8-2016

Understanding the Mechanism and Extent of vlsE Recombination in Borrelia burgdorferi using Next-Generation Sequencing

Surabhi Tyagi

Understanding the Mechanism and Extent of *vlsE* Recombination in *Borrelia burgdorferi* using Next-Generation Sequencing

by

Surabhi Tyagi, B.S.

APPROVED:

_____
Steven J. Norris, Ph.D.
Advisory Professor

_____
Cesar Arias, MD, Ph.D.

_____
Nayun Kim, Ph.D.

_____
Kevin Morano, Ph.D.

_____
David Volk, Ph.D.

APPROVED:

_____
Dean, The University of Texas
Graduate School of Biomedical Sciences at Houston

UNDERSTANDING THE MECHANISM AND EXTENT OF *VLSE* RECOMBINATION IN *BORRELIA BURGDORFERI* USING NEXT-GENERATION SEQUENCING

A

THESIS

Presented to the Faculty of
The University of Texas
Health Science Center at Houston
and
The University of Texas
MD Anderson Cancer Center
Graduate School of Biomedical Sciences
in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE [DOCTOR OF PHILOSOPHY]

by

Surabhi Tyagi, B.S.
Houston, Texas

August 2016

Acknowledgements

I would like to thank my advisor, Dr. Steven Norris, for continuously providing me with guidance and support during the time of research and writing of this thesis. My special thanks to Dr. Diane Edmondson for helping with experiments and other different aspects of my project throughout my research. Without their guidance and persistent help, this project would not have been possible.

I would also like to thank all present and past members of the Norris lab: Dr. Bijay Khajanchi, Liu Gau, Dr. Tao Lin, Evelyn Odhe, Dr. Sabitha Prabhakaran for sharing their knowledge and expertise.

I would like to thank my committee members, Dr. Nayun Kim, Dr. Cesar Arias, Dr. Kevin Morano, and Dr. David Volk, for providing constructive criticism and insightful comments. In addition to being on my committee, Dr. David Volk helped us develop an automated program to process sequences. His knowledge and suggestions were a tremendous help in this project.

Thanks to Dr. George Weinstock and members of his group at the Jackson Laboratory of Genomic Medicine for conducting the PacBio sequencing experiments. Their willingness to collaborate and generously offering their services was greatly appreciated.

Lastly, I would like to thank my family for their support and encouragement.

UNDERSTANDING THE MECHANISM AND EXTENT OF *VLSE* RECOMBINATION IN *BORRELIA BURGDORFERI* USING NEXT-GENERATION SEQUENCING

Surabhi Tyagi, B.S.
Advisory Professor: Steven Norris, Ph.D.

B. burgdorferi, the causative agent of Lyme disease, have an elaborate antigenic variation system that involves varying the sequence of *vlsE.* Previous studies have shown that *vlsE* antigenic variation occurs continuously inside mammalian hosts. Variation has not been shown previously to occur in *in vitro* or in ticks. We hypothesized that the induction of *vlsE* recombination requires contact with dense arrays of host tissue cells and/or ECM components. To test this hypothesis, two methods, quantitative PCR and high-throughput sequencing were used determine the extent and nature of *vlsE* recombination within mouse tissues and *in vitro* model systems. Using these approaches, we were able to detect *vlsE* variants in axenic cultures of *B. burgdorferi* as well as co-cultures with mouse skin and heart tissues; these results were compared with those from mice infected for 7 days. Analysis of PacBio single molecule real-time (SMRT) sequencing indicated the presence of 0.84% to 1.18% variants in pure *in vitro* cultures and 0.79% to 1.22% in tissue explants, as compared 36% to 57% for organisms from mouse bladder tissue 7 days post inoculation. Statistical evaluation of the variants showed that the rate of recombination in tissue explants was not significantly different from the rate of recombination in *in vitro* cultures. Thus, tissue explant co-cultures do not seem to promote a higher recombination rate than in *in vitro* axenic culture. Moreover, high-throughput PacBio sequencing was found to be an effective means of analyzing single molecule sequencing variation in the robust *vlsE* antigenic variation system.

Table of Contents

List of Figures

Chapter 1

Chapter 2

Chapter 3

List of Tables

Chapter 3

List of Abbreviations

B. burgdorferi	Borrelia burgdorferi
Bb	Borrelia burgdorferi
BBK32	Fibronectin-binding protein BBK32
BLAST	Basic Local Alignment Search Tool
BSK II medium	Barbour-Stoenner-Kelly-II medium
Buffer ATL	Animal Tissue Lysis Buffer
CDC	Center of Disease Control and Prevention
C_T	Threshold cycle
DbpA	Decorin-binding protein A
DbpB	Decorin-binding protein B
ECM	Extracellular matrix
ELISA	Enzyme-linked immunosorbent assay
Fn	Fibronectin
GAGs	Glycosaminoglycans
HMM	Hidden Markov model
IgG	Immunoglobulin G
IgM	Immunoglobulin M
IR	Invariable region
kb	Kilobase
lp	Linear plasmid

OspA	Outer surface protein A
OspC	Outer surface protein C
P21	Parental primer, 21 bp long
PacBio	Pacific Biosciences
PCR	Polymerase chain reaction
qPCR	Quantitative polymerase chain reaction
ruvA	Holliday junction ATP-dependent DNA helicase RuvA
ruvB	Holliday junction ATP-dependent DNA helicase RuvB
SCID	Severe combined immunodeficiency
SMRT sequencing	Single molecule real time sequencing
V21	Variant primer, 21 bp long
Vlp	Variable large protein
Vls	Vmp-like sequence
VlsE	Vmp like expression cassette
VMPs	Variable major proteins
VR	Variable region
VSG	Variant-specific glycoprotein
Vsps	Variable-small proteins
ZMWs	Zero mode wave guide

Chapter 1

Introduction

Borrelia Spirochetes

Borrelia is a genus that belongs to the spirochete phylum *Spirochetes* and the order *Spirochaetales* [1]. Like *Leptospira* and *Treponema*, other members of the Spirochete phylum, *Borrelia* are motile bacteria with a distinctive helical shape and an outer membrane. Of all the members of the spirochete phylum, *Borrelia* species are the only ones that require an arthropod vector for vertebrate host transmission. There are currently 36 known *Borrelia* species, and 15 Lyme disease species; the latter group is also known as *Borrelia burgdorferi* sensu lato ("in a general sense"). It was proposed recently that the *Borrelia* genus be reorganized into two genera: *Borreliella* for the Lyme disease group and *Borrelia* for the relapsing fever organisms [2].

Lyme disease *Borrelia*

Of the *B. burgdorferi* sensu lato species, the three that cause the most human disease are *B. garinii, B. afzelii*, and *B. burgdorferi* sensu stricto ("in a strict sense") [1]. *B. burgdorferi* sensu stricto is the predominant causative agent of Lyme disease in the United States, although another species, provisionally called '*Borrelia mayonii*', was recently described [3]. Infections caused by *B. garinii* and *B. afzelii* are frequently found in Europe and Asia.

Approximately 35,000 cases were reported to the Centers for Disease Control and Prevention (CDC) in the year 2013. However, a recent analysis by the CDC indicated that the actual incidence in the U.S. is over 300,000 cases per year [4]. Symptoms of Lyme disease include skin lesions, arthritis, carditis, and neurologic deficits. The name of Lyme disease originates from the town of Lyme, Connecticut, where clusters of pediatric and adult arthritis cases were occurring [5]. These cases were studied and were thought to

represent a new clinical entity, initially named Lyme arthritis. Later studies showed that the syndrome was associated with the bite of *Ixodes* ticks, commonly resulting in a bulls-eye shaped rash called erythema migrans. The causative organism, later called *B. burgdorferi*, was first cultured from ticks in 1981 [6].

Clinical manifestations of Lyme disease

Lyme disease as an illness can be divided into three stages, with each stage presenting different clinical manifestations [7]. The early manifestations of Lyme disease (stage one) include erythema migrans, low-grade fever, malaise and fatigue. Although the lesions are localized, the organism has already begun to disseminate and can be isolated from blood in many cases. The antibody response starts to develop, but is not consistently detected in all patients. At the disseminated stage (stage 2) many patients experience intermittent episodes of migratory musculoskeletal pain and attacks of asymmetric arthritis in large joints, especially the knee. Eventually, the disease becomes persistent (stage 3) and the episodes of arthritis become longer, lasting for months to years. Despite the availability of effective antibiotic treatment for Lyme disease, some patients experience post-treatment Lyme disease syndrome (PTLDS) where they continue to have symptoms. PTLDS is a highly controversial topic in the scientific community; there is a no universally accepted diagnosis and treatment for PTLDS. Moreover, the underlying mechanism of PTLDS is unknown; some scientists believe that the post-infectious autoimmunity is the cause of PTLDS, while others attribute the disease to persistent *B.burgdorferi* infection [8].

Diagnosis and treatment of Lyme disease

The diagnosis of Lyme disease is based primarily on clinical manifestations and history, with support from serologic tests. In the U.S., a two-step laboratory testing process

for Lyme disease is utilized [9]. First, an Enzyme-Linked Immunosorbent Assay (ELISA) is performed to test the antibody response to a *B. burgdorferi* lysate or a mixture of recombinant antigens. Then, those test samples with positive ELISA results are examined by immunoblotting to detect IgM and IgG antibodies against individual *Borrelia* proteins. A positive IgM immunoblot is only useful during the first 4-weeks of infection. IgG antibodies are more reliable but can take 4-6 weeks to develop.

Treatment of Lyme disease consists of an initial 14 to 21 day course of oral doxycycline or amoxicillin [10]. Some patients require an additional 30-day antibiotic treatment. If patients have symptoms such as joint swelling or neuroborreliosis, they are placed on a 4-week course of IV ceftriaxone therapy. These treatments are effective in eradicating symptoms in most cases, but roughly 10% of patients have lingering symptoms called post-treatment Lyme disease syndrome. New oral antibiotics approved by the FDA, including daptomycin, carbomycin, have shown effectiveness against spirochetes in culture, but there are no data yet to support if they work in patients with Lyme disease.

***Borrelia burgdorferi* genome**

B. *burgdorferi (Bb)* has a highly unusual genome, consisting of a linear chromosome and 21 circular and linear plasmids that range from 5 to 56 kilobases in size. Linear plasmid names begin with lp, and circular plasmids begin with cp; both end with a number that indicates the plasmid size in kilobases (e.g., lp54, cp9). *B. burgdorferi* strain B31 has a linear chromosome of approximately 910 kb and 21 plasmids totaling 610 kb [11]. The overall G+C content of the genome is 28.6%. Studies have shown that certain *B. burgdorferi* plasmids are associated with infectivity. For instance, organisms lacking lp28-1 exhibit

intermediate-infectivity (recovery of organisms up to two weeks after inoculation), while the loss of lp25 results in low-infectivity phenotype [12, 13].

The chromosomes of all Borrelia organisms have a high degree of synteny [14]. Although the overall gene content of the plasmids of *B. garinii, B. afzelii* and *B. burgdorferi* is generally well conserved, extensive intraplasmidic recombinations are evident in different species and strains.

Vector tick *Ixodes scapularis*

Lyme disease is transmitted to humans by the bite of infected ticks of the genus *Ixodes*. *I. scapularis* is the vector in the Eastern and North Central U.S., whereas *I. pacificus* is the agent in the western U.S. *I. ricinus* and *I. persulcatus* are the vectors in Europe and Asia respectively [15].

Lyme disease transmission cycle

The enzootic cycle of *B. burgdorferi* begins with eggs laid by adult female ticks in the spring (Figure 1.1). The larvae hatch in the summer and can acquire *Borrelia* when they take their first blood meal from infected small mammals or birds. After feeding, the larvae molt and undergo metamorphosis into nymphal ticks. The ticks take a second blood meal during the following spring. Nymphal ticks are able to transmit *Borrelia* to larger mammals including humans and are responsible for the majority of Lyme disease cases. Humans are an accidental host of Lyme disease and represent a 'dead end' in the transmission process. Next, the nymphal ticks molt into the adult stage. Adult females feed on large mammals, such as deer, and lay eggs in the spring, thus starting the two-year cycle all over again. Adult female ticks can lay up to 3000 eggs and die just a few days afterwards [4].

Differential expression of outer surface lipoproteins in *B. burgdorferi* during the transmission cycle

During its transmission cycle, *B. burgdorferi* goes through many environmental changes as it is transferred between the tick and mammalian hosts. Differential gene expression of lipoproteins plays a major role in the ability of the organism to adapt to different hosts [16]. As an example, outer surface protein A (OspA) is encoded on lp54 and is predominantly expressed in the midgut of an infected unfed tick. When an unfed nymphal tick begins to feed on a mammalian host, the environment in the midgut changes from a pH of 7.4 to pH 6.8 and the temperature changes from ~22 ºC to near 37 ºC. During feeding, OspA expression is down-regulated while OspC expression is up-regulated, and spirochetes migrate from the midgut to the salivary glands. OspC plays an important role in the transmission and initial dissemination of *Borrelia* from the tick in the mammalian host, and its expression peaks after 48 hours of feeding [17].

Decorin-binding proteins DbpA and DbpB, also encoded lp54, are expressed during mammalian infection but not in ticks [18]. These proteins mediate adherence to decorin, a proteoglycan that 'decorates' the surface of collagen fibrils in some tissues. Tissues with high decorin expression, such as skin, joints, and endothelium are thought to function as protective niches for *Borrelia.* Another adhesive lipoprotein is BBK32. This protein binds to glycosaminoglycans (GAGs) and fibronectin (Fn), a complex glycoprotein present in plasma and extracellular matrix (ECM). Studies have shown that the interaction of BBK32 with GAGs and Fn play an important role in the dissemination, colonization and persistence of *Borrelia* in the mammalian host [19].

Overview of antigenic variation systems

Antigenic variation is a mechanism by which an infectious agent can alter the sequence of its antigenic surface proteins. The alteration of surface protein sequence allows the infectious agent to escape the host immune response and thus promote long-term infection [20]. Antigenic variation systems hinder the development of vaccines for many pathogens.

One example of a vector-borne pathogen that undergoes antigenic variation is *Trypanosoma brucei*, a protozoan transmitted to mammals by tsetse flies. *T. brucei* continuously switch their expression of variant-surface glycoproteins (VSGs) surface protein [21]. Of the ~2000 VSG genes and pseudogenes, only one is transcriptionally active while others remain silent. The silent VSG genes, which are located in clusters of large arrays within the subtelomeric region of the 11 Mb chromosomes, can be copied into one of 15 telomeric VSG expression site (ES) through gene conversion. In early stages of infection, an entire silent VSG gene replaces the ES. In the later stages of infection, segmental gene conversion emerges as an additional mechanism of *VSG* switching. Segmental gene conversion of multiple silent VSG genes can result in endless possibilities of new chimeric VSGs.

Neisseria gonorrhoeae, a human pathogen that is the causative agent of the sexually transmitted disease gonorrhea, also has an antigenic variation system that involves variation of their Type IV pilin [21]. Segments of *pilS* silent cassettes are unidirectionally transferred into the *pilE* expression locus through a gene conversion process. The copies of *pilS* sequences are predominantly variable regions while the expressed *pilE* sequence contains both variable and invariable regions. The recombination between *pilS* and *pilE* involves replacement of *pilE* segments by *pilS* sequences, leaving the *pilS* sequence

unchanged and creating a hybrid *pilE* sequence. The length and region of DNA replacement in a recombination event can also vary. Recent data from the Seifert group showed that pilin antigenic variation involves a guanine quartet (G4) structure [22]. Their later studies showed that G-rich sequence can only form a G4 structure when a *pilE* G4-associated sRNA was transcribed [23].

