Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____	_____
Matthew Ezewudo	Date

Genomics studies of population structure and evolution in *Neisseria gonorrhoeae*

By

Matthew Ezewudo
Doctor of Philosophy

Graduate Division of Biological and Biomedical Science

Population Biology, Ecology and Evolution

_____

Timothy D. Read, PhD
Advisor

_____

Michael E. Zwick, PhD
Advisor

_____

William M. Shafer, PhD
Committee Member

_____

David Cutler, PhD
Committee Member

_____

Karen N. Conneely, PhD
Committee Member

_____

Thomas Wingo, MD
Committee Member

Accepted:

_____

Lisa A. Tedesco, PhD

Dean of the James T. Laney School of Graduate Studies

_____

Date

Genomics studies of population structure and evolution in *Neisseria gonorrhoeae*

By

Matthew Ezewudo

M.S., Georgia State University, 2010

Co-Advisor: Timothy D. Read, PhD

Co-Advisor: Michael E. Zwick, PhD

An abstract of

A dissertation submitted to the Faculty of the

James T. Laney School of Graduate Studies of Emory University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Graduate Division of Biological and Biomedical Sciences

Population Biology, Ecology and Evolution

2015

Abstract

Genomics studies of population structure and evolution in *Neisseria gonorrhoeae*

By

Matthew Ezewudo

Improvements in whole genome sequencing technology have created opportunities to find answers to questions relating to evolution within biological populations and the underlying genetic architecture of phenotypes of interest to researchers. This dissertation is aimed at seizing this development, to both develop and refine bioinformatics tools that could interpret high throughput data from next generation sequencing platforms to answer both classical genetics questions and to better characterize microbial populations in the emerging microbial genomics field.

In this body of work, we analyzed genome-wide sequencing data of *Neisseria gonorrhoeae* isolates from across the globe to make inferences as to the nature of evolutionary forces prevalent in the pathogen population. *N. gonorrhoeae* causes gonorrhea, a sexually transmitted infection of public health relevance because the pathogen has shown the ability to evolve resistance to most known antibiotic drugs. It is believed that the transformative nature of *N. gonorrhoeae* underlies this ability. Our analysis suggests an appreciable effect of recombination within its population, and also reconfirmed the presence of previously described horizontally transferred resistance determinants in strains resistant to third generation cephalosporin.

We pointed out some limitations in using the more widespread current sequencing technology platforms to make accurate inferences about the nature of whole genome data. There is also the need for a broader sample set than the collection we assembled, to further characterize this pathogen and similar microbial populations.

Genomics studies of population structure and evolution in *Neisseria gonorrhoeae*

By

Matthew Ezewudo

M.S., Georgia State University, 2010

Co-Advisor: Timothy D. Read, PhD

Co-Advisor: Michael E. Zwick, PhD

A dissertation submitted to the Faculty of the

James T. Laney School of Graduate Studies of Emory University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Graduate Division of Biological and Biomedical Sciences

Population Biology, Ecology and Evolution

2015

Acknowledgements

O na-abu ihe kwuru, ihe a kwudebe ya

To my mother Appolonia Ezewudo; thanks for being my inspiration.

To Chuma and Ify and Ike; for keeping me grounded

To Buchi Osakwe; for challenging me to go all the way

To Ugo Ezeoke and the S4A group; for being family in the best sense

To my colleagues in the PBEE program at Emory; for the worthwhile journey this has been

To Sandeep Joseph, Robert Petit, Ben Rambo-Martin, Mfon Udoh, Chloe Robins, Kajari Mondal, Dahnya Ramachandran, Anne Dodd, Shoshanna Lee, Alex Kotlar; for all the support and assistance in my hours of need

To Bill Shafer, Karen Conneely, Dave Cutler and Thomas Wingo for being my guides through all this

To Tim Read and Mike Zwick for being my teachers and my advisors

I owe everything to you.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

My Ph.D. thesis research is focused at the intersection of the study of genetic variation within bacterial populations and the development and application of next generation sequencing analysis tools. The ultimate goal of my work has been to use the wealth of information obtained from genome sequencing in order to gain insight, and ultimately solve, significant public health challenges. My work has had two main foci. The first aim is to use existing next-generation sequencing to identify genetic variation within strains of bacterial that influence their impact on human disease. The second uses genomics-based approaches to analyze infectious disease pathogen populations. Both efforts are driven by recent developments in bacterial whole-genome sequencing capabilities and the opportunities provided by such genome-wide data sets.

In this chapter, I introduce my study system *Neisseria gonorrhoeae*, within the context of the disease it causes. Then I highlight concepts from bacteria evolutionary biology, and their connection to evolution of antibiotic resistance. Finally, I discuss how analyzing high throughput data from next generation sequencing platforms can improve inferences on the significance of genetic variants underlying various phenotypes of interest such as antibiotic resistance and bacterial virulence.

1.1 *Neisseria Gonorrhoeae:* Overview of pathogen and disease

Neisseria gonorrhoeae, also known as the gonococcus, is a Gram-negative bacterium

responsible for gonorrhea, a highly infectious sexually transmitted infection (Ohnishi et al., 2011). One of eleven members of the *Neisseria* genus, *N.gonorrhoeae*, along with *N. meningitidis* are the two *Neisseria* pathogenic species. *Neisseria* belongs to the very diverse bacteria phylum known as Proteobacteria, which includes a variety of pathogens such as *Escherichia*, *Salmonella*, *Vibrio* and *Helicobacter*. An obligate pathogen, it infects only humans, causing urethritis in men and cervicitis in women.

Gonorrhea, also referred to as "the clap", is an ancient disease (with mentions in the bible; Old Testament; Leviticus 15:1-3) and it remains one of the most prevalent sexually transmitted infections. The first usage of the term "gonorrhea" was in the second century by the Roman physician Aelius Galenus, which implied the 'flow of seed', a description of the symptom of the disease caused by this pathogen. While there were sporadic observations of this bacterium through the millennia that followed, the German physician Albert Neisser, in 1879 first described the pathogen and gave its official name (Bergey & Breed, 1971). He demonstrated that it was present in to be human patients with symptoms. Furthermore he directly demonstrated that when discharges and *N. gonorrhoeae* culture were introduced to urethra of healthy men, they become infected.

According to World Health Organization (WHO) estimates, gonorrhea represent 106 million out of the estimated 498 million new cases of curable sexually transmitted infections globally every year (World Health Organization (WHO), 2012). Gonorrhea occurs in all regions of the globe, with the highest number of cases reported in the WHO Western Pacific Region (42 million cases) and the WHO South-East Asia Region (25.4

million) (Unemo & Shafer, 2014) Left untreated, gonorrhea is associated with high morbidity and negative socioeconomic consequences. In addition there is evidence that gonorrhea is associated with a five-fold higher risk of co-transmission of HIV in infected patients(Cohen et al., 1997; Ohnishi et al., 2011). While a majority of infected women do not display any clinical symptoms, only a small proportion of infected men are asymptomatic. If these infections remain undetected, *N. gonorrhoeae* can ascend to the upper genital tract and result in severe reproductive complications such as pelvic inflammatory diseases, penile edema, endometritis and ectopic pregnancies (Unemo & Shafer, 2014).

The only current therapeutic option for gonorrhea is the use of antimicrobial therapy, since there is no vaccine for prevention. We have had effective antibiotics since the discovery and mass production of penicillin during World War II (Aminov, 2010). However, the choice of therapies has dwindled over time due to the emergence of antimicrobial resistance to most of the classes of drugs previously used to treat the disease and the lack of development of newer antibiotics that could effectively check the pathogen (Ohnishi et al., 2011). *N. gonorrhoeae* is naturally competent for DNA transformation; it has developed an elaborate mechanism which allows it to take up naked DNA from other bacteria within its environment (Aas et al., 2002). This transformative ability is key both in the spread of infection by the pathogen and the evolution of antibiotic resistance within its population (Unemo & Shafer, 2011). The trend toward resistance to current drugs suggests the need for studies aimed at developing therapies based on new targets as well as understanding how resistance to existing

antibiotics develops. For both purposes, I will argue that complete genomic information about the species is required to meet this emerging public health crisis. Furthermore, this kind of knowledge can inform studies of the evolutionary dynamics of other bacteria populations.

1.2 Bacteria evolution and population biology

Evolutionary processes act on genetic variations within the population to perpetuate more fit phenotypes within the context of the organism's environment. How such variations arise and are maintained within biological populations has been at the center of population genetics studies for the past century. After the 1950's, more attention was paid to bacteria and pathogenic populations as subjects of population genetics studies than was the case previously. Studies in this field have expanded, as approaches used by population and evolutionary biologists have increasingly converged with microbiologists, immunologists, clinicians and epidemiologists (Didelot, Bowden, Wilson, Peto, & Crook, 2012a; Levin, Lipsitch, & Bonhoeffer, 1999; Nelson, Whittam, & Selander 1991) to look at problems in infectious diseases. As is the case in eukaryotic populations, the major evolutionary forces affecting gene and variant frequency in bacterial populations include genetic drift and natural selection.

1.2.1 Genetic Drift

Genetic drift is the change in frequency of an allele in a population, caused by random sampling of variants from one generation to the next. The most common model of genetic drift draws from the assumptions of the works of Fisher and Wright (R. A. Fisher, 1922;

S. Wright, 1931). The model consists of an idealized population model, where random sampling of alleles from an finite-sized parental population will result in different proportions of alleles in the next generation. The model assumes that there is no mutation introducing new alleles or natural selection changing the frequency of existing alleles.

The overall effect of genetic drift is the removal of genetic variation from the population, at a rate that is inversely proportional to the population size per generation. Hence, drift is predicted to have a greater impact in smaller-sized populations. Because most bacteria populations are enormous (where infections may contain 10^6-10^{10} cells), genetic drift is expected to play a lesser role in the dynamics of allele frequencies over time (Batut, Knibbe, Marais, & Daubin, 2014). Alleles may be fixed or lost in a population as result of drift and for individual variants there is no directionality to the actions of drift. Were genetic drift the only evolutionary force at work in a finite-sized natural population, all variation would be lost. However, mutation acts to generate new alleles over time and add variation to the population. The balance of mutation and genetic drift on the frequency neutral alleles is elaborated in the neutral theory of molecular evolution proposed by Motoo Kimura (Kimura & Ohta, 1971). Neutral alleles do not affect the evolutionary fitness of the organism, and therefore their frequency should only be influenced by mutation and drift. Thus, Kimura suggests that most molecular evolution in a population results from the action of genetic drift that tends to remove variations and mutations that introduce variations in the first place.

At a minimum, the neutral theory provides a convenient null hypothesis that can be used to test for the departures that are inconsistent with the sole action of mutation and drift. The challenges to the neutral theory of molecular evolution focus largely on the role of natural selection in evolution. For example, the neutral theory predicts that organisms with shorter generation life spans should evolve faster than those with longer generation times. Yet there is strong evidence of similar ranges of heterozygosity across different species with vastly different generation times, which seems inconsistent with predictions from the neutral theory.

A refinement to the neutral theory, by Tomoko Ohta (Ohta, 1973), suggested that most amino-acid mutations are not neutral but are slightly deleterious, and this accounts for the observed deviations from the predictions of the neutral theory. While some precepts of the neutral theory have been called into question (Rocha & Feil, 2010), other works like those of Wagner et al(Wagner, 2008) and Lynch (Lynch, 2007) still suggest a significant role for neutral mutations for eventual adaptation and evolution in a population. The nucleotide composition of the bacterial genome can best be explained if all sites in the genome are under some sort of weak selection, and the idea of purely 'neutral sites' is considered more of an artificial construct (Rocha & Feil, 2010).

Recent widespread use of genomic approaches to study microbial evolution through measuring changes in microbial DNA suggests a pervasive role of positive selection in bacteria evolution (Alam et al., 2014; Joseph et al., 2012; Lefébure & Stanhope, 2009)

1.2.2 Natural Selection

Evolution by natural selection, which leads to adaptation in a particular environment, occurs through the perpetuation of variants underlying more fit phenotypes in a population. While evident everywhere, natural selection is quite difficult to observe in action due to the time scale required to bring about evolution in most eukaryotic organism. Bacteria populations on the other hand, with short generation times and often-difficult environmental forces that provide strong selection pressure, are ideal candidates to demonstrate the work of natural selection and selective adaptation (Dykhuizen, 1990). The classic work of Luria and Delbruck (Luria & Delbrück, 1943) pioneered the experimental evolution approach of studying evolution using bacteria populations and offered insights on the role of selection in evolution. Our newfound ability to generate genome sequence information spanning different prokaryotic organisms has created more opportunities to 'measure' evolution in these organisms (Biek, Pybus, Lloyd-Smith, & Didelot, 2015) by analyzing evolutionary rate parameters gleaned from these genomic data.

Natural selection in bacterial populations could be directional, with either positive or negative selection. Negative or purifying selection acts to remove deleterious alleles, which reduces variation in the population over time. This results in functionally important regions of the genome being more conserved compared to the rest of the genome (I. K. Jordan, Rogozin, Wolf, & Koonin, 2002) Positive selection conversely favors the survival and reproduction of individuals with variants that offer a competitive edge over the rest of the population. Antibiotic resistance in bacteria population is a

classic example of strong positive selection, and in situations called selective sweeps, positively selected variants completely replace the unselected variants across the population (Kaplan, Hudson, & Langley).

Within bacteria, mutations and recombination through horizontal gene transfers are key agents that introduce and reassort variation within the population (Biek et al., 2015). The balance between these processes underlies how natural selection operates in microbial populations.

1.2.2.1 *De novo* point mutations and bacteria evolution

Point mutations are considered the basic material for evolution in bacteria. These changes normally involve substituting one nucleotide base with another on the genome sequence to create Single Nucleotide Variants (SNVs). If a variant increases in frequency above an arbitrary threshold (like 1% or 5%), then that variant may be called a Single Nucleotide Polymorphism (SNP). A synonymous nucleotide change is one that occurs in the coding sequencing of a gene but does not change the encoded amino acid in the protein. This class of variation is generally considered to be "neutral," having little or no effect on biological functions and is therefore invisible to natural selection. This is likely to be an oversimplification (Bailey, Hinz, & Kassen, 2014; Akashi, 1994; Akashi et al., 2006) but I will use this approximation in my own work. Conversely, nucleotide changes on protein coding sequences that lead to change in amino acid encoded in the codon are said to be non-synonymous. Single nucleotide inserts or deletions within a protein coding sequence will lead to a frame shift and the downstream codons will be interpreted differently,

leading to expression of altered amino-acid sequence. Point mutations in non-coding regions of the genome are usually thought of as neutral although they could be significant if they alter a regulatory element (Bryant, Chewapreecha, & Bentley, 2012; Warner, Folster, Shafer, & Jerse, 2007). Non-synonymous changes are of consequence in evolution and could serve as basis for tests for positive selection within a population(McDonald & Kreitman, 1991).

Applying the neutral theory, the number of substitutions found on genome sequences of individuals within a population given a mutation rate for that population could be used as a measure of time - a molecular clock (S. Kumar, 2005) that has elapsed over evolutionary history of the population when traced back to previous common ancestor. This measure serves as a basis to create phylogenetic relationships between individuals sampled from a bacteria population (Didelot & Falush, 2007).

Most non-synonymous mutations are likely to be deleterious and so purifying selection holds the level of spontaneous substitutions to a lower level in bacteria populations (Drake, 1991). In certain instances, as demonstrated experimentally by Sniegowski *et al* (Sniegowski, Gerrish, & Lenski, 1997) using *E.coli* species, strains that have mutator alleles arise to high frequency within the population in association with an adaptive mutation, when the selective cost of the mutator is less than the selective benefit of the adaptive mutation. When faced with novel challenging environments higher *de novo* mutation rates often confer a selective advantage(Tanaka, Bergstrom, & Levin, 2003).

The advent of next-generation sequencing and the wealth of whole genome information it has brought make it feasible to perform deep genome investigations into variations between individuals in a species or clone. Using next-generation sequencing techniques, one could observe on a genomic scale, the function, rate, type and distributions of point mutations in individuals in a bacteria population and draw insights into the evolutionary pressures acting in the population as a result of these (Bryant et al., 2012).

Previously in bacterial functional studies, point mutations underlying phenotypes of interest were identified mostly using laboratory approaches such as complementation analysis (Wassenaar & Gaastra, 2001). Currently, whole-genome sequencing (WGS) and analysis allows for unbiased direct identification of variants that underlie phenotypes, speeding up understanding of the evolution of virulence and antibiotic resistance in pathogens, as a number of recent studies have demonstrated. For example, in the studies performed by Renzoni *et al*, WGS was used to identify SNPs in two isogenic strains of the bacteria pathogen *Staphylococcus aureus;* a parental strain, and a derivative strain that had undergone stepwise *in vitro* selection for resistance to the antibiotic teicoplanin. They identified four SNP differences in three loci that were confirmed experimentally to underlie the resistance phenotype as well as fitness of the derivative strain (Renzoni et al., 2011). Similarly, Cosmas et al in their studies on *Mycobacterium tuberculosis* (Comas et al., 2012) reported the discovery of SNPs that compensate for cost of fitness associated with resistance to the antibiotic rifampicin. They carried out WGS on both laboratory evolved rifampicin isolates and clinical isolates, to understand the difference in fitness between these two sets of strain samples. Mapping the variants detected, to a reference

genome of *M. tuberculosis,* they identified 11 nonsynonymous SNPs from the clinical isolates, suggesting other genomic variations may have arisen outside of the timeline and environment of laboratory experiments to confer fitness advantages to these isolates. The work done by Tomasz's group (Mwangi et al., 2007) demonstrated how the strong selection for resistance to antibiotics could lead to point mutations being fixed in the population. They sequenced the genomes of increasingly resistant *Staphylococcus aureus* isolates obtained from the bloodstream of a patient over a 4-month period. They observed that during this period, the strain accumulated 35-point mutations in 31 loci and this correlated with a steep drop in susceptibility of the strain to the antibiotic vancomycin, from the beginning of the study until the patient's death 12 weeks later. Harris et al (Harris et al., 2010) in their studies of antibiotic resistance in *S. aureus*, performed WGS of 63 strains of Multi-drug resistant *S. aureus* (MRSA) and identified 38 homoplasic (not identified as shared from common ancestor) SNPs, of which 18 were nonsynonymous including ten SNPs previously identified to confer antibiotic resistance. The observation that the variants arose independently suggests frequent independent evolution of resistance in the pathogen in addition to clonal spread of resistant strains. There have been many other similar studies (Chattopadhyay et al., 2009; Mena et al., 2008) that point to evidence of positive selection through independent generation of variants on the same amino acid position in a sequence of interest.

WGS has also made it more feasible to estimate mutation rates for different bacteria species (Biek et al., 2015; Ford et al., 2011). Direct estimation of mutation rates from clinical isolates may better reflect mutation rates in natural populations as compared to

mutation rates estimated in laboratory strains. Furthermore, whole genome sequencing offers the ability to detect most or all of genetic variation within a collection strains. In contrast, marker based approaches only sample a portion of the genome and may be subject to bias. Mutation rates are crucial in understating the evolutionary biology of bacterial species. For instance phylogenetic relationships are inferred using neutral mutations. Furthermore, the idea that mutation rates within a species could vary over time has been demonstrated by Mena et al. (Mena et al., 2008), through sequencing of *Pseudomonas aeruginosa* from early and later or chronic stages of a cystic fibrosis patient. The isolates from the chronic stage had a higher proportion of nonsynonymous SNPs that led to loss of functions of genes involved in the infection stage. This suggests that genes underlying acute infection might be selected against during chronic infection, suggesting hypermutation in long-term infection may improve genetic adaptation.

In summary, point mutations are central to evolution in bacteria population and they arise spontaneously at varying rates across bacterial genomes. Most of the neutral point mutations that arise within a population are simply lost due to genetic drift. Many point mutations are deleterious and are removed by the action of purifying selection over the course of evolutionary time. WGS analysis has enabled a rich approach to understanding both the rates and breadth of these mutations across taxa, and the inference of phylogenetic relationships between sets of isolates. The question of how much of a role *de novo* mutations play in the diversity of bacteria population, especially species that are competent and can acquire genetic materials horizontally and from theoretically any other

bacteria species, is an ongoing one, as is the actual nature of bacteria taxonomy and classification.

1.2.2.2 Impact of recombination in bacteria population

The concept of bacteria 'species' is one that has been beset with ambiguity. In biology, especially in terms of eukaryotes, species refers to a group of organisms that are capable of exchanging genetic materials, which are passed down to their offspring. Bacteria species are generally asexual and haploid and reproduction is through division of their cells in binary fission, hence vertical transmission of genetic features predominates. This idea of passing down identical genetic material during bacteria reproduction is regarded as the 'clone concept' or clonality (Levin et al., 1999).

However, early work done by Zinder et al (Zinder & Lederberg, 1952) demonstrated that bacteria also exchange genetic materials horizontally between individual cells. This work has been extended over the years to highlight sexual DNA exchange as a core process in bacterial evolution. Exchange of genetic material between bacteria is mediated by any one of three processes, collectively termed horizontal gene transfer (HGT). These processes includes: direct uptake of DNA from the surroundings (transformation); phage mediated transduction; and inter-bacteria contact and exchange of DNA (conjugation), resulting in the integration of these external DNA materials in the recipient cell through recombination (Thomas & Nielsen, 2005).

Improvements in DNA sequencing made it obvious that there is a broad extent to these horizontal exchanges of DNA, which are most commonly intra-species events but which rarely can occur between bacteria cells from different species (Anderson & Seifert, 2011). It has become increasingly clear that genetic exchange between bacteria, once thought to be uncommon is now reckoned as a significant force driving evolution of prokaryotes(Didelot & Maiden, 2010). Except for a few lineages of genetic monomorphic pathogen species, it is indeed now considered unusual to not find genetic exchange between bacteria (Achtman, 2008). Sexual DNA transfer appears to pose a challenge to the idea of clonal bacteria and bacteria species in general(Shapiro, 2014). A more nuanced view of the idea of bacteria species, is to treat bacteria populations as not entirely clonal, rather a set of individual genomes capable of exchanging DNA horizontally. There is a varying degree of the ability of different individuals sharing DNA in different ecological niches (Didelot & Maiden, 2010; Shapiro, 2014).

HGT promotes acquisition of novel genes from a pool of so called ‘accessory gene pool’. Within the pangenome (Mira, Martín-Cuadrado, D'Auria, & Rodríguez-Valera, 2010) or the collection of genes among individual bacteria strains in a given population, genes that are present in all the strains are referred to as core genes, while those not found in every member in the collection are regarded as accessory genes. The acquisition of accessory genes within and without a population have been shown in previous studies to impact pathogenic population and lead to emergence of new phenotypes (Hacker & Carniel, 2001). But HGT could also result in homologous recombination, where bacteria import genes or genetic fragments, which recombine with homologous genetic regions in their

genome, akin to gene conversion process in eukaryotes. Unlike homologous recombination in eukaryotes where the exchange of parts of the genome goes both ways, in bacteria it is a one-way exchange with portions of the DNA of one cell replacing the corresponding portion of DNA in the receiving cell (Didelot, Méric, Falush, & Darling, 2012b) This process was first discovered by observing mosaic genes at loci of antibiotic resistant genes in bacteria(Spratt, Bowler, Zhang, Zhou, & Smith, 1992). Spratt et al had observed the DNA sequence of the *penA* locus from 20 *Neisseria* isolates and surmised on evidence of similarity with *Neisseria* commensals, that resistant pathogen strains have acquired portions of the loci from commensals and evolved antibiotic resistance as a result.

The incidence of recombination within bacteria populations has confounded the ability to infer genetic relationships within a lineage or set of strains, just based on differences in their genome sequences (Didelot & Maiden, 2010; Shapiro, 2014). If bacteria were all clonal, without horizontal genetic exchanges, then reproduction by binary fission will parallel replication of the cell's DNA and the genetic history of daughter strains could be traced back to their ancestors given an underlying evolutionary model of a constant rate of substitution over time. Hence, given whole genome sequence data and variants across a group of strains, one could infer if identities observed in polymorphic sites are by descent from a common ancestor or if they occur by chance or convergence (Shapiro, 2014). Recombination muddles this, by introducing "homoplasic" variants across loci that renders the phylogeny on those loci incongruent to the phylogeny based on the entire genome, see Fig 1.1. It is a fundamental challenge in bacteria population studies to

account for the impact of recombination on the phylogenetic relationships of strains and more broadly on evolution within a bacteria population (Didelot & Falush, 2007)

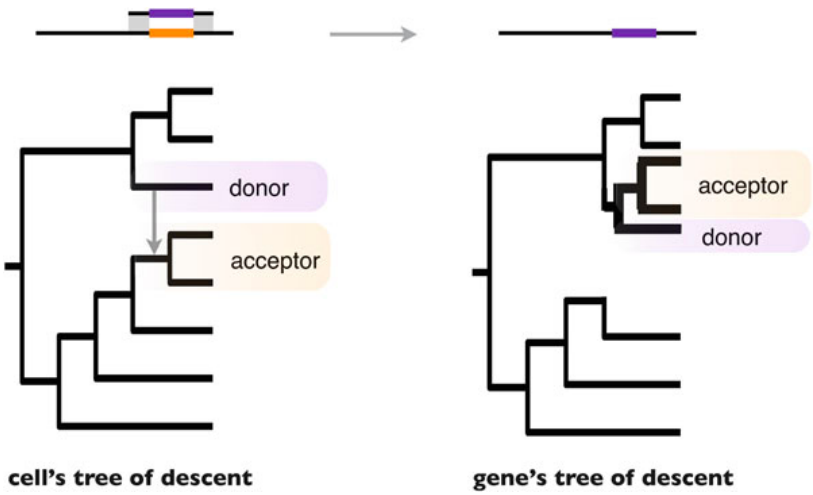


Figure 1.1. Recombination in a locus resulting in replacement of receptor DNA sequence (orange color) with donor DNA sequence (purple color) in the gene, leads to incongruent phylogeny between tree based on the gene, and that based on entire genome. Figure from(Shapiro, 2014).

Inferring the impact of homologous recombination in bacteria population, based on WGS sequence data has been mainly through two broad approaches: first, looking for evidence of unexpected similarities between divergent species, in a comparison where one of the species is hypothesized as the donor and the others as recipients, and movement between donor and recipient hypothesized to cause the similarities. Second, searching closely related bacteria isolates for evidence of imported genetic materials from a distantly related source. Here recombination is identified as regions of higher polymorphisms relative to the background level expected for closely related isolates that evolve clonally (Croucher et al., 2015).

There have been a number of bioinformatics methods developed to test for recombination based on these two approaches. The first uses non-parametric testing of the probability of observing homoplasies given the sequence data, without any recourse to an evolutionary model within the population. Examples of tools applying this approach include RDP3 (Martin et al., 2010), CBrother (Fang, Ding, Minin, Suchard, & Dorman, 2007), and GARD (Kosakovsky Pond, Posada, Gravenor, Woelk, & Frost, 2006). The difficulty with this strategy is the necessity of having DNA sequences for both the donor, recipient and recombinant genotypes. This is quite challenging for bacteria, because of the diversity that could exist within donor populations, some of which might be extant. Hence tractability of source of DNA for all cases is nearly impossible using this approach, especially given our currently still limited genomic sequence data (Croucher et al., 2015).

For the second approach, the methods used are chiefly based on coalescent theory. In brief, this is a statistical description of genealogical history of isolates sampled from a large and constant sized population with no selection or migration – the Fisher Wright population, where mutation introduces variation in the clonal population at a constant rate, which could be used to trace back the evolutionary time posts in the population (McVean, Awadalla, & Fearnhead, 2002). There are a number of tools that use the coalescent approach to estimate recombination and other evolutionary parameters in bacteria population, but two outstanding ones are LDHAT (McVean et al., 2002) and ClonalFrame(Didelot & Falush, 2007).

The algorithm for LDHAT adapted a modification of the infinite site models of the Wright-Fisher population to a finite sized population to compensate for the possibility of numerous mutations within the population that could give the appearance of signals of recombination. The algorithm derives composite likelihood estimation for the recombination rate of the entire sequence $2N_e r$, where N_e is the effective population size and r is recombination rate across the entire sequence length. This estimate is arrived at by first estimating the population mutation rate per site from the coalescent ($\vartheta = 2N_e \mu$; where μ is mutation rate), given the sequence data and observed substitutions across the sites on the sequence. Subsequently for each segregating site, the estimate of the rate of recombination is assessed as the likelihood of observing a particular allele on a given site, given the data and the evolutionary model. The composite estimate for recombination is finally arrived at by combining the likelihood of recombination for each set of comparisons, across all the sites (McVean et al., 2002).

ClonalFrame on the other hand applies a Bayesian Monte Carlo Markov Chain (MCMC), to infer recombination spots as regions within the genomes with elevated levels of polymorphisms, distinct from a simultaneously constructed tree based on a clonal phylogeny of point mutations, outside of the recombination regions. The model assumes a prior coalescent genealogy and also supposes that for each sequence length considered, the number of recombination events, follows a Poisson distribution. It also assumes that the nucleotide substitutions in non-recombined regions are based on the Jukes and Cantor model of substitution (Didelot & Falush, 2007). The approach is based on the idea that recombination in bacteria affects only a contiguous region of DNA sequence, leaving the

rest of the circular chromosome unchanged, unlike crossovers in eukaryotes where the ends of daughter chromatids are swapped after a preceding double stranded break of homologous chromosomes to generate distinct pairs of chromosomes in the next generation (J.-M. Chen, Cooper, Chuzhanova, Férec, & Patrinos, 2007). The method therefore estimates for each branch, the subset of the genome that has not undergone recombination, otherwise known as the "clonal frame" (Milkman & Bridges, 1990).

The pattern of gene flow between individuals within a population also offers clues on how much genetic exchange has taken place within geographical and ecological constraints. Given that physical proximity is essential for any of the HGT processes to occur, if related lineages are more often located in a common environment as compared to more distantly related lineages, then recombination between members of closely related lineages is more likely than recombination between members of different species. Bioinformatics tools like STRUCTURE (Falush, Stephens, & Pritchard, 2003) and BAPS (Tang, Hanage, Fraser, & Corander, 2009) apply the linkage model to trace gene flow in the population. The algorithm for these tools assumes that there are a number of ancestral populations and the genotype of each individual is drawn from these ancestral materials. Some individual genotypes could be exclusively from one ancestral population, while others are admixed. For a recombinogenic population, the admixed individuals will imply strains that have exchanged genetic materials over time, and the signals from analysis using these tools could infer the impact of HGT in a population and its correlation with geography and other ecological barriers (Didelot & Maiden, 2010).

WGS data analysis have provided a platform in recent years to more closely detect recombination in bacteria species and its impact in not only favoring selection of certain phenotypes like virulence and antibiotic resistance, but also in driving speciation within bacteria (Bryant et al., 2012). For instance, in the WGS studies on *S. pneumoniae* by Croucher et al (Croucher et al., 2011), they identified close to 700 recombination events by observing SNP densities and phylogeny in the genomes of 240 isolates of the PMEN1 lineage of the species. One of the recombination hotspots in their studies was a locus that encodes a protein responsible for synthesis of a surface polysaccharide, which is a vaccine target. Changes wrought by recombination in this locus were traced to vaccine escape by the bacterium. Another, a recent studies by Shapiro et al (Shapiro et al., 2012) on *Vibrio cyclitrophicus* suggested how recombination could play a significant role in speciation within bacteria population. The authors studied two strains with identifiable phenotype differences, among the species found in distinct ecological niches that are differentiated by a number of SNPs. These variations were implicated through functional studies in the adaptation of each population. Using population genomic analysis tools, Shapiro et al inferred that there was significantly greater recombination within than between subspecies. This suggests that under different ecological or environmental settings recombination acts to accentuate genetic variation that could ultimately lead to very genetically different bacteria strains. This process could be a fundamental evolutionary pathway for bacteria that plays out in different ecological settings, not the least under antibiotic resistance selection pressure and would suggest that recombination might have a larger than previously understood role in adaptation, especially for highly competent cells like those of this study system --*Neisseria gonorrhoeae*.

1.2.2.3 Estimating positive selection in bacteria population

A major goal of bacterial population genomics studies is to use WGS data to make inferences on signatures of selection within a population, the overall effect of selection and the forces driving evolution by natural selection within the population. There is a whole suite of different statistical tests based on observed sequence data that have been developed to test for selection within a population. Although most of these tests have been designed originally with sexual populations in mind, they could be adapted to detect selection in bacteria populations. The underlying premise for most of these tests is to highlight patterns of genetic variation that have been shaped by selection and contrast them from neutral patterns that are expected by random mutation or genetic drift. (Shapiro, David, Friedman, & Alm, 2009).

A common approach to measure selection in DNA sequences is to estimate the proportion of nonsynonymous nucleotide changes (dN) to that of synonymous nucleotide changes (dS). A $dN/dS > 1$ suggests that more of the nucleotide base changes at a given locus leads to alteration of the amino acids on the sequence and implies positive selection. Conversely, a $dN/dS < 1$ would suggest the action of purifying selection at the given locus, to remove deleterious synonymous mutations. A dN/dS that is equal to or very close to 1 indicates a situation of neutral selection (Z. Z. Yang, 2007). A powerful extension of this measure for individuals of different species is McDonald-Kreitman (MK) test (McDonald & Kreitman, 1991) that uses protein-coding sequences obtained from individuals both within and between species. The key assumption of the MK test is

that in the absence of selection the ratio of nonsynonymous to synonymous changes in a given protein should remain the same for samples both within and between species. So if the measure of polymorphisms within species is represented as nonsynonymous polymorphism (pS) and synonymous polymorphism (pN), then MK derives a value known as the fixation index (FI) which is: (dN/dS)/(pN/pS). FI > 1 for the tested data, suggests positive selection between species, while FI < 1 will suggest negative selection between the species. The different modifications of the MK approach have been used to detect signals of selection in bacteria populations(Simmons et al., 2008), but some of the assumptions of the MK approach especially the assumption that all sites within organisms evolve independently have called to question the suitability of the test in all situations, and the need to relax this assumption for the test (Fay, 2011).

