGB0015*	*Mod Staphylococcus schleiferi*	104	55.6	95.2	not in db	93.0
2014-371	*Streptococcus parasanguinis*	*Streptococcus parasanguinis*	102	55.7	95.2	<- same	87.9
2014-182	*Streptococcus parasanguinis*	*Streptococcus viridans*	99	57.7	95.5	not in db	85.9
2014-286	*Staphylococcus HGB0015*	*Mod staphylococcus schleiferi*	104	57.9	95.2	not in db	93.0
2014-176	*Streptococcus parasanguinis*	*Prevotella species*	99	58.1	95.5	NA	86.0
2014-239	*Streptococcus anginosus*	*Streptococcus anginosus*	98	61.9	96.7	<- same	88.8
2014-229	*Streptococcus anginosus*	*Streptococcus anginosus*	98	65.1	96.7	<- same	88.8

* This is one of three samples with low read coverage (5-7x), which led to low assembly coverage (57-80%) and low ANI coverage (48-63%) for the WGS species.

Table 2.10: Distribution of samples by species adequately determined by WGS.

Sample ID	WGS determination	Initial laboratory determination	Repeat laboratory determination	Culture source
Category 1: Laboratory species changed after repeat (n=12)				
1A: WGS and laboratory species agree (n-6)				
2014-055	*Enterococcus faecalis*	*Streptococcus parasanguinis*	*Enterococcus faecalis*	blood
2014-061	*Myroides odoratimimus*	*Klebsiella oxytoca*	*Myroides odoratimimus*	left leg
2014-065	*Staphylococcus aureus*	*Staphylococcus, coagulase negative (inadequate)*	*Staphylococcus aureus*	urine
2014-202	*Pseudomonas aeruginosa*	*Pseudomonas species, not aeruginosa (inadequate)*	*Pseudomonas aeruginosa*	sputum
2014-290	*Klebsiella pneumoniae*	*Enterobacter cloacae*	*Klebsiella pneumoniae*	right heel
2014-347	*Staphylococcus lugdunensis*	*Staphylococcus, coagulase negative (inadequate)*	*Staphylococcus lugdunensis*	urine
1B: WGS and laboratory species do not agree or fully agree (n=6)				
2014-048	*Enterococcus faecium*	*Skin contamination (inadequate)*	*Enterococcus faecalis*	urine
2014-066	*Staphylococcus epidermidis*	*MRSA*	*Staphylococcus, coagulase negative (inadequate)*	blood
2014-067	*Staphylococcus epidermidis*	*Streptococcus parasanguinis*	*Staphylococcus hominis*	blood
2014-092	*Bacillus licheniformis*	*Actinobaculum (inadequate)*	*Bacillus species, not anthracis (inadequate)*	blood
2014-170	*Corynebacterium HFH0082*	*Enterobacter cloacae*	*Klebsiella pneumoniae*	right knee
2014-237	*Staphylococcus lugdunensis*	*Staphylococcus schleiferi ssp Schleiferi*	*Staphylococcus chromogenes*	scalp ulcer
Category 2: Laboratory species did not change after repeat (n=11)				
2A: WGS and laboratory species still agree (n=3)				
2014-059	*Escherichia coli*	*Escherichia coli*	*Escherichia coli*	abdomen
2014-342	*Klebsiella pneumoniae (adjusted)*	*Klebsiella pneumoniae*	*Klebsiella pneumoniae*	kidney
2014-381	*Klebsiella pneumoniae (adjusted)*	*Many Klebsiella pneumoniae*	*Klebsiella pneumoniae*	pleural fluid
2B: WGS and laboratory species still do not agree (n=5)				
2014-047	*Staphylococcus caprae*	*Staphylococcus capitis*	*Staphylococcus capitis*	right femoral
2014-081	*Enterococcus faecalis, Staphylococcus aureus*	*Enterococcus faecalis*	*Enterococcus faecalis*	blood
2014-114	*Staphylococcus hominis*	*Staphylococcus epidermidis*	*Staphylococcus epidermidis*	blood
2014-258	*Proteus mirabilis*	*Proteus vulgaris*	*Proteus vulgaris*	peg tube
2014-363	*Proteus mirabilis*	*Proteus vulgaris*	*Proteus vulgaris*	peg tube
2C: Species was not adequately determined by WGS (n=3)				
2014-318		*Pantoae agglomerans*	*Pantoae agglomerans*	right knee
2014-386		*Staphylococcus epidermidis*	*Staphylococcus epidermidis*	unspecified
2014-101		*Stenotrophomonas maltophilia*	*Stenotrophomonas maltophilia*	sputum

Chapter 3

Implications of Methicillin-Resistant *Staphylococcus Aureus* (MRSA) Reference Genome Choice for Investigating Clinical Correlates

3.1 Abstract

Methicillin-Resistant *Staphylococcus Aureus* (MRSA) is an opportunistic infectious pathogen of epidemic proportions. Recently, technological advancements have enabled reference guided whole-genome sequencing (WGS) of bacterial genomes. This has uncovered vast heterogeneity in the function of single nucleotide variations (SNVs), which complicates the process of associating clinical phenotypes, genetic profile elements, and antibiotic resistance because SNV based investigations depend critically on the contents of the reference genome. Multiple reference genomes have been utilized throughout the scientific literature. However the contribution of individual reference genomes is unknown for MRSA analyses. We hypothesize that additional sources of variation in DNA sequencing protocols can significantly confound investigations aiming to elucidate pathogen-related genetic virulence associations. Analyses aimed to interrogate this genetic variation, either within or across pathogen strains, must assert assumptions concerning the most optimal reference genome or comparator

strain.

To assess the effect of incorporating multiple reference genomes on investigations aiming associate clinical phenotypes with pathogen virulence factors, we implemented identical WGS analytic pipelines with two different MRSA reference genomes used in literature. DNA Sequences, antibiogram drug resistance, and patient response data were collected for 302 bacterial cultures isolated from routine clinical microbiology laboratory workflows. We describe strategies for multiple reference genome implementations while observing that bacterial reference genome strain can adversely influence clinically relevant claims about pathogenicity which can confound clinical studies. We find that clinical correlations differ amongst reference strain utilization and argue that care must be taken in pursuing studies of genetic variation in complex, highly variable pathogens such as MRSA.

3.2 Introduction

Combatting infectious disease is a critical medical concern among industrialized and developed countries, where hospital and community acquired bacteria account for a significant proportion of global morbidity and mortality. *Staphylococcus aureus* is considered one of the most virulent multi-drug resistant pathogens, which may attribute its virulence and evolutionary capabilities to functional factors present throughout the genome[105–109]. It is well known that genetic elements augment pathogen virulence capabilities to rapidly evolve, adapt, and overcome the most sanitary hospital environments[110]. Great genetic diversity reflects this increase in infection rate and proliferative ability, which is largely due to 1) a substantial lack in efficacious drugs to treat pathogenic infections; 2) the much-recognized ability of infectious pathogens to adapt and develop genetic and biological mechanisms to evade therapeutic interventions, and 3) the wide-spread overuse and improper medical applications of antibiotic treatment[109,111–115]. Efficient and powerful high-throughput screening platforms, i.e. genomic sequencing and antibiotic profiling technologies, has shed light on the degree to which genetic elements influence andor correlate with clinically-relevant outcomes[86,112,116–118]. Specific genetic factors associate with individual phenotypes, however additional work is needed to uncover their complex inter-

play with pathogenicity, environmental pressures, and host-related factors[45,119–121].

Computational, and statistical hurdles impede associating genetic factors for drug resistance with clinical outcomes[122], not the least of which concern the inherent biological complexity. Thus, bacterial infection is complicated, where DNA sequencing based investigations critically depend on DNA sequence protocols and their ability to differentiate signal from noise[90,123–125]. Typically, millions of genome sequencing 'reads' (i.e., relatively small stretches of DNA sequence) captured by sequencers need to be reconciled, ordered and 'assembled' in order to identify differences (i.e., variants) in the genomes and their associations with clinical outcomes[126]. This process can be pursued completely unbiased through '*de novo* assembly' methods, that exclusively utilize reads and no reference genome to reverse engineer particular bacterial genome[127–129]. Unfortunately, *de novo* assembly approaches are computationally and strategically intensive, especially for comparatory investigations[130,131]. Therefore, reference-guided approaches, which match sequencing read(s) to a previously characterized genome, are implemented to aid target pathogen genome reconstruction, which inherently uncovers differences between the target and the reference genomes. These differences may reveal clinically meaningful DNA associations. However This raises the challenge of identifying the most optimal or proximal reference genome which may impact conclusions made about the target genomes in question.

3.2.1 Bacterial Virulence Determination Deficiencies

Clinical and Laboratory Standards Institute (CLSI) antibiogram protocols define gold-standard assays for the clinical practice of determining drug resistance[132]. In attempts to surpass binary drug response indications, high-throughput DNA sequencing instruments have aided antiobiogram protocols by characterizing highly variable individual mutations that contribute to the genetic architecture of drug resistance. A recent study investigating the genotype-phenotype relationship of MRSA antibiotic resistance demonstrated a 99.8 percent correlation between 12 antibiotics based on 193 MRSA samples[124,133,134], which implicated genetic penetrance as a mechanism for antibiotic resistance and pathogenicity. Thus, laboratory or sequencing methods that limit the number of DNA factors, i.e. specific gene regions, mutations, etc, may fail

to account for the full context of DNA functionality inherent in genome. This may potentially confounding research investigations aimed at identifying the biological underpinnings of infectious disease.

3.2.2 Antibiotic Resistance

Decreasing the prevalence of antibiotic resistance by understanding its genetic contributions would benefit the medical community. Over the past two decades, rates of β-lactam targeting methicillin, vancomycin, and multi-drug resistant bacterial phenotypes have unyieldingly increased. For example, resistance to the antibiotic penicillin was observed in clinical isolate strains less than 10 years after its clinical introduction in 1944, and the first MRSA strain was reported only one year after the drug's initial launch in 1975. 15 major antibiotic classes have been utilized clinically to treat pathogenic infections, however none have eluded the emergence of untreatable bacteria[112] (Table 3.1).

Table 3.1: From left to right, columns correspond to the antibiotic class with an example class, the year the drug was discovered, the year the compound was first introduced in clinical practice, the year antibiotic resistance was first observed, the Resistance mechanism of action, and the species the compound is effective for.

Antibiotic Class; Example	Discovery	Clinic Introduction	Resistance Observed	Mechanism of Action	Activity or Target
Sulfadrugs; prontosil	1932	1936	1942	Inhibition of dihydropteroate synthetase	Gram-positive bacteria
β-lactams; penicillin	1928	1938	1945	Inhibition of cell wall biosynthesis	Broad-spectrum activity
Aminoglycosides; streptomycin	1943	1946	1946	Binding of 30S ribosomal subunit	Broad-spectrum activity
Chloramphenicols; chloramphenicol	1946	1948	1950	Binding of 50S ribosomal subunit	Broad-spectrum activity
Macrolides; erythromycin	1948	1951	1955	Binding of 50S ribosomal subunit	Broad-spectrum activity
Tetracyclines; chlortetracycline	1944	1952	1950	Binding of 30S ribosomal subunit	Broad-spectrum activity
Rifamycins; rifampicin	1957	1958	1962	Binding of RNA polymerase β-subunit	Gram-positive bacteria
Glycopeptides; vancomycin	1953	1958	1960	Inhibition of cell wall biosynthesis	Gram-positive bacteria
Quinolones; ciprofloxacin	1961	1968	1968	Inhibition of DNA synthesis	Broad-spectrum activity
Streptogramins; streptogramin B	1963	1998	1964	Binding of 50S ribosomal subunit	Gram-positive bacteria
Oxazolidinones; linezolid	1955	2000	2001	Binding of 50S ribosomal subunit	Gram-positive bacteria
Lipopetides; daptomycin	1986	2003	1987	Depolarization of cell membrane	Gram-positive bacteria
Fidaxomicin (targeting Clostridium difficile)	1948	2011	1977	Inhibition of RNA polymerase	Gram-positive bacteria
Fidaxomicin (targeting Clostridium difficile)	1948	2011	1977	Inhibition of RNA polymerase	Gram-positive bacteria
Diarylquinolines; bedaquiline	1997	2012	2006	Inhibition of F1FO-ATPase	Narrow-spectrum activity (Mycobacterium tuberculosis)

3.3 Overview

Bacterial resistance investigations often leverage statistical associations of genomic variation. These analyses often utilize different reference genomes as a result of heterogeneity in DNA sequencing computational strategies (Table 3.2). We asked the question, does a particular reference genome effect the study of a particular bacterial population? Our investigation explored the impact of utilizing two of most widely used and publicly available MRSA reference genomes, *Staphylococcus aureus* subsp. aureus TW20 (Genbank accession number: FN433596.1) and *Staphylococcus aureus* subsp. aureus HO 5096 0412 (Genbank accession number: HE681097.1) in variant calling practices when assessing genomic DNA sequence variation of clinical MRSA bacterial isolates. These reference genomes were analyzed in conjunction with whole genome sequence data obtained for 302 MRSA clinical isolates from patients in a single hospital system, and 49 MRSA DNA sequences acquired from the National Center for Biotechnology Information (NCBI)[135,136].

Our analyses suggest that the choice of a reference to guide the identification of genetic variations does make a difference, though not necessarily a pronounced one. We argue that the differences in clinically-meaningful associations based on reference choice are an inevitable product of the very pathogen genomic diversity of interest and that newer strategies for ensuring robust claims about associations between the genomic properties of pathogens must be developed and leveraged. Problematic infectious disease control of has caused alarm among clinicians, epidemiologists and public health workers where the multidrug-resistant pathogen has significantly increased global morbidity and mortality rates. Tremendous variation at the genomic level which influences *S. aureus* resistance to every known antibiotic (Table 3.1, data adapted from Lewis (2013)). Thus, understanding the genetic contribution of variability in resistance and pathogenicity phenotypes could increase treatment efficacy, improve clinical outcomes, and highlight functional virulence mechanisms.

Table 3.2: Summary of studies that used difference references, with columns providing the citation, the number of strains studied, the source of the strains, and what reference was used.

Citation	PMID	Sequenced Genomes	RefGenome	Outcomes
Laabei, M(2014) [125]	24717264	Methicillin-resistant Saureus	TW20	A predictive model based on a set of significant single nucleotide polymorphisms (SNPs) and insertion and deletions events (indels) showed a high degree of accuracy in predicting an isolate's toxicity solely from the genetic signature at these sites.
Md Taqeer Alam (2014) [137]	PMC4040999	Vancomycin-intermediate *Staphylococcus aureus* (VISA)	Saureus N315	whole-genome comparison predicted VISA based on the presence of a rare mutationa in a set of candidate genes.
Holmes, A(2013) [138]	23927001	Methicillin-resistant *Staphylococcus aureus* (MRSA)	HO50960412	Single nucleotide polymorphism assay for epidemiological analysis of EMRSA-15 clinical isolates.
Hsu, L(2015) [106]	25903077	Methicillin-resistant *Staphylococcus aureus* (MRSA)	HO50960412 TW20	Competition between clones also has an important role in driving the evolution of nosocomial pathogen populations.
Claudio, K(2012) [134]	22693998	Methicillin-resistant *Staphylococcus aureus* (MRSA)	HO50960412	Revealed a distinct cluster of outbreak isolates and clear separation between these and the nonoutbreak isolates.
Weng, Z(2014) [139]	24969089	Drug-resistant hospital-associated methicillin-resistant Staphylococcusaureus (HA-MRSA)	TW20 T0131 JKD6008	The results suggest that ST239 strains isolated in Hong Kong since the 1990s belong to the Asian clade, present mainly in southern Asia, whereas those that emerged in northern China were of a distinct origin, reflecting the complexity of dissemination and the dynamic evolution of this ST239 lineage.

Table 3.3: Summary (averages across the strains) of variant calls metrics across two separate MRSA reference genomes (Table 3.2). Information compiled from mapped bams using the flagstat command.

Reference Genome	Number of SNVs	Percent Coverage	Average Depth Per Base	Average Base Quality	Percent Mapped Reads	Average Mapped Reads
H050960412 Reference	26080	93.48	175.78	36.60	84.62	5966295
TW20 Reference	16422	88.53	179.75	36.58	87.55	5980804

3.4 Methods

3.4.1 Sample Collection and Clinical Laboratory Processing

302 bacterial cultures collected from a routine workflow in a clinical microbiology laboratory. Specimens were collected from various body sites and cultured overnight. Then, resulting colonies were then inoculated on plates and then taken to a sequencing laboratory for WGS analyses. The clinical laboratory determined the bacterial species following Clinical Laboratory Standards Institute (CLSI) standards with the following guidelines: direct microscopic examination, gram staining, culture

on elective media, and additional biochemical assays. All culture colonies were further analyzed on a BD Phoenix^{TM} Automated Microbiology System to confirm the bacterial species determination and test for antimicrobial susceptibility.

3.4.2 DNA Sequencing

DNA was extracted from the colonies, and WGS was performed on an Illumina HiSeq 2500 sequencer in the rapid run mode. Single-end, 50-bp reads were generated in batches of up to 48 samples per flow cell, with read coverage ranging from 15-247x.

3.4.3 Sequence Data Processing and Variant Calling

All computer analyses were implemented on the Triton Shared Computer Cluster (TSCC) at the San Diego Super Computer Cluster (SDSC) in parallel on 8 or 16 cores within a node to minimize the run times. Our protocol utilized the 'Reddog' short-read length sequencing analysis pipeline (Version V1beta.10.3 070916, Calico Cat) to perform quality based read filtering, specific reference genome FASTA sequence read mapping, and variant calling. The RedDog program is a software package that executes a distributed computational workflow on a specified sample set and was chosen specifically for its streamlined analyses and pipeline reproducibility (https://github.com/katholt/RedDog)[140]. The package includes necessary functions for DNA variant calling including BWA read mapping, Genome Analysis Tool Kit (GATSK) variant detection, and downstream analyses (SNPs only)[141,142].

Additionally, complete genome sequences for 49 MRSA samples were downloaded from the NCBI. These samples were included to act as genetic control samples in our analyses to account for any systemic laboratory or computational biases in our sequencing pipeline.

3.4.4 Statistical Analyses

Variants called against the two reference genomes were utilized to formulate a genetic distance or dissimilarity matrices, which are defined in terms of nucleotide variant content shared across variant positions throughout the genome between each

pair of clinical isolate genomes. Matrices were then subjected to statistical tests to calculate the conditional effects of each particular reference genome, where both tests rely on the nucleotide variant distance matrix to measures the percent dissimilarity of each species when compared to each other species in the dataset (often referred to as a 'Genetic Relationship Matrix'[39]. The first is Mantel's statistical test for matrix equality, which determines the degree of correlation between two square matrices (n rows X n columns), and second being the Generalized Analysis of Molecular Variation (GAMOVA)[143,144] to measure the variance component of each clinical variable.

A genetic distance matrix compares the genetic dissimilarity of reads for each species mapped against the H050960412 reference genome, and then those same reads from each species mapped against the TW20 reference genome. If the MRSA references produced no result altering genetic effects, then each individual isolate would follow similar patterns in their placement relative to MRSA species included in the matrix. Thus, the mapped reference would have no effect on genetic composition. For the implementation of Mantel's test, 1000 Permutations were used to assess the probability that the dissimilarity of the two matrices occurred purely by chance. GAMOVA, implemented in the ADONIS function of the R language (Vegan Package 2.3-3), leveraged the dissimilarity matrix to calculate the proportion of variance explained by clinical variables, the framework allowed use to test the hypothesis that dissimilarity in variant profiles across the genomes of the MRSA isolates as a whole correlated with similarity in clinical outcomes. This analysis is thus complementary to the single locus or gene-specific analyses, since it explores the impact of the phylogenetic relationships between the isolates produced with the use of a specific reference and the clinical outcomes.