The causative agents of relapsing fever, such as *Borrelia hermsii*, also have antigenic variation systems that involve varying variable major proteins (Vmps) [24]. *B. hermsii* plasmids have a single Vmp expression site and ~40 silent gene segments that encode variable large proteins (Vlps) and variable small proteins (Vsps). Gene conversion occurs between a silent *vsp* or *vlp* gene segment and the expression site. Studies have shown that the *vlp* or *vsp* sequence at the expression site is usually replaced completely by a silent *vlp* or *vsp* gene segment [25]. This process involves recombination at upstream and downstream homology sequences. Similar to *B. hermsii*, the antigenic variation system in *Borrelia burgdorferi* involves gene conversion between the *vmp*-like sequence expression site (*vlsE*) and the silent *vls* cassettes, as described below.

The *vls* antigenic variation system

The *vls* antigenic variation system was first described in *B. burgdorferi* in 1997 [26] and is present in all Lyme disease *Borrelia* strains thus far characterized [27]. The gene encoding the 35 kDa outer surface protein VlsE undergoes extensive recombination during mammalian infection. In ticks and in *in vitro* culture, VlsE is expressed, but recombination had not been detected in these environments prior to the research presented in this thesis [28].

The antigenic variation system of *B. burgdorferi* consists of varying the sequence of VlsE, a surface localized lipoprotein [29]. The VlsE protein is encoded by *vlsE*, a gene located on a 28 kb linear plasmid (lp28-1) in *B. burgdorferi* B31. The *vlsE* expression site is downstream of 15 silent cassettes that are not expressed (Figure 1.2 A). There is 90% to 96.1% sequence homology between the 15 silent cassettes and the central cassette region of *vlsE*. Each cassette has 6 variable regions (VRs) and 6 relatively invariant regions (IRs). Most of the sequence differences between *vlsE* and the *vls* silent cassettes occur within the 6 variable regions (Figure 1.2 B). The recombination events include not only nucleotide changes but also codon-length (multiple of threes) indels. There is 76.9% to 91.4% predicted amino acid identity, consistent with the fact that most of the nucleotide differences are non-synonymous. Figure 1.2 C shows an example of gene conversion events, where portions of the central *vlsE* cassette region are replaced by segments of silent cassettes 9, 7 and 10 sequentially. The first *B. burgdorferi* B31 *vlsE* allele described is called *vlsE1.* All the *vlsE* recombination experiments reported herein use clones with the *vlsE1* sequence; this sequence is therefore referred to as the parental sequence.

Previous studies in our laboratory [29], where 1,399 clones isolated from *B. burgdorferi* B31-infected mouse tissues were analyzed, showed that there is an increased frequency and complexity of *vlsE* sequence changes as the days post infection increase. Immunocompetent C3H/HeN and severe combined immunodeficiency (SCID) mice were infected with a B31 clone, and tissues from sacrificed mice were taken at days 4, 7, 10, 14 and 28 days post infection. Recombination events were observed in ~50% of organisms recovered from infected C3H/HeN mice 4 days post inoculation, indicating that the recombination process begins early in the course of infection. *Borrelia* clones with the

'parental' *vlsE* sequence are cleared more rapidly in immunocompetent mice than in SCID mice, indicative of immune elimination of clones with invariant VlsE sequences. By day 28 post-infection, no clones with the *vlsE* parental sequence were found in the immunocompetent mice. In SCID mice, however, *vlsE* variants were isolated in increasing numbers throughout the course of infection, but sequence changes accumulated more gradually and clones with the parental sequence were still detected on day 28 post-infection. Thus, the presence of an adaptive immune response is not required to induce *vlsE* recombination, but non-variant bacteria appear to be eliminated by the immune response. Recombination events were distributed throughout the *vlsE* cassette, showing no preference for a particular region within the cassette.

The length of recombination events ranged from single nucleotide change to almost the entire *vlsE* cassette [29]. Due to high sequence homology among the silent cassettes, a single silent cassette donor could not be identified unambiguously in some instances. Many variant sequences resulted from multiple recombination events and thus resulted in complex sequences with an untraceable lineage. Also, 15% of all the variant clones analyzed had template-independent changes (usually single nucleotide substitutions), meaning that these changes could not be explained as a gene conversion event between *vlsE* and the 15 silent cassettes.

Gene conversion in Chicken immunoglobulin genes

As discussed in the previous sections, many pathogens including *B. burgdorferi* use gene conversion as the mechanism to vary their antigens. Outside of pathogenic microorganisms, gene conversion is used as the mechanism to expand and diversify the B-cell repertoire in Chickens [30]. Diversification of B-cells is essential in allowing the animal

immune system to recognize and target a variety of invading bacteria and viruses. The expansion of antibodies is the result of B-cells recombining their immunoglobulin genes: variable (V), diversity (D) and joining (J). After V(D)J recombination, a unique VλJ segment is generated. In chicken B-cells; the VλJ fragment undergoes gene conversion events, where the upstream pseudo V segments act as the template for the unidirectional, segmental recombination.

Characteristics of VlsE protein structure

VlsE is a surface-exposed lipoprotein that has a molecular weight of 35 kDa [26]. The crystal structure of VlsE [31] shows that the 6 variable regions form random loop structures that comprise most of VlsE membrane-distal surface (Figure 1.3). It has been postulated that the surface exposure of variable regions plays a role in hindering interactions of host antibodies with the bacterial surface; this property may help *Borrelia* evade the host immune response. The invariable regions have a structure made up primarily of alpha helices with some beta-strands. Conserved N- and C-terminal regions flank the central VlsE cassette and also have alpha helical structures.

Mechanism of *vlsE* recombination

Gene conversion, a mechanism that results in a unidirectional exchange between a conserved DNA sequence and the 'targeted' DNA that undergoes sequence alteration, best describes *vlsE* recombination, although non-templated point mutations have also been shown to occur. Previous studies performed by our laboratory show that silent cassettes do not change during recombination, further supporting gene conversion as the primary mechanism of *vlsE* recombination [28]. Comprehensive studies, in which mutants of genes known to be involved in DNA recombination and repair were analyzed, showed that genes

ruvA and *ruvB* are important in *vlsE* recombination [32, 33]. RuvA and RuvB form a complex that acts as the Holliday junction helicase during homologous recombination. *ruvA* and *ruvB* mutants had a significantly lower rate of *vlsE* recombination in C3H/HeN mice. No clones with variant *vlsE* sequences were recovered from *ruvA* or *ruvB* mutants in immunodeficient (C3H/*scid*) mice 28 days post infection, further indicating that the recombination rate was greatly reduced in these mutants. *mutS* mutants also displayed a minor effect on *vlsE* recombination [33]. Mutation of the other genes examined, including *recA*, did not have a significant effect on vlsE recombination rates [32-34]. Other than the involvement of *ruvA* and *ruvB* and the potential involvement of *mutS*, the proteins and recombinases involved in the *vlsE* recombination remain unknown.

vls sequences have an unusually high GC content of ~50% while the rest of the genome has a GC content of ~28% [35]. There is also a high GC skew in *vls* sequences (with a higher G content on the coding strand), which likely contributes to the recombination process. It has been postulated that polymeric G stretches present in the *vlsE* cassette region may promote the formation of secondary structures such as G-quadruplex within *vlsE,* which could help mediate recombination at the expression locus [35]. Attempts to mutate the *vlsE* locus (and thus identify cis-acting elements) have been unsuccessful to date.

Unlike most antigenic variation systems, *vlsE* recombination has not been detected *in vitro* [36] .This finding indicates that the activation of *vlsE* recombination may be triggered by conditions specific to the mammalian host. One of the goals of this project was to evaluate whether a tissue explant model system could provide a 'tissue-like' environment and thus induce a higher number of *vlsE* variants than in pure culture. If this

were the case, the tissue explant system would provide a good *in vitro* model system to study *vlsE* recombination, making the discovery of the trans-acting and cis-acting factors involved in *vlsE* recombination much easier.

In this study, next-generation sequencing was used to generate a large number of *vlsE* sequences. The availability of these data allowed us to thoroughly analyze the occurrence of *vlsE* recombination in mammalian cell-free (axenic) cultures, tissue explant co-cultures and infected mice. Besides examining the rate of recombination, we wanted to analyze individual recombination events to see if any patterns emerged. Patterns in *vlsE* variant sequences could lead to a better understanding of the nature and mechanism of *vlsE* recombination.

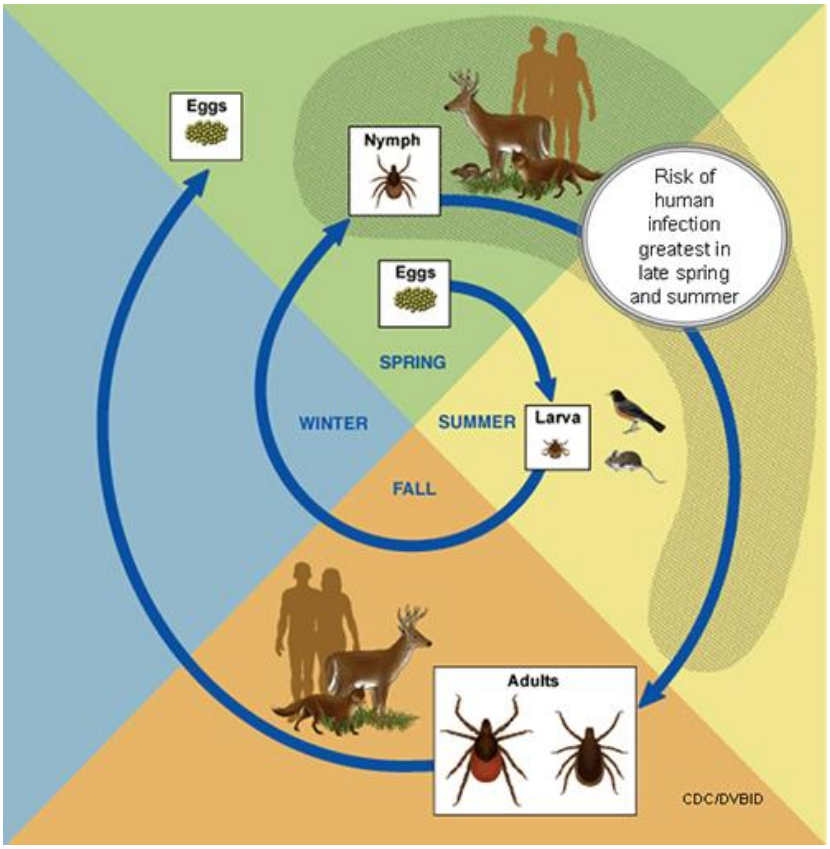


Figure 1.1 Life cycle of *I. scapularis* ticks and its relationship to Lyme disease transmission

Blood-fed females lay eggs in late spring. In the summer, the eggs hatch into larvae, which feed on *Borrelia*-infected small vertebrate hosts. In the spring of the following year, larvae emerge as nymphal ticks. Typically, from May through June, infected nymphal ticks feed and transmit *B. burgdorferi* to mammalian hosts. In the fall of year two, nymphs molt into adults. The females feed on large mammals, mate and lay eggs. Humans are accidental hosts, and are more likely to be infected by nymphs because ticks at this stage are smaller and, therefore, harder to detect and remove. Furthermore, adult ticks do not commonly feed on humans. (Figure from http://www.cdc.gov/ncidod/dvbid/lyme/ld_transmission.htm.)

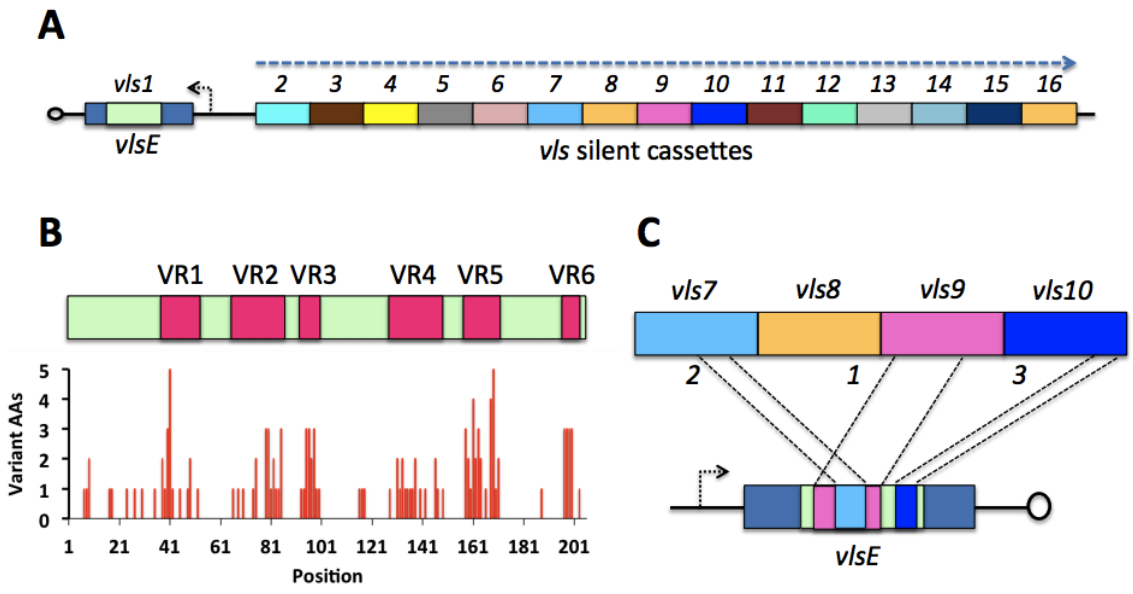


Figure 1.2 Characteristics of the *vls* locus and *vlsE* recombination [27]

A. *vlsE* is located on lp28-1, next to 15 silent cassettes that are highly homologous to each other and to the central cassette region of *vlsE*. **B.** Each of the 16 cassettes has 6 variable and 6 highly conserved invariant regions. The majority of variant or non-parental amino acid sequences are within the variable regions. **C.** This panel shows a hypothetical example of a homologous, unidirectional, random segmental recombination where *vls*9 first donates part its sequence to the central region of *vlsE*. Then *vls*7 donates its sequence, replacing part of the sequence that was donated by *vls9*. Lastly, *vls10* donates its sequence to *vlsE.*

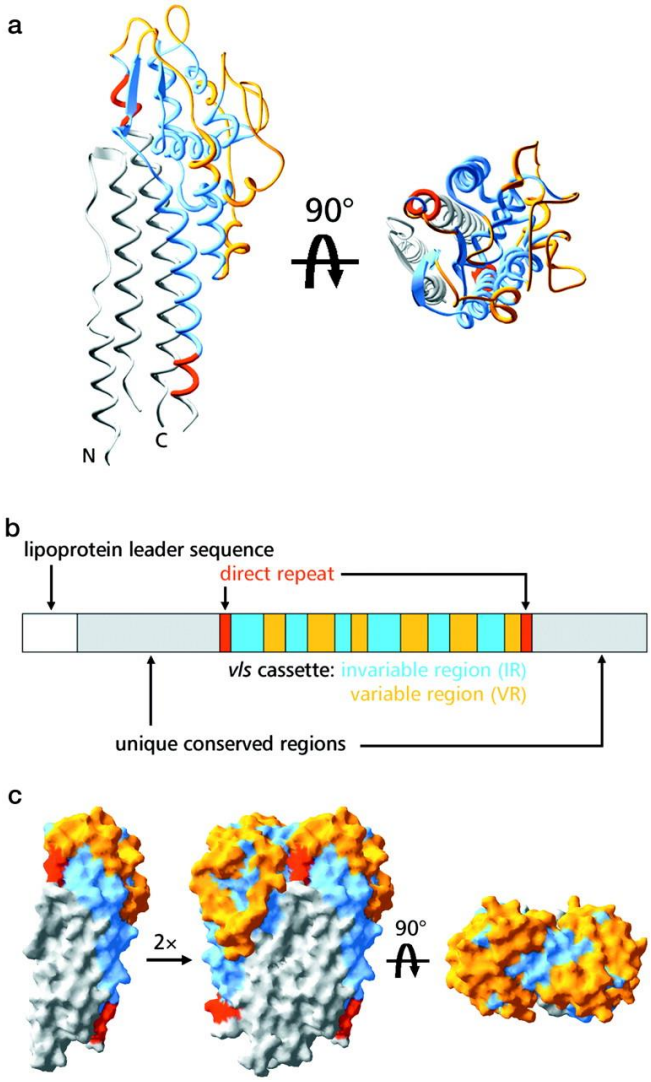


Figure 1.3 VlsE protein structure [29]

a. The conserved N- and C- terminal regions are colored in gray. The variable regions are located at the membrane distal surface and are colored in yellow. The 17-bp direct repeat regions are shown in red.

b. VlsE sequence consists of a lipoprotein leader sequence (white) and conserved N- and C-terminal regions (gray). *vlsE* and the 15 silent cassettes are flanked by 17-bp non-

conserved direct repeat sequence. Each cassette has 6 designated variable regions (yellow) and 6 invariable regions (blue).

c. This image shows the crystal structure of VlsE. The localization of the variable regions to the surface of the protein may act to protect the invariable regions from antibody interaction.