To understand the intricate nature of bacteria populations and evolution of prokaryotic genomes, one has to understand the interplay of the forces that introduce and maintain variation within these populations. Genetic variants are introduced to the population either through point mutations that are passed down, or via horizontal gene transfers from other bacteria cells. For bacteria populations which are normally of large size, with reduced effects of genetic drift relative to selection, variants that are beneficial are generally maintained in the population, while deleterious variants are more effectively removed than in eukaryotes (Shapiro, 2014).

1.3 Bacterial antibiotic resistance

The treatment of bacterial infections with antibiotics has undoubtedly been one of the most significant medical breakthroughs in human history. The development of antibiotics could be traced to the late nineteenth century by acceptance of the 'germ theory of disease' after the work of Pasteur and Koch (F. Winau, Westphal, & Winau, 2004). Subsequently, Ehrlich's small molecule screening approaches and successful characterization of the first anti-trypanosomal and anti-syphilitic drugs introduced the modern era of antimicrobial therapy (F. Winau et al., 2004). This discovery was quickly followed by a sequence of other developments: the development of synthetic antimicrobial agents by using the emerging science of organic chemistry, the discovery of the very potent chemically diverse, non-toxic antibiotics derived from bacteria and fungi and the advent of the medicinal chemistry era where already discovered antibiotic drug scaffolds are tailored or modified to evade antibiotic resistance (G. D. Wright, 2012).

The history of antibiotic drug use has always been confounded by corresponding development of insensitivity to these drugs by target pathogen populations, ultimately eroding their clinical utility (Tenover, 2006). This phenomenon - antibiotic resistance- is becoming more widespread, and in recent decades, a number of pathogens such as *Mycobacterium tuberculosis*, *Acinetobacter baumannii*, *Staphylococcus aureus* and *Haemophilus influenzae* have all developed resistance to a spectrum of antibiotic medications – a situation referred to as multi-drug resistance (MDR)(J. Davies & Davies, 2010) At the same time the pace of development of new antibiotic drugs has fallen off in recent years (Fischbach & Walsh, 2009; Krause, 1992). Furthermore, some of these pathogens that are increasingly multiple-drug resistant and contain many resistance genes

for different antibiotics and are also often associated with increased virulence(S. B. Levy & Marshall, 2004; Livermore, 2004). Some of the multi-drug resistant bacteria are indeed emerging as epidemic and pandemic strains, posing one of the most pressing public health challenges of our time (Croucher et al., 2015; Peirano & Pitout, 2010).

An important step towards solving the antibiotics resistance crisis is understanding the molecular mechanisms driving the emergence of resistance in microbes and placing such processes within the context of the overarching evolutionary process playing out in the pathogen population.

1.3.1 Molecular mechanisms of antibiotic resistance in bacteria

Resistance could be simply defined as the continued growth of microorganisms in the presence of cytotoxic or growth-inhibiting levels of antibiotics. There is no standard dosing concentration across all antibiotic drugs instead each antibiotic drug has an empirically calculated minimum inhibitory concentration (MIC). Resistance by a pathogen is inferred when the organism thrives beyond the MIC concentration level of the drug (Didelot et al., 2012a). Understanding the molecular and biochemical paths involved in the development of resistance in an organism serves to highlight the various genes and underlying evolutionary process behind the emergence of resistant strains.

Most common antimicrobial agents act by inhibiting protein, nucleic acids, or cell wall synthesis in bacteria. Some are also effective in disrupting crucial metabolic pathways in pathogens (Neu, 1992), Table 1.1. The overall effect of antibiotic action is either the

induction of cell death or inhibition of growth on contact with a target in the bacteria cell (Kohanski, Dwyer, & Collins, 2010b). There have been three main molecular mechanisms of antibiotic resistance by pathogens identified from past studies: (1) preventing antibiotic entry into the cell through drug efflux or reduced membrane permeability, (2) preventing the binding of the antibiotic to its target in the cell by target alteration and possible modification of metabolic pathways related to the target and (3) enzymatic modification and degradation of antibiotics (C. Walsh, 2000) .

For instance, the mechanism of resistance to β-lactam antibiotics such as the penicillin and cephalosporin is via enzymatic cleavage of these antibiotics by the hydrolytic actions of β-lactamase enzymes(J. F. Fisher, Meroueh, & Mobashery, 2005) (encoded by resistance genes such as *ampC, TEM, OXA*) found in resistant bacteria. Similarly, the high level of resistance to vancomycin a glycopeptide antibiotic occurs through an intricate process of alteration of the peptidoglycan layers on the cell wall of the resistant bacteria, which is the binding site of the antibiotic. The underlying biochemical mechanism is driven by the expression of the van genes (*vanH, vanA and vanX*) that is triggered by the presence of the drug. The expressed enzymes catalyze an alternate metabolic pathway that leads to modification of the cell wall and reduced affinity for vancomycin binding, hence resistance(Bugg et al., 1991). The *mtrCDE* operon is an active efflux pump found in bacteria that exports macrolide antibiotics and has been implicated in high-level resistance to penicillin antibiotics (Veal, Nicholas, & Shafer, 2002).

These mechanisms of resistance demonstrate that the presence and modifications of so-called resistance genes -- genes involved within the various physiological or biochemical pathways culminating in resistance -- most often drive antibiotic resistance (MacLean, Hall, Perron, & Buckling, 2010). Understanding the source and flow of these genetic resistance determinants in pathogenic as well as non-pathogenic bacteria population will offer a broader picture in the quest to develop more effective antimicrobial therapies.

Molecular Mechanism	**Antibiotic Class**
Interference with cell wall synthesis	**βLactams:** penicillin, cephalosporin, carbapenems, monobactams **Glycopeptides:** vancomycin, teicoplanin
Protein synthesis inhibition	Aminoglycosides, chlorophenicol, clindamycin, tetracyclines, mupirocin and Macrolides
Interference with nucleic acid synthesis	**Inhibit RNA synthesis**: rifampin **Inhibit DNA synthesis**: fluoroquinolones
Inhibition of metabolic pathways	Sulfonamides, folic acid analogues

Table 1.1. Mechanisms of action of antibiotic medications

1.3.1.1 Antibiotic resistance mechanisms in *Neisseria gonorrhoeae*

Antibiotic resistance is rampant within the *Neisseria gonorrhoeae* population (Ohnishi et al., 2011; Unemo & Shafer, 2014). There have been a number of studies that have helped elucidate the underlying molecular mechanisms driving resistance in this pathogen (reviewed by Unemo & Shafer, 2011), see figure 1.2. Broadly speaking, there are two groups of antibiotic drugs currently used in the treatment of gonorrhea: third generation cephalosporin and macrolides.

Cephalosporin inhibits bacteria growth by preventing the biosynthesis of bacteria cell

wall. Studies have shown that the underlying genetic alteration that confers resistance to third generation cephalosporins is a mosaic-like resultant structure of the *pen*A (Ameyama et al., 2002) gene, which encodes the binding protein for this class of antibiotics. Different alleles of the gene are believed to have developed from recombination with portions of DNA transferred horizontally from commensal *Neisseria* species. The reduced affinity for binding of cephalosporins to the penicillin binding protein 2 (PBP2) encoded by the gene translates into decreased susceptibility of *N. gonorrhoeae* to the antibiotic. PBP2 is involved in the development of bacteria cell wall, a key biological process in the pathogen, and hence a target of antibiotic drugs.

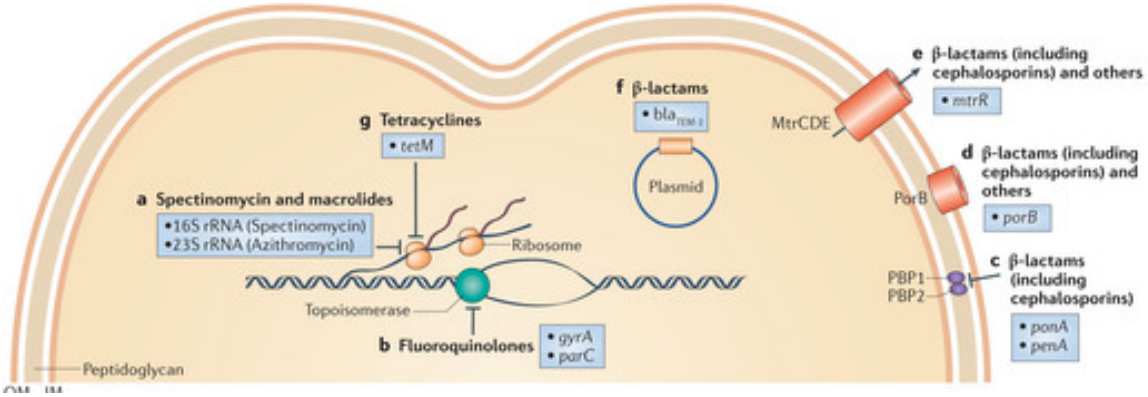


Figure 1.2. Antibiotic resistance mechanisms in *N. gonorrhoeae* (Goire et al., 2014)

The other class of antibiotic drugs used in treating gonorrhea is the macrolides. This class includes such drugs as azithromycin and erythromycin. Macrolides act by inhibiting bacterial protein synthesis by binding to the subunits of bacterial ribosomes. Previous studies have shown that resistance by *N. gonorrhoeae* to macrolides is determined by 4 different alleles of the 23srRNA gene (Chisholm et al., 2009; Palmer, Young, Winter, & Dave, 2008; Starnino, Stefanelli, Neisseria gonorrhoeae Italian Study Group, 2009). This loci is the target for macrolides. It has also been shown that mutations in the *mtrRCDE*

operon, which encode an efflux pump, and penB- a gene that expresses a porin protein on the surface of the bacteria, account for some of the observed resistance to macrolides (Deguchi, Nakane, Yasuda, & Maeda, 2010; Hagman & Shafer, 1995).

There are still some resistance phenotypes within *N. gonorrhoeae* which have not been fully explained by known genetic mechanisms (Unemo & Shafer, 2011). Therefore, the increasing amount of whole genome sequence data available for strains within this pathogen population could enable the characterization of novel variants underlying such antibiotic resistance phenotypes.

1.3.2 Antibiotics resistance in commensal bacteria

The golden era of antibiotics discovery may have commenced after Fleming publicized penicillin, but for millennia, resistance elements had already been circulating in bacteria populations long before any chemically manufactured antibiotic was in use (Clemente et al., 2015). It is also important to note that the vast majority of bacteria are non-pathogenic for humans. Naturally occurring antibiotics produced by microbes, are ancient. It has been estimated that the biosynthetic pathways for erythromycin, streptomycin and vancomycin emerged over 880,610 and 240 million years ago respectively (G. D. G. Wright, 2007). Microbes live their life surrounded by antibiotics and countless other similar molecules. This possibility of natural organisms serving as reservoir - the so-called “resistome” - for the transfer of resistance genes to pathogens was demonstrated by evidence of lateral transfer of *ctx*-m genes, a type of β-lactamase from nonpathogenic bacterium of the genus *Kluyvera* to pathogenic bacteria. *ctx-m* genes had been circulating

in pathogenic bacteria population such as infectious *E.coli* (Bauernfeind, Grimm, & Schweighart, 1990) in clinics but the origin of the gene was unknown. The sequencing analysis work done by Olson et al showed that these resistance genes had 100% identity with chromosomal genes found in *Kluyvera* (Cantón & Coque, 2006; A. B. Olson et al., 2005).

The human microbiome affords a readily observable niche for the dynamics of exchange of resistance elements between commensals and pathogens. Almost every surface of the human body is colonized by a rich and diverse community of commensal microbes--the sum of which makes up the human microbiome-- that have substantial and continuous effects on human and various physiological processes (Dominguez-Bello et al., 2010; Human Microbiome Jumpstart Reference Strains Consortium et al., 2010).

The use of antibiotics to treat infections potentially has unintended consequences on the portion of the microbiome that shares similar environment with the intended target, in addition to disrupting critical biological processes in a specific pathogen. This is especially true, considering the potential role of antibiotics as multi-activity signaling molecules (Yim, Wang, & Davies, 2007). From previous studies, there have been two broad categories of noticeable changes on the commensal microbiota resulting from the use of antibiotics: changes in the relative proportion of different species in the microbiota, with introduction of new species and the decline or eradication of some species (Dethlefsen, Huse, Sogin, & Relman, 2008; Jakobsson et al., 2010) and changes that alter the antibiotic resistome or resistance genes encoded by members of the

microbiota (Shoemaker, Vlamakis, Hayes, & Salyers, 2001; Sommer, Dantas, & Church, 2009). In the latter category, there is noticeable enrichment and exchange of resistance genes within the microbiota, which increases the chances of incorporation of these elements in pathogenic organism and inoculating them against future treatment(Sommer & Dantas, 2011). However, recent findings of antibiotic resistance genes in the microbiome of previously uncontacted Amerindians would suggest that possibly some of these resistance elements are already primed regardless of contact with antibiotics (Clemente et al., 2015)

The expectation of increased abundance of antibiotic resistance elements in microbiota exposed to antibiotics have been borne out by a number of studies, for instance a culture-independent PCR assay showed that there was up to a 10,000-fold increase in the abundance of macrolide resistance genes compared to pre-treatment levels in the biomes of antibiotic-treated patients and persisted for more than two years (Jernberg, Löfmark, Edlund, & Jansson, 2007). The fitness cost of acquiring these resistance elements is offset mostly by compensatory adaptations. The persistence of resistance genes even after the cessation of antibiotic therapy (Jernberg et al., 2007; Sjölund, Tano, Blaser, Andersson, & Engstrand, 2005) is even more problematic, as it will suggest that with increasing exposure to antibiotics our microbiome are evolving to an even more resistant state, basically expanding the reservoir of resistance genes. While there have been *in vitro* experimental demonstrations of lateral transfers of resistant genes between distant genera (Hannan et al., 2010), there has been little *in vivo* experimental support for such transfers (Sommer & Dantas, 2011). This could be because of the difficulties involved in setting

up controlled in vivo studies where the donor and recipients need to have similar microbiota. The challenge will be devising an experimental design that will take into account the diversity in the microbiota of individuals in future studies

1.4 Next generation sequencing (NGS) and big data analysis

Generation of whole genome sequencing data for a wide range of organisms has become commonplace with advances in next generation sequencing technologies, allowing the possibility of in-depth studies in wide-ranging biological populations using information on variations provided by these data (J. Zhang, Chiodini, Badr, & Zhang, 2011).

For instance, in human studies the discovery of genome-wide genetic variation was central to the field of genomics (Chakravarti, 2011; Lander, 1996). Now, recent advances in second-generation sequencing technologies and better methods of targeted enrichment mean the detection of genome-wide patterns of genetic variation will soon be a routine operation (Bhangale, Rieder, & Nickerson, 2008; Fledel-Alon et al., 2009). There have also been numerous population genomics studies (Luikart, England, Tallmon, Jordan, & Taberlet, 2003) of pathogens that pose significant public health challenges, as a result of these improved technologies.

Yet these advances in DNA sequencing have revealed a new bottleneck: the functional classification and interpretation of newly discovered genetic variation. The scale of this problem is enormous. The high throughput and low cost of second-generation sequencing platforms now allow geneticists to routinely perform single experiments that identify tens

of thousands to millions of variant sites in a single individual, but the methods that exist to annotate these variant sites using information from publicly available databases are too slow to be useful for the large sequencing datasets being generated. Because sequence annotation of variant sites is required before functional characterization can proceed, the lack of a high-throughput pipeline to annotate variant sites efficiently can be a major bottleneck in genetics research and clinical applications of genomics technologies (Ezewudo et al., 2012).

1.4.1 Association studies using bacteria NGS data

While identifying and connecting genetic determinants to phenotypes using WGS data has increasingly become the preferred approach for genetic studies in human populations, there is also the emerging possibility of using NGS studies of pathogen population to identify and correlate genetic variants or the underlying genetic architecture of relevant pathogen phenotypes such as antibiotic resistance, virulence and disease outbreak surveillance (P. E. Chen & Shapiro, 2015; Didelot et al., 2012a; Read & Massey, 2014).

Prior to now, the studies and approaches that had been used to identify resistance and virulence determinants in bacteria have mostly been based on genetics analysis of individual loci of interest (Ameyama et al., 2002; Pan & Spratt, 1994; Veal et al., 2002). While individual loci analysis coupled with functional experiments has been helpful in elucidating mechanisms of antibiotic resistance in pathogens, the WGS approach offers more breadth, and the information gleaned from associations between phenotypes and identified variants could paint a better picture of the possible interactions between the

different known genetic variants and possible novel variants underlying phenotypes of interest.

There are two approaches from recent genome-wide studies using WGS to associate genetic variants to phenotypic differences in bacteria. The first approach involves using modifications of the classic case – control association tests using variants detected from WGS of a number of bacteria strain of same species. This approach was taken is the genomics studies by Alam et al(Alam et al., 2014) on 75 strains of *Staphylococcus aureus*, 49 of which were sensitive to vancomycin and 26 were of intermediate resistance to the antibiotic. They used a genome wide association test statistical tool ROADTRIPS (Thornton & McPeek, 2010) that accounts for unknown population or genetic relatedness of the tested samples by using a genome-wide covariance structure. This corrects for the use of bacteria strains that could be of similar lineage and the spurious results population stratification from the sample set could cause. Of the 55,977 SNPs they tested, they found one nonsynonymous mutation in protein *rpoB* to be in significant association with increased vancomycin resistance.

The second approach involves using ancestral reconstruction of phylogenies of individual strains in a sample set to perform statistical tests associating a change in phenotype of interest along branches of the reconstructed phylogeny of the individuals with a change in state or mutations along same branch. Modifications of these approaches are implemented in the bioinformatics tools such as PhyC (Farhat et al., 2013) and PPFS(B. G. Hall, 2014) and in the refined association mapping approach (Sheppard et al., 2013).

In the studies by Farhat *et al* (Farhat et al., 2013) to identify underlying genetic variants associated with antibiotic resistance in *Mycobacterium tuberculosis* they obtained WGS data for 123 strains of the pathogen, which they grouped based on the phenotype of resistance or sensitivity to antibiotics. They based their tests on the concept of evolutionary convergence, and the idea that mutations or variants that give rise to phenotypes of interest will arise independently and repeatedly over time in the evolutionary history of the strains. They were able to establish a significant association between mutations known to cause resistance in the pathogen and resistance phenotype, by correlating presence of these mutations along branches of resistant strains on an ancestrally re-constructed phylogeny of the strains.

Given the growing ease and increase in WGS as well as metagenome sequencing of prokaryotic samples, using whole genome data to identify associations between underlying genetic variants and phenotypes of interest will be of more relevance in bacteria population and evolutionary studies.

1.5 Questions examined in this thesis

From the above discussions, we are left with a number of questions as to the nature of the *N. gonorrhoeae* population, the kinds of data that is used in genomic studies and the tools that could make sense of NGS data. Given the advantages of the sequence annotation tool SeqAnt, over similar tools, how could it be revised to become a platform for addressing similar questions centered on identifying genetic variants of significance associated with

phenotypes of biological interests, in more model organisms, including the plethora of prokaryotic organisms?

N. gonorrhoeae, like a number of other bacterial species, does have an appreciable level of intra-species recombination taking place within its population (Hamilton & Dillard, 2006; M G Lorenz, 1994). How does the impact of recombination in this population correlate with the broader issue of evolution of antibiotic resistance in the pathogen? Given the capability to obtain *N.gonorrhoeae* isolates from different regions of the globe, underscoring different prevalence and treatment approaches to gonorrhea, how does the resistance profile of isolates relate to their geographical location? And what does the admixture pattern from globally assembled strains of the pathogen tell us of the structure of the population and the evolutionary history of the bacterium.

The molecular mechanisms underlying resistance in bacteria and in *N. gonorrhoeae* has been well studied in bacteria literature. The majority of the determinants identified to underlie resistance were discovered based on individual gene studies. What inferences could one make by performing genome-wide search and tests for variants that are associated with antibiotic resistance? Are there novel variants that associate with or augment resistance to antibiotics that could be uncovered using genome-wide association tests?

1.6 Outline of thesis and chapter summaries

I will attempt to answer the questions I raised above in the following chapters of this document. Chapter 2 is an overview of current understanding of the underlying genetic architecture of complex diseases, the utility of using next generation sequencing technology tools to better understand the genetic basis of these disorders and the inherent challenges of using NGS analysis to make strong inferences on the causality of genetic variants (Ezewudo & Zwick, 2013)

In Chapter 3, I follow up on Chapter 2 by presenting work done on revising a NGS annotation web tool, which serves as a bridge to actualize using whole genome sequencing to solve clinical and population genetics questions across a broad spectrum of life forms. It addresses the growing needs of researchers in the field for a platform that offers a broad and regularly updated gene and variants annotation database that efficiently identifies millions of variant positions in a model organism's genome. The tool, which also employs a range of conservation scores to infer the functional relevance of annotated variant positions, is revised to include the capability of annotating prokaryotic genomes and hence could be very useful for pathogen population studies.

The study I present in Chapter 4 is based on population genomics analysis of 76 *N. gonorrhoeae* strains sampled from across the globe, and with varying resistance phenotypes to a suite of antibiotics (Ezewudo et al., 2015). We set out to understand the roles of recombination and positive selection in evolution within this population and the intricate relationships between the structure of the pathogen population, the geographical spread of the isolates and antibiotic resistance phenotypes observed in the sample set. I

used the ClonalFrame tool and BAPS tool to piece the effects of homologous recombination within the pathogen population and the effects of these exchanges on the evolutionary history and population structure of this pathogen. I also queried the WGS used in the studies for the presence of known antibiotics resistance elements that have been previously described in the literature. Our results from this study suggest the pathogen is very cosmopolitan and has appreciable level of recombination occurring in its population.

Chapter 5 is based on genome wide association studies performed on the data from the studies in Chapter 4. I utilized two approaches for this study, first I performed the association tests using QROADTRIPS, which could correct for the genetic relatedness of the strains in the sample set and hence not significantly impacted by population stratification. I also used the approach of PPFS to identify SNPS that are significantly linked to antibiotic resistance phenotypes across branches of the ancestrally reconstructed phylogeny of the strains. We uncovered a number of candidate variants and loci that could be causative of resistance to a number of antibiotic medications.

Finally, in Chapter 6 I summarize the conclusions from the studies I presented in this dissertation. I also touch on further questions emanating from the studies and other inquiries, which could be taken on in the future to build on our results.

Chapter 2

Evaluating Rare Variants in Complex Disorders Using Next-Generation Sequencing

Matthew Ezewudo and Michael E. Zwick

Modified from Current Psychiatry Reports 15(4): 349 doi 10.1007/s11920-013-0349-4

2.1 Introduction

The individual and societal burdens of common, complex neuropsychiatric disorders are truly profound (Eaton et al., 2008). One of the major goals of contemporary biomedical research is to elucidate those disease mechanisms that underlie complex neuropsychiatric disorders like autism spectrum disorder (ASD) or schizophrenia (SZ). The hope is that an understanding of the pathogenesis of these disorders will enable the development of new treatments for those patients already affected, and new preventatives for those who are not. Because susceptibility to neuropsychiatric disorders is influenced both by variation in genes and environmental exposures, both genetic and epidemiological studies can help uncover novel disease mechanisms.

Our review focuses on what genetic studies of complex neuropsychiatric diseases have revealed about their genetic architecture, with a particular emphasis on studies of schizophrenia and autism. We divide the review into four main sections that reflect the technologies, experimental designs, and hypotheses tested in both recent and ongoing

genetic studies of complex neuropsychiatric disorders. The first section discusses the recent history of human genetic studies, which have focused on the contribution of common variation to the risk of complex diseases. The second examines the contributions of exceptionally rare variants with large effects on disease risk. The third section addresses how the rapid development of next-generation sequencing and the targeted enrichment of eukaryotic genomes are contributing to studies of complex traits. Finally the last section focuses on challenges facing the application of next-generation sequencing in both research and clinical translational applications.

2.2 Common genetic variation and complex diseases

For human genetic studies, the decade after the initial sequencing and analysis of a human reference genome has been a revolutionary one (Lander, 1996; Venter, Adams, Myers, Li, & Mural, 2001). The scaffold a reference genome provided allowed us to catalog one class of human genetic variation: single nucleotide polymorphisms (SNPs). Subsequent studies dramatically reduced the cost of genotyping genome-wide collections of hundreds of thousands of SNPs, while at the same time developing a map of the patterns of statistical correlation among common SNP variants, referred to as linkage disequilibrium, in the HapMap project (The International HapMap Consortium, 2005). Furthermore, theoretical predictions suggested that a classic experimental design derived from epidemiology, a case-control association study, would have more statistical power than traditional genetic family-based linkage studies (Risch & Merikangas, 1996). With these technologies in hand, genome-wide studies of complex disorders became feasible. From this point, the conceptual framework and the types of experiments pursued were

driven by both the availability of high-throughput genotyping platforms developed during the HapMap project and assumptions about the genomic architecture of common complex disorders (Zwick, Cutler, & Chakravarti, 2000). The initial application of these technologies focused on an experimental design called the genome-wide association study (GWAS). The human genetic GWAS "industry" set out to test the hypothesis that common variants (those with >5% frequency in the human population) with large effects contributed significantly to the risk of disease. The essential idea was that common disease-causing variants, which were expected to be found at an elevated frequency in cases as compared to matched controls, would either be genotyped directly or be in linkage disequilibrium with common SNPs, thereby allowing them to be discovered. This approach was successful in identifying many novel loci that contribute to a wide variety of complex diseases(Klein, Zeiss, Chew, Tsai, & Sackler, 2005; Visscher, Brown, McCarthy, & Yang, 2012a; Wellcome Trust Case Control Consortium, 2007).

Unfortunately, the results of multiple genome-wide association studies of common neuropsychiatric disorders have been far more modest. Studies of schizophrenia (SZ) revealed only a few loci that exceed genome-wide levels of statistical significance, while the effect sizes of the variants uncovered were remarkably small(International Schizophrenia Consortium et al., 2009; Need et al., 2009; O'Donovan et al., 2008; Shi et al., 2009). Moreover, only a modest amount of the total heritability of SZ has been accounted for, in contrast to other complex traits, such as human height (Lee et al., 2012; McQuillan et al., 2012; J. Yang et al., 2010). Similarly for autism, multiple GWAS identified a few loci of very small effect(Anney et al., 2010; K. Wang et al., 2009; Weiss,

Arking, Gene Discovery Project of Johns Hopkins & the Autism Consortium, Daly, & Chakravarti, 2009). A subsequent meta-analysis suggested that finding any genetic variants with an odds ratio greater than 1.5 for autism is extraordinarily unlikely(Devlin, Melhem, & Roeder, 2011). For neuropsychiatric disorders, therefore, the effect sizes of the variants identified have been disappointingly small, particularly when compared to GWAS of other complex human disease traits.

It is in fact the small effect size of common variants that is the most striking finding from nearly all genome-wide association studies of complex diseases. Together, these studies have soundly rejected the hypothesis that *common variants* with *large effects* underlie the vast majority of complex human diseases. Thus, while one can argue that the GWAS approach has been a success, these studies have revealed that the genetic architecture of most complex diseases is unlike that seen in cystic fibrosis or sickle cell anemia, where common alleles with very large effects account for most of the disease prevalence in human populations. This finding is particularly relevant for complex neuropsychiatric disorders like autism spectrum disorder and schizophrenia. Furthermore, although there are statistically compelling associations between common genetic SNPs and diseases (Manolio, 2010; Manolio, Brooks, & Collins, 2008), single nucleotide polymorphisms (SNPs) alone are unable to account for all the genetic proportion of heritability in complex traits. The fact that a substantial proportion of the estimated heritability of these traits remains unexplained points to other classes of genetic variants that have yet to be discovered (Eichler et al., 2010; Maher, 2008; Manolio et al., 2009).

2.3 Rare genetic variation and complex diseases

In retrospect, perhaps this outcome should have been less surprising. Theoretically, it has long been recognized that both common and rare variation likely contribute to the genetic architecture of complex traits (Barton & Keightley, 2002; Barton & Turelli, 1989; Falconer, 1967; Lande, 2007; Pritchard & Cox, 2002; Zwick et al., 2000). In addition, genome-wide association studies of common variants were pursued for the simple reason that technological advances made this experiment became feasible. Direct sequencing of genomes to identify the contribution of rare variants in large numbers of patient samples faced daunting technological challenges and excessive costs that simply made such studies impractical (Schork, Murray, Frazer, & Topol, 2009; Visscher, Goddard, Derks, & Wray, 2012b; Zwick et al., 2000).

While large-scale genotyping of SNPs for GWAS was underway, similar genome-wide technologies led to the discovery of widespread variations in copy number across the human genome (Iafrate et al., 2004; Perry et al., 2006; Sebat, Lakshmi, Troge, Alexander, & Young, 2004; Sharp et al., 2005).This class of genetic variation, consisting of deletions and duplications larger than 100 kb, was surprisingly frequent. Although clinical geneticists had long recognized that cytologically visible, and usually much larger, chromosomal changes were associated with rare human diseases, the developing technologies allowed the discovery of smaller copy number variants (CNVs) that were not observable using classic cytological approaches. The role of this structural variation in human disease became an immediate focus (Girirajan, Campbell, & Eichler, 2011; Stankiewicz & Lupski, 2010). Soon after, the discovery of an elevated frequency of

CNVs in patients with schizophrenia hinted at an explanation for the great heterogeneity of the disorder (Magri et al., 2010; S. E. McCarthy et al., 2009; Mulle et al., 2010; Stefansson et al., 2008; T. Walsh et al., 2008; Xu et al., 2008). Similar findings also came to light for autism (Glessner et al., 2009; R. A. Kumar et al., 2008; D. Levy et al., 2011; C. R. Marshall et al., 2008; Sebat et al., 2007; Weiss et al., 2008), as well as for both schizophrenia and autism (Moreno-De-Luca et al., 2010). Nevertheless, the apparently pathogenic CNVs discovered to date are in general large and very rare in the population, which means they alone are unable to explain all of the missing heritability.

2.4 Next-generation sequencing, targeted enrichment, and complex diseases

Comprehensive sequencing of human genomes would no doubt be better at capturing the allelic architecture of complex diseases than the genotyping of common variants in GWAS or the detection of rare, large CNVs with methods like array comparative genomic hybridization, but even in the recent past this has been cost-prohibitive. Recent advances in next-generation sequencing (NGS), however, have increased throughput while decreasing costs, so this barrier is eroding quickly (1000 Genomes Project Consortium et al., 2010; D. R. Bentley et al., 2008; Drmanac et al., 2010; Fujimoto et al., 2010; Kidd et al., 2008; Lupski et al., 2010; Tong et al., 2010; Wheeler et al., 2008) see reviews in (Metzker, 2010; Shendure & Ji, 2008). Combining NGS with methods that can enrich for portions of complex eukaryotic genomes (Albert et al., 2007; Gnirke et al., 2009; Hodges et al., 2007; Mondal, Shetty, Patel, Cutler, & Zwick, 2011; Okou et al., 2009; 2007; Porreca et al., 2007; Tewhey et al., 2009)has made it feasible to pursue other types of genetic variation underlying complex disease traits.

The initial application of these technologies focused on whole-exome sequencing, which involves sequencing the 1% of the human genome that codes for proteins, in the context of diseases caused by mutations at single loci (so called Mendelian diseases) (Choi et al., 2009; Ng et al., 2010; 2009). Application of these approaches to schizophrenia uncovered a role for *de novo* mutations in the etiology of the disorder (Girard et al., 2011; Xu et al., 2011). A more recent study suggested that many of the variants contributing to schizophrenia must be very rare and have yet to be discovered (Need et al., 2012). Whole-exome sequencing studies of autism also point to a role for *de novo* mutations in autism phenotypes (Iossifov et al., 2012; Neale et al., 2012; O'Roak et al., 2011; 2012; Sanders et al., 2012). Targeted studies of the X chromosome in males with autism, an attractive target given the 4:1 preponderance of males affected with the disorder, have revealed a number of putative autism susceptibility loci (Chung et al., 2011; Noor et al., 2010; Piton et al., 2011). More recently, a combination of targeted enrichment of the X chromosome exome and next-generation sequencing identified the *AFF2* locus as having a significantly larger number of rare missense mutations in those with autism versus unaffected controls (Mondal et al., 2012).

The clear message from all these studies is that exome sequencing can detect a broader allelic spectrum of complex neuropsychiatric disorders like schizophrenia and autism. As whole-genome sequencing becomes more and more cost effective, the field is bound to move towards this experimental design, which can reduce biases in ascertainment and make it possible to discover the full diversity of genetic variation.

2.5 Challenges facing next-generation sequencing and complex diseases

Next-generation sequencing and genomic enrichment technologies promise to detect both common and rare variants, thereby giving us a better understanding of the genetic architecture of complex diseases, yet there are a number of substantial challenges facing the application of these technologies in both research, and ultimately, the clinic (see reviews in (Kiezun et al., 2012; Lyon & Wang, 2012)). We believe these challenges fall into three main categories: accurate identification of genetic variation, efficient analysis of next-generation sequencing data, and interpreting the functional effects of genetic variation.

2.5.1 Accurate identification of genomic variation

Next-generation sequencing technology platforms (Illumina, Roche 454, ABI SOLiD, Ion Torrent) have higher error rates in individual sequence reads than conventional Sanger sequencing. These errors could be systemic and significant enough to yield false-positive variant calls and associations, as well as obscure actual associations. Making even more errors possible are biases that arise in coverage: GC-rich genomic regions tend to have lower sequencing coverage. Further, enrichment technologies add another layer of possible errors, especially those methods that select sequences by hybridization to a complementary oligonucleotide. The existence of gene families and other repetitive regions imply that multiple genomic regions can be captured and enriched by a single oligonucleotide substrate targeted at a specific region. Finally, errors in mapping sequence against a human genome reference sequence can lead to the misidentification of

genetic variants. This is of particular concern because the human genome reference sequence is an idealized genome, and undetected variation among individuals can lead to spurious outcomes in the mapping and identification of genomic variation (Rosenfeld, Mason, & Smith, 2012).

