Target Capture as an Improved Diagnostic Method for Lyme Disease

A thesis presented

By

Indumathi Prakash

In the Sabeti/Lemieux Lab

to

The Committee on Degrees in Molecular and Cellular Biology

in partial fulfillment of the requirements
for a degree of
Bachelor of Arts Highest Honors (FOR MCB)
in the field of
Molecular and Cellular Biology

Harvard University
Cambridge, Massachusetts

March 28, 2022

Statement of Research

Research activities were conducted under the supervision of Dr. Jacob Lemieux, Instructor of Medicine at Mass General Hospital as well as Dr. Pardis Sabeti, Professor of Organismic and Evolutionary Biology at Harvard. The research was conducted at both Mass General Hospital and the Broad Institute.

I took part in research in the summer of 2020 while I was working on another project with multiplexing tick-borne diseases. This project allowed me to understand more about tick borne diseases and how the lab worked. During the fall of 2020, I started to outline and plan the PCR steps of this current target capture project with Bennett Shaw. During the spring of 2021, I took part in literary research about the different methods currently used in the field as well as different amplification methods. Over the summer, I planned out what I would do for my thesis as well as learning Python, which enabled me to analyze current hybrid capture data from the Sabeti Lab. Dr. Lemieux helped and guided me throughout this process. Unfortunately, I was not able to implement the target capture portion of my research due to COVID-19 and reagent delays but was able to design and plan out the exact steps, which will take place as soon as the reagents arrive.

For the computational portion, I worked directly with Dr. Lemieux. We met once a week to check in on progress. I went through any computational issues I was having and what specifically I should be analyzing and focusing on. I contributed to developing the hypothesis with the check in of my mentor. I wrote the scripts and pipeline independently and analyzed the computational results with the guidance of my mentor.

For bench research, Bennett Shaw trained me with guidance from Dr. Lemieux. He helped me in attaining reagents as well as planning out the PCR runs together. He taught me the right pipetting skills and how to keep samples free from contamination. I learned how to plan out a run on

my own, design an experiment, and analyze the results to make judgment calls on next steps. Arshdeep Singh, Nolan Holbrook, and Gordon Adams also helped in the benchwork for sequencing in terms of picking out reagents and analyzing why steps went wrong. I cite the collaborators above for the help they have given me. I wrote this document myself and received feedback from Dr. Lemieux, Jess Frith, and friends.

Target Capture as an Improved Diagnostic Method for Lyme Disease

Indumathi Prakash
Professor (Pardis Sabeti), Dominic Mao

Abstract

Tick-borne diseases affect thousands of people worldwide and pose an emerging public health threat. Lyme disease, caused by the bacterium *Borrelia burgdorferi*, is the most common vector-borne disease in North America and can cause serious disease if left untreated. The current standard diagnostic, a two-tiered serological testing scheme, is problematic because of false negative results in early disease. Direct molecular detection techniques such as polymerase chain reaction (PCR) and next-generation sequencing (NGS) would allow for rapid diagnosis in early disease but have limited sensitivity for detecting *B. burgdorferi* due to the rarity of spirochetes in the bloodstream.

We aim to use hybridization and target capture followed by NGS to boost the sensitivity of NGS and create a highly sensitive diagnostic technique. To assess the effect of hybridization and target capture on the sensitivity of NGS, we also used computational analysis tools to assess coverage depth across the entire *Borrelia* genome. In our analysis, we first show that the *B. burgdorferi* samples are characterized as positive, as well as that the sequencing of non-target capture runs were of high-quality and can be used as a comparison after target capture. We show that hybridization and target capture improve read depth and coverage of NGS. The workflow built using Python code will allow for future analysis of new target capture data. Target capture with NGS has the potential to improve the early detection of Lyme disease, thereby reducing the number of people diagnosed too late with this potentially devastating disease.

Acknowledgements

First, I will acknowledge Dr. Pardis Sabeti. In the summer of 2020 I reached out to her because of her amazing work on infectious diseases at the Broad Institute. I had been very interested in Lyme disease research because of my past experience suffering with Lyme disease and arthritis as a complication. I was looking at different labs to see if any were working on tick-borne disease research but could not find any. I had previously met Dr. Sabeti when she spoke as a lecturer at the Research Science Institute at MIT, a six week research program, while I was a junior in high school. I looked up the work she was taking part in and one section was tick-borne disease research. When I reached out to her, she was nothing but welcoming and kind. I am forever grateful for the amount of time and energy that Dr. Sabeti has invested in me and everything she has done to encourage me both as a person and a scientist. She has always guided me and supported me in anything I've reached out to her about, always making time despite her busy schedule, and I have grown so much because of her.

Next, I will thank Dr. Jacob Lemieux, who Dr. Sabeti connected me with as the direct mentor on this project. He helped me understand more about lab techniques and the computational aspect of analyzing data. The approach that he showed me is to be independent in research and try to figure out the steps of the process on your own. I have learned so much because of that. I have made many mistakes and because of that, I understand my research and the process so much more. He is always there as a guide when needed and made me feel included in the lab, producing a great working environment. Dr. Lemieux never thinks twice to provide guidance and he has taught me so much in the lab as well as how to approach situations professionally.

I also want to thank the lab members Bennett Shaw, Nolan Holbrook, Arshdeep Singh, Rohan Singh, and Gordon Adams for their support in the lab both scientifically and personally. The lab members were always so willing to help me with retrieving samples, working through some of the steps together, and always answering any questions I had. I very much appreciate having a great lab environment because of them and love coming to Mass General Hospital because of them. Most of all I am thankful for their friendship and compassion, working with me for many hours regardless of how late it was in the lab.

Finally, I will thank Dominic Mao, Irina Cashen, my family, and my friends for their support. Dominic and Irina were always ready to answer the many questions I had over email and always had an encouraging tone. The encouragement I got from everyone was what really helped me in this process and significantly contributed to my well being and learning.

Table of Contents

Chapter I: Introduction

Tick-borne diseases affect thousands of people worldwide and pose a serious problem in early diagnosis. There are a range of tick-borne diseases that infect humans from Powassan virus to Lyme disease, each with their own symptoms and effects on the human body. In America, the incidence of tick-borne diseases in humans rises each year, with the Northeast holding the highest concentration of tick cases. The past year held a record number of 47,743 cases with 33,666 of them attributed to Lyme disease (Kugeler et al.).

Lyme disease, the most common vector-borne disease in North America, can be serious if untreated and undiagnosed early. Many untreated symptoms for months lead to chronic arthritis, facial palsy, Lyme carditis, and other life-threatening conditions. Many of the early signs of Lyme disease are nonspecific, making a diagnosis of Lyme disease difficult (Murray and Shapiro, 2010).

The most common indication of early Lyme disease is a localized skin rash known as erythema migrans (EM). About seventy percent of diagnoses are made using EM, which presents about three to thirty days after the tick bite (Aucott et al., 2009). Identifying the early signs of infection is important, as almost sixty percent of serological testing is negative for patients in the early stages of infection because antibodies need to form in the body (Wormser et al., 2006). In addition, about 10 percent of patients with Lyme disease do not display EM. There needs to be a more sensitive way of diagnosing Lyme disease at the early stages before it spreads to areas such as the nervous system (Steere et al., 2003). This challenge is compounded by the poor performance of existing diagnostic tests for Lyme disease. Therefore, there is an urgent need for improved diagnostics which do not rely on physicians to visually diagnose EM. For these reasons,

we are focusing on molecular methods such as target capture and sequencing, to improve diagnostics.

Current methods of testing

1.1 Serological testing

Lyme disease is caused by infection with the bacterium *Borrelia burgdorferi*, which is a tick-borne spirochete. The standard approach to Lyme disease diagnosis is two-tiered serological testing. This testing approach includes an initial screening assay, optimized for sensitivity, that reflexes, if positive, to a more specific, confirmatory test (John and Taege, 2019). The first test is an ELISA, while the second is a Western Blot. Both tests must be positive to confirm a Lyme diagnosis. If the immunoassay is positive, then the immunoblot detects for *B. burgdorferi* IgM or IgG surface proteins that are more specific. The Western Blot must have ⅔ IgM bands and 5/10 IgG bands to be considered positive. The issue with this testing is false negatives are common early due to the insensitivity of the immunoassay during early disease. Furthermore, there is much error for false positives because the test is not very specific (Lantos et al., 2016). In addition, these serological tests are limited by the time for the body to make antibodies, and as a result, may not be positive at the time of initial clinical symptoms.

1.2 PCR testing

Another way that Lyme can be diagnosed is through nucleic acid amplification. This test identifies Lyme through detecting a specific sequence of DNA of the pathogen. For an accurate diagnostic test, it is important to maintain both specificity and sensitivity.

PCR is specific but not sensitive specifically for Lyme. The problem is that the pathogen DNA lies low in number in the blood and other bodily fluids. Thus, it can be hard to detect the Lyme through PCR (Schutzer et al., 2019). The only current use for *B. burgdorferi* is on synovial fluid, which is usually PCR positive when a patient has Lyme arthritis. In contrast to Lyme disease, PCR is commonly used and can be used for other tick-borne diseases such as Anaplasmosis or Babesiosis because of the ample amount of pathogen nucleic acid in the blood (Moore et al., 2016).

Despite its current limitations in Lyme disease, PCR remains an attractive approach for the diagnosis of Lyme disease, particularly if its sensitivity can be improved. A direct approach to detecting Lyme disease would be an important advance because early detection, where serological testing falls short, would then be possible (Bil-Lula et al., 2015). This is because PCR detects an active microbe and does not need to wait for the body to produce the antibody. With further development, PCR has the potential to work better than the serological two-tier test (Schutzer et al. 2019).

At MGH, the PCR machine and method uses dyes and quenchers for detection. As a first step, a published assay was adapted for use on clinical specimens, specifically the Tokarz assay. The primer and probe sequences used for the Lyme assay were derived from previous work with these pathogens by Tokarz et al. These derived probes are labeled on opposite ends with a reporter fluorescence dye and quencher molecules (Tokarz et al., 2018). Thus, during PCR, when there is an increase in fluorescence detected by one of the channels in the Cobas 480z, which is the PCR machine at MGH, it means that there is amplification of the specific target sequence (Roche Manual). The Taq DNA polymerase has thus cleaved the hybrid product. These assays we use have great performance, but a way to increase sensitivity of these assays and increase detection of

different strains would be through target capture (Livak et al., 1995). The OspA gene, the target of our assay, is known to be highly specific for *Borrelia burgdorferi.* The qPCR can run using a specific real-time qPCR assay which can detect this unique target. It is crucial to improve regular qPCR testing by increasing sensitivity through target capture (Dunn et al., 1990).

1.3 Target Capture

One way to improve diagnostic testing is through target capture and sequencing. The importance of target capture is that there are regions of the pathogen DNA which will be enriched, and this will improve sensitivity. It is used to target a specific gene or region of the human genome. In our case, we have re-purposed the same tool to fish out pathogen genomes from a clinical sample. Target capture can be used in a diagnostic context because we achieve increased sensitivity (Barbour, 2016). Target capture enriches certain regions of the pathogen DNA which increase sensitivity. In addition, sequencing after target capture allows the identification of different strains, which are important to diagnose because they have distinct clinical manifestations (Roy et al., 2019). Target capture will work by collecting/"baiting" target areas of DNA out of the human DNA. This is done by using a microarray and synthesizing DNA oligos onto it. The target DNA will be separated from the total human DNA since complementary strands will attach to the microarray (Jones and Good, 2016). In addition, this can be used in a diagnostic context because in comparison to unbiased sequencing, we achieve increased sensitivity. However, that increased sensitivity is only for what we have designed our probe set to look for. Therefore, if we do not know what we are looking for, this approach is not useful but the probe we use would mean that there would be no need for costly sequencing

(Gnirke et al., 2009). In addition, if improvements to testing are possible, the sensitivity and specificity would increase greatly compared to serological testing or standard PCR. Since the cost of sequencing continues to decrease, this method is increasingly cost-effective (Caboche et al., 2014a). Hybrid capture followed by sequencing is an increasingly cost-efficient method to help diagnose Lyme disease that does not involve culturing the *Borrelia.* Hybrid capture does not contribute to as many sequencing errors that whole genome amplification does (de Bourcy et al., 2014).

Target capture achieves a closer look at the *Borrelia Burgdorferi* genome through enriching a selective section of the genome prior to Next Generation Sequencing. The use of target capture is more cost-effective when compared to Whole Genome Sequencing and higher throughput when compared to multiplex PCR. In addition, target capture has proven to have lower target coverage variance and more accurate SNP calls (Jones and Good, 2016). There have been many uses of target capture in relation to complex diseases such as inflammatory bowel disease (Worthey et al., 2011), infantile mitochondrial disease (Calvo et al., 2012) and autism spectrum disorder (Iossifov et al., 2014). Although the use of target capture has reached beyond humans, and to species such as chimpanzees and *Drosophila Melanogaster*, it is not a commonly used method (Jones and Good, 2016). The drawback of target capture is the design of the capture probe. A recent study helps show that NGS and hybrid capture can reconstruct entire genomes of microorganisms from samples with a low amount of pathogen DNA (Gaudin and Desnues, 2018) which is the situation for *Borrelia Burgdorferi*. Our project aims to use similar methods to sequence low pathogen input from clinical samples. There are certain limitations, such as the need for expertise in library preparation and probe design. This project overcomes this limitation because the library is prepared based on a previous run of target capture at the Broad and changes made to use a more specific probe. *Borrelia* is

a different pathogen than in the study, however another study shows target capture works well for eukaryotic parasites which includes *Borrelia* (Metsky et al., 2019).