Chapter 2

Materials and methods

Culture conditions

B. burgdorferi was grown in Barbour-Stoenner-Kelly (BSK) II medium that was prepared in the laboratory [37]. The cultures were incubated at 37°C with 3% CO_2, and cell densities were counted under darkfield microscopy. Cultures were grown to late log phase or early stationary phase, which has a concentration of 10^7-10^8 cells per ml.

Bacterial strains

B. burgdorferi B31-5A4, a clone that contains all plasmids and has a 'parental' allele of *vlsE1*, is the bacterial strain that was used primarily in these studies. Strain B31 was initially derived by Alan Barbour through a series of limiting dilutions of an extract from ticks from Shelter Island, New York. The B31 strain had undergone three *in vitro* passages when received by our laboratory, and is maintained as stocks at -80°C. B31-5A4 was derived from B31 by two sequential colony isolations, and its plasmid content was determined by polymerase chain reaction (PCR) [12]. This clone has been used previously in animal infectivity studies, *in vitro* cultures and tissue explant experiments. 5A18NP1 is a B31 clone with the 'parental' *vlsE1* allele that lacks two plasmids (lp28-4 and lp56) and has been further modified by disruption of *BBE02* [38]. *BBE02* and *BBQ67* on lp56 encode type 2 restriction-modification enzymes, and their absence increases the transformation rate of *B. burgdorferi*. 5A18NP1 exhibits an increased transformation efficiency and maintains full infectivity in the mouse model [38]. 5A18NP1 was used in one of the mammalian cell-free (axenic) culture studies.

Tissue explant model system

Dr. Diane Edmondson, a fellow lab member, designed the tissue explant model system protocol. Small cubes (1-2 mm^3) of mouse skin, heart, bladder, liver and spleen

tissue were obtained aseptically from C3H/HeN mice. The explants were placed at the air-liquid interface on a hydrated collagen matrix (Gelfoam; Upjohn) in 6 well cluster dishes containing 1 ml BSK II medium per well. Explants were cultured at 37°C in air-3% CO_2 for 1 day prior to the addition of 5 x 10^5 *B. burgdorferi* (B31-5A4) per ml of medium. Starting on Day 1 of co-culture, the medium was removed and then replaced with fresh medium every 2-3 days. At different time points, explants were removed for fixation and staining. Genomic DNA was isolated from parallel samples.

Mouse infection studies

A total of 18 C3H/HeN immunocompetent female mice (6-8 weeks of age) were used for the animal studies. A low passage frozen stock sample (2337C) of B31-5A4 was selected and grown in BSK II medium to 10^7 organisms per ml. A volume of 0.1 ml containing 10^5 organisms was used to inoculate the mice. Groups of 4 mice were sacrificed on days 7, 14, 28, and 90-post infection (Figure 2.1). Five tissues (ear, joint, heart, bladder and skin) were obtained from each of the mice sacrificed. The tissue specimens were cultured for *Borrelia* in BSK II liquid medium with an antibiotic mixture (phosphomycin, rifampicin, amphotericin B) to inhibit the growth of other bacteria. These cultures were grown to 10^7 organisms per ml (approximately 7 days).

DNA preparation

All DNA extractions were done using the DNeasy Blood & Tissue kit (Qiagen). *B. burgdorferi* were cultured in 6 ml liquid BSK II medium and grown to late log phase (~10^7 organisms per ml). *B. burgdorferi* cultures were centrifuged for 15 min at 4700 x g, and the supernatant solution was discarded. Cells were resuspended in one ml of BSK II, vortexed and transferred to 1.5 ml microcentrifuge tubes. The cells were centrifuged at 13200 x g for

6 minutes at 4°C in an Eppendorf 5415R Centrifuge; the supernatant was discarded. The centrifuged pellets were resuspended in 180 µl of buffer ATL and 20 µl of proteinase K, which induces lysis of cells in order to release DNA, and were incubated overnight in a 56°C water bath. The resulting lysates were then subjected to DNA purification according to the manufacturer's instructions and eluted in 150 µl of Tris acetate-EDTA buffer. The concentration of the eluted DNA samples was determined using a NanoDrop spectrophotometer, and the samples were stored at -20°C.

Amplification of *vlsE* cassette

The *vlsE* cassette region was amplified by using *vlsE*-specific primers. Both the forward primer 4120 (5'-TAA GTA GTA CGA CGG GGA AAC CAG-3') and the reverse primer 4066 (5'-CTT TGC GAA CTG CAG ACT CAG CA -3') are located outside of the 17-bp direct repeat region at either end of the *vlsE* central cassette (Figure 1.3B). PCR reactions were performed using Phusion, a high-fidelity DNA polymerase. Five µl of PCR products were analyzed by gel electrophoresis on a 0.80% Agarose gel. Then, the PCR products were purified using Qiagen's PCR Purification kit.

PCR assay to detect *vlsE* recombination

To determine if *vlsE* recombination occurs within tissue explant co-cultures, a PCR assay consisting of two primer pairs was used. Figure 2.1 shows the two primer pairs used for the assay. The 'parental' primer pair has a forward primer sequence that is specific to the 'parental' *vlsE* sequence in variable region (VR) 5, and will amplify any *vlsE* sequence that has an unchanged VR5 region. The 'variant' primer amplifies *vlsE* sequences that have undergone variation in VR5 and match the most common *vls* silent cassette sequence in VR5. Note that the 'variant' primers will not be able to detect the presence of all variants,

specifically those with a VR5 region identical to that of the parental sequence. Both primer pairs have the same reverse primer, 4066, which is specific to the *vlsE* 3' region.

Quantification of parental and variant *vlsE* sequences using Quantitative PCR

A quantitative PCR (qPCR) protocol was developed to quantify the copy number of 'variant' and 'parental' *vlsE* sequence in mammalian cell-free (axenic) cultures, tissue explant co-cultures and infected mice DNA samples. All qPCR experiments were performed using Bio-Rad SYBR Green PCR Master Mix and a Bio-Rad CFX96 Touch Real-Time PCR Detection System. The same primer pairs used in the PCR assay to screen for *vlsE* recombinants (Figure 2.1) were used in the qPCR experiments. To quantify the number of non-variant and variant *vlsE* in a sample, qPCR reactions using both 'variant' and 'parental' primers were performed.

Pacific Biosciences sequencing

Pacific Biosciences (PacBio) methodologies and equipment were used to sequence *vlsE* PCR products obtained from mice, tissue explant and *in vitro* samples. We are grateful to Dr. George Weinstock and his team, who graciously performed the SMRTbell™ procedure (see below) and PacBio sequencing. PacBio uses single molecule real-time (SMRT) sequencing, which utilizes the inherent properties of the DNA polymerase as its sequencing engine. Phospho-linked nucleotides are used to visualize polymerase activity, with each of the four nucleotides (TAGC) attached to a different colored fluorescent label (Figure 2.2). These phospho-linked nucleotides carry their fluorescent label on the terminal phosphate rather than on the base, and DNA polymerase cleaves the fluorescent label as the nucleotides are being incorporated. Sequencing is performed in SMRT cells, each of which contains tens of thousands of zero-mode-wave-guides (ZMWs). Individual

DNA fragments are sequenced simultaneously in the ZMWs, an optical waveguide technology that guides light energy into a small volume, allowing the detection of nucleotides being incorporated by the DNA polymerase one at a time. Once a nucleotide is incorporated and the fluorescent label is cleaved off, the cleaved fluorescent molecule diffuses out of the ZMW detection volume, making the fluorescent signal undetectable.

For amplicon sequencing, SMRTbell™ hairpin adaptors are ligated on to both ends of each amplicon (Figure 2.3). The hairpin adaptors allow circular consensus sequencing (CCS), where the forward and reverse strand of the each amplicon can be sequenced multiple times. The shorter the amplicon, the higher number of repeated sequencing. Post sequencing, a consensus read is generated by averaging the base calls of the multiple reads. If performing multiplex sequencing, where multiple samples can be sequenced in parallel, a barcode sequence is added to the SMRTbell adaptor. For our experiment, a unique barcode was ligated to each of the 8 samples, and then the amplicons were pooled and sequenced. PacBio has a mean read length of 10-15 kb and an error rate of 14% per single read. The error rate can be reduced down to ~1% for consensus reads.

PacBio sequence data workflow

Initially, we attempted to align the raw consensus sequences against *vlsE* and *vls* cassettes using ClustalW. The output was aligned very poorly, most likely due to a combination of high homology among the cassettes, sequencing errors and three-nucleotide indels. After failing to align sequences with other software programs, I decided to manually align reads. Instead of attempting to manually align all reads (~30,000 per sample), we focused on aligning variant reads. First, I filtered out sequences with read lengths less than 601 bp or greater than 680 bp. This read length range was chosen

because most likely anything less than 601 bp would be incomplete reads and reads longer than 680 bp represented incorrect consensus assemblies and therefore would have too many errors.

To identify variant sequences in the mammalian cell-free *in vitro* cultures and tissue explant samples, BLAST searches using each PacBio sequence as the query sequence and the invariant-variant regions (e.g. IR1-VR1) of *vlsE1* and all of the *vls* silent cassettes as the subject sequences were performed to identify PacBio sequences that contained regions matching silent cassettes. All sequences that had a higher percent identity match with an IR-VR sequence of one of the 15 silent cassettes was noted as a potential variant sequence. All of the potential variant sequences were then manually aligned against pre-aligned *vlsE1* and the 15 *vls* cassette sequences.

During the manual alignment process, the base positions of the deletion and insertion errors were recorded and single base indel errors were manually corrected. For ambiguous deletion errors, where different nucleotides are present in the cassettes, an IUPAC code was used to represent the possible choices. Also, the most likely base was determined by looking at the nucleotides and neighboring regions of *vlsE* and the 15 silent cassettes around position as the error.

Automation of alignment and error correction

We collaborated with David Volk, Ph.D. to automate the alignment and error correction process. The Aptaligner software utilizes the FASTQ file, a format that combines the data of raw reads with quality information, as the input file. The program creates a library that consists of all the codons in the *vlsE* parental sequence and the silent cassettes. The raw reads were aligned simultaneously against *vlsE* codon sequence and all possible

silent cassette codons using Hidden-Markov-Modeling (HMM). HMM optimizes the overall read alignment by generating alignment scores for multiple alignment possibilities. Once a possible alignment and its score are generated, the aligned sequence can only be replaced with another aligned sequence of a higher score. The program also counts the number of copies of identical reads and removes the non-unique sequences. Post-alignment, a realignment script condenses equivalent alignments to optimize the subsequent error correction script. Then error corrections are made by comparing codons at each position of the read to the library of cassette codons. The program is designed to record the position and the correction made. After the correction step, the aligned sequences are readjusted to fix any misaligned regions. The realignment and correction steps of this program are currently being optimized.

Analysis of aligned reads

Recombination events in aligned variant sequences were visualized using a Visual Basic macros program [29] developed by Dr. Loic Coutte, a former lab member. This program allowed us to visualize all potential *vls* cassette donors. It also distinguishes between a maximal and minimal recombination event. Each cassette is represented by a different color. Solid color regions indicate regions of the variant sequence that match that cassette (but not the original *vlsE1* sequence). The hatched color indicates that the variant sequence matches both the *vlsE1* and the cassette. A possible **minimum recombination event** represents the largest area that matches a silent cassette region but not the parental *vlsE1* sequence. In the program, this region begins and ends with solid color regions, with hatched regions in all intervening spaces. A potential **maximal recombination event** also includes flanking regions

that match both the silent cassette and *vlsE1*. This region is the same as the minimal recombination event plus the flanking hatched sequences.

	IR5					VR5	
Cassette	**151**	**152**	**153**	**154**	**155**	**156**	**157**
Vlse1	**GCT**	**GCT**	**GCT**	**ATT**	**GGG**	**GAT**	**AAA**
Vsl2	**GCT**	**GCT**	**GCT**	**ATT**	**GGG**	**AAG**	**GGT**
Vls3	**GCT**	**GCT**	**GCT**	**ATT**	**GGG**	**AAG**	**GGT**
Vls4	**GCT**	**GCT**	**GCT**	**ATT**	**GGG**	**AAG**	**GGT**
Vls5	**GCT**	**GCT**	**GCT**	**ATT**	**GGG**	**AAG**	**GGT**
Vls6	**GCT**	**GCT**	**GCT**	**ATT**	**GGG**	**GAT**	**AAA**
Vls7	**GCT**	**GCT**	**GCT**	**ATT**	**GGG**	**AAG**	**GGT**
Vls8	**GCT**	**GCT**	**GCT**	**ATT**	**GGG**	**AAG**	**GGT**
Vls9	**GCT**	**GCT**	**GCT**	**ATT**	**GGG**	**AAT**	**AAA**
Vls10	**GCT**	**GCT**	**GCT**	**ATT**	**GGG**	**AAG**	**GGT**
Vls11	**GCT**	**GCT**	**GCT**	**ATT**	**GGG**	**AAG**	**GGT**
Vls12	**GCT**	**GCT**	**GCT**	**ATT**	**GGG**	**AAG**	**GGT**
Vls13	**GCT**	**GCT**	**GCT**	**ATT**	**GGG**	**AAG**	**GGT**
Vls14	**GCT**	**GCT**	**GCT**	**ATT**	**GGG**	**AAG**	**GGT**
Vls15	**GCT**	**GCT**	**GCT**	**ATT**	**GGG**	**GAT**	**AAA**
Vls16	GCT	GCT	GCT	ATT	GGA	AAG	GGT

The 'Parental' primer is indicated in red for vlse1, vls6, vls9 and vls15. The 'Recombined' primer is shown in green for vls2-vls5, vls7-vls8, and vls10-vls15. Vls16 does not fit either primer due to codon 155.

Figure. 2.1 'Parental' and 'variant' primer sequence used to detect presence of *vlsE* recombinants in both PCR and qPCR assay.

The two primer pairs used to detect parental and variant *vlsE* sequences, were specific to portions of the IR5 and VR5 region. This region was chosen because 12 out of the 15 silent cassettes had the seqeuence "AAGGGT" region in the beginning of their VR5 region. The 'parental' *vlsE* has the sequence of "GATAAA" in the beginning of its VR5 region. This difference allows the detection of most of the *vlsE* variants that have undergone a change in the VR5 region.

