GENOMIC ANALYSES REVEAL PUTATIVELY PATHOGENIC GENES AND FUNCTIONAL ELEMENTS IN BORRELIA BURGDORFERI, THE LYME DISEASE BACTERIUM

by

Che Martin

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

2013

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
Date	Chair of Examining Committee
_____	Laurel Eckhardt
Date	Executive Officer

Jill Bargonetti

Shaneen Singh

Weigang Qiu

_____

Supervisory Committee

THE CITY UNIVERSITY OF NEW YORK

ABSTRACT

GENOMIC ANALYSES REVEAL PUTATIVELY PATHOGENIC GENES AND FUNCTIONAL ELEMENTS IN BORRELIA BURGDORFERI, THE LYME DISEASE BACTERIUM

By

Che Martin

Adviser: Professor Weigang Qiu

B. burgdorferi s.l. (*B. burgdorferi sensu lato*) represents a Gram-negative bacterial species complex that includes causative agents of Lyme disease. As an obligate parasite, *B. burgdorferi*'s persistence in nature depends on its innate ability to exist and survive in two distinct biological environments, the hard-bodied tick (vector) and small vertebrates (hosts), as it progresses through different stages of its enzootic cycle. To accomplish this feat, *B. burgdorferi* heavily depends on its ability to tightly regulate differentially expressed host- and vector-specific genes. However, very little is known about the genes and regulatory elements contributing to the pathogenicity of this increasingly prevalent pathogen. This is primarily due to the fact that *B. burgdorferi* is not a model organism and is difficult to culture and transform. Additionally, we were previously limited in our ability to do comparative genomic studies in this organism due to the unavailability of whole-genome sequences. The recent release of whole-genome sequences from 22 strains spanning 8 different genospecies of *B. burgdorferi* has since made it possible to conduct a comprehensive genome study in this organism. Here, we employ phylogenomics analyses (analysis of the genomes of a group of closely related species) on the core genome of *B. burgdorferi* in order to identify putative genes and functional elements that may contribute to the pathogenicity of this organism. We conducted three main analyses: i) test of positive natural

selection within the coding regions of the core genome replicons, ii) test of evolutionary constraints within the non-coding regions of cp26 and lp54, and iii) structural and phylogenetic comparison of OspA and OspB (Outer Surface Protein A and B), two paralogous virulence-associated lipoproteins. Consequently, we identified three genes putatively involved in adaptive host immunity, fifteen genes putatively involved in adaptive divergence, two new genes putatively under direct transcription RpoS (an alternative regulatory subunit of RNA polymerase), a number of putative *cis* and *trans*-acting regulatory elements, and fourteen fixed differences between OspA and OspB concentrated in the region proximal to the C-terminus barrel domain and the N-terminus globular domain. These findings highlight a number of coding and non-coding sequences that may contribute to the virulence of *Borrelia burgdorferi* in humans. These findings provide a basis for future experimental studies towards the discovery of therapeutic, diagnostic, and preventive approaches to Lyme disease.

ACRONYMS AND EVOLUTIONARY CONCEPTS

Positive selection	Selection in favor of an advantageous mutation
Negative or purifying selection	Selection against a deleterious mutation
Neutral evolution	Rate of evolution without being influenced by the effects of selection
Frequency-dependant selection	The ability of a gene to persist and to be passed on to the next generation depends on it relative frequency
Intergenic- polymorphisms	Count of fixed nucleotide difference between two species in non-coding region of DNA
Intergenic- divergence	Count of fixed nucleotide difference between two species in non-coding region of DNA
K_A	Rate of non-synonymous amino-acid substitution = # non-synonymous substitutions /total possible non-synonymous substitutions
K_S	Rate of synonymous amino-acid substitution = #synonymous substitutions/total possible synonymous substitutions
K_I	Rate of nucleotide substitution
ORF	Open reading frame
IGS	Intergenic spacer (non-coding regions)
PCIBs	Perfectly conserved intergenic blocks
RMSD	Root mean square deviation (measure of distance between atoms)

ACKNOWLEDGEMENTS

This dissertation is dedicated to my wife and children who supported and believed in me while I pursued my Ph.D. I sincerely thank all my family members who have nurtured me over the years; from infancy to adulthood. Many thanks to Dr. Weigang Qiu for his guidance, and to my committee/examination members: Dr. Jill Bargonetti, and Dr. Shaneen Singh, Dr. Chris Braun and Dr. Emmanuel Mongodin for their counsel and time. I would like to extend a special thank you to Dr. Benjamin Ortiz and Dr. Peter Lipkie for their advice and guidance. I would also like to thank both the Department of Biology at Hunter College and The Graduate Center of the City University of New York for the opportunity to pursue my research and study interests under their support. In closing, a gracious thank you to all the members of Dr. Qiu Evolutionary Bioinformatics Lab; it was a pleasure to work along such inspiring, innovative and motivated individuals.