1.4 Sequencing strategy

Next Generation Sequencing (NGS) is a faster way to sequence a whole genome or thousands of genes when compared to the previously used Sanger Sequencing. NGS can detect variants and mutations with the overall steps of DNA fragmentation, library preparation and sequencing (Qin, 2019). It is especially useful in clinical settings because it can examine many targets at the same time. Since PCR is limited by primer and probe diversity, it does not provide the amount of information on genetic diversity and even coverage that next generation sequencing does (Caboche et al., 2014). In order to sequence, an index-tagged library needs to be quantified. This is done by using a qPCR assay which can be used to determine the concentration of each tagged library. We will also test the ratio of pathogen target DNA to background human DNA. The concentration is then used to collect samples accurately for sequencing. Once the index-tagged samples are pooled based on having equimolar concentrations, sequencing can occur using the Agilent multiplex sequencing protocol (Agilent 2011).

When comparing library preparation methods, a crucial step in target capture and NGS, Garcia-Garcia et al. explains the differences between NimbleGen, SureSelect, and NRCEE. Although NimbleGen is cost efficient, it is less efficient at recovering low input of the target sequences. Hence, it seems to be that SureSelect and NRCEE are good options for target- enriched library preparation. When comparing NRCEE and SureSelect, this study shows that NRCEE has higher dispersion which means it has less efficient

fragmentation (García-García et al., 2016). This study increases our confidence in choosing SureSelect, as it shows that it is the most efficient library preparation tested.

Target capture will provide a valuable tool not only to improve Lyme disease diagnostics, but also to discover mechanisms of pathogenesis and survey Lyme genetic diversity. Through comparing the sequencing before and after target capture, the improvement using target capture can be visualized.

1.5 Hypothesis

Target capture and sequencing will increase the sensitivity and specificity of Lyme disease testing.

Aim 1: Establish quantification methods for Borrelia Burgdorferi.

Aim 2: Establish sequencing methods for Borrelia Burgdorferi.

Aim 3: Develop methods for Target Capture.

1.6 Summary

Lyme disease affects thousands of lives; however, the disease is challenging to accurately and efficiently diagnose. The current standard, two-tiered serological testing is problematic because of the lag it takes to make the antibody for the pathogen. Thus, the diagnoses come later than preferred. In addition, the antibody is detected and not the active pathogen DNA meaning that the test cannot differentiate between active and prior infection. The challenge is that the Lyme pathogen is in low concentration in blood, so it is

hard to detect. One way to overcome this is to use SureSelect Agilent hybridization target capture and Next Generation Sequencing in order to create a more sensitive test which can amplify the DNA for detection without losing sensitivity.

Chapter 2: Methods

2.1 Extraction of the *Borrelia Burgdorferi* samples

I made sure to pipette 20ul of Proteinase K into an empty nuclease free tube for each of the five samples. The Proteinase K was used to eliminate contamination as well as prevent degradation of DNA during purification (Wiegers and Hilz, 1971). The Proteinase K was covered with aluminum in order to prevent light degradation. After vortexing each sample to prevent plasma from settling, I pipetted 100µL of the sample as well as 120 ul of PBS into the nuclease free tube. The tubes were vortexed for 10 seconds and 200 ul of buffer Al, a lysis buffer, was added to each tube and the tubes were taken to an incubator at 56 C for 30 minutes which created a dark brown color at the end. After the incubation, 200µL of molecular grade ethanol was pipetted into each of the tubes and vortexed. Using a mini spin column, I pipetted the sample from the nuclease free tube into the top of the min spin column. After, I centrifuged at 6000x g for one minute and when the centrifuge finished, the collection tube was discarded, and the samples were placed in a new collection tube. The same centrifuge process was repeated but 500ul of AW1 was added first before. After the wash, the same centrifuging process was repeated with 500 µL of AW2 added first this time but the centrifuging was now for 3 minutes at 20,000x g. After throwing out the collection tubes, the columns were placed in nuclease free tubes. The wash steps removed contamination from proteins and other contaminants while retaining the DNA bound to the silica membrane column (Pikor et al., 2011). The tubes sat for at least one minute before centrifuging for one minute at 6000x g. Each tube was then ready for the NanoDrop in order to check the concentration of each DNA sample.

2.2 Nano Dropping the DNA samples

Each of the samples were Nano Dropped in order to determine the DNA concentration before PCR was performed. The One Microvolume UV-Vis Spectrophotometer NanoDrop from Thermofisher was used, with the DNA program on the NanoDrop being utilized. A 2µL pipette was used and a water blank was loaded onto the block. This blank reading was run in order to create a baseline. Then, a buffer AE was loaded onto the block and run. This also created a blank reading that was used as a baseline. Finally with the sample, 2µL of the template was loaded onto the block and run. Both ng/ul, 230/260, and 260/280 readings show up on the screen.

2.3 B. *burgdorferi* Tokarz OspA qPCR Culture Confirmation

The qPCR primers and probes are given in the Tokarz 2017 paper. The forward primer is 5'-CCTTCACGTACTCCAGATCCATTG-3', reverse primer is 5'-AAC AGACGGCAAGTA GATC-3', and the OspA probe is 5'-CAACAGTAGCACCGATTGCGAC-3'. The probes are IDT Primetime FAM which is a fluorescently labeled probe used in the PCR. The optimal primer concentration tested by the lab is 300 nm and the probe concentration is 200 nm.

The cycler for the qPCR is the Roche LightCycler 480z which makes it possible not to use the ROX reference dye. Usually, the reference dye is inert and allows for normalization of the samples with a constant fluorescent signal (Agilent 2012 PCR). In a nuclease free tube, I pipetted 10µL of primer and probe as well as 90µL of nuclease free water to make a 100 µL aliquot. Since the probe is light-sensitive and degrades if exposed to too much light, the tube was covered with aluminum foil. When creating the reaction

mix, the Primetime Mastermix, probe, and primers are pipetted and vortexed in a 1.5 mL Eppendorf nuclease-free tube. Using a PCR plate, for each well, 45μL of the reaction mixture was pipetted as well as 5μL of template DNA from each of the samples. For each well, the mixture was pipetted thoroughly up and down in order to ensure mixing. When sealing the plate, a clear optic seal was used as it was essential to make sure that there was a tight seal to prevent contact at the top of the seal as that may interfere with machine imaging.

On the Roche LightCycler 480z, the cycling conditions that have been tried and tested are 2 minutes at 50C. This time is followed by 10 minutes at 95C in order to activate the polymerase, and 40 cycles at 95C for 15s and finally 60C for 1 minute.

2.4 B. *burgdorferi* Sybr Green qPCR Culture Confirmation

The use of the Sybr Green 1 Dye allows the qPCR not to need a probe. The Mastermix was created with a 2X qPCR mix, containing the Sybr green, with a final concentration of 1x. The forward primer with a final concentration of 0.45 μM and the reverse primer with the same final concentration as the forward primer was added to the Mastermix with PCR grade water added as well. For instance, for a 20μL reaction, the Mastermix would be 10 μL 2X qPCR mix, 0.9 μL forward primer, 0.9 μL reverse primer, and 3.2 PCR grade water. In each of the wells, in the PCR plate, 45 μL of the Mastermix is pipetted in as well as 5 μL of the template DNA. For each well, the mixture was pipetted thoroughly up and down in order to ensure mixing. When sealing the plate, a clear optic seal was used, and it was essential to make sure that there was a tight seal and to prevent contact at the top of the seal as that may interfere with machine imaging.

The cycling conditions for the initial denaturation was 94 C for 2 minutes. For forty cycles after the initial denaturation, the denaturation was 94 C for 15 seconds and the annealing, extension, and read fluorescence was 60 C for one minute (Bustin, 2002).

2.5 Qubit to determine the concentration of DNA

The Qubit was another measure of the concentration of DNA which is more accurate than the NanoDrop (Masago et al., 2021). The Qubit needed two assay tubes for the standards and five assay tubes with each tube designated to each of the *Borrelia Burgdorferi* samples. To create the Qubit working solution, I diluted the Qubit reagent 1:200 in the Qubit buffer. Creating the final tubes which are inserted into the Qubit machine, 10 µL of standard one went into the first tube and 190 µL of the working solution was pipetted into the tube. The process was imitated for the second standard from the kit as well as for each of the samples although, for the samples, the 10 µL standard from the kit was 10 µL from the designated sample.

2.6 Next Generation Sequencing with Borrelia *Burgdorferi* Samples

2.6.1 Create Nextera XT DNA Libraries

The first step was to prepare the libraries. The transposomes attach to the double stranded DNA fragments, with a universal overhang, at random sections and create a double-stranded DNA break. Then, there is ligation of the tagged sequences which are partial adaptor sequences that are recognized by the Nextera XT PCR primers. The Nextera XT DNA kit is a kit which allows the examination of small genomes such as the

Borrelia Burgdorferi bacteria and prepares the libraries to be sequencing ready. The tagmentation uses a bead-linked transposome, where the transposomes covering the bead contain Illumina sequencing adaptors. The unfragmented DNA binds to the bead-linked transposome and its saturation allows for DNA normalization. The tagmentation helps fragmentation, normalization, and ligation occur in the adaptors (Illumina Library Preparation).

To a 96-well PCR plate, each of the five wells contained 10 L of Tagment DNA Buffer, 5 µL of DNA from the sample, and 5 µL of Amplicon Tagment Mix. The plate was then sealed and centrifuged at 280x g at 20C for 1 minute. I placed the plate on the thermocycler and used the TAG program from Illumina, stopping at the 10 C step and waiting for no time to start the next step because the transposome is still active. 5 µL of Neutralize Tagment Buffer was added to each well and the sealed plate was centrifuged at 280x g for 1 minute and incubated at room temperature for 5 minutes.

Amplifying the libraries requires the two index primers and twelve cycles of PCR. The i5 and i7 adapters are needed in order to complete the partial adapter sequences from the tagmentation. I added 5 µL of the i7 adapter and 5 µL of the i5 adapter to each well as well as 15 µL of Nextera PCR Master Mix to each well. The sealed plate cycled through 12 times with the precycling conditions of preheat option of 100 C, reaction 50 uL, 72 C for 3 minutes and 95 C for 30 seconds. The 12 cycles were 95 C for 10 seconds, 55 C for 30 seconds, and 72 C for 30 seconds. The post cycling conditions were 72 C for 5 minutes and held at 10 C.

The cleanup of the libraries uses purification beads to eliminate fragments which are less than 200 base pairs so that any residual adapters are removed as well as primer dimers were prevented. The double-sided bead purification selects for the most available

insert size also removing fragments which are too large. The salt and polyethylene glycol (PEG) in the solution forces a certain size of DNA to attach to the beads because of the negatively charged phosphate backbone of DNA. The lower percent of salts and PEG transfers larger sizes of the DNA to the beads. The beads are magnetic, and on the magnetic stand, the beads move to the side of the tube with the magnet. This allows removal of the rest of the supernatant and dissolves the DNA off the beads (Bronner et al., 2009).

After the amplification, the plate was centrifuged at 280x g at 20 C for 1 minute in order to collect the contents at the bottom of the well and 50 µL of the supernatant of each well was transferred to a midi plate. I pipetted 30 µL of the purification beads into each well, sealed the plate and used the plate shaker at 1800 rpm for 2 minutes. After incubating for 5 minutes, placing the plate on the magnetic stand and waiting for the liquid to become clear, I removed the supernatant without disturbing the beads. The wash steps were repeated twice and consisted of adding 200 uL of 80% EtOH, incubating for 30 seconds, and discarding the supernatant. Finishing the wash steps, I added 52.5 µL of resuspension buffer to allow the DNA to dissolve off the beads. The final steps consisted of using the plate shaker at 1800 rpm for two minutes, incubating at room temperature for 2 minutes, waiting on the magnetic stand for the liquid to turn clear, and transferring 50 µL of the supernatant to a 96-well plate. After cleaning up the libraries, it was important to check the library quality using a Bioanalyzer to confirm the library size was in between 250 and 1000 base pairs. I did not need to normalize the libraries because the MiniSeq sequencing system uses onboard denaturation.