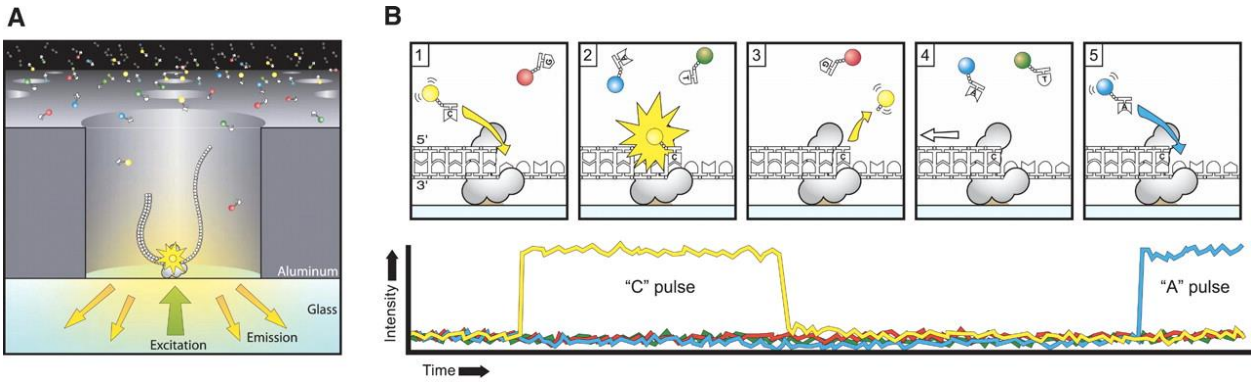


Figure 2.2 PacBio single molecule real-time (SMRT) sequencing [39]

(A) PacBio sequencing occurs within SMRT cells, each of which contains tens of thousands of cylindrical zero-mode-waveguide (ZMW) units. Each ZMW has a single DNA polymerase, which is fixed to the bottom, and a single DNA fragment.

(B) When the DNA polymerase incorporates a nucleotide, a laser light is passed through the bottom of the ZMW cell, which allows the camera to detect the fluorescent signal located the terminal phosphate of the dNTP. The intensity and the color of the signal are recorded in real-time by the PacBio instrument. As the next nucleotide is incorporated, the fluorescent label from the previous nucleotide is cleaved off and diffuses out of the cell.

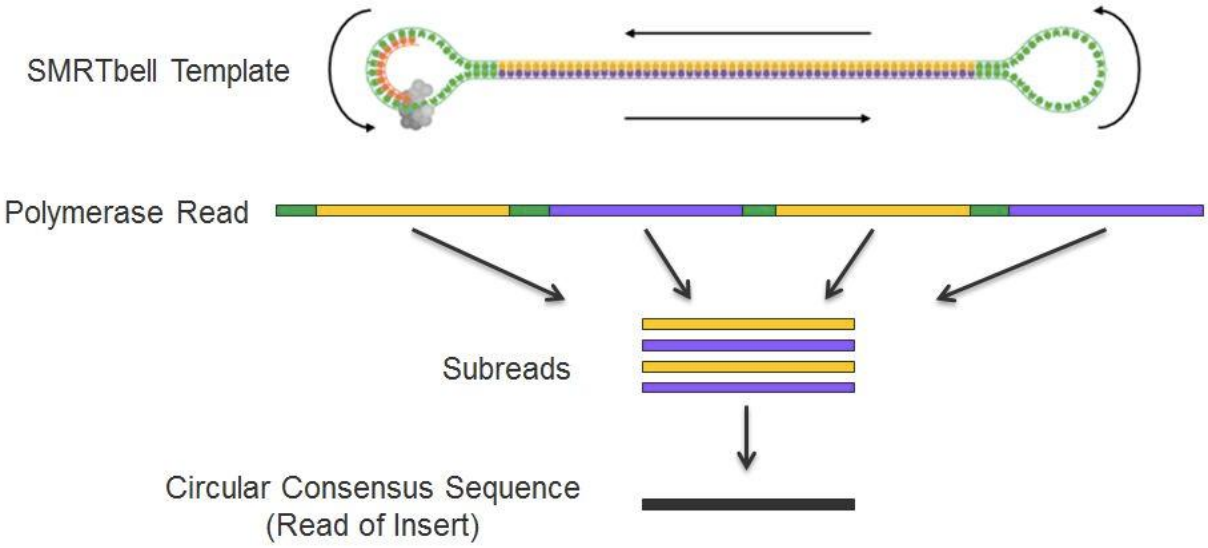


Figure 2.3 SMRTbell amplicon sequencing

In PacBio amplicon sequencing, SMRTbell hairpin adaptors (green) are ligated on to the DNA fragment (yellow-purple). The SMRTbell dumbbell adapters allows the DNA polymerase to sequence the DNA fragment in a circular motion, which in turn results in multiple reads of the forward (yellow) and reverse (purple) strand of the amplicon. These multiple forward and reverse subreads and are used to generate a circular consensus sequence (black). (Figure from http://www.pacb.com/smrt-science/smrt-sequencing/single-molecule-resolution/)

Chapter 3

Results

qPCR results

Standardization of qPCR

The purpose of developing a qPCR assay was to use a sensitive method that would accurately quantitate the number of variants in different samples and allow us find the rate of *vlsE* recombination in *in vitro*, tissue explant and animal infection samples. To standardize a qPCR protocol, the efficacy of three different primer lengths was tested. The primer length of 21 bp was chosen because it produced standard curves with the highest efficiency and R^2 value (Figure 3.1). It also had the cleanest single peak melt curves (no small shoulder peak that suggests the presence of secondary product). The qPCR instrument's thermal gradient feature was used find the optimal annealing temperature.

Serial dilution of 5A4 (parental *vlsE* clone) PCR product was used to generate a standard curve for quantifying non-recombinant *vlsE*. The PCR product of a variant clone, 1396D, was used to construct standard curve for quantifying recombinant *vlsE.* 1:10 serial dilutions between 0.5 ng/µl and 5 x 10^{-10} ng/µl were done for 5A4 and 1396d PCR products. qPCR results yielded highly reproducible standard curves with high squared correlation values (P21 R^2=0.993, V21 R^2=0.979) (Figure 3.1). It is important to note that we were able to detect real variants in mammalian cell-free *in vitro* culture samples using PacBio sequencing (Table 4). Therefore the PCR products from sample 2337C used to construct the standard curve to quantitate copies of parental *vlsE* will contain a small proportion of variant sequences. This might lead to a less precise calculation of the copies of parental *vlsE* in samples.

C_T values represent the cycle number at which the amplification fluorescent signal is above the background fluorescence, i.e. the number of cycles it took to detect a real signal

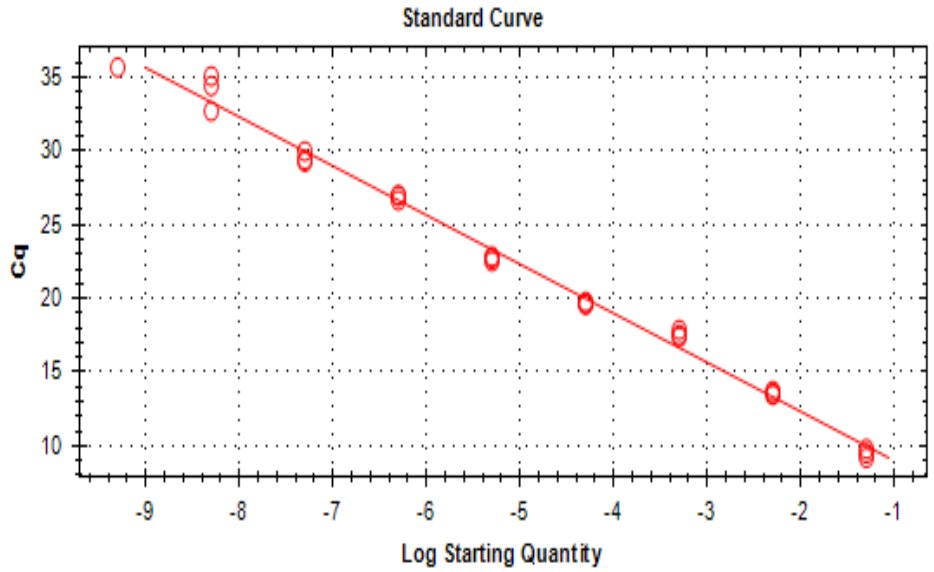
from a sample. The template DNA concentration for all samples was standardized to 5 ng. The C_T values represent the average of 3 technical replicates. Sample 2337C with P21 (parental primer) resulted in a C_T value of 23.04 and a higher C_T value of 27.77 with primer V21 (variant primer) (Table 1). As expected, the reverse was true for the variant clone sample 1396D, which had a C_T value of 31.95 with primer P21 and a lower C_T value of 18.91 with primer V21. TE47, tissue explant co-culture sample, had C_T value of 24.10 with primer P21 and C_T value of 28.28 with primer V21.

qPCR reactions with template DNA from 7 (D7M1B), 14 (D14M1B) and 21 (D21M1B) days post infection mouse bladder tissue were also performed. The results showed that as the days post infection increased, the C_T value increased with P21, and the decreased C_T values with primer V21. These trends indicate that the proportion of spirochetes with parental *vlsE* sequence decrease in mouse infection sample between 7 and 21 day post infection (Table 1).

Absolute quantification

For absolute quantification, the P21 standard curve was used to quantitate the copies of parental *vlsE* and the V21 standard curve was used to quantitate the copies of variant *vlsE* in a sample. First, the standard curve linear equation and C_T values were used to find the quantity in nanograms (ng) and then converted to copy number. The calculated numbers of parental *vlsE* sequences and variant sequences detected using the V5 region primers are provided in Table 2. As expected, 2337C had the highest number of parental *vlsE* sequences and 1396D had the highest number of variant *vlsE* sequences detected. Among the mouse infection samples, the copies of variant *vlsE* increased while the copies of parental *vlsE* decreased over the days post infection.

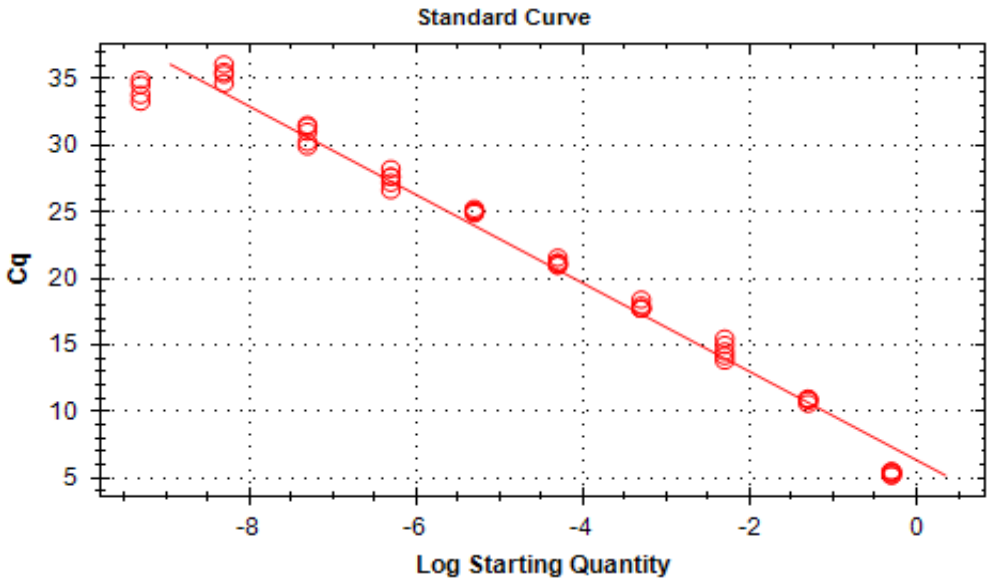
(A) **P21**



E=99.6% R^2=0.993 Slope=-3.332 y-int=5.652

Standard SYBR

(B) **V21**



E=99.8% R^2=0.979 Slope=-3.328 y-int=6.326

Standard SYBR

Figure 3.1 Standard curves for absolute quantification using parental (P21) and variant (V21) primers. The standard curve in panel **(A)** was generated with *vlsE* PCR products from sample 2337C and P21 primer will be used to quantitate the copy number of "parental" *vlsE* in a sample. While the standard curve shown in panel **(B)** was generated using *vlsE* PCR product from sample 1396D and primer V21 will be used to calculate the copies of variant *vlsE* in samples.

Table 1. C_T values with 'parental' primer P21 and variant primer V21.

Sample name	C_T Primer P21	C_T Primer V21
2337C	22.72 ± 0.089	28.41 ± 0.216
1396D	33.6 ± 0.134	21.89 ± 0.008
TE47	24.10 ± 0.81	28.28 ± 0.147
D7M1B	23.44 ± 0.028	26.91 ± 0.269
D14M1B	25.38 ± 0.076	24.47 ± 0.057
D28M1B	24.68 ± 0.128	23.45 ± 0.061

Table 2. QPCR as a measure of *vlsE* recombination. Calculated copy numbers of parental and variant *vlsE* Variable Region 5 (VR5) sequences were determined for in vitro cultures with predominant parental (2337C) and variant (1396D) *vlsE* sequences, a tissue explant culture (TE47), and in mouse bladder tissue 7, 14, and 28 days post inoculation with *B. burgdorferi* B31, clone 2337C (D7M1B, D14M1B, and D28M1B, respectively).

Sample name	**Copies with parental *vlsE* VR5 sequences**	**Copies with variant *vlsE* VR5 sequences**	**Proportion with variant VR5 sequences (%)**
Parental clone (2337C)	$3.49x10^4$	$1.07x10^3$	2.97
Variant clone (1396D)	$1.89x10^1$	$9.77x10^4$	99.98
Tissue explant (TE47)	$1.35x10^4$	$1.17x10^3$	7.97
D7M1B	$2.13x10^4$	$3.02x10^3$	12.42
D14M1B	$2.52x10^4$	$1.64x10^4$	39.42
D28M1B	$9.03x10^3$	$3.31x10^4$	78.57

Pacbio Results:

High-throughput sequencing was the second method used to quantitate the rate of *vlsE* recombination in samples. NGS sequencing generates large amount of data, which would allow us to accurately quantitate the number of variants, specifically those that occur at very low rate (mammalian cell-free samples). Moreover, it allowed us to study individual recombination events within variant sequences. PacBio sequencing was selected as the sequencing platform because it has a long read length (10 kb - 15 kb), and low consensus read error rate (~1%). 454 and Illumina have a comparable error rate but have shorter amplicon read lengths (~ 300-600 bp) and therefore would not be able to sequence the entire *vlsE* cassette. Moreover, 454 and Illumina synthesize DNA from clonally amplified templates, which may result in sequencing bias. We filtered out reads less than 601 bp to eliminate reads that were shorter than the expected length of the *vlsE* cassette plus the upstream and downstream regions (Table 3). Reads greater than 680 bp were also filtered out to remove low quality reads that most likely resulted from misalignment of SMRTbell subreads.

Error analysis

The SMRTbell approach provides multiple reads of both the forward and reverse strands of each individual DNA molecule, and these sequences are then combined into a consensus sequence. This redundancy greatly reduces, but does not eliminate, sequence errors inherent to the PacBio sequencing process. To provide an initial assessment of the presence of sequence errors, preliminary alignments of sequences from the 2337C sample with the reference parental sequence (GenBank U76405) was performed. Because prior Sanger sequencing of a 2337C *vlsE* PCR product matched U76405, it was expected that

most, if not all, of the PacBio sequences would have the same sequence. However, it was found that many of the reads had one or more sequence differences. Therefore a more thorough analysis of the occurrence of sequence errors was done.