The fact that the human genome is very large (3000 Mb) implies that extraordinary accuracy is necessary to identify variant sites. Even modest error rates of 1%, for instance, could impugn the validity of association studies (Glenn, 2011). As a simple example, if we consider only SNPs, we expect approximately 3 million variant sites per genome. As shown in Table 2.1, in a scenario that surely underestimates the possible sources of error, unless algorithms calling variants sites are exceptionally accurate, the result will be an enormous number of false-positive findings.

2.5.2 Efficient analysis of next-generation sequencing data

Assuming we are able to accurately identify variant sites, the next step is functionally annotating those same sites so we can focus on those most likely to contribute to disease. Genomic variation identified with NGS technologies needs to be annotated to establish its type, genomic region, the evolutionary conservation of its site, and whether it has prior characterization. A typical whole-genome association study of a population would yield millions of variants, data that cannot realistically be characterized using public web genome browsers because of the huge effort involved. One early solution to this problem was the open source Sequence Annotator, or SeqAnt(Shetty et al., 2010); there are now a number of similar resources (Chelala, Khan, & Lemoine, 2009; K. Wang, Li, &

Hakonarson, 2010). Researchers will have to rely increasingly on such high-performance annotation tools to analyze the sequencing data generated from large sequencing studies.

2.5.3 Interpreting the functional effects of genetic variation

Ultimately, the goal of association studies is to link genetic variants to phenotypes through statistical tests that show significant connections (effect sizes) between the discovered variant and the phenotype of the disorder. In those cases of Mendelian disorders, where single variants can account for a given disease phenotype, interpreting the functional effects of genetic variation is in many cases easier.

For complex neuropsychiatric disorders, on the other hand, the challenge is far more formidable (Kiezun et al., 2012). When performing statistical testing of association of hundreds of thousands of variants in a genome-wide study, one immediately confronts the multiple testing problem, which is simply that, when performing a very large number of tests, the expected number of findings that exceed a nominal threshold of 0.05 will be substantial. Statistical methods like a Bonferroni correction or permutation can be used to control this issue, such that only very significant association signals are selected and false positives are reduced (Aickin & Gensler, 1996); however, the extent to which false-negative findings are increased by these approaches remains unclear and difficult to determine.

Beyond this, the extent of genomic variation, perhaps much of it having no impact on the patient's phenotype, provides a stark challenge to interpreting the effects of genetic variation. Figure 2.1, for example, shows a summary of single nucleotide variants discovered through targeted sequencing of the genomic region containing the *FMR1* and

AFF2 loci in 144 boys with autism. A striking aspect of the figure is the enrichment for rare variants not seen before in public databases like dbSNP (Figure 2.1). Population genetic models predicted this pattern, and recent genome-wide empirical studies have clearly established the vast excess of rare, previously unseen variants in human populations. As a result, to gain sufficient statistical power to identify genetic variants contributing to complex diseases, very large patient collections, on the order of 10,000, are likely to be required (Kiezun et al., 2012).

Another approach that could help meet the statistical challenges of association studies is biological network and pathway analysis. Increasingly knowledge of biological pathways can enable statistical methods that leverage this information to discern associations between patterns of genetic variation and gene networks or pathways. Holistically testing for pathways and networks between genes across the groups being compared may improve power for associating genes to disease phenotype than the single variant comparison approach (Sun, 2012). Still, these approaches presuppose knowledge of important pathways, and may not be the best way to uncover novel pathways or the action of mutant alleles that act outside of canonical pathways.

Finally, it is worth noting that the ultimate demonstration of causation will almost certainly fall beyond purely statistical methodologies. It may become necessary, and important if we are to understand fundamental disease mechanisms, to perform direct functional testing of variants *in vitro* or in model organisms *in vivo*. These experiments are far lower throughput than the original sequencing at the present time and likely

represent a future bottleneck in our efforts to understand the genetic contribution to complex diseases. Furthermore, as we explore more deeply traits influenced by the actions of many genes, we will also need to more carefully examine the effects of environmental variation on the human traits of interest. In essence, DNA sequence and variation information is context-dependent, and to understand mechanisms of disease, we would ideally perform studies that can take into account both genomic variation information and putative environmental exposures.

2.6 Conclusion

The dramatic increase in the whole-exome and whole genome sequencing of large numbers of individuals has revealed more genetic variation between individuals than was previously suspected, as well as evidence for a higher incidence of rare and private variations in individuals within subpopulations (Keinan & Clark, 2012; Tennessen et al., 2012). In a recent review on genetic variability among humans, Olson emphasized that, although a number of different evolutionary and demographic forces act to influence human genomic variation, population genetics studies and, more recently, deep sequencing point to mutation-selection balance as having the greatest impact on the genetic predisposition to disease (M. V. Olson, 2012).

GWAS studies of neuropsychiatric disorders have unequivocally shown that common variants with large effects do not underlie schizophrenia or autism. While statistical analyses of these complex disorders are consistent with the action of a very large number of common alleles of small effect, they are unable to account for the entire estimated

heritability of the disorders. At the same time, while rare pathogenic CNVs can account for nearly all Mendelian forms of complex neuropsychiatric illnesses, because they are so rare in the general population, alone they cannot explain all of the missing heritability. Now, with the rapid advances and reduced costs of whole-genome sequencing, human geneticists will finally be able to more comprehensively uncover all classes of human genetic variation in large patient populations. Sifting through these enormous datasets will undoubtedly pose a stiff challenge for human geneticists, particularly given the tremendous heterogeneity and complexity underlying neuropsychiatric illnesses like autism and schizophrenia. There is little doubt that better integration of genomics with collaborative studies in physiology, biochemistry, and epidemiology is vital if we are to truly understand disease mechanisms and develop innovative methods of prevention and treatment for these devastating disorders.

2.7 Appendix

The following appendix contains a table and figure already referenced in the text of the chapter.

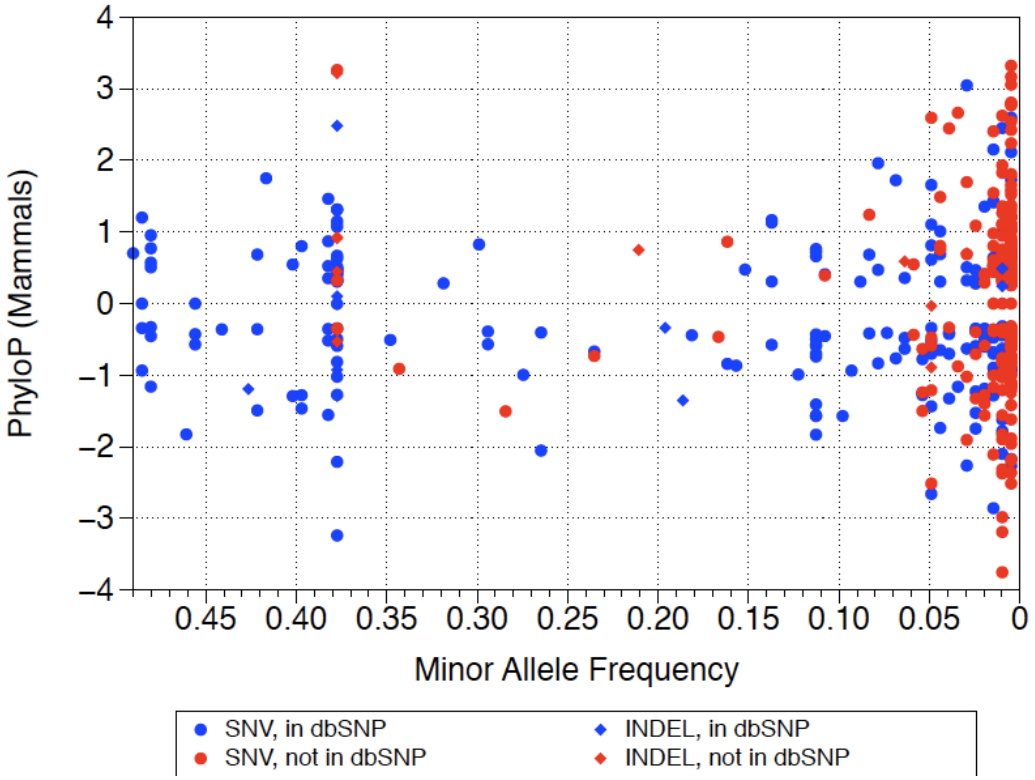


Figure 2.1. Summary of single nucleotide variant (SNV) and insertion/deletion (indel) variation discovered at the FMR1 and AFF2 loci in males with autism spectrum disorder. The frequency of SNVs and indels (minor alleles) in cases is plotted against their level of evolutionary conservation. Most common variation has already been discovered and exists in public databases like dbSNP (blue; circles and diamonds). In contrast, most of the rare variation at both loci was not contained in public databases (red; circles and diamonds).

Error Rate	Expected Number of Errors (3000-Mb human genome)	Expected Number of Variant Sites (per human genome)	Expected Proportion of False-Positive Variant Sites
1.0E-3	3,000,000	3,000,000	0.5
1.0E-4	300,000	3,000,000	0.1
1.0E-5	30,000	3,000,000	0.01
1.0E-6	3,000	3,000,000	0.001
1.0E-7	300	3,000,000	0.0001
1.0E-8	30	3,000,000	0.00001
1.0E-9	3	3,000,000	0.000001

Table 2.1. Expected number of errors in human whole-genome sequencing

Chapter 3

SeqAnt 3.0: Revisions and updates on sequence annotation web application

Thomas Wingo, Matthew Ezewudo, Alex Koltar, Robert Petit, David Cutler, Michael Zwick

Under review for submission to a journal

3.1 Introduction

Improvements in next generation sequencing technology platforms have greatly increased the ease and affordability of generating genome sequence data for a wide range of organisms. There has been an unprecedented drop in the cost per base sequenced, using these platforms (see review in (Shendure & Ji, 2008)). This development parallels improvements in targeted DNA sequencing of functional regions of genomes of complex eukaryotes (see review in (Mamanova et al., 2010), and all of these improvements have increased the chance of finding variants that underlie genetic disorders.

A combination of the aforementioned factors, a better appreciation of the role of rare variants in complex diseases (McClellan & King, 2010; Pritchard & Cox, 2002) and the possibility of identifying these rare genetic variants through direct sequencing (Ezewudo & Zwick, 2013) clearly suggest that direct sequencing will become a mainstay of genetics research. Processing the vast quantities of sequencing data and ascribing meaning to sequence variants remains a challenge for investigators and clinicians alike. The

challenge lies with dealing with the enormous amount of high throughput data from next generation sequencing platforms and efficiently analyzing this data and the possible millions of variants detected, to make sensible functional inference, ultimately linking identified variants to disease phenotypes.

As an answer to the challenge of analyzing next generation sequence data, we had developed SeqAnt (Shetty et al., 2010) a customized and user-friendly web application that speedily annotates extensive genome data within minutes to identify variant types, the region of the genome they fall in, their functional relevance, and how conserved those regions are. There are a number of sequence annotation tools that have been developed in the intervening period since the development of SeqAnt, such as ANNOVAR (K. Wang et al., 2010), SeattleSeq Annotation (Ng et al., 2009) and GEMINI (Paila, Chapman, Kirchner, & Quinlan, 2013). SeqAnt is user friendly, includes reference genomes and annotation tracks for a variety of organisms into *Homo sapiens,* and can be used even in laboratories without dedicated bioinformatics specialists.

Most importantly, performing annotation using this tool takes less than an hour to annotate more than 3 million variant positions from the human genome making it sufficiently fast (Shetty et al., 2010). Subsequent improvements to the annotation tool (Ezewudo et al., 2012) include additional eukaryotic genome tracks, a new conservation score track, as well as an overhaul of the web interface and result files.

Here, we present a major revision of the software. With this version, we have created an extensible framework that accommodates an arbitrary number of site-based information tracks and score tracks, handles non-human genomes, and further improves the speed of annotation processing, while keeping the memory footprint low for the annotation process. These changes put together will enhance the usability of the software while offering an automated and updated database pulled from current annotation and variation tracks from the UCSC genome browser (Sanborn et al., 2011).

3.2 Implementation

3.2.1 Overview

The SeqAnt database is generated from custom genome builds and annotation tracks downloaded and processed from the UCSC genome browser web page. The current SeqAnt databases include the most updated UCSC genome builds for: humans (hg38 and hg19), mouse (mm10), zebra fish (danRer7), housefly (dm6), nematode (ce10) and yeast (sacCer3). There are a set of different tracks for gene annotation, SNP annotation, and conservation scores for each of the genome builds. These tracks are adapted into the new database format either within the Kyoto platform (for gene and SNP annotations) or as randomly accessible binary files (for the sequence and conservation scores data). A detailed description of the different genome builds and associated annotation tracks is shown in Table 3.1. This new revision accepts input files in the Variant Call Format (VCF) or a snpfile format file which contains chromosome positions for all variants within the genome of the model organism that is to be annotated. It could also easily

annotate an input of just a range of base positions for any given genome track within the SeqAnt database. The architecture of SeqAnt which is shown in Figure 3.1 is based on three layers: the user interface that accepts the input files and input command, the annotator that processes the commands and return results to the user, and the database that stores the genomic and annotation data for all the model organism, which is queried by the annotator scripts on the user's command. The tool could also be implemented in the UNIX operating system command line and the scripts are contained in the publicly available distribution on Github.

3.2.2 Database Platforms

The SeqAnt database runs on two platforms: an implementation of the Kyoto DB (KCD)® database and a library of binary files with information that spans the genome of model organisms.

The nature and extent of genome and annotation data being stored in the SeqAnt database determines the type of database used to store the annotation data. Data that does not span the entire genome of model organisms such as SNP, gene annotation, and clinical variation data are referred to as sparse track data and are stored within the KCD database platform for easy retrieval. Genome-wide information data such as the conservation scores and the nucleotide bases for each genomic position are referred to as genome-sized track data and are stored in individual binary files that are randomly accessed when queried in a quick and efficient manner.

3.2.3 Gene Annotation Track

SeqAnt currently uses the Knowngene (for the hg38, hg19, and mm10 genome builds) and refGene (for the other genome builds) UCSC tracks for gene annotation. These tracks detail information on genes belonging to a particular genome build and are obtained by querying the UCSC mysql database for the respective genome builds and extracting information on: the gene names, chromosomal position, Refseq ID, exon positions of the genes to be translated as well as the exon frames. All of this information is stored as a collection of document records in a Kyoto database.

3.2.4 SNP Annotation Tracks

The different variant sites for any particular genome are represented in the UCSC SNP annotation tracks. The hg38 and hg19 currently have the SNP141 track, while the mm10 has the SNP138 track. The remainder of the genomes of non-mammals in SeqAnt does not have a UCSC variation track. The tracks are obtained by querying the UCSC mysql database extracting fields for: the chromosome name, the position on the chromosome where the variant falls in, the allele frequency count for the position, the different alleles in that position and the name of the variant in the public SNP database All of this information is stored as a collection of document records in a Kyoto database platform.

3.2.5 Clinical variations Annotation Tracks

We also implemented the clinical variation (Clinvar) annotation tracks for hg38 and hg19 in this version of the software. Data for these tracks include collated information on SNVs that have been shown in past studies to have clinical relevance. They are displayed in a summary spreadsheet on the NCBI Clinvar website. We obtained the data by

downloading the summary file from NCBI and then parsing the data for relevant fields, which includes the clinical review status of the variant position, the SNP ID, the cytogenic region it falls it, the review status or number of submitters that had worked on same variant and the ID of the phenotype associated with the variant. The information is likewise incorporated into the Kyoto database platform.

3.2.6 Conservation Score Tracks

There are two UCSC conservation tracks implemented in the current version of SeqAnt; phastCons scores and phyloP scores (Siepel et al., 2005). Both are represented in wiggle format files that store large expanse of conservation scores. phastCons scores represent the probability of negative selection on a given site, and uses a hidden Markov model to determine if a particular nucleotide base falls within a region conserved element or a block of conserved region phastCons considers both the individual position and its flanking regions. phyloP scores on the other hand is a measure of the conservation of individual nucleotide positions in the genome, while ignoring the neighboring positions. The absolute values of phyloP scores represent the negative log of the p-values of the positions under a null hypothesis of neutral evolution. So phastCons scores are more suited to detect conserved non-coding regions across genomes in a phylogeny, while phyloP is more suitable for evaluating signatures of selection at particular nucleotide or class of nucleotide sites (Felsenstein & Churchill, 1996; Siepel et al., 2005). For the hg38 genome build, the conservation scores include the phastCons and phyloP scores obtained from multi-genome alignments of 7 vertebrate genomes and the human genome. The hg19 conservation scores consist of phastCons and phyloP scores from multi-alignment

of 99 vertebrate genomes and the human genome. The mm10 conservation scores include phyloP and phastCons scores from multi-alignments between 59 vertebrate genomes and the Mouse genome. The danRer7 build conservation scores include phastCons and phyloP scores from multi-genome alignment of 7 different genomes and the Zebra fish genome. The dm6 conservation scores include phastCons and phyloP scores from alignment of 26 insect genomes with the *D. melanogaster* genome. The ce10 conservation scores include phyloP and phastCons scores from multi-alignments between 6 worm genomes and the *C. elegans* genome. Finally, the sacCer3 build conservation scores include phastCons scores from multi-genome alignment of 7 different Yeast genomes. Data from these different tracks are downloaded from the UCSC browser and preprocessed into binary files that are loaded at runtime within the SeqAnt architecture to speed annotation.

We also implemented C-scores from the combined annotation dependent depletion (CADD) tool, to this version of SeqAnt. The CADD tool is a system of scoring for deleteriousness of variant positions across the human genome. The scores were developed by Kircher et al(Kircher, Witten, Jain, O'Roak, & Cooper, 2014), and represent an integration of a number of conservation score tracks (including GERP, SIFT, Polyphen, phastCons and phyloP), to generate robust inference of the deleteriousness of any given single variant, across the entire genome. We implemented the most recent version-- CADD v1.2, which was predicated on the human genome GRCh37/hg19 build.

3.2.7 Prokaryotic Annotation

Another novel feature of this revision of the software is the inclusion of a prokaryotic annotation capability to the tool. Researchers increasingly use molecular and sequencing approaches to study microbes especially pathogenic bacteria (Ezewudo et al., 2015; Joseph et al., 2012; Vidovic et al., 2014). One of the challenges has been the ability to annotate these prokaryotic genomes to garner information on genetic variants that could underlie different phenotypes of interest. The prokaryotic annotator accepts input file of variant positions in the prokaryotic genome in the VCF format, and an NCBI gene bank format file of the reference genome of the organism and generates an output with annotations for the positions indicated in the input variant file.

3.2.8 Web Interface

The SeqAnt web application has been completely rewritten. It is now a high-performance, single page application (SPA), written in Angular.js, and supported by a Node.js/Express.js high-performance web server. Once compiled, most of the application is held in a single, 1.2-megabyte (MB) javascript file, which is cached by the browser for all future visits. All actions taken in the application, such as switching between pages happen without refresh, with interface performance approaching that of native desktop applications. Additionally, several new features distinguish this revision from SeqAnt 1.0.

The new interface now allows users to register and log in. The application supports uploads in the snpfile format described previously (Shetty et al., 2010) as well as the Variant Call Format (VCF). These formats will have information on all the variants and

chromosome positions that are required to be annotated. Upload progress is displayed in real time by duplex communication using the Node.JS file streaming protocol and XMLHttpRequest transport. The results from the annotation are saved in a zipped folder, which the user can download at any point when logged into the application. The results can also be emailed to the user upon completion. Since the web server now uses an event-driven model, users can log out or close the SeqAnt application as soon as their job is submitted, without interrupting annotation progress. This is a substantial improvement over the original SeqAnt web application (Shetty et al., 2010).

The output from the annotation is a comma delimited text file with a number of result fields spanning information on both the variant type and polymorphism annotation, the gene annotation, information on clinical relevance of annotated position and information on conservation of the given position. Table 3.2 lists out the different result field in the output file of the revised SeqAnt tool. The web application also provides a sample-level graphical summary of the results. A representation of the SeqAnt GUI summary is shown in Figure 3.2.

3.3 Results and Discussions

We performed a number of analyses using the new version the software to test its capabilities. These include: re-running clinically relevant data that had been previously used to establish biological relevance of a rare genetic variant, comparing the efficiency and speed of annotation of the revised SeqAnt and finally using the prokaryotic

annotation feature in the tool to compare genome annotation of a number of prokaryotic genomes from a recent microbial population studies.

FoxP3 mutation associated with inflammatory bowel disease

Previous work done by Okou et al (Okou et al., 2014), identified a novel mutation in the *FOXP3* gene in second-generation family with inflammatory bowel disease. The nonsynonymous X-linked mutation c.694A>C, which is heterozygous in the mother and hemizygous in all the affected sons, was identified running the variants information through SeqAnt using the hg19 genome build. Here, we remapped the variant information to the current hg38 human genome build validate improvements made on this version of SeqAnt.

The annotation results re-confirmed the presence of the mutation in the mother and all the sons as previously described, on chromosome X, with coordinates for the variant position on the hg38 genome build being 49255756. This region is highly conserved, as the phyloP score or the multi-sequence alignment of 100 vertebrates (phylop100way) is 4.197 in the top percentile of all the variants annotated in the sample set. Similarly the CADD score of 24.3 for the mutation is ranked in the top 1% of conserved sites using the newest version (version 1.2) of the CADD scores. These results corroborate the previous findings on the biological significance of this genetic variant.

Big data analysis capability

We implemented analysis of the complete exome sequences of 107 individuals, which include nearly 3 million variant sites, to test the capability of the revised version of this tool to handle large data sets and its performance doing so. The test data we used is available in this public repository: https://github.com/wingolab-org/seq

Our analysis demonstrated that the revised SeqAnt annotates approximately 10,000 variant positions every minute. The complete annotation of the ~3 million variant positions in the input data took under 5 hours. While this speed is a bit reduced in comparison to the previous version, the added detailed annotation information including additional gene annotation names and codes, improved SNP database information and elaborate clinical variation annotation for variants with clinical significance compensates adequately for this.

Annotation of N.gonorrhoeae NGS strains

Previous work by Ezewudo et al (Ezewudo et al., 2015) suggested that some genetic determinants underlie the antibiotics resistance phenotypes present in *N. gonorrhoeae*. We used the prokaryotic annotation feature of SeqAnt to annotate variant positions in genome sequences of 10 isolates used in that study.

We annotated the four strains resistant to the antibiotic cefexime (MUNG4, MUNG15, MUNG14, MUNG20, MUNG4) and four strains sensitive to the antibiotic (MUNG26, ATL0108, ATL0105, ATL0103) from the study using the reference genome FA1090. We analyzed the results from the annotation of each strain for the presence of the mosaic

penA allele, a pattern of mutations on the gene has been shown to underlie resistance to the antibiotic (Ohnishi et al., 2010). Our analysis of the annotation showed the presence of multiple mutations (>12) on the penA locus (NGO1542) for each of the cefexime resistant strains annotated, implying the presence of the mosaic pattern, in contrast with the sensitive strains that have only 3 mutations on the locus. The results here re-confirmed previous observations using molecular approaches about the link between the mosaic patterns on the penA locus and resistance to third generation cephalosporin (Ameyama et al., 2002; Ohnishi et al., 2010; 2011).

3.4 Conclusion

The recent revisions we have made to SeqAnt, has made it currently one of the most user-friendly and efficient NGS annotation tools that is free and publicly available to researchers. We expanded the breadth of model organism genomes that could be annotated using this tool, creating more options for studies using these organisms. We also enhanced the user interface, increasing the ease and accessibility of the tool for non-bioinformatics specialist personnel or laboratories.

The ease of generating NGS is likely to keep improving, with decreased sequencing costs. Analyzing and interpreting the significance of the enormous amount of data from these projects will therefore be of even more importance in the near future. The capacity of SeqAnt to analyze enormous NGS data and present results in an intelligible manner will be quite useful for geneticists and genomics researchers going forward.

3.5 Appendix

The following appendix contains tables and figures already referenced in the text of the chapter.

Genome Build	**Gene Track**	**SNP Track**	**Conservation Scores Track**
hg38 (*Homo sapiens*)	KnownGene	SNP141, Clinvar	phastCons, phyloP, CADD
hg19 (*Homo sapiens*)	KnownGene	SNP141, Clinvar	PhastCons, phyloP, CADD
mm10 (*Mus muculus*)	KnownGene	SNP138	phastCons, phyloP
danRer7 (*Danio rerio*)	RefGene	---	phastCons, phyloP
dm6 (*D. melanogaster*)	RefGene	---	phastCons, phyloP
ce10 (*C.elegans*)	RefGene	---	phastCons, phyloP
sacCer3 (*S. cerevisciae*)	XenoRefGene	---	phastCons

Table 3.1 List of genome tracks for model organisms in SeqAnt database. Tracks were obtained from the UCSC genome browser, the NCBI Clinvar database and the CADD.gs.washington.edu web server.

Field Name	Description
Chr	Chromosome Name
Pos	Chromosome Position
Ref_base	Reference Genome Nucleotide base
Genomic_annotation_code	Interpretation of genomic annotation
Annotation_type	Annotated variant type
Codon number	Number of annotated codon
Codon Position	Position of Codon on transcript
Error_code	Annotation warning messages
Minor_allele	Alternate nucleotide base
New_aa_residue	New amino acid from base change
New_codon_seq	New codon from base change
Ref_aa_residue	Original amino acid residue on reference
Ref_codon_seq	Original codon on reference
Site_type	Annotated site type
Strand	Direction of annotated sequence
Transcript_id	Name of transcript
Snp_Id	Name of annotated SNP
PhastCons	phastCons conservation score
PhyloP	phyloP conservation score
Alt_names	Alternative annotation identification
Snp_features	Additional SNP annotation information

Table 3. 2 Representation of the various fields in the output file from SeqAnt Annotation

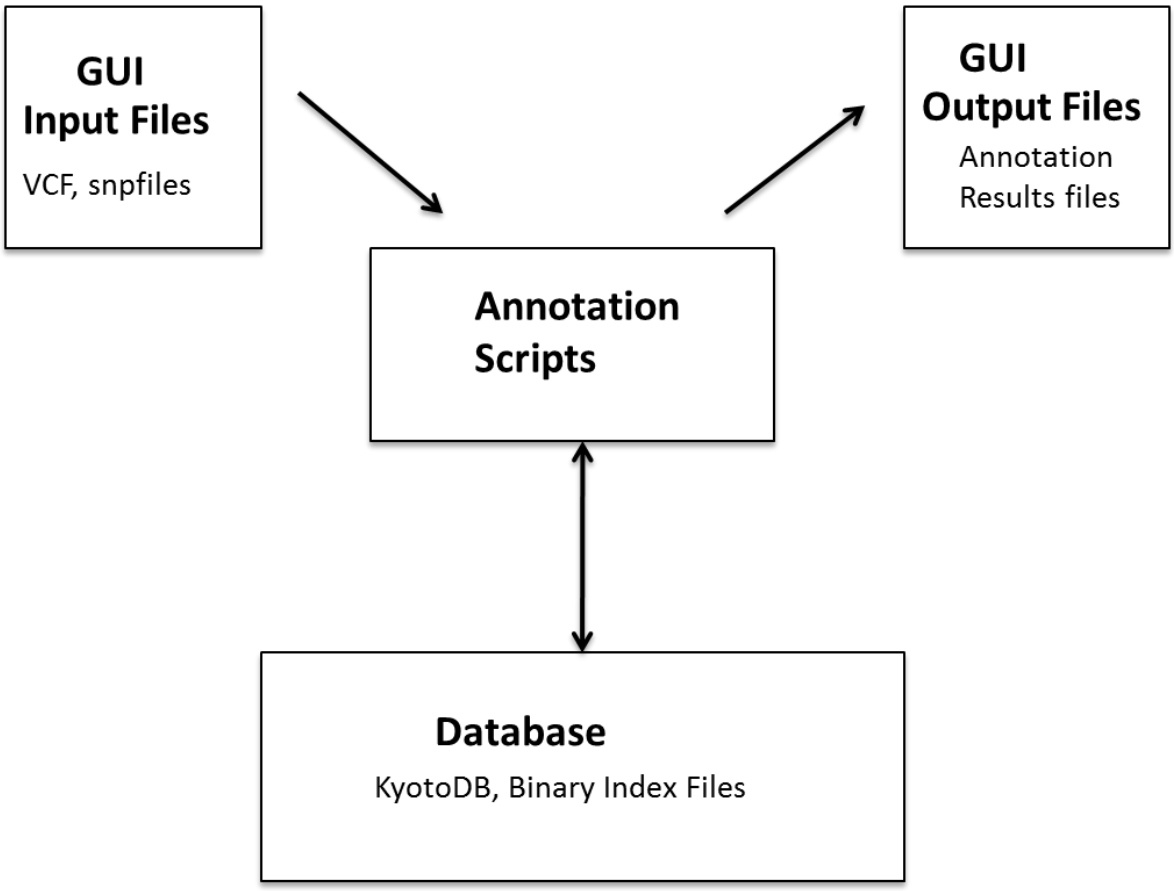


Figure 3.1 SeqAnt Architecture. The binary index files in the database stores conservation scores data while the Kyoto DB platform stores gene and SNP annotation data.

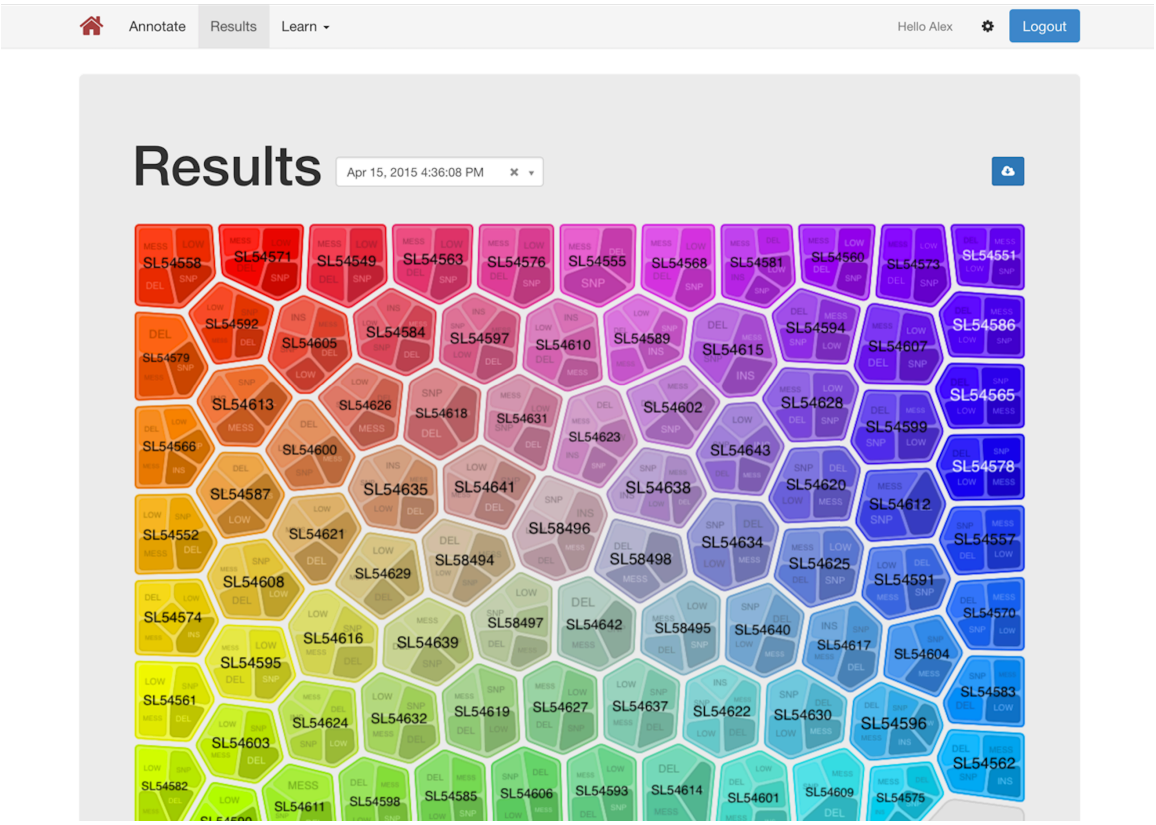


Figure 3.2 Screenshot of SeqAnt webpage display of summary of annotation results

Chapter 4

Population structure of *Neisseria gonorrhoeae* based on whole genome data and its relationship with antibiotic resistance

Ezewudo MN, Joseph SJ, Castillo-Ramirez S, Dean D, del Rio C, Didelot X, Dillon J, Selden RF, Shafer WM, Turingan RS, Unemo M, Read TD

Modified from PeerJ 3:e806 https://dx.doi.org/10.7717/peerj.806

4.1 Introduction

Neisseria gonorrhoeae, a Gram-negative bacterium, causes gonorrhea, the most common bacterial sexually transmitted infection (STIs) causing more than 106 million cases per year globally (World Health Organization (WHO), 2012). The only effective option for treating the disease and stopping its spread has been the use of antimicrobial therapy. Currently, there is no vaccine to prevent infection. Antimicrobial treatment options have diminished over time due to the progressive emergence of antimicrobial resistance (AMR) to drugs previously used to treat gonorrhea and the paucity in the development of newer antibiotics that could effectively eradicate the pathogen (Ohnishi et al., 2011; Unemo & Shafer, 2014).