Using the values from the Qubit for each of the samples, the libraries were diluted to the starting concentration. This made sure that there will be an even read distribution

for all the samples. The average library used was 600 base pairs and in conjunction with the Qubit values, the molarity and RSB volume were calculated (Illumina Custom Protocol).

$$Molarity\ (nM) = \frac{\frac{ng}{\mu l} * 10^6}{\frac{660g}{mol} * average\ library\ size\ (bp)}$$

2.6.2 Denaturing and Diluting Libraries for Sequencing

Creating a multiplexed library pool is cost efficient and each of the samples still has a unique barcode which the MiniSeq can identify (Pomraning et al., 2012). After diluting each of the samples to the starting concentration, I added 10 µl of each diluted library to a tube to create a multiplexed library pool. In order to allow the pooled sample to run through the MiniSeq machine successfully, I had to dilute the library to the loading concentration of 1.4 pM. I combined 2 uL of the library pool with prechilled hybridization buffer in a microcentrifuge tube, centrifuged at 280x g for one minute, transferred 250 uL of the diluted library to another microcentrifuge tube with 250 uL of prechilled hybridization buffer, and centrifuged at 280x g for one minute. To denature the diluted libraries, I placed the tube on the incubator for 2 minutes at the preheated condition. Once heated, I immediately cooled the pool on ice and left it on ice for 5 minutes (Illumina Custom Protocol).

2.6.3 Performing the Run

The reagent cartridge was inserted in the machine and is also where the libraries are inserted. The reagent cartridge thawed for 35 minutes in a 37 C water bath, and I made sure there were no ice crystals in the reservoirs. While waiting for the cartridge, I also thawed the flow cell at room temperature for 30 minutes. I loaded 500 µL of the library into the reservoir and set Local Run Manager on the MiniSeq. After loading both the reagent cartridge and the flow cell, I confirmed the run parameters and started the run.

2.7 Analyzing Target Capture Data using Snakemake

The gathered data from target capture was analyzed using a bioinformatics pipeline, managed by the Snakemake tool. Snakemake is used for automating data analysis workflows and is a simple rule-based scripting language derived from Python. Similarly to GNU make, a file called the Snakefile is used to define the pipeline. It consists of multiple rules, each of which takes in input files and returns output files. Each rule can run shell commands, take advantage of Anaconda packages, and much more to transform the input file to the output file. This tool is intended to provide easily reproducible and scalable analyses, making it ideal for this task. We can take advantage of its support for variables and arrays to easily run our sequences through the pipeline with just a single shell command once the Snakefile has been set up. All subsequent development and analysis were completed on a 2017 Macbook Pro with a 2.3 GHz dual-core Intel i5 processor and 8 GB of RAM.

For the purposes of this project, Snakemake was primarily used to process FASTA files, a text-based file format that represents nucleotide sequences. Relevant sequences

were downloaded from NCBI GenBank and compared against sequences produced by hybrid capture. The first rule in the Snakefile converts the given .fasta files into .fastq files, which is the required format for the rest of the pipeline. This file format is in essence an extension of the fastq format, providing quality scores for each nucleotide in each sequence. This task is accomplished using shell commands and SRAToolkit, a software package provided by the NIH to download, manipulate, and analyze new and existing runs. It provides a binary called fastq-dump that provides the required functionality of transforming various file formats to fastq. The next rule in the pipeline uses BWA, a software package that performs alignment of sequences to a target reference genome. Specifically, this rule uses the mem algorithm provided by BWA. The output provided by bwa creates SAM files, which are again text-based. However, we wish to convert these files to the more efficient binary BAM format, which is more suited to the rest of the tools that we will use. This is achieved by piping the output from BWA (again achieved with shell commands) into the samtools view program. Samtools is another package of binaries that allow for interfacing with sequencing data, which we will use for the rest of the pipeline. The next rule in the pipeline uses the samtools sort program on the generated .bam files to sort the aligned sequences, preparing them for depth analysis. The sorted .bam files are then fed into the samtools depth program, which computes the read depth at each position in the given sequence.

To perform analysis and visualization of the read depths by sequence, the output of samtools depth is written to simple .txt files for easy manipulation. All subsequent work was done in Python, using numpy for mathematical operations and matplotlib for visualization. Per-plasmid graphs were generated using bokeh, a library which produces interactive graphs in .html files. As multiple types of analyses and graphs were generated, I developed reusable functions to process/read the .txt files generated by samtools depth

files into Python dictionaries for further processing. This code reuse across plotting/analysis scripts provided for cleaner code and a much simpler way of processing multiple runs at once (Köster and Rahmann, 2012).

2.8 Creating a Phylogenetic Tree from NGS Data

2.8.1 Chromosome Tree

After the NGS data was collected, I created a phylogenetic tree using a sequence of bioinformatics programs. All processing for this analysis was completed on the Harvard FAS Research Cluster. The specific machine allocated by the cluster for this analysis was a 64-core AMD Opteron server board with 1 TB of memory. This powerful machine was ideal for this task, as all programs run to produce the tree took advantage of parallelism in their design, heavily speeding up tasks.

The first step in producing a phylogenetic tree was to produce annotated sequences, and label them with useful information. The tool used to complete this task was Prokka, open source software that takes in .fasta files and outputs .gff annotated files. It was installed from Anaconda's bioconda channel. Prokka uses several search algorithms in succession to compare given sequences against large reference databases to find matches and annotate them. In order to speed up this computationally intensive task, the genus was provided to Prokka to narrow down the search of reference databases. Prokka was run individually on each of the five .fasta files produced by NGS, producing five .gff files for the next stage of analysis (Seemann, 2014).

Roary is a tool to generate a pan-genome analysis, taking as input annotated assemblies in .gff format. It was also installed from Anaconda's bioconda channel. Roary is quite computationally intensive compared to previous tasks, taking potentially hours to run on a standard desktop machine. In this regard, the additional computational power provided by the research cluster aided in performing this analysis in a timely manner. Roary provides a wide range of outputs for various types of analysis, but it specifically creates a core genome alignment that is used for creating the tree (Page et al., 2015). The final step in creating the tree is using FastTree, which creates approximately-maximum-likelihood trees from genome alignments. The easiest way to install FastTree was downloading and compiling from source. In order to achieve maximal performance, compiling with OpenMP support was required for multi-threaded execution. FastTree produces a phylogenetic tree in Newick format, easily visualizable either online or with a wide variety of downloadable software (Price et al., 2009).

2.8.2 Plasmid Tree

After the NGS data was collected, I was able to add the chromosome reads and plasmid reads from the five samples into the Geneious software which streamlines sequencing analysis. After the samples in Geneious were split into separate folders, the plasmid of interest, cp26 for instance, was taken from the B31 reference genome fasta file which was uploaded into Geneious. The plasmid cp26 was inserted into each of the sample folders and Align/Assemble allowed for a consensus and alignment of the sample plasmid to the reference plasmid. Once all the consensus sequences were created, they were inserted into the same folder with the reference B31 plasmid, and a tree was created showing the relationships and distances between the samples and to the reference.

Chapter 3: Results

3.1 Extraction of cultured Borrelia samples

Using the samples obtained from the culturing of the five Borrelia Burgdorferi samples, I extracted the DNA from the samples in order to make sure the impurities in the samples were removed (Pikor et al., 2011). In addition, the extraction allows for high recovery of the DNA so that it can be used as an input for PCR in the next step. I first extracted whole-blood samples and then applied the same technique to the five cultured samples. Quantifying the DNA after a whole blood extraction involved adding ethanol to buffers AW1 and AW2, as well as improving the pipetting methods from the Challenges section (4.1). This yielded a high amount of DNA and lowered contamination. The NanoDrop from a whole-blood sample provided the values 12.2 ng/uL, 2.17 A260/A280, and 1.40 A260/A230. The DNA concentration is high enough to serve as input into the PCR, meaning there will be amplification. In addition, looking at the curve provided on the NanoDrop, there was only one peak at 260 which indicates that there are nucleic acids present. The lack of double peaks indicates that the sample is pure. On the other hand, the presence of double peaks in the previous NanoDrops with other whole blood samples before changes after Challenges (4.1) were made, indicates a lack of purity (Desjardins and Conklin, 2010). The Qubit is more accurate in terms of the DNA concentration so before the input into PCR, both the A260/280 as well as the A260/A230 values are crucial and serves as a large reason for examining the NanoDrop for indications of purity rather than purely the concentration. The A260/280 value indicates the amount of protein in relation to the nucleic acids since proteins absorb at 280 nm. Thus, a higher 260/280 value is optimal and this number for DNA should at least be around 1.80, with the value 2.17 from the whole blood sample being well above. The A260/230 value measures other

contaminants such as EDTA and should be around 2.0. The value is 1.40 is lower than the optimal value and could be improved with more washes (Pikor et al., 2011).

The five samples used were taken from the B31 *Borrelia Burgdorferi* strain. The skin biopsy samples were collected from U.S. field sites in the summer of 2021. These samples come from the active enrollment phase at MGH. The skin biopsy samples received at MGH were cultured for *Borrelia* using an 8-week incubation protocol. Each of the samples was classified as culture-positive or culture-negative for spirochetes using darkfield microscopy. The confirmation of *B. burgdorferi* in spirochete-positive samples will be performed using PCR after extraction and quantification.

The five cultured samples were extracted and these samples with the improved pipetting methods and addition of ethanol to the buffers had the following results from the NanoDrop:

A

Nanodrop Values for Five Cultured and Extracted Samples

Sample	10071001	10071002	10071003	1008001	10081004
ng/uL	6.3	2.5	2.5	7.0	3.2
A260/A280	1.97	1.82	2.46	1.98	1.77
A260/A230	1.38	1.37	1.81	0.55	2.18

Figure 1. NanoDrop values after extraction with DNA concentration and contamination A.) The first row represents the DNA concentration. The A260/280 values represent amount of protein in relation to the nucleic acids since proteins absorb at 280 nm. The 260/230 value measures other contaminants such as EDTA and salts.

The NanoDrop values of the cultured and extracted *Borrelia*, all have ng/uL values which allow comparison to determine if the samples should proceed to input in the PCR. The small amount of *Borrelia* in the blood creates a difficulty with extracting. The values all are above 2 ng/uL which is lower than the optimal, at least 10 ng/uL but is expected for *Borrelia Burgdorferi* because of its low spirochete amount in blood. In addition, all the A260/280 values for the five samples are either around 1.80 or above with sample 10071003 having a high value of 2.46. The high 260/280 value indicates a low amount of contaminating protein. The 260/230 values should be optimally around 2.0. Only sample 10081004 complies with a value of greater than 2.0. The samples 10071001, 10071002, and 10071003 all have the 260/280 values above 1.20 and could be higher with the addition of wash steps during extraction. The tradeoff here would be that the wash steps would decrease the DNA concentration, which I tried to avoid because *Borrelia* DNA already has a low concentration in blood, and the main purpose of this thesis is to amplify the low amount of DNA in the blood. The sample 10081001 has a very low 260/230 value of 0.55, but this issue can be circumvented with a good primer design in PCR. The five samples are high quality samples to be used as input into the next step of PCR to verify positive *B. burgdorferi.*

3.2 PCR and Qubit of cultured Borrelia samples

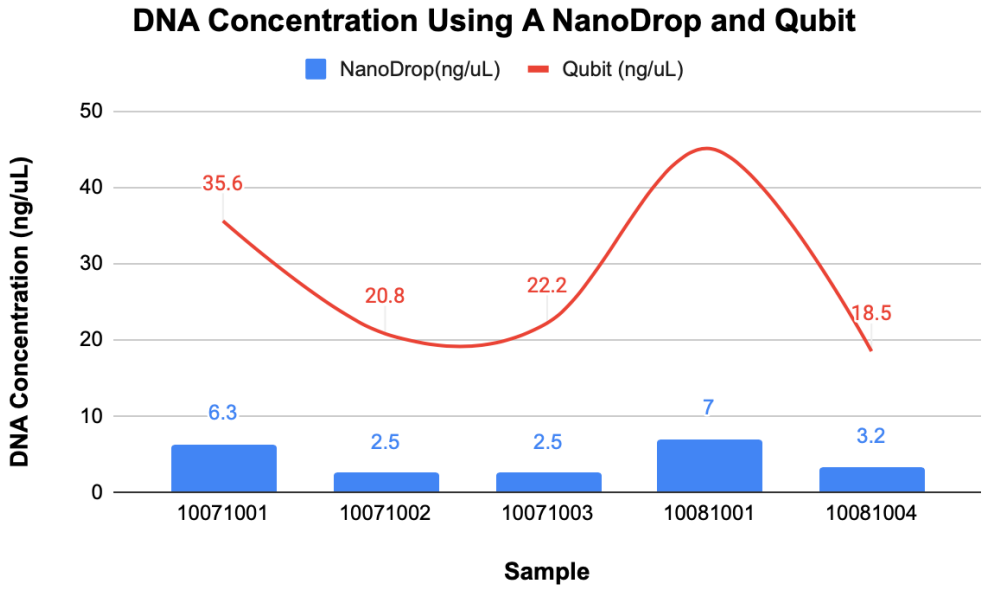
Given the *B. burgdorferi* classification of the previous five cultured samples, the samples can be used as input to PCR. I used the Tokarz assay, primers, and probes in the qPCR for a previously identified positive whole blood sample to ensure that the assay, primers, and probes worked. Originally, the assay did not produce amplification and Challenges 4.2 identified the probe as the issue. In the meantime, the Sybr green dye was used as it does not require a probe.