In this analysis, 127 reads from the 2337C sample were manually aligned with U76405 and the 15 silent cassettes using the program BioEdit to determine the frequency and locations of sequence differences. The total number of sequences with errors at each position in aligned read was graphed. Error analysis of 127 reads from sample 2337C showed that the positions with the highest frequency of errors were deletions of guanine within polymeric G-runs (Figure 3.2). The positions that had errors in 20 sequences or more are labeled above the corresponding bar. The locations of these common sequence errors were consistent in the error analysis of manually aligned reads for the other samples. In order to have a systematic approach when manually aligning sequences, the guanine deletion errors within homopolymer runs were always assigned to the last G in the stretch instead of being equally distributed among all the Gs. Therefore, the number of errors at these terminal G positions are overstated while the number of errors of the non-terminal Gs in homopolymer runs are underreported. These results show that PacBio sequencing errors in homopolymers are not random and instead occur at a higher rate compared to errors at other positions. Fortunately, errors in homopolymer G runs are easily fixed and hence the majority of the sequencing errors can be fixed unambiguously.

The error rate was calculated by dividing the total number of errors found in a sequence by the read length of the aligned sequence (609 bp). To find the error rate of a sample, the number of sequence differences of individual reads was averaged. In general, errors could be distinguished easily from recombination events; nearly all were non-

templated single base insertions or deletions, which were not observed previously in the analysis of 1,496 clones by Sanger sequencing [29]. The error rate for manually aligned reads ranged from 0.74% to3.2% (Table 4). The error rate could be related to the freshness of the samples. The DNA from mouse infection samples was derived from spirochetes grown out from tissue in BSKII medium, while the remaining samples were from stored frozen cultures or DNA. This could explain why D7M1B reads had the lowest error rate (Table 4).

Because most of the sequences contained errors, it was important to do an error analysis that showed the sequence changes are a result of recombination events. First, the mismatch, deletion and insertion error rate was calculated at several non-variant codon positions (same codon sequence across *vlsE* and 15 silent cassettes). For example, the codon at position 49 (Figure 3.4B) is a non-variant codon that has 'AGT' across all 16 cassettes. The mismatch error rate represents the inherent error rate per base present in the sequences. Sample 2337C had 0.023% mismatch error rate, 0.40% deletion error rate, and 0.20% single insertion error rate. Given the mismatch error rate of 0.023%, the odds of a *vlsE* sequence without any mismatch errors is 0.977. I then compared the mismatch error rate against the rate of change of bases at variant codon positions from a *vlsE* sequence to one of the 15 *vls* sequences. For instance, for sample 2337C variants, the rate of change at codon 38 was 11%. 11% is much higher than the calculated mismatch error rate of 0.023%, therefore the sequences with changes at codon 38 are real changes and not result of errors. Overall, this error analysis proves that the probability of having random base changes is extremely low and therefore most of the changes in the variant sequences result from recombination event(s).

FASTQ analysis

In order to evaluate the quality of PacBio consensus sequences, FASTQ files of each sample were analyzed. FASTQ files contain the quality scores associated with each base call for every read. The quality scores were represented as ASCII characters and the ASCII table was used to convert the ASCII characters into a phred or Q score. Essentially, the quality scores represent the probability of a base being called incorrectly. This probability is calculated using the formula $P = 10^{-Q/10}$, where P is the error probability and Q is the quality score.

The majority of the base calls had the ASCII character of K (corresponding to the 11^{th} highest quality score and P_error= 0.07943). This was true for FASTQ files for all the samples. In order to evaluate the quality of individual reads, a quality ratio was generated for every sequence. The quality ratio was calculated by dividing the total number of bases with the ASCII score of K or higher by the total read length. Figure 3.3 shows that there is a negative correlation between the quality ratio and the total number of errors in a read. As the total number of errors in a sequence increases, the quality ratio decreases. The correlation between these two variables indicates that the quality ratio could be used to accurately assess the quality of a read and ultimately help filter out low quality sequences.

Detection of real recombination events

Sequences that appeared to contain recombination events were identified by local BLAST searches of each read against segments of the *vlsE1* cassette region and the 15 *vls* silent cassettes (see Materials and Methods). Those sequences that contained regions that had higher sequence identity with a silent cassette segment than with the parental *vlsE* sequence were considered to have potential recombination events. These were further

verified through manual alignments. Because of their relatively small number, all of the potential variants could be analyzed in the 2337C, 5A18NP1, and the infected tissue explant TE47, TE49, TE50 samples. For day 7 mouse bladder samples, which contained a high number of variants, the first 200 sequences were analyzed by manual alignment. The percentage of variants was 0.84% to 1.18% in pure *in vitro* cultures derived from parent strain B31, and 0.79% to 1.22% in the tissue explant co-culture (Table 4). In contrast, 36% to 57% of the first 200 sequences from day 7 mouse bladder samples were variants. Within the population of variants the number of variants with siblings (identical variant sequences within the same sample) were counted. 11 to 63 variants with siblings were found in the 4 samples analyzed (Table 4). Thus most (67 to 91 percent) of the variant sequences identified in this analysis were unique.

After manually aligning potential variant reads, the reads were entered into a Visual Basic macros program. This program visually displays all potential silent cassette donors that match the non-parental sequence of variant sequences. The length of recombination events ranged from a single nucleotide change to almost the entire cassette. Figure 3.4 shows the Visual Basic macros program output for one variant read from sample 2337C. Figure 3.4A shows the overall pattern of recombination in read 2337C_139710, whereas Figure 3.4B provides a close-up of the variant region. The different colored bars represent silent cassettes 2-15. The solid color represents where the variant sequence matches to a silent cassette, while the surrounding lighter hatched colored region represents where the variant sequence matches both the silent cassette and *vlsE*1. Therefore, the maximal recombination event is inclusive of the surrounding lighter colored regions, while the

minimal recombination event starts at the first change that is unique to a silent cassette (solid color) and ends at the last change that is unique to the same silent cassette.

In many cases, a single *vls* cassette could be deduced as the donor to a recombination event. For example, read 2337C_139710 has non-'parental' sequence within its VR1 region (Figure 3.4A). The putative cassette donor for this variant sequence is *vls3* because it encompasses all the non-'parental' sequence of the aligned variant read in one consecutive segment. The minimal recombination event extends from codon 39 to 42, while the maximal recombination event extends from codon 9 to 68. Figure 3.5 shows the schematic representation of three additional variants from sample 2337C. Read 2337C_42395 has a single well-defined recombination event with a minimal recombination event length of a single nucleotide (Figure 3.5A). The putative cassette donor is *vls*2. Read 2337C_103644 also has a single well-defined recombination event; *vls3* is the putative silent cassette donor and it has a 486-nucleotide long minimal recombination event. In contrast, read 2337C_101310 has several intermittent sequence differences that can be considered multiple recombination events (Figure 3.5B). Furthermore, most of the changes in 2337C_101310 cannot be attributed to a single silent cassette, due to the redundancy that occurs within regions of the silent cassettes.

Analysis of manually aligned variant reads

Variants reads were translated to their protein sequence and then compared against the 'parental' *vlsE* sequence. The sequences with a non-parental amino acid at each position were totaled. Results showed that, as expected, the majority of the non-parental amino acids, or recombination events in variant sequences were within the variable regions

(Figure 3.6). This is consistent with prior studies that also showed that most sequence changes are concentrated in the variable regions [29].

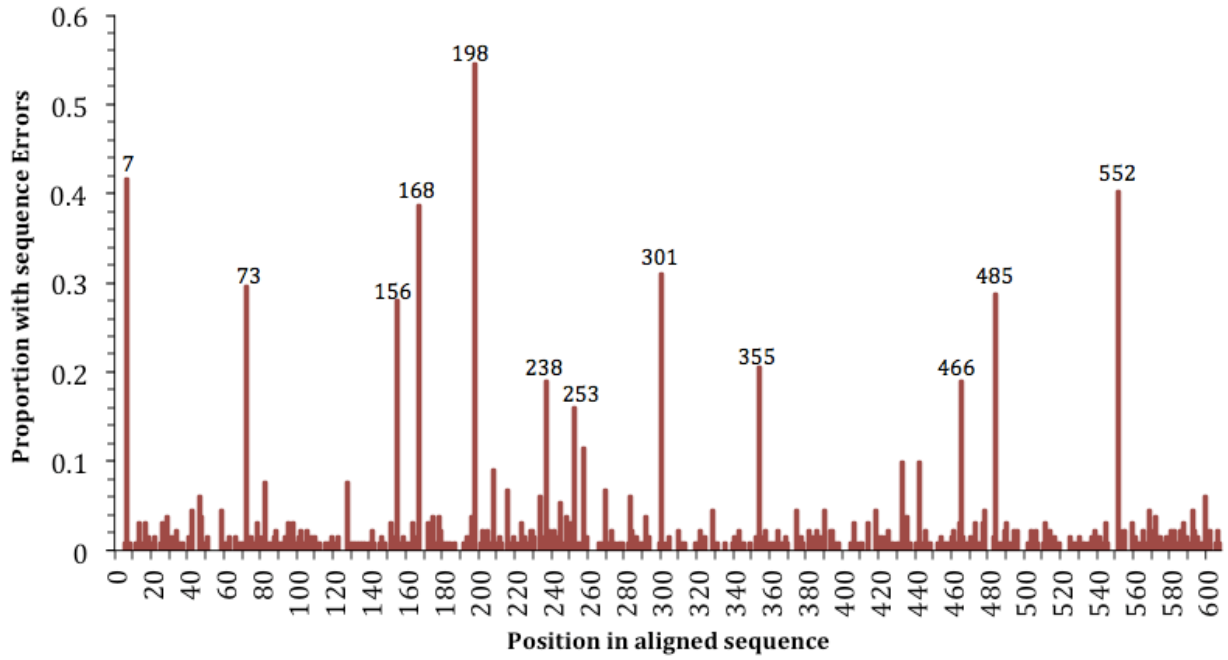
To study the length of individual recombination events, variant sequences with a single well-defined recombination event were further examined. The length of the single recombination event and the putative silent cassette donor was recorded for all variant sequences. The average length of a minimal recombination event of variants from sample 2337C was 48 codons, 33 codons for sample 5A18NP1, 3 codons for sample TE47, and 7 codons for sample D7M1B. The longest length of a minimal recombination event was 172 codons (the entire length of *vlsE* is 203 codons) for variants with single well defined recombination events from both mammalian-cell free *in vitro* samples: 2337C and 5A18NP1. In contrast, the longest minimal recombination event was 35 codons long for both TE47 and D7M1B (Figure 3.7A). The results indicate that variants from mammalian cell-free (axenic) samples had a higher frequency of longer minimal recombination events than the variants from tissue explant co-cultures and infected mice samples (Figure 3.7A).

Figure 3.7B shows the usage of silent cassettes in variant sequences with single recombination events. While most sequences changes in aligned variant reads with a single recombination event could be attributed to one of the 15 silent cassettes, some of the sequences had template-independent changes, or apparent point mutations. It is unclear whether these all these sequence differences represent a real genetic change or a sequencing error.

Table 3. Total number of reads obtained and number of reads that had a read length of ≥ 601 bp and ≤ 680 bp.

Sample name	Total number of reads	Number of reads ≥ 601 bp and ≤ 680 bp
Parental clone (2337C)	35,699	15,184
Parental clone (5A18NP1)	29,259	12,136
Tissue explant (TE47)	29,536	12,762
Tissue explant (TE49)	23,913	11,076
Tissue explant (TE50)	31,861	14,025
D7M1B	24,042	10,036
D7M2B	20,601	8,393
D7M3B	28,981	10,875

(A)



(B)

>*vlsE*1

GAGGGGCTATTAAGGAAGTTAGCGAGTTGTTGGATAAGCTGGTAAAAGCTGTAAAGACAGCTG

AGGGGCTTCAAGTGGTACTGCTGCAATTGGAGAAGTTGTGGCTGATGCTGATGCTGCAAAGGT

TGCTGATAAGGCGAGTGTGAAGGGGATTGCTAAGGGGATAAAGGAGATTGTTGAAGCTGCTGGG

GGGAGTGAAAAGCTGAAAGCTGTTGCTGCTGCTAAAGGGGAGAATAATAAAGGGGCAGGGAAGT

TGTTTGGGAAGGCTGGTGCTGCTGCTCATGGGGACAGTGAGGCTGCTAGCAAGGCGGCTGGTGCT

GTTAGTGCTGTTAGTGGGGAGCAGATATTAAGTGCGATTGTTACGGCTGCTGATGCGGCTGAGC

AGGATGGAAAGAAGCCTGAGGAGGCTAAAAATCCGATTGCTGCTGCTATTGGGGATAAAGATGG

GGGTGCGGAGTTTGGTCAGGATGAGATGAAGAAGGATGATCAGATTGCTGCTGCTATTGCTTTG

AGGGGGATGGCTAAGGATGGAAAGTTTGCTGTGAAGGATGGTGAGAAAGAGAAGGCT

Figure 3.2 Nucleotide positions with the highest number of errors were deletions of guanines in homopolymer G-runs.

(A) This error analysis graph represents the 127 variants sequences from sample 2337C. The total number of errors was calculated at each position in the aligned sequence. The positions with highest number of errors were homopolymer G-runs. The bars that have position numbers above the peak represent these positions. For polymeric G-runs, the deletion error was always assigned to the last G base in the run, and not rationed equally across all of the Gs. Therefore, the error rate for the last G is overstated and under-reported for the other Gs in the G-run.

(B) The homopolymer G-runs within *vlsE1* sequence are underlined and positions with highest number of errors are in red. *vlsE* is highly GC rich and has a total of 12 guanine homopolymer runs. PacBio sequencing appears to have a sequencing error bias in homopolymer runs.

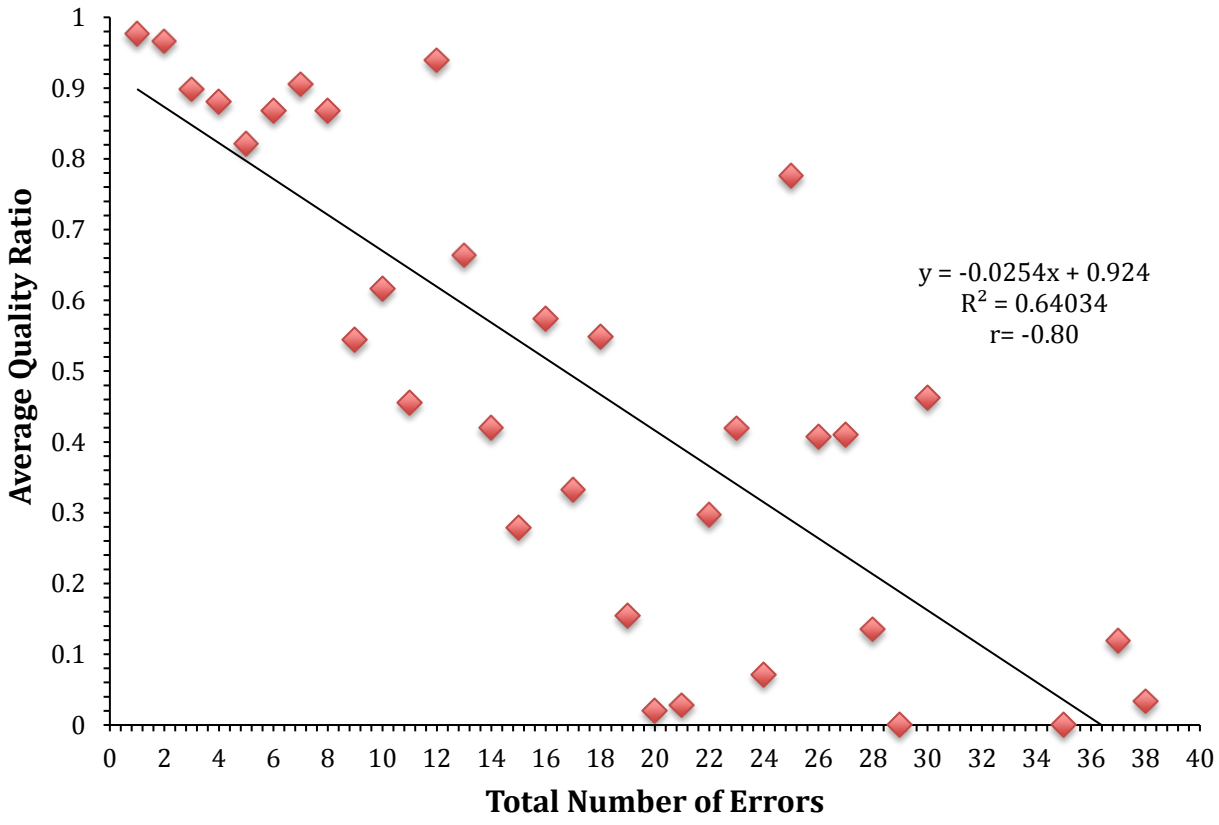


Figure 3.3 There is an inverse correlation between the quality score and total number of errors.