AMR evolution should be considered in the context of the genetic structure of the *N. gonorrhoeae* population. Early work by O'Rourke *et al.* using electrophoretic analysis of enzymes of the pathogen and serological methods suggested that gonococci have a non-

clonal sexual or panmictic population structure (O'Rourke & Stevens, 1993). More recent studies have also suggested high rates of recombination within the *Neisseria* genus (Didelot & Maiden, 2010). High levels of recombination could confound studies of the gonococcal populations, especially if the studies are based on few genetic loci within strains as compared to the entire genomes. Recent multi-genome studies have focused on either a restricted geographic region (Vidovic et al., 2014)(genomes also included in present studies or on a small subset of the *N. gonorrhoeae* population (Grad et al., 2014). Hence, there is a need for studies aimed at understanding the global *N. gonorrhoeae* population structure at the whole genome scale.

Past AMR studies using limited numbers of gonococcal strains from specific geographic regions of the globe have mostly focused on a number of representative genes or genetic regions of the genome to elucidate underlying mechanisms of antibiotic resistance (Hagman & Shafer, 1995; Lindberg, Fredlund, Nicholas, & Unemo, 2007; Ohneck et al., 2011; Thakur, Levett, Horsman, & Dillon, 2014; Tomberg, Unemo, Ohnishi, Davies, & Nicholas, 2013; Unemo & Shafer, 2014; Unemo, Golparian, & Hellmark, 2014; Unemo et al., 2012; Zhao et al., 2009). Extensive genome sequencing studies have yet to be conducted on a diverse collection of strains from different geographical locations and collected over longer time periods. Our approach in this study builds on recent multi-genome studies (Grad et al., 2014; Vidovic et al., 2014), with the goal of using whole genome analysis to elucidate two processes: 1) the population structure and dynamics of *Neisseria gonorrhoeae* and 2) the correlation between this population differentiation and AMR evolution in gonococci. Our genome analysis of strains from multiple sites across

the world offers a geographic diversity of *N. gonorrhoeae* isolates, providing more depth in genome-wide studies of this pathogen and identifying possible sub-populations impacting AMR and evolution within the species.

4.2 Materials and Methods

4.2.1 *Neisseria gonorrhoeae* isolates

Sixty-one *N. gonorrhoeae* isolates of diverse origin were obtained. These included isolates from the Gonococcal Isolation Surveillance Program (GISP) site covering Atlanta, Miami, New York city and North Carolina in the United States (n=21), from Canada (primarily, Saskatchewan) and Chile (n=1) (Vidovic et al., 2014)) (n=24), and from WHO global collaborations; Sweden (n=7), Norway (n=3), Japan (n=2), Austria (n=1), Pakistan (n=1), Philippines (n=1), and Australia (n=1). Phenotypic determination of the minimum inhibitory concentrations (MICs) of all isolates was performed using the agar dilution method or the Etest method (bioMerieux), according to the instructions from the manufacturer. The strains sequenced in this study were tested for resistance to primarily three antibiotics, tetracycline, azithromycin and cefixime, with breakpoints for resistance set at 2, 2.0, and 0.25 μg/mL, respectively, based on the CDC MIC (minimum inhibitory concentration) breakpoints for testing in the GISP protocol (http://www.cdc.gov/std/gisp/gisp-protocol07-15-2010.pdf). Antibiotic resistance profiles of the Canadian strains have been previously reported (Vidovic et al., 2014). Details of

the different isolates with their NCBI accession numbers are presented in Table 4.1.

4.2.2 Sequence generation and assembly

The *N. gonorrhoeae* strains were shotgun (WGS) sequenced using the Illumina HiSeq™ instrument, utilizing libraries prepared from 5 μg of genomic DNA for each sample. The average sequencing coverage was 225. The sequencing reads were filtered using the prinseq-lite algorithm (Schmieder R. et al, 2011) to ensure only sequence reads with average phred score ≥30 were used. The reads for each project were then assembled *de novo*, using the velvet assembler program (Zerbino & Birney, 2008). The optimal kmer length for each assembly prior to assembly was determined using the velvet optimizer algorithm (Gladman & Seemann, 2012). Data was deposited in the NCBI Sequence Read Archive public database (Accession # SRA099559) (Table 4.1). For this study, we included an additional 14 draft genome sequences of *N. gonorrhoeae* strains, downloaded from the NCBI draft genomes database (NCBI Bioproject numbers: PRJNA55649, PRJNA55651, PRJNA55653, PRJNA55905, PRJNA46993, PRJNA55657, PRJNA55655, PRJNA55659, PRJNA55661, PRJNA55663, PRJNA55665, PRJNA55667, PRJNA55669, PRJNA55671, and the reference genome sequence Ref_FA_1090 (NC_002946.2)

4.2.3 Genome-wide phylogeny and pangenome analysis

The assembled genomes were annotated individually using the NCBI PGAP annotation pipeline to give predicted proteome for each of the strains. The orthologs were

determined by OrthoMCL (L. Li, Stoeckert, & Roos, 2003), which uses bi-directional BLASTP scores of all the protein sequences to perform Markov clustering in order to improve sensitivity and specificity of the orthologs. For the OrthoMCL analysis, we used a BLASTP E-value cut-off of 1e-05, and inflation Markov clustering parameter of 1.5. Core genes were defined as the orthologous genes that are shared among all the *N. gonorrhoeae* strains used in this analysis

The nucleotide sequences of all the core genes were concatenated together and core-gene nucleotide alignment was conducted using progressive MAUVE (A. C. E. Darling, Mau, Blattner, & Perna, 2004). Similarly, whole genome amino-acid alignment was also generated by concatenating the deduced amino-acid sequences of all the core genes generated using MUSCLE(Edgar, 2004), and to form a super protein alignment. Homoplasious sites were removed from the whole-genome nucleotide alignment using the Noisy software (Dress et al., 2008). The protein alignments were filtered by GBLOCKS (Talavera & Castresana, 2007)using default settings to remove regions that contained gaps or were highly diverged. A maximum likelihood (ML) tree from the same data set was created using the GTR and JTT substitution models for the nucleotide and protein alignment respectively and the GAMMA evolutionary model (Stamatakis, 2014). The majority rule-consensus tree was generated from 200 bootstrap replicates of the model. Linear regression of the root-to-tip distances against the year of isolation was performed using the Path-O-Gen tool (http://tree.bio.ed.ac.uk/software/pathogen/).

4.2.4 Multi-locus sequence typing (MLST) locus analysis

MLST is a genotyping tool for *Neisseria* based on sequencing of 7 core housekeeping genes (Jolley & Maiden, 2010). There are currently close to 11,000 individual *Neisseria* sequence profiles in the publicly available MLST database (http://pubmlst.org). We utilized a custom python script mlstBLAST.py (http://sourceforge.net/projects/srst/files/mlstBLAST/) to perform a BLAST search of these genes across all the strains in our data set and identified the sequence type (ST) for each strain. Novel alleles of the locus and STs were submitted to the MLST database. A phylogeny of the concatenated DNA sequences of all the *N. gonorrhoeae* STs in the MLST public database was created using the neighbor joining distance matrix approach of the PHYLIP(Felstein, 1989). Mean nucleotide distance for the sequence alignments and MLST genes was computed using MEGA software (Tamura, et al 2013).

4.2.5 Estimating population parameters and homologous recombination

ClonalFrame (Didelot & Falush, 2007) utilizes a statistical framework to reconstruct the clonal genealogy as well as identify the regions along the genomes that has been affected both by recombination and mutation. The model uses a Bayesian approach to predict the phylogenetic relationship in the sample set, given the whole genome sequence alignment data. The input genome alignment data was the core genes (n = 1189) alignment generated from MAUVE. Four independent ClonalFrame runs were performed for 40,000 iterations, with the first 20,000 discarded as burn-in. This allowed the model parameters to converge and each of the 4 runs were checked for the consistency of the estimated parameters as well as the consistency of the topology of the inferred clonal genealogies.

4.2.6 Population structure analysis

The program BAPS (Bayesian Analysis of Population Structure) version 5.3 (Corander & Marttinen, 2006; Tang et al., 2009)was used to infer the underlying population structure of the 76 *N. gonorrhoeae* strains in the sample set. SNPs from the core MAUVE alignment, with gaps removed were converted to a BAPS input file, which is a representation of all the polymorphic loci in the multi-sequence alignment. BAPS applied a Bayesian model to predict the likelihood of a population structure given the input data and non-parametric assumption approach to trace ancestry of the different individuals in the sample set. For the mixture analysis we used the 'Clustering of individuals' approach. We ran a preliminary analysis to evaluate the approximate number of genetically differentiated groups using a vector from 2 to 40 K values, where K is the maximum number of groups. Given that 5 groups was the K value with the best log likelihood, we ran a second analysis using from 3 to 7 K values and again the best K value was 5 groups. We used the 'Admixture based on mixture clustering' module for the admixture analysis. For the analysis; the minimum population and the admixture coefficient for the individuals was then set to 5. For the reference individuals from each population and the admixture coefficient for reference individuals we used the values as described by Castillo-Ramírez et al (Castillo-Ramírez et al., 2012). In addition, population structure analysis of the sample set using the fineSTRUCTURE tool (Lawson, Hellenthal, Myers, & Falush, 2012) was performed. fineSTRUCTURE analysis was a two-step process-1) ChromoPainter algorithm was used to generate the co-ancestry matrix from the genome-wide haplotype data using the linkage model. 2) The fineSTRUCTURE algorithm

performed a model-based clustering using a Bayesian MCMC approach to predict the likelihood of a population structure given the input data and non-parametric assumption approach to trace ancestry of the different individuals in the sample set. The fineSTRUCTURE approach was used to corroborate the findings from the BAPS population structure analysis.

4.2.7 Mapping the movement of DNA between *Neisseria gonorrhoeae* clades

We traced the flow of recombination between strains into five different subgroups in the phylogeny determined from the subgroups of the population defined by the BAPS analysis. We created a BLAST database of the whole genome sequence of all 76 strains in the sample set and included 14 whole genome sequences of all other *Neisseria* species (NC_008767.1, NC_014752.1, NC_017512.1, NC_017516.1, NC_003112.1, NC_010120.1 NC_017501.1, NC_017514.1, NC_017517.1, NC_003116.1, NC_013016.1, NC_017505.1, NC_017515.1, NC_017518.1) that are present in the NCBI database. Next, we performed a BLASTN search for each of the genomic region within the strains identified by ClonalFrame to be under recombination, selecting the best hit within the sequences in the database we created, with an identity of >98%, to be the source of the recombined region. We also removed BLAST matches found in strains from similar subgroups as the source of the recombined region. We used the *migest* package (http://cran.r-project.org/web/packages/migest/) implemented in the R statistical language to create a circular representation of the matrix of relationship between the subpopulations identified by BAPS based on the purported recombination between strains in the different subgroups. We also supplied *migest* with the matrix from BAPS

admixture analysis and recreated the circular flow of recombination across only the subpopulations as defined by BAPS.

4.2.8 Comparison of nucleotide substitution rates

Amino acid sequences were aligned using MUSCLE sequence aligner(Edgar, 2004). The amino acid sequence alignment was converted to nucleotide alignment based on the corresponding gene sequence using PAL2NAL(Suyama, Torrents, & Bork, 2006) and we implemented the YN00 method of the PAML package (Z. Z. Yang, 2007) to calculate the pairwise dN/dS ratios for the strains (Rocha et al., 2006). The contribution of each strain to the overall variation in the dN/dS rates across the sample set was estimated using ANOVA (Analysis of Variance) approaches.

4.2.9 Analysis of positive selection

For the analysis of positive selection within core genes of the strains in the sample set, we first identified and removed core genes that have signals of homologous recombination using three methods of Pairwise Homoplasy Index (PHI), Neighbor Similarity Score (NSS) and the maximum χ^2 method. The three methods are implemented in the PhiPack package (Sawyer, 1989). A window size of 50 nucleotides was used to run the methods in the package, and genes shown to have significant probability of homologous recombination by a majority of the methods were not used for the positive selection analysis. Next, we identified core genes under positive selection using codeml of PAML tool version 4.7 (Z. Z. Yang, 2007). We applied the branch-sites test for positive selection Model A test 2 of the tool, to identify genes under positive selection population groups.

For each of the clades, we performed the Likelihood Ratio Test (LRT) for two hypotheses - the null hypothesis is the existence of neutral selection as implemented in the null model versus the alternative hypothesis implemented in the test model for positive selection. The LRT was performed to a degree of freedom of 1, and we corrected for multiple testing using the False discovery rate approach (FDR)(Benjamini & Hochberg, 1995). We further identified the Gene Ontology (GO) terms and functional characterizations of the genes under positive selection (see Table 4.2) and performed an enrichment test for functionality of these genes using the blast2go test pipeline(Götz et al., 2011).

4.2.10 Confirming known predictors of antibiotic resistance phenotype

We downloaded from NCBI reference DNA sequences of resistance determinants that have been shown in the literature to underlie the resistance phenotype we have observed in our sample set (see Table 4.3), and performed a BLASTN search for each of these DNA sequence regions across all the strains in the database of whole genome sequences. For convenience, the contigs for each assembly were ordered into one pseudocontig after tiling to the reference genome FA1090, using the ABACUS tool (http://abacas.sourceforge.net/).

We selected the top hit (with identity match of 98% or more) for each sequence (strain) in the database and parsed the alignment between the query and the subject sequence in the database for the presence or absence of the underlying resistance genetic mutations as suggested in the literature.

4.3 Results and Discussion

4.3.1 Genome-wide homologous recombination in diverse *N. gonorrhoeae*

We sequenced 61 recent clinical isolates primarily from the US and Canada but also single representatives from other countries, including Japan, Pakistan, Australia, Austria, Philippines, Norway and Sweden using the WGS approach to a high average coverage (average 225-fold read redundancy). *De novo* assemblies based on these data produced a set of contigs that represented draft, unordered representations of the genomes with high sequence quality. A preliminary phylogeographical analysis of the Canadian isolates (n=23) was recently published (Vidovic et al., 2014). For the analysis, we included the 14 *N. gonorrhoeae* draft NCBI genome sequences (12 from the US and 2 from Europe) and the genome sequence of the FA1090 *N. gonorrhoeae* reference strain. The 76 were assigned into 23 previously described MLST STs and four new STs (10931, 10932, 10933, and 10934). The genetic diversity (measured as pairwise nucleotide distances of MLST loci) of the strains in this study was about half that of the *N. gonorrhoeae* strains as a whole (0.001 substitutions per site in our study compared to 0.002 in the large MLST set), and the STs from our sample set were represented across the different clades of a phylogeny of housekeeping genes of *N. gonorrhoeae* strains found in the MLST database (see Fig 5.1). Alignment of the shotgun assembly to reference genome FA1090 (NC_002946.2) yielded 10,962 SNPs in the core region (conserved in all strains). The average per nucleotide diversity in the core genome regions was 0.003.

Homologous recombination is known to play a role in shaping bacteria populations(Didelot & Maiden, 2010). The ClonalFrame tool (Didelot & Falush, 2007) detected 952 independent recombination events, covering more than 50% of the reference genome. The average size of the recombination regions identified was 360 base pairs. The estimate for the ratio of effects of recombination and mutation (*r/m*) for our strain set was 2.2, a relatively high value for bacterial species (Didelot & Maiden, 2010) (and quite similar to the *r/m* estimate of 1.9 based on the whole genome alignment on a less genetically diverse group of *N. gonorrhoeae* strains reported by Grad et al(Grad et al., 2014).

We constructed a maximum likelihood phylogeny of the core genome of the 76 strains using the RAxML program (excluding regions identified as potentially recombinant) (Fig 4.1). This tree had similar topology to the clonal frame that determined by the eponymous software. The tree showed multiple clades but the there was no strong signal of genetic isolation by distance at the continental scale. The rate of the molecular clock was estimated to be 8.93×10^{-6} mutations per year based on the slope of the regression of the root-to-tip divergence with isolation dates (see Fig S2). This value was similar to those obtained in other bacterial studies, ranging from 8.6×10^{-9} to 2.5×10^{-5} (Z. Zhou et al., 2013). However, because the temporal signal was weak in the root-to-tip analysis, we did not use these data for Bayesian phylogeny analysis using the BEAST phylogeny tool (Drummond & Rambaut, 2007).

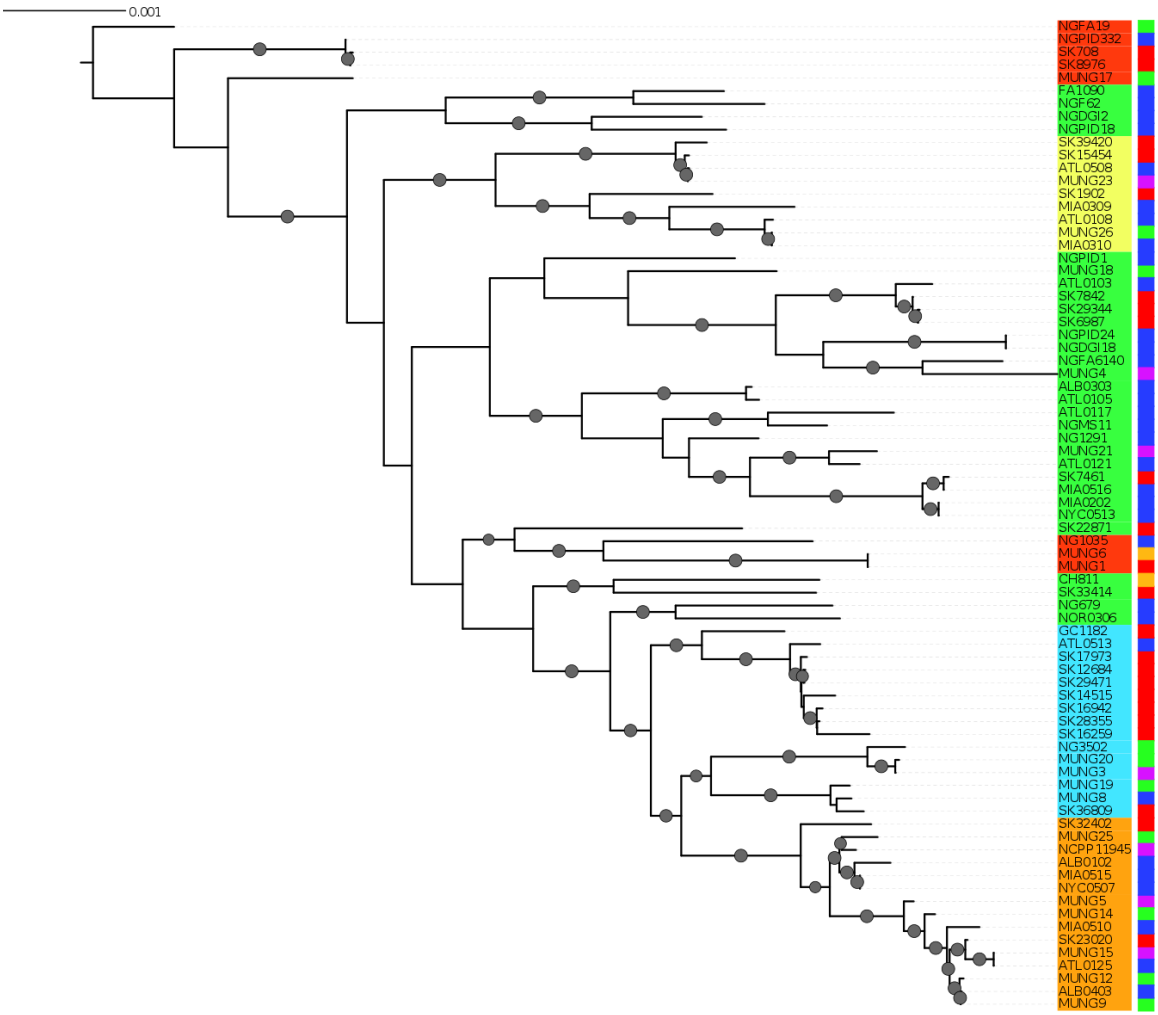


Figure 4.1 Maximum Likelihood phylogeny of strains in sample set. Taxa are colored by the population subgroups defined by BAPS (subgroup1 colored Orange, subgroup 2 Turquoise, subgroup 3 Green, subgroup 4 Yellow and subgroup 5 Red), the tips of the taxa are colored based on location of isolation, Canada is colored red, US is blue, Europe green and Asia purple, tips of strains from Australia and Chile are colored brown

4.3.2 *Neisseria gonorrhoeae* population structure and biogeography

Given that recombination was frequent in these genomes, we sought to evaluate the genetic substructure of the population. We used two complementary methods. BAPS (Tang et al., 2009) predicts the likelihood of a population structure given the input data and uses a non-parametric assumption approach to trace ancestry. fineSTRUCTURE

(Lawson et al., 2012), on the other hand, uses similar methods of predicting population substructure, but to a finer detail and does not assume a prior optimum number of subpopulations (K). The BAPS tool identified 5 subgroups within the *N. gonorrhoeae* population from the strains within the sample set (Fig 4.2). As expected, members with the same subgroup ancestry generally were found near each other when mapped on the ML phylogeny constructed using the non-recombining portion of the genome. On the other hand, fineSTRUCTURE identified 30 genetic subgroups within our sample set. However, individual members of each of the fineSTRUCTURE (supplemental Fig 4.2.1) subgroups belong to the same BAPS subgroup. Each of the five BAPS subgroups contained strains from multiple continents based on geography or location of isolation (Fig 4.1). It was particularly interesting that each BAPS cluster had at least one US strain and one Canadian strain. The BAPS analysis revealed a complex relationship between Group 3 and 5, with the latter group 5 separated into two groups (Fig 4.1). Group 3 strains in clades closely related to group 5 showed significant genetic import from group 5. It is possible that the extent of admixture occurring in group 3 and 5 may have caused misidentification. Also, strains of sequence type (ST) 1901, which is the most abundant ST in our sample set all belong to subgroup 1, hinting at a correlation between BAPS subgrouping and MLST.

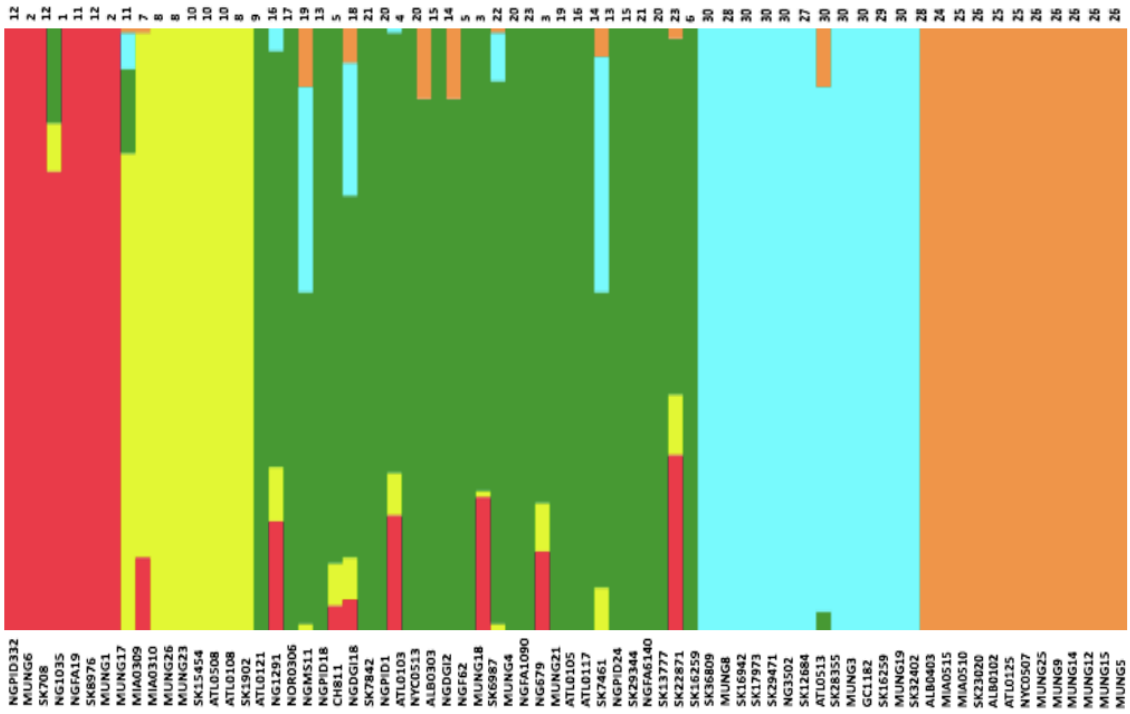


Figure 4.2 Population subgroups from strains or *N. gonorrhoeae* in the sample set defined by BAPS. The names for each strain in the different subgroup are indicated at the bottom of the plot on the x-axis, while the fineSTRUCTURE group labels for each strain is on top of the plot. Each color represents one of the genetically differentiated groups (subgroup1 colored Orange, subgroup 2 Turquoise, subgroup 3 Green, subgroup 4 Yellow and subgroup 5 Red) and each vertical colored bar corresponds to one isolate. When the vertical bars show two colors, each color corresponds to one of the groups. The proportion of every color in the bar reflects the extent of the genetic material coming from the group represented by that particular color.

We assessed patterns of genetic drift effects in the population by estimating the pairwise substitution rates between all the core gene orthologs for the strains and determining the mean dN/dS ratio for each strain. The mean pairwise dN/dS ratios for each strain are shown in Fig 4.3. There was significant variation in the mean dN/dS ratios among the strains (ANOVA p-value = 2.0e-16). The overall mean of the dN/dS estimate was 0.3184, similar to the 0.402 value estimated for the bacterial pathogen *Chlamydia trachomatis* (Joseph et al., 2012). The mean dN/dS ratio for the five subgroups respectively was (0.32412976, 0.33325164, 0.31273103, 0.30952504 and 0.30990092). There is no

significant difference between group mean dN/dS ratios (p-value =0.921, t test of means).

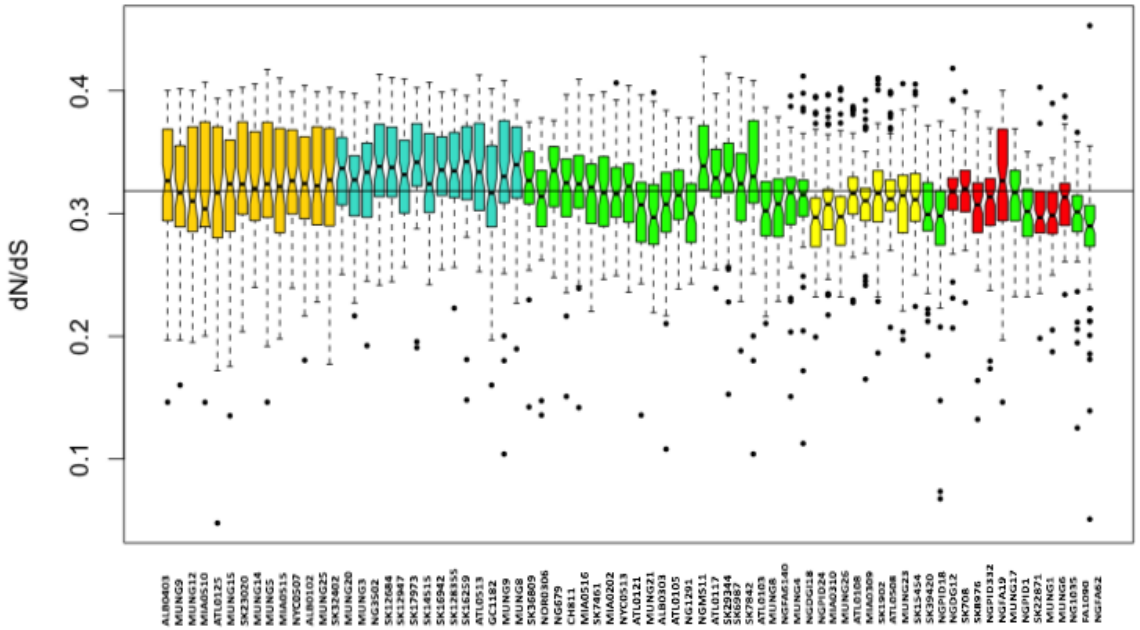


Figure 4.3 Boxplot of mean dN/dS ratio pair-wise comparison of core genes of each of the strains of N. gonorrhoeae in the sample set. The boxplot is colored by subgroups within the Neisseria population, defined by the BAPS tool.

The mean dN/dS ratio for strains from the Canadian region was 0.3279, which was above the overall mean ratio, while that for strains collected in the US was 0.31708, which is below the overall pairwise dN/dS mean ratio for the sample set. This was also a statistically significant difference (p-value 0.0018 for t test of means), suggesting a possible geographical effect within this subset of strains.

4.3.3 Genetic admixture within *N. gonorrhoeae* and with other *Neisseria species*

In order to understand the flow of genetic information between the strains from five

different subgroups defined by the BAPS analysis (Fig 4.2) as well as strains from other *Neisseria* species, we used two independent approaches. The first was to search each of the 952 recombination regions identified by ClonalFrame for a best BLASTN match from another subgroup or *Neisseria* species (We created a blast database of the 76 genomes from this study and representative strains from the *Neisseria* genus). In parallel, we also counted the occurrence of co-ancestry of genetic markers revealed by the BAPS analysis. Both the BAPS and BLAST analyses suggested that group 3 was the most common nexus of homologous recombination between other clades, consistent with its basal phylogenetic status. In the BAPS-based network groups 1 and 2, and to a greater extent, group 5, were primarily DNA donors to group 3 (Fig 4.4B). But this pattern was less visible in the BLAST network (Fig 4.4A). It is notable that more than 90% of the recombination with strains from other *Neisseria* species occurred in groups 2 and 3. Group 5 stood out as a significant source of genetic exchange into strains in group 3.

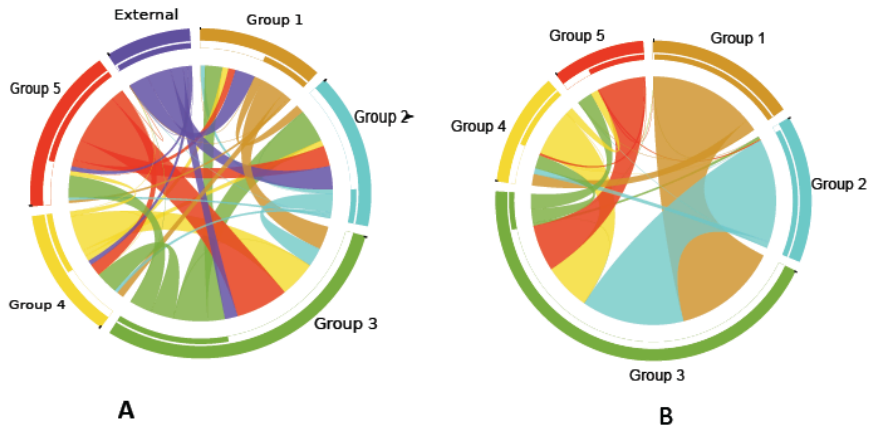


Figure 4.4 Pathways for exchange of genetic materials between populations. (A) Recombination pattern traced from BLAST results of similarity of recombined regions between the subgroups defined by BAPS of *N. gonorrhoeae*. The clade showed as external represents strains from other Neisseria species. (B) Exchange of genetic materials among subgroups within the sample set as defined by BAPS admixture analysis. Colored base sub-sectors of the circle for each subgroup in the diagram

represents outflow of genetic material while blank or white colored sub-sectors represents inflow of genetic materials to the subgroups.

The genetic relatedness of the strains in the sample set or the purported sharing of genetic materials across the different subgroups shown by the BAPS figure paralleled the pattern revealed by the BLAST clonal frame analysis (p-value = 0.048, Mantel test for comparing the distance matrices of the five populations between both methods). The exchange of genetic materials from other *Neisseria* species was not accounted for in the BAPS admixture analysis. Based on the BLAST analysis, the proportion of DNA transferred within *N. gonorrhoeae* compared to arriving from *Neisseria* strains the species was 729 out of 849 intra-specific genetic events. This finding is line with the "fuzzy species" concept of Fraser *et al* (Fraser, Hanage, & Spratt, 2007): while *N. gonorrhoeae* is not sexually isolated, DNA flow seemed predominantly through intra-specific exchanges.

4.3.4 Genes under positive selection

Of the 1189 core genes, we identified 352 genes as likely to contain past recombination histories using the PHIPACK tests (Sawyer, 1989). Thirty-one genes within the subset of 837 non-recombining core genes were found to be under positive selection using the tests implemented by the PAML software (Materials and Methods). BAPS subgroup 5 had the highest number of core genes under selection (14) followed by subgroup 3 (7). While we found no significant enrichment of genes under positive selection in any of the functional classes in the Gene Ontology (GO) database, the functions of the best match proteins from genes under positive selection can be broadly classified to genes involved in DNA

or RNA synthesis of gene expression, membrane or transport proteins, and, to a lesser extent, genes involved in metabolic pathways in the bacterial cell (Supplement Data S2 spreadsheet). Of the 352 genes found to have signals of recombination, we found no significant enrichment of the genes in any of the functional classes in the GO database. The functions of these genes could broadly be classified into 2 groups: genes encoding membrane and transport proteins; and those involved in metabolic pathways in the cell.

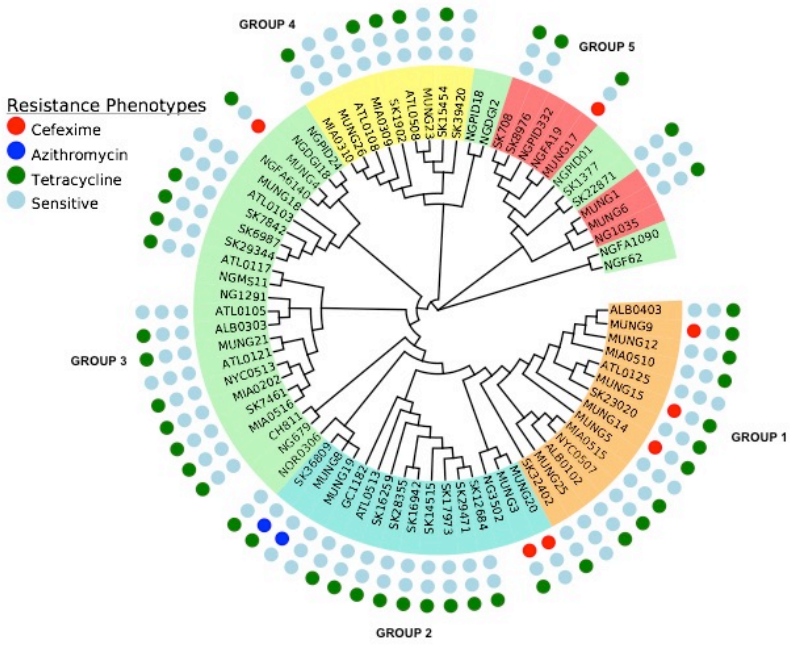


Figure 4.5 Representation of antibiotic resistance profile of *N. gonorrhoeae* strains across different subgroups of the population.