CONTENTS

TABLE OF FIGURES

LIST OF TABLES

CHAPTER I – IDENTIFYING PUTATIVE GENES INVOLVED IN *BORRELIA burgdorferi* HOST DEFENSE VIA POSITIVE SELECTION ANALYSIS

INTRODUCTION

Borrelia burgdorferi sensu lato

DIVERSITY AND GENOMICS

B. burgdorferi s.l. (*B. burgdorferi sensu lato*) is a Gram-negative, spirochetal bacterial species complex. All species in this group are tick-borne obligate parasites of vertebrates (Brisson et al. 2012, Biesiada et al. 2012). Globally, this *Borrelia* complex is classified into at least 18 named and putative species. Each *B. burgdorferi s.l.* species consists of a large number of intra-specific and often co-existing genomic groups distinguished by molecular phylogeny. For example, *B. burgdorferi s.s.* (*B. burgdorferi sensu stricto)* contains at least twenty different intra-specific strains (Figure 1-1) (Margos et al. 2011, Brisson and Dykhuizen 2004).

The *B .burgdorferi* genome consists of two replicon sets, a conserved linear chromosome with a total length of approximately 900 kilobases, and an additional ~900 kilobases of circular and linear plasmids whose composition vary both within strains and between species (Glöckner et al. 2006, Casjens et al. 2000). The least variable subset of *B .burgdorferi s.l.* genomic content is considered the core genome and consists of three replicons; the main chromosome, linear plasmid lp54, and circular plasmid cp26. These replicons contain approximately 750, 62, and 26 orthologous gene families respectively across *B. burgdorferi s.l.* species. Of all the identified open reading frames (ORFs) in *B .burgdorferi s.l.* core genome, about half of them from the main chromosome code for uncharacterized hypothetical proteins, and more than 80% of

plasmid-borne ORFs code for unknown proteins with no apparent homologs outside the *Borrelia* genus (Glöckner et al. 2006, Casjens et al. 2000). Given our limited experiment-based understanding of *Borrelia*, comparative genomics provides a valuable approach for identifying genes and gene-regulatory elements putatively contributing to virulence and host-parasite interactions in this species complex. The recent release of thirteen *B .burgdorferi s.s.* and nine *B .burgdorferi s.l.* whole genome sequences brings the total number of completed or draft genome sequences to at least 24 (Casjens 2000, Casjens, Fraser-Liggett, et al. 2011, Casjens, Mongodin, et al. 2011, Fraser et al. 1997, Schutzer et al. 2011a, Schutzer et al. 2012, Schutzer et al. 2011b). This expanded dataset facilitates a comprehensive comparative genomic study with substantially improved statistical power over previous studies with a small number of genomes (Sukarna, 2010. *Molecular and Evolutionary Properties of Non-Coding Regions in Bacteria using Comparative Genome Bioinformatics*. PhD dissertation The City University of New York, (Eddy 2005). Although comparative genomics does not directly predict the molecular or cellular functions of ORFs, it can identify genes contributing to species adaptation by scanning for signals of positive natural selection (Ellegren 2008) (Luikart et al. 2003, Storz 2005). Here, we employ such analyses towards the identification of putatively functional ORFs within the core genome of *B .burgdorferi.*