This graph analyzes variant sequences from sample 2337C. The quality ratio was calculated by dividing the total number of base calls with ASCII quality score of K or greater by the total read length. The quality ratios for sequences with the same number of errors were averaged. There is a negative correlation between the number of errors and the average quality ratio. Sequences with higher numbers of errors tend to have a lower quality ratio, while sequences with low numbers of errors tend to have a higher quality ratio. This correlation suggests that FASTQ will be useful for identification and filtering of low quality reads.

Table 4. Percent variants detected in samples and the calculated error rate. The error rate of each sample was calculated by averaging the error rate of individual variant reads. The error rate of a single sequence is the total number of errors divided by the read length. Sibling sequences are sequences that have the same *vlsE* variant sequence.
*Number of reads from TE49 and TE50 samples were additionally filtered for high quality scores.

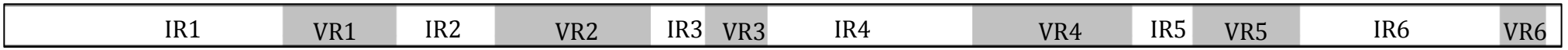
Sample name	Number of reads ≥ 601 and ≤ 680	Number of variants	Percent of variants	Error rate	Number of sequences with siblings
2337C	15,184	127	0.84%	1.58%	14
5A18NP1	12,136	143	1.18%	2.5%	14
TE47	12,762	102	0.80%	3.2%	11
TE49	11,076	127	1.02%	0.89%	63
TE50	14,025	126	1.22%	0.74%	57
D7M1B-1st 200 reads	200 (1st 200 reads ≥ 601 and ≤ 680)	114	57%	0.96%	17
D7M2B-1st 200 reads	200 (1st 200 reads ≥ 601 and ≤ 680)	72	36%	0.92%	21
D7M3B-1st 200 reads	200 (1st 200 reads ≥ 601 and ≤ 680)	105	52.5%	1.01%	14

Table 5. Chi-square test using *in vitro* and tissue explant samples
*The Day 7 Mouse samples are excluded from this analysis.

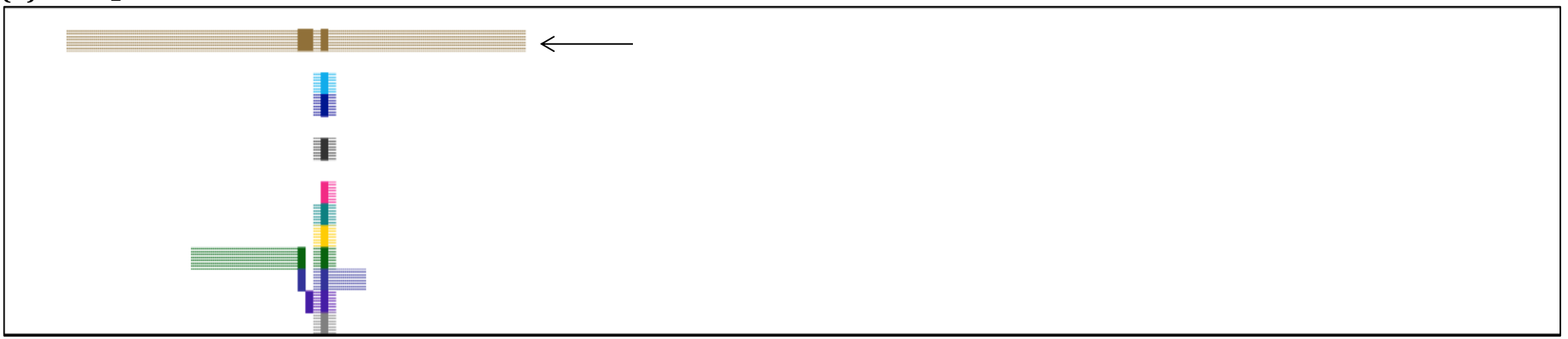
Item	Group-1 (*in vitro* culture)	Group-2 (tissue explant)	Total
Observed numbers (O)	0.99%	0.94%	1.93E-02
Expected numbers (E)	1.0%	1.01%	2.02E-02
O - E	-0.02%	-0.08%	-9.46E-04
$(O-E)^2$	3.64E-08	5.70E-07	6.06E-07
(O-E)2 / E	3.61E-06	5.63E-05	5.99E-05

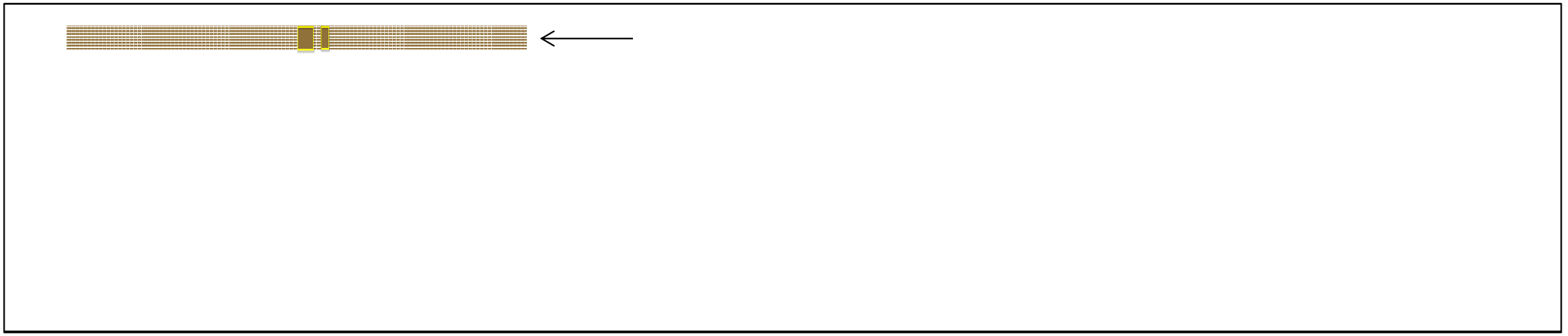
Critical Chi-square statistic for DF = # of groups -1= 2-1 =1 and 5% significance level

Conclusion:
As the calculated chi-square statistic (5.99E-05) is less than the Critical Chi-square statistic (3.84), the Null Hypothesis can't be rejected.
Thus, it can be concluded that the number of variants found tissue explant samples (Group-2) is not statistically different the number of variants found in in vitro culture samples (Group-1).



(A) 2337C_139710





(B)

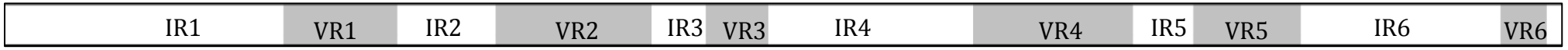
VR1

	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51
1	GCT	GAT	GCT	GAT	GCT	GCA	AAG	GTT	GCT	GAT	AAG	GCG	AGT	GTG	AAG
2	OOO	OOO	OOO	OOO	OOO	OOO	OOO	OOO	OOO	OOO	OOO	OOO	OOO	OOO	OOO
3	GCT	GAT	GAT	AAT	GCT	GCG	AAG	GTT	GCT	GAT	AAG	GCG	AGT	GTG	AAG
4	GCT	AAT	GCT	GGT	GCT	GCA	AAG	GTT	GCT	GAT	AAG	GCG	AGT	GTG	ACG
5	GAT	AAT	OOO	OOO	GCT	GCG	AAG	GCT	GCT	GAT	AAG	GCG	AGT	GTG	ACG
6	GAT	AAT	OOO	OOO	GCT	GCG	AAG	GCT	GCT	GAT	AAG	GAT	AGT	GTG	ACG
7	GCT	GAT	GCT	OOO	GCT	OOO	AAG	GTT	GCT	GAT	AAG	GCG	AGT	GTG	ACG
8	GAT	AAT	GCT	GGT	GCT	GCG	AAG	GCT	GCT	GAT	AAG	GAT	AGT	GTG	AAG
9	GAT	AAT	OOO	GAT	GCT	OOO	AAG	GTT	GCT	GAT	AAG	GCG	AGT	GTG	ACG
10	GAT	AAT	OOO	GAT	OOO	GCG	AAG	GCT	GCT	GAT	AAG	GCG	AGT	GTG	ACG
11	GCT	AAT	GCT	GGT	GCT	GCG	AAG	GCT	GCT	GAT	AAG	GCG	AGT	GTG	ACG
12	GCT	GAT	OOO	OOO	GCT	GCG	AAG	GCT	GCT	GAT	AAG	GAT	AGT	GTG	AAG
13	GCT	GAT	GAT	GCT	GCT	GCG	AAG	GCT	GCT	GAT	AAG	GAG	AGT	GTG	AAG
14	GCT	AAT	GAT	GCT	GCT	GCG	AAG	GTT	GCT	GAT	AAG	GAG	AGT	GTG	ACG
15	GAT	AAT	AAT	AAT	GCT	GCG	AAG	GCT	GCT	GAT	AAG	GCG	AGT	GTG	ACG
16	GCT	GAT	AAT	AGT	GCT	GCG	AAG	GCT	GCT	GAT	GAG	GCG	AGT	GTG	ACG
2337C_148680	GCT	GAT	GAT	AAT	GCT	GCG	AAG	GTT	GCT	GAT	AAG	GCG	AGT	GTG	AAG

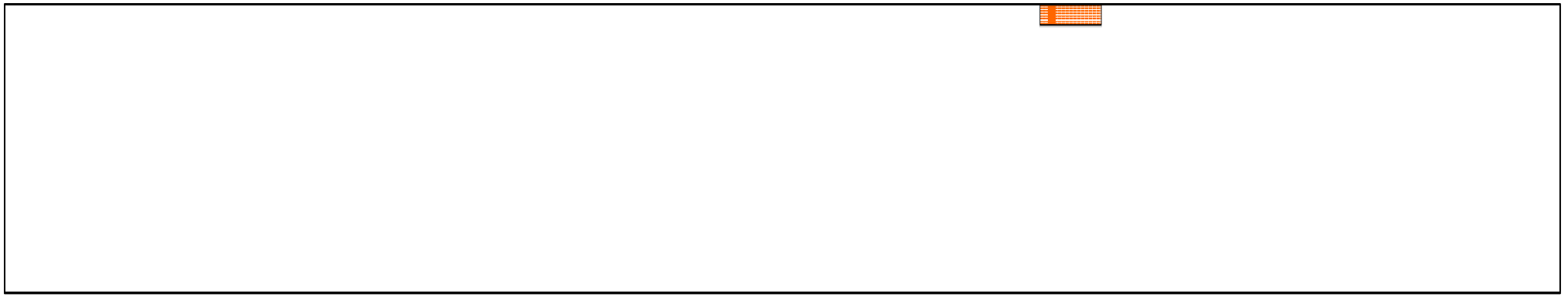
Figure 3.4 Visual representation of a *vlsE* variant and all potential silent cassette donor

(A) The variant read 2337C_139710 has a single recombination event within VR1 with the minimal recombination event length of 4 codons. The putative cassette donor is *vls*3 because it contains the longest contiguous matching segment.

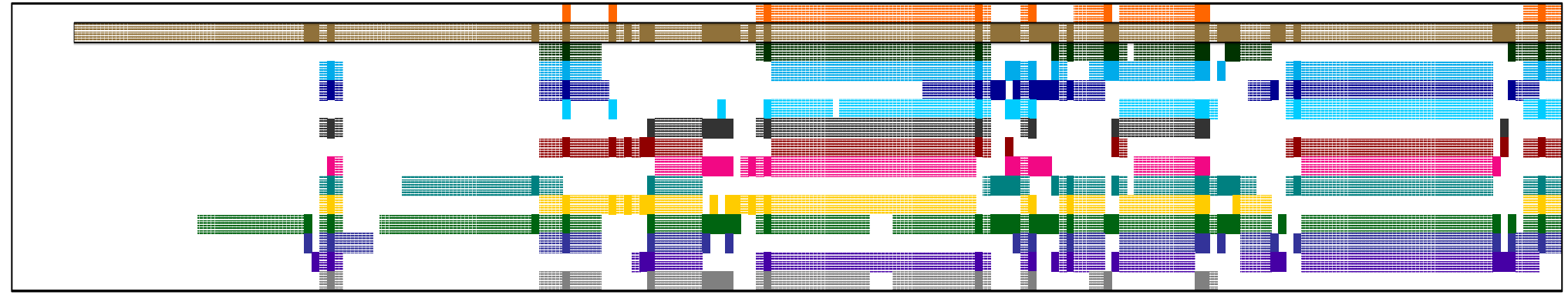
(B) The bottom panel shows the magnified VR1 region of 2337C_139710 recombination pattern and where it matches *vls3*. The top table shows the *vlsE*, 15 silent cassettes and the 2337C_139710 sequence. The brown solid colored regions represent where the change in 2337C_139710 is present in *vls3*. The lighter colored hatched region adjacent to solid colored region are identical to *vlsE*1, the vls3 silent cassette and the variant sequence. The yellow boxed regions indicate where 2337C_139710 matches *vls3* sequence. The recombination event most likely used *vls*3 as a template, because it is the only silent cassette that matches all three codon changes.

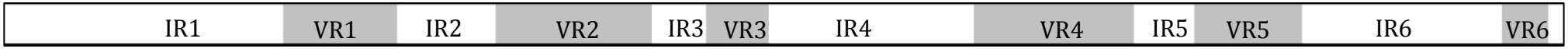


(A) 2337C_127488



(B) 2337C_103644





(C) 2337C_101310

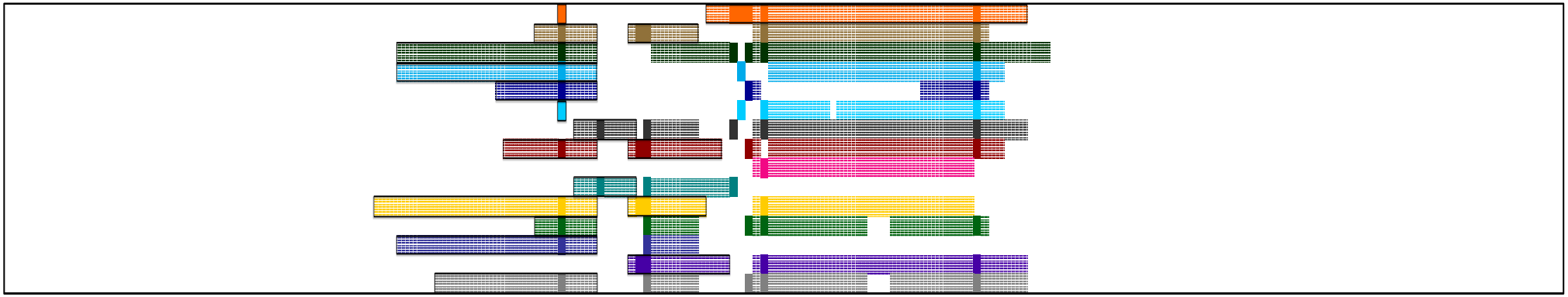


Figure 3.5 Visual representation of three *vlsE* variants and all potential silent cassette donors.