3.5 Results

3.5.1 Strain Similarities as a Function of the Reference Used

After mapping reads for each isolate to either MRSA reference genome and SNP variants were called, we assessed the number of variant sites called per clinical isolate in addition to comparing variant calling metrics to determine mapping quality

of each reference respectively. Variant calling metrics included base depth, percentage of the reference genome covered, number of mapped reads, etc. A summary of these analyses is provided in Table 3.3, where it can be seen that although many of the mapping statistics are similar, a greater number of variants were called with confidence using H050960412 reference. This increase is mostly likely a function of the genetic distance of this particular reference genome in relation to the cohort of clinical isolates, i.e. phylogenetic relationship which was reflected in the outcome of Mantel's test. Phylogenetic divergence often comes with (or assumes) differences at the genomic level, such that mapping reads on to the reference would inevitably expose differences in the nucleotide content of the target and reference sequences; i.e., the more phylogenetic divergence between the target and reference, the more nucleotide differences. Annotating the two references for gene content lead to 736 genes being identified with the H050960412 reference genome and 885 genes being identified on the TW20 reference genome. The two references had 722 genes in common (80.3%). Probably due to the greater number of variants found in the H050960412 reference genome, there were more variants found in these 722 genes in the H050960412 reference genome (Figure 3.1; Supplementary Figure 3.4.

Figure 3.2 provides dendrograms that reflect the genetic similarity and clustering of the MRSA clinical isolate genomes based on whether the H050960412 reference was used to identify genomic variants from each isolate in contrast to whether the TW20 reference was used. The figure demonstrates that the clustering is not identical when the different references are used. Another graphical device, a heatmap, was utilized to display the GRMs and they also exhibit a difference in isolate representation (Figure 3.3). Principal components analyses also suggested that clinical isolate genomes clustered with greater more densely when the TW20 reference genome was used as opposed to the H050960412 reference genome (Supplementary Figure 3.5). A Mantel's matrix equality test of the GRMs utilized in the dendogram analysis was correlated at .508 with a permutation p-value of .001 (correlation = 0.5077231, at p(x) = .001). This result suggests that the use of the two references generates different relationships between the MRSA clinical isolate genomes.

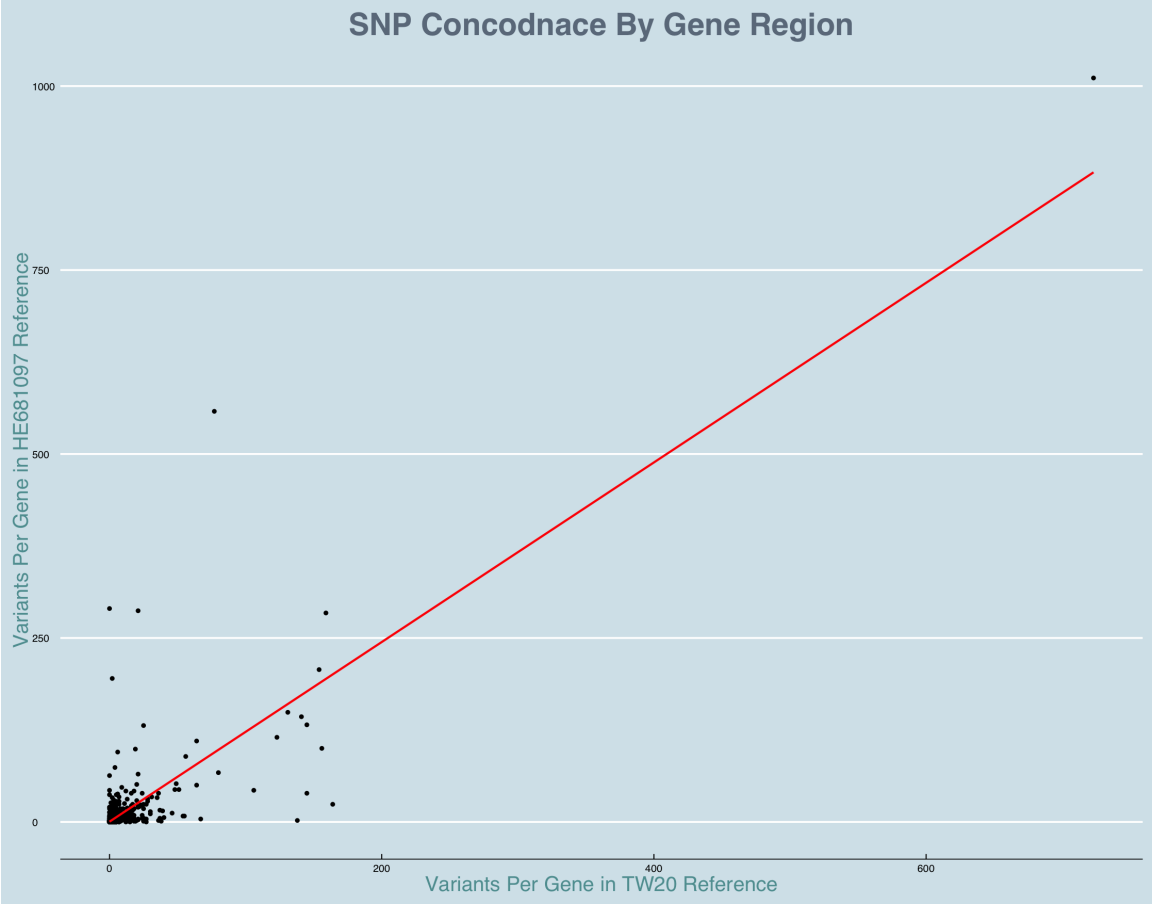

Figure 3.1: Gene Region SNV Frequency Correlation By MRSA Reference Genome

3.5.2 Clinical Associations with Clinical Isolate Genomes

We next utilized the GAMOVA statistical test to determine if the use of different reference genomes affected the relationships between isolate genome similarity, clinical outcomes, and individual isolate drug resistance profiles. Table 3.4 and 3.5 presents the results of the analyses involving the clinical outcomes and Table 3.6 and 3.7 presents the results of the analyses involving the drug resistance phenotypes. These tables suggest that the association strength and their statistical significance differs somewhat depending on what reference is used, but the overall trends are the same: whether the strain resulted in a pathogenic outcome, the age at which the infection was diagnosed, the antibiotics chosen to treat the infection and the diagnosis at admission are all associated with isolate genome similarities. However, the diagnosis at admission was only significantly associated if the less polymorphic TW20

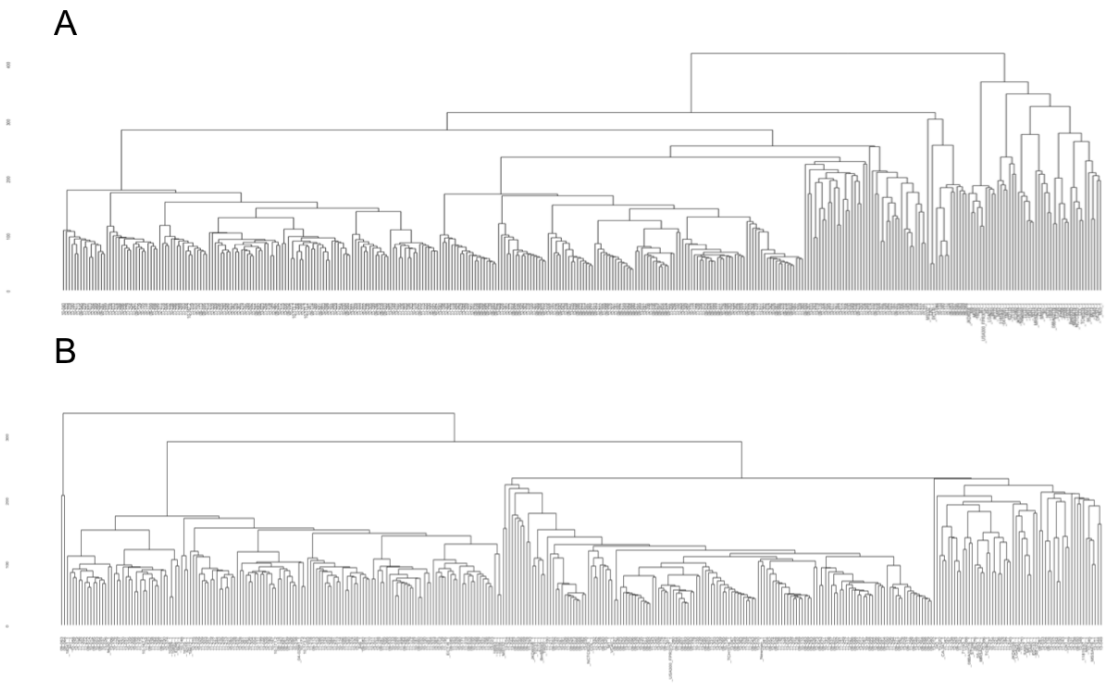

Figure 3.2: Genetic Variant Based Sample Phylogeny By Mapped Reference. SNP variants were utilized to create a genetic distance based dendogram tree. This tree provides the clustering of each clinical isolate sample for each of the two reference genomes mapped against.

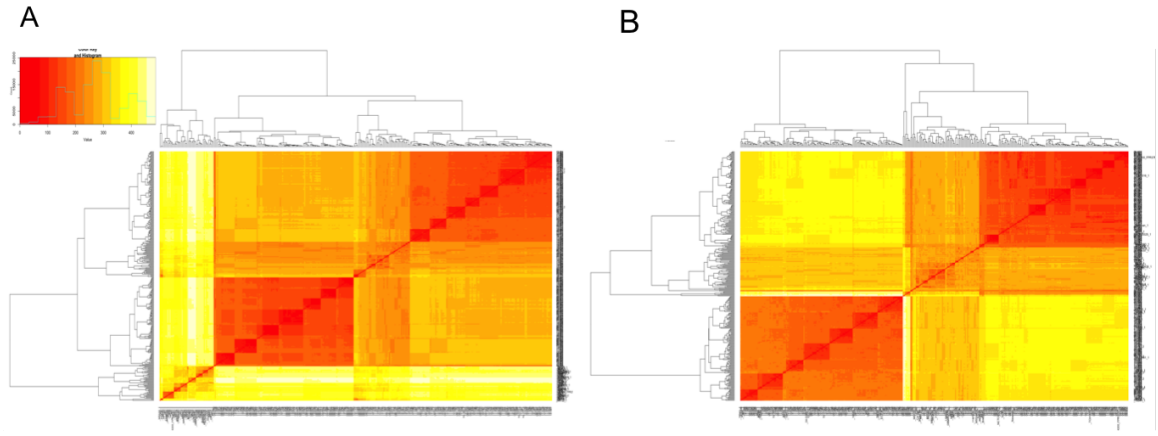

Figure 3.3: Genetic Variant Heatmap Visualization For Species Distance. Clustering of clinical strains when using two different references.

genome was used as a reference.

An analyses involving the drug resistance profiles of each isolate suggest greater concordance between the results, as the resistance profiles associated with antibiotics

penicillin, clindamycin, erythromycin and oxacillin were found to be associated with isolate genome similarity, but to different degrees depending on the reference used. This was the case despite the fact the isolates exhibited great variation in the drug resistance profiles (Supplementary Figure 3.6).

Table 3.4: ANOVA statistics for for variance explained for clinical covariates in clinical isolates acquired from Scripps Green hospital mapped against the H050960412 reference genome.

	Df	SumsOfSqs	MeanSqs	F.Model	R2	Pr(>F)	Significance
Pathogenic	1	104434	104434	4.6335	0.04019	0.003996	**
Age of Infection	1	179205	179205	7.9509	0.06897	0.000999	***
Antibiotic Duration	1	16172	16172	0.7175	0.00622	0.621379	
Hospital Time (Days)	1	20353	20353	0.903	0.00783	0.508492	
Vancomycin MIC	1	33526	33526	1.4875	0.0129	0.148851	
Antibiotic Chosen	33	990018	30001	1.3311	0.38101	0.021978	*
Admission Diagnosis	29	758810	26166	1.1609	0.29203	0.151848	
Residuals	22	495856	22539	0.19083			
Total	89	2598374	1				

Table 3.5: ANOVA statistics for for variance explained for clinical covariates in clinical isolates acquired from Scripps Green hospital mapped against the TW20 reference genome.

	Df	SumsOfSqs	MeanSqs	F.Model	R2	Pr(>F)	Significance
Pathogenic	1	99089	99089	5.0068	0.03986	0.0037	**
Age of Infection	1	182876	182876	9.2404	0.07356	0.0002	***
Antibiotic Duration	1	16084	16084	0.8127	0.00647	0.464554	
Hospital Time (Days)	1	18768	18768	0.9483	0.00755	0.460954	
Vancomycin MIC	1	25607	25607	1.2939	0.0103	0.222778	
Antibiotic Chosen	33	936206	28370	1.4335	0.37658	0.007999	**
Admission Diagnosis	29	772023	26621	1.3451	0.31054	0.043296	*
Residuals	22	435398	19791	0.17514			
Total	89	2486050	1				

Table 3.6: The ANOVA outcome statistics for variance explained when the antibiogram drug response data was compared to the SNP dissimilarity matrix derived from variant calls against the H050960412 reference genome.

	Df	SumsOfSqs	MeanSqs	F.Model	R2	Pr(>F)	
Penicillin	2	101126	50563	2.2743	0.01561	0.012987	*
Clindamycin	2	1329223	664612	29.8936	0.2052	0.000999	***
Erythromycin	2	125353	62677	2.8191	0.01935	0.008991	**
Oxacillin	2	120651	60326	2.7134	0.01863	0.004995	**
Tetracycline	2	26501	13250	0.596	0.00409	0.942058	
Trimethoprim	2	29388	14694	0.6609	0.00454	0.777223	
Rifampin	2	40461	20230	0.9099	0.00625	0.495504	
Levofloxacin	2	58410	29205	1.3136	0.00902	0.171828	
Residuals	209	4646613	22233	0.71732			

3.6 Discussion

Pathogenic infections are increasingly becoming a public health issue, given the ability of pathogens to evolve quickly and develop antibiotic treatment resistances.

Table 3.7: The ANOVA outcome statistics for variance explained for variance explained when the antibiogram drug response data was compared to the SNP dissimilarity matrix derived from variant calls against the TW20 reference genome.

	Df	SumsOfSqs	MeanSqs	F.Model	R2	Pr(>F)	
Penicillin	2	95759	47879	2.384	0.01505	0.030969	*
Clindamycin	2	1649395	824697	41.061	0.25917	0.000999	***
Erythromycin	2	124387	62193	3.097	0.01954	0.008991	**
Oxacillin	2	113804	56902	2.833	0.01788	0.011988	*
Tetracycline	2	28451	14226	0.708	0.00447	0.727273	
Trimethoprim	2	30881	15441	0.769	0.00485	0.583417	
Rifampin	2	47602	23801	1.185	0.00748	0.320679	
Levofloxacin	2	76198	38099	1.897	0.01197	0.100899	
Residuals	209	4197668	20085	0.65958			
Total	225	6364145	1				

Identifying specific pathogen strains, genes and genetic variants that contribute to communicable nature of a pathogen, its virulence and treatment resistances are therefore of tremendous importance. Unfortunately, genetically-mediated resistance mechanisms exploited by pathogens are complex and hard to decipher. This challenge is further complicated by the computational and statistical manners in which genetic variations are identified, catalogued and tested. The traditional method of identifying variants in pathogen genomes, which involves mapping DNA sequence obtained from target pathogen genomes onto a chosen reference genome, can produce undesired sources of variation if the reference genome mapped against differs in terms of gene content, structural variations, overall organization and specific nucleotide content. Our study demonstrates the consequences of reference genome choice in relating genomic variation to clinical outcomes and drug resistance profiles.

We find obvious and expected differences in two reference genomes we chose for study, but also differences in the outcomes of association studies involving the clinical isolates when variants are identified with each of two references. Although not pronounced, our study suggests SNVs identified from a particular reference genome can impact an ability to identify associations between clinical isolate genomes and clinical outcomes, however this effect was not drastic. Differential reference use to formulate genetic distance matrices as a measure of similarity can effect the phylogenetic clustering of a population of bacterial isolates, which can be important in investigations attempting to compare variant mutations from vastly different bacterial species. We also find that variant call frequencies in annotated gene regions have a strong correlation, which indicates a general uniformity between reference genomes for MRSA.

3.6.1 Study Limitations

We recognize that our study has obvious limits. A more comprehensive study would have explored the use of many different potential reference genomes. We were constrained by what was available in the public domain. In addition, we did not explore the utility of, e.g., multiple sequence alignments, and other analyses to identify points of divergence between the references that could have been overcome by perhaps combining their use. We also did not consider structural variant analyses or copy number variant analyses. Finally, our analyses were also constrained by the relatively small number of patients for which we had clinical outcome and phenotype data and the fact that we did not explicitly consider host-genome and other host-related factors that might interact with pathogen-related factors to affect clinical outcomes[145].

3.6.2 Conclusions

Despite these limitations, our results do suggest that care should be taken when interpreting pathogen genome background and clinical outcome associations. There may be, however, ways to mitigate the problems associated with a choice of a reference genome. First, one could simply look at the sensitivity of the results when using different references, more or less as we did. Second, one could try to combine genome information from different references into a unique read mapping framework, much in the way strategies like those implemented in the program GenomeMapper[146]. Third, one could do away with reference mapping altogether and rely on assembly each genome of interest *de novo*, although the problem of reconciling the various positions in the target genomes subjected to *de novo* assembly might be problematic despite the high quality of multiple sequence alignment tools[93,147].

3.7 Acknowledgements

Chapter 3, in full is currently being prepared for submission for publication, Quarless Q., Liu X.*, Pfeiffer W.*, Lee J., Oliveira G., and Schork N. "Implications of Methicillin-Resistant Staphylococcus Aureus (MRSA) Reference Genome Choice for Investigating Clinical Correlates". The dissertation author is the primary researcher

and author on this paper.

3.8 Supplementary Figures

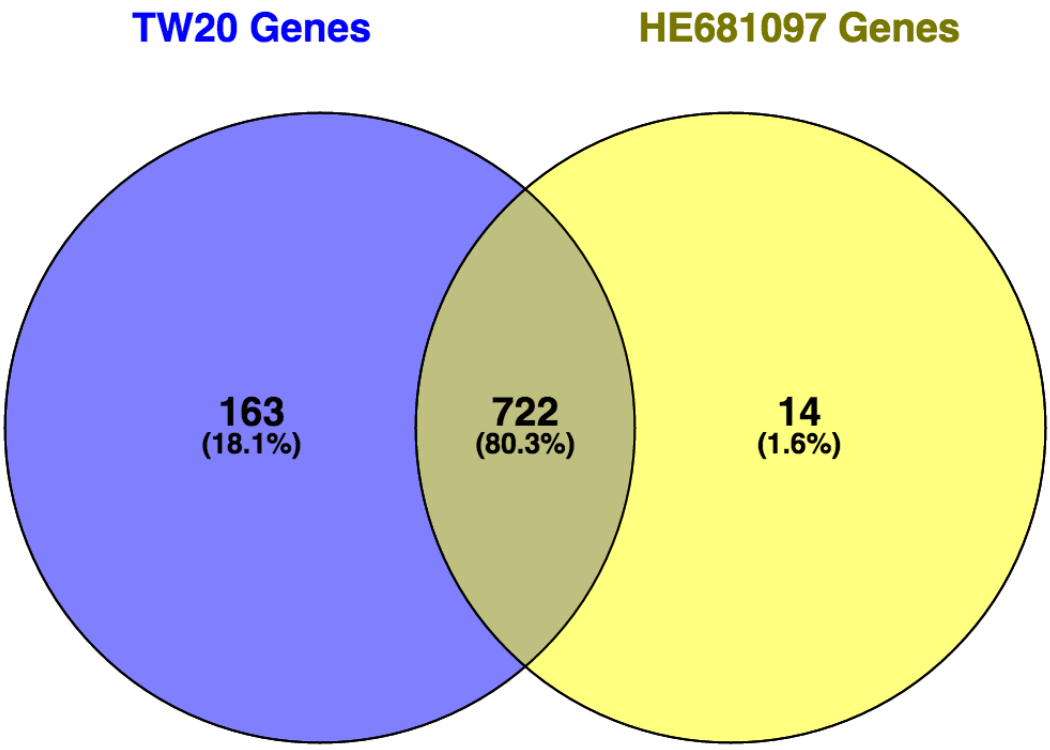

Figure 3.4: Venn Diagram of Gene Regions Shared By Each MRSA Reference Genome. Annotations were downloaded the NCBI gene tracks.