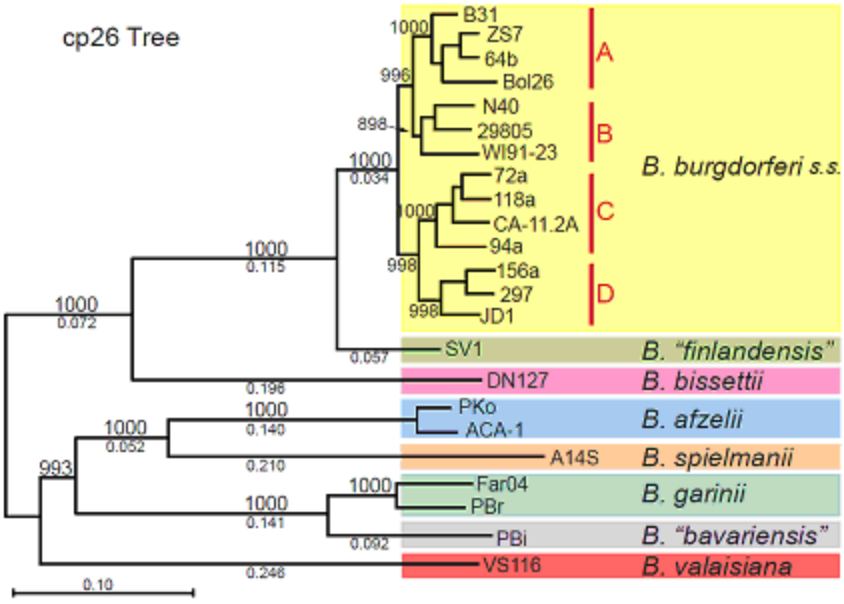


Figure 1-1: *B .burgdorferi s.l.* phylogeny reconstructed with high statistical confidence (bootstrap values greater than 90%) using SNPs across the plasmid cp26 in these genomes (Mongodin et al. 2013).

LYME DISEASE

Several species of the *B .burgdorferi s.l.* complex have been identified as causatives agents of Lyme disease. Lyme disease is a tick-borne disease that is increasing in prevalence throughout the United States, Europe and East Asia (Bacon et al. 2008, Margos et al. 2011, Stanek et al. 2012). Clinical manifestations include cardiac, joint, skin, and neurological abnormalities (Halperin 2012, Steere and Glickstein 2004). Three species, *B. garinii, B. afzelii, and B. burgdorferi s.s.* cause the majority of Lyme disease worldwide. In North America, Lyme disease is predominately caused by *B .burgdorferi s.s.* At least twenty different intra-specific clonal groups (strains) of *B .burgdorferi s.s.* coexists in Europe and North America, some of which are more likely than others to cause disseminated forms of Lyme disease (Dykhuizen et al. 2008, Brisson et al. 2011). Two main, and possibly co-occurring selective mechanisms cited as contributing to the high genetic diversity in natural *B. burgdorferi s.s.* are: i) host-niche

preferences and ii) negative frequency-dependant selection due to immune escape (Brisson and Dykhuizen 2004, Haven et al. 2011).

MECHANISM OF PATHOGENICITY

B .burgdorferi is an obligate parasite of vertebrates and maintained in natural populations through complex enzootic cycles consisting of ticks and mammalian vertebrates (Samuels 2011, Radolf et al. 2012). Differently expressed genes are crucial to *Borrelia* ability to survive in its tick vector, and successfully infect its mammalian host. Studies have highlighted three lipoprotein genes; *ospC* [outer surface protein C], *ospA* [outer surface protein A], and *dbpA* [decorin binding protein A] which are differentially expressed during tick to vertebrate transition, and directly contribute to *Borrelia's* ability to invade its vertebrate hosts via the RpoS-RpoN regulatory pathway (Samuels 2011, Xu, McShan, and Liang 2008).

OspA and OspC are among the best studied lipoproteins in *B .burgdorferi.* It has been noted that OspC production significantly increases, and OspA synthesis significantly decreases as *Borrelia* migrates from the midgut to salivary gland in its tick vector and is transmitted to its vertebrate host (Fingerle et al. 2002, Samuels 2011). Studies reveal that OspA plays a key role in anchoring the spirochete to the midgut of the tick and act as a protective agent against vertebrate host antibodies after a blood meal (Pal et al. 2004, Battisti et al. 2008). The lipoprotein OspC has been shown to be required for tick to vertebrate transmission and for the establishment of host infection (Fingerle et al. 2007, Grimm et al. 2004). Though the specific function OspC is still not fully understood, research shows that OspC binds to both mammalian plasminogen (Lagal et al. 2006) and a protein in the tick salivary gland (Ramamoorthi et al. 2005). Additionally, OspC synthesis is shut down early in mammalian infections to avoid adaptive

immune responses as it is an immunodominant antigen (Crother et al. 2004, Liang et al. 2002, Liang, Yan, et al. 2004).