(A) The variant read 2337C_127488 has a single recombination event within VR1 with the minimal recombination event length of a single nucleotide. The putative cassette donor is *vls*3 (orange) because it is the only cassette that matches that sequence change.

(B) Read 2337C_103644 has a single, long, well-defined recombination event that has a minimal recombination event that extends from VR1 to VR6, and has a length of 486 nucleotides. The putative donor is *vls*3 (brown) because it is the only silent cassette that matches the entire range of the recombination event.

(C) Variant read 2337C_101310 has multiple recombination events. The deduced number of recombination events is 4 because there are 4 intermittent segments that match silent cassettes. A single putative cassette donor cannot be determined for 3 of the 4 sequence changes. *vls*2 is the most likely donor for the right-hand recombination event spanning from VR3 to VR4.

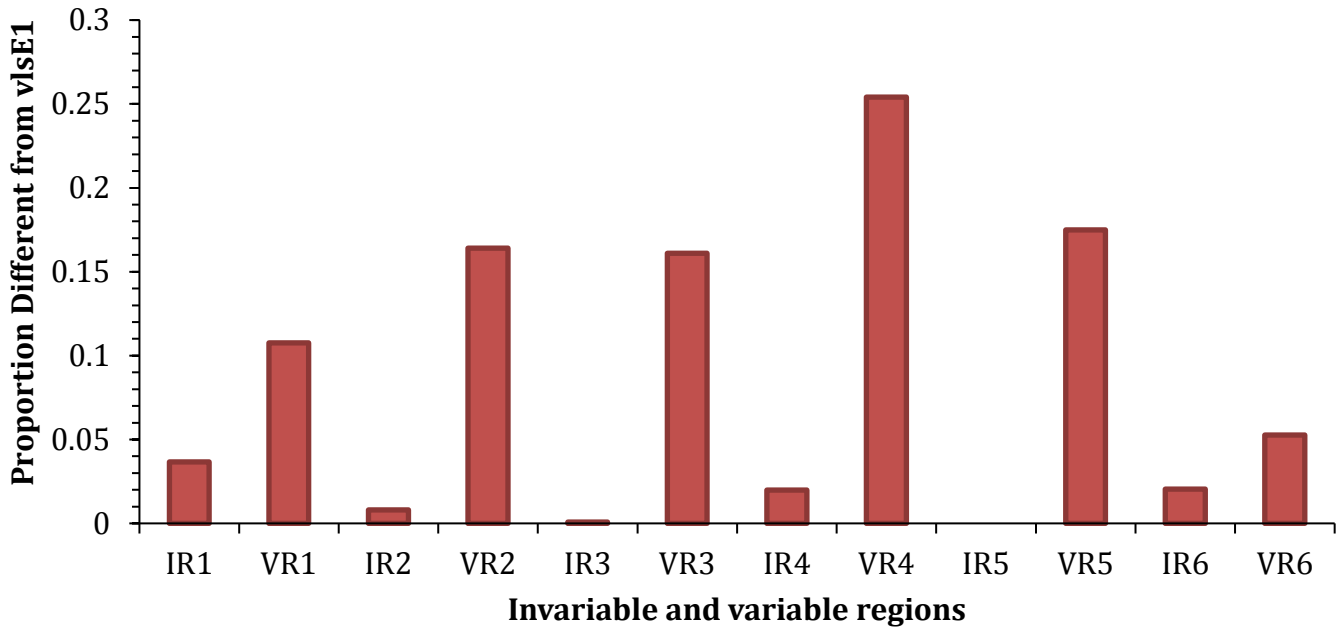
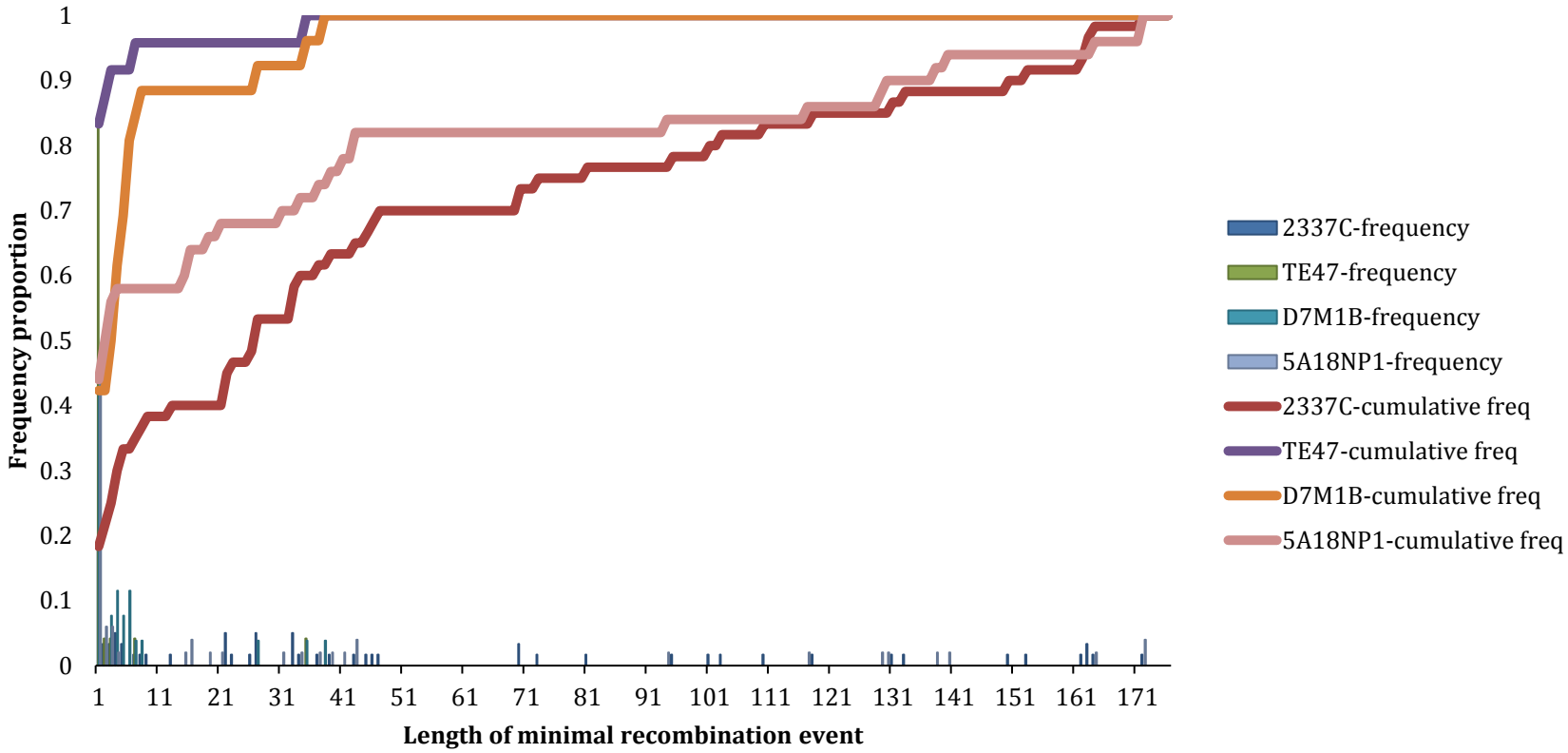


Figure 3.6 Majority of non-'parental' amino acid sequences in variant sequences were within the expected lengths of the variable regions.

The variant nucleotide sequences were translated into amino acid sequences and then compared against the parental *vlsE* amino acid sequence. The majority of the positions with non-parental amino acid sequence were within the variable regions.

(A)


Frequency proportion
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1
11
21
31
41
51
61
71
81
91
101
111
121
131
141
151
161
171
Length of minimal recombination event
2337C-frequency
TE47-frequency
D7M1B-frequency
5A18NP1-frequency
2337C-cumulative freq
TE47-cumulative freq
D7M1B-cumulative freq
5A18NP1-cumulative freq

(B)

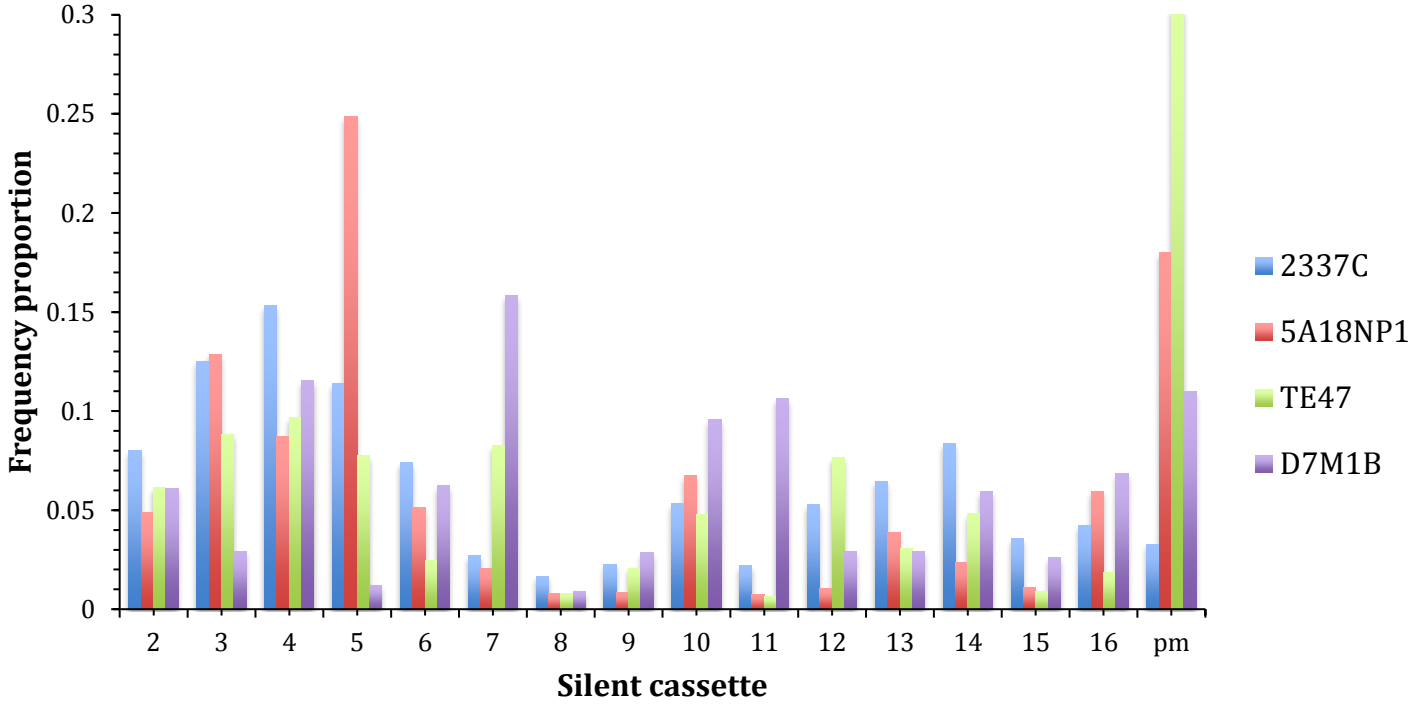


Figure 3.7 Lengths of minimum recombination events in variant sequences with a single, well-defined recombination event.

(A) For each sample, the length of the minimal recombination event in variant sequences with a well defined single recombination event was graphed against the frequency proportion. The cumulative frequency of sequences that have a minimal recombination event length that is less than or equal to a set length is shown by the line graph.

(B) Along with the length of the minimal recombination event, the deduced cassette donor for variant sequences with a single recombination event were counted and graphed. Some sequences with a single recombination event had a single nucleotide change as point mutations (pm).

Chapter 4

Discussion

Both qPCR and high throughput sequencing were effective methods in detecting and quantifying *vlsE* variants. This is the first time *vlsE* recombination has been shown to occur in *in vitro* mammalian cell-free cultures and tissue explant co-cultures. In both methods, the difference in the number of variants found in axenic in vitro samples and tissue co-culture samples was not statistically significant. The results from the previous chapter suggest that not only is there a difference in the rate of recombination among axenic culture, tissue explant co-culture and infected mice samples, but there is also a difference in the nature and extent of antigenic variation.

qPCR results

We were able to standardize a qPCR protocol that allowed us to use absolute quantification to compare the rate of *vlsE* recombination in axenic, tissue explant co-culture and animal infection samples. Two primer pairs were used, one for detecting parental *vlsE* and the other for detecting variant *vlsE*. Both primer pairs are specific for detecting whether the VR5 region has undergone a change. The variant primer detects variants with a change in VR5 region; therefore the parental primer and not the variant primer will amplify any variants with an unchanged VR5 region. Given that 12 of 15 silent cassettes have a VR5 sequence that differs from the parental sequence at the priming site, a positive PCR result would be expected in ~80% of clones with a recombination event in the region. In addition, recombination events occurring in other portions of the *vlsE* cassette region would not be detected by this method. Therefore this qPCR method provides an indication of the extent of recombination in a *B. burgdorferi* population, but does not provide an estimate of, for

example, the proportion of clones that have a recombination event in any location within the *vlsE* cassette.

The optimized qPCR assay had linear standard curves with R^2 values $\geq$ 0.979, high amplification efficiency ($\geq$ 99%), and consistency across replicates. Using absolute quantification standard curves, we found that the percentage of variants was 2.97% in sample 2337C, 99.98% in 1396D, 7.97% in TE47, 12.42% in D7M1B, 39.42% in D14M1B, 78.57% in D28M1B (Table 3.2). Comparing the percent of variants found in samples 2337C and TE47, the difference was not statistically significant. Similar to the results observed in prior studies, the qPCR assay showed an increase in the accumulation of variants between day 7 and day 28 post-infection in mice. However, the percentage of variants calculated by qPCR absolute quantification in sample 2337C and TE47 was higher than the percentage of variants found by high-throughput sequencing. This is most likely the result of the inherent differences between the two methodologies.

PacBio sequencing results

High-throughput sequencing of *vlsE* PCR products generated a large amount of data with approximately 30,000 reads per sample. Despite the presence of sequencing errors, we were able to manually align sequences against 16 pre-aligned cassettes. The error analysis showed that we were able distinguish real *vlsE* variants from sequence errors. The majority of sequencing errors are within invariant regions of *vlsE* and therefore could be unambiguously 'repaired'. Most of these errors are deletions of guanine within a homopolymer G-run. This demonstrates that like many other sequencing platforms PacBio may also have difficulties sequencing GC-

rich, repetitive sequences or stretches of homopolymer DNA. Several other studies have also reported that deletion errors are twice as common as substitution and insertion errors in PacBio sequencing [40].

