Evolutionary Analyses on the Core Genome of *Borrelia burgdorferi sensu lato:* Elucidating the Genomics of Virulence

by

James Haven

A dissertation submitted to the Graduate Faculty in Biology in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York

2011

This manuscript has been read and accepted for the Graduate Faculty in Biology in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

__Weigang Qiu_____

Full Name

_____ __HIDDEN_____

Date Chair of Examining Committee

__Laurel Eckhardt_____

Full Name

_____ ___HIDDEN_____

Date Executive Officer

ABSTRACT

Adaptive Evolution in *Borrelia burgdorferi*

by

James Haven

Advisor: Dr. Weigang Qiu

The availability of multiple genomes of closely related pathogen strains makes it possible to identify genome-wide variations associated with strain-specific phenotypes such as pathogenicity and virulence. One main challenge of gene-trait associative mapping in bacterial species is finding a way to minimize the effect of linkage among loci due to pervasive clonal population structures. A second concern is to distinguish selective sequence variations from random, selectively neutral differences among strains. Here we identified adaptive, strain-specific nucleotide polymorphisms (SSNPs) on the core genome of *Borrelia burgdorferi*, the Lyme disease pathogen. We minimized the linkage effect by comparing the genomes of seven isolates representing four genospecies (*B. burgdorferi sensu stricto, B. bissettii, B. afzelii*, and *B. garinii*) and four clonal groups of a single species (A, C, E, and K clones of *B. burgdorferi sensu stricto).* Identification of selective nucleotide polymorphisms was achieved by applying codon-based, tests of positive selection based on rates of synonymous (K_S) and nonsynonymous (K_A) substitutions. We then tested for the presence of positive selection at 824 gene loci on the main chromosome, 68 loci on the linear plasmid lp54, and 26 loci on the circular plasmid cp26. Consequently, we identified 28 genes under positive selection without regard for lineage, 12 genes associated with genospecies divergence, and 7 genes associated with the adaptive divergence of B31, a highly invasive strain. We checked results by

excluding loci with high alignment uncertainties, mapping positively selected sites on protein structure models, and evaluating the possibility of false positives. Cell envelope genes are significantly over-represented among the positively selected genes. Additional categories of interest are DNA metabolism, transcription, cell division, and regulation. Focused analyses on copy number variation of established immune elicitors and a survey of intraspecific recombination support a prominent role for adaptive evolution in the maintenance of the *B. burgdorferi* pathogen cycle. These findings highlight immune escape as a driver of positive natural selection via surface protein variation and possibly pathogen replication dynamics.

ACKNOWLEDGEMENTS

I thank the *Borrelia* sequencing team of Sherwood R. Casjens, John J. Dunn, Benjamin J. Luft, Claire M. Fraser, Weigang Qiu, and Steven E. Schutzer, working under grants from the Lyme Disease Association and National Institutes of Health (AI37256 and AI49003), for access to unpublished sequence information. Special thanks to my mentor Dr. Weigang Qiu for expert guidance on the *Borrelia burgdorferi sensu lato* complex including project design, data processing, data analysis, and manuscript preparation, Dr. Shaneen Singh for expert guidance on protein structure informatics, Dr. Ewa Wywial for significant contributions to chapter II including protein modeling and manuscript preparation, Levy Vargas for significant contributions to chapter III including data processing and recombination modeling, Dr. Stephane Boissinot for excellent criticisms on project design for chapter I and the Appendix, Dr. David Lahti for expert guidance on relaxed selection and the Appendix, and Dr. Oliver Attie, Yozen Hernandez, and William McCaig for technical support. Additional financial support for this work came from grants GM083722 (to WQ), GM60665 (JH), and RR03037 (to Hunter College) from the National Institutes of Health.

TABLE OF CONTENTS

List of Tables

List of Figures

CHAPTER I – Detecting Positive Selection in *Borrelia burgdorferi*

1. INTRODUCTION

1.1.1 Comparative Bacterial Genomics

Comparative bacterial genomics is a burgeoning field fueled by great progress in DNA sequencing technology. Thirty years ago, a single *E. coli* genome of 4.6-Mbp would require a thousand years to sequence (Binnewies, Motro et al. 2006). At the present time, the same genome can be sequenced in a day or less (Margulies, Egholm et al. 2005; Shendure, Porreca et al. 2005). Technological improvements have lead to the sequencing of over a thousand bacterial genomes in varying stages of completion (http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi). The knowledge harnessed from these data is invaluable, forming the bases of significant contributions in evolution and ecology (Mongodin, Emerson et al. 2005), epidemiology (van Belkum, Struelens et al. 2001), biotechnology (Diaz 2008), and vaccine development (Mora, Donati et al. 2006). There are two major hurdles in genomic research: assembly (the piecing together of sequence reads), and efficient data management and analysis. The relatively new field of bioinformatics has played a major role in this research area, facilitating the development of novel algorithms for genome assembly (Ewing and Green 1998), gene prediction (Delcher, Harmon et al. 1999), homologous gene identification (Altschul, Madden et al. 1997), ortholog and paralog designation (Li, Stoeckert et al. 2003), gene annotation (Huang, Adams et al. 1997), genome visualization software (http://gmod.org/wiki/Main_Page), and application programming interfaces (API) for flexible implementation of analysis protocols (www.bioperl.org). Parallel progress in other fields such as computational phylogenetics and protein structure informatics has

fostered interdisciplinary research in the form of computational analysis pipelines. A typical analysis schema is shown in figure 1.1. The crux of this particular pipeline is the detection of adaptively evolving genes and assessing the effects of non-neutral substitutions on protein structure and function. Other applications of computational pipelines are the identification of synteny (gene order) among species (Ovcharenko, Nobrega et al. 2004), the reconstruction of genome-based phylogenies (Wu and Eisen 2008), and the discovery of regulatory elements (Doh, Zhang et al. 2007); to name a few. In the following sections, we detail a computational study on the core genomes of seven strains of *Borrelia burgdorferi* sensu lato. The results have clear implications in biomedical research and evolutionary theory pertaining to pathogenic bacteria.

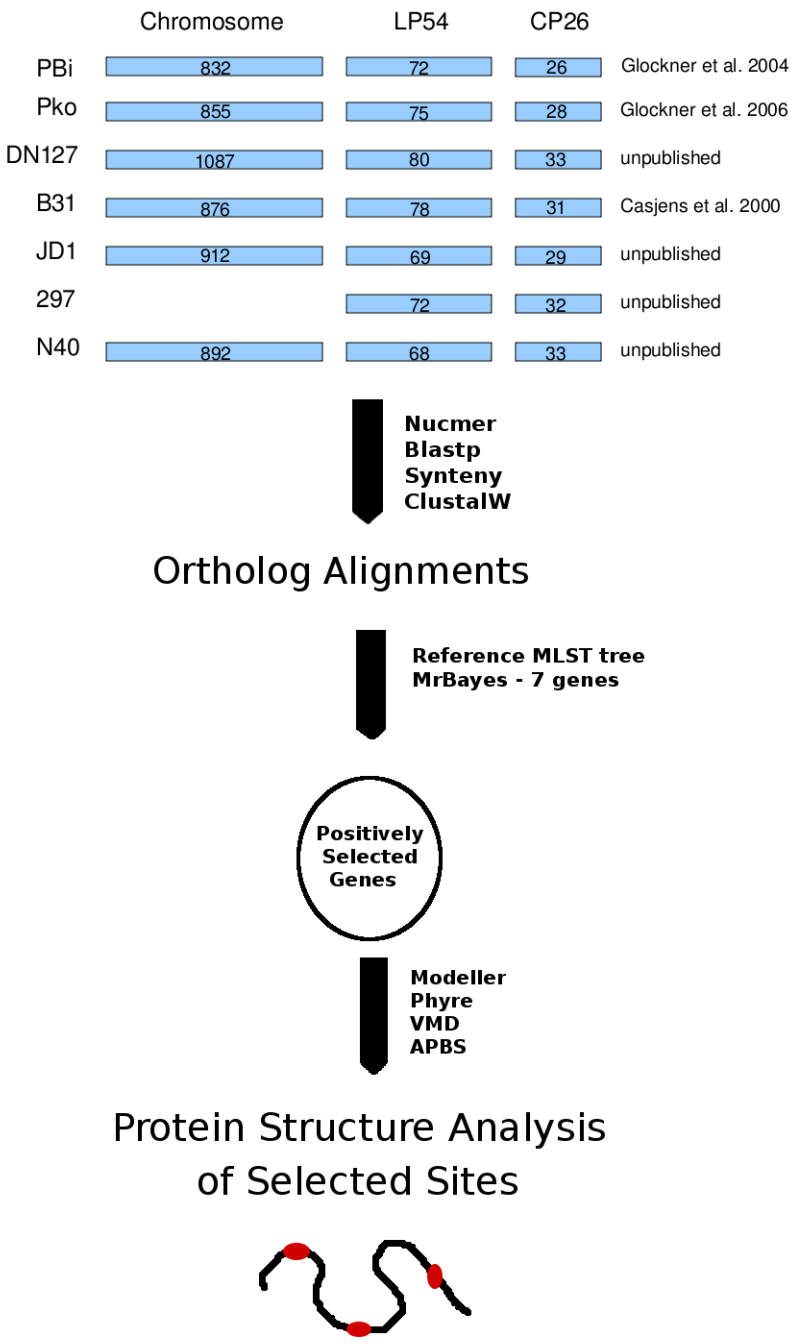


Figure 1.1. Computational Analysis Pipeline Model. The steps involved in our analysis pipeline as read from top to bottom, detailing the genomic data, phylogenetic processing, and structural analysis. Top portion illustrates the three genomic regions used in the analysis as rectangles. On the left of each rectangle is the *Borrelia* strain name. On the right of each rectangle is the data origin's reference. Inside each rectangle is the number of ORFs present in each genomic region. Text adjacent to arrows denotes software used for data processing.

1.1.2 Pathogen Genomics

Pathogen genomics research generally poses two questions. First, what are the genes involved in virulence and immune escape, and second, how do these genes evolve? One approach involves comparing the genome content from two strains of a bacterial species, one being pathogenic, and one non-pathogenic. The genome material that is unique to one strain may be the cause of pathogenicity (Hayashi, Makino et al. 2001). This approach does not suffice for pathogenic bacteria with no known non-pathogenic strains. Furthermore, pathogen-unique genomic content may simply be the result of neutral evolution. A different approach analyzes nucleotide substitution patterns in the core genome, and requires genome content from multiple strains (Chen, Hung et al. 2006; Lefebure and Stanhope 2007). This method uses phylogenetic structure to estimate the synonymous (silent) and non-synonymous (replacement) substitution rates. The ratio of non-synonymous to synonymous substitution rates is known as *Ka/Ks*, and is indicative of the strength of natural selection acting on a gene (McDonald and Kreitman 1991). If *Ka/Ks* is significantly greater than 1, one can assume that positive selection is present. Pathogen genes involved in host interactions are frequently under positive selection. Common examples are antigenic variation in surface exposed proteins (Gupta, Ferguson et al. 1998) and paralogous copy number variation among host immune elicitors (Yogev, Rosengarten et al. 1991). These proteins are engaged by the host immune system, and the pathogen is subject to removal by host macrophages. This scenario leads to the accelerated fixation of non-synonymous mutations in the pathogen host-interacting gene; a phenomenon known as frequency-dependent selection, which is characterized by the favoring of rare mutants. Screening a bacterial genome for positively selected genes will

yield a pool of genes, some of which illicit the host immune response. Suchlike genes are often sought as candidates for vaccine development through a computational approach known as ‘reverse vaccinology’ (Rappuoli 2001). Furthermore, their evolutionary histories give insight into host-pathogen co-evolution. There are, of course, positively selected pathogen genes not directly involved in host interaction, and we discuss this in the section, “Positively selected non-cell envelope genes”.

Spirochetes of the genus *Borrelia burgdorferi* sensu lato comprise a fascinating group of pathogens for evolutionary study. They exhibit niche specialization of multiple host species (Brisson and Dykhuizen 2004), which is likely driven by diversifying selection. Strains of *B. burgdorferi* have a varied propensity for invasiveness (high virulence or hematogeneous dissemination) (Wormser, Brisson et al. 2008) and different forms of Lyme disease (Wilske, Preac-Mursic et al. 1993), such as Lyme dermatitis and Neuroborreliosis. *B. burgdorferi sensu lato* have strong geographic structure across North America and Eurasia (Piesman and Gern 2004), and a probable European origin for *B. burgdorferi sensu stricto* (Margos, Gatewood et al. 2008). These features pose interesting questions such as 'what genes are involved in the colonization of different hosts?', 'what genes influence virulence?', 'what is the migration route of *B. burgdorferi*?', and 'are there instances of horizontal gene transfer within and between *B. burgdorferi* genospecies?'. Although this study focuses on a subset of these questions, all of them can be partly addressed using a comparative genomic approach, which underscores the utility of comparative genomic research methods.

1.1.3 Borrelia burgdorferi sensu lato

1.1.3.1 Lyme disease: etiology, geography, and virulence

Borrelia burgdorferi sensu lato is a bacterial species complex representing a group of spirochetes, several members of which cause Lyme Disease; a tick-borne illness with a wide variety of clinical manifestations including neurologic, cardiac, skin, and joint abnormalities (Steere, Grodzicki et al. 1983). Lyme disease is the most prevalent vector-borne disease in the United States and Europe as well as becoming an important emerging infection in East Asia (Piesman and Gern 2004). The predominant species in the United States is *B. burgdorferi sensu stricto*, which is endemic in the Northeastern states and is transmitted primarily by the arthropod vector species, *Ixodes scapularis*. Another major group, termed the 25015 or the DN127 group, also known as *B. bissettii*, has been found in California, Colorado, and the southern United States (Postic, Ras et al. 1998; Norris, Johnson et al. 1999). To date, all human-derived *B. burgdorferi* isolates in North America are members of *B. burgdorferi sensu stricto*. In Eurasia, the predominant pathogenic species are *B. garinii* and *B. afzelii*, with *B. burgdorferi sensu stricto* being less represented (Baranton, 1998).

1.1.3.2 Genomics and adaptation

Several investigators have reported the completed genomes of the *B. burgdorferi* strains B31, N40, JD1, and 297(Fraser, Casjens et al. 1997; Casjens, Palmer et al. 2000; Qiu, Schutzer et al. 2004) Casjens – personal communication, all of which are Northeastern U.S. isolates, and the partial genomes of *B. garinii* PBi and *B. afzelii* PKo from Europe (Glockner, Lehmann et al. 2004; Glockner, Schulte-Spechtel et al. 2006). A typical *B. burgdorferi* genome consists of a ~900 kilobase linear chromosome that is conserved in both gene content and gene order among *B. burgdorferi* species, and an additional ~900 kilobases of linear and circular plasmids that vary in composition even

among strains of the same *B. burgdorferi* species (Casjens, Palmer et al. 2000; Glockner, Schulte-Spechtel et al. 2006). While about half of the open-reading frames (ORFs) on the main chromosome code for uncharacterized hypothetical proteins, more than 80% of plasmid-borne ORFs code for unknown proteins that have no apparent homologs outside the *Borrelia* genus (Casjens, Palmer et al. 2000). For such a phylogenetically remote and poorly understood bacterium, comparative genomics is a valuable means of genome annotation. Although comparative genomics does not directly predict the molecular or cellular functions of ORFs, it can identify genes that contribute to the ecological adaptation of the species by scanning for signals of positive natural selection (Storz 2005; Ellegren 2008). In pathogenic species, genes involved in host-pathogen interactions or virulence, such as tissue attachment, immune evasion, and alteration of host pathways, are often under adaptive evolution. Indeed, besides acquiring virulence factors through horizontal gene transfers (Pallen and Wren 2007), bacteria can gain pathogenicity through "patho-adaptive" mutations (Sokurenko, Hasty et al. 1999). The comparative-genomic approach of identifying virulence factors in bacterial species through genome-wide scans of positive selection has been applied to the study of uropathogenic strains of *Escherichia coli* and multiple *Streptococcus* species (Chen, Hung et al. 2006; Lefebure and Stanhope 2007). The link between positive selection and virulence is expected to be particularly strong in *Borrelia species*, since they have an obligate dependence to tick and vertebrate hosts.

Here we report results of a positive-selection scan of the least variable core of the *B. burgdorferi* genome, which includes the main chromosome, the linear plasmid lp54 and the circular plasmid cp26. These three replicons are well conserved in their presence,

gene content, and gene order among all known *B. burgdorferi* isolates, allowing unambiguous identification of sequence orthology at most loci. Previously, we identified genes likely under positive selection by scanning the *B. burgdorferi* genome for hypervariable ORFs with elevated ratios of nonsynonymous to synonymous substitution rates (K_A/K_S) (Qiu, Schutzer et al. 2004). Many well-known *Borrelia*-host interaction genes tested positive in the variability and K_A/K_S tests, including *ospC* (coding for outer surface protein C), *dbpA* (coding for decorin-bind protein A), and BbCRASP-1 (coding for *B. burgdorferi* complement resistance acquired surface protein 1). For the present study, we expanded the number of *B. burgdorferi* genomes for comparison from three to seven, thus improving the statistical power of tests of positive selection. We further improved the statistical power by using likelihood ratio tests, which are able to identify codon positions under positive selection ($K_A/K_S > 1$) against the null hypothesis of absence of positively selected sites ($K_A/K_S \leq 1$). More importantly, lineage-specific tests of positive selection make it possible to distinguish among various selective processes, such as the adaptive divergence between genospecies, and the presence of patho-adaptive mutations in strains likely to cause disseminative Lyme disease.

1.1.3.3 Mechanisms of virulence

One of the most interesting aspects of *B. burgdorferi sensu lato* is the abundance and variable expression of cell envelope genes (Fraser, Casjens et al. 1997; Casjens, Palmer et al. 2000; Miller, von Lackum et al. 2003; Ohnishi, Schneider et al. 2003; Brooks, Vuppala et al. 2006; Bykowski, Woodman et al. 2008). Examples include outer surface proteins A, B, and C, the vlsE gene cassette, the Erp gene family, and paralogous gene family 54 (PFam54). This repertoire of genes enables the spirochete to persist in

multiple environments, such as a variety of vertebrate hosts and an *Ixodid* tick (Stevenson 2002; Hovius, van Dam et al. 2007). These genes function in host-pathogen interactions, which include attachment, host manipulation, and immune escape. For example, outer surface proteins A & B have apparent roles in spirochete persistence in the tick by attaching the spirochete to the tick midgut (Schwan, Piesman et al. 1995; Pal, de Silva et al. 2000). Furthermore, there is strong evidence that *Borrelia* manipulate the expression of a tick receptor, TROSPA; TROSPA or 'tick receptor for ospA' has high expression levels in the midgut of infected ticks versus uninfected ticks (Pal, Li et al. 2004).

Immediately upon infection of the vertebrate host, the spirochete encounters the innate immune response in the form of the complement cascade. In order to successfully overcome this assault, *Borrelia sensu lato* has evolved a series of complement neutralizing genes, which bind the host molecules factor H and factor H-like 1 (Kurtenbach, De Michelis et al. 2002; Stevenson, El-Hage et al. 2002). These *Borrelia* molecules are termed complement regulator acquiring surface proteins (CRASPs) and ospE/F related proteins (Erp). Interestingly, BbCRASP-1, a CRASP molecule, strongly binds human factor H (Kraiczy, Rossmann et al. 2006). BbCRASP-1 resides on linear plasmid 54 within a tandem array of CRASP homologues comprising a subset of a large gene family known as PFam54 (Casjens, Palmer et al. 2000). PFam54 is of great interest due to its homology with BbCRASP-1, and its expression patterns during host infection (Gilmore, Howison et al. 2008).