In regard to antibiotic resistance and selection, the most interesting gene found to be under positive selection was *porB* (N. H. Smith, Maynard Smith, & Spratt, 1995), which has been shown to be involved in mechanisms of resistance to penicillin, macrolides, cephalosporin and tetracycline (Unemo & Shafer, 2014). *porB* exhibited signals of

selection in subgroups 1, 3 and 5 - the groups that harbored most of the antibiotic resistant strains in our sample set (Fig 4.5). *comA*, which encodes a membrane protein necessary for competence of *N. gonorrhoeae,* was also found to be under selection in a handful of strains that make up subgroup 5. This finding is of interest in regards to the potential for DNA uptake in these strains, since they appear to be primarily DNA donors, rather than recipients in genetic exchanges (Fig 4.4). Other genes putatively under selection included a stress response gene, a gene encoding a chaperone protein of the HscA family and a number of proteins: ribosyl transferase, RNA polymerase, and an *arsR* family transcriptional regulator, which were all linked to gene expression. Genes under positive selection in subgroup 3 were also mainly involved in gene expression or DNA metabolism, including DNA helicase and tRNA pseudo uridine synthase. Most of the genes with known functions, identified to be under positive selection in subgroup 1 were either membrane-associated or transport proteins.

4.3.5 Analysis of known genetic predictors for AMR phenotypes

A substantial amount of research effort over the past 10 years has been devoted to understanding the genetic basis of drug resistance in *N. gonorrhoeae* (Garvin et al., 2008; MD et al., 2014; Unemo & Shafer, 2011; Veal et al., 2002; World Health Organization (WHO), 2012). Since there is a an increasing interest in the direct attribution of resistance phenotypes based on genome sequencing, we attempted to ascertain how knowledge of existing variants could be applied to the *N. gonorrhoeae* genomes in this study. We searched for variants known to underlie resistance to 3 antibiotics classes within our

study (Table 4.3). In terms of subgroup distribution, tetracycline resistance was found in each of the 5 population subgroups, azithromycin resistance was present in only 2 of the strains tested (SK36809 and MUNG8) and restricted to subgroup 2 (Fig 4.5), and cefixime resistance was found in subgroup 1 and subgroup 2. We identified genes responsible for resistance to the drugs tested in this works using literature searches and the CARD antimicrobial resistance database (McArthur et al., 2013).

The *tetM* resistant determinant, which confers high-level resistance to tetracycline, is borne on plasmids and is transferred either through conjugation or transformation (Knapp, Johnson, Zenilman, Roberts, & Morse, 1988; Morse, Johnson, Biddle, & Roberts, 1986; Turner, Gough, & Leeming, 1999). It was found in only 5 of the 10 strains with high-level resistance to tetracycline (MIC equal or greater than 16 µg/ml). Strain SK1902, one of the 5 strains with the *tetM* determinant, had a significantly higher MIC (> 256 µg/ml) than the rest (see attached Supplement S1). Other strains with reduced susceptibility or chromosomally-mediated resistance to tetracycline, that is without the *tetM* determinant do have other corresponding chromosomal mutations on one or more of the resistance loci: *mtrR* (including its promoter), *penB*, *rpsJ*. Only one strain (ATL0508) within the sample set exhibits resistance to tetracycline in the laboratory, without the presence of any of the known resistance determinants of the tetracycline resistance phenotype.

Different “mosaic” *penA* alleles are thought to have developed from recombination with portions of DNA transferred horizontally from commensal *Neisseria* and or *N.*

meningitidis and underlie decreased susceptibility or resistance to cephalosporin by preventing their binding action on the encoded mosaic PBP2 (Ameyama et al., 2002). The mosaic *penA* XXXIV (Ohnishi, 2011; Grad 2014; Unemo & Shafer, 2014) had the best positive predictive value of all the known resistance determinants we searched for within our dataset, being present in 6/7 of the strains resistant to cefixime. This result echoed the observations made by Grad et al (Grad et al., 2014)in their epidemiologic study of *N. gonorrhoeae* strains. The other loci (i.e., mutations in the *mtrR*, *mtrCDE* operon promoter region and *penB* gene) also proven to enhance the MICs of cephalosporin (Unemo & Shafer, 2011; Warner, Shafer, & Jerse, 2008) did not have a similar predictive property within strains in our data set. These variants were seen in 2 out of 7 and 3 out of 7 cefixime resistant strains, respectively. MUNG17 is the only strain in the sample set that has an elevated MIC (0.38μg/mL) to cefixime that we could not find any of the known resistance determinants within its genome sequence (See attached supplement Data S1).

Resistance to azithromycin can be mediated by mutations in the previously mentioned *penB* and *mtr* operon genes as well as mutations found in the 4 different alleles of the 23S rRNA gene that inhibits protein synthesis (Chisholm et al., 2009; Palmer, Young, Winter, & Dave, 2008; Starnino, Stefanelli, Neisseria gonorrhoeae Italian Study Group, 2009). The 23S rRNA mutation allele was found in one (SK36809) of the two strains with the azithromycin resistance phenotype. The other azithromycin resistant strain, MUNG8, did not have the 23S rRNA resistance determinant or any of the other known mutations in the *mtrR* or *penB* loci (See attached Supplementary Data S1).

4.4 Conclusions

Our study suggested that *N. gonorrhoeae* globally is made up of at least five genetic subpopulations. That individual strains from the subpopulations are from diverse geographical locations confirms the cosmopolitan nature of the pathogen. This suggested a population structure with multiple waves of rapid international expansion. Subgroup 3 strains may be the nexus for gene exchange within the species. Groups 1 and 2 might be the most recently branched and contain a higher proportion of resistant isolates to more currently used antibiotics. Given the importance of the antibiotic resistant phenotype, these may be emerging lineages that are expanding within *N. gonorrhoeae.* It will require a more extensive study with a broader number of strains to ascertain this suggested evolutionary trend. Our analysis confirms earlier studies that showed an appreciable effect of recombination within the population. This could be playing a role in the evolution of AMR in the bacterium, as strains with resistance phenotypes to currently used antibiotics are mostly within similar population sub-groupings.

Although most of the known predictors that underlie the observed resistance phenotypes were accounted for in the strains we studied, they could not explain some of the phenotypes of several strains. These findings suggested that a broader genome search of a large number of whole genomes from strains of this pathogen could yield candidate novel variants that may explain some of the “missing” antibiotic resistance phenotypes we have observed.

In general, large genome sequencing studies examining a high number of temporally and geographically diverse *N. gonorrhoeae* isolates are essential to elucidate the evolution and diversity of the *N. gonorrhoeae* species as well as associations between genomic content, antibiotic resistance and clinical outcome of treatment.

4.5 Appendix

The following appendix contains tables and figures already referenced in the text of the chapter.

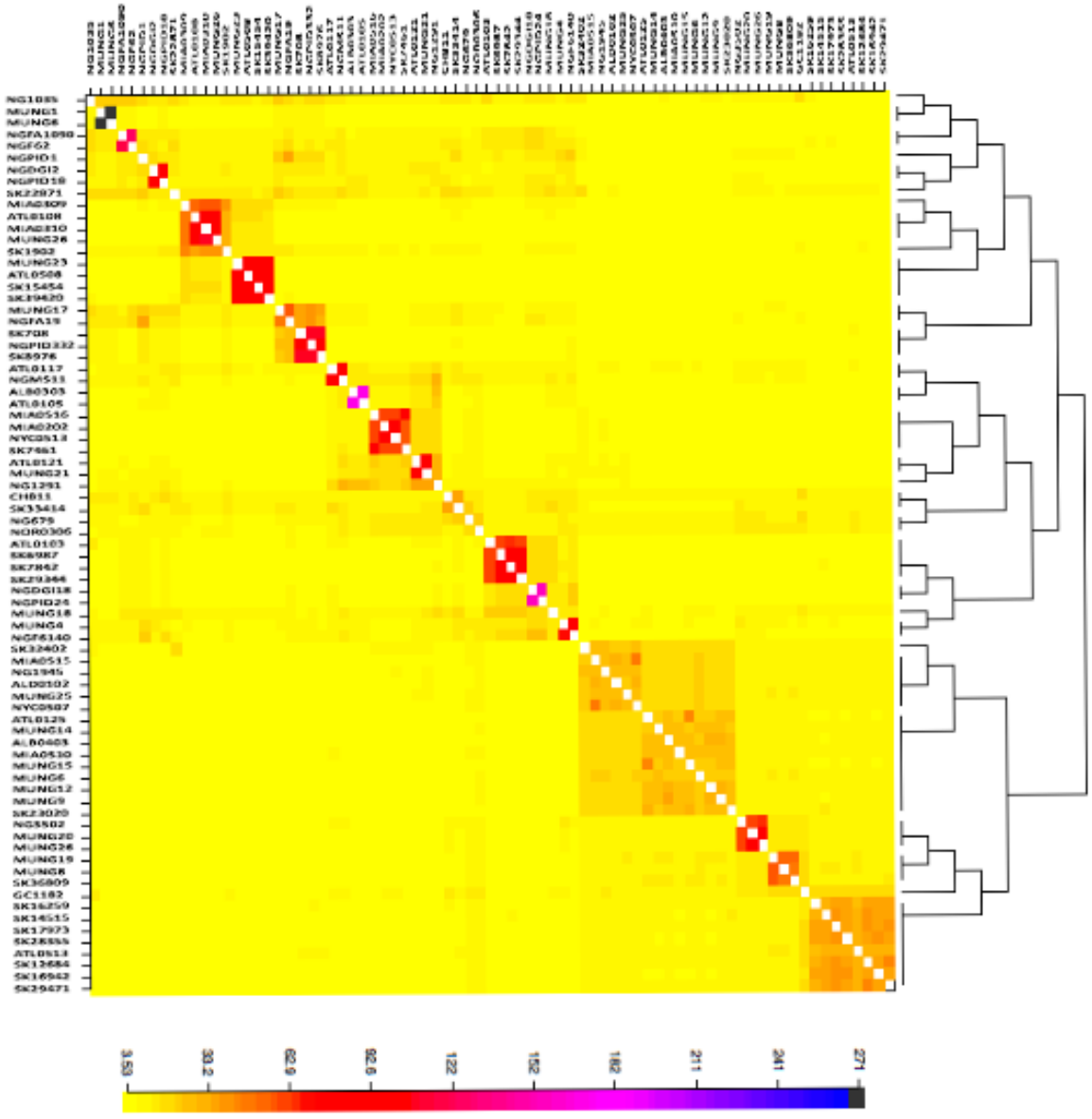


Figure 4.2.1 fineSTRUCTURE plot of strains in the sample set. The heat map represents pairwise relatedness between the different stains, superimposed on the left is the fineSTRUCTURE grouping of the strains into 30 different subgroups

Strain Name	Location	Date	MLST	Azithromycin (MIC)	Cefixime (MIC)	Tetracycline(MIC)
CH811	Chile	1982	1583	0.25	0.008	2
GC1-182	Canada	1982	1583	0.5	0.008	4
SK708	Canada	2006	1594	1	0.016	0.5
SK1902	Canada	2006	10935	0.25	0.002	256
SK6987	Canada	2006	10010	1	0.016	4
SK7461	Canada	2008	1901	0.5	0.032	8
SK7842	Canada	2006	10010	1	0.016	8
SK8976	Canada	2006	1594	0.06	0.004	2
SK12684	Canada	2006	31129	0.5	0.016	8
SK33414	Canada	2007	1928	0.25	0.008	4
SK14515	Canada	2005	1893	0.25	0.016	2
SK15454	Canada	2007	1585	0.06	0.004	2
SK16259	Canada	2007	1893	0.125	0.008	4
SK16942	Canada	2005	1893	0.125	0.016	2

SK17973	Canada	2006	1893	1	0.016	8
SK22871	Canada	2007	8122	0.125	0.004	4
SK23020	Canada	2006	1901	0.25	0.125	16
SK28355	Canada	2007	1893	0.25	0.016	4
SK29344	Canada	2007	10010	0.125	0.008	4
SK29471	Canada	2005	1893	0.25	0.016	2
SK32402	Canada	2007	8153	0.5	0.016	4
SK36809	Canada	2007	8126	2	0.008	8
SK39420	Canada	2008	1585	0.5	0.016	0.5
ALB0303	USA	2011	1588	0.03	0.015	16
ALB0403	USA	2011	1901	1	0.125	4
ATL0103	USA	2011	10931	0.5	0.015	0.25
ALB0102	USA	2011	1901	0.25	0.06	2
ATL0105	USA	2011	1588	0.06	0.015	0.25
ATL0108	USA	2011	1584	0.03	0.015	0.25
ATL0117	USA	2011	10932	0.125	0.015	16

ATL0121	USA	2011	1902	0.5	0.03	1
ATL0125	USA	2011	1901	0.25	0.015	1
ATL0508	USA	2011	1585	0.06	0.015	16
ATL0513	USA	2011	1893	0.25	0.03	2
MIA0202	USA	2011	1901	0.5	0.03	2
MIA0309	USA	2011	1931	0.125	0.015	16
MIA0310	USA	2011	1584	0.03	0.015	16
MIA0510	USA	2011	1901	1	0.03	2
MIA0515	USA	2011	1901	0.25	0.03	16
MIA0516	USA	2011	1901	0.5	0.06	8
NOR0306	USA	2011	1583	0.25	0.015	2
NYC0507	USA	2011	1901	0.25	0.06	2
NYC0513	USA	2011	1901	0.25	0.06	4
MUNG1	Canada	1991	10934	0.125	<0.016	0.25
MUNG3	Japan	2003	7363	0.25	0.5	2
MUNG4	Japan	1996	1590	0.5	0.25	4

MUNG5	Philippines	1992	1901	0.25	<0.016	1
MUNG6	Australia	2001	10008	0.125	<0.016	16
MUNG8	USA	2001	8127	2	<0.016	0.5
MUNG9	Sweden	2010	1901	0.5	1	2
MUNG12	Norway	2010	1901	0.5	0.25	4
MUNG14	Norway	2010	1901	0.5	0.25	4
MUNG15	Austria	2011	1901	0.25	1	2
MUNG17	Sweden	2010	1892	1	0.5	2
MUNG18	Norway	2010	10933	0.125	<0.016	2
MUNG19	Sweden	2010	1580	>256	<0.016	2
MUNG20	Sweden	2013	7363	0.25	0.5	2
MUNG21	Pakistan	2008	1902	1	0.032	2
MUNG23	Sweden	1998	1585	0.064	<0.016	0.125
MUNG25	Sweden	1998	1901	0.125	<0.016	0.5
MUNG26	Sweden	1999	1584	0.064	<0.016	0.5

Table 4.1 Location and date of collection of the *N. gonorrhoeae* strains including

Sequence Types and MICs of the different strains to the antibiotics azithromycin, cefixime and tetracycline The MIC breakpoint value for azithromycin resistance is 2 μg/mL, for cefixime 0.25 μg/mL, and for tetracycline 2 μg/mL, based on the CDC breakpoints for antibiotic testing.

Gene	**Clades Present**	**Gene ID (reference genomeFA 1090)**
PorB	1,3,5	NGO1812
Acetate kinase 2	5	NGO1521
Primosomal replication protein	3	NGO0582
DNA Helicase	3,5	NGO1196
Hypothetical protein	5	NGO0880
Hypothetical Protein	5	NGO1847
Hypothetical protein	5	NGO1948
ComA	5	NGO0276
Chaperone Protein HscA	5	NGO0829
tRNA-ribosyltransferase	5	NGO0294
RNA polymerase Subunit β	5	NGO1850
ArsR family transcriptional regulator	5	NGO1562
Hypothetical protein	5	NGO0165
PriB	5	NGO0582
ABC transporter subunit	3	NGO2088
Hypothetical protein	3	NGO1984
tRNA pseudouridine synthase B	3	NGO0642
Prolyl endopeptidase	1	NGO0026

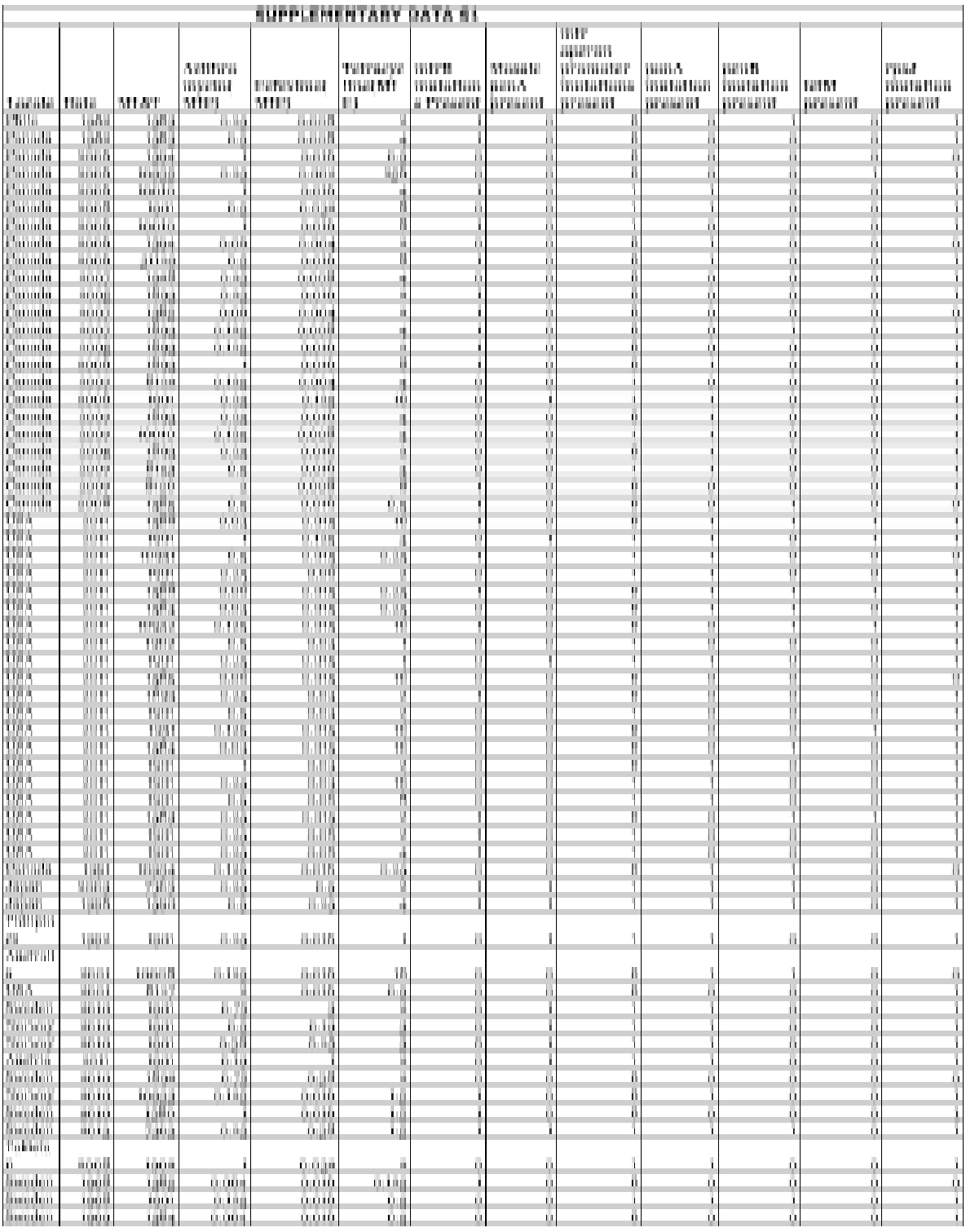
Apo-lipoprotein N-acyltransferase	1	NGO0289
Sodium dependent transporter	1	NGO2096
Phage associated protein	1	NGO1012
Hypothetical Protein	4	NGO0914

Table 4.2 Core genes of N. gonorrhoeae under positive selection in the different clades of the phylogeny of strains in the sample set.

Gene name	**FA1090 Reference locus_tag/Gene Bank ID**	**Genetic Mutations**	**Resistance Phenotype**	**References**
*mtr*R	NGO1366	G45D, A39T (glycine and aspartate substitutions)	Decreased susceptibility to macrolides and beta-lactams	PMID: 18761689
mtrCDE promoter	NGO1366	Single nucleotide deletion on reference genome position 1327932	Decreased susceptibility to macrolides and beta-lactams	PMID: 18761689
*pen*B	NGO1812	G101K, A102D (glycine and alanine substitutions)	Decreased susceptibility to third-generation cephalosporins	PMID: 17420216
Mosaic *pen*A	NGO1542	Mosaic pattern amino acid substitutions form position 294 to end of gene	Decreased susceptibility to third-generation cephalosporins	PMID: 20028823
*rps*J	NGO1841	V57M	Decreased susceptibility to tetracycline	PMID:16189114
23S rRNA	AF450080	C2611T (Cystine to Threonine substitution)	Decreased susceptibility to Azithromycin	PMID: 12183262
*tet*M	N/A	Horizontally transferred determinant on plasmids	Resistance to tetracycline (MIC >= 16μg/ml)	PMID: 21349987

Table 4.3 Known antibiotic resistance determinants in sample set. The description includes the PubMed reference ID and associated resistance phenotypes of these determinants *in N. gonorrhoeae*. Cephalosporin antibiotics include cefixime and ceftriaxone, while macrolides include erythromycin and azithromycin.

Supplementary Data



SUPPLEMENTAY DATA S2		
Protein	**GO ID**	**Protein Function**
acetate kinase	GO:0005737	cytoplasm
acetate kinase	GO:0006085	acetyl-CoA biosynthetic process
acetate kinase	GO:0008776	acetate kinase activity
acetate kinase	GO:0016310	phosphorylation
acetate kinase	GO:0005524	ATP binding
acetate kinase	GO:0000287	magnesium ion binding
acetate kinase	GO:0006082	organic acid metabolic process
bacterial regulatory arsr family protein	GO:0006355	regulation of transcription, DNA-templated
bacterial regulatory arsr family protein	GO:0003700	sequence-specific DNA binding transcription factor activity
bacterial regulatory arsr family protein	GO:0003677	DNA binding
doxx family protein	GO:0047134	protein-disulfide reductase activity
doxx family protein	GO:0055114	oxidation-reduction process
nadh-ubiquinone plastoquinone oxidoreductase chain 6 family protein	GO:0055114	oxidation-reduction process
nadh-ubiquinone plastoquinone oxidoreductase chain 6 family protein	GO:0008137	NADH dehydrogenase (ubiquinone) activity
membrane protein	GO:0015288	porin activity
membrane protein	GO:0005886	plasma membrane
membrane protein	GO:0009279	cell outer membrane
membrane protein	GO:0055085	transmembrane transport
membrane protein	GO:0006811	ion transport
membrane protein	GO:0046930	pore complex
tetratricopeptide repeat protein	GO:0005515	protein binding
dna-directed rna beta subunit	GO:0003899	DNA-directed RNA polymerase activity
dna-directed rna beta subunit	GO:0003677	DNA binding
dna-directed rna beta subunit	GO:0006351	transcription, DNA-templated
atp-dependent helicase	GO:0003676	nucleic acid binding
atp-dependent helicase	GO:0003676	nucleic acid binding
atp-dependent helicase	GO:0008026	ATP-dependent helicase activity
atp-dependent helicase	GO:0008026	ATP-dependent helicase activity
atp-dependent helicase	GO:0006200	ATP catabolic process
atp-dependent helicase	GO:0006200	ATP catabolic process
atp-dependent helicase	GO:0005524	ATP binding
atp-dependent helicase	GO:0005524	ATP binding
periplasmic protein	GO:0003899	DNA-directed RNA polymerase activity
periplasmic protein	GO:0032774	RNA biosynthetic process
periplasmic protein	GO:0006508	proteolysis
periplasmic protein	GO:0008237	metallopeptidase activity
amino acid abc transporter permease	GO:0016021	integral component of membrane
amino acid abc transporter permease	GO:0006810	transport
amino acid abc transporter permease	GO:0005886	plasma membrane
amino acid abc transporter permease	GO:0005215	transporter activity
rmuc family protein	GO:0006310	DNA recombination
abc transporter family protein	GO:0015688	iron chelate transport
abc transporter family protein	GO:0006200	ATP catabolic process
abc transporter family protein	GO:0005524	ATP binding
abc transporter family protein	GO:0015623	iron-chelate-transporting ATPase activity
neurotransmitter symporter family protein	GO:0016021	integral component of membrane
neurotransmitter symporter family protein	GO:0005328	neurotransmitter:sodium symporter activity
neurotransmitter symporter family protein	GO:0006836	neurotransmitter transport
neurotransmitter symporter family protein	GO:0055085	transmembrane transport
prolyl oligopeptidase family protein	GO:0070008	serine-type exopeptidase activity
prolyl oligopeptidase family protein	GO:0006508	proteolysis
prolyl oligopeptidase family protein	GO:0004252	serine-type endopeptidase activity
queuosine biosynthesis protein	GO:0008270	zinc ion binding
queuosine biosynthesis protein	GO:0005524	ATP binding
queuosine biosynthesis protein	GO:0016879	ligase activity, forming carbon-nitrogen bonds
queuosine biosynthesis protein	GO:0016787	hydrolase activity
queuosine biosynthesis protein	GO:0008616	queuosine biosynthetic process
competence protein	GO:0016021	integral component of membrane
competence protein	GO:0030420	establishment of competence for transformation
competence protein	GO:0008152	metabolic process
competence protein	GO:0016787	hydrolase activity
competence protein	GO:0006810	transport
competence protein	GO:0005886	plasma membrane
apolipoprotein n-acyltransferase	GO:0006807	nitrogen compound metabolic process
apolipoprotein n-acyltransferase	GO:0016021	integral component of membrane
apolipoprotein n-acyltransferase	GO:0016410	N-acyltransferase activity
apolipoprotein n-acyltransferase	GO:0016810	hydrolase activity, acting on carbon-nitrogen bonds
apolipoprotein n-acyltransferase	GO:0042158	lipoprotein biosynthetic process
apolipoprotein n-acyltransferase	GO:0005886	plasma membrane
queuine trna-ribosyltransferase	GO:0008479	queuine tRNA-ribosyltransferase activity
queuine trna-ribosyltransferase	GO:0046872	metal ion binding
queuine trna-ribosyltransferase	GO:0008616	queuosine biosynthetic process
uroporphyrinogen decarboxylase	GO:0005737	cytoplasm
uroporphyrinogen decarboxylase	GO:0004853	uroporphyrinogen decarboxylase activity

Chapter 5

Genome-wide tests for antibiotic resistance-associated variants within *Neisseria gonorrhoeae*

Matthew Ezewudo, Timothy D. Read

Under preparation for submission to a peer reviewed journal

5.1 Introduction

Neisseria gonorrhoeae is a Gram negative human pathogen responsible for gonorrhea, an important sexually transmitted infection (STI) of humans. Gonorrhea is one of the most prevalent STIs, representing 106 million out of the estimated 498 million new cases of curable non-viral STIs globally every year(World Health Organization (WHO), 2012). The only effective option for treating the disease and stopping its spread has been the use of antimicrobial therapy; there is no vaccine to prevent the pathogen. Antimicrobial treatment options have diminished over time due to the emergence of antimicrobial resistance (AMR) to all of the classes of drugs previously used to treat gonorrhea and the paucity in the development of newer antibiotics that could effectively eradicate the pathogen (Ohnishi et al., 2011; World Health Organization (WHO), 2012). One of the biggest challenges in this field is to understand the underlying molecular and evolutionary mechanisms that drive resistance in *N. gonorrhoeae*.

A number of studies have helped to elucidate the underlying molecular mechanisms driving the resistance of this pathogen to the major classes of antibiotic drugs used to

treat the condition (reviewed by Unemo & Shafer, 2011). Broadly speaking, two groups of antibiotics currently used in the treatment of gonorrhea are third generation cephalosporin and macrolides.

Cephalosporins inhibit bacteria by preventing the biosynthesis of peptidoglycan and accordingly the cell wall of the bacteria. Studies have shown that the main underlying genetic determinant for resistance to third generation cephalosporin such as cefixime and ceftriaxone, is a mosaic *pen*A gene (Ameyama et al., 2002; Lindberg et al., 2007). This gene encodes a re-modeled (mosaic) penicillin binding protein 2 (PBP2), the lethal target for this class of antibiotics that is poorly acylated by beta-lactam antibiotics. Furthermore, an overexpression of the MtrC-MtrD-MtrE efflux pump, which actively exports beta-lactams and other antibiotics, and specific mutations in the PorB1b porin gene, which decrease the influx of the antibiotics, further enhance the MICs of the cephalosporins (Unemo & Shafer, 2014).

Macrolides, which include such drugs as azithromycin and erythromycin, act by inhibiting bacterial protein synthesis via binding to the 50S subunit of bacterial ribosomes. Resistance to macrolides frequently involves specific mutations in the macrolide target, i.e. 23S rRNA. As with beta-lactams, the presence and overproduction of the MtrC-MtrD-MtrE efflux pump can reduce gonococcal susceptibility to macrolides (Deguchi, Nakane, Yasuda, & Maeda, 2010; Hagman & Shafer, 1995).

Over the past 10 years, substantial research effort has been devoted to understanding the

genetic basis of drug resistance in *N. gonorrhoeae* (Garvin et al., 2008; MD et al., 2014; Unemo & Shafer, 2011; Veal et al., 2002; World Health Organization (WHO), 2012). Since there is an increasing interest in direct attribution of resistance phenotypes based on genome sequencing, we attempted to build on knowledge of existing variants underlying resistance in the *N. gonorrhoeae* genomes in this study. This is because while past studies of this population in specific regions have focused on a number of representative genes or regions of the genome of *N. gonorrhoeae*, no extensive studies has been conducted on a global collection of strains. Our approach of genome-wide analysis of multiple international strains, represents an initial step towards providing increased depth in genomic regions covered, impacting the evolution of antibiotic resistance within the species.

5.2 Materials and Methods

5.2.1 *Neisseria gonorrhoeae* isolates

Our sample set, which has been previously described (Ezewudo et al., 2015), consists of 61 *N. gonorrhoeae* isolates of diverse origin. These included isolates from the Atlanta, Georgia Gonococcal Isolation Surveillance Program (GISP) site covering certain cities in the United States (n=21), Canada (n=24), Sweden (n=7), Norway (n=3), Japan (n=2), Austria (n=1), Pakistan (n=1), the Philippines (n=1), and Australia (n=1). The strains sequenced in this study were tested for resistance to two antibiotic drugs, azithromycin and cefixime, with MIC (minimum inhibitory concentration) breakpoints for resistance

set at 2.0, and 0.25 μg/mL, respectively, based on the CDC MIC breakpoints for testing in the GISP protocol. Phenotypic determination of the MICs across a number of different antibiotic classes was performed using the agar dilution method or the Etest method (bioMerieux), according to the instructions from the manufacturer. Details of the different isolates and their NCBI accession numbers is shown on Table 4.1.

5.2.2 Sequence generation

The *N. gonorrhoeae* strains were whole genome sequenced (WGS) using the Illumina HiSeq™ instrument, utilizing libraries prepared from 5 μg of genomic DNA for each sample. The sequencing reads' quality were filtered using the prinseq-lite algorithm (Schmieder R. et al, 2011) to ensure only sequence reads with average phred score ≥30 were used (Fig 5.1). The sequence reads data were deposited in the NCBI Sequence Read Archive public database (Accession # SRA099559).

5.2.3 Variants calling

We used two approaches to detect variants in the whole genome sequences of the strains. The first was through mapping the reads for each strain to the reference genome sequence Ref_FA_1090 (NC_002946.2) using the BWA aligner too l(H. Li & Durbin, 2009). We subsequently used the pipeline implemented in the Samtools package (H. Li et al., 2009) to identify variants in each of the strains and to generate a list of variants position files in the VCF format for each strain. The merged VCF file corresponding to variant positions of all the strains served as the source for the input files required by the QROADTRIPS tool (Thornton & McPeek, 2010).

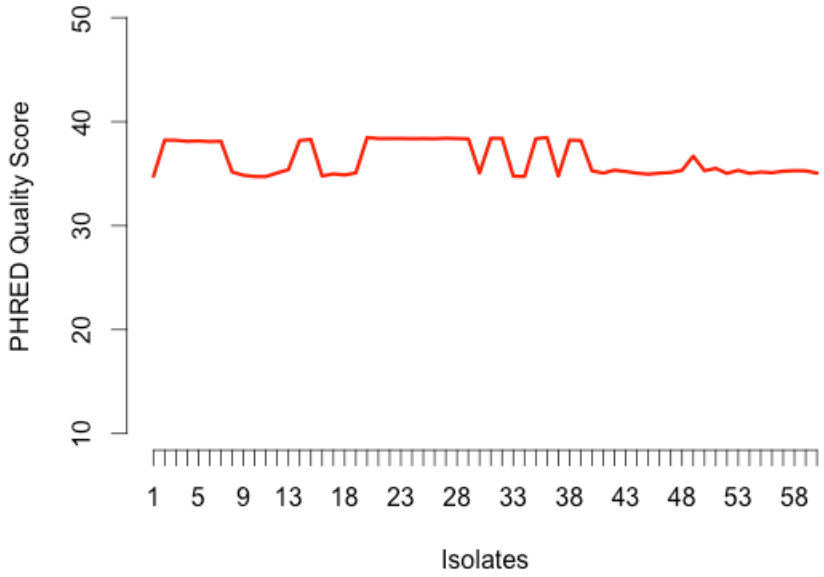


Figure 5.1 Phred quality score of sequencing reads of the *N. gonorrhoeae* isolates used for this project.