Looking at the standard curve for the Sybr green dye assay, the slope of the amplification measures the efficiency. The slope also measures how accurate and reproducible the results are. A slope closest to -3.32 indicates the efficiency. All the standard curve slopes for the five samples have a value (-3.247, -3.744, -3.421, -3.557, -3.512) within about 10% of -3.32. Thus, the amplification was efficient and accurate. When examining the amplification plots, all the samples have an original baseline which transitions into an exponential region and finally a plateau signaling a reduction in amplification. In addition, the dilutions are equally spread out and do not overlap which points towards the lack of a primer dimer (Jaton and Greub, 2007). Examining the melt curve, there is one peak shown indicating one amplicon.

When looking at the amplification curves after performing a PCR on each of the samples, we see that there is amplification for each of the dilutions. We dilute 10 times, and each dilution is 10^-1 times the previous sample. Thus, we can characterize each of the cultured samples as positive.

Knowing the samples are characterized as positive means the next step is sequencing. Right before sequencing, it is important to determine the most accurate DNA concentrations of the five samples. Since the NanoDrop is not very accurate when looking at DNA concentrations, the Qubit is used and the values 35.6, 20.8, 22.2, 45.1, and 18.5 ng/uL for the samples respectively were quantified. When comparing the NanoDrop to the Qubit and inverse Ct value, the same trend among the samples exists as seen in **Figure 2**. The Qubit concentration values are higher than the NanoDrop values and the Qubit values are used to dilute the sequencing values to the starting concentration before library pooling.

A



B

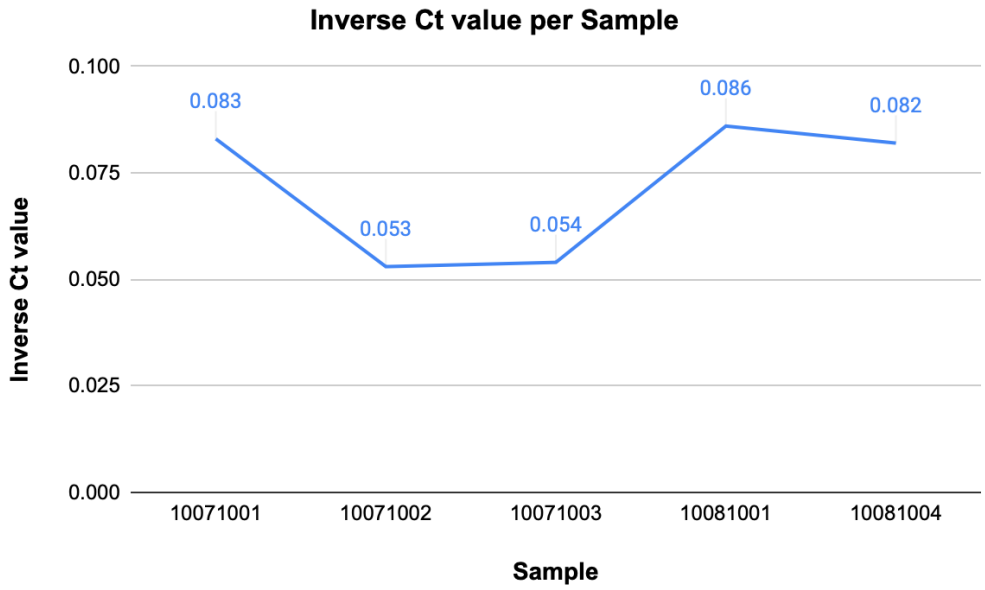


Figure 2. DNA concentration trend comparison between three quantification methods including NanoDrop, Qubit, and qPCR. A.) NanoDrop was used after extraction to quantify the DNA concentration of the *Borrelia burgdorferi.* The Qubit, which is a more accurate measure, is shown in conjunction with the NanoDrop. The trend between the samples are similar regardless of the differences in concentration **B.)** The Ct value output from the qPCR shows a similar trend in comparison to plot A. The inverse Ct is shown because the Ct is inversely proportional to the DNA concentration.

3.3 Sequencing of the five Borrelia Samples

3.3.1 Sequencing Challenges leading to Kapa Analysis

The original sequencing did not show amplification and in order to check if the library existed in the sample after preparation, a Kapa PCR was performed. To prevent the libraries from not clustering, the Kapa output concentrations were taken and recalculated when diluting the libraries to the starting concentration. It was important not to keep the libraries highly concentrated or the sequencing run would not complete. The first Kapa PCR showed no amplification, but with the replacement of the reagents, the second Kapa PCR had a standard curve slope of -3.547 which is close to the 100 percent efficiency of -3.32. Hence, the results are efficient and there is an exponential curve in the amplification plot indicating that the library remains intact.

3.3.2 The Quality of the Sequencing Run

The quality of the run is used to determine if the output from sequencing is reliable enough to use as a comparison in Hybrid Capture. The probability of selecting the right basecall, an overall quality value of sequencing, is 95%. The probability decreases with increased cluster density of the flow cell. There is a decrease in confidence in the calls because clusters which are close together can obscure the signal. Another reason is that over time, the molecules within a cluster are phasing, where one of the cycles is less efficient and is behind or ahead one cycle. In a long read, the difference in cycles accumulates and the molecules in the cluster are out of phase.

The cluster density generally remains in the range of 170 - 220 K/mm2 but our value of 90 K/mm2 was much lower. We under clustered and erred on the side of caution

when loading the samples by overestimating their concentrations, since over clustering would crash the run resulting in a lack of data. The conservative nature of the clustering means there is less data than expected but not enough to take away from the quality of the output.

The barcode.txt file contains all the barcodes the Mini-Seq identified on the i5 and i7 independent of the barcode I instructed for the machine to look for. The barcodes with reads in the millions do correspond with the i5 and i7 for each sample which I input. There were millions of reads per sample which points towards a high amount of data, and the reads amount was similar for each sample.

3.3.3 Analyzing Phylogenetic Trees

The coverages for the five samples across all the plasmids point towards the plasmid lp54 as the plasmid with the highest coverage. Creating a tree to understand the relationship between lp54 in the samples, the plasmid lp54 on the sample 10071003 is most aligned with the plasmid ip54 on the reference B31 chromosome. The plasmid cp36 is known to contain the OspC gene which is one of the main B. burgdorferi antigens used for transmission to animals from ticks. The cp36 tree emphasizes that the plasmid cp36 on the sample 10071003 is most aligned with the plasmid cp36 on the reference B31 chromosome. On the other hand, plasmid cp36 for samples 1007001 and 1008004 are both the least aligned with cp36 on the reference B31 chromosome. In the overall chromosome, the sample 10081001 is most aligned with the B31 reference genome. The next step is to examine hybrid capture and interpret if I am capturing the genome from either clinical samples or low input samples.

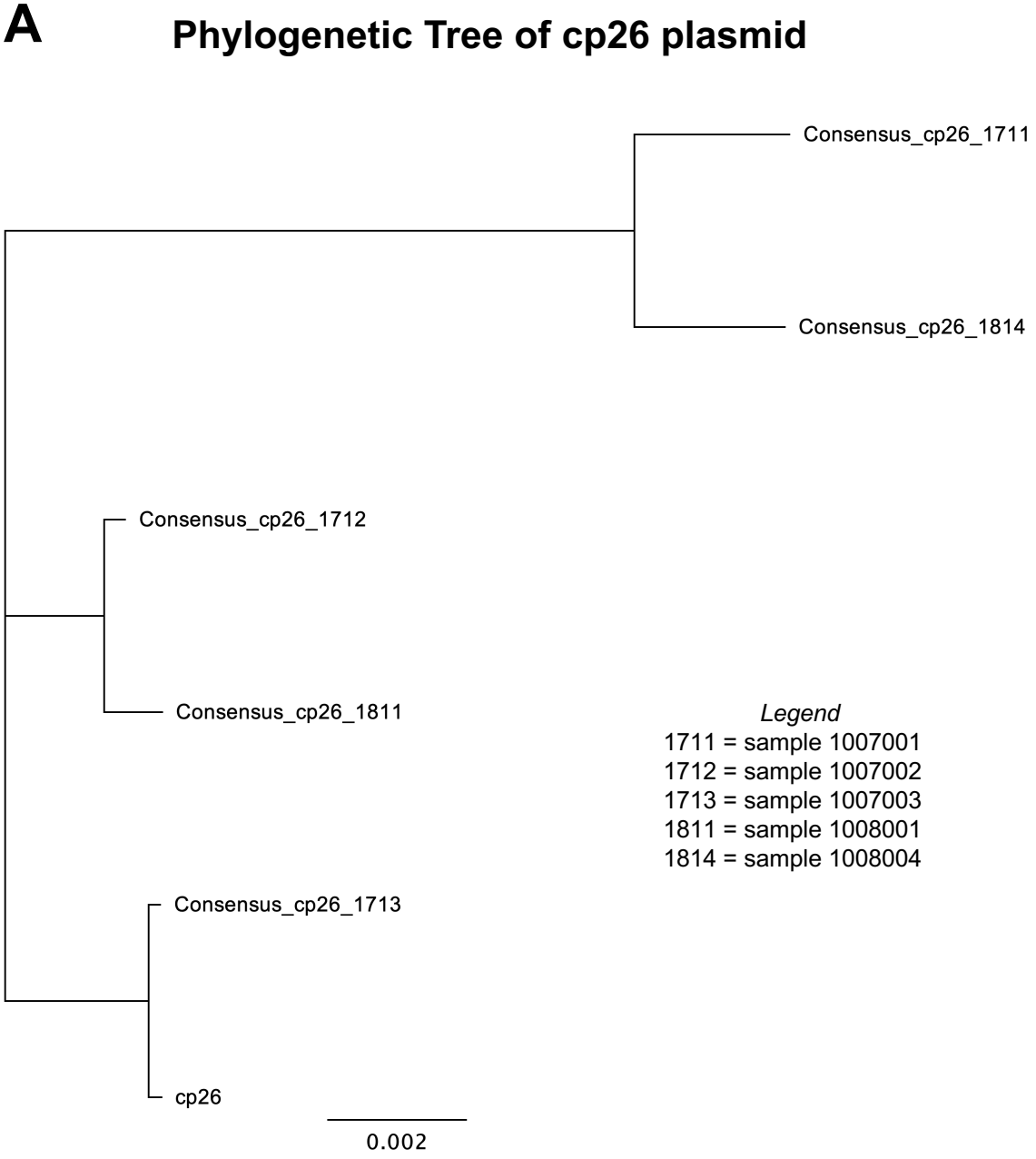


Figure 3A. Phylogenetic tree of plasmid cp26 and *Borrelia burgdorferi* samples. A.) Geneious was used to create a phylogenetic tree comparing the cp26 plasmid of all samples to the cp26 plasmid from the B31 Borrelia burgdorferi strain.