1.2. MATERIALS AND METHODS

1.2.1 Strains, genomes, and ortholog identification

Table 1.1 lists the biological sources, geographic origins, and GenBank accessions of genome sequences of the strains used in the present study. For each of the three orthologous replicons, we identified orthologous ORF sets by first finding all homologs of each ORF using all-against-all BLAST (Altschul, Madden et al. 1997). Homologous ORFs were clustered using the MCL algorithm (Enright, Van Dongen et al. 2002). Within each homolog cluster, orthologs were distinguished from paralogs by visual inspection of gene orders on a synteny map of the seven genomes custom designed by the authors (http://diverge.hunter.cuny.edu/~weigang/synteny_map/). Small ORFs (< 150 bp) are excluded from the study. Protein multiple sequence alignments were constructed using ClustalW 1.83 (Thompson, Higgins et al. 1994). Codon alignments were derived from protein alignment templates using PERL scripts. DNA and protein sequence alignments of ortholog families are available from the online synteny browser.

Table 1.1. *Borrelia* strain origin

Strain	**Geographic Origin**	**Biological Source**	**Strain Reference**	**Main Chromosome**	**lp54**	**cp26**
B. burgdorferi B31	New York, U.S.	Tick	(Burgdorfer, Barbour et al. 1982)	NC_001318	NC_001857	NC_001903
B. burgdorferi N40	New York, U.S.	Tick	(Barthold, Moody et al. 1988)			
B. burgdorferi JD1	Massachusetts, U.S.	Tick	(Piesman, Mather et al. 1987)			
B. burgdorferi 297	Connecticut, U.S.	Human CSF	(Steere, Grodzicki et al. 1983)	(Not Sequenced)		
B. bissettii DN127	California, U.S.	Tick	(Bissett and Hill 1987)			
B. afzelii PKo	Europe	Human skin	(Wilske, Preac-Mursic et al. 1993)	NC_008277	NC_008564	NC_008274
B. garinii PBi	Germany	Human CSF	(Wilske, Preac-Mursic et al. 1993)	NC_006156	NC_006129	NC_006128

1.2.2 Strain phylogeny

The molecular phylogeny of *ospA* sequences (from lp54) was used to represent the overall phylogenetic relationships among the seven strains. Sequences at other loci on the chromosome and plasmids (e.g., the ribosomal intergenic locus *rrsl-rrlA* and concatenated sequences at six chromosomal loci: bb0057, bb0160, bb0243, bb0545, bb0622, bb0809) yielded similar tree topologies (data not shown). The *ospA* gene tree was estimated using MrBayes (v3.0B4) with three different rates at three codon positions (Huelsenbeck and Ronquist 2001). A consensus tree including posterior branch support probabilities was obtained by running four chains of Markov Chain Monte Carlo simulations until convergence (after 2 million generations).

1.2.3 Positive selection schemes in Borrelia burgdorferi

We applied both lineage dependent and lineage independent tests of positive selection to our dataset under three schemes, summarized in figure 1.2. The first aims to identify all *Borrelia* genes potentially involved in host-pathogen interactions. Suchlike genes are typically under strong adaptive pressure in one or more lineages. We first tested each orthologous ORF family, consisting of both within and between-population sequences, for the presence of amino acid sites under positive selection using the CODEML program (site model) of the PAML package (v4.0) (Yang, Nielsen et al. 2000). The key PAML "site model" parameters include "runmode = 0" for a user-defined tree, incorporating the *ospA* gene tree for all ORF sets; "model = 0" for a single *Ka/Ks* estimate across all branches; "NSsites = 0 1 2" for a single selective pressure among sites ("M0"), variable selective pressure among sites with either negative selection or neutral evolution ("M1a"), and variable selective pressure with negative selection, neutral

evolution, and positive selection ("M2a"). We then compared the log likelihoods of the M2a and M1a models. ORF families were considered to have evolved under positive selection if there was significant improvement in the log likelihood of the positive selection model (M2a) over that of the nearly neutral model (M1a). We refer to this scheme as 'lineage independence'.

The second scheme aims to identify those genes under adaptive pressure in lineages leading to speciation. These genes may permit colonization of different hosts, and can include genes identified under 'lineage independence'. In the second set of analyses of positive selection, we tested the hypothesis that divergence among *B. burgdorferi* genospecies is adaptive. We used the "branch-site" model implemented in the CODEML program of the PAML (v4.0) package to test the presence of positive

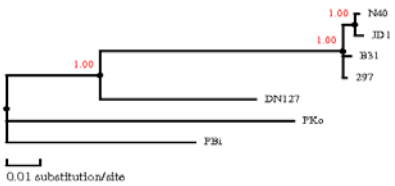
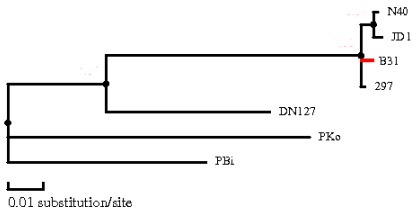
Evolution Model		PAML Test Specifications
Lineage Independence		Site Models: M0 (Negative Selection), M1a (Nearly Neutral), M2a (Positive Selection) Test statistics: $2ln(L_{M2a}/L_{M1a})$, with χ^2 (df=2)
Genospecies Divergence		Branch-Site Model: Genospecies as foreground branches. Test statistics: $2ln(L_{BS}/L_{M2a})$, with χ^2 (df=1)
B31 Divergence		Branch-Site Model: B31 lineage as foreground branch. Test statistics: $2ln(L_{BS}/L_{M2a})$, with χ^2 (df=1)

Figure 1.2. Strain phylogeny and tests of positive selection. Three evolutionary tests were performed on each orthologous ORF family. ORF families were assumed to evolve

following a single strain phylogeny, represented by a maximum-likelihood tree based on *ospA* sequences. (1st row) In the "Lineage Independent" test, codon sites evolve with homogeneous selective pressure (K_A/K_S values) among branches, regardless. We used the PAML Site Model to identify positively selected codon sites by comparing the likelihoods of the Negative Selection model (M0), the Nearly Neutral model (M1a, containing negative selected and neutral sites), and the Positive Selection model (M2a, containing negative selected, neutral, as well as positively selected sites). (2nd row) In the "Genospecies divergence" test, codon sites evolve under different selective pressures among branches, depending on whether they represent between-genospecies divergence ("foreground" branches, colored red) or within-population diversification ("background" branches, in black). We used the PAML Branch Site Model to identify positively selected sites associated with genospecies divergence by comparing the likelihoods of the Branch Site model with the Positive Selection model (M2a). (3rd row) In the "B31-specific" test, codon sites evolve differently on the B31 lineage ("foreground" branches, colored red) from all other branches ("background" branches, in black). We used the PAML Branch Site Model to identify positively selected sites associated with B31 evolution by comparing the likelihoods of the Branch Site model with the Positive Selection model (M2a).

selection on lineages leading to individual genospecies and an absence of positive selection within genospecies (Yang, Nielsen et al. 2000; Zhang, Nielsen et al. 2005). This test design was based on the observation that the majority of evolutionary sequence changes among the seven strains occurred among genospecies (figure 1.2). For ORFs that tested positive in the site model, a branch-based test helps to distinguish whether the positive selection occurred as diversifying selection within populations, or as adaptive divergence between genospecies. For ORFs testing negative in the site model, the branch-based test can reveal lineage-specific adaptive changes obscured by the branch-averaging effect of the site model. The test was performed by labeling all between-genospecies branches as "foreground" branches with $K_A/K_S > 1$, while setting within-population branches as "background" branches with $K_A/K_S \leq 1$ (figure 1.2). We refer to this scheme as 'genospecies divergence'.

The third scheme seeks to identify genes under adaptive pressure in strain B31, which belongs to a group of clones categorized as ospC type A; a class of strains prone to causing invasive forms of Lyme disease. In the third positive selection analysis, we assume hematogenous dissemination is an adaptive phenotype of *B. Burgdorferi*, and propose genes under positive selection in these invasive lineages may contribute to hematogeneous dissemination. Within our dataset, the virulent strains (defined as having a high propensity to disseminate) are B31, a North American strain, and PBi and PKo, two European strains. We chose to test adaptive sequence evolution in the B31 lineage given the extensive study and clear statistical associations between North East American strain types and hematogenous dissemination (Wormser, Liveris et al. 1999; Wang, Ojaimi et al. 2002; Dykhuizen, Brisson et al. 2008). Furthermore, the *B. burgdorferi* genospecies is the only genospecies in our dataset represented by more than one strain, including low virulent strains, which aids in comparative testing. We used the PAML branch-site model to identify ORFs showing significant adaptive nucleotide substitutions on the branch leading to B31. The B31 branch was labeled as the "foreground" branch with $K_A/K_S > 1$, while the other, within-population branches, were set as "background" branches with $K_A/K_S \leq 1$. We refer to this scheme as 'B31 divergence'.

The assessment of statistical significance in genome-wide positive selection scans can be a confusing matter. The question is whether one should control for testing multiple genes or testing multiple hypotheses (lineages). It has been argued given the expectation that some genes must be under positive selection, one need not control for multiple gene testing (Yang 2006). We follow this rationale and apply a Bonferroni correction (0.05/3=0.016) controlling for our three hypotheses. The *P*-values of the likelihood

differences were obtained by using the "pchisq" function of the statistical package R (http://r-project.org), assuming a χ^2 (*d.f.* = 2) distribution of the log likelihood ratios.

1.2.4 Structural validation of positive selection

Mapping sites presumed to be under positive selection onto a protein model is a powerful way to understand the molecular mechanism of natural selection and can affirm the occurrence of positive natural selection. A homology model of the BBA73 protein was generated by Phyre 0.2 (Bennett-Lovsey, Herbert et al. 2008) and evaluated using the Prosa web server (Wiederstein and Sippl 2007). Structural visualizations were done in VMD 1.8.6 (Humphrey, Dalke et al. 1996). File preparation for electrostatic continuum calculations used PDB2PQR 1.3.0 (Dolinsky, Czodrowski et al. 2007). Poisson-Boltzmann electrostatic continuum calculations were done using APBS 0.5.1 (Holst, Kozack et al. 1994; Baker, Sept et al. 2001). Crystallographic structures of ospC and BbCRASP-1 were visualized using APBS and VMD as described above.

1.3. RESULTS

1.3.1 Strain phylogeny, genome synteny, and copy-number variations

The evolutionary relationships among the *B. burgdorferi* strains were approximated by a phylogeny based upon *ospA* sequences (figure 1.2). The main inconsistency among gene trees based on different loci is the occurrence of discordant topologies among the four *B. burgdorferi sensu stricto* strains. Recombination between genospecies is probably rare, although plasmid exchanges could occur within populations (Dykhuizen, Polin et al. 1993; Qiu, Schutzer et al. 2004). An incongruent within-species tree may result from lineage sorting, recombination, or both. For the purpose of testing positive selection in nucleotide substitutions, however, the use of the *ospA*-based tree

versus other gene trees did not significantly affect test results (data not shown). This is presumably because within-species divergence time contributes only a minor portion of the total tree length.

ORFs on the main chromosomes, cp26, and lp54 plasmids are virtually all syntenic in these strains. Using combined evidence from BLAST searches, MCL clustering, and the genome synteny, we identified 918 sets of orthologous ORF families, including 824 families on the chromosome, 68 families on the lp54 plasmids, and 26 families on the cp26 plasmids. Genome synteny maps and ORF alignments are available for viewing and download at the *Borrelia* Genome Browser (http://diverge.hunter.cuny.edu/synteny_browser). The lp54 synteny map illustrates paralog copy-number variation in the PFam54 cluster (figure 1.3). These genes represent a suite of cell surface molecules, which have been implicated in immune escape.

1.3.2 Tests of positive selection

Twenty-eight ORFs are positively selected under 'lineage independence' table 1.2). Twelve ORFs are positively selected under the branch site model for 'genospecies divergence' (table 1.2). Seven ORFs are positively selected in the branch site model for 'B31 divergence' (table 1.2), of which five were also selected in 'genospecies divergence'. Nearly all of these ORFs reached a corrected significance value below 0.01. As expected, the three tests yielded distinct numbers of positively selected ORFs, with lineage independence detecting the most, and B31 divergence detecting the least. This is likely a result of the number of branches examined, which differs among the tests. This suggests positive natural selection may be uniformly distributed throughout the *Borrelia* strain phylogeny. Examples are illustrated in figure 1.7A-B where several positively

selected sites exhibit different amino acids in multiple lineages; however, it is unclear whether some of these substitutions are adaptive or neutral.

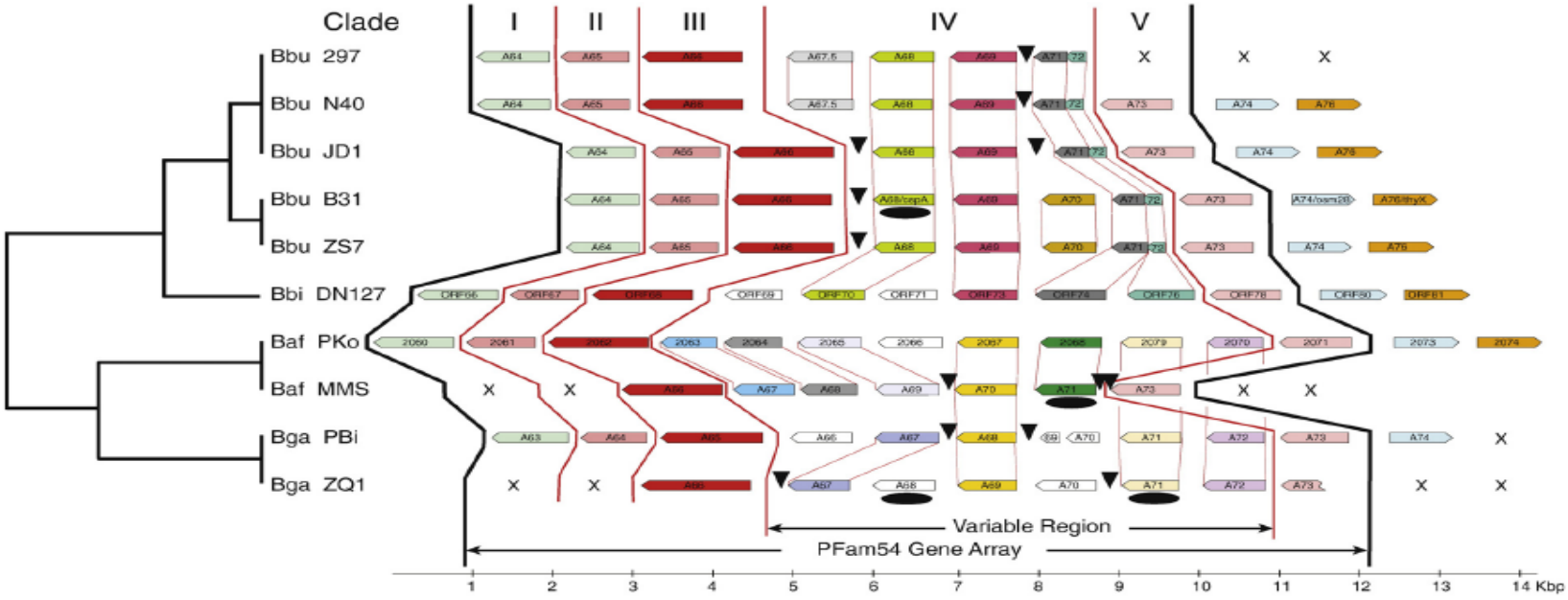


Figure 1.3. Copy number variations of BbCrasp-1 homologs on lp54 plasmids. Each arrow represents an ORF. Orthologs are shown in the same color. The genomic region of lp54 between the bbA66 and bbA73 orthologous loci has the highest degree of copy number variation in the *Borrelia* core genome. This region consists of a variable number of BbCRASP-1 homologs arranged as a single operon and may be co-transcribed. Loss and gain of orthologs can be inferred using DN127, PBi, or PKo as outgroups.

Table 1.2. Positively selected open reading frames

Orth Fam	B31	ORF Annotation	Uniform Divergence	Genospecies Divergence	B31 Divergence
31	BB0003	UTP--glucose-1-phosphate uridylyltransferase subfamily			****
54	BB0024	hypothetical protein			*
111	BB0072	membrane protein, putative	**		
202	BB0158	antigen, S2, putative			**
93	BB0210	surface-located membrane protein 1 (*lmp1*) {*Mycoplasma hominis*}	*		
188	BB0217	phosphate ABC transporter, permease protein (*pstA*) {*Synechocystis* PCC6803}	**		
343	BB0304	UDP-N-acetylmuramoylalanyl-D-glutamyl-2,6-diaminopimelate--D-alanyl-D-alanine ligase (*murF*)			****
374	BB0345	hypothetical protein			*
387	BB0364	conserved hypothetical protein {*Bacillus subtilis*}		**	
404	BB0388	DNA-directed RNA polymerase (*rpoC*) {*Escherichia coli*}			*
425	BB0416	pheromone shutdown protein (*traB*) {*Enterococcus faecalis*}	***		
429	BB0420	sensory transduction histidine kinase/response regulator {*Synechocystis* PCC6803}	***		
435	BB0432	hypothetical protein	*		
440	BB0441	ribonuclease P protein component (*rnpA*)		**	
452	BB0455	DNA polymerase III, delta subunit superfamily	**	****	
96	BB0532	hypothetical protein			**
502	BB0533	phnP protein (*phnP*) {*Escherichia coli*}	*		
510	BB0546	hypothetical protein	**	****	
80	BB0553	hypothetical protein	*		
73	BB0670	purine-binding chemotaxis protein (*cheW-3*)		*	
272	BB0737	histidine phosphokinase/phophatase, putative {*Mycobacterium leprae*}	*		
142	BB0797	DNA mismatch repair protein (*mutS*) {*Haemophilus influenzae*}	**		

3	BBA05	antigen, S1	***		
969	BBA24	decorin binding protein A (*dbpA*)	****		
972	BBA32	hypothetical protein	*		
986	BBA50	hypothetical protein	****		
988	BBA52	outer membrane protein		**	
2	BBA57	P45-13	****	****	
10	BBA60	surface lipoprotein P27	**		
12	BBA65	lipoprotein, putative	**		
13	BBA66	antigen, P35, putative	****	****	
996	BBA68	complement regulator-acquiring surface protein 1 (*CRASP-1*)	****		
997	BBA69	surface protein, putative	**		
998	A067.5 (N40)	hypothetical protein	**		
999	BGA68 (PBi)	hypothetical protein	***		
776	BBA71	hypothetical protein	****		
690	BBA73	antigen, P35, putative	****	****	
18	BBB08	lipoprotein, putative		**	
951	BBB13	hypothetical protein		*	
20	BBB14	hypothetical protein		*	
23	BBB19	outer surface protein C (*ospC*)	****		
28	BBB28	putative ankyrin repeat protein	**		

*: $0.01<P<0.05$
**: $10^{-3}<P<10^{-2}$
***: $10^{-4}<P<10^{-3}$
****: $P<10^{-4}$
Loci annotated as “Cell envelope” in the JCVI database are shaded gray.