The other approach we used to call polymorphic sites in the data was an alignment -free method implemented in the kSNP tool (Gardner & Hall, 2013). kSNP searches the genomes in the sample set for the presence of SNPs flanked by a given length of repeating bases (Kmers). It only identifies variants present in more than one genome in the sample set; otherwise the base call is attributed to sequencing error. The list of SNP positions identified by kSNP serves as input data for the PPFS method (B. G. Hall, 2014) of identifying variants associated with phenotypes of interest.

5.2.4 Nucleotide diversity analysis

We estimated the genetic relatedness of the *N. gonorrhoeae* strains we have in our sample set by identifying the pairwise and average nucleotide distances between the

strains. We used the MEGA software (Tamura, et al 2013) to compute these nucleotide distances for the sequence alignments and compared with those of the Multi-Locus Sequence Typing (MLST) (Jolley & Maiden, 2010) housekeeping genes.

5.2.5 Pangenome Analysis and Test for flexible genes underlying resistance

We used the OrthoMCL tool to perform all by all blast searches of all the predicted open reading frames present in the genomes of strains in the sample set. While orthologs are defined as matching genes from a different strain, paralogs are closely matching genes found in the same strain. Orthologs and paralogs from all the strains were grouped into clusters, the entirety of which makes up the pangenome. The core genes are orthologs present in all the strains and are functionally relevant to the species, while the flexible genes are found in only a subset of the pangenome.

We used a hypergeometric approach to test for the significance of associating the presence of a given accessory gene to the phenotype of antibiotic resistance across two antibiotic drugs: cefixime and azithromycin.

5.2.6 Genome-wide association analyses

We used two software packages for the genome-wide association analysis, to identify genetic variants that are significantly associated with antibiotic resistance phenotype: PPFS and QROADTRIPS.

PPFS

Predicting Phenotypes From SNPs(PPFS) (B. G. Hall, 2014) is a recently published approach that accepts as its input data, several output files from kSNP. The output of kSNP includes phylogenetic trees built using either parsimony, neighbor joining, or maximum likelihood methods, as well as FASTA format files of SNP sites across the different strains in the sample set. kSNP determines how many SNPs are common among strains on each node and branch of the phylogenetic tree. The SNPs predicted by kSNP are then presented to PPFS for analysis.

PPFS analysis takes place in two steps: First, the tool determines the set of SNPs from all the SNPs in the dataset that could be used to accurately infer a given phenotype. This subset of SNPs is referred to as 'diagnostic SNPs'. The diagnostic SNPs are selected using a chi-square test, the null hypothesis being that SNPs are randomly distributed between individuals with a phenotype of interest. P-values for the test, suggests the probability of observing the test statistic or more extreme values for the test, if the null hypothesis were true. The consensus diagnostics SNPs are selected over a number of iterations to test the accuracy of phenotype prediction by the set of SNPs chosen. Accuracy, which is the ratio of the accurate predictions to the overall predictions, is estimated by comparing the prediction of phenotypes in a subset of the sample set using the diagnostic SNPs. An accuracy level of 90% is the level encouraged, using the tool. Secondly, the ancestral relationship between all the diagnostic SNPs is reconstructed using a maximum likelihood phylogeny approach and the history of change of state across the branches from the internal nodes of the phylogeny is traced. The chi-square test statistic is again used to assess the null hypothesis that the occurrence of a diagnostic

SNP along branches where there is a change in phenotype is random and unrelated to causality of antibiotics resistance. So for this subsequent chi-square test, diagnostic SNPs, that occur frequently on branches or nodes of an ancestrally reconstructed phylogeny where there is a change of state of the phenotype of interest will have smaller the p-values, which represents the probability of observing the test statistic or a more extreme value if the null hypothesis in this case is true.

QROADTRIPS

QROADTRIPS is a statistical tool in the ROADTRIPS (Thornton & McPeek, 2010) module, designed for performing case-control association analysis of genetic variants within a population with partially or unknown population and pedigree structure (Alam et al., 2014). QROADTRIPS is able to incorporate quantitative phenotypic data (in this case MICs) to test for associations with genetic variants, unlike ROADTRIPS, which is based on categorical phenotypes only. The p-values estimated for the associations were corrected for multiple testing using the Bonferroni method, and loci with variants significantly associated with the resistance phenotype were flagged as candidates for causation of antibiotics resistance. The input data included genotypes from all SNPs detected in each WGS *N. gonorrhoeae* genome compared to the FA 1090 reference genome, using Samtools (H. Li et al., 2009) SNP calling pipeline. The phenotypes were MIC values from AMR assays.

5.3 Results and Discussion

5.3.1 Nucleotide diversity of sample set is lower than overall *Neisseria* diversity

A comparison of the mean nucleotide distance at the seven housekeeping loci between strains in our sample set on one hand, and those of the different sequence types found in the MLST database (Table 5.1) suggests a low diversity among the strains we used for this study (average nucleotide distance = 0.001). This value is not significantly different from the mean nucleotide distance for all the sequence types in the MLST data (0.002), suggesting genetic similarity within the species across these loci, regardless of where the strains were collected. A phylogeny of all the *N. gonorrhoeae* sequence types in the MLST database inclusive of sequence types in our sample set is shown in Fig 5.2.

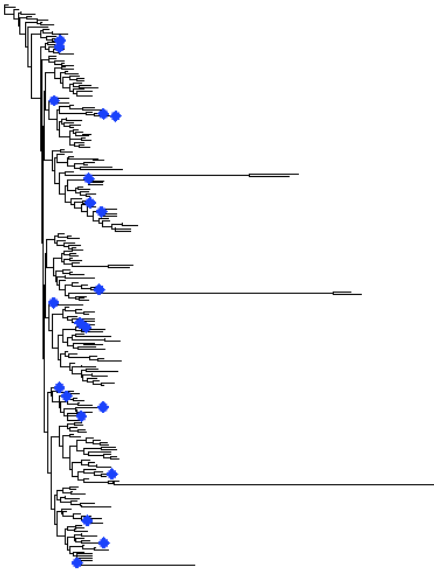


Fig 5.2 Neighbor-Joining phylogeny of all the sequence types in the MLST database, taxa colored blue are Sequence types present in our sample set

5.3.2 Novel genetic variants potentially associated with drug resistance in *N. gonorrhoeae*

The previous findings of the possibility that some *N.gonorrhoeae* strains have an antibiotic resistance phenotype that could not be explained by known resistance genes and mutations (Unemo & Shafer, 2011) led us to search for novel genetic determinants underlying resistance in the species.

The results from our analysis of the pangenome of *N. gonorrhoeae* follow the expected partitioning of core and flexible genes see Fig 5.3. Our approach to determine whether any non-core accessory gene showed a strong association with strains that were resistant to the antibiotic drugs tested yielded no significant association between any flexible gene and antibiotic resistance phenotype ($p > 0.05$ using Fishers exact test with Bonferroni

correction).

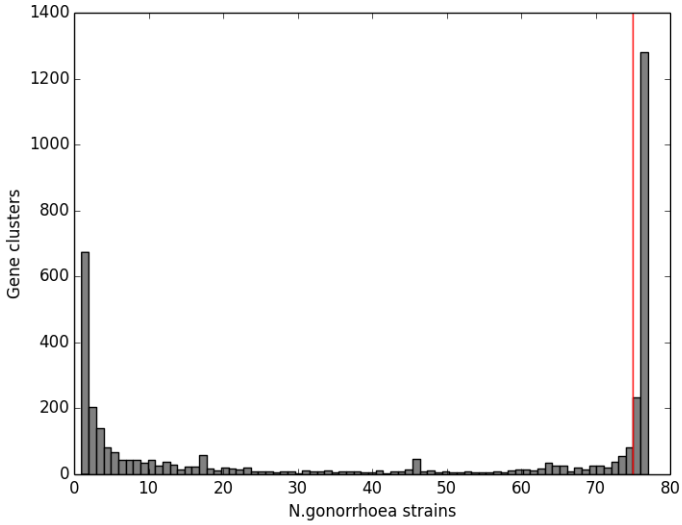


Fig 5.3 Pangenome of strains represented in sample set. Genes to the left of the red line represents the flexible genes while those to the right represent the extended core genes.

The other strategy was to apply genome wide association (GWA) algorithms to attempt to link SNPs and small insertions and deletions (indels) within the core genome to a phenotype.

SNP Calling analysis

Our first GWA method was PPFS (B. G. Hall, 2014), which flags variants in an ancestrally reconstructed phylogenic tree that are highly correlated with the branches where there is change in phenotypes on the phylogeny, as candidates for causation of antibiotics resistance in the pathogen. The alignment-free tool kSNP which we used at the initial phase of this method to identify SNPs within strains in the sample set detected 21,706 variant sites, which were fed to the PPFS tool for analysis. The accuracy level for selecting diagnostic SNPs was set at 90%.

For our second approach, we used QROADTRIPS (Alam et al., 2014; Thornton &

McPeek, 2010) which applies a classical regression method to associate variants with antibiotic resistance but compensates for the effects of population structure. For this method, we started with 17,560 SNPs detected using the SNP calling pipeline on the mapped reads of all the strains in the sample set. The association test was performed using the level of antibiotic resistance as a continuous variable. Given the number of SNPs tested, the threshold for significance was set at 2.8×10^{-6} (correcting for multiple testing using the Bonferroni approach of dividing the desired experiment-wise Type I error rate (α=0.05) by the number of tests performed (m=17,560)).

Phenotypic analysis

The first phenotype we tested for was resistance to azithromycin. Both GWA methods identified a combined number of 122 variants purported to be linked to resistance to this antibiotic (see table 5.2). For the other antibiotic drug –cefixime, QROADTRIPS did not identify SNPs that were significantly linked to the resistance phenotype, but PPFS flagged 31 diagnostic SNPs that were potentially associated with cefixime resistance (table 5.3). For both phenotypes there were 16 non-synonymous SNPS within characterized loci of the reference genome and called by either method to be significantly associated with the phenotype (table 5.4)

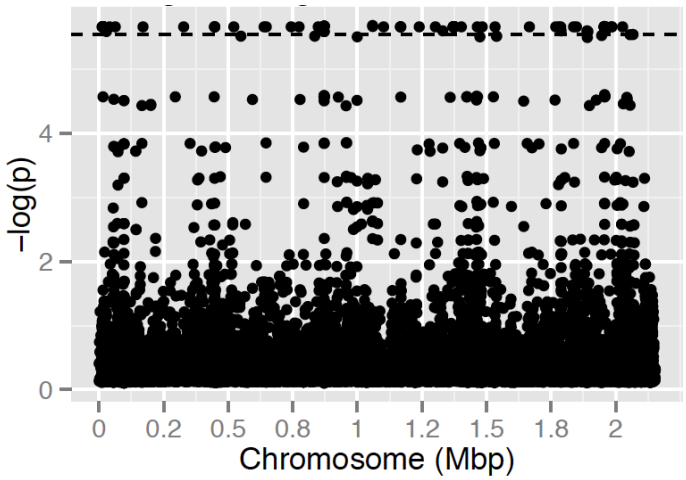


Fig 5.4 Manhattan plot of the natural log of the p-values of the SNPs within strains in the data set. SNP points above the dotted line are significantly associated to resistance to

Of these 16 non-synonymous SNPs, 4 (reference genome positions: 14776, 1516686, 1768055, 1957743) were corroborated to be appreciably linked to antibiotic resistance by both methods with a high degree of confidence ($p < 1 \times 10^{-5}$) see table 5.4. These 4 SNPs which were corroborated by both GWA methods offer the most confidence of their potential association with antibiotic resistance. Also, the characteristics of the loci where these non-synonymous variants arise tend to support their potential for being causative of antibiotic resistance. For instance, SNP positions 14776 and 178055 occur on DNA isomerase and methyl-transferase proteins respectively. These are enzymes involved in metabolic pathways such as activation or deactivation of DNA synthesis and are potential targets of antibiotic medications. The non-synonymous mutation on position 1516686 of the reference genome occurs within the NGO1542 locus that codes for penicillin binding protein 2 (PBP2). Mosaic *penA,* the gene encoding PBP2, have been previously indicated to be involved in cephalosporin resistance (Ameyama et al., 2002). Also, a number of mutations (SNP positions 1516409 – 1538106 on the reference genome) within this locus –NGO1542 were significantly linked to cefixime resistance by the PPFS method and to an appreciable extent the QROADTRIPS method, indicative of the mosaic *penA* pattern

(Table 5.3). Finally, the SNP on position 1957743, which was significantly associated with resistance to azithromycin by both methods, falls within a hypothetical gene (locus tag NGO1982) that has not been fully characterized in previous literature. Given that it is the only non-synonymous SNP within a hypothetical gene locus corroborated by both methods, it possibly could play a role in resistance to azithromycin, and this could be verified using functional and complementation analysis in the bacteria genetics laboratory. For the other 12 non-synonymous SNPs in Table 5.4 called by either of the methods, while they were not significantly corroborated by both methods, they could also serve as candidates for further analysis to experimentally validate their link to the respective antibiotics resistance phenotype.

5.3.2.1 Comparison between PPFS and QROADTRIPS methods

For the sets of SNPs tested by the two association test methods (21,706 and 17,560 for PPFS and QROADTRIPS respectively), both methods identified 6631 similar variants on the reference genome. kSNP identified 15153 unique SNPs not identified by our QROADTRIPS variant calling pipeline, while our SNP calling pipeline for QROADTRIPS identified 6608 unique SNPs not called using kSNP.

Comparing the SNP calling results from both methods, 14 out of the 16 diagnostic SNPs identified by PPFS as associated with azithromycin resistance were also flagged by QROADTRIPS as associated with the phenotype. In addition, a majority of the SNPs identified as significantly associated with azithromycin resistance via the QROADTRIPS SNP calling pipeline, were flagged as diagnostic SNPs by PPFS with p-values < 0.001

(Table 5.2). Similarly, for the cefixime antibiotics resistance phenotype, for two-thirds of the SNPs called by both pipelines, there is agreement between the two methods on the significance of association between the SNPs and cefixime resistance (Table 5.3). Both of these observations suggest an agreement by both methods on the approach to identify SNPs associated with antibiotics resistance, even though there is some discrepancy in the orders of magnitude of the p-values obtained using the two methods (Table 5.2, 5.3 and 5.4).

5.4 Conclusion

We identified 4 primary variants that were corroborated by two different methods of genome wide association testing to be significantly associated with antibiotics resistance to azithromycin and cefixime (Tables 5.4). Both approaches appear to reconfirm the importance of the *penA* loci (that expresses the PBP2) as a predictor of resistance to third generation cephalosporin. In addition, we flagged potential novel mutations found in genes that express a range of proteins, potentially significantly associated to resistance to the various antibiotic drugs tested. These genes are candidates for further study using experimental bacteria genetics, to confirm their involvement in antibiotics resistance. We also need to further investigate possible epistatic interactions of these genes with other known resistance determinants and how their varying levels of expression could correlate with the resistance phenotype.

We applied two different tools -- PPFS and QROADTRIPS-- for the genome-wide association tests for novel antibiotic variants in strains within the sample set. PPFS

identified variants with lower p-values compared to QROADTRIPS. This is in part because of the PPFS approach of first detecting diagnostic SNPs that are associated to a particular phenotype as predictive SNPs for the phenotype and taking the further step of assessing the significance of change in state of this now smaller set of SNPs in relation to change in the phenotype along the branches of an ancestrally reconstructed phylogeny of all the strains tested. This could also be attributable to the fact that QROADTRIPS applies a quantitative approach when correlating the phenotype (MICs for the different drugs tested) as a continuous variable to the underlying genetic variation versus the qualitative approach of the PPFS method, which is based on a threshold and categorization of the phenotypes as discrete variables. Also, the kSNP approach of detecting SNPs in the sample set without recourse to a particular reference genome increases the pool of SNPs within the set from accessory genes, which may have been missed otherwise. One limitation of the PPFS method seem to be the absence of multiple testing correction for the identification of diagnostic SNPs, but it is partly compensated for by the iterative nature of selecting the consensus set of diagnostic SNPs, and a threshold of acceptable measure of accuracy for arriving at this set of diagnostic SNPs. Going forward, carrying out complementation tests on the candidate resistant determinants identified through both methods will be a helpful step in corroborating the accuracy of the 2 different approaches. Also a different prokaryotic genome-wide phenotype association test such as PhyC (Farhat et al., 2013) could be used on the same data to provide better insight on the nature of the underlying SNPs in the sample set.

Previous antibiotics resistance studies on *N. gonorrhoeae* have been performed on a

small number of loci. Only recently have there been extensive population genomics studies of the pathogen using WGS; for example, Grad *et al* (Grad et al., 2014) applied WGS analysis of more than 200 *N. gonorrhoeae* strains within the United States to make inferences on the epidemiology of resistance to the antibiotic cefexime. We analyzed data that was collected from a wider spread of geographical locations, and tested resistance to a larger suite of antibiotic drugs, to investigate the evolution of antibiotics resistance within the pathogen population. This work points to the need for a broader collection of strains sampled from locations worldwide and isolated at different time points that will serve as materials for a comprehensive genomic studies of evolution of antibiotic resistance in the *N. gonorrhoeae* population.

5.5 Appendix

The following appendix contains tables already referenced in the text of the chapter.

MLST Locus	**Nucleotide diversity in *Neisseria* spp**	**Nucleotide diversity in *N. gonorrhoeae***	**Nucleotide diversity in our Sample**
abcZ	0.084	0.003	0.002
adK	0.052	0.001	0.001
aroE	0.169	0.002	0.001
Pgm	0.089	0.001	0.001
fumC	0.035	0.003	0.002
Gdh	0.116	0.001	0.001
pdhC	0.089	0.001	0.001

Table 5.1 comparison of mean nucleotide distances in housekeeping genes of strains in MLST database and strains in our sample set

SNP position	chi-test p-value for diagnostic SNPs	chi-test p-value for causal SNPs	QROADTRIPS p-value
12319	NA	NA	2.19E-06
12332	NA	NA	2.19E-06
12435	0.0015443	7.85E-34	2.19E-06
12606	0.0015443	NA	2.19E-06
13388	NA	NA	2.19E-06
13708	NA	NA	2.19E-06
13922	NA	NA	2.19E-06
13928	NA	NA	2.19E-06
13955	NA	NA	2.19E-06
14270	0.0043299	NA	2.19E-06
14324	0.6455	NA	2.19E-06
14393	NA	NA	2.19E-06
14401	NA	NA	2.19E-06
14776	0.0015443	7.85E-34	2.19E-06
14813	0.0015443	NA	2.19E-06
15347	NA	NA	2.19E-06
15362	NA	NA	2.19E-06
15481	0.0015443	NA	2.19E-06
15606	0.0015443	NA	2.19E-06
15636	0.0015443	NA	2.19E-06
15823	0.002744	NA	2.19E-06
15944	NA	NA	2.19E-06
15972	NA	NA	2.19E-06
15974	NA	NA	2.19E-06
16041	NA	NA	2.19E-06
16050	NA	NA	2.19E-06
16074	0.0015443	7.85E-34	2.19E-06
16088	0.68578	NA	2.19E-06
16182	0.011231	NA	2.19E-06
40517	0.0045409	NA	2.19E-06
62627	0.0015443	NA	2.19E-06
170126	0.0043299	NA	2.19E-06
327499	NA	NA	2.19E-06
447334	NA	NA	2.19E-06
447336	NA	NA	2.19E-06
447481	NA	NA	2.19E-06
447489	NA	NA	2.19E-06
521548	0.0039151	NA	2.19E-06
638993	NA	NA	2.19E-06
647226	0.0015443	NA	2.19E-06
743830	NA	NA	2.19E-06
774339	0.0093725	NA	2.19E-06
848476	NA	NA	2.19E-06

871260	NA	NA	2.19E-06
871267	NA	NA	2.14E-06
871302	0.059171	NA	2.14E-06
871357	NA	NA	2.14E-06
871941	NA	NA	2.14E-06
871944	NA	NA	2.14E-06
1058899	0.0015443	NA	2.08E-06
1059613	0.98213	NA	2.16E-06
1101551	0.0015443	NA	2.19E-06
1166653	0.0015443	NA	2.19E-06
1167797	NA	NA	2.19E-06
1167916	0.0015443	NA	2.19E-06
1239564	0.0015443	NA	2.19E-06
1302153	0.0043299	NA	2.19E-06
1330716	0.78413	NA	2.53E-06
1371326	0.0015443	NA	2.19E-06
1371374	0.0015443	NA	2.19E-06
1371440	0.0015443	NA	2.19E-06
1371803	0.0015443	NA	2.19E-06
1371917	0.0015443	7.85E-34	2.19E-06
1371974	0.0015443	NA	2.19E-06
1371995	0.0015443	NA	2.19E-06
1372293	NA	NA	2.19E-06
1372484	NA	NA	2.19E-06
1372749	NA	NA	2.19E-06
1373320	NA	NA	2.19E-06
1373356	0.0015443	NA	2.19E-06
1373506	0.0015443	7.85E-34	2.19E-06
1373653	0.0015443	NA	2.19E-06
1373945	0.0015443	7.85E-34	2.19E-06
1374038	0.0015443	NA	2.19E-06
1374161	NA	NA	2.19E-06
1374164	NA	NA	2.19E-06
1374176	0.0015443	NA	2.19E-06
1374200	0.0015443	NA	2.19E-06
1374287	0.041593	NA	2.19E-06
1374378	NA	NA	2.19E-06
1374472	NA	NA	2.19E-06
1374619	0.002744	NA	2.19E-06
1374696	0.0015443	NA	2.19E-06
1374748	NA	NA	2.19E-06
1374751	NA	NA	2.19E-06
1402714	0.059171	NA	2.19E-06
1464451	NA	NA	2.19E-06
1483576	0.0015443	NA	2.19E-06

1533228	NA	NA	2.19E-06
1666211	0.0057358	NA	2.19E-06
1713181	0.0015443	7.85E-34	NA
1768055	0.0015443	7.85E-34	2.19E-06
1780498	0.010964	NA	2.19E-06
1787211	NA	NA	2.19E-06
1837063	0.0015443	NA	2.19E-06
1848254	NA	NA	2.19E-06
1892797	0.51564	NA	2.59E-06
1955412	0.0015443	NA	2.19E-06
1955703	0.0045409	NA	2.19E-06
1955928	0.0029302	NA	2.19E-06
1956147	0.0015443	7.85E-34	2.19E-06
1956159	0.0015443	7.85E-34	1.44E-04
1956174	0.0015443	NA	2.19E-06
1956217	0.0015443	NA	2.19E-06
1956252	0.0045409	NA	2.19E-06
1956453	0.0015443	NA	2.19E-06
1956492	0.0015443	7.85E-34	2.19E-06
1956510	0.0015443	NA	2.19E-06
1956700	NA	NA	2.19E-06
1957060	0.0015443	7.85E-34	NA
1957099	0.0015443	1.03E-12	4.83E-04
1957243	0.0043299	NA	2.16E-06
1957277	0.0015443	1.30E-17	2.70E-05
1957476	NA	NA	2.16E-06
1957521	0.0015443	1.03E-12	1.44E-04
1957609	NA	NA	2.19E-06
1957620	NA	NA	2.19E-06
1957743	0.0015443	7.85E-34	2.19E-06
1983110	0.0015443	NA	2.19E-06
1986130	NA	NA	2.19E-06
2027045	0.0015443	NA	2.19E-06
2056067	0.059171	NA	2.90E-06

Table 5.2 List of SNPs associated with resistance to azithromycin by both association test methods. The table depicts position of the SNPs on the reference genome and the chi-square p-value for test of randomness of the SNPs relative to the resistance phenotype of strains in sample set and the corresponding QROADTRIPS p-value.

SNP position	chi-test p-value for diagnostic SNPs	chi-test p-value for causal SNPs	QROADTRIPS p-value
205677	0.00014937	7.85E-34	NA
379862	0.00014937	1.30E-17	NA
421792	0.00014937	7.85E-34	NA
427464	0.0031895	0.9607	0.114332
749288	0.00027512	1.97E-09	0.730586
764262	0.00014937	0.9343	NA
887030	1.43E-06	1.30E-17	NA
963415	0.00014937	7.85E-34	NA
1089378	1.43E-06	1.30E-17	NA
1127405	0.00014937	7.85E-34	NA
1173580	1.43E-06	3.60E-12	NA
1270489	0.00014937	1.30E-17	NA
1406914	0.00014937	0.8678	NA
1436301	0.00014937	7.85E-34	NA
1452474	0.00014937	7.85E-34	NA
1467455	0.00014937	1.30E-17	NA
1515665	0.000014385	3.20E-11	0.0771767
1516253	0.00051788	1.30E-17	NA
1516364	0.00016463	1.30E-17	0.804895
1516409	0.000040213	1.30E-17	4.40E-05
1516586	0.00016463	1.30E-17	0.00129052
1516686	0.00016463	1.30E-17	0.00049408
1516709	0.00016463	1.30E-17	0.00039415
1516799	0.00051788	1.30E-17	0.00017035
1516823	0.00016463	1.30E-17	0.00017035
1538106	0.014117	1.30E-17	NA
1678634	0.00037474	7.26E-06	0.0120541
1834823	0.00014937	7.85E-34	NA
1865101	0.00014937	7.85E-34	NA
1885133	0.00036157	0.8111	0.444821
1885377	3.48E-06	0.8517	0.716525

Table 5.3 List of SNPs associated with resistance to cefixime. The table depicts position of the SNPs on the reference genome and the chi- square p-value for test of randomness of the SNPs relative to the resistance phenotype of strains in sample set and the corresponding QROADTRIPS p-value.

Ref Genome position	PPFS causal p-value	QROAD TRIPS causal P-value	FA1090 Reference locus_tag	Protein Description	Amino Acid Change	Phenotype
14776	7.85E-34	2.19E-06	NGO0017	tri-iso phospahate isomerase	A146T	Azithromycin Resistance
16088	NA	2.19E-06	NGO0019	protein-L-isoaspartate O-methyltransferase	A30E	Azithromycin Resistance
963415	7.84E-34	NA	NGO0992	Hypothetical Protein	M192T	Cefixime Resistance
1059613	NA	2.16E-06	NGO1097	Phage associated protein	Q438L	Azithromycin Resistance
1089378	1.29E-17	NA	NGO1149	Succinyl sulfhydrolase	G251S	Cefixime Resistance
1127405	7.84E-34	NA	NGO1183	ribosyl glycinamide synthase	H1039R	Cefixime Resistance
1371326	NA	2.19E-06	NGO1406	glycine cleavage system aminomethyltransferase	S324R	Azithromycin Resistance
1373356	NA	2.19E-06	NGO1408	Hypothetical protein	D6N	Azithromycin Resistance
1374200	NA	2.19E-06	NGO1410	Hypothetical protein	I87M	Azithromycin Resistance

1374619	NA	2.19E-06	NGO1411	Hypothetical protein, ion transport	A12T	Azithromycin Resistance
1483576	NA	2.19E-06	NGO1514	Hypothetical protein	L96S	Azithromycin Resistance
1516686	1.29E-17	4.94E-04	NGO1542	Penicillin binding protein 2	A160V	Cefixime Resistance
1666211	NA	2.19E-06	NGO1710	Trans-acylase protein	Q622*	Azithromycin Resistance
1768055	7.85E-34	2.19E-06	NGO1795	DNA methyl-transferase	G145R	Azithromycin Resistance
1957243	NA	2.16E-06	NGO1981	Hypothetical protein	T119N	Azithromycin Resistance
1957743	7.85E-34	2.19E-06	NGO1982	Hypothetical protein	L19M	Azithromycin Resistance

Table 5.4 SNP identities, p-values of association by either association test methods, the protein ID and name of gene where the SNPs occur. Reference gene used for annotation was FA 1090 (NCBI accession: NC_002946.2)

Chapter 6

Summary and future directions

This dissertation has focused on the study of genetic variation within bacterial populations and the development and application of next generation sequencing analysis tools. I have applied these methods to answer questions about evolutionary processes in *Neisseria gonorrhoeae*, the causative organism of the sexually transmitted infection referred to as gonorrhea. The wide prevalence and large impact of gonorrhea in human populations makes this work highly relevant to public health. My studies also intersect with the recent developments in whole genome sequencing. How can we best make sense of the enormous quantities of data that we can now generate at low cost? How can we use this information to address classical questions from microbial genetics at a higher resolution using these data? And finally, how can we effectively combine phenotypic or clinical information form bacterial isolates with microbial genomics in order to address important questions related to public health? These are the significant and substantial problems addressed by this dissertation that led to the advances and discoveries I have described in detail in the previous chapters.

My research started with how NGS has helped address one of the major questions in genetics studies, about the nature of the genetic architecture underlying complex diseases in humans. Furthermore, I addressed how deep sequencing could uncover novel rare variants and tie a phenotype of interest to genetic variants that contribute to its etiology. The advent of NGS in human genetics studies also parallels the application of the same

sequencing technologies to perform genomics analysis of microbial populations, an emerging field that offers broad opportunities for both clinical microbiologists and microbial population geneticists. Both developments require bioinformatics tools that could make efficient and accurate inferences on the biological significance of millions of variants generated through NGS. The developments we made in SeqAnt(Shetty et al., 2010), a next generation sequencing analysis tool, address this issue. Over the past 3 years we worked to progressively add relevant functionalities and user-friendly capabilities to this tool, which has extensive capacity to annotate millions of variant positions for a wide range of model organisms with record speed and accuracy. This is because of the various modifications we implemented on the underlying databases and the architecture of the software. Given that the tool now has added capacity to annotate bacteria whole genomes, it will be relevant going forward for most current research in genetics and pathogen genomics, dealing with the enormous amount of data from next-generation sequencing technology platforms, as evidenced in its ability to annotate the mosaic *penA* alleles in strains in our sample set that are resistant to cefixime.

I subsequently addressed the issue of antibiotic resistance in pathogenic bacteria, which continues to be a huge public health problem. There is preponderance of studies(Grad et al., 2014; Kohanski, DePristo, & Collins, 2010a; Schmieder & Edwards, 2012; Vidovic et al., 2014) aimed at trying to understand and mitigate this problem. *N. gonorrhoeae* is a species system that is representative of this challenge; a pathogen linked to an ancient disease and has a history of antibiotics resistance that parallels the development of different classes of antibiotic medications in the past century(Ohnishi et al., 2011; Unemo

& Shafer, 2011; 2014). The studies I carried out using this system therefore could serve as a basis to understand the roles played by different evolutionary forces, in the emergence and persistence of resistance in pathogen populations. Also, the *N. gonorrhoeae* isolates that make up the sample set used in our study were collected from different geographical regions -- with differing prevalences and treatment regimens to the disease – from across the globe. Other *N. gonorrhoeae* genomics studies were mostly based on isolates collected within a smaller geographic region(Grad et al., 2014; Vidovic et al., 2014).

The bioinformatics methods we used in these studies were based on the analysis of whole genome sequence data from next generation sequencing platforms, yielding results from observations tested across the entire genome sequence and not just representative loci or genomic regions. One key observation from our studies was the significant impact of recombination within the pathogen population. A measure of the effects of recombination to mutation (r/m) is 2.2; this is appreciably greater than the ratio found in most other bacteria species. This ratio, which is distinct from relative recombination rate (Didelot & Maiden, 2010) -- the measure of the rate of recombination to that of mutation-- actually reflects the ratio of substitutions introduced by gene-conversion type recombination in bacteria genomes to that introduced through point mutations. (The actual value for ratio of recombination rate to mutation estimated using the Gubbins tool(Croucher et al., 2015) for our dataset is 0.098). We showed that random sequencing error in itself would not be enough to significantly skew the r/m ratio, by simulating base call errors (1 base call error

per 1000 nucleotide bases) into the multiple whole genome sequence alignment of the strains in the sample set. The r/m ratio obtained from this simulated data is a value of 1.9; this is understandable because random sequencing errors will not introduce multiple variants blocks across regions of the whole genome alignment, which is required to infer recombination signals using the ClonalFrame or Gubbins tools. How this appreciable recombination taking place in the population ties directly to evolution of resistance to antibiotics is not immediately clear from our studies, but the presence of most strains resistant to cefixime antibiotic on one super-clade in the phylogeny hints at possible horizontal acquisition of the underlying resistance determinant. Going forward, performing similar analysis with a larger sample size of strains with different antibiotics resistance phenotypes will likely confirm any HGT breakpoints in the phylogeny and the source of the resistance determinants.

Another key finding from the studies was the determination that there were possibly 5 ancestral subpopulations within the pathogen population. The strains in the sample set which showed the most admixture were mostly from one ancestral subpopulation and were those with earlier collection dates, which will suggest that over time individual strains acquire genetic materials from other strains from a different lineage. The observation that there was no correlation between geographical location of sampling and membership of a particular ancestral subpopulation and stratification within the population, could suggest a recent expansion of the pathogen across different geographical niches. It is instructive that previous studies on the structure of human population have suggested 5 subpopulations that parallels the 5 major geographic regions

in the world (Rosenberg, Pritchard, Weber, & Cann, 2002). It is possible, given that *Neisseria gonorrhoeae* is exclusively a human pathogen, that its population diversification could have tracked the human population expansion. Previous studies on pathogenic bacteria population structure (Achtman, 2004) have shown how epidemic factors could impact the nature of microbial populations, even highly recombinant ones like those of *Neisseria meningiditis*. It will be instructive for future studies to focus on how epidemiological factors drive the expansion of this pathogen population against the backdrop of other evolutionary forces in the population.