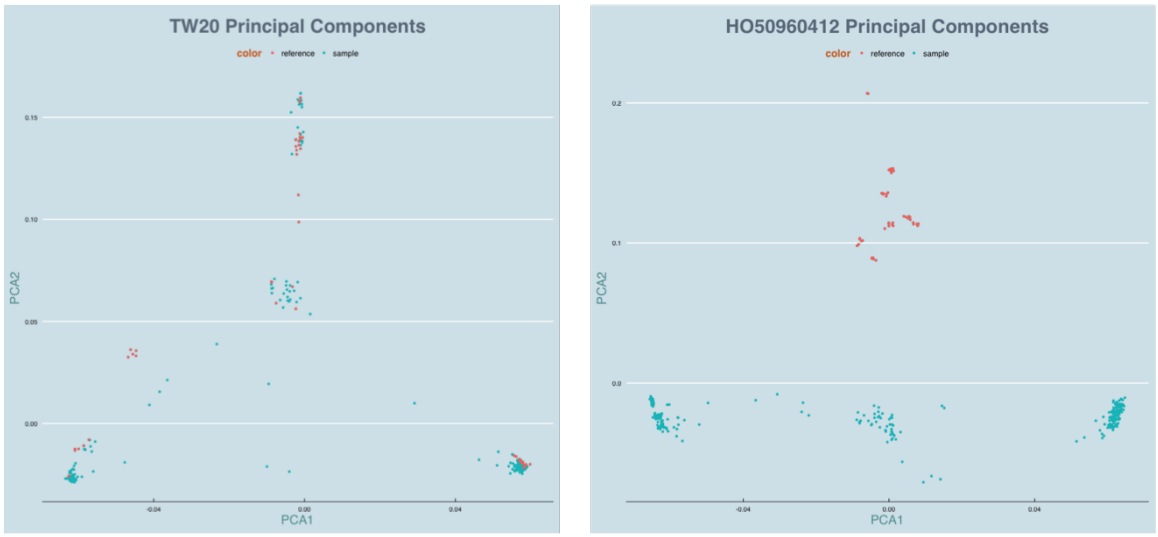

Figure 3.5: Principal component analysis of samples by MRSA reference genomes

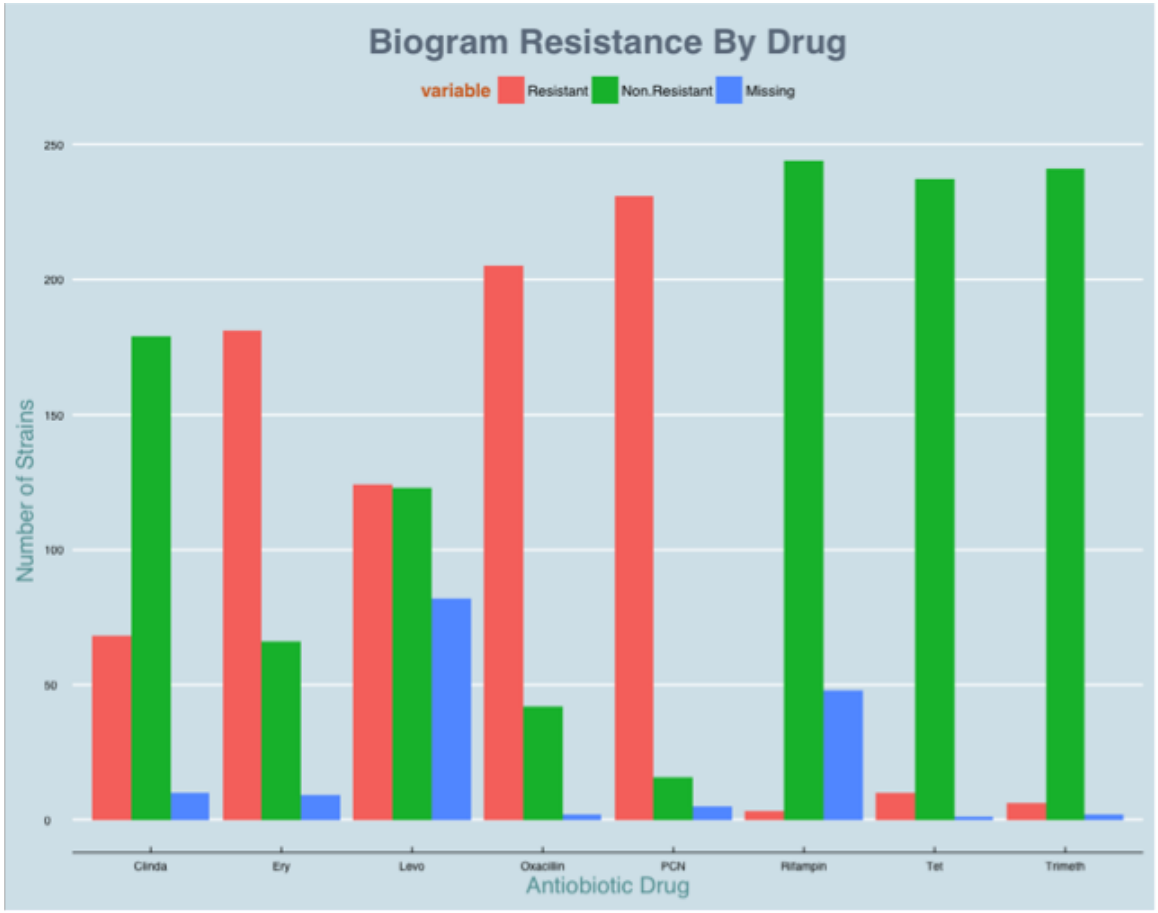

Figure 3.6: Clinical Isolate Antibiogram Drug Response Data

Chapter 4

Comprehensive Gene Expression-Based Mediator-Wide Association Study of Alzheimer's Disease

4.1 Abstract

Alzheimer's Disease (AD) is a complex, multifactorial condition that plagues approximately 5.5 million Americans who collectively require hundreds of billions of dollars' worth of health care. Identifying the factors contributing to AD is not trivial given its complexities. Mediator-wide Association Studies (mWAS), which search for molecular and subclinical phenotypes that mediate the relationship between a genetic variant and a disease, have the potential to shed light on factors contributing to AD that can leverage genetic data from existing studies. We pursued a large-scale mWAS for AD that exploited: 1. three very large cohorts (International Genomics of Alzheimer's Project (IGAP), the Alzheimer's Disease Genetics Consortium 1 (ADGC1) and ADGC2 cohorts) that collectively totaled 11576 to cases and 10796 controls (ADGC1: 9549 cases and 7683 control, ADGC2: 2027 cases and 3133 control); 2. Comprehensive gene expression and genetic data within the Genotype-Tissue Expression (GTEx) database obtained on 44 different human tissues in hundreds of

individuals; and 3. Proven analytical methods for conducting mWAS that also considered potential heterogeneity in the relationships between the genetic variants, gene expression levels and AD across and within the cohorts. We ultimately found evidence for a few weak associations, but little consistency in mWAS results between the cohorts beyond tests involving the APOE gene, despite minimal statistical evidence for heterogeneity in effects sizes among the individual subcohorts that make up the total ADGC1 and ADGC2 cohorts. We did see compelling and replicated evidence suggesting that expression levels in the APOE gene are associated with AD and likely mediate the relationship between APOE genetic variants and AD susceptibility.

4.2 Introduction

Alzheimer's Disease (AD) is a chronic neuropsychiatric disorder caused by amyloid-plaque and neurofibrillary tangle build-up in the cortical and hippocampal regions in the brain[148–150]. AD is the most common cause of dementia in many countries, as roughly 5.5 million individuals are affected in the United States alone and 44 million are affected worldwide, and the costs associated with caring for people with AD amounts to hundreds of billions of dollars and continues to rise[151,152]. Currently, no reliable and proven therapeutic interventions exist to alter the disease progression of AD. Thus, two critical challenges for the biomedical community are to identify factors that contribute to AD susceptibility and progression, and to identify ways of mitigating the effects of those factors to prevent and treat AD effectively[153].

Unfortunately, AD is a complex, multifactorial disease with many genetic and non-genetic influences whose individual contributions and interactions have been difficult to sort out[150,152,154–156]. Genome wide association studies (GWAS) have been pursued to identify genetic variants that might impact AD and, outside of the well-known Apolipoprotein ϵ (APOE) gene, few variants of strong effect have been identified and replicated[70,156–159]. This is unfortunate, since the identification of genetic variants contributing to AD could lead to insights into the molecular processes that contribute to AD pathogenesis and reveal points for pharmacotherapeutic intervention. However, recent extensions of GWAS that consider the role that an intermediate phenotype (IP) might play in a causal pathophysiologic chain leading from a genetic

variant to AD clinical manifestations have promise[160]. Mediator-wide association studies (MWAS) leverage GWAS data along with sources of information that provide knowledge of relationships between genetic variants and IPs to systematically test IPs for potential causal associations with a disease of interest, such as AD, that could be considered drug targets15,16. Unfortunately, the choice of IPs to study and their measurement on thousands of individuals can be difficult. As a result, aspects of the implementation of MWAS initiatives can be impractical and challenging[161,162]. Unfortunately, the choice of IPs to study and their measurement on thousands of individuals can be difficult. As a result, aspects of the implementation of MWAS initiatives can be impractical and challenging.

Clever statistical analysis methods have been developed to overcome the need to directly measure an IP on individuals for an MWAS. Consider the fact that it is possible to leverage knowledge of associations between genetic variants and IPs to build predictive models of those IPs that can then be used to impute or assign predicted IP values to individuals with genetic variants typed on them as part of a GWAS. In this way, the predicted IP values can then be tested for association with a disease of interest with the GWAS data. it has been further shown that an MWAS can be conducted using only summary statistic information resulting from a GWAS[163–165]. If done properly under many testable assumptions, the principles behind Mendelian Randomization (MR) tests can be invoked and thereby lead to statistical hypothesis tests in an MWAS setting that can be used to investigate causal links between the genetic variants, an IP, and a disease of interest[164,166,167].

We pursued an MWAS of AD using state-of-the-field epidemiological data from multiple cohorts, gene expression-based IP databases and state-of-the-field statistical methods for relating the IPs, SNPs and AD diagnosis. We obtained GWAS summary statistic data from the International Genomics of Alzheimer's Project (IGAP) as well as raw GWAS genotype and phenotype data from the Alzheimer's Disease Genetics Consortium 1 (ADGC1) and ADGC2 cohorts[156,158]. Together the ADGC cohorts totaled 11,272 to cases and 10,419 controls (ADGC1: 9549 cases and 7683 control, ADGC2: 1723 cases and 2736 control). We also used information in the GTEx database about relationships between SNPs and the expression levels of genes in 44 human tissues from 449 donors[168]. Finally, we leveraged state-of-the-field statistical

analysis methods, many associated with MR Base, a resource for Mendelian Randomization and MWAS tools, to carry out relevant analyses integrating the AD GWAS and gene expression data[164]. These statistical analysis methods included methods for assessing evidence for heterogeneity in the strength of the MWAS results within and across the AD cohorts.

Our results provide evidence that suggests that APOE gene expression mediates the relationship between APOE genetic variants and AD. We found evidence that other genes may have expression levels that influence AD susceptibility that are affected by genetic variants, but the evidence for these other genes was not as compelling as for the APOE gene. We did not find evidence for heterogeneity in the effects of SNPs and gene expression levels on AD within or across the cohorts. We ultimately believe that analyses like ours can shed light on potentially modifiable factors contributing to AD, but must be approached with a sensitivity to their limitations.

4.3 Methods

4.3.1 ADGC Data Processing

ADGC Data Processing. We obtained and analyzed genotype and phenotype data from the ADGC, which has been previously described21. The ADGC is made up of 26 individual cohorts divided into an initial set of cohorts and data (ADGC1; 13 cohorts) and a replication or follow-up set of cohorts and data (ADGC2; 13 cohorts). For each individual cohort within the broader collection of cohorts in ADGC1 and ADGC2, we generated imputed genotype input files from the available genotype data for GWAS and MWAS analyses. Essentially, we leveraged the Plink bioinformatic suite (Version 1.90b3v) to convert ADGC IMPUTE2/SNPTEST Oxford-format files containing genotype imputation probabilities available for hard genotype calls in binary Plink format. We used Plink for data processing and applied the default uncertainty threshold for imputing variants to remove any imputation calls with certainty less than 0.1. We also applied the following additional Plink filtration parameters to remove genotypes that were problematic from the imputation protocol: variants above

the maximum missing genotype rate of 0.02, variants violating the Hardy-Weinberg equilibrium threshold of 0.000001, variants with minor allele frequency values below the threshold of 1 percent, and highly correlated genotype calls arising from strong linkage disequilibrium. Individual samples missing more than 10% of genotype calls were also removed. We also removed individuals with missing AD phenotype status, sex status, or APOE $\epsilon 4$ allele designation.

4.3.2 Summary Data from the International Genomics of Alzheimer's Project (IGAP)

We obtained publicly available summary statistics data from the published meta-analysis of the AD GWAS published by IGAP investigators[156]. Details about the study participants' ascertainment, data quality control, and association analyses are described in the study9. We ultimately used the stage 1 IGAP summary statistics dataset, which consists of 7,055,881 single nucleotide polymorphisms (SNPs) from 17,008 Alzheimer's disease cases and 37,154 controlsÂăof European ancestry. For each SNP, we leveraged the following information in our analyses from the available IGAP resources: Chromosome of the SNP (Build 37, Assembly Hg19), Position of the SNP (Build 37, Assembly Hg19), SNP rsID or chromosome position if the rsID was not available, reference allele (i.e., coded allele), Non reference allele (i.e., non-coded allele), overall estimated effect size for the effect allele (i.e., the Beta weight provided), overall standard error for effect size estimate (denoted as the SE in the dataset), and the meta-analysis GWAS p-value for the stage 1 IGAP analyses.

4.3.3 GTEx Genotype Expression Data

As noted, we considered gene expression levels as potential IPs for AD in our analyses. We obtained information about SNP-gene expression level relationships from the Genotype-Tissue Expression Project (GTEx) publicly available resource[168]. The GTEx resource provides expression quantitative trait loci (eQTLs) based on correlation analysis between genotype and tissue-specific RNA expression on 44 different tissues obtained on 449 individuals with genotype information. We focused on primarily cis-acting eQTLS. Details about the expression QTLs (eQTLs) analysis can be

found in multiple publications by the GTEx Consortium[168]. For each SNP identified as an eQTL in the GTEx database, we obtained the following information: SNP rsID, Chromosome of the SNP (Build 37, Assembly Hg19), Position of the SNP (Build 37, Assembly Hg19), Reference allele (coded allele), Non reference allele (non-coded allele), the gene whose expression levels were associated with a SNP in given tissue (IP), the overall estimated effect size for the associated allele (Beta), the overall standard error for effect size estimate (SE), and the p-value reflected the statistical significance of the association. Based on this, we extracted 187,263 independent cis-eQTL SNPs for 27,094 genes across the 44 tissues.

The intersection between the independent cis-eQTL SNPs from the GTEx consortium and the IGAP AD GWAS was found by matching rsID. A total of 77,854 SNPs were found in common, and used as instruments in the MR analysis. The exposure and outcome datasets were harmonized to ensure each SNP corresponded to the same effect allele. We tested the causal effect of cis-eQTLs for 22,878 genes across 44 tissues on AD using the analysis tools described below.

4.3.4 Mendelian Randomization Analysis

For each eQTL from GTEx for which the associated tissue specific expression data were available to build a predictive model of the expression level, and for which the associated SNP was also genotyped or imputed in the IGAP and ADGC cohorts, we pursued Mendelian Randomization (MR) tests. MR tests, as noted, test the hypothesis that the IP of interest âĂŞ in this case the expression level of a gene âĂŞ is causally associated with the phenotype (AD) based on its association with a SNP and that SNP's relationship to AD. MR tests have their roots in instrumental variables analyses[169]. Ultimately, in the context of instrumental variables analyses, the genetic variants (SNPs) associated with an IP are used as instrumental variables, the IP is considered the "exposure" variable, and the disease is the outcome variable. To conduct MR tests we used, in part, two analytical techniques and associated software available from the MR-Base resource: the two sample MR test and associated R package âĂIJTwoSampleMRâĂİ and the âĂIJMRInstrumentsâĂİ analysis suite developed by Hemani et al. The MR estimate of the causal effect between the

expression levels and AD was calculated as the Wald ratio as discussed by Hemani et al. We considered the Fixed effects meta-analysis method (delta method for standard errors) also available from MR-Base. To control for multiple comparisons we used a Bonferroni correction based on the number individual mediation tests we performed. We note that this is conservative given that many of the tests are correlated due to linkage disequilibrium and relationships between expression values of genes within and across tissues.

4.3.5 ADGC Cohort-Based Meta-Analyses

To assess the robustness of GWAS association metrics from each individual ADGC1 and ADGC2 subcohort we analyzed using the Metasoft meta-analysis package (version v2.0.0)[169]. Briefly, plink case/control association statistics for each individual ADGC dataset, ADGC1 and ADGC2 respectively, were converted to the standard Metasoft input format. The resulting files were then analyzed in the Metasoft using default run parameters, with filtering out of GWAS SNVs that reported either a standard error of 0 or a plink unadjusted p-value of "INF". To determine the robustness of mWAS results, the metap (Version 0.8) was utilized in the R language to perform meta-weight analyses of the cohort based results for ADGC1 and ADGC2. To assess the meta-significance of MWAS results, beta scores and standard error values from MRbase outputs for each individual cohort were loaded into metasoft. This analysis only consisted of fixed effect MR tests, which included 130,659 MWAS tests across 90,383 unique RSID identifiers. Metasoft excludes entries where only 2 or less cohorts have entries, which left 130,641 resulting entries.

4.3.6 Heterogeneity Analysis

To explore evidence that the strength of the association in the GWAS and MWAS analyses varied across ADGC cohorts, we used Cochran's Q test as implemented in the software package, derived from the Metasoft package[169].