1.3.3 Functions of positively selected loci

We used an updated JCVI annotation pipeline to interrogate the functions of positively selected ORFs (figure 1.4). Among ORF families with known functions, the most abundant (15 loci) consist of ORFs coding for known or putative outer surface proteins, a proportion significantly over-represented on the basis of genome average ($p < 10^{-4}$, Fisher's Exact test). These "cell envelope" loci included *ospC*, BbCrasp-1, *dbpA*, *lmp1*, and loci coding for P27, P35, S1, and S2 antigens (table 1.2). Among these "cell envelope" loci, 12 were positive under lineage independence, including BB0072, *lmp1*, BBA05, *dbpA*, BBA57, BBA60, BBA65, BBA66, *BbCRASP-1*, BBA69, BBA067.5, BBA73, and *ospC*. Five surface-protein loci, with mostly unknown functions, were positive under genospecies divergence: BBA52, BBA57, BBA66, BBA73, and BBB08. BB0003, BB0304, and BB0158 are 'cell envelope' loci significantly associated with adaptive divergence of B31; BB0158 is not annotated under 'cell envelope' in JCVI, but is defined in NCBI (as of 8/24/09) as a homolog of the p23-like cell envelope protein of *Borrelia hermsii*.

The remaining major categories of positively selected ORFs are 'DNA metabolism/transcription', 'regulatory', and 'unknown', which contain 4, 3, and 14 ORFs respectively (figure 1.4). A fourth category we termed 'miscellaneous' contains a variety of selected ORFs; this category cannot be decomposed into sub-categories containing more than one ORF. The category 'DNA metabolism/transcription' contains rpoC, rnpA, DNA Polymerase III-delta, and MutS. The last three are selected under 'lineage independence', while rpoC is selected under B31 divergence. The category 'regulatory' contains traB (a pheromone shutdown protein), a histidine kinase response regulator, and

a histidine kinase/phosphatase. All three are selected under lineage independence. The category ‘unknown’ contains all selected ORFs annotated as hypothetical; 3 are selected under B31 divergence (BB0024, BB0345, BB0532), 4 are selected under genospecies divergence (BB0364, BB0546, BBB13, BBB14), and 8 are selected under lineage independence (BB0432, BB0546, BB0553, BBA32, BBA50, BBA067.5, BGA68, BBA71). The category ‘miscellaneous’ contains pstA (a permease gene), phnP (involved in alkyl-phosphate uptake), CheW-3 (a chemotaxis adaptor homolog), and P45-13 (involved in energy metabolism). pstA, phnP, and p45-13 are selected under ‘lineage independence’, while CheW-3 and p45-13 are selected under ‘genospecies divergence’.

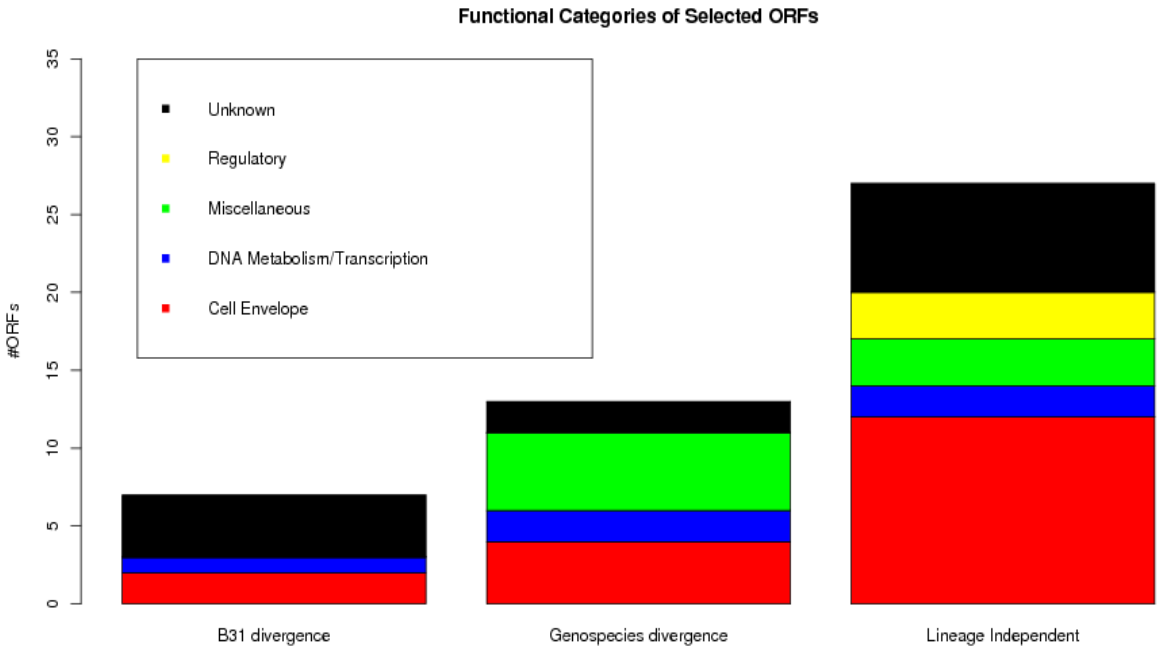


Figure 1.4. Functional categories of positively selected ORF families. Colored bars show the number of ORFs in each of five functional categories (see legend) for the positive selection tests including “Lineage Independence”, “Genospecies Divergence”, and “B31 divergence”. Functional categories were assigned on the basis of JCVI annotations as of 10/7/09.

1.3.4 Structural examinations of selected sites

In order to assess both the occurrence and nature of positive selection, we examined the three-dimensional protein structures of two highly selected loci in our dataset: BBA73 and ospC; both loci participate in host-pathogen interactions (Grimm, Tilly et al. 2004; Gilmore, Howison et al. 2008). Figures 1.5 and 1.6 illustrate a three-dimensional model and a crystallographic structure, respectively, detailing the location of positively selected amino acid sites. In both cases the locations of adaptive sites correspond with previous experimental work on these molecules (Kumaran, Eswaramoorthy et al. 2001; Cordes, Roversi et al. 2005). In BBA73, adaptive sites are found near putative interfaces for dimer formation and ligand binding. In ospC, adaptive sites are found in a putative ligand binding interface.

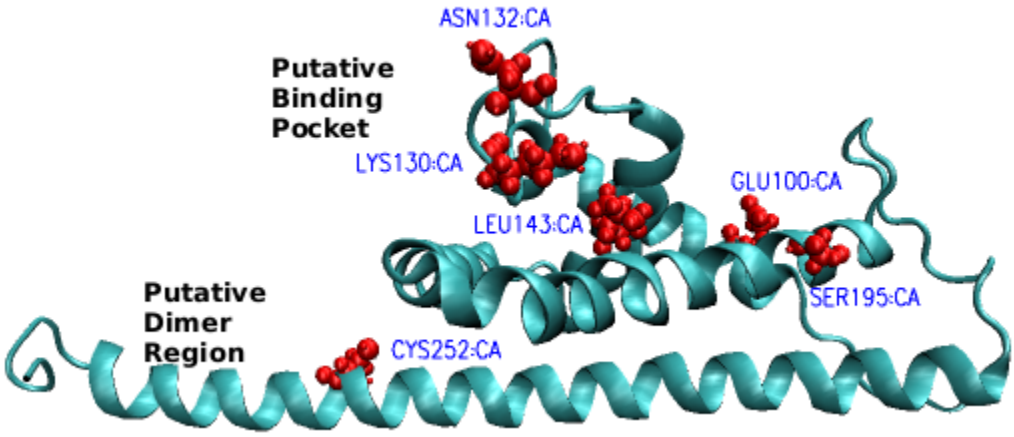


Figure 1.5. Positively selected sites on a BBA73 homology model. A homology model of BBA73, generated by the Phyre server (Bennett-Lovsey, 2008) is shown graphically as implemented in VMD [(Humphrey, 1996). Positively selected sites are shown in CPK form (red). The PROSA evaluation z-score is -5.25. All sites shown in this figure were found to be under positive selection in both the site and branch-site models of PAML.

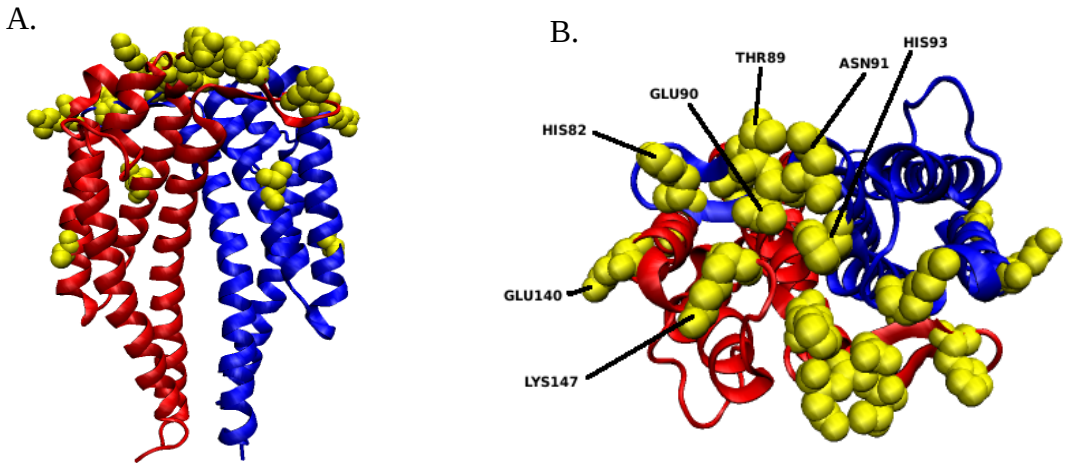


Figure 1.6. Positively selected sites on OspC. Secondary structure visualizations of the ospC homodimer from strain B31 are presented using VMD. OspC monomers are colored red and blue. **(A)** Lateral view of ospC with positively selected sites (PAML site model) shown as Van Der Waals radii in yellow. **(B)** Top view. Membrane distal end of ospC (as shown in A). Positively selected sites (shown as Van Der Waals radii, in yellow) are clustered. Labels for positive sites of one monomer are shown in black.

1.4 DISCUSSION

In this study, we examine positive selection on the core genome of the pathogenic bacterium, *Borrelia sensu lato*, in hopes that variant pathogenicities and host preferences apply detectable selection pressure on core genes. The identification of these genes may provide a greater understanding on the maintenance of *Borrelia* genospecies, vaccine efforts in Lyme Disease, and pathogen evolution in general. Our results support the expectation of adaptive evolution in cell envelope genes, and highlight the adaptive potential of genes involved in DNA metabolism, replication, and repair. Herein we address the methodology of identifying gene-trait associations in bacteria, limitations of positive selection tests, and the structure, function, and adaptive potential of positively selected *Borrelia* genes.

1.4.1 Population structure and associative mapping in bacteria

The statistical power to detect genomic elements underlying a phenotype (e.g., virulence) in a bacterial species depends critically on genetic population structure, especially the recombination rate and degree of clonality (Falush and Bowden 2006). At one hypothetical extreme, in which a species consists solely of clonal lineages without recombination between clones, any genomic differences between a virulent and a benign strain would manifest as a complete association with virulence, including neutral differences such as synonymous mutations. Such artificial association is caused by repetitive sampling of isolates having the same genomic types. At the other extreme, associative mapping of virulence in a freely recombining species requires no special consideration in the sampling of isolates. Estimating the rate of recombination between clonal groups in a bacterial population, using methods such as multilocus sequence typing (MLST) at housekeeping loci (Cooper and Feil 2004), is therefore a prerequisite for designing rational strategies when sampling strains for gene-trait associative mapping in bacteria.

The useful bacterial strains have a population structure between complete clonality and panmix (Smith, Smith et al. 1993). Thus, the sampling of isolates should not be random, but instead take phylogenetic relatedness of clonal complexes into consideration. In these cases, a useful analytical framework consists of phylogenetic comparative methods, which avoid the problem of phylogenetic autocorrelation (e.g., repeated sampling of closely related species) by estimating the number of evolutionary *changes*, rather than the number of presence or absence of an evolutionary *character* in samples (Harvey and Purvis 1991). One advantage of using the comparative approach for

the study of gene-trait associations in bacteria is that phylogeny-based methods are inherently hierarchical and there is no need for first defining artificial taxonomical units (e.g., genospecies), as long as a robust phylogeny can be estimated based on, for example, concatenated housekeeping gene sequences. Nevertheless, sampling of different MLST clonal complexes is more economical than repetitive sampling of isolates belonging to the same MLST clonal complex. This economy occurs because loci are more likely in linkage disequilibrium among clonal complexes than between clones, so that identified loci are more likely to be targets of natural selection than the results of hitchhiking.

Initially the population structure of *B. burgdorferi* was found to be strictly clonal (Dykhuizen, Polin et al. 1993). Subsequent sampling of isolates coexisting within the same natural populations revealed genetic exchange mediated by plasmid transfers at a rate three times that of point mutations (Qiu, Schutzer et al. 2004). *B. burgdorferi* population structure thus appears to fit the "epidemic" model of bacterial population structure, in which a small set of genomic types rapidly grows into largely clonal bacterial groups from an ancestral, recombining population (Attie, Bruno et al. 2007). As plasmids exchanged among different clonal groups before their adaptive radiation, identification of strain-specific genomic variations in *B. burgdorferi* is best achieved by sampling isolates representing different clonal complexes, as identified by their *ospC* sequences.

1.4.2 Types I and II errors in positive selection tests

Clearly, not all strain or taxa-specific polymorphisms are adaptive. Tests of positive natural selection are designed to distinguish between neutral variations (e.g.,

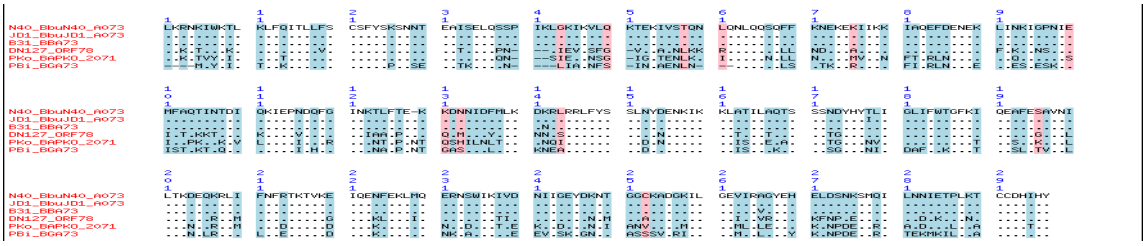
synonymous substitutions), selective constraints (negative selection), and changes conferring selective advantages (positive selection). Footprints of positive natural selection at the molecular level are often detected on the basis of deviation from neutral expectations of allelic frequency distributions within natural populations and by comparisons of within-species polymorphisms with between-species divergence. Our results of positively selected genes were based primarily on likelihood ratio tests of $K_A/K_S>1$ as implemented in PAML. PAML has been a popular choice for detecting genome-wide positively selected genes and amino acid sites, including the identification of the genetic basis of bacterial virulence (Chen, Hung et al. 2006; Lefebure and Stanhope 2007). Nevertheless, concerns have been raised on the use of a PAML and K_A/K_S-based approach in producing both under-estimation (Type I error) and over-estimation (Type II error) of positive selection (Hughes 2007). The possible sources of such errors in PAML tests in our study are discussed below.

First, the limited genomic scope of our dataset may under-detect selective targets. We studied only nucleotide polymorphisms in the coding sequences of the core *B. burgdorferi* genome. We did not study many other possible genomic variations underlying strain differences, including strain-specific nucleotide polymorphisms (SSNPs) on other plasmids, the gain and loss of genes and plasmids, and SSNPs in noncoding regions.

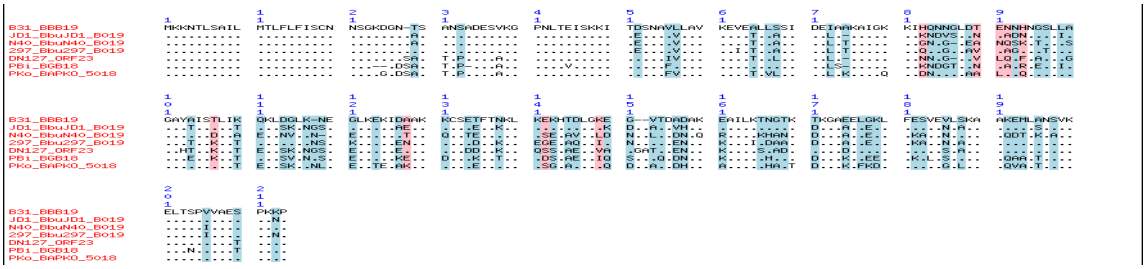
Second, the limited number of strains we compared may result in both under- and over-estimation of positively selected targets. The five-to-seven orthologous sequences (depending on the gene) used here approached the lower limit of sample sizes required by PAML tests. Inclusion of orthologous sequences from more genomes (but not from the

same *ospC* clonal complexes) will allow the observation of more synonymous and nonsynonymous substitutions, resulting in more precise estimates of K_A/K_S ratios. We do not believe, however, that the list of genes identified as being under positive selection is greatly affected by the small sample sizes because of our strain choice. For genes under lineage independence (e.g., *ospC*), four *B. burgdorferi sensu stricto* strains appear to be sufficient to show high within-genospecies variability (figure 1.7B). In fact, we were able to detect genes associated with adaptive genospecies divergence using the strains of four genospecies (figure 1.7A; table 1.2). However, it is not known whether some variations represent fixed differences between genospecies as only one genospecies (*B. burgdorferi sensu stricto*) was represented by multiple clonal complexes; *B. afzelii*, *B. bissettii*, and *B. garinii* are all represented by one strain only. Similarly, genes associated with B31-specific selection may not prove true if sequences

A.



B.



C.