B **Phylogenetic Tree of lp54 plasmid**

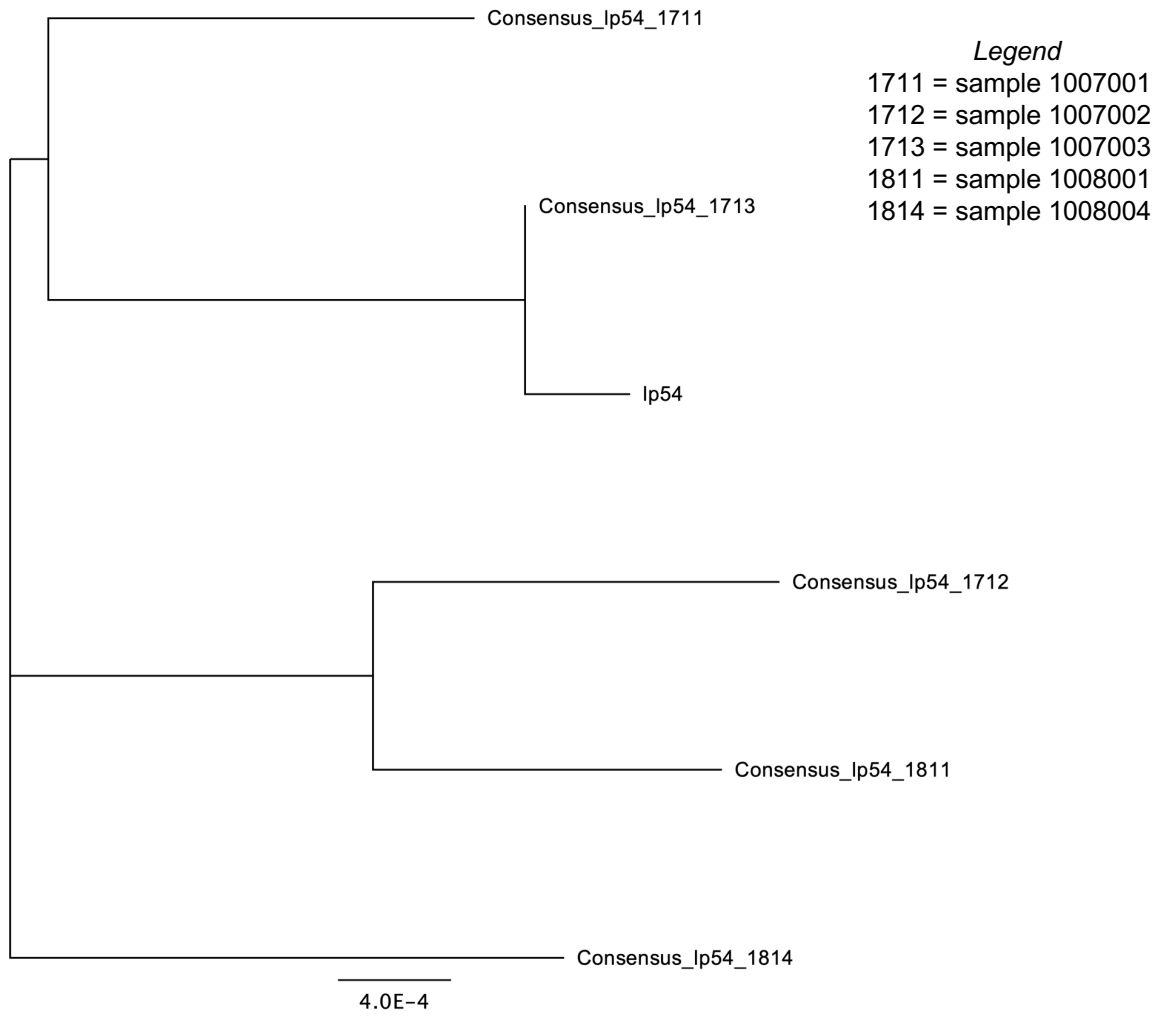
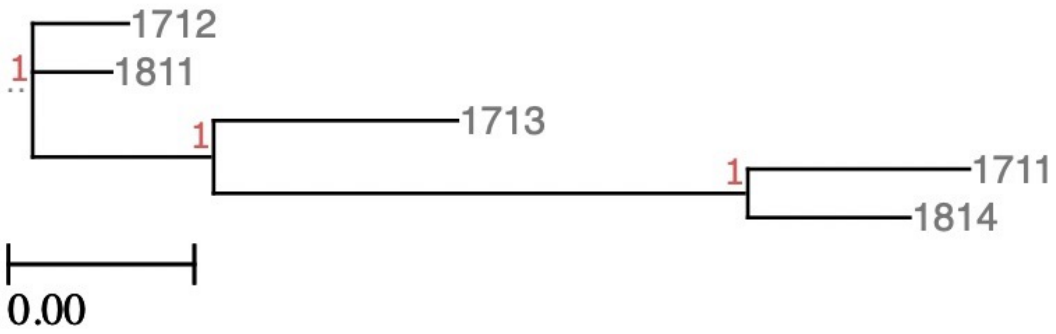


Figure 3B. Phylogenetic tree of plasmid lp54 and *Borrelia burgdorferi* samples. A.) Geneious was used to create a phylogenetic tree comparing the lp54 plasmid of all samples to the lp54 plasmid from the B31 Borrelia burgdorferi strain.

C **Phylogenetic Tree of Chromosome**



Legend
1711 = sample 1007001
1712 = sample 1007002
1713 = sample 1007003
1811 = sample 1008001
1814 = sample 1008004

Figure 3C. Phylogenetic tree of B31 *Borrelia burgdorferi* chromosome strain and chromosomes of *Borrelia burgdorferi* samples. A.) Geneious was used to create a phylogenetic tree comparing the chromosomes of all samples to the chromosome of the B31 Borrelia burgdorferi strain.

3.4 Analyzing Existing Target Capture Data

The goal was to look at three different papers and determine which could be the most effective method in terms of increasing coverage. The papers by Walter and Carpi had a link to the data used on the SRA. This data included both methods and the non-hybrid capture runs came from data in Broad from the Sabeti lab. The samples used from these samples were from uncultured blood which is different from the five cultured samples used in the sequencing experiments done at MGH.

The first thing to look at was the read depth coverage across the genome, comparing the reads from the runs with hybrid capture to the runs without hybrid capture. After being processed by the Snakemake pipeline as described in the Methods section, the runs were read into a Python script. For each type of run, ten different randomly selected runs were averaged together and compared. Using plotting libraries, interactive graphs were generated for each of the 22 plasmids found in the runs for various metrics. Additionally, a paired t-test was performed that showed statistical significance across every plasmid.

The first plot shown represents the genome position vs read depth for hybrid capture (red) and no hybrid capture (blue) of one of the 22 plasmids in the genome. The second plot represents the ratio between hybrid capture and no hybrid capture. Looking at the plots, the runs with hybrid capture had much higher read depth than the runs without hybrid capture throughout almost the entire genome. This means that the hybrid capture must affect genome coverage and increase it. Finding this is very helpful because it helps address the initial problem with PCR in diagnosing Lyme disease, that the test is not sensitive enough to pick up the parasite DNA. This can be solved when there is higher genome coverage, as can be done by hybrid capture methods.

A

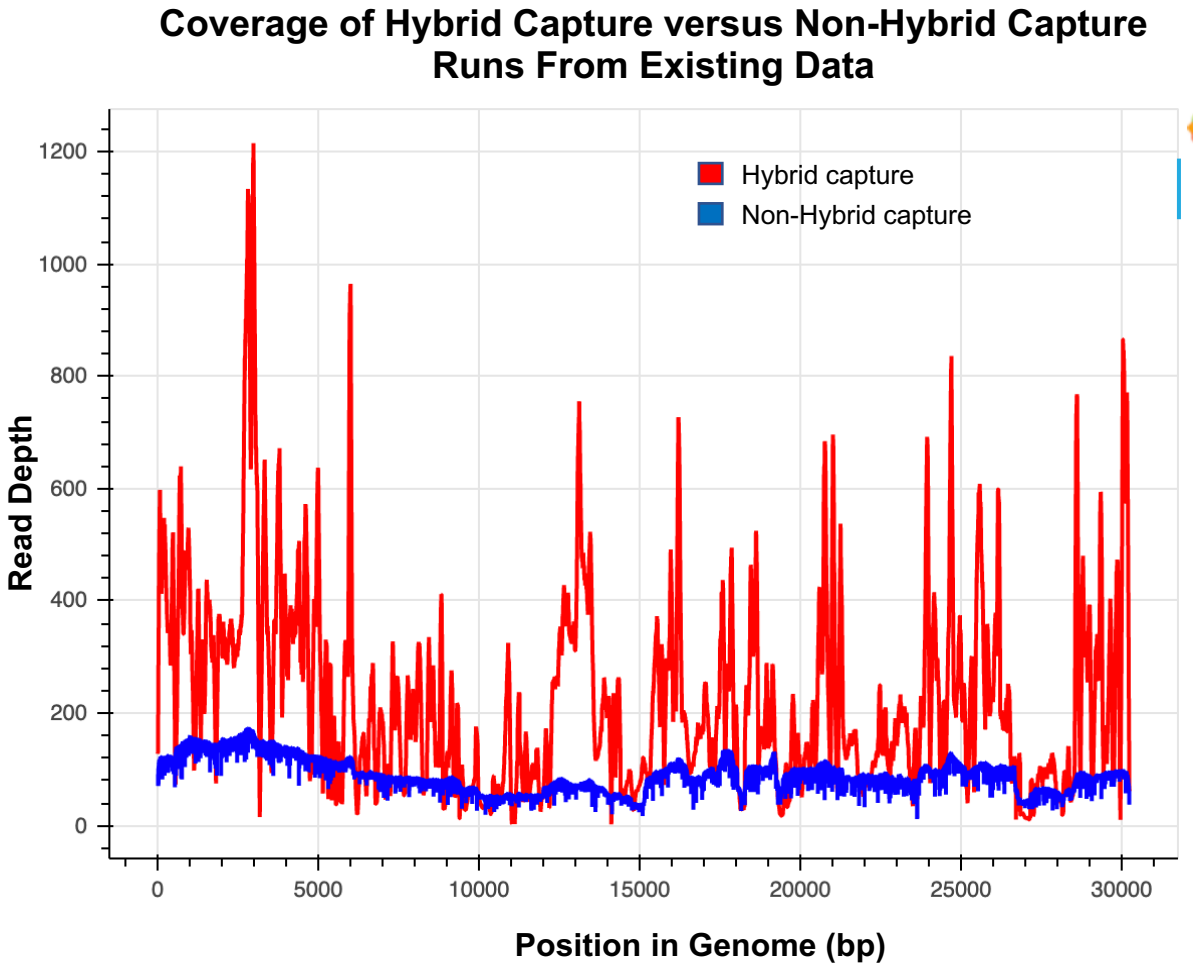


Figure 4A. Comparing coverage of Hybrid Capture runs to Non-Hybrid Capture runs from existing Sabeti Lab data A.) Using Snakemake, a Bokeh graph was created showing the genome coverage through read depths for hybrid capture runs versus non-hybrid capture runs. The red depicts the hybrid capture runs while the blue depicts the non-hybrid capture runs.

B

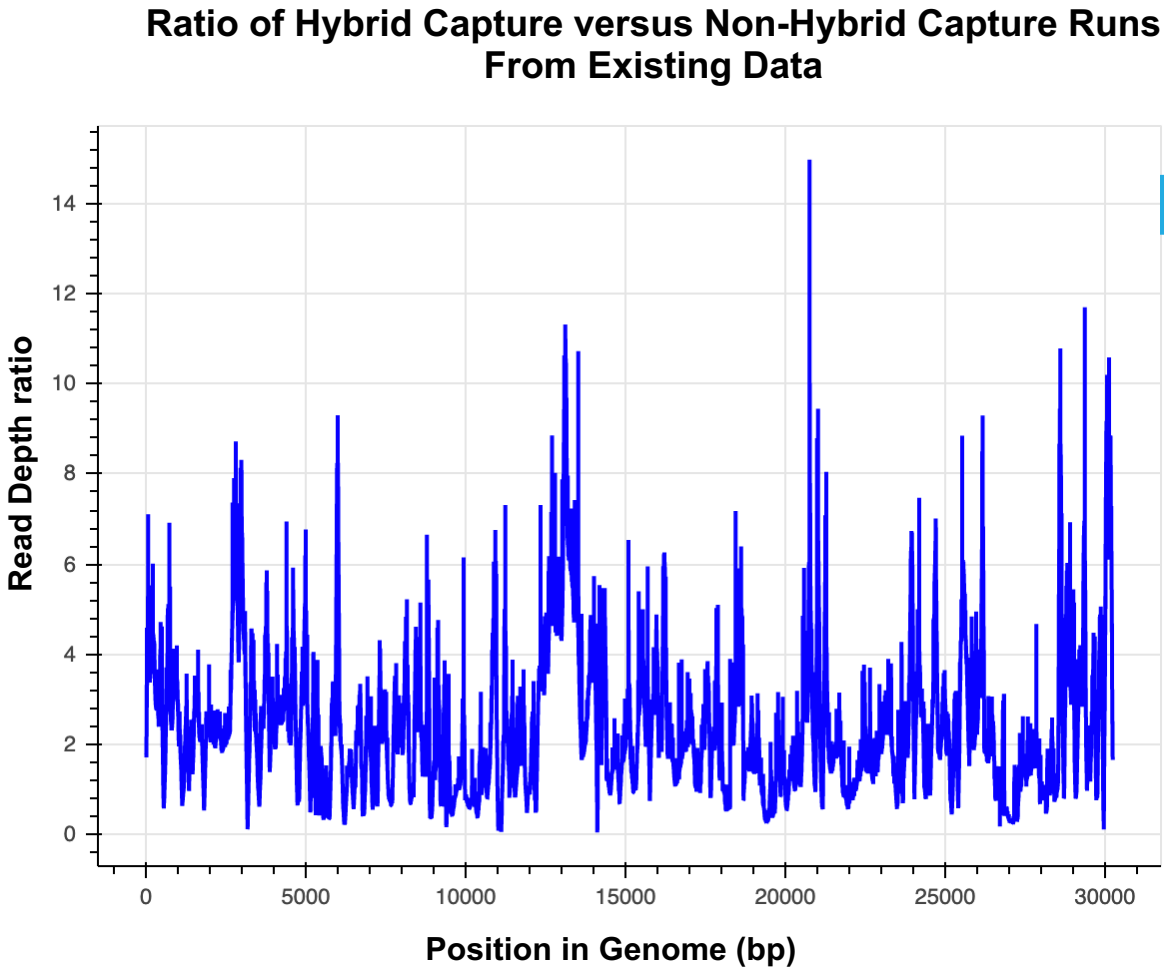


Figure 4B. Comparing coverage through ratio of Hybrid Capture runs to Non-Hybrid Capture runs from existing Sabeti Lab data A.) Using Snakemake, a Bokeh graph was created showing the genome coverage ratio through read depths for hybrid capture runs versus non-hybrid capture runs. The majority of the ratio is above 1.