Given the possibility of yet undiscovered variants underlying or augmenting antibiotic resistance in *N. gonorrhoeae*, I tested for these novel resistance variants in our sample set. I identified a number of potential candidate variants that are significantly associated with resistance to azithromycin and cefixime. I used both a conventional GWAS-type statistical test approach and an approach based on correlating change of state on branches of ancestrally reconstructed phylogeny of all the strains in the sample set to a change in resistance phenotype along same branch, to make the inferences. The putative variants would be subject to complementation analysis in bacteria genetics laboratory to confirm their functional relevance.

As is the case with any scientific inquiry, my research in seeking to answer one set of questions, has raised some more relevant questions for future studies. Although the samples we collected for analysis in the studies offered ample breadth in terms of geographical locations covered, stronger inference could be achieved with greater depth

of coverage and an increase in the number of isolates available for analysis, especially with regards to the association tests for novel resistance determinants and the possibility of detecting rare variants within the population underlying resistance phenotypes. Collaborations like the recent studies by Grad et al (Grad et al., 2014)that involved sequencing of ~200 *N.gonorrhoeae* strains for epidemiological studies of the pathogen point to the future of studies in this field, because as data from these collaborations are more readily available to the public, researchers can more easily design experiments to answer microbial population related questions.

The Illumina platform, which generated millions of paired end short reads of an average 100 base length per read, was used to generate the data analyzed in this dissertation. The average coverage and sequencing reads' quality scores (225 and Q30 respectively) offer confidence in the reliability of the nucleotide base calls made during sequencing. Also, the similar values obtained when comparing nucleotide diversity of *N. gonorrhoeae* housekeeping genes using strains in our sample set, against those in the MLST database (Table 5.1) offers validation as to the accuracy of the base calls in the sequencing project. Still, Illumina technology does come with inherent challenges of context-specific errors, and the length of the reads creates a very difficult challenge for determining the accurate nucleotide sequence in genomic regions of long tandem repeats (Carneiro et al., 2012). The advent of pacBio sequencing which applies real time polymerase synthesis sequencing approach begins to address the above problems, by generating longer reads of thousands of base pairs in length and with sufficient depth of coverage attains 99.9% consensus accuracy (Ferrarini et al., 2013). Currently, the cost per nucleotide using this

technology is still exorbitant, but with the trend in the industry of decreasing sequencing costs, pacBio sequencing seems the approach for future analysis like this, which will offer better resolution on problematic genomic regions, and longer reads for easier *de novo* assembly.

In this study we predicted the core genome size of *N.gonorrhoeae* to be 1189 and the number of accessory genes to be 2141. With an increased number of isolates, we could more confidently define the pangenome of *N. gonorrhoeae,* which is the representation of every possible gene within the population. With just 76 strains in our sample set, we likely have not captured the breadth of *N. gonorrhoeae* pangenome; a larger sample size would bring to the forefront accessory genes that could be tied to relevant phenotypes in the population. This expanded information will allow more precise estimates of how often taxas in the phylogeny (strains in the sample set) acquire or shed genes through HGT. A measure of the relative rate of HGT for each strain could be an important value to correlate with the antibiotic resistance phenotype, giving a snapshot to the relationship between HGT and antibiotic resistance within the population.

Finally, functional analysis in the bacterial genetics laboratory is an important component to piece together the mechanisms of and the causative nature of the novel variants shown to be associated with antibiotic resistance using statistical tests. Using knock-down or knock-out approaches, bacteria geneticists could establish that the detected variants actually play a role in antibiotic resistance and are not just linked to the phenotype by happenstance. Because many variants may be found in strong linkage disequilibrium,

follow-on direct manipulative experiments, perhaps using CRISPR-CAS9 technologies for example, would allow direct determination of the effects of specific variants that I have implicated as being causative.

In conclusion, it is my hope that the work presented in this thesis serves as a step further in understanding the *N. gonorrhoeae* population, and the evolutionary process within this population and possibly other microbial populations that drives the emergence of antibiotic resistance. Gonorrhea and antibiotic resistant gonococcus is a menacing public health threat, which requires many hands on deck to curtail.

Bibliography

1000 Genomes Project Consortium, Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., et al. (2010). A map of human genome variation from population-scale sequencing. *Nature*, *467*(7319), 1061–1073. doi:10.1038/nature09534

Aas, F. E., Wolfgang, M., Frye, S., Dunham, S., Løvold, C., & Koomey, M. (2002). Competence for natural transformation in Neisseria gonorrhoeae: components of DNA binding and uptake linked to type IV pilus expression. *Molecular Microbiology*, *46*(3), 749–760.

Achtman, M. (2004). Population structure of pathogenic bacteria revisited. *International Journal of Medical Microbiology*, *294*(2-3), 67–73.

Achtman, M. (2008). Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens. *Microbiology*, *62*, 53–70. doi:10.1146/annurev.micro.62.081307.162832

Aickin, M., & Gensler, H. (1996). Adjusting for multiple testing when reporting research results: the Bonferroni vs Holm methods. *American journal of public health*, *86*(5), 726–728.

Akashi, H. (1994). Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. *Genetics*, *136*(3), 927–935.

Akashi, H., Ko, W.-Y., Piao, S., John, A., Goel, P., Lin, C.-F., & Vitins, A. P. (2006). Molecular evolution in the Drosophila melanogaster species subgroup: frequent parameter fluctuations on the timescale of molecular divergence. *Genetics*, *172*(3), 1711–1726. doi:10.1534/genetics.105.049676

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

Albert, T. J., Molla, M. N., Muzny, D. M., Nazareth, L., Wheeler, D., Song, X., et al. (2007). Direct selection of human genomic loci by microarray hybridization. *Nature methods*, *4*(11), 903–905. doi:10.1038/nmeth1111

Ameyama, S., Onodera, S., Takahata, M., Minami, S., Maki, N., Endo, K., et al. (2002). Mosaic-like structure of penicillin-binding protein 2 Gene (penA) in clinical isolates of Neisseria gonorrhoeae with reduced susceptibility to cefixime. *Antimicrobial Agents and Chemotherapy*, *46*(12), 3744–3749.

Aminov, R. I. (2010). A brief history of the antibiotic era: lessons learned and challenges for the future. *Frontiers in Microbiology*, *1*, 134. doi:10.3389/fmicb.2010.00134

Anderson, M. T., & Seifert, H. S. (2011). Opportunity and means: horizontal gene transfer from the human host to a bacterial pathogen. *MBio*.

Anney, R., Klei, L., Pinto, D., Regan, R., Conroy, J., Magalhaes, T. R., et al. (2010). A genome-wide scan for common alleles affecting risk for autism. *Human Molecular Genetics*, *19*(20), 4072–4082. doi:10.1093/hmg/ddq307

Bailey, S. F., Hinz, A., & Kassen, R. (2014). Adaptive synonymous mutations in an experimentally evolved Pseudomonas fluorescens population. *Nature communications*, *5*, 4076. doi:10.1038/ncomms5076

Barton, N. H., & Keightley, P. D. (2002). Understanding quantitative genetic variation. *Nature reviews. Genetics*, *3*(1), 11–21. doi:10.1038/nrg700

Barton, N. H., & Turelli, M. (1989). Evolutionary quantitative genetics: how little do we know? *Annual review of genetics*, *23*, 337–370. doi:10.1146/annurev.ge.23.120189.002005

Batut, B., Knibbe, C., Marais, G., & Daubin, V. (2014). Reductive genome evolution at both ends of the bacterial population size spectrum. *Nature reviews. Microbiology*, *12*(12), 841–850. doi:10.1038/nrmicro3331

Bauernfeind, A., Grimm, H., & Schweighart, S. (1990). A new plasmidic cefotaximase in a clinical isolate of Escherichia coli. *Infection*, *18*(5), 294–298.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B …*.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, *456*(7218), 53–59. doi:10.1038/nature07517

Bergey, D. H., & Breed, R. S. (1971). *Bergey's Manual of Determinative Bacteriology*. Williams and Wilkins.

Bhangale, T. R., Rieder, M. J., & Nickerson, D. A. (2008). Estimating coverage and power for genetic association studies using near-complete variation data. *Nature Publishing Group*, *40*(7), 841–843. doi:10.1038/ng.180

Biek, R., Pybus, O. G., Lloyd-Smith, J. O., & Didelot, X. (2015). Measurably evolving pathogens in the genomic era. *Trends in ecology & evolution*. doi:10.1016/j.tree.2015.03.009

Bryant, J., Chewapreecha, C., & Bentley, S. D. (2012). Developing insights into the mechanisms of evolution of bacterial pathogens from whole-genome sequences. *Future Microbiology*, *7*(11), 1283–1296. doi:10.2217/fmb.12.108

Bugg, T. D., Wright, G. D., Dutka-Malen, S., Arthur, M., Courvalin, P., & Walsh, C. T. (1991). Molecular basis for vancomycin resistance in Enterococcus faecium BM4147: biosynthesis of a depsipeptide peptidoglycan precursor by vancomycin resistance proteins VanH and VanA. *Biochemistry*, *30*(43), 10408–10415.

Cantón, R., & Coque, T. M. (2006). The CTX-M beta-lactamase pandemic. *Current opinion in microbiology*, *9*(5), 466–475. doi:10.1016/j.mib.2006.08.011

Carneiro, M. O., Russ, C., Ross, M. G., Gabriel, S. B., Nusbaum, C., & DePristo, M. A. (2012). Pacific biosciences sequencing technology for genotyping and variation discovery in human data. *BMC Genomics*, *13*, 375. doi:10.1186/1471-2164-13-375

Castillo-Ramírez, S., Corander, J., Marttinen, P., Aldeljawi, M., Hanage, W. P., Westh, H., et al. (2012). Phylogeographic variation in recombination rates within a global clone of methicillin-resistant Staphylococcus aureus. *Genome biology*, *13*(12), R126–R126. doi:10.1186/gb-2012-13-12-r126

Chakravarti, A. (2011). Genomic contributions to Mendelian disease. *Genome Research*, *21*(5), 643–644. doi:10.1101/gr.123554.111

Chattopadhyay, S., Weissman, S. J., Minin, V. N., Russo, T. A., Dykhuizen, D. E., & Sokurenko, E. V. (2009). High frequency of hotspot mutations in core genes of Escherichia coli due to short-term positive selection. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(30), 12412–12417. doi:10.1073/pnas.0906217106

Chelala, C., Khan, A., & Lemoine, N. R. (2009). SNPnexus: a web database for

functional annotation of newly discovered and public domain single nucleotide polymorphisms. *Bioinformatics*, *25*(5), 655–661. doi:10.1093/bioinformatics/btn653

Chen, J.-M., Cooper, D. N., Chuzhanova, N., Férec, C., & Patrinos, G. P. (2007). Gene conversion: mechanisms, evolution and human disease. *Nature reviews. Genetics*, *8*(10), 762–775. doi:10.1038/nrg2193

Chen, P. E., & Shapiro, B. J. (2015). The advent of genome-wide association studies for bacteria. *Current opinion in microbiology*, *25*, 17–24. doi:10.1016/j.mib.2015.03.002

Chisholm, S. A., Neal, T. J., Alawattegama, A. B., Birley, H. D. L., Howe, R. A., & Ison, C. A. (2009). Emergence of high-level azithromycin resistance in Neisseria gonorrhoeae in England and Wales. *The Journal of antimicrobial chemotherapy*, *64*(2), 353–358. doi:10.1093/jac/dkp188

Choi, M., Scholl, U. I., Ji, W., Liu, T., Tikhonova, I. R., Zumbo, P., et al. (2009). Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(45), 19096–19101. doi:10.1073/pnas.0910672106

Chung, R.-H., Ma, D., Wang, K., Hedges, D. J., Jaworski, J. M., Gilbert, J. R., et al. (2011). An X chromosome-wide association study in autism families identifies TBL1X as a novel autism spectrum disorder candidate gene in males. *Molecular Autism*, *2*(1), 18–18. doi:10.1186/2040-2392-2-18

Clemente, J. C., Pehrsson, E. C., Blaser, M. J., Sandhu, K., Gao, Z., Wang, B., et al. (2015). The microbiome of uncontacted Amerindians. *Science Advances*, *1*(3), e1500183–e1500183. doi:10.1126/sciadv.1500183

Cohen, M. S., Hoffman, I. F., Royce, R. A., Kazembe, P., Dyer, J. R., Daly, C. C., et al. (1997). Reduction of concentration of HIV-1 in semen after treatment of urethritis: implications for prevention of sexual transmission of HIV-1. AIDSCAP Malawi Research Group. *Lancet*, *349*(9069), 1868–1873.

Comas, I., Borrell, S., Roetzer, A., Rose, G., Malla, B., Kato-Maeda, M., et al. (2012). Whole-genome sequencing of rifampicin-resistant Mycobacterium tuberculosis strains identifies compensatory mutations in RNA polymerase genes. *Nature Publishing Group*, *44*(1), 106–110. doi:10.1038/ng.1038

Corander, J., & Marttinen, P. (2006). Bayesian identification of admixture events using multilocus molecular markers. *Molecular ecology*, *15*(10), 2833–2843. doi:10.1111/j.1365-294X.2006.02994.x

Croucher, N. J., Harris, S. R., Fraser, C., Quail, M. A., Burton, J., van der Linden, M., et al. (2011). Rapid pneumococcal evolution in response to clinical interventions. *Science*, *331*(6016), 430–434. doi:10.1126/science.1198545

Croucher, N. J., Page, A. J., Connor, T. R., Delaney, A. J., Keane, J. A., Bentley, S. D., et al. (2015). Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Research*, *43*(3), e15–e15. doi:10.1093/nar/gku1196

Darling, A. C. E., Mau, B., Blattner, F. R., & Perna, N. T. (2004). Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Research*, *14*(7), 1394–1403. doi:10.1101/gr.2289704

Davies, J., & Davies, D. (2010). Origins and evolution of antibiotic resistance. *Microbiology and molecular biology reviews : MMBR*, *74*(3), 417–433. doi:10.1128/MMBR.00016-10

Deguchi, T., Nakane, K., Yasuda, M., & Maeda, S.-I. (2010). Emergence and spread of drug resistant Neisseria gonorrhoeae. *The Journal of urology*, *184*(3), 851–8– quiz 1235. doi:10.1016/j.juro.2010.04.078

Dethlefsen, L., Huse, S., Sogin, M. L., & Relman, D. A. (2008). The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. *PLoS biology*, *6*(11), e280. doi:10.1371/journal.pbio.0060280

Devlin, B., Melhem, N., & Roeder, K. (2011). Do common variants play a role in risk for autism? Evidence and theoretical musings. *Brain research*, *1380*, 78–84. doi:10.1016/j.brainres.2010.11.026

Didelot, X., & Falush, D. (2007). Inference of bacterial microevolution using multilocus sequence data. *Genetics*, *175*(3), 1251–1266. doi:10.1534/genetics.106.063305

Didelot, X., & Maiden, M. C. J. (2010). Impact of recombination on bacterial evolution. *Trends in Microbiology*, *18*(7), 315–322. doi:10.1016/j.tim.2010.04.002

Didelot, X., Bowden, R., Wilson, D. J., Peto, T. E. A., & Crook, D. W. (2012a). Transforming clinical microbiology with bacterial genome sequencing. *Nature reviews. Genetics*, *13*(9), 601–612. doi:10.1038/nrg3226

Didelot, X., Méric, G., Falush, D., & Darling, A. E. (2012b). Impact of homologous and non-homologous recombination in the genomic evolution of Escherichia coli. *BMC Genomics*, *13*, 256. doi:10.1186/1471-2164-13-256

Dominguez-Bello, M. G., Costello, E. K., Contreras, M., Magris, M., Hidalgo, G., Fierer, N., & Knight, R. (2010). Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(26), 11971–11975. doi:10.1073/pnas.1002601107

Drake, J. W. (1991). Spontaneous mutation. *Annual review of genetics*, *25*, 125–146. doi:10.1146/annurev.ge.25.120191.001013

Dress, A. W. M., Flamm, C., Fritzsch, G., Grünewald, S., Kruspe, M., Prohaska, S. J., & Stadler, P. F. (2008). Noisy: identification of problematic columns in multiple sequence alignments. *Algorithms for molecular biology : AMB*, *3*, 7. doi:10.1186/1748-7188-3-7

Drmanac, R., Sparks, A. B., Callow, M. J., Halpern, A. L., Burns, N. L., Kermani, B. G., et al. (2010). Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science*, *327*(5961), 78–81. doi:10.1126/science.1181498

Drummond, A. J., & Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC evolutionary biology*, *7*, 214. doi:10.1186/1471-2148-7-214

DYKHUIZEN, D. E. (1990). *Experimental studies of natural selection in bacteria* (Vol. 21, pp. 373–398). Annual Review of Ecology and Systematics.

Eaton, W. W., Martins, S. S., Nestadt, G., Bienvenu, O. J., Clarke, D., & Alexandre, P. (2008). The burden of mental disorders. *Epidemiologic Reviews*, *30*(1), 1–14. doi:10.1093/epirev/mxn011

Edgar, R. C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, *5*, 113–113. doi:10.1186/1471-2105-5-113

Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H., & Nadeau, J. H. (2010). Missing heritability and strategies for finding the underlying causes of

complex disease. *Nature reviews. Genetics*, *11*(6), 446–450. doi:10.1038/nrg2809

Ezewudo, M. N., Joseph, S. J., Castillo-Ramírez, S., Dean, D., Del Rio, C., Didelot, X., et al. (2015). Population structure of Neisseria gonorrhoeae based on whole genome data and its relationship with antibiotic resistance. *PeerJ*, *3*, e806–e806. doi:10.7717/peerj.806

Ezewudo, M., & Zwick, M. E. (2013). Evaluating rare variants in complex disorders using next-generation sequencing. *Current psychiatry reports*, *15*(4), 349. doi:10.1007/s11920-013-0349-4

Ezewudo, M., Ramachandran, D., Mondal, K., Zwick, M. E., Bose, P., & Patel, V. (2012). *SeqAnt 2012: Recent Developments in Next-Generation Sequencing Annotation*. (H. perez-Sanchez, Ed.) (77 ed.). Intech.

Falconer, D. S. (1967). The inheritance of liability to diseases with variable age of onset, with particular reference to diabetes mellitus. *Annals of human genetics*, *31*(1), 1–20.

Falush, D., Stephens, M., & Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, *164*(4), 1567–1587.

Fang, F., Ding, J., Minin, V. N., Suchard, M. A., & Dorman, K. S. (2007). cBrother: relaxing parental tree assumptions for Bayesian recombination detection. *Bioinformatics*, *23*(4), 507–508. doi:10.1093/bioinformatics/btl613

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

Fay, J. C. (2011). Weighing the evidence for adaptation at the molecular level. *Trends in Genetics*, *27*(9), 343–349. doi:10.1016/j.tig.2011.06.003

Felsenstein, J., & Churchill, G. A. (1996). A Hidden Markov Model approach to variation among sites in rate of evolution. *Molecular Biology and Evolution*, *13*(1), 93–104.

Felstein, J. (1989). PHYLIP-Phylogeny Inference Package. *Cladistics*, *5*(2), 163–166. doi:10.1111/j.1096-0031.1989.tb00562.x

Ferrarini, M., Moretto, M., Ward, J. A., Šurbanovski, N., Stevanović, V., Giongo, L., et al. (2013). An evaluation of the PacBio RS platform for sequencing and de novo assembly of a chloroplast genome. *BMC Genomics*, *14*, 670. doi:10.1186/1471-2164-14-670

Fischbach, M. A., & Walsh, C. T. (2009). Antibiotics for emerging pathogens. *Science*.

Fisher, J. F., Meroueh, S. O., & Mobashery, S. (2005). Bacterial resistance to beta-lactam antibiotics: compelling opportunism, compelling opportunity. *Chemical reviews*, *105*(2), 395–424. doi:10.1021/cr030102i

Fisher, R. A. (1922). On the dominance ratio. *Proceedings of the Royal Society of Edinburgh*, *42*, 321–341.

Fledel-Alon, A., Wilson, D. J., Broman, K., Wen, X., Ober, C., Coop, G., & Przeworski, M. (2009). Broad-scale recombination patterns underlying proper disjunction in humans. *PLoS genetics*, *5*(9), e1000658. doi:10.1371/journal.pgen.1000658

Ford, C. B., Lin, P. L., Chase, M. R., Shah, R. R., Iartchouk, O., Galagan, J., et al. (2011). Use of whole genome sequencing to estimate the mutation rate of Mycobacterium tuberculosis during latent infection. *Nature Publishing Group*, *43*(5), 482–486. doi:10.1038/ng.811

Fraser, C., Hanage, W. P., & Spratt, B. G. (2007). Recombination and the nature of bacterial speciation. *Science*, *315*(5811), 476–480. doi:10.1126/science.1127573

Fujimoto, A., Nakagawa, H., Hosono, N., Nakano, K., Abe, T., Boroevich, K. A., et al. (2010). Whole-genome sequencing and comprehensive variant analysis of a Japanese individual using massively parallel sequencing. *Nature Publishing Group*, *42*(11), 931–936. doi:10.1038/ng.691

Gardner, S. N., & Hall, B. G. (2013). When Whole-Genome Alignments Just Won't Work: kSNP v2 Software for Alignment-Free SNP Discovery and Phylogenetics of Hundreds of Microbial Genomes. (H. Tang, Ed.)*PLoS ONE*, *8*(12), e81760. doi:10.1371/journal.pone.0081760.s004

Garvin, L. E., Bash, M. C., Keys, C., Warner, D. M., Ram, S., Shafer, W. M., & Jerse, A. E. (2008). Phenotypic and Genotypic Analyses of Neisseria gonorrhoeae Isolates That Express Frequently Recovered PorB PIA Variable Region Types Suggest that Certain P1a Porin Sequences Confer a Selective Advantage for Urogenital Tract Infection. *Infection and Immunity*, *76*(8), 3700–3709. doi:10.1128/IAI.00265-08

Girard, S. L., Gauthier, J., Noreau, A., Xiong, L., Zhou, S., Jouan, L., et al. (2011). Increased exonic de novo mutation rate in individuals with schizophrenia. *Nature Publishing Group*, *43*(9), 860–863. doi:10.1038/ng.886

Girirajan, S., Campbell, C. D., & Eichler, E. E. (2011). Human copy number variation and complex genetic disease. *Annual review of genetics*, *45*, 203–226. doi:10.1146/annurev-genet-102209-163544

Gladman, S., & Seemann, T. (2012, December 22). Velvet Optimiser. *vicbioinformatics.com*. Retrieved May 30, 2014, from http://www.vicbioinformatics.com/software.velvetoptimiser.shtml

Glenn, T. C. (2011). Field guide to next-generation DNA sequencers. *Molecular ecology resources*, *11*(5), 759–769. doi:10.1111/j.1755-0998.2011.03024.x

Glessner, J. T., Wang, K., Cai, G., Korvatska, O., Kim, C. E., Wood, S., et al. (2009). Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature*, *459*(7246), 569–573. doi:10.1038/nature07953

Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E. M., Brockman, W., et al. (2009). Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature biotechnology*, *27*(2), 182–189. doi:10.1038/nbt.1523

Goire, N., Lahra, M., Chen, M., Donovan, B., Fairley, C., Guy, R., Kaldor, J, Regan, J. Nissen, M., Sloots, T., Whiley, D. (2014) Molecular approaches to enhance surveillance of gonococcal antimicrobial resistance. Nature reviews Microbiology (12) 223-229

Götz, S., Arnold, R., Sebastián-León, P., Martín-Rodríguez, S., Tischler, P., Jehl, M.-A., et al. (2011). B2G-FAR, a species-centered GO annotation repository. *Journal of Gerontology*, *27*(7), 919–924. doi:10.1093/bioinformatics/btr059

Grad, Y. H., Kirkcaldy, R. D., Trees, D., Dordel, J., Harris, S. R., Goldstein, E., et al. (2014). Genomic epidemiology of Neisseria gonorrhoeae with reduced susceptibility to cefixime in the USA: a retrospective observational study. *The Lancet Infectious Diseases*, *14*(3), 220–226. doi:10.1016/S1473-3099(13)70693-5

Hacker, J., & Carniel, E. (2001). Ecological fitness, genomic islands and bacterial pathogenicity. A Darwinian view of the evolution of microbes. *EMBO Reports*, *2*(5),

376–381. doi:10.1093/embo-reports/kve097
Hagman, K. E., & Shafer, W. M. (1995). Transcriptional control of the mtr efflux system of Neisseria gonorrhoeae. *Journal of bacteriology*, *177*(14), 4162–4165.
Hall, B. G. (2014). SNP-Associations and Phenotype Predictions from Hundreds of Microbial Genomes without Genome Alignments. (M. E. Zwick, Ed.)*PLoS ONE*, *9*(2), e90490. doi:10.1371/journal.pone.0090490.s007
Hamilton, H. L., & Dillard, J. P. (2006). Natural transformation of Neisseria gonorrhoeae: from DNA donation to homologous recombination. *Molecular Microbiology*, *59*(2), 376–385. doi:10.1111/j.1365-2958.2005.04964.x
Hannan, S., Ready, D., Jasni, A. S., Rogers, M., Pratten, J., & Roberts, A. P. (2010). Transfer of antibiotic resistance by transformation with eDNA within oral biofilms. *FEMS immunology and medical microbiology*, *59*(3), 345–349. doi:10.1111/j.1574-695X.2010.00661.x
Harris, S. R., Feil, E. J., Holden, M. T. G., Quail, M. A., Nickerson, E. K., Chantratita, N., et al. (2010). Evolution of MRSA during hospital transmission and intercontinental spread. *Science*, *327*(5964), 469–474. doi:10.1126/science.1182395
Hodges, E., Xuan, Z., Balija, V., Kramer, M., Molla, M. N., Smith, S. W., et al. (2007). Genome-wide in situ exon capture for selective resequencing. *Nature Publishing Group*, *39*(12), 1522–1527. doi:10.1038/ng.2007.42
Human Microbiome Jumpstart Reference Strains Consortium, Nelson, K. E., Weinstock, G. M., Highlander, S. K., Worley, K. C., Creasy, H. H., et al. (2010). A catalog of reference genomes from the human microbiome. *Science*, *328*(5981), 994–999. doi:10.1126/science.1183605
Iafrate, A. J., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y., et al. (2004). Detection of large-scale variation in the human genome. *Nature Genetics*, *36*(9), 949–951. doi:10.1038/ng1416
International Schizophrenia Consortium, Purcell, S. M., Wray, N. R., Stone, J. L., Visscher, P. M., O'Donovan, M. C., et al. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, *460*(7256), 748–752. doi:10.1038/nature08185
Iossifov, I., Ronemus, M., Levy, D., Wang, Z., Hakker, I., Rosenbaum, J., et al. (2012). De novo gene disruptions in children on the autistic spectrum. *Neuron*, *74*(2), 285–299. doi:10.1016/j.neuron.2012.04.009
Jakobsson, H. E., Jernberg, C., Andersson, A. F., Sjölund-Karlsson, M., Jansson, J. K., & Engstrand, L. (2010). Short-term antibiotic treatment has differing long-term impacts on the human throat and gut microbiome. *PLoS ONE*, *5*(3), e9836. doi:10.1371/journal.pone.0009836
Jernberg, C., Löfmark, S., Edlund, C., & Jansson, J. K. (2007). Long-term ecological impacts of antibiotic administration on the human intestinal microbiota. *The ISME Journal*, *1*(1), 56–66. doi:10.1038/ismej.2007.3
Jolley, K. A., & Maiden, M. C. J. (2010). BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics*, *11*, 595. doi:10.1186/1471-2105-11-595
Jordan, I. K., Rogozin, I. B., Wolf, Y. I., & Koonin, E. V. (2002). Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Research*, *12*(6), 962–968. doi:10.1101/gr.87702

Joseph, S. J., Didelot, X., Rothschild, J., de Vries, H. J. C., Morré, S. A., Read, T. D., & Dean, D. (2012). Population genomics of Chlamydia trachomatis: insights on drift, selection, recombination, and population structure. *Molecular Biology and Evolution*, *29*(12), 3933–3946. doi:10.1093/molbev/mss198

Kaplan, N. L., Hudson, R., Langley, C. H. (1989). The "Hitchhiking effect" Revisited. *Genetics,* 123(4), 887 - 899.

Keinan, A., & Clark, A. G. (2012). Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science*, *336*(6082), 740–743. doi:10.1126/science.1217283

Kidd, J. M., Cooper, G. M., Donahue, W. F., Hayden, H. S., Sampas, N., Graves, T., et al. (2008). Mapping and sequencing of structural variation from eight human genomes. *Nature*, *453*(7191), 56–64. doi:10.1038/nature06862

Kiezun, A., Garimella, K., Do, R., Stitziel, N. O., Neale, B. M., McLaren, P. J., et al. (2012). Exome sequencing and the genetic basis of complex traits. *Nature Publishing Group*, *44*(6), 623–630. doi:10.1038/ng.2303

KIMURA, M., & OHTA, T. (1971). Protein Polymorphism as a Phase of Molecular Evolution. *Nature*, *229*(5285), 467–469. doi:10.1038/229467a0

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., & Cooper, G. M. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Genetics*.

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J. Y., & Sackler, R. S. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science*.

Knapp, J. S., Johnson, S. R., Zenilman, J. M., Roberts, M. C., & Morse, S. A. (1988). High-level tetracycline resistance resulting from TetM in strains of Neisseria spp., Kingella denitrificans, and Eikenella corrodens. *Antimicrobial Agents and Chemotherapy*, *32*(5), 765–767.

Kohanski, M. A., DePristo, M. A., & Collins, J. J. (2010a). Sublethal antibiotic treatment leads to multidrug resistance via radical-induced mutagenesis. *Molecular cell*, *37*(3), 311–320. doi:10.1016/j.molcel.2010.01.003

Kohanski, M. A., Dwyer, D. J., & Collins, J. J. (2010b). How antibiotics kill bacteria: from targets to networks. *Nature Publishing Group*, *8*(6), 423–435. doi:10.1038/nrmicro2333

Kosakovsky Pond, S. L., Posada, D., Gravenor, M. B., Woelk, C. H., & Frost, S. D. W. (2006). GARD: a genetic algorithm for recombination detection. *Bioinformatics*, *22*(24), 3096–3098. doi:10.1093/bioinformatics/btl474

Krause, R. M. (1992). The origin of plagues: old and new. *Science*.

Kumar, R. A., KaraMohamed, S., Sudi, J., Conrad, D. F., Brune, C., Badner, J. A., et al. (2008). Recurrent 16p11.2 microdeletions in autism. *Human Molecular Genetics*, *17*(4), 628–638. doi:10.1093/hmg/ddm376

Kumar, S. (2005). Molecular clocks: four decades of evolution. *Nature reviews. Genetics*, *6*(8), 654–662. doi:10.1038/nrg1659

Lande, R. (2007). The maintenance of genetic variability by mutation in a polygenic character with linked loci. *Genetical research*, *89*(5-6), 373–387. doi:10.1017/S0016672308009555

Lander, E. S. (1996). The new genomics: global views of biology. *Science*, *274*(5287), 536–539.

Lawson, D. J., Hellenthal, G., Myers, S., & Falush, D. (2012). Inference of population structure using dense haplotype data. *PLoS genetics*, *8*(1), e1002453. doi:10.1371/journal.pgen.1002453

Lee, S. H., DeCandia, T. R., Ripke, S., Yang, J., Schizophrenia Psychiatric Genome-Wide Association Study Consortium (PGC-SCZ), International Schizophrenia Consortium (ISC), et al. (2012). Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nature Publishing Group*, *44*(3), 247–250. doi:10.1038/ng.1108

Lefébure, T., & Stanhope, M. J. (2009). Pervasive, genome-wide positive selection leading to functional divergence in the bacterial genus Campylobacter. *Genome Research*, *19*(7), 1224–1232. doi:10.1101/gr.089250.108

Levin, B. R., Lipsitch, M., & Bonhoeffer, S. (1999). Population biology, evolution, and infectious disease: convergence and synthesis. *Science*, *283*(5403), 806–809.

Levy, D., Ronemus, M., Yamrom, B., Lee, Y.-H., Leotta, A., Kendall, J., et al. (2011). Rare de novo and transmitted copy-number variation in autistic spectrum disorders. *Neuron*, *70*(5), 886–897. doi:10.1016/j.neuron.2011.05.015

Levy, S. B., & Marshall, B. (2004). Antibacterial resistance worldwide: causes, challenges and responses. *Nature medicine*.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Journal of Gerontology*, *25*(14), 1754–1760. doi:10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Journal of Gerontology*, *25*(16), 2078–2079. doi:10.1093/bioinformatics/btp352

Li, L., Stoeckert, C. J., & Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research*, *13*(9), 2178–2189. doi:10.1101/gr.1224503

Lindberg, R., Fredlund, H., Nicholas, R., & Unemo, M. (2007). Neisseria gonorrhoeae isolates with reduced susceptibility to cefixime and ceftriaxone: association with genetic polymorphisms in penA, mtrR, porB1b, and ponA. *Antimicrobial Agents and Chemotherapy*, *51*(6), 2117–2122. doi:10.1128/AAC.01604-06

Livermore, D. M. (2004). The need for new antibiotics. *Clinical Microbiology and Infection*.

Luikart, G., England, P. R., Tallmon, D., Jordan, S., & Taberlet, P. (2003). The power and promise of population genomics: from genotyping to genome typing. *Nature reviews. Genetics*, *4*(12), 981–994. doi:10.1038/nrg1226

Lupski, J. R., Reid, J. G., Gonzaga-Jauregui, C., Rio Deiros, D., Chen, D. C. Y., Nazareth, L., et al. (2010). Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. *The New England journal of medicine*, *362*(13), 1181–1191. doi:10.1056/NEJMoa0908094

Luria, S. E., & Delbrück, M. (1943). Mutations of Bacteria from Virus Sensitivity to Virus Resistance. *Genetics*, *28*(6), 491–511.