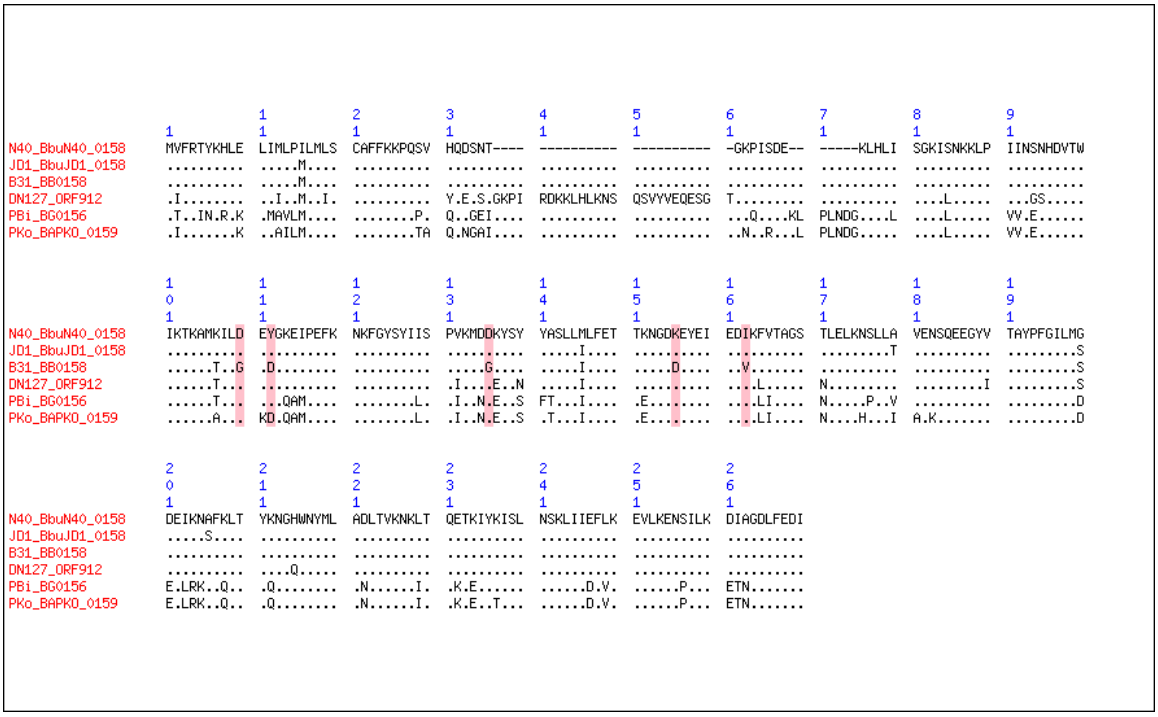


Figure 1.7. Positively selected sites in sample alignments. Neutrally evolving (K_A/K_S =1) and positively selected sites (K_A/K_S >1) identified by PAML tests are shaded blue and red respectively (unshaded sites are negatively selected, conserved sites). **(A).** BBA73, a BbCRASP-1 homolog. This ORF family tested positive in the "lineage independent" test and the likelihood score was improved further in the "genospecies divergence" model. Neutrally evolving sites vary primarily between two amino acid states, while 11 positively selected sites are polymorphic, with three or four amino acid states. Virtually all amino acid variability is between genospecies and there is little amino acid variability among the four *B. burgdorferi sensu stricto* strains. **(B).** OspC (BBB19). This ORF family tested positive in the "lineage independent" model but negative in the "genospecies divergence" and "B31 divergence" models. Neutral and positively selected sites vary both between and within-genospecies. The *ospC* locus is known to be under strong diversifying selection within local *B. burgdorferi* populations. **(C).** BB0158, an antigen termed 'S2'. Five positively selected sites associated with B31 divergence are closely clustered. It tested negative in both the "lineage independent" and "genospecies-divergence" models.

from additional *B. burgdorferi sensu stricto* clonal complexes share the same types of nucleotide polymorphisms.

Third, alignment uncertainty causes over-estimation of positive selection due to inflated nonsynonymous substitution rates in poorly aligned regions. In our study, most orthologous ORFs were aligned with little ambiguity because of the close evolutionary relationship among the seven strains. Most regions of poor alignment, e.g., regions containing a large number of gaps, appeared in the N- and C- terminus. Computational uncertainties in base calls and ORF calls may have caused such artifacts of high variability. Being aware of this pitfall, we examined alignment at each locus that was significant in the evolutionary tests and removed those loci where high K_A/K_S ratios were due to high amino acid variability at poorly aligned terminal regions of the ORF sequences.

Fourth, the more serious sources of false-positive (Type II) errors are the nonselective causes of high K_A/K_S ratios, especially the strong effect of genetic drift in obligate parasite species (Hughes 2007). *B. burgdorferi* is a non-free-living, vector-borne parasite that, to complete a life cycle, must be transmitted from a reservoir host to the tick vector as well as the reverse migration from the tick to a new host. Thus, the effective population size of *B. burgdorferi* is small relative to that of free-living bacterial species, weakening the purifying selection responsible for removing deleterious nonsynonymous mutations from the population. However, genes with the highest K_A/K_S ratios, such as *ospC* and *dbpA*, are clearly involved in host-parasite interactions and their variability is maintained by positive selection (Grimm, Tilly et al. 2004). Independent evidence based on allele frequency distributions within natural populations showed that natural selection in *ospC* takes the form of frequency-dependent selection for immune escape antigens or host adaptation (Brisson and Dykhuizen 2004). Therefore, the high K_A/K_S ratios at

plasmid loci appear to be maintained by positive selection rather than by nearly neutral deleterious mutations. On the other hand, chromosome alleles that have a low recombination rate relative to that of the plasmid loci, are more likely to be fixed with deleterious nonsynonymous mutations. Investigation of the within-population frequencies of these SSNPs would help to determine whether they are maintained by positive selection or by genetic drift.

Lastly, the presence of recombination and horizontal gene transfer are other causes of high K_A/K_S ratios due to the introduction of a large number of homoplasies by recombination events (Anisimova, Nielsen et al. 2003). However, simulation studies showed that the likelihood ratio tests of positive selection are robust when recombination rates are relatively low.

To summarize the statistical evaluation of our test results, while the present study is limited in genomic scope, used a small number of strains, and did not correct for recombination, the likelihood ratio test used for identifying loci contributing to within-population diversification was nevertheless robust because of strain choice and alignment validations. Further sequencing of genomes representing more distinct clonal types will help to reduce the Types I and II errors in identifying positive selection during genospecies divergence and strain-specific polymorphisms.

1.4.3 Structural mapping of two positively selected genes

Protein structure and function offer independent means of validating positive selection. The BBA73 locus exhibits one of the strongest signals of positive selection between genospecies (table 1.2) and is a paralog of *BbCRASP-1* (BBA68), which codes for a protein that binds to and neutralizes human complement factor H or Factor H-like 1

proteins (Kraiczy, Rossmann et al. 2006). We used a crystal structure of BbCrasp-1 (PDB identifier: 1W33) (Cordes, Roversi et al. 2005) to predict the three-dimensional structure of the BBA73 protein. Figure 1.5 displays a homology model of BBA73. Recent studies suggest that the template molecule (BBA68 protein) functions as a homodimer and that ligand binding occurs in a centralized cleft formed after dimerization of the CRASP monomers (Cordes, Roversi et al. 2005). It is reasonable to assume the BBA73 product also functions as a dimer. Two positively selected sites (Lysine 130 and Asparagine 132) are located at the entrance of the putative binding cleft. Another positively selected site, Cysteine 252, is located deep within the putative binding cleft. Thus, structural examination of BBA73 provides appreciable functional support for positive selection, going beyond the comparison of nucleotide substitutions.

OspC is a major surface antigen of *B. burgdorferi sensu lato* and is among the most highly selected genes in our analysis. Crystal structures of ospC have been reported for strains B31 and N40 (Kumaran et al. 2001, Eicken et al. 2001). These structures are highly similar homodimers and are almost entirely helical. Kumaran and colleagues have identified a strong, negatively charged surface patch on the 'membrane distal end' of ospC, which is present only in invasive *Borrelia* strains, suggesting this patch may serve as a binding site of an unknown host factor (Kumaran et al. 2001). Interestingly, our phylogenetic analysis strongly supports this suggestion. Figure 1.6 displays the secondary structure of the B31 ospC homodimer, with positively selected sites mapped onto the structure. As can be seen clearly, the abundance of positive sites maps directly onto the 'membrane distal end'. Taken together, the electrostatics and positive site data suggest

this location in ospC is of great functional importance, most likely serving as a docking platform for an unknown host ligand.

1.4.4 Positively selected, non-cell envelope genes

The positively selected, non-cell envelope genes include ORFs involved in DNA metabolism (BB0455, BB0797), transcription (BB0388, BB0441), regulation (BB0416, BB0420, BB0737), transport (BB0217), chemotaxis (BB0670), energy metabolism (BB0364, BBA57), central intermediary metabolism (BB0533), and cell division (BBB13). The adaptive potential of positive selection on a number of these genes is plausible; selection for replication timing, speed, or efficiency can be achieved through genes involved in DNA metabolism and repair, cell division, and regulation. A possible driver of this selection is immune escape. An increase in the mutation rate will result in greater nucleotide diversity, and heightened diversity is known to facilitate immune escape in viruses such as HIV and Influenza via epitope alteration (Richman, Wrin et al. 2003; Koelle, Cobey et al. 2006). Therefore, *Borrelia* in-host persistence can be enhanced in a similar manner and at a cost of accumulating deleterious substitutions (Giraud, Matic et al. 2001). In an alternative scenario, the need to clear viral infections may drive positive selection in bacteria. Csaba Pal and colleagues (Pal, Macia et al. 2007) experimentally showed the selection of mutator strains of *Pseudomonas fluorescens* via MutS knockouts (a DNA repair gene) in the presence of viruses. These mutator strains were under positive selection because they were able to clear viral infections. Note the *Borrelia* MutS gene is under positive selection (Table 1.2). There is strong evidence *Borrelia* has coevolved with viruses (Eggers and Samuels 1999), suggesting *Borrelia* may have followed a similar course as in the *P. fluorescens* experiment.

Another driver of positive selection in non-cell envelope genes is the modulation of the pathogen generation time. Reducing virulence until a transmission event is common in vector-borne pathogens where replication is restricted in the vector and explosive in the host (Ewald 1994). Interestingly, BBB13, a *Borrelia* cell division gene, is under positive selection (table 1.2), which may influence bacterial replication. The within-host population dynamics of the bacterial pathogen is a critical component of its persistence and is regulated by its intrinsic growth rate, the host immune response, and quorum sensing. Quorum sensing is population density-dependent gene expression, and in bacterial pathogens quorum sensing can enable the timely expression of virulence factors, optimizing bacterial persistence in the host (de Kievit and Iglewski 2000). Quorum sensing has been identified in *B. burgdorferi* (Stevenson and Babb 2002).

It is important to note that positive selection in *Borrelia* non-cell envelope genes might be false, attributable to deleterious mutations and weak purifying selection. Given the apparent low rate of recombination in *Borrelia* (and consequently weak recombinational DNA repair), one might assume deleterious substitutions are common. However, significant positive selection exists in transcriptional genes in the pathogenic species, *Streptococcus,* and DNA metabolism genes in uropathogenic *E. coli*. Both of these species are freely-recombining and are thus capable of recombinational DNA repair (Guttman and Dykhuizen 1994; Lefebure and Stanhope 2007). Hence, findings of positive selection in similar categories of 'housekeeping' genes in *Borrelia*, *Streptococcus*, and uropathogenic *E. coli* are likely the result of a common selection pressure shared among bacterial pathogens.

1.4.5 Hematogenous dissemination as an adaptation

Hematogenous dissemination (HD) in Lyme disease refers to invasive and highly virulent clinical manifestations, whereby spirochetes migrate to extant body areas from the tick bite via the bloodstream. Several studies have indicated a significant statistical relationship between the strain type and HD (Wormser, Liveris et al. 1999; Jones, Glickstein et al. 2006; Wormser, Brisson et al. 2008). These studies have also pointed out the ability of most *B. burgdorferi* strains to disseminate, including the strains with the greatest propensity for dissemination. Several studies have attempted to identify the shared gene(s) involved in HD using micro-array and proteomic based approaches (Ojaimi, Mulay et al. 2005; Nowalk, Gilmore et al. 2006). Our study is the first attempt to identify a culprit gene(s) using a phylogenetic approach, which was conducted with our "B31 divergence" test. The phylogenetic approach assumes HD is an adaptive feature originating by positive selection on gene(s) in highly virulent strains. This assumption warrants a discussion. A pathogen typically evolves in a direction to maximize its R_0; the basic reproductive ratio (Ewald 1994; Lenski and May 1994). The R_0 value is the average number of secondary transmissions per infection. It serves as a threshold value to predict whether the introduction of a pathogen into a population of hosts will reach epidemic proportions. R_0 values greater than one indicate an epidemic, whereas R_0 values less than one indicate the decay of the pathogen population. Under this model, pathogens typically evolve towards an intermediate level of virulence (Fenner and Woodroofe 1965). Such virulence is driven by a pathogen population size (within-host) suitable for effective transmission, while not imposing a disease cost that would impair transmission too greatly. There is evidence *Borrelia* loosely follows this model, where strains of high

virulence (denoted as RST1) are least common in ticks, compared to medium and low virulent strains; RST2 and RST3, respectively (Wormser, Brisson et al. 2008).

Genes that were found to be under positive selection in our "B31 divergence" test are not present among potential HD-influential genes reported in the micro-array and proteomic studies mentioned above. There are, at least, several reasons for this: 1) our phylogenetic study is restricted to the core *Borrelia* genome; 2) the study by Ojaimi and colleagues is restricted to differences in gene regulation between strains; and 3) the study by Nowalk et al. is restricted to membrane proteins of a single invasive strain. The fact that HD is not restricted to a particular genospecies or phylogenetic clade strongly suggests multiple mechanisms can contribute to HD. Some possible mechanisms are prolonged immune escape, variable strain generation times, variable chemotactic abilities, and an affinity to specific host tissues.

The 'B31 divergence' genes under positive selection fall into several categories: 1) 'cell envelope' (BB0003, BB0158, and BB0304); 'DNA metabolism' (BB0388); and 3) 'hypothetical' (BB0024, BB0345, BB0532). This positively selected gene set for strain B31 is not dissimilar from our global positive selection results. Although this set may represent bona-fide targets of HD-influence, such targets are probably not shared causes among *Borrelia* genotypes exhibiting HD. For instance, RST3 is generally the least virulent *Borrelia* genotype, however, ospC type I, a member of RST3, is highly virulent (Wormser, Brisson et al. 2008). As prescribed in the previous section, there is a significant and interesting trend of positive selection in 'housekeeping' genes among pathogenic bacteria, which suggests pathogen replication dynamics plays a significant role in virulence. Therefore, instead of focusing efforts in identifying HD targets, our

attention may be better placed on understanding the evolutionary and ecological determinants of HD, so as to enact policy that may deter its emergence.

Ultimately HD is probably a short-term strategy of high multiplicity within hosts at a cost of transmission efficiency. This is reflected by the comparably low numbers of high virulence RST1 strains among ticks (Wormser, Brisson et al. 2008), which indicate RST1 strains are the poorest performers in natural environments. The mechanism behind this transmission cost is unclear, because *Borrelia* cycles through an arthropod and multiple distinct vertebrate hosts (Brisson, Dykhuizen et al. 2008). Future studies examining this cost may be invaluable in controlling Lyme disease.

1.5. CONCLUSION

This study identified loci on the core *B. burgdorferi* genome that show strain-specific nucleotide polymorphisms associated with adaptive evolution among genospecies, including a high virulence strain. Although there are substantial sources of error in underestimation and overestimation of positive selection tests, we validated test results through alignment checking, appropriate strain choice, and functional and structural analyses. More than half of the genes showing evidence of positive selection code for cell surface proteins, which support our approach. The approach of using likelihood ratio tests with PAML improves on a previous comparative genomics study of *B. burgdorferi* high-density nucleotide polymorphisms because of its ability to reveal positive selection on specific branches of a gene tree as well as on individual amino acids. In the future, an expanded analysis including genomes of more *Borrelia* clonal complexes will improve the statistical power to identify genetic variations associated with selective evolution in two ways. First, the incorporation of additional genomes

results in a larger sample size so that selective targets are more likely to show significance in associative mapping. Second and more importantly, each sampling of a new clonal complex reduces the average size of linkage groups, thereby making it more likely that the identified loci are indeed targets of positive selection and not variations due to the effect of pervasive genetic hitchhiking in bacterial species.

Importantly, several functional categories containing adaptively evolving genes are represented among adaptive gene categories in the pathogenic bacterial species, *E. coli* and *Streptococcus;* the categories are 'cell envelope', 'DNA metabolism', and 'transcription'. (Chen, Hung et al. 2006; Lefebure and Stanhope 2007). Taken together, these results suggest pathogenicity among distinct bacterial pathogens is under similar evolutionary pressures. Unlike the 'cell envelope' gene category, finding adaptive genes in the categories 'DNA metabolism' and 'transcription' is unexpected. This finding points towards pathogen replication dynamics as an adaptive mechanism, which may influence pathogen transmission and virulence.

Hematogenous dissemination occurs in almost all strains of *B. burgdorferi sensu lato*, although certain strains are more prone to invasiveness. Our study is the first attempt to identify HD determinants in a phylogenetic context. We propose cell envelope genes, and genes involved in DNA metabolism and transcription potentially influence HD. We argue HD can probably result through multiple mechanisms, and may be an inherent ability of *Borrelia* to maximize its within-host population size.

Chapter II – The Pfam54 Virulence Cassette

2.1 INTRODUCTION

2.1.1 Innate immunity

Innate immunity is the first line of defense against invading pathogens. It comprises non-specific immune functions and does not confer long-term immunity. An example of innate immunity is the complement system. The host complement cascade is a near-instantaneous and powerful assault on an invading pathogen. Its mechanisms include degrading a microbial plasma membrane, tagging a pathogen for removal by immune cells, and clearing antigen-antibody complexes (Alberts 2002). Only those pathogens capable of evading the complement system may establish a persistent infection. Hence, pathogens have evolved a variety of mechanisms to escape the complement cascade. It is beyond the scope of this thesis to describe the different escape mechanisms employed by bacterial pathogens, but a comprehensive review is given by Zipfel and colleagues (Zipfel, Wurzner et al. 2007).

2.1.2 Complement regulator acquiring surface proteins

Members of the genus *Borrelia* harbor genes that encode complement regulator acquiring surface proteins (CRASPs) (Kraiczy, Rossmann et al. 2006). These proteins have the ability to bind host complement regulators, providing a disguise for the *Borrelia* spirochete. In particular, there exists a paralogous gene cassette known as Pfam54, which resides on the core linear plasmid, lp54 (Casjens, Palmer et al. 2000). Within this cassette is a gene (bba68) encoding BbCRASP-1, which has been shown to bind the human complement regulator factor H, and thus enable infection in humans (Kraiczy, Skerka et

al. 2001). BbCRASP-1 has a novel fold, which is a homodimer of alpha helical bundles (Cordes, Roversi et al. 2005).

We begin our survey of the Pfam54 virulence cassette by assembling a genetic dataset of all bba68 homologues present in the seven *Borrelia* strains under study. We follow with a phylogenetic analysis that involves gene tree reconstruction, estimating natural selection pressure, and protein structure modeling and electrostatics. The integration of these approaches reveal key nucleotide substitutions likely responsible for the high affinity of BbCRASP-1 for human factor H. Given our current understanding of the *Borrelia* life cycle, we contend this affinity is probably incidental and not driven by a specialization for the human host.

2.3 MATERIALS AND METHODS

2.3.1 Data Assembly

The Pfam54 gene array resides near one end of the lp54 plasmid; a core DNA replicon in *Borrelia burgdorferi sensu lato* (Casjens, Palmer et al. 2000). DNA sequences of this segment were retrieved from GenBank representing ten strains from the following genospecies: *B. burgdorferi sensu stricto*, *B. garinii*, *B. afzelii,* and *B. bissettii*. The strains are termed B31, N40, 297, JD1, ZS7, DN127, PKo, MMS, PBi, and ZQ1. Included in this dataset are Pfam54 homologues, which were identified in seven genomes (B31, N40, JD1, 297, PBi, PKo), using the Blastp program of the BLAST package (Altschul, Madden et al. 1997). BBA68, also known as *cspA*, was used as the search query.

2.3.2 Phylogenetic Tree Reconstruction

A phylogenetic tree of all sequences in our dataset was constructed using MrBayes 3 (Huelsenbeck and Ronquist 2001). The DNA sequences of the open reading frames (ORFs) were translated into protein sequences, which were aligned using ClustalW, version 1.83 (Thompson, Higgins et al. 1994). The MrBayes protein tree was built using the Jones amino acid matrix (Jones, Taylor et al. 1992), running 10^6 Markov Chain Monte Carlo (MCMC) generations. The average standard deviation of split frequencies fell below 0.015.