4.4 Results

4.4.1 GWAS Analyses

Figure 4.1

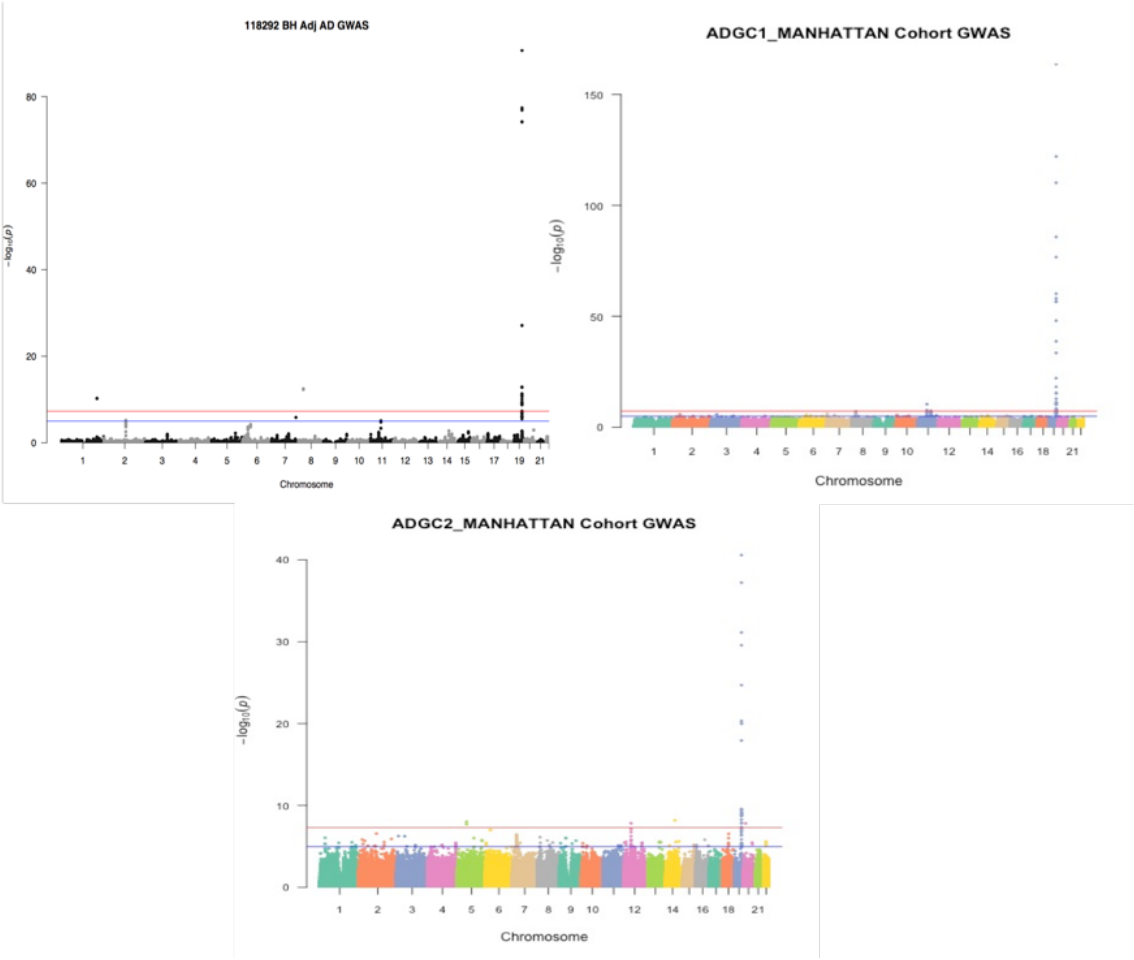

Figure 4.1: Manhattan plots for IGAP, ADGC1, and ADGC2 GWAS results. ADGC1 and ADGC2 results were calculated by the metasoft meta analysis software, where each dataset contains the corresponding individual cohorts.

Figure 4.1 provides the Manhattan plots associated with the GWAS results for the three cohorts we analyzed. Note that IGAP GWAS results simply reflect the summary statistics obtained from IGAP. The ADGC1 and ADGC2 plots reflect meta-analysis p-values based on re-analysis of GWAS for each component cohort within the ADGC1 and ADGC2 cohorts. Tables 4.1, 4.2, and 4.3 provide a list of the top hits of the GWAS across the ADGC and IGAP cohorts, with support

Figure 4.2

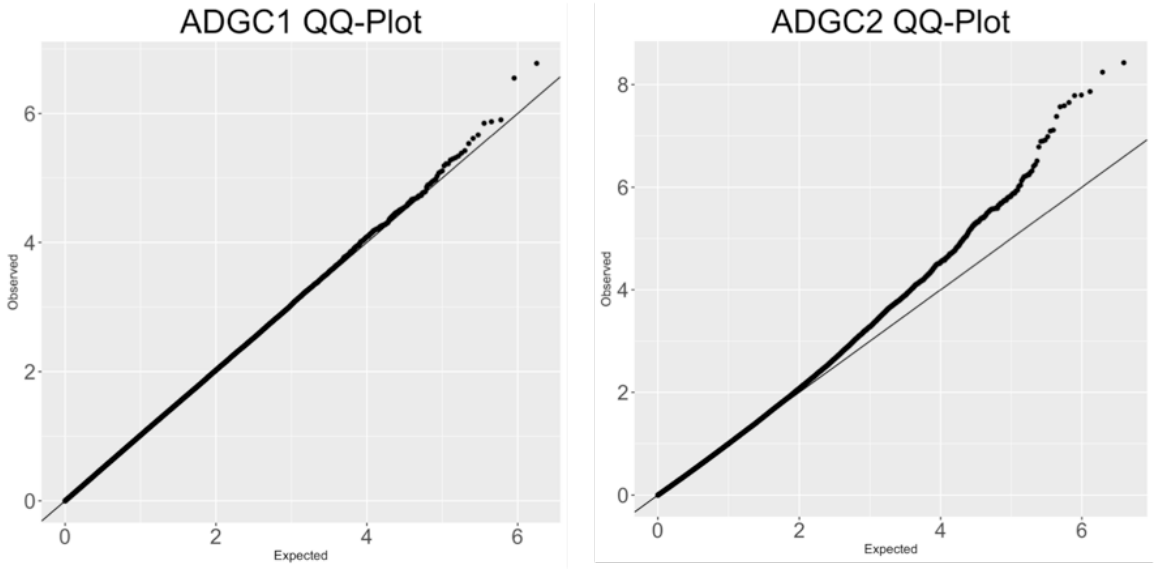

Figure 4.2: ADGC1 and ADGC2 Heterogeneity P-value QQ-Plots. Plots contain the plotted p-value statistic for the Cochran statistic (Q) calculated by metasoft. Although the ADGC2 QQ-plot appears to demonstrate inflation, bonferroni adjustment of the p-values returns zero SNVs which remain significant. The same is true for ADGC1. This indicates that there was no observable heterogeneity effect in either cohort from the ADGC data, which could have downstream effects on Mrbase analyses.

for the associations for each of the cohorts separately. Additionally, we tested each SNP for heterogeneity of association strength with AD among the different cohorts across the individual ADGC1 and ADGC2 subgroups for GWAS results. Figure 4.2 provides q:q plots displaying the observed p-values against expected p-values (i.e., if no heterogeneity exists among the ADGC cohorts). It can be seen that little evidence for heterogeneity exists in SNP association strength, as no GWAS entry surpassed the Bonferroni significance threshold, which gives us confidence the MWAS analyses are not likely to suffer from heterogeneity as well.

Gene Expression-Based mWAS Analyses. We conducted MWAS analyses for the IGAP cohort and each individual ADGC cohorts for all SNPs associated with gene expression values in the different tissues available from the GTEx database, but limited to only genes whose expression levels could be imputed or assigned to individuals in the three GWAS cohorts based on the availability of relevant SNP information. This resulted in 118,292 tests for IGAP, 353,756 tests across 53,665

Table 4.1: Top GWAS meta-analysis results derived from the metasoft program for ADGC1.

ADGC1 Top 20 Meta-Analysis GWAS Results									
RSID	Chromsome	Position	Num Meta-analyzed Studies	PVALUE FE (Metasoft)	BETA FE (Metasoft)	STD FE (Metasoft)	I_SQUARE (Metasoft)	Q (Metasoft)	PVALUE Q (Metasoft)
rs2075650	19	45395619	7	1.56E-164	1.01075	0.0369741	81.2219	31.9521	1.67E-05
rs4420638	19	45422946	4	8.39E-123	1.12172	0.0475967	90.9811	33.2636	2.83E-07
rs429358	19	45411941	2	5.66E-111	1.4296	0.0638677	0	0.5986	4.39E-01
rs769449	19	45410002	2	1.57E-86	1.37085	0.0695289	0	2.96E-05	9.96E-01
rs157580	19	45395266	10	1.81E-77	-0.50507	0.0271093	56.1231	20.5119	1.50E-02
rs71352238	19	45394336	2	6.22E-61	1.0493	0.0637172	0	0.937504	3.33E-01
rs405509	19	45408836	10	7.65E-59	-0.41114	0.0254193	47.2925	17.0754	4.75E-02
rs439401	19	45414451	9	1.94E-57	-0.442135	0.0276787	27.7828	11.0777	1.97E-01
rs157581	19	45395714	2	2.62E-57	0.877809	0.0550169	0	0.402422	5.26E-01
rs6859	19	45382034	10	8.41E-49	0.373075	0.0254104	60.2141	22.6211	7.11E-03
rs34878901	19	45402477	5	1.67E-39	-0.453861	0.0345101	0	2.07306	7.22E-01
rs41290120	19	45382675	8	2.72E-34	-0.988403	0.0809445	63.0659	18.9527	8.34E-03
rs10402271	19	45329214	8	7.35E-23	0.281806	0.0286301	29.993	9.99901	1.89E-01
rs7412	19	45412079	2	6.10E-19	-0.924255	0.103963	0	0.747576	3.87E-01
rs1160985	19	45403412	3	3.00E-16	-0.416806	0.0509963	86.1594	14.4502	7.28E-04
rs2965101	19	45237812	9	5.11E-16	-0.225402	0.0277967	42.7708	13.9789	8.23E-02
rs8106922	19	45401666	2	1.18E-13	-0.423675	0.0571028	3.90855	1.04068	3.08E-01
rs412776	19	45379516	5	4.60E-12	0.342644	0.0495346	0	2.86085	5.81E-01
rs8103315	19	45254168	7	1.97E-11	0.294913	0.0439642	56.0497	13.6518	3.38E-02
rs1114832	19	45636201	11	2.81E-11	0.260182	0.0390889	46.6563	18.7464	4.36E-02

unique entries for ADGC1, and 838,716 tests across 87,182 unique entries for ADGC2. Meta-analysis was used to combine the MWAS results from the individual ADGC1 and ADGC2 cohorts as with the GWAS data. Supplementary Figure 1 provides q:q plots for the GWAS heterogeneity tests for the ADGC1 and ADGC2 subcohorts to see if the might be reason to believe the MWAS results were likely to show variation because of differences in the SNP association strengths across the cohorts. This suggested that although there appears to be a slightl levels of inflation. Tables 4.4 and 4.5 provide a list of the most significant MWAS associated genes and gives the SNP, the gene whose expression level is associated with that SNP, and the p-value resulting from the MR tests for each of the two datasets (IGAP, ADGC). These tables again suggest that outside the APOE gene and top associated SNVs on chromosome 19, i.e. PVRL2, Bin1, and CEACAM19, etc, little consistency exists between the MR test results even though within some of the cohorts strong MR associations exist.

Table 4.2: Top GWAS meta-analysis results derived from the metasoft program for ADGC2.

ADGC1 Top 20 Meta-Analysis GWAS Results									
RSID	Chromsome	Position	Num Meta-analyzed Studies	PVALUE FE (Metasoft)	BETA FE (Metasoft)	STD FE (Metasoft)	I_SQUARE (Metasoft)	Q (Metasoft)	PVALUE Q (Metasoft)
rs2075650	19	45395619	7	1.56E-164	1.01075	0.0369741	81.2219	31.9521	1.67E-05
rs4420638	19	45422946	4	8.39E-123	1.12172	0.0475967	90.9811	33.2636	2.83E-07
rs429358	19	45411941	2	5.66E-111	1.4296	0.0638677	0	0.5986	4.39E-01
rs769449	19	45410002	2	1.57E-86	1.37085	0.0695289	0	2.96E-05	9.96E-01
rs157580	19	45395266	10	1.81E-77	-0.50507	0.0271093	56.1231	20.5119	1.50E-02
rs71352238	19	45394336	2	6.22E-61	1.0493	0.0637172	0	0.937504	3.33E-01
rs405509	19	45408836	10	7.65E-59	-0.41114	0.0254193	47.2925	17.0754	4.75E-02
rs439401	19	45414451	9	1.94E-57	-0.442135	0.0276787	27.7828	11.0777	1.97E-01
rs157581	19	45395714	2	2.62E-57	0.877809	0.0550169	0	0.402422	5.26E-01
rs6859	19	45382034	10	8.41E-49	0.373075	0.0254104	60.2141	22.6211	7.11E-03
rs34878901	19	45402477	5	1.67E-39	-0.453861	0.0345101	0	2.07306	7.22E-01
rs41290120	19	45382675	8	2.72E-34	-0.988403	0.0809445	63.0659	18.9527	8.34E-03
rs10402271	19	45329214	8	7.35E-23	0.281806	0.0286301	29.993	9.99901	1.89E-01
rs7412	19	45412079	2	6.10E-19	-0.924255	0.103963	0	0.747576	3.87E-01
rs1160985	19	45403412	3	3.00E-16	-0.416806	0.0509963	86.1594	14.4502	7.28E-04
rs2965101	19	45237812	9	5.11E-16	-0.225402	0.0277967	42.7708	13.9789	8.23E-02
rs8106922	19	45401666	2	1.18E-13	-0.423675	0.0571028	3.90855	1.04068	3.08E-01
rs412776	19	45379516	5	4.60E-12	0.342644	0.0495346	0	2.86085	5.81E-01
rs8103315	19	45254168	7	1.97E-11	0.294913	0.0439642	56.0497	13.6518	3.38E-02
rs1114832	19	45636201	11	2.81E-11	0.260182	0.0390889	46.6563	18.7464	4.36E-02

4.5 Discussion

The complexities surrounding the pathobiology of AD make it difficult to identify genetically-mediated factors that might not only contribute to the disease, but also act as potential targets for intervention and treatment. Although progress has been made in characterizing aspects of the subclinical manifestations of AD using, e.g., post-mortem brain samples and sophisticated neuroimaging techniques, these studies often suffer from small sample sizes or a lack of integration with molecular phenotyping[170]. Obtaining molecular phenotypes associated with AD that could reveal drug targets on large numbers of living humans is extremely difficult for practical and ethical reasons (e.g., brain biopsies are notoriously problematic and risky). Therefore, practical alternatives are needed. GWAS initiatives have identified a few genes that have acted as entry points into the pathobiology of AD, but have not necessarily revealed many druggable factors[171,172]. As a result, we considered the use

Table 4.3: Top GWAS meta-analysis results derived from the metasoft program for IGAP.

ADGC1 Top 20 Meta-Analysis GWAS Results									
RSID	Chromsome	Position	Num Meta-analyzed Studies	PVALUE FE (Metasoft)	BETA FE (Metasoft)	STD FE (Metasoft)	I_SQUARE (Metasoft)	Q (Metasoft)	PVALUE Q (Metasoft)
rs2075650	19	45395619	7	1.56E-164	1.01075	0.0369741	81.2219	31.9521	1.67E-05
rs4420638	19	45422946	4	8.39E-123	1.12172	0.0475967	90.9811	33.2636	2.83E-07
rs429358	19	45411941	2	5.66E-111	1.4296	0.0638677	0	0.5986	4.39E-01
rs769449	19	45410002	2	1.57E-86	1.37085	0.0695289	0	2.96E-05	9.96E-01
rs157580	19	45395266	10	1.81E-77	-0.50507	0.0271093	56.1231	20.5119	1.50E-02
rs71352238	19	45394336	2	6.22E-61	1.0493	0.0637172	0	0.937504	3.33E-01
rs405509	19	45408836	10	7.65E-59	-0.41114	0.0254193	47.2925	17.0754	4.75E-02
rs439401	19	45414451	9	1.94E-57	-0.442135	0.0276787	27.7828	11.0777	1.97E-01
rs157581	19	45395714	2	2.62E-57	0.877809	0.0550169	0	0.402422	5.26E-01
rs6859	19	45382034	10	8.41E-49	0.373075	0.0254104	60.2141	22.6211	7.11E-03
rs34878901	19	45402477	5	1.67E-39	-0.453861	0.0345101	0	2.07306	7.22E-01
rs41290120	19	45382675	8	2.72E-34	-0.988403	0.0809445	63.0659	18.9527	8.34E-03
rs10402271	19	45329214	8	7.35E-23	0.281806	0.0286301	29.993	9.99901	1.89E-01
rs7412	19	45412079	2	6.10E-19	-0.924255	0.103963	0	0.747576	3.87E-01
rs1160985	19	45403412	3	3.00E-16	-0.416806	0.0509963	86.1594	14.4502	7.28E-04
rs2965101	19	45237812	9	5.11E-16	-0.225402	0.0277967	42.7708	13.9789	8.23E-02
rs8106922	19	45401666	2	1.18E-13	-0.423675	0.0571028	3.90855	1.04068	3.08E-01
rs412776	19	45379516	5	4.60E-12	0.342644	0.0495346	0	2.86085	5.81E-01
rs8103315	19	45254168	7	1.97E-11	0.294913	0.0439642	56.0497	13.6518	3.38E-02
rs1114832	19	45636201	11	2.81E-11	0.260182	0.0390889	46.6563	18.7464	4.36E-02

Table 4.4: Top MWAS meta-analysis results derived from MRbase for ADGC.

ADGC Top Hit Sorted Table

					ADGC Metasoft RESULTS									IGAP MRBASE RESULTS					
rsID	GENE	EFFECT ALLELE	OTHER ALLELE	GTEX TISSUE	CASES	CONTROLS	Num Meta-analyzed Studies	PVALUE FE (Metasoft)	BETA FE (Metasoft)	STD FE (Metasoft)	I_SQUARE (Metasoft)	Q (Metasoft)	PVALUE Q (Metasoft)	PVALUE FE (MRBase)	MR Beta (MRBase)	SE (MRBase)	GWAS PVAL (MRBase)	GWAS BETA (PLINK)	GWAS SE (PLINK)
rs439401	APOE	T	C	Skin Sun Exposed Lower leg	9247	8488	20	9.73E-36	-1.08513	0.086957	29.4885	26.9459	0.10592	1.84E-11	1.0466585	0.15580127	3.55E-79	-0.3609	0.0192
rs439401	APOC1P1	T	C	Adrenal Gland	9247	8488	20	1.64E-28	-0.6854	0.061882	21.7315	24.2754	0.1858	4.57E-08	0.6643109	0.12150456	3.55E-79	-0.3609	0.0192
rs6859	PVRL2	A	G	Artery Tibial	7826	7202	21	6.44E-17	-1.52	0.181888	10.362	22.312	0.32383	2.49E-06	1.9543336	0.41503275	3.31E-96	0.334	0.016
rs6859	TOMM40	A	G	Esophagus Mucosa	7826	7203	21	1.97E-16	-1.76005	0.214023	8.76947	21.9225	0.34473	4.58E-06	2.2592867	0.49297546	3.31E-96	0.334	0.016
rs440277	CTB-129P6.4	G	A	Pancreas	5086	2658	6	1.05E-08	0.386839	0.067591	0	1.01367	0.96146	1.39E-07	-0.2413446	0.04582024	3.03E-15	0.1411	0.0179
rs440277	PVRL2	G	A	Pancreas	5086	2658	6	1.05E-08	0.386839	0.067591	0	1.01367	0.96146	4.07E-10	-0.2329544	0.03726396	3.03E-15	0.1411	0.0179
rs440277	CTB-129P6.4	G	A	Liver	5086	2658	6	3.06E-08	0.332506	0.060039	0	0.9962	0.96287	1.24E-05	-0.3052324	0.06985398	3.03E-15	0.1411	0.0179
rs440277	PVRL2	G	A	Liver	5086	2658	6	3.06E-08	0.332506	0.060039	0	0.9962	0.96287	2.21E-08	-0.2038611	0.03643994	3.03E-15	0.1411	0.0179
rs584007	APOE	A	G	Skin Not Sun Exposed Suprapubic	1257	2364	8	5.51E-07	-1.29818	0.259235	16.5523	8.38849	0.29959	8.02E-08	1.4028712	0.2614004	1.06E-82	-0.3683	0.0191
rs5167	APOC1	T	G	Lung	9624	8775	22	5.52E-07	0.58432	0.116696	2.27791	21.4895	0.42942	1.57E-04	-0.5281253	0.13972167	2.27E-11	-0.1122	0.0168
rs636147	MS4A6A	C	T	Whole Blood	2265	3487	10	5.61E-07	1.56159	0.312066	20.7463	11.3559	0.2521	2.17E-05	-0.6269056	0.14760602	1.11E-08	-0.0897	0.0157
rs584007	APOE	A	G	Skin Not Sun Exposed Suprapubic	1257	2364	8	4.41E-06	-0.71917	0.156648	4.87947	7.35908	0.39247	8.02E-08	1.4028712	0.2614004	1.06E-82	-0.3683	0.0191
rs714948	CEACAM19	C	A	Cells Transformed fibroblasts	5086	2661	6	3.07E-05	0.297025	0.071255	37.4835	7.99789	0.15635	1.18E-09	-0.2429774	0.0399435	6.26E-13	-0.1877	0.0261
rs714948	CEACAM19	C	A	Thyroid	5086	2661	6	3.36E-05	0.319083	0.076928	37.3643	7.98267	0.15719	1.46E-09	-0.2563971	0.0423865	6.26E-13	-0.1877	0.0261
rs10846537	DDX55	C	T	Adipose Subcutaneous	1306	3220	10	5.33E-05	-0.50527	0.125049	0	2.07392	0.99024	7.22E-01	0.0134872	0.03784839	7.22E-01	-0.0056	0.0157
rs6710467	BIN1	G	A	Pancreas	8938	7621	19	8.04E-05	-0.15652	0.039692	28.5061	25.177	0.12012	1.29E-06	0.1551462	0.03204903	4.23E-08	-0.126	0.023

Table 4.5: Top MWAS meta-analysis results derived from MRbase for IGAP.