TG insertions (which would result in frameshifts) are found in single locations of two *vls* silent cassettes, *vls14* and *vls16*. Otherwise, all other relative indels in the *B. burgdorferi* B31 *vls* cassette regions occur in multiples of three nucleotides, thus preserving the open reading frame [26]. This same pattern has been found in all Lyme disease *Borrelia vls* loci examined [27]. Furthermore, a single-base indel was found in only one of 1,399 *vlsE* sequences examined by Coutte et al. [29] by Sanger sequencing. These findings provide further confidence that the single nucleotide indels found in the PacBio sequences represent sequence errors.

In this study, manual alignments and sequence corrections were performed on 486 PacBio sequences to permit evaluation of the frequency and nature of *vlsE* recombinations that occurred under the conditions examined. This analysis provided a view of the previously undetected *vlsE* variation in axenic cultures, as well as in tissue explants and in infected mice.

Analyses of axenic cultures samples (2337C and 5A18NP1) revealed that approximately 1% of the sequences were variant *vlsE*. According to our hypothesis, we expected to see a significantly higher number of variants within the tissue explant co-culture samples than the number of variants detected in axenic culture. Looking at the number of variants found in samples, there was not a significant difference between the number of variants found in tissue explant co-culture samples (TE47, TE49, TE50) and the *in vitro* mammalian cell free samples (2337C, 5A18NP1). Thus

the data analyzed to date indicate that the recombination rate *in vitro* is not enhanced by the presence of mammalian tissue. It is possible that not all variant sequences were found using the method of local BLAST to compare the raw reads against IR-VR regions. Also, alteration of the tissue explant system may result in higher recombination rates. Therefore, additional analysis is needed to reinforce this initial finding.

To find the number of variants in mouse infection samples (mouse bladder tissue 7 days post inoculation), 200 sequences were manually aligned. Fifty-seven percent of those 200 sequences were found to have alterations. These results are in agreement with prior studies, which showed ~50% of clones recovered from mice at 7 days post infection had a variant *vlsE* sequence [27]. The similarity in the number of variants found in day 7 mouse infection samples in this study and former studies further help to support PacBio sequencing as a valid method for quantitating the rate of *vlsE* recombination.

Across all variant sequences, the majority of the changes occurred within the proposed variable regions. This result supports prior studies presented in Coutte et al. [27]. The conserved regions of VlsE may preserve the protein structure or functionality, serving as a 'scaffold' for the variable regions.

Analysis of variants with single well-defined recombination events showed that the mammalian cell-free *in vitro* samples had more variants with longer minimal recombination than the tissue explant co-culture and infected mouse samples. This distinction could be indicative of a mechanistic difference in *vlsE* recombination when in *in vitro* (axenic samples) and when in the presence of mammalian cells

(tissue explant co-cultures and mouse infected tissue samples). A different set of proteins may be involved when *vlsE* recombination occurs in an *in vitro* mammalian cell free environment that results in longer minimal recombination events. Interestingly, like the variant reads from axenic *in vitro* sample, the variant *vlsE* sequences recovered from *ruv*AB mutants also had long minimal recombination events. This observation further supports the proposition that additional proteins involved in recombination may be expressed and active in the mammalian host than in an axenic environment.

Some of the single recombination events were template-independent point mutations. It is not clear whether these are real point mutations caused by genetic change or sequencing errors. The average quality ratio for variants with a point mutation as the single well defined recombination event was lower than the average quality ratio for the remaining variants with a single recombination event (data not shown). Therefore, it is likely that some the point mutations are sequencing errors.

VlsE has not been shown to have a function other than helping *B. burgdorferi* evade host immune response by varying its sequence. It is also possible that the proteins required for an increase *vlsE* recombination rate are activated at specific environmental conditions such as pH, temperature, and osmolarity. Previous studies have shown that *vlsE* recombination can occur within SCID mice, and therefore the adaptive immune response does not play a role in the induction of *vlsE* recombination [33]. It is also unlikely that host proteins are internalized by *B. burgdorferi* and are used in *vlsE* recombination because that would require the host proteins to penetrate the bacterial membrane. Interaction between specific host

proteins and the surface exposed VlsE may be required to activate the proteins that play a role in increasing the rate of *vlsE* recombination in the mammalian host.

Conclusion and future perspectives

In this study, we developed a protocol using PacBio sequencing to identify real *vlsE* variants in samples from mammalian cell-free in vitro cultures, tissue explant co-cultures and mouse infection studies. High-throughput sequencing generated large amounts of data that allowed us to detect rare recombination events. *vlsE* recombination has been shown to occur in random segmental changes that can include a single base change to almost the entire cassette. Therefore, the study of *vlsE* variants is sensitive to sequencing errors. Despite the PacBio sequencing results being hampered by errors, we were able to unambiguously fix a majority of the errors.

There seems to be an intrinsic low level of *vlsE* recombination occurring in axenic cultures, which is enhanced when in mammalian host. We found that the length of a recombination event in variant sequences with a single well defined recombination event were longer in mammalian cell-free samples than in tissue explant and mouse infection samples. This may be the result of a different set of proteins being activated during *vlsE* recombination in *in vitro* and in *in vivo.* We also developed a qPCR assay that is a sensitive and relatively simple means of assessing the frequency of *vlsE* recombination in a sample. The qPCR results corroborate the PacBio sequencing results, as both methods detected the occurrence of *vlsE* variation in axenic cultures and showed increases during mouse infection. Also, both approaches indicated that there was not a significant difference between the number

of variants found in the mammalian cell-free and the tissue explant co-culture samples.

It might be possible to increase the sensitivity of the qPCR assay by standardizing tissue weight prior to DNA extraction. For the next sequencing experiments, we will be looking into whether there is a difference in the error rate between the SMRTbell sequencing of the sense and antisense strand of a single molecule. If there is a significant difference, averaging the strand with the lower error rate will result in a higher quality consensus read.

The remaining samples from mouse infection studies will be sequenced and analyzed. This includes joint, bladder, heart, skin, and ear tissue samples from days 7, 14, 28, 60, and 90 post infection. To further understand the effect of passage number on the number of variants detected in mammalian cell-free *in vitro* culture, we will be sending clone samples that have undergone 1-10 passages. We also plan to sequence another *B. burgdorferi* gene with the same GC content as *vlsE.* This approach will allow us to better measure the error rate in PacBio sequencing. We will be using the automation program to align sequences and make error corrections.

Furthermore, the protocol developed in this study using PacBio high-throughput sequencing and qPCR to study *vlsE* recombination could be applied to study other antigenic variation systems. For example, *Neisseria gonorrhoeae*, the causative agent of the sexually transmitted disease gonorrhea, varies its *pilE* sequence. Other examples are the previously described *T. brucei* system, the Vlp family of lipoproteins in *Mycoplasma hyrohini*, and Msp2 antigenic variation in *Anaplasma phagocytophilum* [41].

Chapter 5

Bibliography

1. Tilly, K., P.A. Rosa, and P.E. Stewart, *Biology of infection with Borrelia burgdorferi.* Infectious Disease Clinics of North America, 2008. **22**(2): p. 217-234.
2. Adeolu, M. and R.S. Gupta, *A phylogenomic and molecular marker based proposal for the division of the genus Borrelia into two genera: the emended genus Borrelia containing only the members of the relapsing fever Borrelia, and the genus Borreliella gen. nov. containing the members of the Lyme disease Borrelia (Borrelia burgdorferi sensu lato complex).* Antonie Van Leeuwenhoek, 2014. **105**(6): p. 1049-72.
3. Pritt, B.S., et al., *Identification of a novel pathogenic Borrelia species causing Lyme borreliosis with unusually high spirochaetaemia: a descriptive study.* Lancet Infect Dis, 2016. **DOI S1473-3099(15)00464-8 [pii] 10.1016/S1473-3099(15)00464-8**.
4. Mead, P.S., *Epidemiology of Lyme Disease.* Infect Dis Clin North Am, 2015. **29**(2): p. 187-210.
5. Steere, A.C., et al., *Lyme arthritis: an epidemic of olioarticular arthritis in children and adults in three Connecticut communities.* Arthritis Rheum., 1977. **20**: p. 7-17.
6. Schwan, T.G., W.J. Simpson, and P.A. Rosa, *Laboratory confirmation of Lyme disease.* Can. J. Infect. Dis., 1991. **2**(2): p. 64-9.
7. Sanchez, J.L., *Clinical Manifestations and Treatment of Lyme Disease.* Clinics in Laboratory Medicine, 2015. **35**(4): p. 765-778.
8. Bogdos, M., S. Giannopoulos, and M. Kosmidou, *The conflict on posttreatment Lyme disease syndrome: a clinical mini review.* Neuroimmunology and Neuroinflammation, 2016. **10**(3).
9. Bacon, R.M., et al., *Serodiagnosis of Lyme disease by kinetic enzyme-linked immunosorbent assay using recombinant VlsE1 or peptide antigens of Borrelia burgdorferi compared with 2-tiered testing using whole-cell lysates.* J. Infect. Dis., 2003. **187**(8): p. 1187-99.
10. Arvikar SL, S.A., *Diagnosis and treatment of Lyme arthritis.* Infectious Disease Clinics of North America, 2015. **29**(2): p. 269-280.
11. Fraser, C.M., et al., *Genomic sequence of a Lyme disease spirochaete, Borrelia burgdorferi.* Nature, 1997. **390**(6660): p. 580-586.
12. Purser, J.E. and S.J. Norris, *Correlation between plasmid content and infectivity in Borrelia burgdorferi.* Proc. Natl. Acad. Sci. USA, 2000. **97**(25): p. 13865-13870.
13. Labandeira-Rey, M., E. Baker, and J. Skare, *Decreased infectivity in Borrelia burgdorferi strain B31 is associated with loss of linear plasmid 25 or 28-1.* Infect. Immun., 2001. **69**: p. 446-455.
14. Norris, S.J., *Encyclopedia of Microbiology*. 3 ed, ed. M. Schaechter. Vol. 1. 2009, San Diego: Elsevier.
15. Steere, A.C., *Lyme disease.* N. Engl. J. Med., 1989. **321**: p. 586-596.
16. Yang, X., et al., *Interdependence of environmental factors influencing reciprocal patterns of gene expression in virulent Borrelia burgdorferi.* Mol. Microbiol., 2000. **37**(6): p. 1470-9.

17. Schwan, T.G. and J. Piesman, *Temporal changes in outer surface proteins A and C of the Lyme disease- associated spirochete, Borrelia burgdorferi, during the chain of infection in ticks and mice.* J. Clin. Microbiol., 2000. **38**(1): p. 382-8.
18. Samuels, D.S. and J.D. Radolf, *Borrelia: Molecular Biology, Host Interaction and Pathogenesis*. 2010, Hethersett, Norwich, UK: Caister Academic Press.
19. Norris, S.J., et al., *Pathobiology of Lyme disease Borrelia*, in *Borrelia: Molecular and Cellular Biology*, D.S. Samuels and J.D. Radolf, Editors. 2010, Caister Academic Press: Hethersett, Norwich, UK. p. 299-331.
20. van der Woude, M.W. and A.J. Bäumler, *Phase and antigenic variation in bacteria.* Clin. Microbiol. Rev., 2004. **17**(3): p. 581-611.
21. Cornelis Vink, G.R., H. Steven Seifert, *Mircrobial antigenic variation mediated by homologous DNA recombination.* FEMS Microbiol 2012: p. 917-948.
22. Cahoon, L.A. and H.S. Seifert, *An alternative DNA structure is necessary for pilin antigenic variation in Neisseria gonorrhoeae.* Science, 2009. **325**(5941): p. 764-767.
23. Cahoon, L.A., Seifert HS., *Transcription of a cis-acting, Noncoding, Small RNA Is Required for Pilin Antigenic Variation in Neisseria gonorrhoeae.* PLoS Pathog, 2013.
24. Norris, S.J., *Antigenic variation with a twist - the Borrelia story.* Mol. Microbiol., 2006. **60**(6): p. 1319-22.
25. Dai, Q., et al., *Antigenic variation by Borrelia hermsii occurs through recombination between extragenic repetitive elements on linear plasmids.* Mol. Microbiol., 2006. **60**(6): p. 1329-43.
26. Zhang, J.R., et al., *Antigenic variation in Lyme disease borreliae by promiscuous recombination of VMP-like sequence cassettes.* Cell, 1997. **89**(2): p. 275-285.
27. Norris, S.J., *vls antigenic variation systems of Lyme disease Borrelia: Eluding host immunity through both random, segmental gene conversion and framework heterogeneity.* Microbiol Spectr, 2014. **2**(6): p. doi: 10.1128/microbiolspec.MDNA3-0038-2014.
28. Zhang, J.R. and S.J. Norris, *Genetic variation of the Borrelia burgdorferi gene vlsE involves cassette-specific, segmental gene conversion.* Infect. Immun., 1998. **66**(8): p. 3698-3704.
29. Coutte, L., et al., *Detailed analysis of sequence changes occurring during vlsE antigenic variation in the mouse model of Borrelia burgdorferi infection.* PLoS Pathog., 2009. **5**(2): p. e1000293.
30. Kurosawa, K. and K. Ohta, *Genetic Diversification by Somatic Gene Conversion.* Genes, 2011(2): p. 48-58.
31. Eicken, C., et al., *Crystal structure of Lyme disease variable surface antigen VlsE of Borrelia burgdorferi.* J. Biol. Chem., 2002. **277**(24): p. 21691-21696.
32. Dresser, A.R., P.O. Hardy, and G. Chaconas, *Investigation of the genes involved in antigenic switching at the vlsE locus in Borrelia burgdorferi: an essential role for the RuvAB branch migrase.* PLoS Pathog., 2009. **5**(12): p. e1000680.
33. Lin, T., et al., *Central role of the Holliday junction helicase RuvAB in vlsE recombination and infectivity of Borrelia burgdorferi.* PLoS Pathog., 2009. **5**(12): p. e1000679.

34. Liveris, D., et al., *Borrelia burgdorferi vlsE antigenic variation is not mediated by RecA.* Infect. Immun., 2008. **76**(9): p. 4009-4018.
35. Walia, R. and G. Chaconas, *Suggested role for G4 DNA in recombinational switching at the antigenic variation locus of the Lyme disease spirochete.* PLoS One, 2013. **8**(2): p. e57792.
36. Zhang, J.R. and S.J. Norris, *Kinetics and in vivo induction of genetic variation of vlsE in Borrelia burgdorferi.* Infect. Immun., 1998. **66**(8): p. 3689-3697.
37. Barbour, A.G., *Isolation and cultivation of Lyme disease spirochetes.* Yale J. Biol. Med., 1984. **57**(4): p. 521-525.
38. Kawabata, H., S.J. Norris, and H. Watanabe, *BBE02 disruption mutants of Borrelia burgdorferi B31 have a highly transformable, infectious phenotype.* Infect. Immun., 2004. **72**(12): p. 7147-7154.
39. Rhoads, A. and K. Fai Au, *PacBio Sequencing and its Applications.* Genomics, Proteomics & Bioinformatics, 2015. **13**(5): p. 278-289.
40. Laehnemann D, B.A., McHardy AC., *Denoising DNA deep sequencing data—high-throughput sequencing errors and their correction.* Briefings in Bioinformatics, 2015. **17**: p. 159-179.
41. Palmer, G.H., T. Bankhead, and H.S. Seifert, *Antigenic Variation in Bacterial Pathogens.* Microbiolspec, 2016. **4**(1).

VITA

Surabhi Tyagi was born in India on September 2nd 1991 to Deepa and Aditya Tyagi. Raised in Austin Texas, she graduated from McNeil High School and attended University of Texas at Austin in the fall of 2009. She studied Biology with a concentration in Computational Biology and received her Bachelor of Science degree in May 2013. She moved to Houston in the summer of 2013 and enrolled in The University of Texas Health Science Center Graduate School of Biomedical Science at Houston.