2.3.3 Branch site test of positive selection

A subtree of our BbCRASP-1 homolog tree, as described above, consisting of orthologs of bba68 and the PBi outgroup gene, bga66, was analyzed for positively selected codons (omega > 1) using the branch site model in PAML 4 (Yang, Nielsen et al. 2000; Zhang, Nielsen et al. 2005). The ancestral branch of the *sensu stricto* clade was set as the foreground branch (putatively under positive selection). Two distinct runs of the branch site model, allowing omega to vary in one run, and fixing omega to 1 in the second run, produced negative log likelihood scores that were compared in a nested likelihood ratio test, using one degree of freedom.

2.3.4 Homology modeling and electrostatics

Structural comparisons of BbCRASP-1, bba69, and the bba68 ortholog from *B. bissettii* (DN127_lp54_ORF70) were conducted using homology modeling and electrostatics. The crystal structure of BbCRASP-1 (PDB: 1W33), coded by bba68, which shares 50% sequence identity with bba69 and greater than 90% sequence identity with DN127_lp54_ORF70, served as a template for modeling the target structures of bba69 and DN127_lp54_ORF70. Homology models were built using the automated

Phyre web server (Bennett-Lovsey, Herbert et al. 2008) and PROSA (Wiederstein, M. and M. J. Sippl 2007) structural evaluations were performed. A homodimer configuration of the targets was assumed due to the established homodimer state of their template relative, BbCRASP-1 (), and achieved via CE super-positioning (Shindyalov and Bourne 1998) of two copies of a target homology model onto the BbCRASP-1 homodimer. Electrostatic profiles of all three homodimers were generated and visualized using GRASP (Nicholls, Sharp et al. 1991).

2.4 RESULTS

2.4.1 The BbCRASP-1 homolog tree

The homolog tree of BbCRASP-1 exhibits five major clades (figure 2.1) (Wywial, Haven et al. 2009). Four clades represent distinct lineages containing the strain B31 orthologs: bba64, bba65, bba66, and bba73. The remaining clade contains bba68 (BbCRASP-1) and its closest homologs, which include homologs that do not reside on lp54. It is the largest clade and is unbalanced (no 1:1 orthology) among the *Borrelia* strains examined in this study (a genome synteny diagram of Pfam54 is given in the previous chapter; figure 1.3). A *B. afzelii* sequence was chosen as an outgroup because of its distant relation to the four, core Pfam54 homologs, and because it does not have orthologs in any other strain. Therefore, this sequence is likely an ancestral Pfam54 gene, which was lost to all other genospecies. This tree is congruent to an earlier, smaller tree produced by Hughes and colleagues (Hughes, Nolder et al. 2008).

2.4.2 Positive selection in the ancestral BbCRASP-1

The bba68 ortholog subtree (figure 2.1) was examined for positive selection on the ancestral branch of *B. burgdorferi sensu stricto*. Our motivation for examining this

branch stems from the fact that all human Lyme disease cases in the United States are caused by *B. burgdorferi sensu stricto* (Piesman and Gern 2004), and the structural state of BbCRASP-1 within this clade may be a contributing factor. Positive selection was detected along the ancestral branch with statistical significance of P=0.025. The positively selected amino acid sites were mapped onto the crystal structure of BbCRASP-1 (PDB ID: 1W33), and a corresponding, annotated alignment is presented in figure 2.2. The alignment illustrates the sequence locations of the putative binding cleft and dimerization region of BbCRASP-1. Shown are positively selected codons manifesting multiple transversion mutations, which are unlikely under neutral evolution. In particular, several codons in the annotated regions are under positive selection: Lysine 99 and Tyrosine 113 within the putative binding cleft, and Alanine 257 in the dimerization region. The most significant amino acid property change is on Lysine 99, which manifests one transition and two transversions; GCT (Valine) to AAG (Lysine).

2.4.3 Homology modeling supports adaptive evolution in BbCRASP-1

The protein structure models of bba69 and DN127_lp54_ORF70, and the crystal structure of BbCRASP-1 (bba68) were contoured with electrostatic profiles and compared visually (figure 2.3). The strongest distinction among these structures is the positively charged area within the putative binding cleft in BbCRASP-1. This charge differential corresponds to the positively selected sites within region 1 (figure 2.3). Strikingly, there is a strong positive charge within the cleft of BbCRASP-1, not present in bba69 or DN127_lp54_ORF70, which represent the closest bba68 paralog and ortholog, respectively.

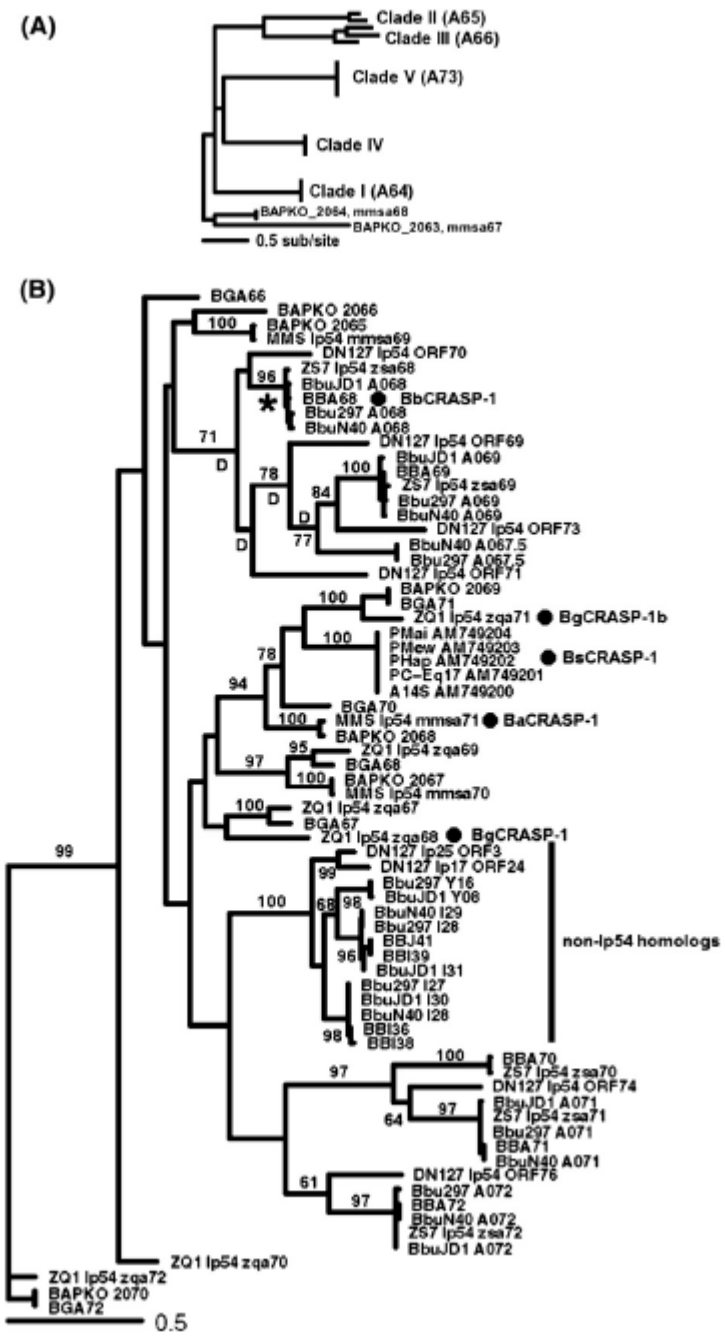


Figure 2.1. The BbCRASP-1 homolog trees. A. MrBayes protein tree depicting the five major clades. This tree is rooted using the *B. afzelii* homolog sequence (see Materials and Methods for explanation). B. The full MrBayes protein tree of BbCRASP-1 homologs using the same outgroup sequence as in A. The '*' denotes the ancestral branch of the sensu stricto BbCRASP-1 tree, which is labeled as the foreground branch in the branch site test of positive selection. Adapted from Wywial, Haven et al. 2009.

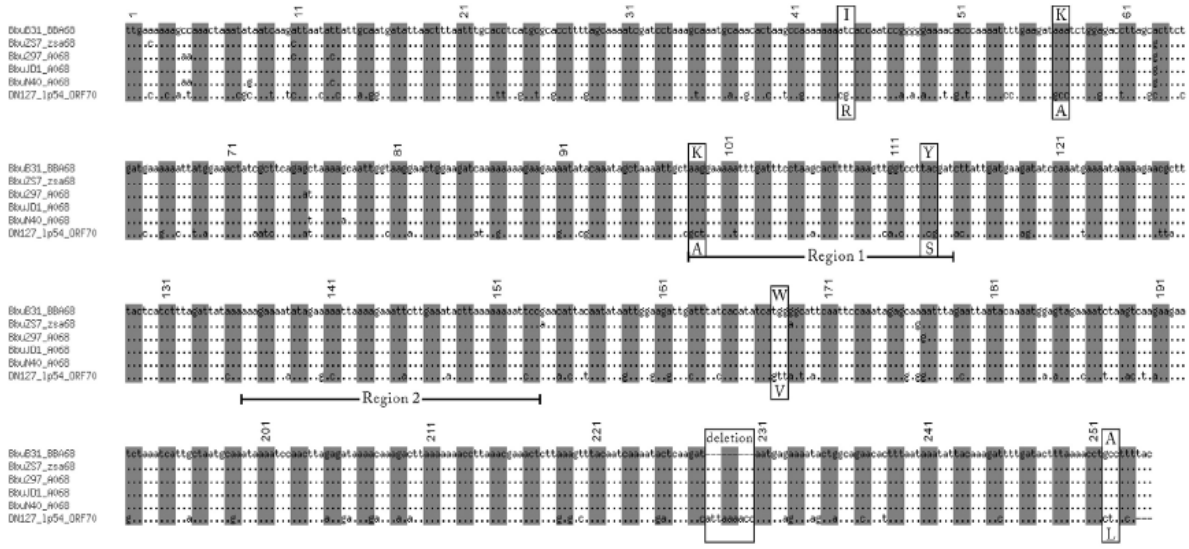


Figure 2.2. Annotated alignment of positive selection on bba68 orthologs. Shown is a codon alignment of bba68 homologs from B. burgdorferi sensu stricto and B. bissettii. Regions 1 and 2 form the putative binding pocket on BbCRASP-1. Boxes with amino acid letters denote positively selected sites. Adapted from Wywial, Haven et al. 2009.

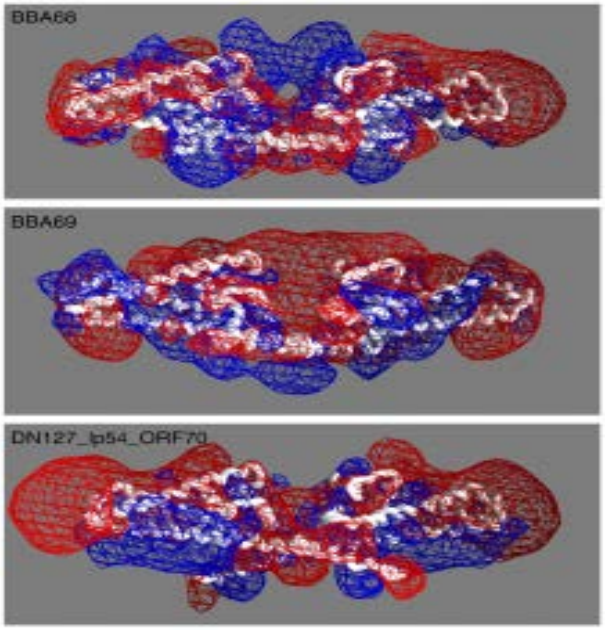


Figure 2.3. Electrostatic profiles of select BbCRASP-1 homologs. Bba68 is a crystal structure, while bba69 and DN127_lp54_ORF70 are homology models. The putative binding cleft is in the top-center of the molecule in each panel. Only bba68 (which codes BbCRASP-1) is positively charged. Structures are represented as C-alpha backbone traces with the binding cleft pointed

upwards. Electrostatic potentials were calculated and visualized in GRASP. PROSA evaluation z-scores for bba69 and DN127_lp54_ORF70 were -6.41 and -7.11, respectively. Adapted from Wywial, Haven et al. 2009.

2.5 DISCUSSION

2.5.1 Constancy and variability in the Pfam54 clades

The major Pfam54 clades in figures 2.1 and 1.3 show the strict orthologous nature of clades I, II, III, and V; each genome contributes one gene to each clade. However, clade IV presents multiple copies from each genome, and is the apparent result of frequent gene duplications. Genes from each clade are differentially expressed during the infection cycle, and considering the ubiquitous presence of all five clades in every genome, it is practical to assume all clades are necessary for survival in a natural environment.

The variability in gene content of clade IV suggests this clade may be a focal point for strain advantage. For example, if all competing strains possess the same set of surface proteins then there is no competitive advantage with respect to evading host immunity, but if one strain contains an additional surface protein (or lacks one surface protein), then natural selection may increase the basic reproductive ratio of that particular strain. In support of this notion, the intraspecific sequence divergence of Pfam54 genes is very low, strengthening the role of genome content over nucleotide composition in escaping host innate immunity. One way to test this hypothesis is to track the prevalence of strains with distinct clade IV repertoires in a natural environment over time while assuming strain dynamics are governed by frequency dependent selection. However, it is unclear how much time (years) would be required.

2.5.2 The affinity of BbCRASP-1 to human factor H

A strain-comparative, structural examination of positive selection on the crystallographic structure of BbCRASP-1 is quite revealing. BbCRASP-1 functions as a homodimer facilitated by the last 10 residues at the carboxy terminus (Cordes, Roversi et al. 2005). Once dimerized, the protein possesses a centralized cleft, which has been postulated as the factor H binding site. Previous analyses involving site-directed mutagenesis and positive selection have identified several key areas on the protein (Cordes, Kraiczy et al. 2006). Thus far, positive selection analyses on this molecule have not involved a lineage, or clade based hypothesis. We conducted a branch-site analysis hypothesizing positive selection in the *sensu stricto* ancestral node, while using the *B. garinii* BbCRASP-1 homologue, BGA66, as an outgroup. This hypothesis is motivated by the fact that no *B. bissettii* strains have been found to infect a human in the United States (Piesman and Gern 2004). Therefore, if the BbCRASP-1 sequence moiety providing affinity to human factor H is adaptive within *sensu stricto* (most likely incidentally), then positive selection should have occurred on the specified, ancestral branch. Furthermore, as mentioned earlier, BbCRASP-1 lies within the variable region of Pfam54 genes, and may be adaptively evolving for the purpose of evading host innate immunity. As shown in figure 2.3, positive selection within the putative binding cleft corresponds to a unique positive charge on BbCRASP-1. Furthermore, the ligand (factor H) possesses negatively charged SCR domains, which are probable binding sites for BbCRASP-1 – factor H interaction (Kotarsky, Hellwage et al. 1998; Lukacik, Roversi et al. 2004). Given our understanding that humans are not reservoir hosts for *Borrelia*, natural selection would be severely limited in facilitating a co-evolutionary state between

the two species. It is far more likely that *Borrelia* have adapted to factor H homologues within reservoir species such as the white-footed mouse, and these homologues are interchangeable with human factor H.

2.6 CONCLUSION

The Pfam54 virulence cassette contains five major lineages, one of which is variable (clade IV) among *Borrelia* strains. This variability may contribute to the virulence potential of a genospecies. Within clade IV in *sensu stricto* resides BbCRASP-1, an established binder of human factor H. Positive selection has been identified for the first time (in a lineage dependent manner) on this molecule, and the selective sites manifest multiple transversions, significant amino acid charge changes, and a positive electrostatic profile unique to BbCRASP-1. The positive selection occurring on the putative binding cleft of BbCRASP-1 is the likely result of adaptation to reservoir hosts, such as the white-footed mouse, and is incidentally compatible with human factor H due to the homology of factor H proteins among vertebrates.

Chapter III - Pervasive Localized Recombination in Lyme-Disease Pathogens Revealed by Population Genomic Analysis of Con-specific Strains

3.1 INTRODUCTION

3.1.1 Clonal frames and recombination in Borrelia burgdorferi

Borrelia burgdorferi sensu lato (*Bbsl*) is a bacterial species complex including agents of Lyme Disease, a tick-borne illness with a wide variety of clinical manifestations including neurologic, cardiac, skin, and joint abnormalities (Steere, Coburn et al. 2004). Lyme disease is a rising emerging infectious disease worldwide, as climate changes and tick habitats expands (Kurtenbach and Hanincova 2006). *Bbsl* isolates are classified into genomic groups ("genospecies") based on their DNA sequences. Over ten *Bbsl* genospecies, often geographically distinct, have been identified worldwide (Kurtenbach and Hanincova 2006). *B. burgdorferi sensu stricto* (*Bbss*) is the predominant genospecies in the Northeastern and Midwestern United States and exists in Eurasia with much less intensity.

The genetic structure of local *Bbsl* populations is largely clonal, in which the number of theoretically possible multilocus genotypes vastly exceeds the number of genotypes actually observed from the field (Bunikis, Garpmo et al. 2004). However, a variety of evolutionary mechanisms other than a lack of recombination may lead to clonal population structures in bacteria (Maynard Smith, 1993). Previously, using multilocus sequencing of multiple isolates within the same clonal complexes, we showed that the rate of recombination might in fact be higher than the mutation rate (Qiu, Schutzer et al. 2004). We have argued further that clonality of *Bbsl* populations in the Northeastern US is a consequence of rapid population growth rather than an actual lack of recombination

(Qiu, Bruno et al. 2008). Recently, recombinant *Bbss* genotypes have been reported independently in Midwestern and Western US and Europe where the *Bbss* populations appear to be historically stable (Margos, Gatewood et al. 2008).

Here, we used recently sequenced con-specific genomes (Schutzer, Fraser-Liggett et al. 2010) to investigate the nature of recombination in *Bbsl*, such as the frequency, hotspots, and sizes of horizontal gene transfer events. We report results of a linkage analysis among fourteen *Bbss* strains in the core, most conserved parts of their genomes, consisting of the main chromosome, the linear plasmid lp54 and the circular plasmid cp26. We found that recombination in *Bbsl* almost exclusively takes the form of localized gene conversion in which mostly small (<1 kilobase) DNA fragments are laterally transferred among strains. While surface protein loci appear to be recombination hotspots, such appearance is more likely a result of selective maintenance of sequence diversity at these loci. We conclude that recombination among co-existing strains is the main and adequate mechanism of *Bbsl* acquiring beneficial genetic diversity necessary for its adaptive evolution in nature.

3.2 MATERIALS AND METHODS

3.2.1 Strains, genomes, and orthologs

Genome sequences of twenty-three *B. burgdorferi sensu lato* (*Bbsl*) isolates were downloaded from GenBank and we focused on fourteen genomes of a single genospecies, *B. burgdorferi sensu stricto* (*Bbss*) (Casjens, Palmer et al. 2000; Glockner, Lehmann et al. 2004; Schutzer, Fraser-Liggett et al. 2010). The fourteen *Bbss* genomes represent a majority of clonal groups in North America (B31, N40, 297, JD1, 64b, 156a, 118a, 94a, 72a, 29805, CA-11.2A, and WI91-23) and two clonal groups in Europe (ZS7 and Bol26)

(Qiu, Bruno et al. 2008; Margos, Gatewood et al. 2008; Travinsky, Bunikis et al. 2010). The non-*Bbss* genomes included two *B. afzelii* strains (PKo and ACA-1), three *B. garinii* strains (PBi, PBr, and Far04), one *B.spielmanii* strain (A14S), one *B. bisettii* strain (DN127), and one *B. valaisian* strain (VS116), and a strain from an unnamed genospecies (SV1). Except for DN127, all non-*Bbss* strains are European isolates. ORFs were identified by using GLIMMER with a minimum length of 150 bases (Delcher, Harmon et al. 1999). For each of the three replicons (main chromosome, lp54, and cp26), we identified orthologous ORFs by first clustering them into homologous families using all-against-all BLASTp (Altschul, Madden et al. 1997) followed by MCL (Enright, Van Dongen et al. 2002). Orthologs were subsequently distinguished from paralogs and validated by visual inspection of the gene order on a customized synteny map. Protein sequence alignments were constructed using ClustalW 1.83 (Thompson, Higgins et al. 1994). Codon alignments were derived from protein alignments using customized PERL scripts based on the BioPerl programming library (http://bioperl.org). DNA and protein sequence alignments of ortholog families are available from the website http://borreliagenome.org. Whole-plasmid sequence alignments were obtained by using ClustalW-MPI (Li, 2003).