Chapter 4: Challenges

4.1 Extraction of cultured Borrelia *Burgdorferi* samples

When initially extracting the DNA from a whole blood sample, and Nano Dropping, the DNA concentration was low as well as there was high contamination when examining the A260/280 and A260/230 numbers. In order to lower the level of contamination, when pipetting during the extraction step when the buffers and samples are transferred to the mini spin column, it is important not to touch the tip of the pipette to the rim of the mini-spin column. This prevents there from being residue at the edge of the rim that can cause contamination. When Nano Dropping again after caution with pipetting, there was still contamination although the contamination had decreased. The DNA concentration was still low as well so, I realized that when pipetting, the size of pipette used was the 20ul pipette for a 5ul amount. Switching this pipette to a 10ul helped with accuracy because the accuracy of the pipette is decreased when the amount pipetted was closer to the lowest value possible in the pipette. Thus, when Nano Dropping after matching the sample size of the pipette to the size of the pipette closest to the sample amount, the concentration of DNA increased. When extracting with Borrelia, the pipetting technique was especially important because of the low amount of DNA in the blood. Although through correcting pipetting techniques the contamination decreased slightly, there was still a relatively high amount of contamination. The addition of 100% molecular grade ethanol to both the AW1 and AW2 buffers, 160mL and 130mL respectively, in the wash steps of the extraction decreased the contamination significantly. When the AW1 and AW2 buffers are eluted through the mini-spin columns, the nonpolar state of the ethanol prevents the DNA from

eluting as well into the collection tubes. Thus, the ethanol ensures that the buffers do not interact with the DNA, so it remains in the top capture tube of the mini spin column and is able to be washed again (Pikor et al., 2011).

4.2 PCR of cultured Borrelia *Burgdorferi* samples

I found that there was no amplification of the DNA when initially using the Tokarz assay. One way I tried to identify the issue was through the process of elimination of the major components of the PCR which are the primers, probe, and Mastermix. When running the multiplex qPCR in the lab with the same Mastermix, but neither the same primers and probe, the qPCR had an amplified result. The Mastermix must be high quality. The second component to test would be the primers. Through a Sybr green dye 1 assay, which does not require a probe, there was amplification shown in the PCR results. The amplification pointed to the primers working in the original Tokarz ospA assay. Thus, the probe was either not an accurate sequence or the probe was too old of a reagent to use for accurate results from the qPCR. Running the Tokarz assay again with a new probe with the same sequence as used previously, there was amplification. The probe was hence old and in the process of degrading, but the Tokarz sequences were accurate. Regardless, I decided to use the Sybr assay and the Tokarz assay to double check that there was amplification.

4.3 Sequencing of cultured Borrelia *Burgdorferi* samples

The NanoDrop from quantifying the concentrations after extraction were neither as recent nor as accurate in DNA concentration as a Qubit (Masago et al., 2021). Thus,

when using the Qubit, the concentrations were 35.6 ng/uL, 18.5 ng/uL, 45.1 ng/uL, 22.2 ng/uL, 20.8 ng/uL. These values are much higher than the initial NanoDrop values which were all below 20 ng/uL. In addition, the values are between 20 - 50 ng/uL which is optimal for the MiniSeq. After running the MiniSeq manually without using the Local Run Manager, I needed to create a sample sheet and use Illumina's Basespace CLI to generate .fastq files from the run data. Basespace showed that the Nextera XT libraries did not cluster on the MiniSeq run. I did not quantify the libraries before diluting the libraries to the starting concentration using either a Bioanalyzer or a TapeStation. In order to verify that the Nextera libraries were complete, I could do a PCR reaction using i5 and i7 specific primers with the libraries as template DNA. If I observed a band on a gel (or smear), it meant that the i5 and i7 were on each side of the libraries. If there were not observable amplicons, this would mean that the libraries were not complete. The Kapa PCR is specific to viable library molecules and won't amplify anything that isn't viable, showing if there is a valid library to use. The Kapa PCR showed an amplification and therefore we were able to use the concentrations from the Kapa to dilute down the starting concentrations in the sequencing protocol in order to get clustering in the MiniSeq.

4.4 Analyzing Phylogenetic trees

Producing the phylogenetic trees using Prokka, Roary, and FastTree came with several challenges from an operational viewpoint. Initially, the target machine was the 2017 dual-core MacBook Pro used for the target capture data analysis. However, Prokka was exceptionally difficult to install and run properly on this machine. The software is heavily

dependent on BioPerl and trying to install that component on its own ran into several compiler and linker issues that were difficult to solve. Installing through Anaconda led to dependency issues that could only be fixed by editing the master Perl script that ran Prokka, an unsustainable situation for continued use. Attempts to instead run the software on a more 2021 M1 ARM Macbook Pro were stymied by a lack of Intel AVX hardware support. Attempting to install and run Roary was also unsuccessful due to issues with BioPerl installations not working properly. While FastTree worked out of the box after compilation, it needed the output from Roary.

These problems were all solved by using the Harvard FAS Research Cluster, which provided a powerful Linux machine as described in the methods section. Dependency installation was smooth using Anaconda, although Roary and Prokka could not be in the same conda environment due to conflicting dependencies- a feature not mentioned by documentation. In order to ease the process of creating trees from the source .fasta files, I created a shell script to run Prokka, Roary, and FastTree, switching between environments as needed.

Chapter 5: Discussion

5.1 Extraction and Quantification

Starting by extracting whole blood samples before the cultured five samples allows for refining the extraction method. The first few extractions with whole blood yielded a low amount of DNA and high level of contamination. After choosing the right pipette size per sample, pipetting away from the rim of the spin column, and adding ethanol to the extraction buffers to prevent DNA elution into the collection tubes, the DNA concentration increased and contamination from other proteins decreased. The one contamination value that did not decrease was the A260/230 value which is the contamination from other salts. The reason I did not concentrate on improving this value to the optimal 2.0 value is because additional washes could fix the problem, but the issue is the loss of DNA which is crucial to preserve for *Borrelia* as discussed above.

Having the extraction methodology improved and shown to yield results through the whole blood samples guaranteed the success of the cultured samples. The reason I did not want to experiment with the cultured samples originally was because the quantity of the cultured samples would not allow for many trials, costing time and money while whole-blood samples had many aliquots to experiment with. In addition, the ultimate goal of the project is to use whole-blood samples which have the low concentration of *Borrelia Burgdorferi* while the cultured samples are guaranteed to have concentrated spirochete DNA (Mitchell et al., 1993). The spirochetes are in the skin and tend not to be in the blood except in very low amounts. When taking a sample from the skin, there is a larger amount of *Borrelia Burgdorferi* material. Thus, when culturing from the skin biopsies, the culturing is conducive to the growth of *Borrelia.* Even though I

used cultured skin biopsies which would have more material than the whole blood samples, as well as the fact that culturing is not time-efficient to use as a clinical method, using the cultured samples and proving the extraction, PCR, sequencing, and target capture works will allow for the next step of implementing whole-blood samples.

Examining the NanoDrop values of the extracted cultured samples, the DNA concentration is high enough to proceed to PCR with the protein contamination also being low. The salts contamination is higher but that can be overlooked due to the PCR primer design (Jaton and Greub, 2007). Hence, the next step was to use PCR to identify the positive or negative concentration of the five samples.

5.2 PCR and Quantification

Although using the Tokarz assay for PCR did not originally work, the Sybr green dye assay did not need the use of a probe (which was the issue with the Tokarz assay). The Sybr green dye is the dye that binds to the double stranded product and helps detect amplification. In general, there is an advantage to using the Sybr green dye because the assay setup is reduced without the need for a probe, meaning less chance for error. The per-assay costs are also decreased. The most important issue is time, because earlier diagnosis is the best solution to Lyme disease diagnostic issues. Although Sybr green assays are time and cost efficient when compared to the Tokarz assay, the drawback to using an intercalating dye is the nonspecific reaction products. Hence, it is important to look at the melt curve analysis when using the Sybr green assay (Vitzthum et al., 1999). The melt curve analysis showed one peak, confirming that there was a specific product (also confirmed by the lack of double peak).

The output of the PCR gave Ct values which are inversely related to the DNA concentration in the samples. Both the NanoDrop values and the Qubit values are needed. The Qubit uses fluorometry and emits fluorescence when examining DNA binding while the NanoDrop uses spectrometry with two different wavelengths of light passing from the DNA to show absorbance. Even though the Qubit values are more accurate in concentration than the NanoDrop values and will be the values used in the sequencing starting concentration calculation, the NanoDrop values are crucial to determine the purity of the sample since other protein contamination can terminate the PCR reaction (Masago et al., 2021). **Figure 2** exemplifies how all three quantification methods, PCR, NanoDrop, and Qubit, show the same trend in samples despite the different values. The PCR Ct value trend matching with the Qubit trend confirms the use of the Qubit values in the sequencing calculations since there is no discrepancy.

5.3 Sequencing and Hybrid Capture

The first sequencing run of the samples produced no result after the run and showed that the Nextera XT libraries did not cluster. Using a Kapa PCR to reaffirm existence of the libraries, there was amplification in the second Kapa run. One reason for sequencing to fail is high concentrations. After completing the Kapa run with diluted concentrations, those same concentrations were used in the recalculation of the starting concentrations before pooling in the sequencing run. I combined the Kapa with a TapeStation to get the average fragment size, in order to get an exact quantity of the libraries by adjusting the quantity from the standard curve with the Kapa which assumes the fragments are 452bp. Once the new concentrations were used, there was good

quality data with under clustering when examining the MultiQC report. The under clustering is favorable over over clustering because there is still high-quality data available in the former.

Furthermore, the production of .fastq files containing the reads from all of the samples allowed the creation of Phylogenetic trees. These trees were made showing both the plasmids cp26 and lp54. Both plasmids showed the highest percentage coverage across the samples and the plasmids. The plasmid cp26 in particular is found on the surfaces of bacteria in the tick's salivary glands while feeding (Kumaran et al,. 2001). The OspC synthesizing during feeding is because the OspC protein is important to infection in the host, as a major antigen on the spirochete surface (Walter et al., 2017). Thus, focusing on this plasmid is important in its prevalence as a gene. It also attains high diversity because of the high level of recombination. Regardless, there are highly conserved regions of OspC including the C-terminus (Metsky et al., 2019). Similarly, lp54 had the highest coverage amongst all samples and is the largest plasmid (53 Kb). It is also known to be highly conserved as well. The plasmids show that there is high-quality NGS data which has to be compared to and used with the hybrid capture data (Dunn et al., 1990). The next step would be to perform the hybrid capture and see if more of the genome is captured from clinical samples or low input samples.

Before I implemented the target capture data, I found data from the Sabeti Lab to show that there was a difference between the hybrid capture and non-hybrid capture runs. Figure **4A** and **4B** exemplify how there is an increase in coverage when using hybrid capture.

Chapter 6: Future Works

Although I was not able to complete the target capture step in the lab due to COVID-19 and reagent delays, I was able to plan out my reagents and my experimental plan. Target capture can be thought of as capturing the target sequences (“prey”), out of a “pool” of human DNA, using microbial sequences (“bait”). One way to start this process is to synthesize the oligonucleotides on a microarray. A microarray allows only complementary strands to attach, meaning we can isolate the DNA we are targeting from the pool. After a PCR step and a transcription step, the single stranded RNA hybrid product is used as the “bait” target sequence to fish out of the “pool.” This “pool” is prepared through PCR amplification, shearing and ligation of DNA. Capturing the hybrids by magnetic bead extraction allows subsequent analysis on a next-generation sequencing instrument and in this case the MiniSeq instrument I used for sequencing. It would be important to find samples with tick-pathogens of Lyme since the control in this experiment would be the Lyme samples which do not go through the hybrid capture protocol. For the Lyme samples that do go through the target capture protocol, the first step is to shear the DNA. The shearing should increase the amount of target sequence. The DNA shearing will be useful because enzymatic shearing causes blunt ends at each end of the fragment which are primer-binding sites used for amplification. Thus, shearing increases the number of binding sites (Poptsova et al., 2014). The next step will be to ligate and tag the DNA with adaptors. This adaptor will add onto the genomic DNA and allow it to be found by a primer. The adaptor- tagged library is then amplified using PCR (Gnirke et al., 2009). This is now the prepared pool of DNA. On the other end, there will be a target DNA chosen from a library available from Agilent

Technologies. The target DNA will go through in vitro transcription with biotin-UTP. The in vitro transcription and the biotin-UTP labeling will allow the target, or “bait”, to be more sensitive than if randomly primed. Now that we have both the “bait” and the “pool,” we can use hybridization with the microarray which will capture the complementary strands to the target sequence. The next thing is to finally use universal primers to PCR amplify the target sequence (Gnirke et al., 2009). We will use Agilent Technologies software to identify the optimal sequences to serve as whole genome bait. Similarly, we plan to use Agilent design tools which will help synthesize a custom target capture library to capture the complete *B. burgdorferi* sequence to capture the target from. Something we will take into consideration is to use RNA baits since they increase efficiency and stability over DNA: DNA hybrids (Gaudin & Desnues, 2018). The target capture method is efficient because hybridization is fast, and the result of this process is increased sensitivity and inclusion of many genetic variants (Gnirke et al. 2009).