Some studies suggest that presentation of OspC and OspA on the spirochete cell surface may be inversely correlated (Samuels 2011), however, these observations are not consistent universally. It should be noted that constitutive OspC production is observed in a null *ospAB* mutant suggesting some regulatory connection between these two lipoproteins (He et al. 2008).

The exact role of DbpA in B .*burgdorferi* is also not well defined. Recent findings suggest that DbpA is involved host cell adherence, and plays a crucial role in the spirochete ability to maintain persistent and disseminated infection in vertebrate hosts (Weening et al. 2008, Shi et al. 2008).

Other *Borrelia* lipoproteins cited to be functionally important include: CspA, a complement-regulator acquiring surface protein required for resistance to the host complement system (Brooks et al. 2005, Kenedy et al. 2009); Erps (OspE/F and Elp), genes noted to play a role in dissemination and host colonization (Kenedy and Akins 2011, Hefty et al. 2001) and VlsE, a variable surface antigen with a distinct role in immune evasion (Rogovskyy and Bankhead 2013)

The functionality of OspA, OspC and DbpA as well as the other lipoproteins mentioned above are consistent with the hypothesis that in *B .burgdorferi* lipoproteins constitute an important membrane component for protection against host innate immunity (Xu, McShan, and Liang 2008) and hence underscores their importance as virulence genes crucial to *B .burgdorferi* infectious life cycle. The identification of other lipoproteins and genes which contribute to B

.*burgdorferi* virulence via comparative genomics would have significant implications in biomedical research towards discovery of therapeutic targets of Lyme disease.

PATHOGEN COMPARATIVE GENOMICS

In this post-genomic era, the field of comparative genomics has experienced extensive progress due to the increased availability of sequenced genomes. As a result, Comparative genomics and bioinformatics have become essential tools with significant contributions to many disciplines of research including biotechnology (Díaz et al. 2008), epidemiology (van Belkum et al. 2001) and vaccine design (Mora et al. 2006). One field of research where the application of comparative genomics has grown increasingly popular is the study of pathogens. With the increased amounts of available pathogen genomes, it is now possible to compare sequences of bacterial strains of the same species, or those of a related species. Comparisons between bacterial genomes with different or similar pathogenic profiles permits for the identification of genes associated with pathogenicity. In these types of analyses, estimates of non-synonymous (K_A) and synonymous (K_S) substation rates are used to infer non-neutral evolution in genes. Under this model, neutral evolution is inferred when the ration of $K_A/K_S = 1$, while positive and purifying (negative) selection is inferred when $K_A/K_S > 1$ and $K_A/K_S < 1$ respectively (Yang 2007). For example, genes involved in host-pathogen interactions or virulence usually exhibit elevated levels of non-synonymous substitution and are often under adaptive evolution (Wywial et al. 2009).

In *Escherichia coli* and *Streptococcus,* both model organisms, comparative genomics methods were employed to identify virulent genes via genome-wide scans across multiple strains and species for positive selection (Chen et al. 2006, Lefébure and Stanhope 2007). Notably, the Lyme disease pathogen has not been subjected to such levels of rigorous intra, and inter-species

evolutionary comparative analysis. Additionally *B. burgdorferi* is not a model organism, and the limited means for genetic manipulation in this organism underscores the importance of comparative genomics in the identification of virulence, and functionally important genes. Results of such analyses will prove valuable to our understanding of *B. burgdorferi* pathogenicity.

MATERIALS AND METHODS

GENOME SEQUENCES AND GENE ORTHOLOGOUS DATA

The present study is based on three previously sequenced *B. burgdorferi s.l.* genomes including *B. burgdorferi s.s.* B31, *B. garinii* PBi and *B. afzelii* PKo (Fraser et al. 1997, Casjens et al. 2000, Glöckner et al. 2004, Glöckner et al. 2006) and 23 more recently completed genomes (Casjens, Fraser-Liggett, et al. 2011, Schutzer et al. 2012, Schutzer et al. 2011a, Casjens, Mongodin, et al. 2011). In total, our sample of genomes encompasses eight *B. burgdorferi s.l.* species including, in addition to the three pathogenic species listed above, *B. bavariensis*, *B. spielmanii*, *B. valaisiana*, *B. bissettii*, and *B. finlandensis* (Table 1).