IGAP Top Hit Sorted Table

					ADGC Metasoft RESULTS									IGAP MRBASE RESULTS					
rsID	GENE	EFFECT ALLELE	OTHER ALLELE	GTEX TISSUE	CASES	CONTROLS	Num Meta-analyzed Studies	PVALUE FE (Metasoft)	BETA FE (Metasoft)	STD FE (Metasoft)	I_SQUARE (Metasoft)	Q (Metasoft)	PVALUE Q (Metasoft)	PVALUE FE (MRBase)	MR Beta (MRBase)	SE (MRBase)	GWAS PVAL (MRBase)	GWAS BETA (PLINK)	GWAS SE (PLINK)
rs439401	APOE	T	C	Skin_Sun_Exposed_Lower_leg	9247	8488	20	9.73E-36	-1.08513	0.0869569	29.4885	26.9459	0.105918	1.84E-11	1.0466585	0.1558	3.55E-79	-0.3609	0.0192
rs440277	PVRL2	G	A	Pancreas	5086	2658	6	1.05E-08	0.386839	0.0675908	0	1.01367	0.961456	4.07E-10	-0.2329544	0.03726	3.03E-15	0.1411	0.0179
rs714948	CEACAM19	C	A	Cells_Transformed_fibroblasts	5086	2661	6	3.07E-05	0.297025	0.0712548	37.4835	7.99789	0.156352	1.18E-09	-0.2429774	0.03994	6.26E-13	-0.1877	0.0261
rs714948	CEACAM19	C	A	Thyroid	5086	2661	6	3.36E-05	0.319083	0.076928	37.3643	7.98267	0.157193	1.46E-09	-0.2563971	0.04239	6.26E-13	-0.1877	0.0261
rs440277	PVRL2	G	A	Liver	5086	2658	6	3.06E-08	0.332506	0.0600389	0	0.996197	0.962872	2.21E-08	-0.2038611	0.03644	3.03E-15	0.1411	0.0179
rs439401	APOC1P1	T	C	Adrenal_Gland	9247	8488	20	1.64E-28	-0.6854	0.061882	21.7315	24.2754	0.1858	4.57E-08	0.6643109	0.1215	3.55E-79	-0.3609	0.0192
rs584007	APOE	A	G	Skin_Not_Sun_Exposed_Suprapubic	1257	2364	8	5.51E-07	-1.29818	0.259235	16.5523	8.38849	0.299586	8.02E-08	1.4028712	0.2614	1.06E-82	-0.3683	0.0191
rs584007	APOE	A	G	Skin_Not_Sun_Exposed_Suprapubic	1257	2364	8	4.41E-06	-0.719165	0.156648	4.87947	7.35908	0.392473	8.02E-08	1.4028712	0.2614	1.06E-82	-0.3683	0.0191
rs440277	CTB-129P6.4	G	A	Pancreas	5086	2658	6	1.05E-08	0.386839	0.0675908	0	1.01367	0.961456	1.39E-07	-0.2413446	0.04582	3.03E-15	0.1411	0.0179
rs714948	CEACAM19	C	A	Colon_Transverse	5086	2661	6	0.00016451	0.32911	0.0873415	34.9014	7.68066	0.174736	4.18E-07	-0.2788557	0.05511	6.26E-13	-0.1877	0.0261
rs714948	CEACAM19	C	A	Artery_Tibial	5086	2661	6	0.00018656	0.439065	0.117505	34.671	7.65357	0.17639	5.02E-07	-0.3672175	0.07307	6.26E-13	-0.1877	0.0261
rs10403682	CTB-171A8.1	C	T	Brain_Caudate_basal_ganglia	2949	906	4	0.180933	0.0858978	0.064204	68.0376	9.38602	0.024575	5.47E-07	-0.123938	0.02474	4.41E-10	-0.1508	0.0242
rs714948	CEACAM19	C	A	Adipose_Subcutaneous	5086	2661	6	0.00027071	0.442713	0.121564	33.9516	7.57021	0.181567	1.26E-06	-0.3754284	0.07748	6.26E-13	-0.1877	0.0261
rs6710467	BIN1	G	A	Pancreas	8938	7621	19	8.04E-05	-0.156516	0.0396916	28.5061	25.177	0.120118	1.29E-06	0.1551462	0.03205	4.23E-08	-0.126	0.023
rs714948	CEACAM19	C	A	Lung	5086	2661	6	0.00027746	0.510827	0.140512	33.9019	7.56452	0.181925	1.40E-06	-0.4338827	0.08992	6.26E-13	-0.1877	0.0261
rs714948	CEACAM19	C	A	Brain_Caudate_basal_ganglia	5086	2661	6	0.00030704	0.222602	0.0616747	33.6947	7.54088	0.18342	1.88E-06	-0.1891471	0.03969	6.26E-13	-0.1877	0.0261
rs35103166	BIN1	T	C	Whole_Blood	804	224	2	0.516715	-0.363049	0.5599	75.548	4.08964	0.043147	2.10E-06	-0.4560672	0.09615	2.13E-08	-0.0934	0.0167
rs714948	CEACAM19	C	A	Artery_Coronary	5086	2661	6	0.00036626	0.339356	0.0952371	33.3224	7.49878	0.186109	2.22E-06	-0.2865329	0.06055	6.26E-13	-0.1877	0.0261
rs6859	PVRL2	A	G	Artery_Tibial	7826	7203	21	6.44E-17	-1.52	0.181888	10.362	22.312	0.323825	2.49E-06	1.9543336	0.41503	3.31E-96	0.334	0.016
rs9473119	RP11-385F7.1	G	A	Brain_Cerebellum	918	2177	7	0.878051	0.0099679	0.064963	2.03156	6.12442	0.409398	2.56E-06	0.0916076	0.01948	4.59E-08	-0.0938	0.0172

of emerging statistical analysis strategies rooted in MR tests and MWAS concepts, in which GWAS data are repurposed to accommodate the imputation or assignment of predicted molecular (or intermediate) phenotypes (IPs) to a cohort of individuals based on their genetic profiles. These imputed or assigned molecular phenotypes can then be tested for association with a trait. In this light, one can avoid costly and problematic measurement of factors in less accessible tissues (like the brain) and still obtain evidence implicating a molecular factor in the pathobiology of the disease (at least if certain assumptions uphold).

We pursued an MWAS of AD that considered gene expression levels as candidate IPs. We leveraged available GWAS data on three very large cohorts. We used the SNP-gene expression relationships in 44 tissues described in the GTEx database as IPs in our MWAS as well as state-of-the-field MR test analytical methods to test each SNP-gene expression pair's relationship to AD. Unfortunately, we did not find a great deal of consistency in the MWAS and individual MR test results across the three cohorts, despite the fact we explicitly tested for heterogeneity in effect sizes and found little evidence for heterogeneity. The exceptions involved the APOE gene, an obvious factor in AD susceptibility, and a few other genes that exhibited marginally significant associations in one or more of the cohorts, as noted in Table 4.5.

There are limitations to our study, however, despite the large sample sizes of

the GWAS cohorts used and the state-of-the-field GTEx information and analytical methods. For example, as noted, our ability to assess associations between predicted gene expression and AD depends critically on how well we can predict gene expression values from genetic variants through resources such as the GTEx database. This in turn depends on the sample sizes in the GTEx database, the reliability of the gene expression assays, and the how comprehensive the genotyping of the samples is. Many research groups are seeking to improve and extend both databases like GTEx[168] as well as analytical methods for predicting gene expression values34, both of which future work can take advantage of in studies of AD. In addition, the gene expression data we used from GTEx was confined to 44 tissues whose relevance to AD may not be the strongest. Despite this fact, we believe strategies such as MWAS and individual MR tests have promise in elucidating both the pathobiology and drug targets complex diseases like AD and could be pursued with other, perhaps more compelling, IPs, such as protein levels, metabolite levels, lipid levels, or other factors once databases for these factors are constructed in an analogous manner to GTEx.

4.6 Acknowledgements

Chapter 4, in full is currently being prepared for submission for publication, Quarless Q., Mitra I., ADGC GROUP, Schellenberg G., and Schork N. "Comprehensive Gene Expression-Based Mediator-Wide Association Study of Alzhei-mer's Disease". The dissertation author is the primary researcher and author on this paper.

Chapter 5

Conclusions and Discussion

5.1 Summary

We have pursued a number of data analyses designed to identify and characterize settings in which genetic effects on a human disease are context-dependent. We focused on three broad settings. First, we considered issues surrounding the use of DNA sequencing to facilitate clinical diagnoses and prognoses of patients infected with MRSA. These issues considered how well sequencing could lead to appropriate diagnoses as opposed to standard clinical culture-based systems. Secondly, we considered how the reference genome choice impacts not only the identification of variants on the MRSA genome, but also the association between genomic variation within strains of MRSA. This was specifically for samples obtained from patients being treated in a hospital for MRSA infection and patient outcomes. And lastly, we considered how context-dependent variants associate with AD through tools which leverage Mendelain Randomization. We assessed these effects by testing whether or not a number of variants associated with AD through intermediate phenotypes. In this particular case we utilized particular gene expression levels to help identify groups of variants associated with AD. We found evidence for context specificity in both settings. We briefly summarize our findings for each of these activities below.

5.1.1 MRSA Whole Genome Sequencing vs. Standard Clinical Identification Methods

We developed and implemented an end-to-end *de novo* assembly pipeline for whole genome sequencing (WGS) and characterizing sequence variants for strains of MRSA isolated from hospital patients. The purpose of this investigation was to establish baseline accuracy for WGS based pathogen identification compared to gold-standard clinical practices. In our pipeline, a successful identification was defined as $\geq$95% identity with an entry in the NCBI non-redundant bacterial sequence database across $\geq$25% of that species' genome. Our study demonstrated that out of 350 samples where WGS was implemented, roughly 10% (38 samples) showed discrepancies with the standard clinical identification methods. Despite this discordance, WGS was able to provide accurate and high resolution species identification with faster turnaround times, increased cost effectiveness, and additionally was able identifying a number a polymicrobial infections which would have been undetected through conventional methods. Thus, DNA sequencing protocols have the potential to be more accurate in MRSA pathogen identification than standard methodologies.

5.1.2 MRSA Reference Genome Implications for Investigating Clinical Correlates

Genetic variation, i.e., the genetic background or mutational profile, of infectious bacteria is understood to mechanistically decrease the efficacy of antimicrobial treatments while increasing pathogenicity. Although many mutations exhibit a binary response in their ability to promote disease or virulence effects, heterogeneity in the genetic background of infectious pathogens has been observed through numerous whole genome sequencing investigations. This heterogeneity can exist with respect to the either the presentation of clinical phenotypes stemming from host-pathogen interactions or through differential outcomes for pathogenic virulence factors, where a reference genome from a similar species of pathogen or comparator strain is utilized for analytic purposes. Our investigation was designed to quantify the degree to which the choice of a reference genome for identifying MRSA genetic variants impacts as-

sessments of the correlations between in patient outcomes and MRSA genomic profile. We found that choice of reference genome can significantly impact the correlation of MRSA strains with clinical outcomes and antibiotic drug responses profiles. This suggests that the context within which MRSA genomes are evaluated could impact inferences about pathogen virulence.

5.1.3 Associating Alzheimer's Disease Factors Through Intermediate Phenotypes

The identification of genetic variants that influence susceptibility to AD is intricate and complex for a number of reason as the previous section makes clear. Most variants which associate with AD have weak to moderate effects, thus, this complicates identifying disease contributing variants without large sample sizes. However, through statistics its is possible to exploit biological links that exists between genetics variants and AD, then the power to detect the effects of those variants, which may increase power. We leveraged the use of mediator variables and a predicted intermediate phenotype, gene expression levels, to identify a relationship between AD and a gene previously not thought to be associated with AD. We also found that there are a group of potential genes that could impact AD using the same methodology and further that some of these genes have a heterogeneous relationship with AD. Thus, we were able to identify groups of variants whose association with AD can best be brought to light by their impact on gene expression, this creating a context within which their effects occur.

5.2 Limitations

Although we used what we feel are state-of-the-art genomic and clinical phenotype data sets, we recognize that inherent limitations exist that may impact both the generalizability of the results of our analyses as well as direct interpretation of our findings. We discuss some of the more salient limitations in isolation below.

5.3 Small sample sizes

The sample sizes that we used for our studies were quite large relative to other studies considered infectious disease and genetic association studies. However, since the goal of our studies was to identify interactions and context-specific effects of variants that might be subtle, even larger sample sizes would have been desirable.

5.3.1 Better and More Sophisticated Clinical Data and Outcomes

For our MRSA studies exploring the clinical utility of genomic sequencing relative to the traditional culture-based pathogen identification methods, we would have benefited from having more sophisticated outcome measures on a larger collection of patients. For example, longitudinal data, host-specific factor characterization, medication use, prior health issues, etc. information on the patients would have helped put into our results into an even more compelling clinical context.

5.3.2 More reference genomes for MRSA project

Our analysis of the MRSA genomes and the choice of a reference genome focused on the differences in interpretation when two different reference genomes are used. Many other MRSA reference genomes have been discussed in the literature and it would have been ideal had we been able to consider those genomes as well.

5.3.3 Additional ADGC Clinical Data to Refine AD Diagnosis

Our analysis of the ADGC data considered genotypes that were assigned or imputed to individuals that did not have those genotypes. This imputation strategy was pursued by the original ADGC investigative team. Unfortunately, the reliability of the genotyping resulting from imputation strategies depends critically on the imputation strategy itself. Although we have no reason to believe that our results are not-trustworthy due to the imputation, it would have been better to use directly genotyped variants rather than imputed variants in our analyses.

5.3.4 Use of Quantitative Phenotypes for Association

We focused on AD diagnosis as our primary dependent variable, but we could have benefited from an analysis of a subclinical phenotype, like cognitive decline or cognitive score to reduce problems associated with criteria for AD diagnosis.

5.3.5 Greater Diversity of Individuals in our AD Study

Ironically, our studies of the context dependency of variants that influence AD focused on individuals of largely European descent. It would have been ideal to explore the effect of the variants on AD susceptibility in other genetic backgrounds. This would allow us to test the hypothesis that greater context-dependency occurs, whereby the variants in different ancestral populations modifies the influence of the APOE4 allele on other variants.

5.3.6 GTEX Database Limitations

We used the GTEX database to build models relating genetic variants to gene expression values. Although state-of-the-field, the GTEX database only contains information on a few hundred individuals, which could compromise the power to identify variant-gene expression associations. This could easily impact our ability to relate predicted gene expression values to AD based on GTEX-derived models.

5.3.7 Prediction Modeling

As noted above, the assignment of predicted gene expression values to individuals in the ADGC dataset depends critically on the reliability of the models relating genetic variants to gene expression values. There are many ways to build such models. A study comparing the different methods and how they might impact the results of mediator analyses like ours would help put into context how sensitive our results are to model choice.

5.4 Future Directions

It is quite likely that as genomic and related technologies advance and become more efficient, they will be used in a wide variety of settings in which context will matter to an even greater degree. For example, we believe that at some point pathogen sequencing will become routine in clinical settings. If this is will be the case, then making sure the clinical impact of a particular pathogen is not in doubt will be crucial. The reference-guided characterization of variants in pathogen genomes used to date is complicated and potentially problematic, as our analyses suggest: if the mere choice of a reference for variant identification and genome characterization can impact interpretation, then the downstream clinical diagnostic and prognostic interpretation of the pathogen is likely to be even more affected. Therefore, alternative strategies, such as de novo assembly of pathogen genomes, are needed. In addition, greater attention to host-related factors, such as a compromised immune system, co-infection, or debilitating disease, will be necessary to understand the likely impact that an infection may have on an individual's health. These host-related factors can also be characterized with genetic and related technologies. Thus, greater integration of data sources will be required for the effective use of pathogen sequencing in clinical practice.

In the context of the discovery of genetic factors that influence common chronic conditions such as Alzheimer's Disease (AD), not only will better strategies for identifying variants that might be associated with AD be necessary, but also better characterization of their likely functional impact will be necessary. The mere identification of an associated variant - even in a context dependent manner - will not be as useful as understanding the functional impact of the variant. Thus, more sophisticated high-throughput functional assays, such as those leveraging emerging CRISPR-based technologies[173], will be essential for validating and putting into even more basic biological contexts the effects of genetic variants. Such studies can consider the influence of genetic background by modifying the sequence of relevant constructs using cells or organoids with different genetic backgrounds and then determining if the background influenced the in vitro activity of the variant of interest. In addition, although we explored the use of gene expression levels as a 'mediator' between genetic factor effects

and the clinical phenotype of AD, there are many other possibly mediators that could have been used; for example, protein levels, metabolite levels, cognition scores, etc. Future work will leverage information about these possible mediators to link genetic variants to AD and other complex diseases using technology similar to what we used. Ultimately, all of biology and life depends on the interplay of genetic, environmental and stochastic factors. Believing that individual genetic factors will have robust and completely reproducible effects within every genetic background and environment flies in the face of this fact. Therefore, methodology that can accommodate and characterize the context-specific effects of genetic factors is absolutely essential for moving biology in a genomic era forward. When taken in this light, we hope that our work will motivate future research.

Bibliography

[1] Wen-Hua Wei, Gibran Hemani, and Chris S Haley. Detecting epistasis in human complex traits. *Nat Rev Genet*, 15(11):722–733, nov 2014.

[2] Patrick C Phillips. Epistasis [mdash] the essential role of gene interactions in the structure and evolution of genetic systems. *Nat Rev Genet*, 9(11):855–867, nov 2008.

[3] Joshua S Bloom, Ian M Ehrenreich, Wesley T Loo, Thuy-Lan Vo Lite, and Leonid Kruglyak. Finding the sources of missing heritability in a yeast cross. *Nature*, 494(7436):234–7, February 2013.

[4] Christopher H Chandler, Sudarshan Chari, and Ian Dworkin. Does your gene need a background check? How genetic background impacts the analysis of mutations, genes, and evolution. *Trends in genetics : TIG*, 29(6):358–66, June 2013.

[5] Sudarshan Chari and Ian Dworkin. The Conditional Nature of Genetic Interactions: The Consequences of Wild-Type Backgrounds on Mutational Interactions in a Genome-Wide Modifier Screen. *PLoS Genetics*, 9(8):e1003661, August 2013.

[6] Wen Huang, Stephen Richards, Mary Anna Carbone, Dianhui Zhu, Robert R H Anholt, Julien F Ayroles, Laura Duncan, Katherine W Jordan, Faye Lawrence, Michael M Magwire, Crystal B Warner, Kerstin Blankenburg, Yi Han, Mehwish Javaid, Joy Jayaseelan, Shalini N Jhangiani, Donna Muzny, Fiona Ongeri, Lora Perales, Yuan-Qing Wu, Yiqing Zhang, Xiaoyan Zou, Eric A Stone, Richard A Gibbs, and Trudy F C Mackay. Epistasis dominates the genetic architecture of Drosophila quantitative traits. *Proceedings of the National Academy of Sciences of the United States of America*, 109(39):15553–15559, sep 2012.