3.2.2 Linkage analysis

We calculated recombination rates between all pairs of SNP sites by using the program LDhat (version 2.1) [McVean, 2002 #822] (McVean, Awadalla et al. 2002). SNP pairs showing >2.0 or <2.0 log likelihood scores in a likelihood test included in LDhat were regarded as significant for the presence or absence of recombination between two SNPs, respectively. Results of linkages analysis were formatted by using customized

PERL scripts and plotted using the software packages Circos (Krzywinski, Shein et al. 2009) and R (http://r-project.org).

3.2.3 Coalescence simulations

An event of genetic recombination between two homologous DNA strands may result in either an exchange of strands surrounding a single break point ("crossover") or incorporation of a DNA tract from one strand into the other ("gene conversion") (Wiuf and Hein 2000). To understand mechanisms underlying the observed patterns of linkage disequilibrium among *Bbss* strains, we generated simulated 10,000 nucleotide-long sequence alignments under various models of recombination. The sequences were generated using the coalescence sampler, *ms* (Hudson 2002) followed by the program Seq-Gen (Rambaut and Grassly 1997). A sample of fifteen sequences was generated under each of the three models of evolution: an absence of recombination, presence of crossover (with the use of *–r* option in *ms*), and presence of gene conversion (the *–c* and *–l* options in *ms*). In each model, the sequences were generated based on the Jukes-Cantor model of nucleotide substitutions (Jukes and Cantor 1969) and a per-site scaling factor of 0.005, resulting in approximately 200 SNPs or a density of about 2%. This SNP density matches that of empirically collected plasmid and ORF alignments (see Results).

3.3 RESULTS

3.3.1 Genomic and ortholog alignments

ORFs on the main chromosomes, cp26, and lp54 plasmids are virtually all syntenic in the sampled *Bbsl* genomes. There are two large-scale genome variations on the core genome, including the variable left chromosome end and the PFam54 gene cluster on lp54 plasmids (Huang, Robertson et al. 2004; Wywial, Haven et al. 2009). We

obtained a total of 73,451-base long genomic alignments of the cp26 plasmids and lp54 plasmids (the PFam54 gene cluster excluded) and a total of 837 alignments of orthologous ORF families on the three replicons with a total length of 989,679 nucleotides (Table 3.1).

Table 3.1 Genomic and orthologous alignments

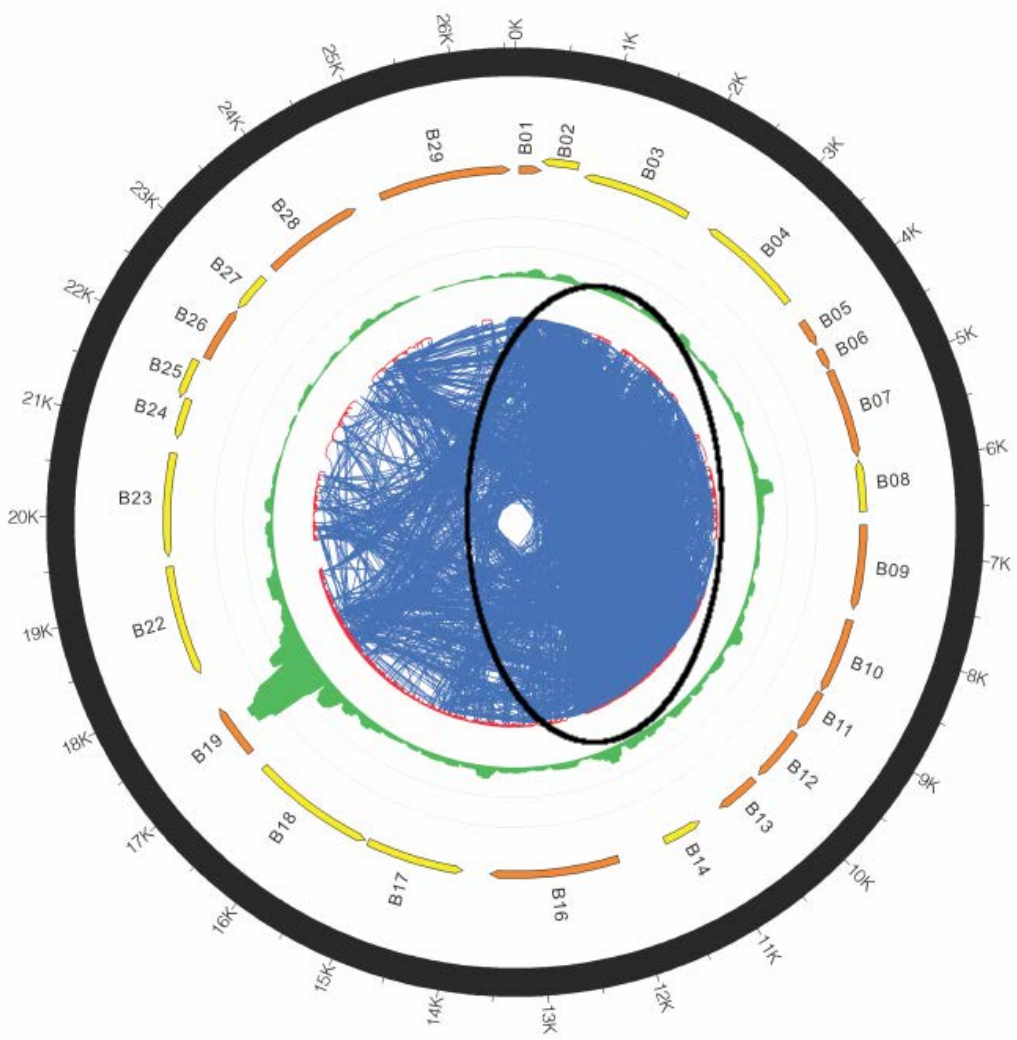
	Genomic Alignments			Orthologous ORF Alignments			
	Length (bases)	No. SNPs(a)	SNP density	No. ortholog families	Total alignment length (bases)	No. SNPs	SNP density
cp26	26,591	1,267	4.76%	26	26,451	826	3.12%
lp54(b)	46,860	870	1.86%	61	53,383	661	1.24%
Main chromosome	N.A.(c)	N.A.	N.A.	750	909,845	12,311	1.35%
Total	73,451	2,137	2.90%	837	989,679	13,798	1.39%

(a)Two-state sites only.
(b)PFam54 gene array excluded for lack of synteny
(c)Not available.

3.3.2 Genome-wide localized recombination

We first estimated recombination rates between pairs of SNP sites on individual replicons by using LDhat. Figure 3.1 shows the results for the cp26 genomic alignment. Pairs of SNPs showing significant recombination (red lines) were all localized within 500 bp to each other. The *ospC* locus and its surrounding regions showed the highest levels of sequence polymorphism and appeared to be a hotspot for localized recombination. Nevertheless, localized recombination footprints were found throughout the cp26 plasmid (figure 3.1). Results for the lp54 plasmid and the main chromosome (not shown) were

similar in showing localized recombination covering the entire replicon (figure 3.2A). In contrast, all SNP pairs situated more than 500 bp from each other have significantly low recombination rates (blue lines) (figure 3.1). Conspicuous among the tightly linked SNPs on cp26 is a contiguous region encompassing BBB01 through BBB14 (figure 3.1). On one side of this linkage block, the lack of linkage in the BBB22-BBB29 region appeared to be associated with low sequence polymorphisms or SNP density (figure 3.1). On the other side of the linkage block, however, the linkage breakdown appeared to be associated with high recombination rates in the *ospC*-flanking region (figure 3.1). No apparent contiguous linkage blocks were found on lp54 or the main chromosome.



B01 acylphosphatase
B02 unknown
B03 *resT*
B04 *chbC*
B05 *chbA*
B06 *chbB*
B07 α3-β1 integrin-binding protein
B08 lipoprotein
B09 lipoprotein
B10 PF62
B11 PF50
B12 PF32, plasmid partition
B13 PF49, plasmid partition
B14 unknown
B16 *oppAIV*
B17 *guaB*
B18 *guaA*
B19 *ospC*
B22 guanine/xanthine permease
B23 guanine/xanthine permease
B24 unknown
B25 unknown
B26 unknown
B27 lipoprotein
B28 ankyrin repeat protein
B29 PTS component

Figure 3.1. Localized recombination and nucleotide polymorphisms on cp26. Four data tracks are shown, representing (starting from outside) the circular plasmid cp26 (black, B31 coordinates), ORFs (orange and yellow, indicating opposite coding directions), average nucleotide diversity (green, 120-base window size and 3-base window step; lower and upper gray lines indicating scales of 0.15 and 0.3 differences per alignment site, respectively), and recombination rates between pairs of SNPs (a red line linking two SNPs with a high recombination rate and a blue line linking two SNPs with a low recombination rate. Recombination rates were obtained by using LDhat (see Materials and Methods). Localized recombination occurs throughout the plasmid and boosts local sequence diversity, e.g., at BBB08, BBB14, and BBB19 loci. The region

surrounding ospC has the highest sequence diversity and the highest recombination rates. Linkage among SNPs increases with their distances to ospC, creating a multilocous clonal frame between BBB01 and BBB14 (marked within oval).

3.3.3 A cross-plasmid linkage block

We further investigated genome-wide linkage in *Bbsl* by examining SNPs across all replicons. A joint analysis of SNPs on cp26 and lp54 revealed significant linkage between the two replicons, the strongest being the genetic association between the BBB01-BBB14 region with *dbpA* (coding for decorin-binding protein A) and BBA36 (coding for a putative lipoprotein) (figure 3.2A). This cross-plasmid linkage was unexpected because there was no significant linkage among SNPs within lp54 itself (figure 3.2A, upper-right quadrant). Joint analysis between all three replicons (cp26, lp54, and the main chromosome) was performed and no additional significant genome-wide linkage blocks were found.

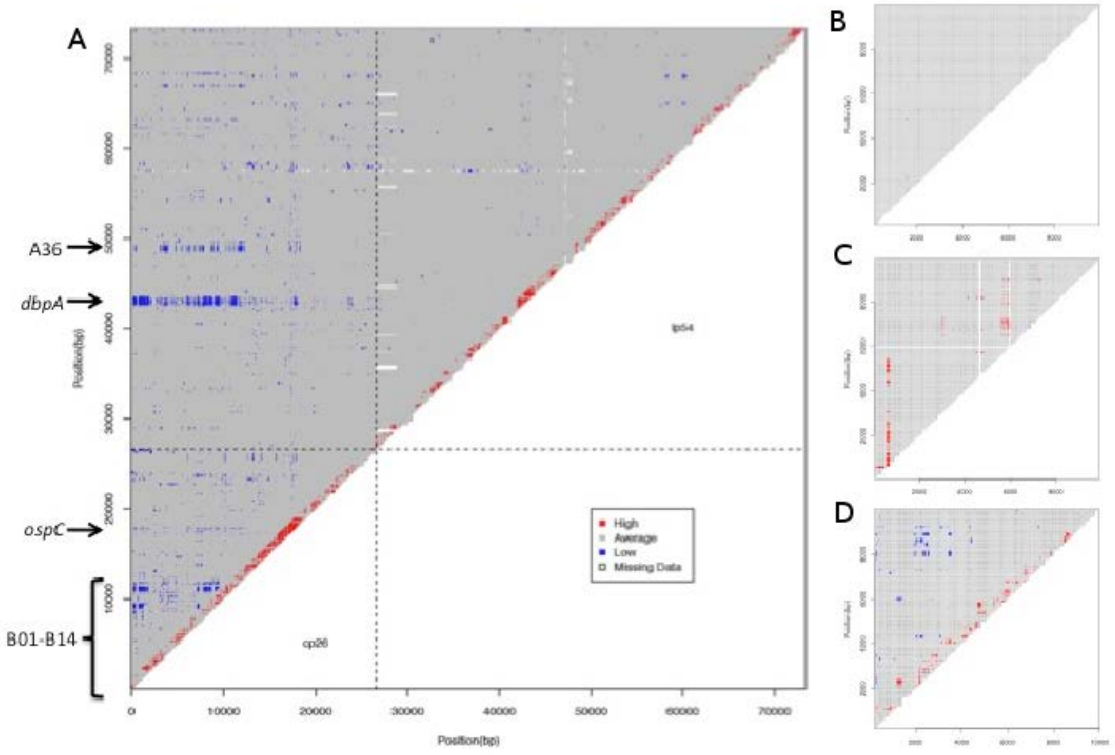


Figure 3.2. Coalescence simulations. Recombination patterns based on simulated sequences (Panels B, C, & D) were used to test the mechanism underlying the observed recombination pattern on cp26 and lp54 (Panel A). The observed pattern (Panel A) matches most closely with the simulated pattern with gene conversion (Panel D). Each panel depicts pair-wise recombination rates among SNP sites were obtained by using LDHat [McVean, 2002 #822]. Red, blue, and gray dots represent SNP pairs having, respectively, significantly high, significantly low, and average recombination rates. **(A)** Recombination rates among 1,487 SNPs on cp26 and lp54 plasmids from fourteen *Bbss* strains. Gene conversion occurs throughout the lp54 plasmid (upper-right quadrant) as well, with *dbpA* being a hotspot. A BBB01-BBB14 linkage block on cp26 is strongly associated with polymorphisms at *dbpA* and BBA36 (coding for a putative lipoprotein) on lp54. **(B)** Simulation without recombination. A sample of fifteen alleles, each 10,000-nucleotide long, was obtained by using *ms* [Hudson, 2002] and Seq-Gen [Rambaut and Grassly, 1997]. **(C)** Simulation with crossover. The crossover parameter ($\rho=2N_e r$) was set to 2.5, assuming a per-site crossover rate (r) of 10^{-7} and an effective population size (N_e) of 1250. **(D)** Simulation with gene conversion. The gene-conversion parameter ($c=2N_e g$) was set to 250, which corresponded to an effective population size (N_e) of 1250 and a per-site gene conversion rate (g) of 10^{-1}. A tract length of 100 bases was used.

3.3.4 Coalescence simulations

Localized high recombination rates suggest that gene conversion, rather than crossover, may be the main mechanism of recombination in *Bbsl*. Results of coalescence simulations, shown in figure 3.2 (B, C, and D), lent strong support for the hypothesis of gene conversion. Panel A is a heat map representing pair-wise recombination rates among SNPs on a concatenated alignment of cp26 and lp54 sequences. The graph shows localized recombination throughout the plasmid and, more visibly, the BBB01-BBB14 linkage block (figure 3.2A). In a coalescence-simulated sequence sample assuming no recombination, we found no such signals of localized recombination or linkage blocks (figure 3.2B). When crossover recombination was introduced into the simulation model, patches of recombination appeared in which SNP pairs located on opposite sides of a breakpoint showed as red spots, as one would expect (figure 3.2C). In the final simulation model, we introduced a high rate of gene conversion and recreated a linkage pattern similar to the one based on the genomic sample, in which recombining loci were

localized while linked loci were distantly located (figure 3.2D). The simulated sample, however, differed from the genomic sample in its lack of gene-conversion hotspots as well as a lack of contiguous blocks of strongly linked SNPs (figure 3.2A and 3.2D).

3.4 DISCUSSION

3.4.1 Recombination in B. burgdorferi: a genomic perspective

Conflicting views have been presented on the nature and frequency of recombination in *Bbsl*. Initially, *Bbsl* was found to be strictly clonal based on congruence among phylogenetic trees at multiple loci (Dykhuizen, Polin et al. 1993). The study, however, was based on inter-specific comparisons and a subsequent sampling of intra-specific isolates revealed lateral genetic exchanges between coexisting strains at a rate three times that of point mutations (Qiu, Schutzer et al. 2004). Earlier, evidence for lateral DNA transfers was established at individual loci such as *ospC* and Erp genes as well as on the cp32 plasmids (Livey, Gibbs et al. 1995; Stevenson, Casjens et al. 1998). Most recently, large-scale multilocus sequence typing of *Bbss* isolates from across the United States and Europe revealed significant geographic isolation, a likely European origin, and a mostly clonal population structure of this genospecies (Margos, Gatewood et al. 2008). Nevertheless, incongruence between individual gene trees and the overall multilocus phylogeny indicated that many *Bbss* isolates from the Midwestern US, Western US, and Europe contained horizontally transferred genes (Qiu, Bruno et al. 2008). We show here, based on this first genome-wide scan of recombination rates in *Bbsl*, that recombination is indeed the main source of genome diversity in *Bbsl*, particularly at surface antigen loci. With plenty of localized sequence diversity at antigen loci maintained by recombination and natural selection, a strong genome-wide clonality

in *Bbsl* thus does not seem to be a severe impediment to its adaptive diversification in the wild. While it has long been proposed that recombination in *Borrelia* and other bacterial species usually takes the form of "localized sex" (Smith, Dowson et al. 1991), our approach of a genome-wide scan of recombination using con-specific strains of a single bacterial species provides a corroboration to the genome-wide pervasiveness of gene conversion in bacteria, including a species considered to be highly clonal.

3.4.2 Selective retention of gene-conversion footprints at surface-protein loci

While surface-protein loci appear to be recombination hotspots in the *Bbsl* genome they are perhaps not inherently predisposed to recombination. This is because localized linkage equilibrium among SNPs occurs, in fact, throughout the genome (figure 3.1 & 3.2). It is known that nucleotide polymorphisms are retained at surface-antigen loci like *ospC* by balancing natural selection (Qiu, Dykhuizen et al. 2002). We confirmed here that *ospC* and *dbpA* had the highest significant proportion of nonsynonymous polymorphisms among all ORF loci (data not shown). In contrast, we found mostly synonymous SNPs at other regions affected by gene conversion, including loci flanking *ospC* and *dbpA* (data not shown). More directly, we observed at least three independent gene-conversion events at the *guaA*-*ospC*-BBB22 loci in a single *Bbss* genome (figure 3.1). These indirect and direct evidence of high gene-conversion rates at genome regions close to *ospC* and *dbpA* led us to believe that balancing selection alone may account for the high observed gene-conversion rates at surface-antigen loci and it is not necessary to assume them to be inherently prone to gene conversion. In other words, these surface-protein loci may have a higher *effective*, but not actual, gene-conversion rate.