Orthologous ORF sets used in our analyses were previously identified from 24 sequenced *B. burgdorferi s.l.* genomes (Haven et al. 2011). Here, ORFs were identified using GLIMMER3 (Delcher et al. 2007). Orthologous ORFs were then determined by clustering them into homologous protein families via an all-against-all BLASTp (Camacho et al. 2009), followed by MCL (Enright, Van Dongen, and Ouzounis 2002), a Markov cluster algorithm which assigns proteins into families based on precomputed sequence similarity information. This resulted in the identification of 750 orthologous protein-coding loci on the main chromosome (Figure S1), 26 loci on the cp26 plasmid, and 62 loci on the lp54 plasmid (Figure 1-2 and 1-3) (Haven et al.

2011). Short (<150 bases) ORFs with limited phylogenetic presence (available in less than 6 species) were removed from the analysis. Preliminary study of our ORF data revealed inconsistencies in the start-codon positions of orthologous ORFs from cp26, lp54, and the main chromosome. To address this, start codon positions were manually checked and synchronized for cp26 and lp54 based on the majority consensus of predicted start-codons for each orthologous ORF set. After the synchronization, customized Perl scripts; based on BioPerl (Stajich et al. 2002) were used to extract orthologous ORFs from whole genome sequences (Appendix 1). Nucleotide sequences of orthologous ORF were aligned according to alignment of translated protein sequences using the program CLUSTALW (Larkin et al. 2007). It should be noted that synchronization of the start-codon positions for main chromosome is currently in progress. Hence the main chromosome data used in the current study were not synchronized.

Strain	Genospecies	Biological Source[a]	Main Accession	Cp26 Accession	Lp54 Accession	Genome Report
B31	*Bb sensu stricto*	*I. scapularis*	AE000783.1	AE000792.1	AE000790.2	FRASER *et al.* 1997
64b	*Bb sensu stricto*	Human	PRJNA 28633	CP001422.1	CP001421.1	SCHUTZER *et al.* 2011
ZS7	*Bb sensu stricto*	*I. ricinus*	CP001205.1	CP001212.1	CP001199.1	SCHUTZER *et al.* 2011
JD1	*Bb sensu stricto*	*I. scapularis*	CP002312.1	CP002316.1	CP001652.1	SCHUTZER *et al.* 2011
CA-11.2A	*Bb sensu stricto*	*I. pacificus*	PRJNA28629	CP001484.1	CP001473.1	SCHUTZER *et al.* 2011
N40	*Bb sensu stricto*	*I. scapularis*	CP002228.1	CP002239.1	CP001651.1	SCHUTZER *et al.* 2011
72a	*Bb sensu stricto*	Human	PRJNA21003	CP001375.1	CP001370.1	SCHUTZER *et al.* 2011
156a	*Bb sensu stricto*	Human	PRJNA19835	CP001271.1	CP001257.1	SCHUTZER *et al.* 2011
WI91-23	*Bb sensu stricto*	Bird	PRJNA28627	CP001446.1	CP001447.1	SCHUTZER *et al.* 2011
118a	*Bb sensu stricto*	Human	PRJNA21001	CP001535.1	CP001542.1	SCHUTZER *et al.* 2011
297	*Bb sensu stricto*	Human	PRJNA29361	CP002268.1	CP001653.1	SCHUTZER *et al.* 2011
29805	*Bb sensu stricto*	*I. scapularis*	PRJNA28621	CP001550.1	CP001554.1	SCHUTZER *et al.* 2011
Bol26	*Bb sensu stricto*	*I. ricinus*	PRJNA19837	CP001568.1	CP001571.1	SCHUTZER *et al.* 2011
94a	*Bb sensu stricto*	Human	PRJNA20999	CP001493.1	CP001500.1	SCHUTZER *et al.* 2011
SV1	*B. finlandensis*	*I. ricinus*	PRJNA28631	CP001522.1	CP001524.1	CASJENS *et al.* 2011
DN127	*B. bissettii*	*I. pacificus*	PRJNA29363	CP002747.1	CP002761.1	SCHUTZER *et al.* 2011
PKo	*B. afzelii*	Human	CP002933.1	CP002934.1	CP002950.1	CASJENS *et al.* 2011
ACA-1	*B. afzelii*	Human	PRJNA19841	CP001250.1	CP001247.1	CASJENS *et al.* 2011
PBi	*B. bavariensis*	Human	CP000013.1	CP000014.1	CP000015.1	GLOCKNER *et al.* 2004
PBr	*B. garinii*	Human	PRJNA28625	CP001305.1	CP001308.1	CASJENS *et al.* 2011
Far04	*B. garinii*	Bird	PRJNA29573	CP001319.1	CP001318.1	CASJENS *et al.* 2011
VS116	*B. valaisiana*	*I. ricinus*	*ABCY02000001.1*	CP001432.1	CP001433.1	SCHUTZER *et al.* 2011
A14S	*B. spielmani*	*I. ricinus*	*PRJNA28635*	CP001467.1	CP001469.1	SCHUTZER *et al.* 2011