[7] Or Zuk, Eliana Hechter, Shamil R Sunyaev, and Eric S Lander. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences of the United States of America*, 109(4):1193–8, January 2012.

[8] The Wellcome, Trust Case, and Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–78, June 2007.

[9] Hakon Hakonarson, Struan F a Grant, Jonathan P Bradfield, Luc Marchand, Cecilia E Kim, Joseph T Glessner, Marcella Devoto, Hui-Qi Qu, and Constantin Polychronakos. A genome-wide association study identifies KIAA0350 as a type 1 diabetes gene. *Nature*, 448(7153):591–4, August 2007.

[10] Guillaume Lettre, Cameron D Palmer, Taylor Young, Kenechi G Ejebe, Hooman Allayee, Emelia J Benjamin, Franklyn Bennett, Donald W Bowden, Aravinda Chakravarti, Al Dreisbach, Deborah N Farlow, James G Wilson, Richard R Fabsitz, Stacey B Gabriel, Sekar Kathiresan, and Eric Boerwinkle. Genome-wide association study of coronary heart disease and its risk factors in 8,090 African Americans: the NHLBI CARe Project. *PLoS genetics*, 7(2):e1001300, January 2011.

[11] John A Todd, Neil M Walker, Jason D Cooper, Deborah J Smyth, Kate Downes, Vincent Plagnol, Rebecca Bailey, Sergey Nejentsev, Sarah F Field, Felicity Payne, Christopher E Lowe, Jeffrey S Szeszko, Jason P Hafler, Lauren Zeitels, Jennie H M Yang, Adrian Vella, Sarah Nutland, Helen E Stevens, Helen Schuilenburg, Gillian Coleman, Meeta Maisuria, William Meadows, Luc J Smink, Barry Healy, Oliver S Burren, Alex A C Lam, Nigel R Ovington, James Allen, Ellen Adlem, Hin-Tak Leung, Chris Wallace, Joanna M M Howson, Cristian Guja, Constantin Ionescu-Tirgoviste, Matthew J Simmonds, Joanne M Heward, Stephen C L Gough, David B Dunger, Linda S Wicker, and David G Clayton. Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nat Genet*, 39(7):857–864, jul 2007.

[12] Matthew Holderfield, Marian M Deuker, Frank McCormick, and Martin McMahon. Targeting RAF kinases for cancer therapy: BRAF-mutated melanoma and beyond. *Nat Rev Cancer*, 14(7):455–467, jul 2014.

[13] Tiangui Huang, Michael Karsy, Jian Zhuge, Minghao Zhong, and Delong Liu. B-raf and the inhibitors: from bench to bedside. *Journal of Hematology & Oncology*, 6(1):30, 2013.

[14] Caitriona Holohan, Sandra Van Schaeybroeck, Daniel B Longley, and Patrick G Johnston. Cancer drug resistance: an evolving paradigm. *Nat Rev Cancer*, 13(10):714–726, oct 2013.

[15] Konstantina G Yiannopoulou and Sokratis G Papageorgiou. Current and future treatments for AlzheimerâĂŹs disease. *Therapeutic Advances in Neurological Disorders*, 6(1):19–33, jan 2013.

[16] Wayne N Frankel and Nicholas J Schork. Who's afraid of epistasis? *Nat Genet*, 14(4):371–373, dec 1996.

[17] Teri A Manolio, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia A Hindorff, David J Hunter, Mark I McCarthy, Erin M Ramos, Lon R Cardon, Aravinda Chakravarti, Judy H Cho, Alan E Guttmacher, Augustine Kong, Leonid Kruglyak, Elaine Mardis, Charles N Rotimi, Montgomery Slatkin, David Valle, Alice S Whittemore, Michael Boehnke, Andrew G Clark, Evan E Eichler, Greg Gibson, Jonathan L Haines, Trudy F C Mackay, Steven A McCarroll, and Peter M Visscher. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, oct 2009.

[18] Guo-Bo Chen. On the reconciliation of missing heritability for genome-wide association studies, jul 2016.

[19] Leonard D Pysh. Two alleles of the AtCesA3 gene in Arabidopsis thaliana display intragenic complementation 1. 102:1434–1441, 2015.

[20] José Manuel Pérez-Pérez, Héctor Candela, and José Luis Micol. Understanding synergy in genetic interactions. *Trends in Genetics*, 25(8):368–376, sep 2016.

[21] Trudy F C Mackay. Epistasis and quantitative traits: using model organisms to study gene-gene interactions. *Nat Rev Genet*, 15(1):22–33, jan 2014.

[22] William Bateson and Gregor Mendel. *Mendel's Principles of Heredity: A Defence, with a Translation of Mendel's Original Papers on Hybridisation*. Cambridge University Press, Cambridge, 007 2009.

[23] Atsushi Yoshiki and Kazuo Moriwaki. Mouse phenome research: implications of genetic background. *ILAR journal / National Research Council, Institute of Laboratory Animal Resources*, 47(2):94–102, January 2006.

[24] Marylyn D Ritchie, Emily R Holzinger, Ruowang Li, Sarah A Pendergrass, and Dokyoon Kim. Methods of integrating data to uncover genotype-phenotype interactions. *Nat Rev Genet*, 16(2):85–97, feb 2015.

[25] Katharine P Hummel, Douglas L Coleman, and Priscilla W Lane. The Influence of Genetic Background on Expression of Mutations at the Diabetes Locus in the Mouse . I . C57BL / KsJ and C57BL / 6J Strains. 13:1–13, 1972.

[26] D L Coleman. The Influence of Genetic Background on the Expression of the Obese. 293:287–293, 1973.

[27] Manuel C Lemos, Brian Harding, Anita A C Reed, Jeshmi Jeyabalan, Gerard V Walls, Michael R Bowl, James Sharpe, Sarah Wedden, Julie E Moss, Allyson Ross, Duncan Davidson, and Rajesh V Thakker. Genetic background influences embryonic lethality and the occurrence of neural tube defects in Men1 null mice : relevance to genetic modifiers. 2008.

[28] Jackson Laboratory. Genetic Background: Understanding its importance in mouse-based biomedical research. 2006.

[29] Xavier Montagutelli. Effect of the Genetic Background on the Phenotype of Mouse Mutations. pages 101–105, 2000.

[30] Bert Vogelstein, Nickolas Papadopoulos, Victor E Velculescu, Shibin Zhou, Luis a Diaz, and Kenneth W Kinzler. Cancer genome landscapes. *Science (New York, N.Y.)*, 339(6127):1546–58, March 2013.

[31] J Luo, N Solimini, and S. Elledge. Principles of Cancer Therapy: Oncogene and Non-oncogene Addiction. *Cell*, 6(6):823–837, 2009.

[32] Charles Lu, Mingchao Xie, Michael C Wendl, Jiayin Wang, Michael D Mclellan, Mark D M Leiserson, Kuan-lin Huang, Matthew A Wyczalkowski, Reyka Jayasinghe, Tapahsama Banerjee, Jie Ning, Piyush Tripathi, Qunyuan Zhang, Beifang Niu, Kai Ye, Heather K Schmidt, Robert S Fulton, Joshua F Mcmichael, Prag Batra, Cyriac Kandoth, Maheetha Bharadwaj, Daniel C Koboldt, Christopher A Miller, Krishna L Kanchi, James M Eldred, David E Larson, John S Welch, Ming You, Bradley A Ozenberger, Ramaswamy Govindan, Matthew J Walter, Matthew J Ellis, Elaine R Mardis, Timothy A Graubert, John F Dipersio, Timothy J Ley, Richard K Wilson, Paul J Goodfellow, Benjamin J Raphael, Feng Chen, Kimberly J Johnson, and Jeffrey D Parvin. germline variants across 12 cancer types. *Nature Communications*, 6:1–13, 2015.

[33] Qian Zhu, Peng Zhu, Yilei Zhang, Jie Li, Xuejun Ma, Ning Li, Qi Wang, Xiujuan Xue, Le Luo, Zizhao Li, Huijun Z Ring, Brian Z Ring, and Li Su. Analysis of Social and Genetic Factors Influencing Heterosexual Transmission of HIV within Serodiscordant Couples in the Henan Cohort. *PloS one*, 10(6):e0129979, January 2015.

[34] Laura. Goetz, Liliana. Uribe-Bruce, Danjuma Quarless, Libeger. Ondrej, and Schork Nicholas. Admixture and Clinical Phenotype Variation. *Frontiers in Genetics*, 2014.

[35] Indrani Halder, Kevin E Kip, Suresh R Mulukutla, Aryan N Aiyer, Oscar C Marroquin, Gordon S Huggins, and Steven E Reis. Biogeographic ancestry, self-identified race, and admixture-phenotype associations in the Heart SCORE Study. *American journal of epidemiology*, 176(2):146–155, July 2012.

[36] N Kato. Ethnic differences in genetic predisposition to hypertension. *Hypertension Research*, pages 574–581, 2012.

[37] Greg Gibson. Rare and common variants: twenty arguments. *Nature reviews. Genetics*, 13(2):135–45, February 2011.

[38] Michael W Lutz, Scott S Sundseth, Daniel K Burns, Ann M Saunders, Kathleen M Hayden, James R Burke, Kathleen A Welsh-bohmer, Allen D Roses, and Disease Neuroimaging. A genetics-based biomarker risk algorithm for predicting risk of Alzheimer's disease. 2:30–44, 2016.

[39] Jian Yang, S Hong Lee, Michael E Goddard, and Peter M Visscher. GCTA : A Tool for Genome-wide Complex Trait Analysis. *The American Journal of Human Genetics*, 88(1):76–82, 2011.

[40] Jian Yang, Teri A Manolio, Louis R Pasquale, Eric Boerwinkle, Neil Caporaso, Julie M Cunningham, Mariza de Andrade, Bjarke Feenstra, Eleanor Feingold, M Geoffrey Hayes, William G Hill, Maria Teresa Landi, Alvaro Alonso, Guillaume Lettre, Peng Lin, Hua Ling, William Lowe, Rasika A Mathias, Mads Melbye, Elizabeth Pugh, Marilyn C Cornelis, Bruce S Weir, Michael E Goddard, and Peter M Visscher. Genome partitioning of genetic variation for complex traits using common SNPs. *Nat Genet*, 43(6):519–525, jun 2011.

[41] R.A. Fisher. The correlation between relatives on the supposition of mendelian inheritance. 52:399–433, 1918.

[42] T Vogwill, M Kojadinovic, and R C Maclean. Epistasis between antibiotic resistance mutations and genetic background shape the fitness effect of resistance across species of Pseudomonas. 2016.

[43] Genomic Basis for Methicillin Resistance in Staphylococcus aureus. *Infection & chemotherapy*, 45(2):117–36, jun 2013.

[44] Nelson E Martins, Vitor G Faria, Luis Teixeira, Sara Magalhães, and Élio Sucena. Host adaptation is contingent upon the infection route taken by pathogens. *PLoS pathogens*, 9(9):e1003601, January 2013.

[45] Mark Loeb. Host genomics in infectious diseases. *Infection & chemotherapy*, 45(3):253–9, September 2013.

[46] H A Crosby, J Kwiecinski, and A R Horswill. Chapter One - Staphylococcus aureus Aggregation and Coagulation Mechanisms, and Their Function in HostâĂŞPathogen Interactions. volume Volume 96, pages 1–41. Academic Press, 2016.

[47] Michelle E Mulcahy and Rachel M McLoughlin. Host-Bacterial Crosstalk Determines *Staphylococcus aureus* Nasal Colonization. *Trends in Microbiology*, 24(11):872–886, oct 2016.

[48] R Sen, L Nayak, and R K De. A review on hostâĂŞpathogen interactions: classification and prediction. *European Journal of Clinical Microbiology & Infectious Diseases*, 35(10):1581–1599, 2016.

[49] Julie G In, Jennifer Foulke-Abel, Mary K Estes, Nicholas C Zachos, Olga Kovbasnjuk, and Mark Donowitz. Human mini-guts: new insights into intestinal physiology and host-pathogen interactions. *Nat Rev Gastroenterol Hepatol*, 13(11):633–642, nov 2016.

[50] Thorhildur Halldorsdottir and Elisabeth B Binder. Gene ÃŮ Environment Interactions: From Molecular Mechanisms to Behavior. *Annual Review of Psychology*, jan 2016.

[51] O. VasemÃď gi A. Taavitsainen J. Tourunen A. Saloniemi I Vuorisalo, T. Arjamaa. High Lactose Tolerance in North Europeans: A Result of Migration, Not In Situ Milk Consumption. 55 no. 2:2154–2161, 2012.

[52] Andrew Szilagyi. Adaptation to Lactose in Lactase Non Persistent People: Effects on Intolerance and the Relationship between Dairy Food Consumption and Evalution of Diseases. *Nutrients*, 7(8):6751–6779, aug 2015.

[53] Harald Brüssow. Minireview Nutrition , population growth and disease : a short history of lactose. 15:2154–2161, 2013.

[54] Hudson Reddon, Jean-Louis Guéant, and David Meyre. The importance of gene–environment interactions in human obesity. *Clinical Science*, 130(18):1571–1597, 2016.

[55] Ruoxu Dou, Kimmie Ng, Edward L. Giovannucci, JoAnn E. Manson, Zhi Rong Qian, and Shuji Ogino. Vitamin d and colorectal cancer: molecular, epidemiological and clinical evidence. *British Journal of Nutrition*, 115(9):1643–1660, 005 2016.

[56] Tamara L Wall, Susan E Luczak, and Susanne Hiller-Sturmhöfel. Biology, Genetics, and Environment: Underlying Factors Influencing Alcohol Metabolism. *Alcohol Research : Current Reviews*, 38(1):59–68, 2016.

[57] Jason H Moore and Scott M Williams. Traversing the conceptual divide between biological and statistical epistasis : systems biology and a more modern synthesis. pages 637–646, 2005.

[58] EPIQâĂŤefficient detection of SNPâĂŞSNP epistatic interactions for quantitative traits. *Bioinformatics*, 30(12):i19–i25, jun 2014.

[59] Futao Zhang, Eric Boerwinkle, and Momiao Xiong. Epistasis analysis for quantitative traits by functional regression model. *Genome Research*, 24(6):989–998, jun 2014.

[60] Snehit Prabhu and Itsik Pe'er. Ultrafast genome-wide scan for SNPâĂŞSNP interactions in common complex disease. *Genome Research*, 22(11):2230–2240, nov 2012.

[61] Erdem Bangi, Dan Garza, and Marc Hild. In vivo analysis of compound activity and mechanism of action using epistasis in Drosophila. pages 55–68, 2011.

[62] Yaping Wang, Donghui Li, and Peng Wei. Powerful TukeyâĂŹs One Degree-of-Freedom Test for Detecting GeneâĂŞGene and GeneâĂŞEnvironment Interactions. *Cancer Informatics*, 14(Suppl 2):209–218, jun 2015.

[63] Sana Suri, Verena Heise, Aaron J Trachtenberg, and Clare E Mackay. Neuroscience and Biobehavioral Reviews The forgotten APOE allele : A review of the evidence and suggested mechanisms for the protective effect of APOE2. *Neuroscience and Biobehavioral Reviews*, 37(10):2878–2886, 2013.

[64] C William Rebeck, Joel S Reiter, Dudley K Strickland, Bradley Hyman, and Red Cross. Apolipoprotein E in Sporadic Alzheimer's Disease : Allelic Variation and Receptor Interactions. *Neuron*, 11:575–580, 1993.

[65] Walter A Kukull, Richard Mayeux, Richard H Myers, and Margaret A Pericakvance. Effects of Age, Sex, and Ethnicity on the Association Between Apolipoprotein E Genotype and Alzheimer Disease. *JAMA*, 02118, 2016.

[66] Sara Goodwin, John D Mcpherson, and W Richard Mccombie. Coming of age : ten years of next- generation sequencing technologies.

[67] Heidi Chial. Mendelian Genetics : Patterns of Inheritance and Single âĂŚ Gene Disorders Autosomal Dominant Single âĂŚ Gene Diseases. (2008):1–7, 2008.

[68] Rita M Cantor, Kenneth Lange, and Janet S Sinsheimer. Prioritizing GWAS Results : A Review of Statistical Methods and Recommendations for Their Application. *The American Journal of Human Genetics*, 86(1):6–22, 2010.

[69] Christopher S Coffey, Patricia R Hebert, Marylyn D Ritchie, Harlan M Krumholz, J Michael Gaziano, Paul M Ridker, Nancy J Brown, Douglas E Vaughan, and Jason H Moore. An application of conditional logistic regression and multifactor dimensionality reduction for detecting gene-gene Interactions on risk of myocardial infarction : The importance of model validation. 10:1–10.

[70] Danielle Welter, Jacqueline Macarthur, Joannella Morales, Tony Burdett, Peggy Hall, Heather Junkins, Alan Klemm, Paul Flicek, Teri Manolio, Lucia Hindorff, and Helen Parkinson. The NHGRI GWAS Catalog , a curated resource of SNP-trait associations. 42(December 2013):1001–1006, 2014.

[71] Wei-yin Loh. Variable Selection for Classification and Regression in Large p , Small n Problems. 205:133–157, 2012.

[72] Hongkai Ji and X Shirley Liu. primer Analyzing 'omics data using hierarchical models. *Nature Publishing Group*, 28(4):337–340, 2010.

[73] Sonia Castillo-lluva, Lourdes Hontecillas-prieto, and Jian Hua Mao. Missing Heritability of Complex Diseases: Enlightenment by Genetic Variants From Intermediate Phenotypes. pages 664–673, 2016.

[74] Evan E Eichler, Jonathan Flint, Greg Gibson, Augustine Kong, Suzanne M Leal, Jason H Moore, and Joseph H Nadeau. Missing heritability and strategies for finding the underlying causes of complex disease. *Nature reviews. Genetics*, 11(6):446–50, June 2010.

[75] Sang Hong Lee, Naomi R Wray, Michael E Goddard, and Peter M Visscher. Estimating missing heritability for disease from genome-wide association studies. *American journal of human genetics*, 88(3):294–305, March 2011.

[76] Evan E Eichler, Jonathan Flint, Greg Gibson, Augustine Kong, Suzanne M Leal, Jason H Moore, and Joseph H Nadeau. Missing heritability and strategies for finding the underlying causes of complex disease. *Nature reviews. Genetics*, 11(6):446–50, June 2010.

[77] Tomasz M Ignac, Alexander Skupin, Nikita A Sakhanenko, and David J Galas. Discovering Pair-Wise Genetic Interactions : An Information Theory-Based Approach. 9(3):1–14, 2014.

[78] Joshua S Bloom, Iulia Kotenko, Meru J Sadhu, Sebastian Treusch, Frank W Albert, and Leonid Kruglyak. effects to quantitative trait variation in yeast. *Nature Communications*, 6:1–6, 2015.

[79] Anastasia Baryshnikova, Michael Costanzo, Chad L Myers, Brenda Andrews, and Charles Boone. Genetic Interaction Networks : Toward an Understanding of Heritability. (June):1–23, 2013.

[80] Onofre Combarros, Cornelia M Van Duijn, Naomi Hammond, Olivia Belbin, Alejandro Arias-vásquez, Mario Cortina-borja, Michael G Lehmann, Yurii S Aulchenko, Maaike Schuur, Heike Kölsch, Reinhard Heun, Gordon K Wilcock, Kristelle Brown, Patrick G Kehoe, Rachel Harrison, Eliecer Coto, Victoria Alvarez, Panos Deloukas, Ignacio Mateo, Rhian Gwilliam, Kevin Morgan, Donald R Warden, A David Smith, and Donald J Lehmann. Replication by the Epistasis Project of the interaction between the genes for IL-6 and IL-10 in the risk of Alzheimer ' s disease. 6:1–9, 2009.