Lynch, M. (2007). The evolution of genetic networks by non-adaptive processes. *Nature reviews. Genetics*, *8*(10), 803–813. doi:10.1038/nrg2192

Lyon, G. J., & Wang, K. (2012). Identifying disease mutations in genomic medicine settings: current challenges and how to accelerate progress. *Genome medicine*, *4*(7),

58. doi:10.1186/gm359
M G Lorenz, W. W. (1994). Bacterial gene transfer by natural genetic transformation in the environment. *Microbiological reviews*, *58*(3), 563.
MacLean, R. C., Hall, A. R., Perron, G. G., & Buckling, A. (2010). The population genetics of antibiotic resistance: integrating molecular mechanisms and treatment contexts. *Nature reviews. Genetics*, *11*(6), 405–414. doi:10.1038/nrg2778
Magri, C., Sacchetti, E., Traversa, M., Valsecchi, P., Gardella, R., Bonvicini, C., et al. (2010). New copy number variations in schizophrenia. *PLoS ONE*, *5*(10), e13422. doi:10.1371/journal.pone.0013422
Maher, B. (2008, November 6). Personal genomes: The case of the missing heritability. *Nature*, pp. 18–21. doi:10.1038/456018a
Mamanova, L., Coffey, A. J., Scott, C. E., Kozarewa, I., Turner, E. H., Kumar, A., et al. (2010). Target-enrichment strategies for next-generation sequencing. *Nature methods*, *7*(2), 111–118. doi:10.1038/nmeth.1419
Manolio, T. A. (2010). Genomewide association studies and assessment of the risk of disease. *The New England journal of medicine*, *363*(2), 166–176. doi:10.1056/NEJMra0905980
Manolio, T. A., Brooks, L. D., & Collins, F. S. (2008). A HapMap harvest of insights into the genetics of common disease. *The Journal of clinical investigation*, *118*(5), 1590–1605. doi:10.1172/JCI34772
Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., et al. (2009). Finding the missing heritability of complex diseases. *Nature*, *461*(7265), 747–753. doi:10.1038/nature08494
Marshall, C. R., Noor, A., Vincent, J. B., Lionel, A. C., Feuk, L., Skaug, J., et al. (2008). Structural variation of chromosomes in autism spectrum disorder. *The American Journal of Human Genetics*, *82*(2), 477–488. doi:10.1016/j.ajhg.2007.12.009
Martin, D. P., Lemey, P., Lott, M., Moulton, V., Posada, D., & Lefeuvre, P. (2010). RDP3: a flexible and fast computer program for analyzing recombination. *Journal of Gerontology*, *26*(19), 2462–2463. doi:10.1093/bioinformatics/btq467
McArthur, A. G., Waglechner, N., Nizam, F., Yan, A., Azad, M. A., Baylay, A. J., et al. (2013). The comprehensive antibiotic resistance database. *Antimicrobial Agents and Chemotherapy*, *57*(7), 3348–3357. doi:10.1128/AAC.00419-13
McCarthy, S. E., Makarov, V., Kirov, G., Addington, A. M., McClellan, J., Yoon, S., et al. (2009). Microduplications of 16p11.2 are associated with schizophrenia. *Nature Publishing Group*, *41*(11), 1223–1227. doi:10.1038/ng.474
McClellan, J., & King, M. C. (2010). Genetic heterogeneity in human disease. *Cell*.
McDonald, J. H., & Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. *Nature*, *351*(6328), 652–654. doi:10.1038/351652a0
McQuillan, R., Eklund, N., Pirastu, N., Kuningas, M., McEvoy, B. P., Esko, T., et al. (2012). Evidence of inbreeding depression on human height. *PLoS genetics*, *8*(7), e1002655–e1002655. doi:10.1371/journal.pgen.1002655
McVean, G., Awadalla, P., & Fearnhead, P. (2002). A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics*, *160*(3), 1231–1241.
MD, Y. H. G., MD, R. D. K., PhD, D. T., PhD, J. D., PhD, S. R. H., PhD, E. G., et al. (2014). Genomic epidemiology of. *The Lancet Infectious Diseases*, *14*(3), 220–226.

doi:10.1016/S1473-3099(13)70693-5
Mena, A., Smith, E. E., Burns, J. L., Speert, D. P., Moskowitz, S. M., Perez, J. L., & Oliver, A. (2008). Genetic adaptation of Pseudomonas aeruginosa to the airways of cystic fibrosis patients is catalyzed by hypermutation. *Journal of bacteriology*, *190*(24), 7910–7917. doi:10.1128/JB.01147-08
Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nature reviews. Genetics*, *11*(1), 31–46. doi:10.1038/nrg2626
Milkman, R., & Bridges, M. M. (1990). Molecular evolution of the Escherichia coli chromosome. III. Clonal frames. *Genetics*, *126*(3), 505–517.
Mira, A., Martín-Cuadrado, A. B., D'Auria, G., & Rodríguez-Valera, F. (2010). The bacterial pan-genome:a new paradigm in microbiology. *International Microbiology*, *13*(2), 45–57.
Mondal, K., Ramachandran, D., Patel, V. C., Hagen, K. R., Bose, P., Cutler, D. J., & Zwick, M. E. (2012). Excess variants in AFF2 detected by massively parallel sequencing of males with autism spectrum disorder. *Human Molecular Genetics*, *21*(19), 4356–4364. doi:10.1093/hmg/dds267
Mondal, K., Shetty, A. C., Patel, V., Cutler, D. J., & Zwick, M. E. (2011). Targeted sequencing of the human X chromosome exome. *Genomics*, *98*(4), 260–265. doi:10.1016/j.ygeno.2011.04.004
Moreno-De-Luca, D., SGENE Consortium, Mulle, J. G., Simons Simplex Collection Genetics Consortium, Kaminsky, E. B., Sanders, S. J., et al. (2010). Deletion 17q12 is a recurrent copy number variant that confers high risk of autism and schizophrenia. *The American Journal of Human Genetics*, *87*(5), 618–630. doi:10.1016/j.ajhg.2010.10.004
Morse, S. A., Johnson, S. R., Biddle, J. W., & Roberts, M. C. (1986). High-level tetracycline resistance in Neisseria gonorrhoeae is result of acquisition of streptococcal tetM determinant. *Antimicrobial Agents and Chemotherapy*, *30*(5), 664–670.
Mulle, J. G., Dodd, A. F., McGrath, J. A., Wolyniec, P. S., Mitchell, A. A., Shetty, A. C., et al. (2010). Microdeletions of 3q29 confer high risk for schizophrenia. *The American Journal of Human Genetics*, *87*(2), 229–236. doi:10.1016/j.ajhg.2010.07.013
Mwangi, M. M., Wu, S. W., Zhou, Y., Sieradzki, K., de Lencastre, H., Richardson, P., et al. (2007). Tracking the in vivo evolution of multidrug resistance in Staphylococcus aureus by whole-genome sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(22), 9451–9456. doi:10.1073/pnas.0609839104
Neale, B. M., Kou, Y., Liu, L., Ma'ayan, A., Samocha, K. E., Sabo, A., et al. (2012). Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature*, *485*(7397), 242–245. doi:10.1038/nature11011
Need, A. C., Ge, D., Weale, M. E., Maia, J., Feng, S., Heinzen, E. L., et al. (2009). A genome-wide investigation of SNPs and CNVs in schizophrenia. *PLoS genetics*, *5*(2), e1000373. doi:10.1371/journal.pgen.1000373
Need, A. C., McEvoy, J. P., Gennarelli, M., Heinzen, E. L., Ge, D., Maia, J. M., et al. (2012). Exome sequencing followed by large-scale genotyping suggests a limited role for moderately rare risk factors of strong effect in schizophrenia. *The American*

Journal of Human Genetics, *91*(2), 303–312. doi:10.1016/j.ajhg.2012.06.018
Neu, H. C. (1992). The crisis in antibiotic resistance. *Science*, *257*(5073), 1064–1073.
Nelson, K., Whittam, T.S., Selander, R., K. (1991) Nucleotide polymorphism and evolution in the glyceraldehyde-3-phosphate dehydrogenase gene (*gapA*) in natural populations of *Salmonella* and Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America, 88(15), 6667-6671*
Ng, S. B., Buckingham, K. J., Lee, C., Bigham, A. W., Tabor, H. K., Dent, K. M., et al. (2010). Exome sequencing identifies the cause of a mendelian disorder. *Nature Publishing Group*, *42*(1), 30–35. doi:10.1038/ng.499
Ng, S. B., Turner, E. H., Robertson, P. D., Flygare, S. D., Bigham, A. W., Lee, C., et al. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, *461*(7261), 272–276. doi:10.1038/nature08250
Noor, A., Whibley, A., Marshall, C. R., Gianakopoulos, P. J., Piton, A., Carson, A. R., et al. (2010). Disruption at the PTCHD1 Locus on Xp22.11 in Autism spectrum disorder and intellectual disability. *Science Translational Medicine*, *2*(49), 49ra68–49ra68. doi:10.1126/scitranslmed.3001267
O'Donovan, M. C., Craddock, N., Norton, N., Williams, H., Peirce, T., Moskvina, V., et al. (2008). Identification of loci associated with schizophrenia by genome-wide association and follow-up. *Nature Genetics*, *40*(9), 1053–1055. doi:10.1038/ng.201
O'Roak, B. J., Deriziotis, P., Lee, C., Vives, L., Schwartz, J. J., Girirajan, S., et al. (2011). Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nature Genetics*, *43*(6), 585–589. doi:10.1038/ng.835
O'Roak, B. J., Vives, L., Girirajan, S., Karakoc, E., Krumm, N., Coe, B. P., et al. (2012). Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature*, *485*(7397), 246–250. doi:10.1038/nature10989
O'Rourke, M., & Stevens, E. (1993). Genetic structure of Neisseria gonorrhoeae populations: a non-clonal pathogen. *Journal of general microbiology*, *139*(11), 2603–2611.
Ohneck, E. A., Zalucki, Y. M., Johnson, P. J. T., Dhulipala, V., Golparian, D., Unemo, M., et al. (2011). A novel mechanism of high-level, broad-spectrum antibiotic resistance caused by a single base pair change in Neisseria gonorrhoeae. *MBio*, *2*(5). doi:10.1128/mBio.00187-11
Ohnishi, M., Golparian, D., Shimuta, K., Saika, T., Hoshina, S., Iwasaku, K., et al. (2011). Is Neisseria gonorrhoeae Initiating a Future Era of Untreatable Gonorrhea?: Detailed Characterization of the First Strain with High-Level Resistance to Ceftriaxone. *Antimicrobial Agents and Chemotherapy*, *55*(7), 3538–3545. doi:10.1128/AAC.00325-11
Ohnishi, M., Watanabe, Y., Ono, E., Takahashi, C., Oya, H., Kuroki, T., et al. (2010). Spread of a chromosomal cefixime-resistant penA gene among different Neisseria gonorrhoeae lineages. *Antimicrobial Agents and Chemotherapy*, *54*(3), 1060–1067. doi:10.1128/AAC.01010-09
Ohta, T. (1973). Slightly deleterious mutant substitutions in evolution. *Nature*, *246*(5428), 96–98. doi:10.1038/246096a0
Okou, D. T., Locke, A. E., Steinberg, K. M., Hagen, K., Athri, P., Shetty, A. C., et al. (2009). Combining microarray-based genomic selection (MGS) with the Illumina Genome Analyzer platform to sequence diploid target regions. *Annals of human*

genetics, *73*(Pt 5), 502–513. doi:10.1111/j.1469-1809.2009.00530.x

Okou, D. T., Mondal, K., Faubion, W. A., Kobrynski, L. J., Denson, L. A., Mulle, J. G., et al. (2014). Exome sequencing identifies a novel FOXP3 mutation in a 2-generation family with inflammatory bowel disease. *Journal of pediatric gastroenterology and nutrition*, *58*(5), 561–568. doi:10.1097/MPG.0000000000000302

Okou, D. T., Steinberg, K. M., Middle, C., Cutler, D. J., Albert, T. J., & Zwick, M. E. (2007). Microarray-based genomic selection for high-throughput resequencing. *Nature methods*, *4*(11), 907–909. doi:10.1038/nmeth1109

Olson, A. B., Silverman, M., Boyd, D. A., McGeer, A., Willey, B. M., Pong-Porter, V., et al. (2005). Identification of a progenitor of the CTX-M-9 group of extended-spectrum beta-lactamases from Kluyvera georgiana isolated in Guyana. *Antimicrobial Agents and Chemotherapy*, *49*(5), 2112–2115. doi:10.1128/AAC.49.5.2112-2115.2005

Olson, M. V. (2012). Human genetic individuality. *Annual review of genomics and human genetics*, *13*, 1–27. doi:10.1146/annurev-genom-090711-163825

Paila, U., Chapman, B. A., Kirchner, R., & Quinlan, A. R. (2013). GEMINI: integrative exploration of genetic variation and genome annotations. *PLoS computational biology*, *9*(7), e1003153–e1003153. doi:10.1371/journal.pcbi.1003153

Palmer, H. M., Young, H., Winter, A., & Dave, J. (2008). Emergence and spread of azithromycin-resistant Neisseria gonorrhoeae in Scotland. *Journal of Antimicrobial Chemotherapy*, *62*(3), 490–494. doi:10.1093/jac/dkn235

Pan, W., & Spratt, B. G. (1994). Regulation of the permeability of the gonococcal cell envelope by the mtr system. *Molecular Microbiology*, *11*(4), 769–775.

Peirano, G., & Pitout, J. D. D. (2010). Molecular epidemiology of Escherichia coli producing CTX-M beta-lactamases: the worldwide emergence of clone ST131 O25:H4. *International journal of antimicrobial agents*, *35*(4), 316–321. doi:10.1016/j.ijantimicag.2009.11.003

Perry, G. H., Tchinda, J., McGrath, S. D., Zhang, J., Picker, S. R., Cáceres, A. M., et al. (2006). Hotspots for copy number variation in chimpanzees and humans. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(21), 8006–8011. doi:10.1073/pnas.0602318103

Piton, A., Gauthier, J., Hamdan, F. F., Lafrenière, R. G., Yang, Y., Henrion, E., et al. (2011). Systematic resequencing of X-chromosome synaptic genes in autism spectrum disorder and schizophrenia. *Molecular Psychiatry*, *16*(8), 867–880. doi:10.1038/mp.2010.54

Porreca, G. J., Zhang, K., Li, J. B., Xie, B., Austin, D., Vassallo, S. L., et al. (2007). Multiplex amplification of large sets of human exons. *Nature methods*, *4*(11), 931–936. doi:10.1038/nmeth1110

Pritchard, J. K., & Cox, N. J. (2002). The allelic architecture of human disease genes: common disease-common variant... or not?

Read, T. D., & Massey, R. C. (2014). Characterizing the genetic basis of bacterial phenotypes using genome-wide association studies: a new direction for bacteriology. *Genome medicine*, *6*(11), 109. doi:10.1186/s13073-014-0109-z

Renzoni, A., Andrey, D. O., Jousselin, A., Barras, C., Monod, A., Vaudaux, P., et al. (2011). Whole genome sequencing and complete genetic analysis reveals novel pathways to glycopeptide resistance in Staphylococcus aureus. *PLoS ONE*, *6*(6),

e21577–e21577. doi:10.1371/journal.pone.0021577
Risch, N., & Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science*, *273*(5281), 1516–1517.
Rocha, E. P. C., & Feil, E. J. (2010). Mutational patterns cannot explain genome composition: Are there any neutral sites in the genomes of bacteria? *PLoS genetics*, *6*(9), e1001104–e1001104. doi:10.1371/journal.pgen.1001104
Rocha, E. P. C., Smith, J. M., Hurst, L. D., Holden, M. T. G., Cooper, J. E., Smith, N. H., & Feil, E. J. (2006). Comparisons of dN/dS are time dependent for closely related bacterial genomes. *Journal of theoretical biology*, *239*(2), 226–235. doi:10.1016/j.jtbi.2005.08.037
Rosenberg, N. A., Pritchard, J. K., Weber, J. L., & Cann, H. M. (2002). Genetic structure of human populations. *Science*.
Rosenfeld, J. A., Mason, C. E., & Smith, T. M. (2012). Limitations of the human reference genome for personalized genomics. *PLoS ONE*, *7*(7), e40294. doi:10.1371/journal.pone.0040294
Sanborn, J. Z., Benz, S. C., Craft, B., Szeto, C., Kober, K. M., Meyer, L., et al. (2011). The UCSC Cancer Genomics Browser: update 2011. *Nucleic Acids Research*, *39*(Database issue), D951–9. doi:10.1093/nar/gkq1113
Sanders, S. J., Murtha, M. T., Gupta, A. R., Murdoch, J. D., Raubeson, M. J., Willsey, A. J., et al. (2012). De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature*, *485*(7397), 237–241. doi:10.1038/nature10945
Sawyer, S. (1989). Statistical tests for detecting gene conversion. *Molecular Biology and Evolution*, *6*(5), 526–538.
Schmieder, R., & Edwards, R. (2012). Insights into antibiotic resistance through metagenomic approaches. *Future Microbiology*, *7*(1), 73–89. doi:10.2217/fmb.11.135
Schork, N. J., Murray, S. S., Frazer, K. A., & Topol, E. J. (2009). Common vs. rare allele hypotheses for complex diseases. *Current Opinion in Genetics & Development*, *19*(3), 212–219. doi:10.1016/j.gde.2009.04.010
Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., et al. (2007). Strong association of de novo copy number mutations with autism. *Science*, *316*(5823), 445–449. doi:10.1126/science.1138659
Sebat, J., Lakshmi, B., Troge, J., Alexander, J., & Young, J. (2004). Large-scale copy number polymorphism in the human genome. *Science*.
Shapiro, B. J. (2014). Signatures of natural selection and ecological differentiation in microbial genomes. *Advances in Experimental Medicine and Biology*, *781*, 339–359. doi:10.1007/978-94-007-7347-9_17
Shapiro, B. J., David, L. A., Friedman, J., & Alm, E. J. (2009). Looking for Darwin's footprints in the microbial world. *Trends in Microbiology*, *17*(5), 196–204. doi:10.1016/j.tim.2009.02.002
Shapiro, B. J., Friedman, J., Cordero, O. X., Preheim, S. P., Timberlake, S. C., Szabó, G., et al. (2012). Population genomics of early events in the ecological differentiation of bacteria. *Science*, *336*(6077), 48–51. doi:10.1126/science.1218198
Sharp, A. J., Locke, D. P., McGrath, S. D., Cheng, Z., Bailey, J. A., Vallente, R. U., et al. (2005). Segmental Duplications and Copy-Number Variation in the Human Genome.

The American Journal of Human Genetics, *77*(1), 78–88. doi:10.1086/431652

Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nature biotechnology*, *26*(10), 1135–1145. doi:10.1038/nbt1486

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

Shetty, A. C., Athri, P., Mondal, K., Horner, V. L., Steinberg, K. M., Patel, V., et al. (2010). SeqAnt: a web service to rapidly identify and annotate DNA sequence variations. *BMC Bioinformatics*, *11*, 471–471. doi:10.1186/1471-2105-11-471

Shi, J., Levinson, D. F., Duan, J., Sanders, A. R., Zheng, Y., Pe'er, I., et al. (2009). Common variants on chromosome 6p22.1 are associated with schizophrenia. *Nature*, *460*(7256), 753–757. doi:10.1038/nature08192

Shoemaker, N. B., Vlamakis, H., Hayes, K., & Salyers, A. A. (2001). Evidence for extensive resistance gene transfer among Bacteroides spp. and among Bacteroides and other genera in the human colon. *Applied and Environmental Microbiology*, *67*(2), 561–568. doi:10.1128/AEM.67.2.561-568.2001

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, *15*(8), 1034–1050. doi:10.1101/gr.3715005

Simmons, S. L., Dibartolo, G., Denef, V. J., Goltsman, D. S. A., Thelen, M. P., & Banfield, J. F. (2008). Population genomic analysis of strain variation in Leptospirillum group II bacteria involved in acid mine drainage formation. *PLoS biology*, *6*(7), e177–e177. doi:10.1371/journal.pbio.0060177

Sjölund, M., Tano, E., Blaser, M. J., Andersson, D. I., & Engstrand, L. (2005). Persistence of resistant Staphylococcus epidermidis after single course of clarithromycin. *Emerging infectious diseases*, *11*(9), 1389–1393. doi:10.3201/eid1109.050124

Smith, N. H., Maynard Smith, J., & Spratt, B. G. (1995). Sequence evolution of the porB gene of Neisseria gonorrhoeae and Neisseria meningitidis: evidence of positive Darwinian selection. *Molecular Biology and Evolution*, *12*(3), 363–370.

Sniegowski, P. D., Gerrish, P. J., & Lenski, R. E. (1997). Evolution of high mutation rates in experimental populations of E. coli. *Nature*, *387*(6634), 703–705. doi:10.1038/42701

Sommer, M. O. A., & Dantas, G. (2011). Antibiotics and the resistant microbiome. *Current opinion in microbiology*, *14*(5), 556–563. doi:10.1016/j.mib.2011.07.005

Sommer, M. O. A., Dantas, G., & Church, G. M. (2009). Functional characterization of the antibiotic resistance reservoir in the human microflora. *Science*, *325*(5944), 1128–1131. doi:10.1126/science.1176950

Spratt, B. G., Bowler, L. D., Zhang, Q. Y., Zhou, J., & Smith, J. M. (1992). Role of interspecies transfer of chromosomal genes in the evolution of penicillin resistance in pathogenic and commensal Neisseria species. *... of molecular evolution*, *34*(2), 115–125. Retrieved from http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=1556747&retmode=ref&cmd=prlinks

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. doi:10.1093/bioinformatics/btu033

Stankiewicz, P., & Lupski, J. R. (2010). Structural variation in the human genome and its role in disease. *Annual review of medicine*, *61*, 437–455. doi:10.1146/annurev-med-100708-204735

Starnino, S., Stefanelli, P., Neisseria gonorrhoeae Italian Study Group. (2009). Azithromycin-resistant Neisseria gonorrhoeae strains recently isolated in Italy. *The Journal of antimicrobial chemotherapy*, *63*(6), 1200–1204. doi:10.1093/jac/dkp118

Stefansson, H., Rujescu, D., Cichon, S., Pietiläinen, O. P. H., Ingason, A., Steinberg, S., et al. (2008). Large recurrent microdeletions associated with schizophrenia. *Nature*, *455*(7210), 232–236. doi:10.1038/nature07229

Sun, Y. V. (2012). Integration of biological networks and pathways with genetic association studies. *Human genetics*, *131*(10), 1677–1686. doi:10.1007/s00439-012-1198-7

Suyama, M., Torrents, D., & Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research*, *34*(Web Server issue), 609–612. doi:10.1093/nar/gkl315

Talavera, G., & Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic Biology*, *56*(4), 564–577. doi:10.1080/10635150701472164

Tanaka, M. M., Bergstrom, C. T., & Levin, B. R. (2003). The evolution of mutator genes in bacterial populations: the roles of environmental change and timing. *Genetics*, *164*(3), 843–854.

Tang, J., Hanage, W. P., Fraser, C., & Corander, J. (2009). Identifying currents in the gene pool for bacterial populations using an integrative approach. *PLoS computational biology*, *5*(8), e1000455. doi:10.1371/journal.pcbi.1000455

Tennessen, J. A., Bigham, A. W., O'Connor, T. D., Fu, W., Kenny, E. E., Gravel, S., et al. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*, *337*(6090), 64–69. doi:10.1126/science.1219240

Tenover, F. C. (2006). Mechanisms of antimicrobial resistance in bacteria. *The American journal of medicine*.

Tewhey, R., Warner, J. B., Nakano, M., Libby, B., Medkova, M., David, P. H., et al. (2009). Microdroplet-based PCR enrichment for large-scale targeted sequencing. *Nature biotechnology*, *27*(11), 1025–1031. doi:10.1038/nbt.1583

Thakur, S. D., Levett, P. N., Horsman, G. B., & Dillon, J.-A. R. (2014). Molecular epidemiology of Neisseria gonorrhoeae isolates from Saskatchewan, Canada: utility of NG-MAST in predicting antimicrobial susceptibility regionally. *Sexually transmitted infections*, *90*(4), 297–302. doi:10.1136/sextrans-2013-051229

The International HapMap Consortium. (2005). A haplotype map of the human genome. *Nature*, *437*(7063), 1299–1320. doi:10.1038/nature04226

Thomas, C. M., & Nielsen, K. M. (2005). Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nature reviews. Microbiology*, *3*(9), 711–721. doi:10.1038/nrmicro1234

Thornton, T., & McPeek, M. S. (2010). ROADTRIPS: case-control association testing with partially or completely unknown population and pedigree structure. *The*

American Journal of Human Genetics, *86*(2), 172–184. doi:10.1016/j.ajhg.2010.01.001

Tomberg, J., Unemo, M., Ohnishi, M., Davies, C., & Nicholas, R. A. (2013). Identification of amino acids conferring high-level resistance to expanded-spectrum cephalosporins in the penA gene from Neisseria gonorrhoeae strain H041. *Antimicrobial Agents and Chemotherapy*, *57*(7), 3029–3036. doi:10.1128/AAC.00093-13

Tong, P., Prendergast, J. G. D., Lohan, A. J., Farrington, S. M., Cronin, S., Friel, N., et al. (2010). Sequencing and analysis of an Irish human genome. *Genome biology*, *11*(9), R91. doi:10.1186/gb-2010-11-9-r91

Turner, A., Gough, K. R., & Leeming, J. P. (1999). Molecular epidemiology of tetM genes in Neisseria gonorrhoeae. *Sexually transmitted infections*, *75*(1), 60–66.

Unemo, M., & Shafer, W. M. (2011). Antibiotic resistance in Neisseria gonorrhoeae: origin, evolution, and lessons learned for the future. *Annals of the New York Academy of Sciences*, *1230*(1), E19–E28. doi:10.1111/j.1749-6632.2011.06215.x

Unemo, M., & Shafer, W. M. (2014). Antimicrobial resistance in Neisseria gonorrhoeae in the 21st century: past, evolution, and future. *Clinical Microbiology Reviews*, *27*(3), 587–613. doi:10.1128/CMR.00010-14

Unemo, M., Golparian, D., & Hellmark, B. (2014). First three Neisseria gonorrhoeae isolates with high-level resistance to azithromycin in Sweden: a threat to currently available dual-antimicrobial regimens for treatment of gonorrhea? *Antimicrobial Agents and Chemotherapy*, *58*(1), 624–625. doi:10.1128/AAC.02093-13

Unemo, M., Golparian, D., Nicholas, R., Ohnishi, M., Gallay, A., & Sednaoui, P. (2012). High-Level Cefixime- and Ceftriaxone-Resistant Neisseria gonorrhoeae in France: Novel penA Mosaic Allele in a Successful International Clone Causes Treatment Failure. *Antimicrobial Agents and Chemotherapy*, *56*(3), 1273–1280. doi:10.1128/AAC.05760-11

Veal, W. L., Nicholas, R. A., & Shafer, W. M. (2002). Overexpression of the MtrC-MtrD-MtrE efflux pump due to an mtrR mutation is required for chromosomally mediated penicillin resistance in Neisseria gonorrhoeae. *Journal of bacteriology*, *184*(20), 5619–5624.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., & Mural, R. J. (2001). The sequence of the human genome. *Science*.

Vidovic, S., Caron, C., Taheri, A., Thakur, S. D., Read, T. D., Kusalik, A., & Dillon, J.-A. R. (2014). Using Crude Whole-Genome Assemblies of Neisseria gonorrhoeae as a Platform for Strain Analysis: Clonal Spread of Gonorrhea Infection in Saskatchewan, Canada. *Journal of Clinical Microbiology*, *52*(10), 3772–3776. doi:10.1128/JCM.01502-14

Visscher, P. M., Brown, M. A., McCarthy, M. I., & Yang, J. (2012a). Five years of GWAS discovery. *The American Journal of Human Genetics*, *90*(1), 7–24. doi:10.1016/j.ajhg.2011.11.029

Visscher, P. M., Goddard, M. E., Derks, E. M., & Wray, N. R. (2012b). Evidence-based psychiatric genetics, AKA the false dichotomy between common and rare variant hypotheses. *Molecular Psychiatry*, *17*(5), 474–485. doi:10.1038/mp.2011.65

Wagner, A. (2008). Neutralism and selectionism: a network-based reconciliation. *Nature reviews. Genetics*, *9*(12), 965–974. doi:10.1038/nrg2473

Walsh, C. (2000). Molecular mechanisms that confer antibacterial drug resistance. *Nature*, *406*(6797), 775–781. doi:10.1038/35021219

Walsh, T., McClellan, J. M., McCarthy, S. E., Addington, A. M., Pierce, S. B., Cooper, G. M., et al. (2008). Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science*, *320*(5875), 539–543. doi:10.1126/science.1155174

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, *38*(16), e164. doi:10.1093/nar/gkq603

Wang, K., Zhang, H., Ma, D., Bucan, M., Glessner, J. T., Abrahams, B. S., et al. (2009). Common genetic variants on 5p14.1 associate with autism spectrum disorders. *Nature*, *459*(7246), 528–533. doi:10.1038/nature07999

Warner, D. M., Folster, J. P., Shafer, W. M., & Jerse, A. E. (2007). Regulation of the MtrC-MtrD-MtrE efflux-pump system modulates the in vivo fitness of Neisseria gonorrhoeae. *The Journal of infectious diseases*, *196*(12), 1804–1812. doi:10.1086/522964

Warner, D. M., Shafer, W. M., & Jerse, A. E. (2008). Clinically relevant mutations that cause derepression of the Neisseria gonorrhoeaeMtrC-MtrD-MtrE Efflux pump system confer different levels of antimicrobial resistance and in vivofitness. *Molecular Microbiology*, *70*(2), 462–478. doi:10.1111/j.1365-2958.2008.06424.x

Wassenaar, T. M., & Gaastra, W. (2001). Bacterial virulence: can we draw the line? *FEMS Microbiology Letters*.

Weiss, L. A., Arking, D. E., Gene Discovery Project of Johns Hopkins & the Autism Consortium, Daly, M. J., & Chakravarti, A. (2009). A genome-wide linkage and association scan reveals novel loci for autism. *Nature*, *461*(7265), 802–808. doi:10.1038/nature08490

Weiss, L. A., Shen, Y., Korn, J. M., Arking, D. E., Miller, D. T., Fossdal, R., et al. (2008). Association between microdeletion and microduplication at 16p11.2 and autism. *The New England journal of medicine*, *358*(7), 667–675. doi:10.1056/NEJMoa075974

Wellcome Trust Case Control Consortium. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, *447*(7145), 661–678. doi:10.1038/nature05911

Wheeler, D. A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., et al. (2008). The complete genome of an individual by massively parallel DNA sequencing. *Nature*, *452*(7189), 872–876. doi:10.1038/nature06884

Winau, F., Westphal, O., & Winau, R. (2004). Paul Ehrlich — in search of the magic bullet. *Microbes and Infection*, *6*(8), 786–789. doi:10.1016/j.micinf.2004.04.003

World Health Organization (WHO). (2012). Global action plan to control the spread and impact of antimicrobial resistance in *Neisseria gonorrhoea*.

Wright, G. D. (2012). The origins of antibiotic resistance. *Handbook of experimental pharmacology*, (211), 13–30. doi:10.1007/978-3-642-28951-4_2

Wright, G. D. G. (2007). The antibiotic resistome: the nexus of chemical and genetic diversity. *Audio, Transactions of the IRE Professional Group on*, *5*(3), 175–186. doi:10.1038/nrmicro1614

Wright, S. (1931). Evolution in Mendelian Populations. *Genetics*, *16*(2), 97–159.

Xu, B., Roos, J. L., Dexheimer, P., Boone, B., Plummer, B., Levy, S., et al. (2011). Exome sequencing supports a de novo mutational paradigm for schizophrenia. *Nature Publishing Group*, *43*(9), 864–868. doi:10.1038/ng.902

Xu, B., Roos, J. L., Levy, S., van Rensburg, E. J., Gogos, J. A., & Karayiorgou, M. (2008). Strong association of de novo copy number mutations with sporadic schizophrenia. *Nature Publishing Group*, *40*(7), 880–885. doi:10.1038/ng.162

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, *42*(7), 565–569. doi:10.1038/ng.608

Yang, Z. Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, *24*(8), 1586–1591. doi:10.1093/molbev/msm088

Yim, G., Wang, H. H., & Davies, J. (2007). Antibiotics as signalling molecules. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *362*(1483), 1195–1200. doi:10.1098/rstb.2007.2044

Zerbino, D. R., & Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, *18*(5), 821–829. doi:10.1101/gr.074492.107

Zhang, J., Chiodini, R., Badr, A., & Zhang, G. (2011). The impact of next-generation sequencing on genomics. *Journal of Genetics and Genomics*, *38*(3), 95–109. doi:10.1016/j.jgg.2011.02.003

Zhao, S., Duncan, M., Tomberg, J., Davies, C., Unemo, M., & Nicholas, R. A. (2009). Genetics of chromosomally mediated intermediate resistance to ceftriaxone and cefixime in Neisseria gonorrhoeae. *Antimicrobial Agents and Chemotherapy*, *53*(9), 3744–3751. doi:10.1128/AAC.00304-09

Zhou, Z., McCann, A., Litrup, E., Murphy, R., Cormican, M., Fanning, S., et al. (2013). Neutral Genomic Microevolution of a Recently Emerged Pathogen, Salmonella enterica Serovar Agona. *PLoS genetics*, *9*(4), e1003471. doi:10.1371/journal.pgen.1003471

Zinder, N. D., & Lederberg, J. (1952). Genetic exchange in Salmonella. *Journal of bacteriology*.

Zwick, M. E., Cutler, D. J., & Chakravarti, A. (2000). Patterns of genetic variation in Mendelian and complex traits. *Annual review of genomics and human genetics*, *1*, 387–407. doi:10.1146/annurev.genom.1.1.387