Table 1: *B. burgdorferi sensu lato* genomes

CP26 PLASMID MAP

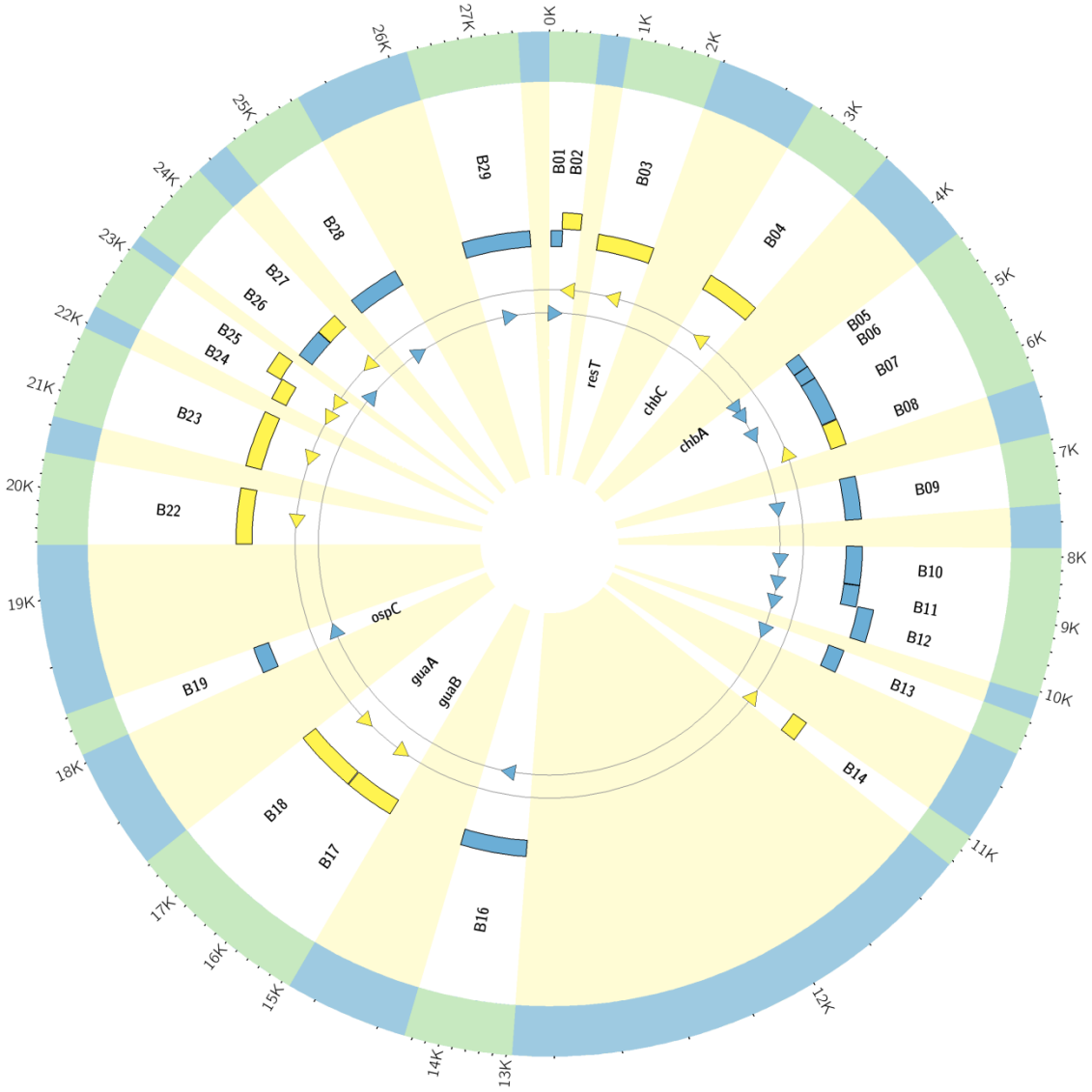


Figure 1-2: cp26 plasmid map illustrating ORFs included in our analysis. ORFs names are depicted next to their representative ORFs, while gene names (where available) are depicted in the innermost circle along their respective ORFs. ORFs are colored either yellow or blue based on their directionality (illustrated by arrowheads). ORFs in the reverse direction are donated by the color yellow, while forward direction ORFs in blue. Intergenic regions are depicted in khaki. Relative intergenic positions along the scale are shown in sky-blue while ORF positions are represented in green.