3.4.3 Genome-wide linkage and a model of Borrelia evolution

We observed a strong, cross-replicon linkage group encompassing BBB01-BBB14 on cp26 and BBA24 and BBA36 on lp54 among the *Bbss* genomes (figure 3.2). On cp26, this linkage block is located opposite to *ospC*, suggesting that it is likely a result of breakdown of linkage at loci close to *ospC* (figure 3.1). Less obvious is the mechanism behind the association of this linkage block with two surface-protein loci on lp54, an entirely separate plasmid (figure 3.2). One possible cause is epistatic interactions among these loci, in which certain alleles of *dbpA* or BBA36 may function with only a subset of alleles within the BBB01-BBB14 linkage block. Such cross-relicon epistasis is unlikely considering that SNPs in the BBB01-BBB14 linkage block are mostly synonymous. More probably, polymorphisms at *dbpA* and BBA36 and those within the BBB01-BBB14 linkage block are phylogenetically correlated, both having accumulated independently within individual clonal lineages. We conclude that *ospC* plays an outsized role in initiating and maintaining new clonal lineages within *Bbsl* populations. Sequence diversity at *ospC*, by mutation and mostly by recombination, is preferably maintained by natural selection, as evidenced by heavy footprints of recombination and polymorphism at this locus (figure 3.1). It could be that there is a threshold of sequence diversity at *ospC* beyond which a strain becomes ecologically stabilized within a *Bbsl* population, with a half-life long enough for additional adaptive variations to accumulate in a run-away fashion at *ospC* and other surface-antigen loci (*e.g. dbpA* and BBA36). Since these additional selective variations accumulate quickly in individual *ospC*-anchored lineages, genome-wide linkage and clonality develops as a result of independent lineage evolution.

3.4.4 Evolutionary and practical implications

Although uncoupled from reproduction and often infrequent, homologous recombination and horizontal gene transfer are widespread in bacteria and important for species adaptation (Levin and Bergstrom 2000; Smith, Feil et al. 2000). In fact, recombination rate is a critical parameter in predicting the rate of adaptation in bacteria (Levin and Cornejo 2009). To understand whether sequence clusters in natural bacterial populations are maintained by natural selection or a lack of recombination, one approach is to study the process of bacterial genome divergence in the absence of natural selection (Fraser, Hanage et al. 2007). If recombination between strains is frequent, intra-specific subgroups are unstable. On the other hand, subgroups diverge irreversibly if recombination is rare. A simulation study based on neutral sequence variations and constant population sizes suggested that the threshold recombination rate above which bacterial subgroups become transient is about a quarter to twice of the mutation rate (Fraser, Hanage et al. 2007). Under this model, *B. burgdorferi* may be considered a sexual species; this would require a reliable recombination-to-mutation ratio. The large number of sequence clusters stably coexisting within local *Bbsl* populations is probably due to natural selection and rapid population expansion, rather than to a lack of recombination.

In practice, our results of pervasive gene conversion in *Bbsl* suggest cautions in estimating phylogenetic relations among *Bbsl* clonal groups as well as in identifying adaptive mutations at individual ORF loci. Presence of recombination violates nucleotide substitution models underlying likelihood-based methods of phylogenetic reconstruction and estimation of synonymous and nonsynonymous evolutionary rates (Anisimova,

Nielsen et al. 2003). It is therefore perhaps best to first screen for and exclude sequences affected by gene conversion before performing multilocus phylogenetic reconstruction. Other tests of positive natural selection are less affected by the presence of recombination, such as those based on comparisons between intra-specific and inter-specific rates of evolution (Jensen, Thornton et al. 2008).

3.5 CONCLUSION

Sequencing the genomes of multiple sympatric bacterial strains proves to be a powerful approach for uncovering the genetic mechanisms of recombination and mutation, estimating their rates, and identifying targets of natural selection. Genetic discontinuity of *Bbsl* in the form of co-existing clonal groups is a result of natural selection for antigenic diversity and the non-equilibrium status of many of its natural populations, rather than a lack of recombination. Our model of gene conversion and adaptive evolution at surface-protein loci helps to understand the ecological processes underlying *Bbsl* evolution in nature and will lead to more accurate inferences on its inter- and intra-specific phylogenies and improved methods identifying genes associated with human-virulent strains.

Appendix – Proposing Relaxed Selection in the *Borrelia* Pathogen System

ABSTRACT

A.1 INTRODUCTION

A.1.1 Relaxed selection

Vector-borne pathogens present a case of relaxed selection whereby a pathogen is required to alternate between a vertebrate host and an arthropod vector, during which natural selection on specific genes is reduced or eliminated. In such radically different environments the pathogen cannot easily evolve suboptimal or non-specific host-interaction factors in order to simultaneously exploit both the vector and host. The pathogen will likely require sets of genes that are dedicated to each environment. An interesting question arises with this strategy: how are genes that operate solely in one environment maintained during their tenure in the alternate environment? Such genes will alternate between natural selection and neutral evolution, and are prone to deleterious mutations. This strategy may carry a significant cost to the reproductive success of the pathogen, and may consequently affect its virulence.

Relaxed selection addresses the persistence of traits after the weakening or removal of a specific selective pressure (Lahti, Johnson et al. 2009). Examples include vision retention in the absence of strong light, and the retention of a protective trait (ex. armor) in a prey species following the decline or elimination of a predator species. In order to understand the mechanism(s) behind relaxed selection, it is necessary to develop theoretical models involving allele maintenance and population dynamics, at a minimum. Such models must then be confronted with real data. A recent review by Lahti and colleagues presents the history of relaxed selection, synthesizes the results of relevant

studies, and provides a general framework for modeling relaxed selection. The maintenance of the infection cycle of *B. burgdorferi* presents a very specific case of relaxed selection.

A.1.2 The evolutionary study of pathogens

The evolutionary study of vector-borne pathogens requires a particular focus on their complex transmission cycles in order to better understand disease emergence and the virulence spectrum. Pathogen genotypes and population sizes determine disease outbreaks, and are influenced by many factors that affect transmission such as host diversity, host population sizes, host behavior, host immune responses, and climate; to name a few (Kurtenbach, Hanincova et al. 2006; Koelle 2009). With respect to vector-borne pathogens there are additional, critical factors such as host reservoir competency and the dilution effect (LoGiudice, Ostfeld et al. 2003), vector population dynamics, vector immune responses, vector behavior, and the extrinsic incubation period of the pathogen (within-vector latency) (Ewald 1994; Brisson, Dykhuizen et al. 2008; Tsao 2009). The study of such complex phenomenon must draw from multiple areas including evolutionary theory, epidemiology, population genetics, molecular biology, phylogenetics, and comparative genomics. Comparative genomics is a recent contributor to infectious disease research and has great potential in evaluating the forces shaping pathogen evolution, namely, gene duplication and loss, mutation, recombination, genetic drift, and natural selection (Lefebure and Stanhope 2007).

When attempting to examine the virulence spectrum in a pathogenic species, a proper starting point is the tradeoff between pathogen growth and transmissibility, or more simply, the transmission cost (Ewald 1994). Transmission is the cornerstone of

pathogen evolution and is detailed in the basic reproductive value, R_0 (Lenski and May 1994). One basic transmission cost is host fitness reduction. A pathogen exhibiting maximum growth within a host (and generally greater virulence) will impose a heavy disease burden, which may subsequently harm its transmissibility, and will decrease its frequency with respect to more 'moderate' pathogen genotypes. A general approach to understanding the virulence spectrum is to examine the population frequencies of pathogen genotypes with virulence variants. A 'host fitness based' transmission cost should select for genotypes with intermediate virulence; a famous study of an Australian rabbit species suffering from Myxomatosis is a canonical example (Fenner and Woodroofe 1965). This theory is complicated in vector-borne and/or multi-host pathogens because host morbidity may not significantly affect pathogen transmissibility (Ewald 1994), or may result in suboptimal virulence (Woolhouse, Taylor et al. 2001).

A.1.3 Single-stage versus multi-stage genes

The utility of a vector raises difficulties in pathogen adaptation, particularly if the host is a mammal and the vector an arthropod. The pathogen must, at the very least, be adapted to both a vector and a host. These environments can be very different from each other (vertebrate vs. invertebrate), and the pathogen will likely require three sets of genes: 1) genes that permit colonization of the host; 2) genes that permit colonization of the vector; and 3) genes that can function in both environments. A gene set under natural selection in one environment may be neutrally evolving once the pathogen transitions to an alternate environment, and is thus prone to deleterious substitutions (from here forward, a gene that functions during one part of the transmission cycle is termed a

'single-stage gene'). We use these categories to model the evolution of single-stage genes in a population of spirochetes.

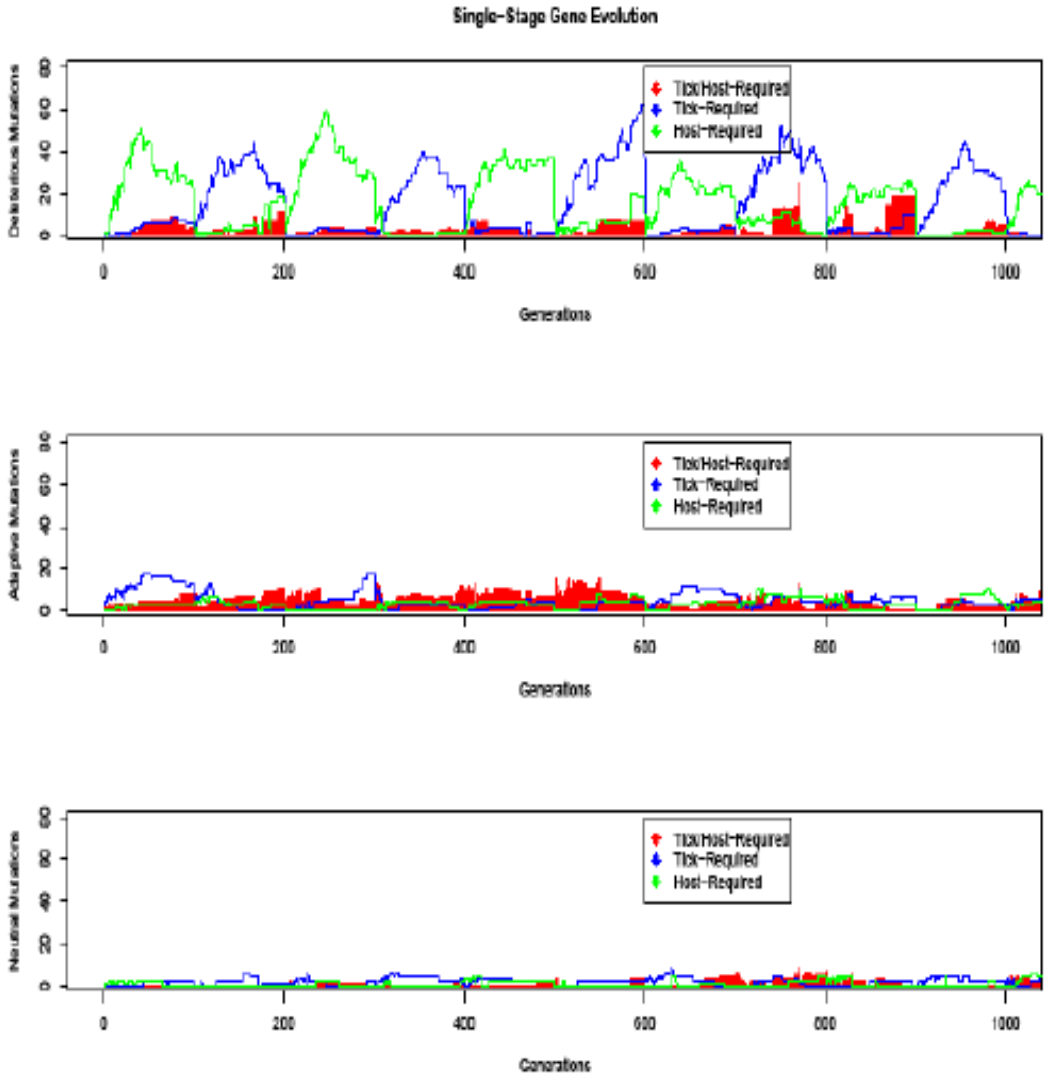


Figure A.1. An illustrative, stochastic simulation of single stage gene evolution. Panels from top to bottom display mutations that are deleterious, adaptive, and neutral in a population of spirochetes cycling between tick and host environments. Top panel exhibits cycling patterns of deleterious mutations on single stage genes (blue & green lines). Remaining lines on all panels exhibit flat-line patterning. This figure was illustrated using R graphics.

A.2 MATERIALS AND METHODS

A.2.1 Perl simulation

A simulation was carried out using the model variables described in section A.1.3 and the Perl programming language. The simulation begins with an initial population of 100 spirochetes, each with a genome of three genes: two single-stage genes and one multi-stage gene. The evolution of each gene is dependent on the life-cycle phase of the spirochete. The population cycles between tick and host environments for 1000 generations while undergoing mutation, growth, natural selection, and genetic drift. Mutation is implemented stochastically at probability of 0.1 for each gene. The effect of the mutation can be positive, negative, or neutral at the respective probabilities, 0.05, 0.90, and 0.05. Genetic drift is implemented by sampling 100 spirochetes from the growing population during each cycle. Sampling probabilities were generated using the Perl library, Math::Random.

A.3 RESULTS

A.3.1 Simulation study

Figure A.1 displays the results of a simulation study that demonstrates the expected effects of relaxed selection on single-stage genes (shown in blue and green), and these effects are in stark contrast with the evolution of a gene under continuous natural selection (shown in red). The model exhibits cyclical dynamics for single-stage genes, which is expected, and can now be extended to examine the effects of a wide range of cost-attenuators.

A.4 DISCUSSION

A.4.1 Evaluating the feasibility of studying relaxed selection in Borrelia

The cost of extreme pathogen generalism may be mitigated by recombinational DNA repair, constraints on the mutation rate, constraints on environment-specific growth rates (latency), or robust (mutation-tolerant) protein structure of single-stage genes. It is unclear how much of an effect any of these would have in maintaining the infection cycle of a vector-borne pathogen. Therefore, the appropriate next step is to incorporate one or more of these cost-attenuators into the model. When applicable, a cost-attenuator should be parameterized using real data, which would shed light on whether the *Borrelia* pathogen is a suitable system for studying relaxed selection in vector-borne pathogens. For instance, specifying a low mutation rate may be enough to negate the cyclical dynamics observed in figure 3.1. Furthermore, in the presence of multiple cost-attenuators, it is possible that the effects of single-stage gene evolution leave no detectable mark on the genome. Of great importance is the consideration of epistasis among single-stage and multi-stage genes. Traits are generally the end product of multiple genes functioning cooperatively. Furthermore, these genes may be involved in other traits unstudied. Such interactions can greatly affect the maintenance of a trait (Lahti, Johnson et al. 2009). More information is needed to assess the significance of epistasis in single-stage genes of vector-borne pathogens, particularly those genes that function to escape the host or vector immune responses. Ultimately, more theoretical work is needed in order to fine-tune the expectations of single-stage gene evolution before undertaking practical studies.

BIBLIOGRAPHY

Alberts, B. J., Alexander, Ed. (2002). Molecular Biology of the Cell, New York and London: Garland Science.

Altschul, S. F., T. L. Madden, et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res 25(17): 3389-3402.

Anisimova, M., R. Nielsen, et al. (2003). "Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites." Genetics 164(3): 1229-1236.

Attie, O., J. F. Bruno, et al. (2007). "Co-evolution of the outer surface protein C gene (ospC) and intraspecific lineages of *Borrelia burgdorferi* sensu stricto in the northeastern United States." Infect Genet Evol 7(1): 1-12.

Baker, N. A., D. Sept, et al. (2001). "Electrostatics of nanosystems: application to microtubules and the ribosome." Proc Natl Acad Sci U S A 98(18): 10037-10041.

Barthold, S. W., K. D. Moody, et al. (1988). "Experimental Lyme arthritis in rats infected with *Borrelia burgdorferi*." J Infect Dis 157(4): 842-846.

Bennett-Lovsey, R. M., A. D. Herbert, et al. (2008). "Exploring the extremes of sequence/structure space with ensemble fold recognition in the program Phyre." Proteins 70(3): 611-625.

Binnewies, T. T., Y. Motro, et al. (2006). "Ten years of bacterial genome sequencing: comparative-genomics-based discoveries." Funct Integr Genomics 6(3): 165-185.

Bissett, M. L. and W. Hill (1987). "Characterization of *Borrelia burgdorferi* strains isolated from Ixodes pacificus ticks in California." J Clin Microbiol 25(12): 2296-2301.

Brisson, D. and D. E. Dykhuizen (2004). "ospC diversity in *Borrelia burgdorferi*: different hosts are different niches." Genetics 168(2): 713-722.

Brisson, D., D. E. Dykhuizen, et al. (2008). "Conspicuous impacts of inconspicuous hosts on the Lyme disease epidemic." Proc Biol Sci 275(1631): 227-235.

Brooks, C. S., S. R. Vuppala, et al. (2006). "Identification of *Borrelia burgdorferi* outer surface proteins." Infect Immun 74(1): 296-304.

Bunikis, J., U. Garpmo, et al. (2004). "Sequence typing reveals extensive strain diversity of the Lyme borreliosis agents *Borrelia burgdorferi* in North America and *Borrelia afzelii* in Europe." Microbiology 150(Pt 6): 1741-55.

Burgdorfer, W., A. G. Barbour, et al. (1982). "Lyme disease-a tick-borne spirochetosis?" Science 216(4552): 1317-1319.

Bykowski, T., M. E. Woodman, et al. (2008). "*Borrelia burgdorferi* complement regulator-acquiring surface proteins (BbCRASPs): Expression patterns during the mammal-tick infection cycle." Int J Med Microbiol 298 Suppl 1: 249-256.

Casjens, S., N. Palmer, et al. (2000). "A bacterial genome in flux: the twelve linear and nine circular extrachromosomal DNAs in an infectious isolate of the Lyme disease spirochete *Borrelia burgdorferi*." Mol Microbiol 35(3): 490-516.

Chen, S. L., C. S. Hung, et al. (2006). "Identification of genes subject to positive selection in uropathogenic strains of Escherichia coli: a comparative genomics approach." Proc Natl Acad Sci U S A 103(15): 5977-5982.

Cooper, J. E. and E. J. Feil (2004). "Multilocus sequence typing--what is resolved?" Trends Microbiol 12(8): 373-377.
Cordes, F. S., P. Kraiczy, et al. (2006). "Structure-function mapping of BbCRASP-1, the key complement factor H and FHL-1 binding protein of *Borrelia burgdorferi*." Int J Med Microbiol 296 Suppl 40: 177-184.
Cordes, F. S., P. Roversi, et al. (2005). "A novel fold for the factor H-binding protein BbCRASP-1 of *Borrelia burgdorferi*." Nat Struct Mol Biol 12(3): 276-277.
de Kievit, T. R. and B. H. Iglewski (2000). "Bacterial quorum sensing in pathogenic relationships." Infect Immun 68(9): 4839-4849.
Delcher, A. L., D. Harmon, et al. (1999). "Improved microbial gene identification with GLIMMER." Nucleic Acids Res 27(23): 4636-4641.
Diaz, E., Ed. (2008). Microbial Biodegradation: Genomics and Molecular Biology, Caister Academic Press. .
Doh, S. T., Y. Zhang, et al. (2007). "Non-coding sequence retrieval system for comparative genomic analysis of gene regulatory elements." BMC Bioinformatics 8: 94.
Dolinsky, T. J., P. Czodrowski, et al. (2007). "PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations." Nucleic Acids Res 35(Web Server issue): W522-525.
Dykhuizen, D. E., D. Brisson, et al. (2008). "The propensity of different *Borrelia burgdorferi* sensu stricto genotypes to cause disseminated infections in humans." Am J Trop Med Hyg 78(5): 806-810.
Dykhuizen, D. E., D. S. Polin, et al. (1993). "*Borrelia burgdorferi* is clonal: implications for taxonomy and vaccine development." Proc Natl Acad Sci U S A 90(21): 10163-10167.
Eggers, C. H. and D. S. Samuels (1999). "Molecular evidence for a new bacteriophage of *Borrelia burgdorferi*." J Bacteriol 181(23): 7308-7313.
Ellegren, H. (2008). "Comparative genomics and the study of evolution by natural selection." Mol Ecol 17(21): 4586-4596.
Enright, A. J., S. Van Dongen, et al. (2002). "An efficient algorithm for large-scale detection of protein families." Nucleic Acids Res 30(7): 1575-1584.
Ewald, P. (1994). Evolution of Infectious Disease. New York, Oxford University Press, Inc.
Ewing, B. and P. Green (1998). "Base-calling of automated sequencer traces using phred. II. Error probabilities." Genome Res 8(3): 186-194.
Falush, D. and R. Bowden (2006). "Genome-wide association mapping in bacteria?" Trends Microbiol 14(8): 353-355.
Fenner, F. and G. M. Woodroofe (1965). "Changes in the Virulence and Antigenic Structure of Strains of Myoma Virus Recovered from Australian Wild Rabbits between 1950 and 1964." Aust J Exp Biol Med Sci 43: 359-370.
Fraser, C. M., S. Casjens, et al. (1997). "Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*." Nature 390(6660): 580-586.
Fraser, C., W. P. Hanage, et al. (2007). "Recombination and the nature of bacterial speciation." Science 315(5811): 476-80.