[81] James M Bullock, Christopher Medway, Mario Cortina-borja, James C Turton, Jonathan A Prince, Carla A Ibrahim-verbaas, Maaike Schuur, Monique M Breteler, Cornelia M Van Duijn, Patrick G Kehoe, Rachel Barber, Eliecer Coto, Victoria Alvarez, Panos Deloukas, Naomi Hammond, Onofre Combarros, Ignacio Mateo, Donald R Warden, Michael G Lehmann, Olivia Belbin, Kristelle Brown, Gordon K Wilcock, Reinhard Heun, Heike Kölsch, A David Smith, Donald J Lehmann, and Kevin Morgan. Neurobiology of Aging Discovery by the Epistasis Project of an epistatic interaction between the GSTM3 gene and the HHEX / IDE / KIF11 locus in the risk of Alzheimer's disease. *Neurobiology of Aging*, 34(4):1309.e1–1309.e7, 2013.

[82] Eloy Rodríguez-rodríguez, Ignacio Mateo, Jon Infante, Javier Llorca, Inés García-gorostiaga, José Luis Vázquez-higuera, and Pascual Sánchez-juan. Interaction between HMGCR and ABCA1 cholesterol-related genes modulates Alzheimer ' s disease risk. *Brain Research*, 1280:166–171, 2009.

[83] D R Warden and A D Smith. Synergy between the C2 allele of transferrin and the C282Y allele of the haemochromatosis gene (HFE) as risk factors for developing Alzheimer's disease. pages 261–266, 2004.

[84] J S K Kauwe, S Bertelsen, K Mayo, C Cruchaga, R Abraham, P Hollingworth, D Harold, M J Owen, J Williams, S Lovestone, J C Morris, A M Goate, R Abraham, P Hollingworth, D Harold, J Williams, S Lovestone, and Morris Jc. Suggestive Synergy Between Genetic Variants in TF and HFE as Risk Factors for Alzheimer's Disease. (December):955–959, 2009.

[85] Cassandra E Murcray, Juan Pablo Lewinger, and W James Gauderman. Practice of Epidemiology Gene-Environment Interaction in Genome-Wide Association Studies. 169(2):219–226, 2009.

[86] Xavier Didelot, Rory Bowden, Daniel J Wilson, Tim E a Peto, and Derrick W Crook. Transforming clinical microbiology with bacterial genome sequencing. *Nature reviews. Genetics*, 13(9):601–12, sep 2012.

[87] David a Relman. Microbial genomics and infectious diseases. *The New England journal of medicine*, 365(4):347–57, jul 2011.

[88] Filipe J Ribeiro, Dariusz Przybylski, Shuangye Yin, Ted Sharpe, Sante Gnerre, Amr Abouelleil, Aaron M Berlin, Anna Montmayeur, Terrance P Shea, Bruce J Walker, Sarah K Young, Carsten Russ, Chad Nusbaum, Iain Maccallum, and David B Jaffe. Finished bacterial genomes from shotgun sequence data. pages 2270–2277, 2012.

[89] Vien Thi Minh Le and Binh An Diep. Selected Insights from Application of Whole Genome Sequencing for Outbreak Investigations. 19(5):432–439, 2014.

[90]

[91] Claudio U. KÃűser, Ole B. Schulz-Trieglaff, Geoffrey P. Smith, and Sharon J. Peacock. Rapid whole-genome sequencing for investigation of a neonatal mrsa outbreak. *New England Journal of Medicine*, 366(24):2267–2275, 2012. PMID: 22693998.

[92] Sandra Reuter, Matthew J Ellington, Edward J P Cartwright, Claudio U Köser, M Estée Török, Theodore Gouliouris, Simon R Harris, Nicholas M Brown, Matthew T G Holden, Mike Quail, Julian Parkhill, Geoffrey P Smith, Stephen D Bentley, and Sharon J Peacock. Rapid Bacterial Whole-Genome Sequencing to Enhance Diagnostic and Public Health Microbiology. *JAMA internal medicine*, 173(15):1397–1404, aug 2013.

[93] Daniel R Zerbino and Ewan Birney. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, 18(5):821–829, may 2008.

[94] David Hernandez, Patrice François, Laurent Farinelli, Magne Østerås, and Jacques Schrenzel. *De novo* bacterial genome sequencing : Millions of very short reads assembled on a desktop computer. pages 802–809, 2008.

[95] Henrik Hasman, Dhany Saputra, Thomas Sicheritz-Ponten, Ole Lund, Christina Aaby Svendsen, Niels Frimodt-MÃÿller, and Frank M. Aarestrup. Rapid whole-genome sequencing for detection and characterization of microorganisms directly from clinical samples. *Journal of Clinical Microbiology*, 52(1):139–146, 2014.

[96] Surekha Y Asangi, J Mariraj, and M S Sathyanarayan. Speciation of clinically significant Coagulase Negative Staphylococci and antibiotic resistant patterns in a tertiary care hospital . 2(3):735–739, 2011.

[97] A.M. Day, J.A.T. Sandoe, J.H. Cove, and M.K. Phillips-Jones. Evaluation of a biochemical test scheme for identifying clinical isolates of enterococcus faecalis and enterococcus faecium. *Letters in Applied Microbiology*, 33(5):392–396, 2001.

[98] Maria L. G. Quiloan, John Vu, and John Carvalho. Enterococcus faecalis can be distinguished from enterococcus faecium via differential susceptibility to antibiotics and growth and fermentation characteristics on mannitol salt agar. *Frontiers in Biology*, 7(2):167–177, 2012.

[99] Johan Goris, Konstantinos T. Konstantinidis, Joel A. Klappenbach, Tom Coenye, Peter Vandamme, and James M. Tiedje. DnaâĂŞdna hybridization values and their relationship to whole-genome sequence similarities. *International Journal of Systematic and Evolutionary Microbiology*, 57(1):81–91, 2007.

[100] Michael Richter and Ramon Rosselló-Móra. Shifting the genomic gold standard for the prokaryotic species definition. *Proceedings of the National Academy of Sciences of the United States of America*, 106(45):19126–19131, nov 2009.

[101] Sushim Kumar Gupta, Babu Roshan Padmanabhan, Seydina M Diene, Rafael Lopez-Rojas, Marie Kempf, Luce Landraud, and Jean-Marc Rolain. ARG-ANNOT, a New Bioinformatic Tool To Discover Antibiotic Resistance Genes in Bacterial Genomes. *Antimicrobial Agents and Chemotherapy*, 58(1):212–220, jan 2014.

[102] Johan Goris, Konstantinos T. Konstantinidis, Joel A. Klappenbach, Tom Coenye, Peter Vandamme, and James M. Tiedje. DnaâĂŞdna hybridization values and their relationship to whole-genome sequence similarities. *International Journal of Systematic and Evolutionary Microbiology*, 57(1):81–91, 2007.

[103] Stephen F. Altschul, Thomas L. Madden, Alejandro A. SchÃď ffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, 1997.

[104] Michael Richter and Ramon RossellÃş-MÃşra. Shifting the genomic gold standard for the prokaryotic species definition. *Proceedings of the National Academy of Sciences*, 106(45):19126–19131, 2009.

[105] Mark Woolhouse, Catriona Waugh, Meghan Rose Perry, and Harish Nair. Global disease burden due to antibiotic resistanc,, state of the evidence. 6(1):1–5, 2016.

[106] Li-yang Hsu, Simon R Harris, Monika A Chlebowicz, Jodi A Lindsay, Tse-hsien Koh, Prabha Krishnan, Thean-yen Tan, Pei-yun Hon, Warren B Grubb, Stephen D Bentley, Julian Parkhill, Sharon J Peacock, and Matthew T G Holden. Evolutionary dynamics of methicillin-resistant Staphylococcus aureus within a healthcare system. pages 1–13, 2015.

[107] Ethan R Wyrsch, Piklu Roy Chowdhury, Toni A Chapman, Ian G Charles, Jeffrey M Hammond, and Steven P Djordjevic. Genomic Microbial Epidemiology Is Needed to Comprehend the Global Problem of Antibiotic Resistance and to Improve Pathogen Diagnosis , 2016.

[108] Richard R. Watkins and Robert A. Bonomo. Overview: Global and local impact of antibiotic resistance. *Infectious Disease Clinics of North America*, 30(2):313 – 322, 2016. Antibiotic Resistance: Challenges and Opportunities.

[109] Louis Stokes and Cleveland V A Medical. Rising Threat of Infections Unfazed by Antibiotics. pages 2–3, 2010.

[110] Keiichi Hiramatsu, Teruyo Ito, Sae Tsubakishita, Takashi Sasaki, Fumihiko Takeuchi, Yuh Morimoto, Yuki Katayama, Miki Matsuo, Kyoko Kuwahara-Arai, Tomomi Hishinuma, and Tadashi Baba. Genomic Basis for Methicillin Resistance in Staphylococcus aureus. *Infection & chemotherapy*, 45(2):117–36, jun 2013.

[111] Gerard D Wright. The antibiotic resistome: the nexus of chemical and genetic diversity. 5(March):175–186, 2007.

[112] Kim Lewis. Platforms for antibiotic discovery. *Nat Rev Drug Discov*, 12(5):371–387, may 2013.

[113] Martin Blaser. Stop the killing of beneficial bacteria. pages 7–8, 1940.

[114] San Francisco Medical, San Francisco, and San Francisco. Curtailing antibiotic use in agriculture. 176(January):9–11, 2002.

[115] Qiu Ying Lau, Yoke Yan Fion Tan, Vanessa Chai Yin Goh, David Jing Qin Lee, Fui Mee Ng, Esther H Q Ong, Jeffrey Hill, and Cheng San Brian Chia. An FDA-Drug Library Screen for Compounds with Bioactivities against Meticillin-Resistant Staphylococcus aureus (MRSA). *Antibiotics*, 4(4):424–434, dec 2015.

[116] Ramy K Aziz and Victor Nizet. Pathogen microevolution in high resolution. *Science translational medicine*, 2(16):16ps4, January 2010.

[117] Shailesh V Date, Zora Modrusan, Michael Lawrence, J Hiroshi Morisaki, Karen Toy, Ishita M Shah, Janice Kim, Summer Park, Min Xu, Li Basuino, Liana Chan, Deborah Zeitschel, Henry F Chambers, Man-wah Tan, Eric J Brown, Binh An Diep, and Wouter L W Hazenbos. Global Gene Expression of Methicillin-resistant Staphylococcus aureus USA300 During Human and Mouse Infection. 209, 2014.

[118] R J Scheffler, S Colmer, H Tynan, a L Demain, and V P Gullo. Antimicrobials, drug discovery, and genome mining. *Applied microbiology and biotechnology*, 97(3):969–78, February 2013.

[119] Claire Chewapreecha, Pekka Marttinen, Nicholas J Croucher, Susannah J Salter, Simon R Harris, Alison E Mather, William P Hanage, David Goldblatt, Francois H Nosten, Claudia Turner, Paul Turner, Stephen D Bentley, and Julian Parkhill. Comprehensive identification of single nucleotide polymorphisms associated with beta-lactam resistance within pneumococcal mosaic genes. *PLoS genetics*, 10(8):e1004547, August 2014.

[120] Keiichi Hiramatsu, Teruyo Ito, Sae Tsubakishita, Takashi Sasaki, Fumihiko Takeuchi, Yuh Morimoto, Yuki Katayama, Miki Matsuo, Kyoko Kuwahara-Arai, Tomomi Hishinuma, and Tadashi Baba. Genomic Basis for Methicillin Resistance in Staphylococcus aureus. *Infection & chemotherapy*, 45(2):117–36, June 2013.

[121] Joseph Osmundson, Scott Dewell, and Seth a Darst. RNA-Seq reveals differential gene expression in Staphylococcus aureus with single-nucleotide resolution. *PloS one*, 8(10):e76572, jan 2013.

[122] Timothy D Read and Ruth C Massey. Characterizing the genetic basis of bacterial phenotypes using genome-wide association studies : a new direction for bacteriology. pages 1–11, 2014.

[123] a Holmes, G McAllister, P R McAdam, S Hsien Choi, K Girvan, a Robb, G Edwards, K Templeton, and J R Fitzgerald. Genome-wide single nucleotide polymorphism-based assay for high-resolution epidemiological analysis of the methicillin-resistant Staphylococcus aureus hospital clone EMRSA-15. *Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases*, July 2013.

[124] Matthew T G Holden, Jodi a Lindsay, Craig Corton, Michael a Quail, Joshua D Cockfield, Smriti Pathak, Rahul Batra, Julian Parkhill, Stephen D Bentley, and Jonathan D Edgeworth. Genome sequence of a recently emerged, highly transmissible, multi-antibiotic- and antiseptic-resistant variant of methicillin-resistant Staphylococcus aureus, sequence type 239 (TW). *Journal of bacteriology*, 192(3):888–92, February 2010.

[125] Maisem Laabei, Mario Recker, Justine K Rudkin, Mona Aldeljawi, Zeynep Gulay, Tim J Sloan, Paul Williams, Jennifer L Endres, Kenneth W Bayles, Paul D Fey, Vijaya Kumar Yajjala, Todd Widhelm, Erica Hawkins, Katie Lewis, Sara Parfett, Lucy Scowen, Sharon J Peacock, Matthew Holden, Daniel Wilson, Timothy D Read, Jean Van Den Elsen, Nicholas K Priest, Edward J Feil, Laurence D Hurst, Elisabet Josefsson, and Ruth C Massey. Predicting the virulence of MRSA from its genome sequence. pages 1–12, 2014.

[126] Nicholas J Loman, Raju V Misra, Timothy J Dallman, Chrystala Constantinidou, Saheer E Gharbia, John Wain, and Mark J Pallen. Performance comparison of benchtop high-throughput sequencing platforms. *Nature biotechnology*, 30(5):434–9, May 2012.

[127] Monya Baker. De novo genome assembly : what every biologist should know.

[128] Phillip E C Compeau, Pavel A Pevzner, and Glenn Tesler. primer How to apply de Bruijn graphs to genome assembly. *Nature Biotechnology*, 29(11):987–991, 2011.

[129] Mark J P Chaisson, Richard K Wilson, and Evan E Eichler. Genetic variation and the de novo assembly of human genomes. *Nature Publishing Group*, 16(11):627–640, 2015.

[130] Can Alkan, Saba Sajjadian, and Evan E Eichler. Limitations of next-generation genome sequence assembly. *Nature methods*, 8(1):61–65, jan 2011.

[131] L U Bingxin, Zeng Zhenbing, and S H I Tieliu. Comparative study of de novo assembly and genome-guided assembly strategies for transcriptome reconstruction based on RNA-Seq. 56(2):143–155, 2013.

[132] CLSI. Performance Standards for Antimicrobial Susceptibility; Twenty-Fifth Informational Supplement. *CLSI document M100-S25. Wayne, PA: Clinical and Laboratory Standards Institute; 2015*, (2):313–322, jun.

[133] Edward J. Torok M Este Holden Matthew T G Harris, Simon R., Nicholas M Brown, Amanda L Ogilvy-Stuart, Matthew J Ellington, Michael a Quail, Stephen D Bentley, Julian Parkhill, and Sharon J Peacock. Whole-genome sequencing for analysis of an outbreak of meticillin-resistant Staphylococcus aureus: a descriptive study. *The Lancet infectious diseases*, 13(2):130–6, 2013.

[134]

[135] Matthew T G Holden, Jodi a Lindsay, Craig Corton, Michael a Quail, Joshua D Cockfield, Smriti Pathak, Rahul Batra, Julian Parkhill, Stephen D Bentley, and Jonathan D Edgeworth. Genome sequence of a recently emerged, highly transmissible, multi-antibiotic- and antiseptic-resistant variant of methicillin-resistant Staphylococcus aureus, sequence type 239 (TW). *Journal of bacteriology*, 192(3):888–92, February 2010.

[136] Matthew T G Holden, Li-yang Hsu, Kevin Kurt, Lucy A Weinert, Alison E Mather, Simon R Harris, Birgit Strommenger, Franziska Layer, Wolfgang Witte, Herminia De Lencastre, Robert Skov, Henrik Westh, Jonathan Edgeworth, Ian Gould, Vanya Gant, Jonathan Cooke, Giles F Edwards, Paul R Mcadam, Kate E Templeton, Angela Mccann, Edward J Feil, Lyndsey O Hudson, Zhemin Zhou, Santiago Castillo-ramÄś, Mark C Enright, Francois Balloux, David M Aanensen, Brian G Spratt, J Ross Fitzgerald, Julian Parkhill, Mark Achtman, and Stephen D Bentley. A genomic portrait of the emergence , evolution , and global spread of a methicillin-resistant Staphylococcus aureus pandemic. pages 653–664, 2013.

[137] Md Tauqeer Alam, Robert A Petit, Emily K Crispell, Timothy A Thornton, Karen N Conneely, Yunxuan Jiang, Sarah W Satola, and Timothy D Read. Dissecting Vancomycin-Intermediate Resistance in Staphylococcus aureus Using Genome-Wide Association. *Genome Biology and Evolution*, 6(5):1174–1185, may 2014.

[138] a Holmes, G McAllister, P R McAdam, S Hsien Choi, K Girvan, a Robb, G Edwards, K Templeton, and J R Fitzgerald. Genome-wide single nucleotide polymorphism-based assay for high-resolution epidemiological analysis of the methicillin-resistant Staphylococcus aureus hospital clone EMRSA-15. *Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases*, July 2013.

[139] Zheng Wang, Haokui Zhou, Hui Wang, Hongbin Chen, K K Leung, Stephen Tsui, and Margaret Ip. Comparative genomics of methicillin-resistant Staphylococcus aureus ST239: distinct geographical variants in Beijing and Hong Kong. *BMC Genomics*, 15(1):529, jun 2014.

[140] Pope B. J. Edwards, D. J. and K. E. Holt. RedDog: comparative analysis pipeline for large numbers of bacterial isolates using high-throughput sequences. *In Preparation*, 2015.

[141] Heng Li and Richard Durbin. Fast and accurate short read alignment with burrowsâĂŞwheeler transform. *Bioinformatics*, 25(14):1754, 2009.

[142] Mark a DePristo, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony a Philippakis, Guillermo del Angel, Manuel a Rivas, Matt Hanna, Aaron McKenna, Tim J Fennell, Andrew M Kernytsky, Andrey Y Sivachenko, Kristian Cibulskis, Stacey B Gabriel, David Altshuler, and Mark J Daly. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics*, 43(5):491–8, May 2011.

[143] Nathan Mantel. The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27(2 Part 1):209–220, 1967.

[144] Caroline M Nievergelt, Ondrej Libiger, and Nicholas J Schork. Generalized analysis of molecular variance. *PLoS genetics*, 3(4):e51, apr 2007.

[145] Jessica Carrière, Nicolas Barnich, and Hang Thi Thu Nguyen. *Exosomes: From Functions in Host-Pathogen Interactions and Immunity to Diagnostic and Therapeutic Opportunities*, pages 1–37. Springer Berlin Heidelberg, Berlin, Heidelberg.

[146] Korbinian Schneeberger, Jörg Hagmann, Stephan Ossowski, Norman Warthmann, Sandra Gesing, Oliver Kohlbacher, and Detlef Weigel. Simultaneous alignment of short reads against multiple genomes. *Genome Biology*, 10(9):R98, 2009.

[147] Robert C. Edgar. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797, 2004.

[148] Na Zhao, Chia-Chen Liu, Wenhui Qiao, and Guojun Bu. Apolipoprotein E, Receptors, and Modulation of Alzheimer's Disease. *Biological Psychiatry*, apr 2017.

[149] Anil Kumar, Arti Singh, and Ekavali. A review on Alzheimer's disease pathophysiology and its management: an update. *Pharmacological Reports*, 67(2):195–203, apr 2015.

[150] Perry G Ridge, Kaitlyn B Hoyt, Kevin Boehme, Shubhabrata Mukherjee, Paul K Crane, Jonathan L Haines, Richard Mayeux, Lindsay A Farrer, Margaret A Pericak-vance, Gerard D Schellenberg, John S K Kauwe, Disease Genetics, and Consortium Adgc. Neurobiology of Aging Assessment of the genetic variance of late-onset Alzheimer's disease. 41, 2016.