LP54 PLASMID MAP

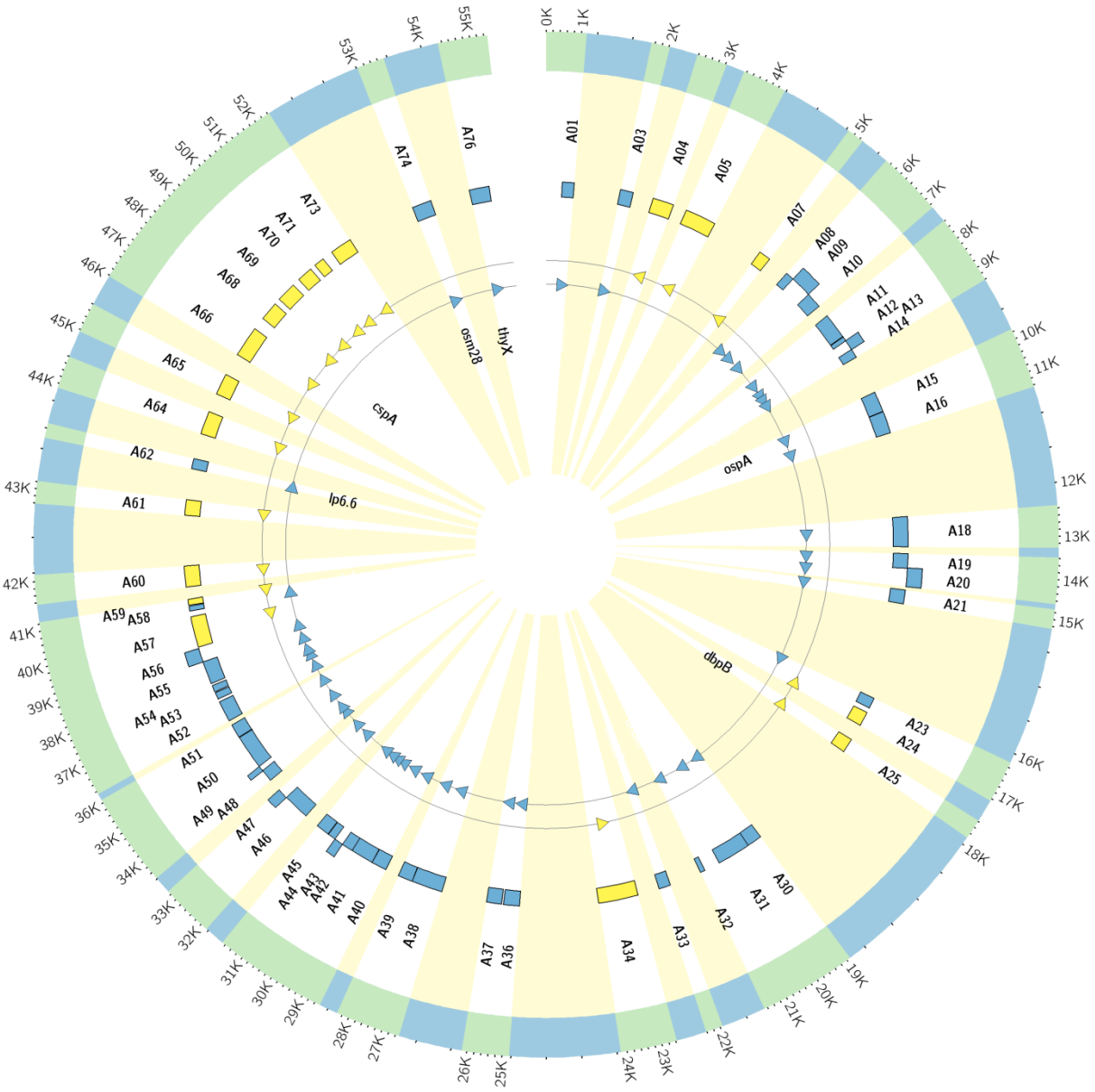


Figure 1-3: lp54 Plasmid map illustrating ORFs included in our analysis. As on the previous page, ORFs names are depicted next to their representative ORFs. Gene names (where available) are depicted in the innermost circle along their respective ORFs. ORFs are colored either yellow or blue based on their directionality (illustrated by arrowheads). ORFs in the reverse direction are donated by the color yellow, while forward direction ORFs are blue. Intergenic regions are depicted in khaki. Relative intergenic positions along the scale are shown in sky-blue while ORF positions are represented in green.