Gilmore, R. D., Jr., R. R. Howison, et al. (2008). "*Borrelia burgdorferi* expression of the bba64, bba65, bba66, and bba73 genes in tissues during persistent infection in mice." Microb Pathog 45(5-6): 355-360.
Giraud, A., I. Matic, et al. (2001). "Costs and benefits of high mutation rates: adaptive evolution of bacteria in the mouse gut." Science 291(5513): 2606-2608.
Glockner, G., R. Lehmann, et al. (2004). "Comparative analysis of the Borrelia garinii genome." Nucleic Acids Res 32(20): 6038-6046.
Glockner, G., U. Schulte-Spechtel, et al. (2006). "Comparative genome analysis: selection pressure on the Borrelia vls cassettes is essential for infectivity." BMC Genomics 7: 211.
Grimm, D., K. Tilly, et al. (2004). "Outer-surface protein C of the Lyme disease spirochete: a protein induced in ticks for infection of mammals." Proc Natl Acad Sci U S A 101(9): 3142-3147.
Gupta, S., N. Ferguson, et al. (1998). "Chaos, persistence, and evolution of strain structure in antigenically diverse infectious agents." Science 280(5365): 912-915.
Guttman, D. S. and D. E. Dykhuizen (1994). "Clonal divergence in Escherichia coli as a result of recombination, not mutation." Science 266(5189): 1380-1383.
Harvey, P. H. and A. Purvis (1991). "Comparative methods for explaining adaptations." Nature 351(6328): 619-624.
Hayashi, T., K. Makino, et al. (2001). "Complete genome sequence of enterohemorrhagic Escherichia coli O157:H7 and genomic comparison with a laboratory strain K-12." DNA Res 8(1): 11-22.
Holst, M., R. E. Kozack, et al. (1994). "Protein electrostatics: rapid multigrid-based Newton algorithm for solution of the full nonlinear Poisson-Boltzmann equation." J Biomol Struct Dyn 11(6): 1437-1445.
Hovius, J. W., A. P. van Dam, et al. (2007). "Tick-host-pathogen interactions in Lyme borreliosis." Trends Parasitol 23(9): 434-438.
Huang, X., M. D. Adams, et al. (1997). "A tool for analyzing and annotating genomic sequences." Genomics 46(1): 37-45.
Huang, W. M., M. Robertson, et al. (2004). "Telomere exchange between linear replicons of *Borrelia burgdorferi*." J Bacteriol 186(13): 4134-41.
Hudson, R. R. (2002). "Generating samples under a Wright-Fisher neutral model of genetic variation." Bioinformatics 18(2): 337-8.
Huelsenbeck, J. P. and F. Ronquist (2001). "MRBAYES: Bayesian inference of phylogenetic trees." Bioinformatics 17(8): 754-755.
Hughes, A. L. (2007). "Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level." Heredity 99(4): 364-373.
Hughes, J. L., C. L. Nolder, et al. (2008). "*Borrelia burgdorferi* surface-localized proteins expressed during persistent murine infection are conserved among diverse Borrelia spp." Infect Immun 76(6): 2498-2511.
Humphrey, W., A. Dalke, et al. (1996). "VMD: visual molecular dynamics." J Mol Graph 14(1): 33-38, 27-38.
Jensen, J. D., K. R. Thornton, et al. (2008). "Inferring selection in partially sequenced regions." Mol Biol Evol 25(2): 438-46.
Jones, D. T., W. R. Taylor, et al. (1992). "The rapid generation of mutation data matrices from protein sequences." Comput Appl Biosci 8(3): 275-282.

Jones, K. L., L. J. Glickstein, et al. (2006). "*Borrelia burgdorferi* genetic markers and disseminated disease in patients with early Lyme disease." J Clin Microbiol 44(12): 4407-4413.

Jukes, T. H. C., C.R., Ed. (1969). The Evolution of Protein Molecules. Mammalian Protein Metabolism. New York, Academic Press.

Koelle, K. (2009). "The impact of climate on the disease dynamics of cholera." Clin Microbiol Infect 15 Suppl 1: 29-31.

Koelle, K., S. Cobey, et al. (2006). "Epochal evolution shapes the phylodynamics of interpandemic influenza A (H3N2) in humans." Science 314(5807): 1898-1903.

Kotarsky, H., J. Hellwage, et al. (1998). "Identification of a domain in human factor H and factor H-like protein-1 required for the interaction with streptococcal M proteins." J Immunol 160(7): 3349-3354.

Kraiczy, P., E. Rossmann, et al. (2006). "Binding of human complement regulators FHL-1 and factor H to CRASP-1 orthologs of *Borrelia burgdorferi*." Wien Klin Wochenschr 118(21-22): 669-676.

Kraiczy, P., C. Skerka, et al. (2001). "Immune evasion of *Borrelia burgdorferi* by acquisition of human complement regulators FHL-1/reconectin and Factor H." Eur J Immunol 31(6): 1674-1684.

Krzywinski, M., J. Schein, et al. (2009). "Circos: an information aesthetic for comparative genomics." Genome Res 19(9): 1639-45.

Kumaran, D., S. Eswaramoorthy, et al. (2001). "Crystal structure of outer surface protein C (OspC) from the Lyme disease spirochete, *Borrelia burgdorferi*." EMBO J 20(5): 971-978.

Kurtenbach, K., S. De Michelis, et al. (2002). "Host association of *Borrelia burgdorferi* sensu lato--the key role of host complement." Trends Microbiol 10(2): 74-79.

Kurtenbach, K., K. Hanincova, et al. (2006). "Fundamental processes in the evolutionary ecology of Lyme borreliosis." Nat Rev Microbiol 4(9): 660-669.

Lahti, D. C., N. A. Johnson, et al. (2009). "Relaxed selection in the wild." Trends Ecol Evol 24(9): 487-496.

Lefebure, T. and M. J. Stanhope (2007). "Evolution of the core and pan-genome of Streptococcus: positive selection, recombination, and genome composition." Genome Biol 8(5): R71.

Lenski, R. E. and R. M. May (1994). "The evolution of virulence in parasites and pathogens: reconciliation between two competing hypotheses." J Theor Biol 169(3): 253-265.

Levin, B. R. and C. T. Bergstrom (2000). "Bacteria are different: observations, interpretations, speculations, and opinions about the mechanisms of adaptive evolution in prokaryotes." Proc Natl Acad Sci U S A 97(13): 6981-5.

Levin, B. R. and O. E. Cornejo (2009). "The population and evolutionary dynamics of homologous gene recombination in bacterial populations." PLoS Genet 5(8): e1000601.

Li, L., C. J. Stoeckert, Jr., et al. (2003). "OrthoMCL: identification of ortholog groups for eukaryotic genomes." Genome Res 13(9): 2178-2189.

Li, K. B. (2003). "ClustalW-MPI: ClustalW analysis using distributed and parallel computing." Bioinformatics 19(12): 1585-6.

Livey, I., C. P. Gibbs, et al. (1995). "Evidence for lateral transfer and recombination in OspC variation in Lyme disease Borrelia." Mol Microbiol 18(2): 257-69.
LoGiudice K, Ostfeld RS, et al. (2003). “The ecology of infectious diseases: effects of host diversity and community composition on Lyme disease risk.” Proc Natl Acad Sci USA 100(2):567-71.
Lukacik, P., P. Roversi, et al. (2004). "Complement regulation at the molecular level: the structure of decay-accelerating factor." Proc Natl Acad Sci U S A 101(5): 1279-1284.
Margos, G., A. G. Gatewood, et al. (2008). "MLST of housekeeping genes captures geographic population structure and suggests a European origin of *Borrelia burgdorferi*." Proc Natl Acad Sci U S A 105(25): 8730-8735.
Margulies, M., M. Egholm, et al. (2005). "Genome sequencing in microfabricated high-density picolitre reactors." Nature 437(7057): 376-380.
McDonald, J. H. and M. Kreitman (1991). "Adaptive protein evolution at the Adh locus in Drosophila." Nature 351(6328): 652-654.
McVean, G., P. Awadalla, et al. (2002). "A coalescent-based method for detecting and estimating recombination from gene sequences." Genetics 160(3): 1231-41.
Miller, J. C., K. von Lackum, et al. (2003). "Temporal analysis of *Borrelia burgdorferi* Erp protein expression throughout the mammal-tick infectious cycle." Infect Immun 71(12): 6943-6952.
Mongodin, E. F., J. B. Emerson, et al. (2005). "Microbial metagenomics." Genome Biol 6(10): 347.
Mora, M., C. Donati, et al. (2006). "Microbial genomes and vaccine design: refinements to the classical reverse vaccinology approach." Curr Opin Microbiol 9(5): 532-536.
Nicholls, A., K. A. Sharp, et al. (1991). "Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons." Proteins 11(4): 281-296.
Norris, D. E., B. J. Johnson, et al. (1999). "Population genetics and phylogenetic analysis of Colorado *Borrelia burgdorferi*." Am J Trop Med Hyg 60(4): 699-707.
Nowalk, A. J., R. D. Gilmore, Jr., et al. (2006). "Serologic proteome analysis of *Borrelia burgdorferi* membrane-associated proteins." Infect Immun 74(7): 3864-3873.
Ohnishi, J., B. Schneider, et al. (2003). "Genetic variation at the vlsE locus of *Borrelia burgdorferi* within ticks and mice over the course of a single transmission cycle." J Bacteriol 185(15): 4432-4441.
Ojaimi, C., V. Mulay, et al. (2005). "Comparative transcriptional profiling of *Borrelia burgdorferi* clinical isolates differing in capacities for hematogenous dissemination." Infect Immun 73(10): 6791-6802.
Ovcharenko, I., M. A. Nobrega, et al. (2004). "ECR Browser: a tool for visualizing and accessing data from comparisons of multiple vertebrate genomes." Nucleic Acids Res 32(Web Server issue): W280-286.
Pal, C., M. D. Macia, et al. (2007). "Coevolution with viruses drives the evolution of bacterial mutation rates." Nature 450(7172): 1079-1081.
Pal, U., A. M. de Silva, et al. (2000). "Attachment of *Borrelia burgdorferi* within Ixodes scapularis mediated by outer surface protein A." J Clin Invest 106(4): 561-569.
Pal, U., X. Li, et al. (2004). "TROSPA, an Ixodes scapularis receptor for *Borrelia burgdorferi*." Cell 119(4): 457-468.

Pallen, M. J. and B. W. Wren (2007). "Bacterial pathogenomics." Nature 449(7164): 835-842.
Piesman, J. and L. Gern (2004). "Lyme borreliosis in Europe and North America." Parasitology 129 Suppl: S191-220.
Piesman, J., T. N. Mather, et al. (1987). "Seasonal variation of transmission risk of Lyme disease and human babesiosis." Am J Epidemiol 126(6): 1187-1189.
Postic, D., N. M. Ras, et al. (1998). "Expanded diversity among Californian borrelia isolates and description of Borrelia bissettii sp. nov. (formerly Borrelia group DN127)." J Clin Microbiol 36(12): 3497-3504.
Qiu, W. G., D. E. Dykhuizen, et al. (2002). "Geographic uniformity of the Lyme disease spirochete (*Borrelia burgdorferi*) and its shared history with tick vector (Ixodes scapularis) in the Northeastern United States." Genetics 160(3): 833-49.
Qiu, W. G., S. E. Schutzer, et al. (2004). "Genetic exchange and plasmid transfers in *Borrelia burgdorferi* sensu stricto revealed by three-way genome comparisons and multilocus sequence typing." Proc Natl Acad Sci U S A 101(39): 14150-14155.
Qiu, W. G., J. F. Bruno, et al. (2008). "Wide distribution of a high-virulence *Borrelia burgdorferi* clone in Europe and North America." Emerg Infect Dis 14(7): 1097-104.
Rambaut, A. and N. C. Grassly (1997). "Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees." Comput Appl Biosci 13(3): 235-8.
Rappuoli, R. (2001). "Reverse vaccinology, a genome-based approach to vaccine development." Vaccine 19(17-19): 2688-2691.
Richman, D. D., T. Wrin, et al. (2003). "Rapid evolution of the neutralizing antibody response to HIV type 1 infection." Proc Natl Acad Sci U S A 100(7): 4144-4149.
Schwan, T. G., J. Piesman, et al. (1995). "Induction of an outer surface protein on *Borrelia burgdorferi* during tick feeding." Proc Natl Acad Sci U S A 92(7): 2909-2913.
Schutzer, S. E., C. M. Fraser-Liggett, et al. (2010)."Whole Genome Sequences of Thirteen Isolates of *Borrelia burgdorferi*." J Bacteriology [Epub ahead of print]
Shendure, J., G. J. Porreca, et al. (2005). "Accurate multiplex polony sequencing of an evolved bacterial genome." Science 309(5741): 1728-1732.
Shindyalov, I. N. and P. E. Bourne (1998). "Protein structure alignment by incremental combinatorial extension (CE) of the optimal path." Protein Eng 11(9): 739-747.
Smith, J. M., C. G. Dowson, et al. (1991). "Localized sex in bacteria." Nature 349(6304): 29-31.
Smith, J. M., N. H. Smith, et al. (1993). "How clonal are bacteria?" Proc Natl Acad Sci U S A 90(10): 4384-4388.
Smith, J. M., E. J. Feil, et al. (2000). "Population structure and evolutionary dynamics of pathogenic bacteria." Bioessays 22(12): 1115-22.
Sokurenko, E. V., D. L. Hasty, et al. (1999). "Pathoadaptive mutations: gene loss and variation in bacterial pathogens." Trends Microbiol 7(5): 191-195.
Steere, A. C., R. L. Grodzicki, et al. (1983). "The spirochetal etiology of Lyme disease." N Engl J Med 308(13): 733-740.
Steere, A. C., J. Coburn, et al. (2004). "The emergence of Lyme disease." J Clin Invest 113(8): 1093-101.

Stevenson, B., S. Casjens, et al. (1998). "Evidence of past recombination events among the genes encoding the Erp antigens of *Borrelia burgdorferi*." Microbiology 144 (Pt 7): 1869-79.

Stevenson, B. (2002). "*Borrelia burgdorferi* erp (ospE-related) gene sequences remain stable during mammalian infection." Infect Immun 70(9): 5307-5311.

Stevenson, B. and K. Babb (2002). "LuxS-mediated quorum sensing in *Borrelia burgdorferi*, the lyme disease spirochete." Infect Immun 70(8): 4099-4105.

Stevenson, B., N. El-Hage, et al. (2002). "Differential binding of host complement inhibitor factor H by *Borrelia burgdorferi* Erp surface proteins: a possible mechanism underlying the expansive host range of Lyme disease spirochetes." Infect Immun 70(2): 491-497.

Storz, J. F. (2005). "Using genome scans of DNA polymorphism to infer adaptive population divergence." Mol Ecol 14(3): 671-688.

Thompson, J. D., D. G. Higgins, et al. (1994). "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice." Nucleic Acids Res 22(22): 4673-4680.

Travinsky, B., J. Bunikis, et al. "Geographic differences in genetic locus linkages for *Borrelia burgdorferi*." Emerg Infect Dis 16(7): 1147-50.

Tsao, J. I. (2009). "Reviewing molecular adaptations of Lyme borreliosis spirochetes in the context of reproductive fitness in natural transmission cycles." Vet Res 40(2): 36.

van Belkum, A., M. Struelens, et al. (2001). "Role of genomic typing in taxonomy, evolutionary genetics, and microbial epidemiology." Clin Microbiol Rev 14(3): 547-560.

Wang, G., C. Ojaimi, et al. (2002). "Disease severity in a murine model of lyme borreliosis is associated with the genotype of the infecting *Borrelia burgdorferi* sensu stricto strain." J Infect Dis 186(6): 782-791.

Wiederstein, M. and M. J. Sippl (2007). "ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins." Nucleic Acids Res 35(Web Server issue): W407-410.

Wilske, B., V. Preac-Mursic, et al. (1993). "An OspA serotyping system for *Borrelia burgdorferi* based on reactivity with monoclonal antibodies and OspA sequence analysis." J Clin Microbiol 31(2): 340-350.

Wiuf, C. and J. Hein (2000). "The coalescent with gene conversion." Genetics 155(1): 451-62.

Woolhouse, M. E., L. H. Taylor, et al. (2001). "Population biology of multihost pathogens." Science 292(5519): 1109-1112.

Wormser, G. P., D. Brisson, et al. (2008). "*Borrelia burgdorferi* genotype predicts the capacity for hematogenous dissemination during early Lyme disease." J Infect Dis 198(9): 1358-1364.

Wormser, G. P., D. Liveris, et al. (1999). "Association of specific subtypes of *Borrelia burgdorferi* with hematogenous dissemination in early Lyme disease." J Infect Dis 180(3): 720-725.

Wu, M. and J. A. Eisen (2008). "A simple, fast, and accurate method of phylogenomic inference." Genome Biol 9(10): R151.

Wywial, E., J. Haven, et al. (2009). "Fast, adaptive evolution at a bacterial host-resistance locus: the PFam54 gene array in *Borrelia burgdorferi*." Gene 445(1-2): 26-37.
Yang, Z. (2006). "On the varied pattern of evolution of 2 fungal genomes: a critique of Hughes and Friedman." Mol Biol Evol 23(12): 2279-2282.
Yang, Z., R. Nielsen, et al. (2000). "Codon-substitution models for heterogeneous selection pressure at amino acid sites." Genetics 155(1): 431-449.
Yogev, D., R. Rosengarten, et al. (1991). "Molecular basis of Mycoplasma surface antigenic variation: a novel set of divergent genes undergo spontaneous mutation of periodic coding regions and 5' regulatory sequences." EMBO J 10(13): 4069-4079.
Zhang, J., R. Nielsen, et al. (2005). "Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level." Mol Biol Evol 22(12): 2472-2479.
Zipfel, P. F., R. Wurzner, et al. (2007). "Complement evasion of pathogens: common strategies are shared by diverse organisms." Mol Immunol 44(16): 3850-3857.