[151] Perry G Ridge, Mark T W Ebbert, and John S K Kauwe. Genetics of Alzheimer's Disease. 2013, 2013.

[152] Caroline Van Cauwenberghe, Christine Van Broeckhoven, and Kristel Sleegers. The genetic landscape of Alzheimer disease: clinical implications and perspectives, may 2016.

[153] Maurílio De Souza Cazarim I, Julio Cesar, Moriguti Ii, Abayomi Tolulope, Ogunjimi Iii, Leonardo Régis, and Leira Pereira. Perspectives for treating Alzheimer's disease A review on promising pharmacological substances Perspectives. 134(4), 2016.

[154] Yuetiva Deming, Zeran Li, Manav Kapoor, Oscar Harari, Jorge L Del-Aguila, Kathleen Black, David Carrell, Yefei Cai, Maria Victoria Fernandez, John Budde, Shengmei Ma, Benjamin Saef, Bill Howells, Kuan-lin Huang, Sarah Bertelsen, Anne M Fagan, David M Holtzman, John C Morris, Sungeun Kim, Andrew J Saykin, Philip L De Jager, Marilyn Albert, Abhay Moghekar, Richard O'Brien, Matthias Riemenschneider, Ronald C Petersen, Kaj Blennow, Henrik Zetterberg, Lennart Minthon, Vivianna M Van Deerlin, Virginia Man-Yee Lee, Leslie M Shaw, John Q Trojanowski, Gerard Schellenberg, Jonathan L Haines, Richard Mayeux, Margaret A Pericak-Vance, Lindsay A Farrer, Elaine R Peskind, Ge Li, Antonio F Di Narzo, John S K Kauwe, Alison M Goate, and Carlos Cruchaga. Genome-wide association study identifies four novel loci associated with Alzheimer's endophenotypes and disease modifiers. *Acta Neuropathologica*, 133(5):839–856, 2017.

[155] Rahul S Desikan, Chun Chieh Fan, Yunpeng Wang, Andrew J Schork, Howard J Cabral, L Adrienne Cupples, Wesley K Thompson, Lilah Besser, Walter A Kukull, Dominic Holland, Chi-Hua Chen, James B Brewer, David S Karow, Karolina Kauppi, Aree Witoelar, Celeste M Karch, Luke W Bonham, Jennifer S Yokoyama, Howard J Rosen, Bruce L Miller, William P Dillon, David M Wilson, Christopher P Hess, Margaret Pericak-Vance, Jonathan L Haines, Lindsay A Farrer, Richard Mayeux, John Hardy, Alison M Goate, Bradley T Hyman, Gerard D Schellenberg, Linda K McEvoy, Ole A Andreassen, and Anders M Dale. Genetic assessment of age-associated Alzheimer disease risk: Development and validation of a polygenic hazard score. *PLOS Medicine*, 14(3):e1002258, mar 2017.

[156] Jean-Charles Lambert, Carla A Ibrahim-Verbaas, Denise Harold, Adam C Naj, Rebecca Sims, Celine Bellenguez, Gyungah Jun, Anita L DeStefano, Joshua C Bis, Gary W Beecham, Benjamin Grenier-Boley, Giancarlo Russo, Tricia A Thornton-Wells, Nicola Jones, Albert V Smith, Vincent Chouraki, Charlene Thomas, M Arfan Ikram, Diana Zelenika, Badri N Vardarajan, Yoichiro Kamatani, Chiao-Feng Lin, Amy Gerrish, Helena Schmidt, Brian Kunkle, Melanie L Dunstan, Agustin Ruiz, Marie-Therese Bihoreau, Seung-Hoan Choi, Christiane Reitz, Florence Pasquier, Paul Hollingworth, Alfredo Ramirez, Olivier Hanon, Annette L Fitzpatrick, Joseph D Buxbaum, Dominique Campion, Paul K Crane, Clinton Baldwin, Tim Becker, Vilmundur Gudnason, Carlos Cruchaga, David Craig, Najaf Amin, Claudine Berr, Oscar L Lopez, Philip L De Jager, Vincent Deramecourt, Janet A Johnston, Denis Evans, Simon Lovestone, Luc Letenneur, Francisco J Moron, David C Rubinsztein, Gudny Eiriksdottir, Kristel Sleegers, Alison M Goate, Nathalie Fievet, Matthew J Huentel-

man, Michael Gill, Kristelle Brown, M Ilyas Kamboh, Lina Keller, Pascale Barberger-Gateau, Bernadette McGuinness, Eric B Larson, Robert Green, Amanda J Myers, Carole Dufouil, Stephen Todd, David Wallon, Seth Love, Ekaterina Rogaeva, John Gallacher, Peter St George-Hyslop, Jordi Clarimon, Alberto Lleo, Anthony Bayer, Debby W Tsuang, Lei Yu, Magda Tsolaki, Paola Bossu, Gianfranco Spalletta, Petroula Proitsi, John Collinge, Sandro Sorbi, Florentino Sanchez-Garcia, Nick C Fox, John Hardy, Maria Candida Deniz Naranjo, Paolo Bosco, Robert Clarke, Carol Brayne, Daniela Galimberti, Michelangelo Mancuso, Fiona Matthews, European Alzheimer's Disease Initiative (EADI), Genetic (GERAD), Environmental Risk in Alzheimer's Disease, Alzheimer's Disease Genetic Consortium (ADGC), Cohorts for Heart (CHARGE), Aging Research in Genomic Epidemiology, Susanne Moebus, Patrizia Mecocci, Maria Del Zompo, Wolfgang Maier, Harald Hampel, Alberto Pilotto, Maria Bullido, Francesco Panza, Paolo Caffarra, Benedetta Nacmias, John R Gilbert, Manuel Mayhaus, Lars Lannfelt, Hakon Hakonarson, Sabrina Pichler, Minerva M Carrasquillo, Martin Ingelsson, Duane Beekly, Victoria Alvarez, Fanggeng Zou, Otto Valladares, Steven G Younkin, Eliecer Coto, Kara L Hamilton-Nelson, Wei Gu, Cristina Razquin, Pau Pastor, Ignacio Mateo, Michael J Owen, Kelley M Faber, Palmi V Jonsson, Onofre Combarros, Michael C O'Donovan, Laura B Cantwell, Hilkka Soininen, Deborah Blacker, Simon Mead, Thomas H Mosley Jr, David A Bennett, Tamara B Harris, Laura Fratiglioni, Clive Holmes, Renee F A G de Bruijn, Peter Passmore, Thomas J Montine, Karolien Bettens, Jerome I Rotter, Alexis Brice, Kevin Morgan, Tatiana M Foroud, Walter A Kukull, Didier Hannequin, John F Powell, Michael A Nalls, Karen Ritchie, Kathryn L Lunetta, John S K Kauwe, Eric Boerwinkle, Matthias Riemenschneider, Merce Boada, Mikko Hiltunen, Eden R Martin, Reinhold Schmidt, Dan Rujescu, Li-San Wang, Jean-Francois Dartigues, Richard Mayeux, Christophe Tzourio, Albert Hofman, Markus M Nothen, Caroline Graff, Bruce M Psaty, Lesley Jones, Jonathan L Haines, Peter A Holmans, Mark Lathrop, Margaret A Pericak-Vance, Lenore J Launer, Lindsay A Farrer, Cornelia M van Duijn, Christine Van Broeckhoven, Valentina Moskvina, Sudha Seshadri, Julie Williams, Gerard D Schellenberg, and Philippe Amouyel. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat Genet*, 45(12):1452–1458, dec 2013.

[157] Johanna Jakobsdottir, Sven J van der Lee, Joshua C Bis, Vincent Chouraki, David Li-Kroeger, Shinya Yamamoto, Megan L Grove, Adam Naj, Maria Vronskaya, Jose L Salazar, Anita L DeStefano, Jennifer A Brody, Albert V Smith, Najaf Amin, Rebecca Sims, Carla A Ibrahim-Verbaas, Seung-Hoan Choi, Claudia L Satizabal, Oscar L Lopez, Alexa Beiser, M Arfan Ikram, Melissa E Garcia, Caroline Hayward, Tibor V Varga, Samuli Ripatti, Paul W Franks, Göran Hallmans, Olov Rolandsson, Jan-Håkon Jansson, David J Porteous, Veikko Salomaa, Gudny Eiriksdottir, Kenneth M Rice, Hugo J Bellen, Daniel Levy, Andre G Uitterlinden, Valur Emilsson, Jerome I Rotter, Thor As-

pelund, Cohorts for Heart Consortium, Aging Research in Genomic Epidemiology, Alzheimer's Disease Genetic Consortium, Genetic Consortium, Environmental Risk in Alzheimer's Disease, Christopher J O'Donnell, Annette L Fitzpatrick, Lenore J Launer, Albert Hofman, Li-San Wang, Julie Williams, Gerard D Schellenberg, Eric Boerwinkle, Bruce M Psaty, Sudha Seshadri, Joshua M Shulman, Vilmundur Gudnason, and Cornelia M van Duijn. Rare Functional Variant in TM2D3 is Associated with Late-Onset Alzheimer's Disease. *PLOS Genetics*, 12(10):e1006327, oct 2016.

[158] Adam C Naj, Gyungah Jun, Gary W Beecham, Li-San Wang, Badri Narayan Vardarajan, Jacqueline Buros, Paul J Gallins, Joseph D Buxbaum, Gail P Jarvik, Paul K Crane, Eric B Larson, Thomas D Bird, Bradley F Boeve, Neill R Graff-Radford, Philip L De Jager, Denis Evans, Julie A Schneider, Minerva M Carrasquillo, Nilufer Ertekin-Taner, Steven G Younkin, Carlos Cruchaga, John S K Kauwe, Petra Nowotny, Patricia Kramer, John Hardy, Matthew J Huentelman, Amanda J Myers, Michael M Barmada, F Yesim Demirci, Clinton T Baldwin, Robert C Green, Ekaterina Rogaeva, Peter St George-Hyslop, Steven E Arnold, Robert Barber, Thomas Beach, Eileen H Bigio, James D Bowen, Adam Boxer, James R Burke, Nigel J Cairns, Chris S Carlson, Regina M Carney, Steven L Carroll, Helena C Chui, David G Clark, Jason Corneveaux, Carl W Cotman, Jeffrey L Cummings, Charles DeCarli, Steven T DeKosky, Ramon Diaz-Arrastia, Malcolm Dick, Dennis W Dickson, William G Ellis, Kelley M Faber, Kenneth B Fallon, Martin R Farlow, Steven Ferris, Matthew P Frosch, Douglas R Galasko, Mary Ganguli, Marla Gearing, Daniel H Geschwind, Bernardino Ghetti, John R Gilbert, Sid Gilman, Bruno Giordani, Jonathan D Glass, John H Growdon, Ronald L Hamilton, Lindy E Harrell, Elizabeth Head, Lawrence S Honig, Christine M Hulette, Bradley T Hyman, Gregory A Jicha, Lee-Way Jin, Nancy Johnson, Jason Karlawish, Anna Karydas, Jeffrey A Kaye, Ronald Kim, Edward H Koo, Neil W Kowall, James J Lah, Allan I Levey, Andrew P Lieberman, Oscar L Lopez, Wendy J Mack, Daniel C Marson, Frank Martiniuk, Deborah C Mash, Eliezer Masliah, Wayne C McCormick, Susan M McCurry, Andrew N McDavid, Ann C McKee, Marsel Mesulam, Bruce L Miller, Carol A Miller, Joshua W Miller, Joseph E Parisi, Daniel P Perl, Elaine Peskind, Ronald C Petersen, Wayne W Poon, Joseph F Quinn, Ruchita A Rajbhandary, Murray Raskind, Barry Reisberg, John M Ringman, Erik D Roberson, Roger N Rosenberg, Mary Sano, Lon S Schneider, William Seeley, Michael L Shelanski, Michael A Slifer, Charles D Smith, Joshua A Sonnen, Salvatore Spina, Robert A Stern, Rudolph E Tanzi, John Q Trojanowski, Juan C Troncoso, Vivianna M Van Deerlin, Harry V Vinters, Jean Paul Vonsattel, Sandra Weintraub, Kathleen A Welsh-Bohmer, Jennifer Williamson, Randall L Woltjer, Laura B Cantwell, Beth A Dombroski, Duane Beekly, Kathryn L Lunetta, Eden R Martin, M Ilyas Kamboh, Andrew J Saykin, Eric M Reiman, David A Bennett, John C Morris, Thomas J Montine, Alison M Goate, Deborah Blacker, Debby W Tsuang, Hakon Hakonarson, Wal-

ter A Kukull, Tatiana M Foroud, Jonathan L Haines, Richard Mayeux, Margaret A Pericak-Vance, Lindsay A Farrer, and Gerard D Schellenberg. Common variants in MS4A4/MS4A6E, CD2uAP, CD33, and EPHA1 are associated with late-onset Alzheimer's disease. *Nature genetics*, 43(5):436–441, may 2011.

[159] P M Visscher, Matthew A Brown, M I McCarthy, and J Yang. Five years of GWAS discovery. *American Journal of Human Genetics*, 90(1):7–24, jan 2012.

[160] Samantha L Rosenthal and M Ilyas Kamboh. Late-Onset Alzheimer's Disease Genes and the Potentially Implicated Pathways. *Current Genetic Medicine Reports*, 2(2):85–101, mar 2014.

[161] Eric R Gamazon, Heather E Wheeler, Kaanan P Shah, Sahar V Mozaffari, Keston Aquino-Michaels, Robert J Carroll, Anne E Eyler, Joshua C Denny, GTEx Consortium, Dan L Nicolae, Nancy J Cox, and Hae Kyung Im. A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet*, 47(9):1091–1098, sep 2015.

[162] Zhihong Zhu, Futao Zhang, Han Hu, Andrew Bakshi, Matthew R Robinson, Joseph E Powell, Grant W Montgomery, Michael E Goddard, Naomi R Wray, Peter M Visscher, and Jian Yang. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet*, 48(5):481–487, may 2016.

[163] The Telomeres Mendelian Randomization Collaboration. Association between telomere length and risk of cancer and non-neoplastic diseases: A mendelian randomization study. *JAMA Oncology*, feb 2017.

[164] Gibran Hemani, Jie Zheng, Kaitlin H Wade, Charles Laurin, Benjamin Elsworth, Stephen Burgess, Jack Bowden, Ryan Langdon, Vanessa Tan, James Yarmolinsky, Hashem A Shihab, Nicholas Timpson, David M Evans, Caroline Relton, Richard M Martin, George Davey Smith, Tom R Gaunt, and Philip C and Haycock. MR-Base: a platform for systematic causal inference across the phenome using billions of genetic associations. *bioRxiv*, dec 2016.

[165] Bogdan Pasaniuc and Alkes L Price. Dissecting the genetics of complex traits using summary association statistics. *Nat Rev Genet*, 18(2):117–127, feb 2017.

[166] Philip C Haycock, Stephen Burgess, Kaitlin H Wade, Jack Bowden, Caroline Relton, and George Davey Smith. Best (but oft-forgotten) practices: the design, analysis, and interpretation of Mendelian randomization studies. *The American Journal of Clinical Nutrition*, 103(4):965–978, apr 2016.

[167] Brandon L Pierce and Stephen Burgess. Efficient Design for Mendelian Randomization Studies: Subsample and 2-Sample Instrumental Variable Estimators. *American Journal of Epidemiology*, 178(7):1177–1184, oct 2013.

[168] John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, Barbara Foster, Mike Moser, Ellen Karasik, Bryan Gillard, Kimberley Ramsey, Susan Sullivan, Jason Bridge, Harold Magazine, John Syron, Johnelle Fleming, Laura Siminoff, Heather Traino, Maghboeba Mosavel, Laura Barker, Scott Jewell, Dan Rohrer, Dan Maxim, Dana Filkins, Philip Harbach, Eddie Cortadillo, Bree Berghuis, Lisa Turner, Eric Hudson, Kristin Feenstra, Leslie Sobin, James Robb, Phillip Branton, Greg Korzeniewski, Charles Shive, David Tabor, Liqun Qi, Kevin Groch, Sreenath Nampally, Steve Buia, Angela Zimmerman, Anna Smith, Robin Burges, Karna Robinson, Kim Valentino, Deborah Bradbury, Mark Cosentino, Norma Diaz-Mayoral, Mary Kennedy, Theresa Engel, Penelope Williams, Kenyon Erickson, Kristin Ardlie, Wendy Winckler, Gad Getz, David DeLuca, Daniel MacArthur, Manolis Kellis, Alexander Thomson, Taylor Young, Ellen Gelfand, Molly Donovan, Yan Meng, George Grant, Deborah Mash, Yvonne Marcus, Margaret Basile, Jun Liu, Jun Zhu, Zhidong Tu, Nancy J Cox, Dan L Nicolae, Eric R Gamazon, Hae Kyung Im, Anuar Konkashbaev, Jonathan Pritchard, Matthew Stevens, Timothee Flutre, Xiaoquan Wen, Emmanouil T Dermitzakis, Tuuli Lappalainen, Roderic Guigo, Jean Monlong, Michael Sammeth, Daphne Koller, Alexis Battle, Sara Mostafavi, Mark McCarthy, Manual Rivas, Julian Maller, Ivan Rusyn, Andrew Nobel, Fred Wright, Andrey Shabalin, Mike Feolo, Nataliya Sharopova, Anne Sturcke, Justin Paschal, James M Anderson, Elizabeth L Wilder, Leslie K Derr, Eric D Green, Jeffery P Struewing, Gary Temple, Simona Volpi, Joy T Boyer, Elizabeth J Thomson, Mark S Guyer, Cathy Ng, Assya Abdallah, Deborah Colantuoni, Thomas R Insel, Susan E Koester, A Roger Little, Patrick K Bender, Thomas Lehner, Yin Yao, Carolyn C Compton, Jimmie B Vaught, Sherilyn Sawyer, Nicole C Lockhart, Joanne Demchok, and Helen F Moore. The Genotype-Tissue Expression (GTEx) project. *Nat Genet*, 45(6):580–585, jun 2013.

[169] Buhm Han and Eleazar Eskin. Random-Effects Model Aimed at Discovering Associations in Meta-Analysis of Genome-wide Association Studies. *The American Journal of Human Genetics*, 88(5):586–598, apr 2011.

[170] Marina Boccardi, Martina Bocchetta, Rossana Ganzola, Nicolas Robitaille, Alberto Redolfi, Simon Duchesne, Clifford R Jack Jr., Giovanni B Frisoni, George Bartzokis, John G Csernansky, Mony J de Leon, Leyla DeToledo-Morrell, Ronald J Killiany, Stéphane Lehéricy, Nikolai Malykhin, Johannes Pantel, Jens C Pruessner, Hilkka Soininen, and Craig Watson. Operationalizing protocol differences for EADC-ADNI manual hippocampal segmentation. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, 11(2):184–194, apr 2017.

[171] Lon S Schneider, Francesca Mangialasche, Niels Andreasen, Howard Feldman, Ezio Giacobini, Roy Jones, Valentina Mantua, Patrizia Mecocci, Luca Pani,

Bengt Winblad, and Miia Kivipelto. Clinical trials and late-stage drug development for Alzheimer's disease: an appraisal from 1984 to 2014. *Journal of internal medicine*, 275(3):251–283, mar 2014.

[172] Schott JM Barnes J, Bartlett JW, Fox NC. Targeted recruitment using cerebrospinal fluid biomarkers: Implications for Alzheimer's disease therapeutic trials. *Journal of Alzheimer's Disease*, 34(2):431–7, 2013.

[173] Le Cong, F Ann Ran, David Cox, Shuailiang Lin, Robert Barretto, Naomi Habib, Patrick D Hsu, Xuebing Wu, Wenyan Jiang, Luciano A Marraffini, and Feng Zhang. Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science*, 339(6121):819 LP – 823, feb 2013.