For every DNA sample that is to be sequenced, one library is prepared. Library preparation creates short DNA fragments of about 150 to 200 base pairs. Using the SureSelect library prep kit, samples will be prepared. The samples will now be adaptor-tagged, and PCR amplified. There is now a prepared DNA library (Agilent 2011).

Once the library is prepared, the library will be hybridized using the Agilent SureSelect protocol. The prepared DNA hybrids will be captured on magnetic beads. In order to purify the DNA, the beads and gDNA samples are put into a plate which collects and washes the bead-bound DNA. The next step for the now captured libraries will be to use indexing primers to amplify the library using PCR. Once the library is amplified, the Bioanalyzer High Sensitivity DNA assay will be used to analyze the

captured DNA. Once we have pooled samples, we will then sequence the libraries on Illumina instruments and quantify the amount of *B. burgdorferi* DNA present before and after the target capture (Agilent 2011).

A

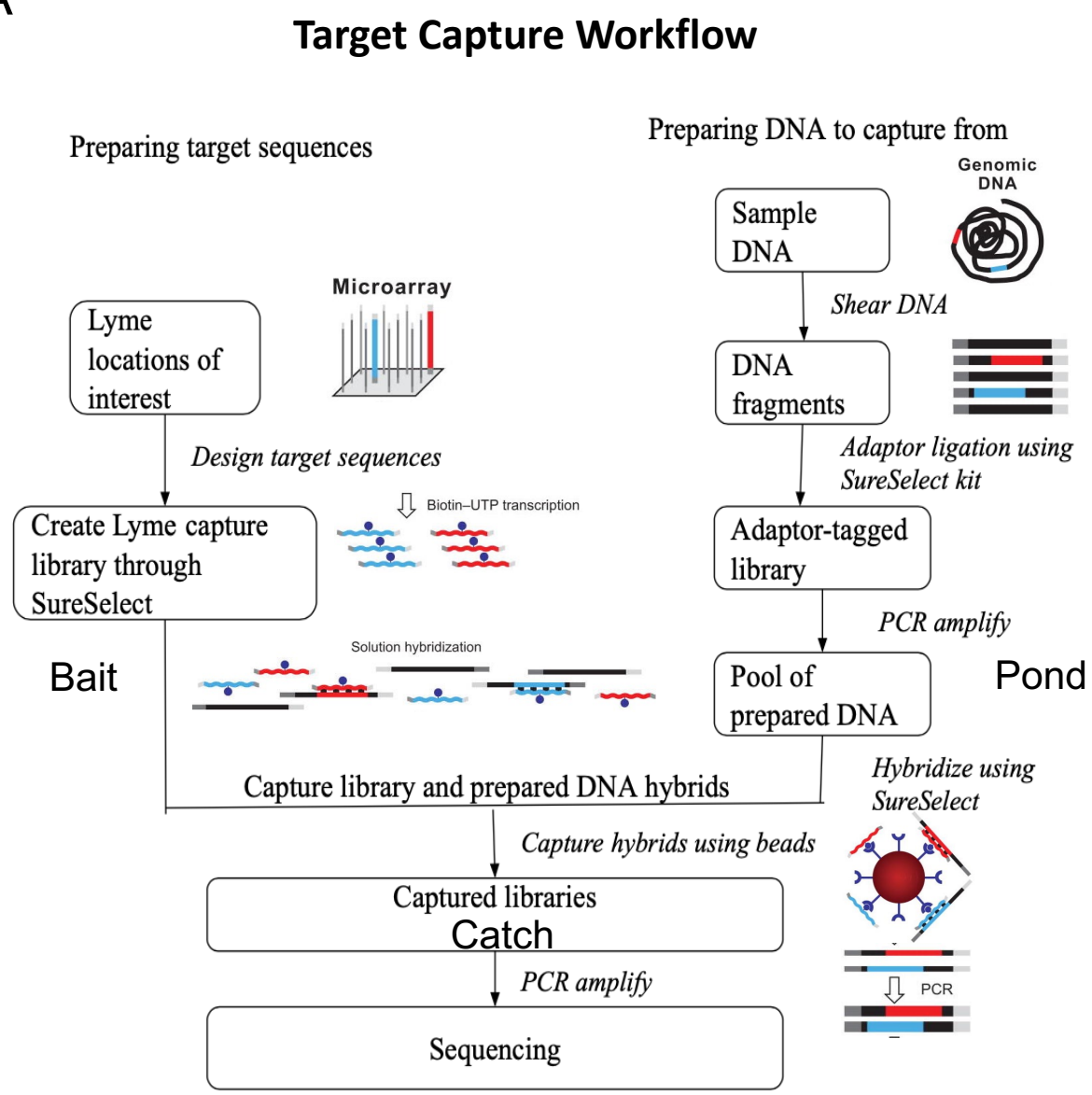


Figure 5. The Target Capture Workflow A.) Combining the Illumina SureSelect Protocol and the in-solution hybrid capture described and imaged by Gnirke et al., 2009. This workflow is to be implemented with the main three sections of "Bait," "Pond," and the "Catch."

References

Aucott, J., Morrison, C., Munoz, B., Rowe, P.C., Schwarzwalder, A., and West, S.K. (2009). Diagnostic challenges of early Lyme disease: Lessons from a community case series. BMC Infectious Diseases *9*, 79.

Barbour, A.G. (2016). Multiple and Diverse vsp and vlp Sequences in Borrelia miyamotoi, a Hard Tick-Borne Zoonotic Pathogen. PLoS One *11*, e0146283.

Becker, N.S., Rollins, R.E., Nosenko, K., Paulus, A., Martin, S., Krebs, S., Takano, A., Sato, K., Kovalev, S.Y., Kawabata, H., et al. (2020). High conservation combined with high plasticity: genomics and evolution of Borrelia bavariensis. BMC Genomics *21*, 702.

Bil-Lula, I., Matuszek, P., Pfeiffer, T., and Woźniak, M. (2015). Lyme Borreliosis--the Utility of Improved Real-Time PCR Assay in the Detection of Borrelia burgdorferi Infections. Adv Clin Exp Med *24*, 663–670.

de Bourcy, C.F.A., De Vlaminck, I., Kanbar, J.N., Wang, J., Gawad, C., and Quake, S.R. (2014). A Quantitative Comparison of Single-Cell Whole Genome Amplification Methods. PLoS One *9*, e105585.

Bronner, I.F., Quail, M.A., Turner, D.J., and Swerdlow, H. (2009). Improved Protocols for Illumina Sequencing. Curr Protoc Hum Genet *0 18*, 10.1002/0471142905.hg1802s62.

Bustin, S. (2002). Quantification of mRNA using real-time reverse transcription PCR (RT-PCR): trends and problems. Journal of Molecular Endocrinology *29*, 23–39.

Caboche, S., Audebert, C., and Hot, D. (2014). High-Throughput Sequencing, a Versatile Weapon to Support Genome-Based Diagnosis in Infectious Diseases: Applications to Clinical Bacteriology. Pathogens *3*, 258–279.

Calvo, S.E., Compton, A.G., Hershman, S.G., Lim, S.C., Lieber, D.S., Tucker, E.J., Laskowski, A., Garone, C., Liu, S., Jaffe, D.B., et al. (2012). Molecular diagnosis of infantile mitochondrial disease with targeted next-generation sequencing. Sci Transl Med *4*, 118ra10.

Desjardins, P., and Conklin, D. (2010). NanoDrop Microvolume Quantitation of Nucleic Acids. J Vis Exp 2565.

Dunn, J.J., Lade, B.N., and Barbour, A.G. (1990). Outer surface protein A (OspA) from the Lyme disease spirochete, Borrelia burgdorferi: high level expression and purification of a soluble recombinant form of OspA. Protein Expr Purif *1*, 159–168.

Earnhart, C.G., Rhodes, D.V.L., Smith, A.A., Yang, X., Tegels, B., Carlyon, J.A., Pal, U., and Marconi, R.T. (2014). Assessment of the potential contribution of the highly conserved C-terminal motif (C10) of Borrelia burgdorferi outer surface protein C (OspC) in transmission and infectivity. Pathog Dis *70*, 176–184.

García-García, G., Baux, D., Faugère, V., Moclyn, M., Koenig, M., Claustres, M., and Roux, A.-F. (2016). Assessment of the latest NGS enrichment capture methods in clinical context. Sci Rep *6*, 20948.

Gaudin, M., and Desnues, C. (2018). Hybrid Capture-Based Next Generation Sequencing and Its Application to Human Infectious Diseases. Front Microbiol *9*, 2924.

Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E.M., Brockman, W., Fennell, T., Giannoukos, G., Fisher, S., Russ, C., et al. (2009a). Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol *27*, 182–189.

Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E.M., Brockman, W., Fennell, T., Giannoukos, G., Fisher, S., Russ, C., et al. (2009b). Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol *27*, 182–189.

Iossifov, I., O'Roak, B.J., Sanders, S.J., Ronemus, M., Krumm, N., Levy, D., Stessman, H.A., Witherspoon, K.T., Vives, L., Patterson, K.E., et al. (2014). The contribution of de novo coding mutations to autism spectrum disorder. Nature *515*, 216–221.

Jaton, K., and Greub, G. (2007). [PCR in microbiology: from DNA amplification to results interpretation]. Rev Med Suisse *3*, 931–932, 934–938.

John, T.M., and Taege, A.J. (2019). Appropriate laboratory testing in Lyme disease. CCJM *86*, 751–759.

Jones, M.R., and Good, J.M. (2016). TARGETED CAPTURE IN EVOLUTIONARY AND ECOLOGICAL GENOMICS. Mol Ecol *25*, 185–202.

Köster, J., and Rahmann, S. (2012). Snakemake—a scalable bioinformatics workflow engine. Bioinformatics *28*, 2520–2522.

Kugeler, K.J., Schwartz, A.M., Delorey, M.J., Mead, P.S., and Hinckley, A.F. Estimating the Frequency of Lyme Disease Diagnoses, United States, 2010–2018 - Volume 27, Number 2—February 2021 - Emerging Infectious Diseases journal - CDC.

Lantos, P.M., Auwaerter, P.G., and Nelson, C.A. (2016). Lyme Disease Serology. JAMA *315*, 1780–1781.

Livak, K.J., Flood, S.J., Marmaro, J., Giusti, W., and Deetz, K. (1995). Oligonucleotides with fluorescent dyes at opposite ends provide a quenched probe system useful for detecting PCR product and nucleic acid hybridization. Genome Research *4*, 357–362.

Masago, K., Fujita, S., Oya, Y., Takahashi, Y., Matsushita, H., Sasaki, E., and Kuroda, H. (2021). Comparison between Fluorimetry (Qubit) and Spectrophotometry (NanoDrop) in the Quantification of DNA and RNA Extracted from Frozen and FFPE Tissues from Lung Cancer Patients: A Real-World Use of Genomic Tests. Medicina (Kaunas) *57*, 1375.

Metsky, H.C., Siddle, K.J., Gladden-Young, A., Qu, J., Yang, D.K., Brehio, P., Goldfarb, A., Piantadosi, A., Wohl, S., Carter, A., et al. (2019). Capturing sequence diversity in metagenomes with comprehensive and scalable probe design. Nat Biotechnol *37*, 160–168.

Mitchell, P.D., Reed, K.D., Vandermause, M.F., and Melski, J.W. (1993). Isolation of Borrelia burgdorferi from skin biopsy specimens of patients with erythema migrans. Am J Clin Pathol *99*, 104–107.

Moore, A., Nelson, C., Molins, C., Mead, P., and Schriefer, M. (2016). Current Guidelines, Common Clinical Pitfalls, and Future Directions for Laboratory Diagnosis of Lyme Disease, United States. Emerg Infect Dis *22*, 1169–1177.

Murray, T.S., and Shapiro, E.D. (2010). Lyme Disease. Clin Lab Med *30*, 311–328.

Page, A.J., Cummins, C.A., Hunt, M., Wong, V.K., Reuter, S., Holden, M.T.G., Fookes, M., Falush, D., Keane, J.A., and Parkhill, J. (2015). Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics *31*, 3691–3693.

Pikor, L.A., Enfield, K.S.S., Cameron, H., and Lam, W.L. (2011). DNA extraction from paraffin embedded material for genetic and epigenetic analyses. J Vis Exp 2763.

Pomraning, K.R., Smith, K.M., Bredeweg, E.L., Connolly, L.R., Phatale, P.A., and Freitag, M. (2012). Library Preparation and Data Analysis Packages for Rapid Genome Sequencing. Methods Mol Biol *944*, 1–22.

Poptsova, M.S., Il'icheva, I.A., Nechipurenko, D.Y., Panchenko, L.A., Khodikov, M.V., Oparina, N.Y., Polozov, R.V., Nechipurenko, Y.D., and Grokhovsky, S.L. (2014). Non-random DNA fragmentation in next-generation sequencing. Sci Rep *4*, 4532.

Price, M.N., Dehal, P.S., and Arkin, A.P. (2009). FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. Molecular Biology and Evolution *26*, 1641–1650.

Qin, D. (2019). Next-generation sequencing and its clinical application. Cancer Biol Med *16*, 4–10.

Replogle, A.J., Sexton, C., Young, J., Kingry, L.C., Schriefer, M.E., Dolan, M., Johnson, T.L., Connally, N.P., Padgett, K.A., and Petersen, J.M. (2021). Isolation of Borrelia miyamotoi and other Borreliae using a modified BSK medium. Sci Rep *11*, 1926.

Roy, S., Hartley, J., Dunn, H., Williams, R., Williams, C.A., and Breuer, J. (2019). Whole-genome Sequencing Provides Data for Stratifying Infection Prevention and Control Management of Nosocomial Influenza A. Clinical Infectious Diseases *69*, 1649–1656.

Schutzer, S.E., Body, B.A., Boyle, J., Branson, B.M., Dattwyler, R.J., Fikrig, E., Gerald, N.J., Gomes-Solecki, M., Kintrup, M., Ledizet, M., et al. (2019). Direct Diagnostic Tests for Lyme Disease. Clinical Infectious Diseases *68*, 1052–1057.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics *30*, 2068–2069.

Steere, A.C., Dhar, A., Hernandez, J., Fischer, P.A., Sikand, V.K., Schoen, R.T., Nowakowski, J., McHugh, G., and Persing, D.H. (2003). Systemic symptoms without erythema migrans as the presenting picture of early Lyme disease. Am J Med *114*, 58–62.

Tilly, K., Casjens, S., Stevenson, B., Bono, J.L., Samuels, D.S., Hogan, D., and Rosa, P. (1997). The Borrelia burgdorferi circular plasmid cp26: conservation of plasmid structure and targeted inactivation of the ospC gene. Molecular Microbiology *25*, 361–373.

Tokarz, R., Mishra, N., Tagliafierro, T., Sameroff, S., Caciula, A., Chauhan, L., Patel, J., Sullivan, E., Gucwa, A., Fallon, B., et al. (2018). A multiplex serologic platform for diagnosis of tick-borne diseases. Sci Rep *8*, 3158.

Vitzthum, F., Geiger, G., Bisswanger, H., Brunner, H., and Bernhagen, J. (1999). A quantitative fluorescence-based microplate assay for the determination of double-stranded DNA using SYBR Green I and a standard ultraviolet transilluminator gel imaging system. Anal Biochem *276*, 59–64.

Walter, K.S., Carpi, G., Caccone, A., and Diuk-Wasser, M.A. (2017). Genomic insights into the ancient spread of Lyme disease across North America. Nat Ecol Evol *1*, 1569–1576.

Wiegers, U., and Hilz, H. (1971). A new method using “proteinase K” to prevent mRNA degradation during isolation from HeLa cells. Biochem Biophys Res Commun *44*, 513–519.

Wormser, G.P., Dattwyler, R.J., Shapiro, E.D., Halperin, J.J., Steere, A.C., Klempner, M.S., Krause, P.J., Bakken, J.S., Strle, F., Stanek, G., et al. (2006). The clinical assessment, treatment, and prevention of lyme disease, human granulocytic anaplasmosis, and babesiosis: clinical practice guidelines by the Infectious Diseases Society of America. Clin Infect Dis *43*, 1089–1134.

Worthey, E.A., Mayer, A.N., Syverson, G.D., Helbling, D., Bonacci, B.B., Decker, B., Serpe, J.M., Dasu, T., Tschannen, M.R., Veith, R.L., et al. (2011). Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. Genet Med *13*, 255–262.

(2001). Crystal structure of outer surface protein C (OspC) from the Lyme disease spirochete, Borrelia burgdorferi. The EMBO Journal *20*, 971–978.

Illumina Custom Protocol.

Roche cobas x 480 Operators Manual | Manualzz.

Tagmentation | Reduce library prep time with on-bead tagmentation.

(OPTIONAL)
APPENDIX

A.1: Plotting script for depth ratio plot

```
 import bokeh

from bokeh.models import Panel, Tabs
from bokeh.plotting import figure, output_file, show
from operator import add

import numpy as np
import math

# Hybrid runs
runs = ['SRR2034333.1', 'SRR2034334.1', 'SRR2034335.1', 'SRR2034336.1', 'SRR2034337.1',
'SRR2034338.1',
'SRR2034339.1', 'SRR2034340.1', 'SRR2034341.1', 'SRR2034342.1']

# Normal method runs
# norm_runs = ['SRR9616106', 'SRR9616113', 'SRR9616115', 'SRR9616121', 'SRR9616125',
'SRR9616129', 'SRR9616130',
# 'SRR9616133', 'SRR9616136', 'SRR9616137']

norm_runs = ['10071001.l10071001.000H3LCYN.1', '10071002.l10071002.000H3LCYN.1',
'10071003.l10071003.000H3LCYN.1',
'10081001.l10081001.000H3LCYN.1', '10081004.l10081004.000H3LCYN.1']

# Function to parse .txt version of .bam file
def read_bam(filename):
    # Open file, read in lines
    path = "../depths/" + filename + ".txt"
    reads = open(path)

    lines = reads.readlines()
    reads.close()

    # Dictionary to differentiate between plasmids
    chromosomes = {}
    current = ""

    # Loop through lines. If we see a new plasmid, change to that. Regardless, add next
depth reading to plasmid.
    for line in lines:
        temp = line.split()

        if current != temp[0]:
            current =  temp[0]
```

```
            chromosomes[current] = []

        chromosomes[current].append(int(temp[2]))

    return chromosomes

# Lists of dictionaries for each run, hybrid and normal.
depths = [read_bam(run) for run in runs]
norm_depths = [read_bam(run) for run in norm_runs]
tab_list = []
print(depths[0].keys())
print(norm_depths[0].keys())
# Iterate through each plasmid
for key in depths[0]:
    # Find average depth for hybrids
    pos = [run[key] for run in depths]
    total = [sum(x) for x in zip(*pos)]
    average = [x/10 for x in total]

    # Set up ndarrays for plotting
    y = np.array(average)
    x = np.arange(start=1, stop=y.size + 1, step=1)

    # Find aerage depth for normal runs
    norm_pos = [run[key] for run in norm_depths]
    norm_total = [sum(x) for x in zip(*norm_pos)]
    norm_average = [x/5 for x in norm_total]

    # Set up ndarrays for plotting
    norm_y = np.array(norm_average)
    norm_x = np.arange(start=1, stop=norm_y.size + 1, step=1)

    # Find percentage of the time hybrid does better vs percentage normal does better.
    hybrid = 0
    normal = 0
    ratio = []
    for h, n in zip(y, norm_y):
        if h > n:
            hybrid += 1
        elif n > h:
            normal += 1
        if n == 0 and h != 0:
            ratio.append(h)
        elif n==0 and h==0:
            ratio.append(1)
        else:
            ratio.append(h/n)

    print("HYBRID %: " + str(hybrid/len(y)))
```

```
    print("NORMAL %: " + str(normal/len(norm_y)))

    nd_ratio = np.array(ratio)

    p = figure(plot_width=600, plot_height=400, title='Ratio of average read depth of
hybrid capture vs non-hybrid capture samples',
        x_axis_label='Read depth ratio', y_axis_label='Position in plasmid')
    p.line(x, nd_ratio, line_width=2, color='blue')

    # p.line(x, y, line_width=2, color='red')
    # p.line(norm_x, norm_y, line_width=2, color='blue')

    tab = Panel(child=p, title=key)
    tab_list.append(tab)

tabs = Tabs(tabs=tab_list)

show(tabs)
output_file("plots/ratio.html")
```

A.2: Plotting script for simple read depth

```
import matplotlib.pyplot as plt
from matplotlib import gridspec
import numpy as np
import math

reads = open('results.txt', 'r')
lines = reads.readlines()[1:]
reads.close()

chromosomes = {}
current = ""

for line in lines:
    temp = line.split()

    if current != temp[0]:
        current =  temp[0]
        chromosomes[current] = []

    chromosomes[current].append(temp[2])

N = len(chromosomes)
cols = 4
rows = int(math.ceil(N / cols))

gs = gridspec.GridSpec(rows, cols)
fig = plt.figure()

for key, n in zip(chromosomes, range(N)):
    pos = chromosomes[key]
    y = np.array(pos)
    x = np.arange(start=1, stop=y.size + 1, step=1)

    ax = fig.add_subplot(gs[n])
    ax.plot(x, y)
    ax.yaxis.set_ticks([0, 50, 100])

plt.savefig('plots/depth.svg', dpi=1200)

SAMPLES = ["10071001.l10071001.000H3LCYN.1.fasta"]
```

A.3 Script to perform per-plasmid paired t-test

```
import bokeh

from bokeh.models import Panel, Tabs
from bokeh.plotting import figure, output_file, show
from operator import add
from  scipy.stats import ttest_ind

import numpy as np
import math

# Hybrid runs
runs = ['SRR2034333.1', 'SRR2034334.1', 'SRR2034335.1', 'SRR2034336.1', 'SRR2034337.1',
'SRR2034338.1',
'SRR2034339.1', 'SRR2034340.1', 'SRR2034341.1', 'SRR2034342.1']

# Normal method runs
norm_runs = ['SRR9616106', 'SRR9616113', 'SRR9616115', 'SRR9616121', 'SRR9616125',
'SRR9616129', 'SRR9616130',
'SRR9616133', 'SRR9616136', 'SRR9616137']

# Function to parse .txt version of .bam file
def read_bam(filename):
    # Open file, read in lines
    path = "../depths/" + filename + ".txt"
    reads = open(path)

    lines = reads.readlines()
    reads.close()

    # Dictionary to differentiate between plasmids
    chromosomes = {}
    current = ""

    # Loop through lines. If we see a new plasmid, change to that. Regardless, add next
depth reading to plasmid.
    for line in lines:
        temp = line.split()

        if current != temp[0]:
            current =  temp[0]
            chromosomes[current] = []

        chromosomes[current].append(int(temp[2]))

    return chromosomes

# Lists of dictionaries for each run, hybrid and normal.
```

```
depths = [read_bam(run) for run in runs]
norm_depths = [read_bam(run) for run in norm_runs]
tab_list = []

# Iterate through each plasmid
for key in depths[0]:
    # Find average depth for hybrids
    pos = [run[key] for run in depths]
    total = [sum(x) for x in zip(*pos)]
    average = [x/10 for x in total]

    # Find average depth for normal runs
    norm_pos = [run[key] for run in norm_depths]
    norm_total = [sum(x) for x in zip(*norm_pos)]
    norm_average = [x/10 for x in norm_total]

    # Run paired t-test
    # res = ttest_ind(average, norm_average)

    print("PLASMID: " + key)
    # print("T-STATISTIC: " + str(res[0]), "P-VALUE: " + str(res[1]))
    print(ttest_ind(average, norm_average))
    print("================================")
```

A.4: Snakefile for pipelined data processing

```
rule all:
    input:
        "plots/quals.svg"

rule sra_fastq:
    input:
        "data/samples/selfgen/10081004.l10081004.000H3LCYN.1.fasta"
    shell:
        "./seqtk/seqtk seq -F '!' {input} >
data/samples/selfgen/10081004.l10081004.000H3LCYN.1.fastq"
        # "./sratoolkit/bin/fastq-dump --fasta {input}"

rule bwa_map:
    input:
        "data/refgen.fna",
        "data/samples/selfgen/10081004.l10081004.000H3LCYN.1.fastq"
    output:
        "mapped_reads/10081004.l10081004.000H3LCYN.1.bam"
    shell:
        "bwa mem {input} | samtools view -Sb - > {output}"


rule samtools_sort:
    input:
        "mapped_reads/10081004.l10081004.000H3LCYN.1.bam"
    output:
        "sorted_reads/10081004.l10081004.000H3LCYN.1.bam"
    shell:
        "samtools sort -T {input} "
        "-O bam {input} > {output}"


rule samtools_index:
    input:
        "sorted_reads/10071001.l10071001.000H3LCYN.1.bam"
    output:
        "sorted_reads/10071001.l10071001.000H3LCYN.1.bam.bai"
    shell:
        "samtools index {input}"


rule bcftools_call:
    input:
        fa="data/refgen.fna",
        bam=expand("sorted_reads/{sample}.bam", sample=SAMPLES),
        bai=expand("sorted_reads/{sample}.bam.bai", sample=SAMPLES)
    output:
        "calls/all.vcf"
```

```
    shell:
        "samtools mpileup -g -f {input.fa} {input.bam} | "
        "bcftools call -mv - > {output}"


rule plot_quals:
    input:
        "calls/all.vcf"
    output:
        "plots/quals.svg"
    script:
        "scripts/plot-quals.py"

#./samtools/samtools depth sorted_reads/....bam -a -o depths/....txt
```