INFERENCE OF GENE GAIN/LOSS

Under the principle of parsimony, the ancestral character state of a sequence or gene is reconstructed from extant character states by a criterion which requires the fewest possible evolutionary changes (citation: Page & Holmes). Here if two nodes have the same character state, then their ancestral node is most parsimoniously assumed to have the same state. In the current study, evolutionary gains and losses of orthologous ORFs in cp26 and lp54 were reconstructed manually based on the parsimony principle and a phylogeny of sequenced genomes. This maximum-likelihood phylogenetic tree was reconstructed with high statistical confidence (bootstrap values greater than 90%) from an un-gapped alignment of twenty-two *B. burgdorferi s.l.* chromosomal conserved blocks. These blocks incorporated approximately 95.4% of the potential genetic information of the chromosomes (Mongodin et al. 2013).

EVOLUTIONARY ANALYSIS AND DETECTION OF POSITIVE NATURAL SELECTION

Both intra (within–species) and inter-specific (between-species) evolutionary analyses were conducted on orthologous ORFs via customized PERL and UNIX pipelines (illustrated in the Appendix). CODEML; a program of the PAML package (Yang 2007) was used to test for the presence of elevated K_A/K_S ratio in intra-specific ORFs (within-species analyses) and amino acid sites under positive selection in individual, inter-specific, orthologous ORF families (between-species analyses). The key PAML "site model" parameters include, "runmode = 0" for user-defined tree, which we used the chromosomal SNP tree (Mongodin et al. 2013); "model = 0" for uniform selective pressure across all branches; "NSsites = 0 1 2" for uniform 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 sites,

and positively selective sites ("M2a"). The log likelihoods of the M2a and M1a models were compared. ORF families showing significant (P≤0.001 by χ^2 test with 2 degrees of freedom) improvement in the log likelihood values of the positive selection model ("M2a") over that of the nearly neutral model ("M1a") were considered to have evolved under positive selection (Appendix I).

Being fully aware that poor alignments and uncertainty in ORFs calls (unsynchronized start-codon positions; main chromosome) could result in PALM falsely reporting high sequence variability. Sequence alignments of ORFs identified as significant in either the within or between-species analysis were manually inspected, and ORFs with poor alignments were omitted from our results.

RESULTS

GENE GAINS AND LOSSES ON lp54

Parsimony analysis revealed extensive gene gains and losses on the lp54 plasmid (Figure1-4). Here, PF54; the largest paralogous gene family in *B. burgdorferi s.l.,* showed the most variability both within *B. burgdorferi s.s.* strains, and between *B. burgdorferi s.l.* species. Within *B. burgdorferi s.s., a67.5* (orange arrows), an apparent species-specific gene gain due to its absence in between-species strains, seem to be independently lost in both JD1 and Clade A isolates. The alternative explanation, which requires a67.5 being gained in three independent lineages, is much less evolutionarily parsimonious. Another interesting observation within *B. burgdorferi s.s* is an apparent gain of the gene *a70* (*blue arrows*) in Clade A strains only, most likely through a duplication within the PFam54 gene array. Aside from *a66* and *a73*, there are no clear orthologous gene groups between *B. burgdorferi s.s* and other species. Truncated orthologs

of *B. bissettii a72* is also observed in both *B. finlandesis SV1* and all *B. burgdorferi s.s.* strains. The species group consisting of *B. bavariensis* (PBi) and *B. garinii* (PBr and Far04) share a relatively similar PF54 synteny, while the species group consisting of *B. spielmanii* A14S and *B. afzelii* (ACA1 and PKo) share another distinct set of orthologs. Only one ortholog set (*pink arrows*) is shared between these two species groups. Outside PF54, OspB [*a16*] (green arrows) is present in all *B. burgdorferi s.l.* species, with the exception of *B. garinii* where we observe an apparent gene loss. On cp26, gene composition was conserved across all species.

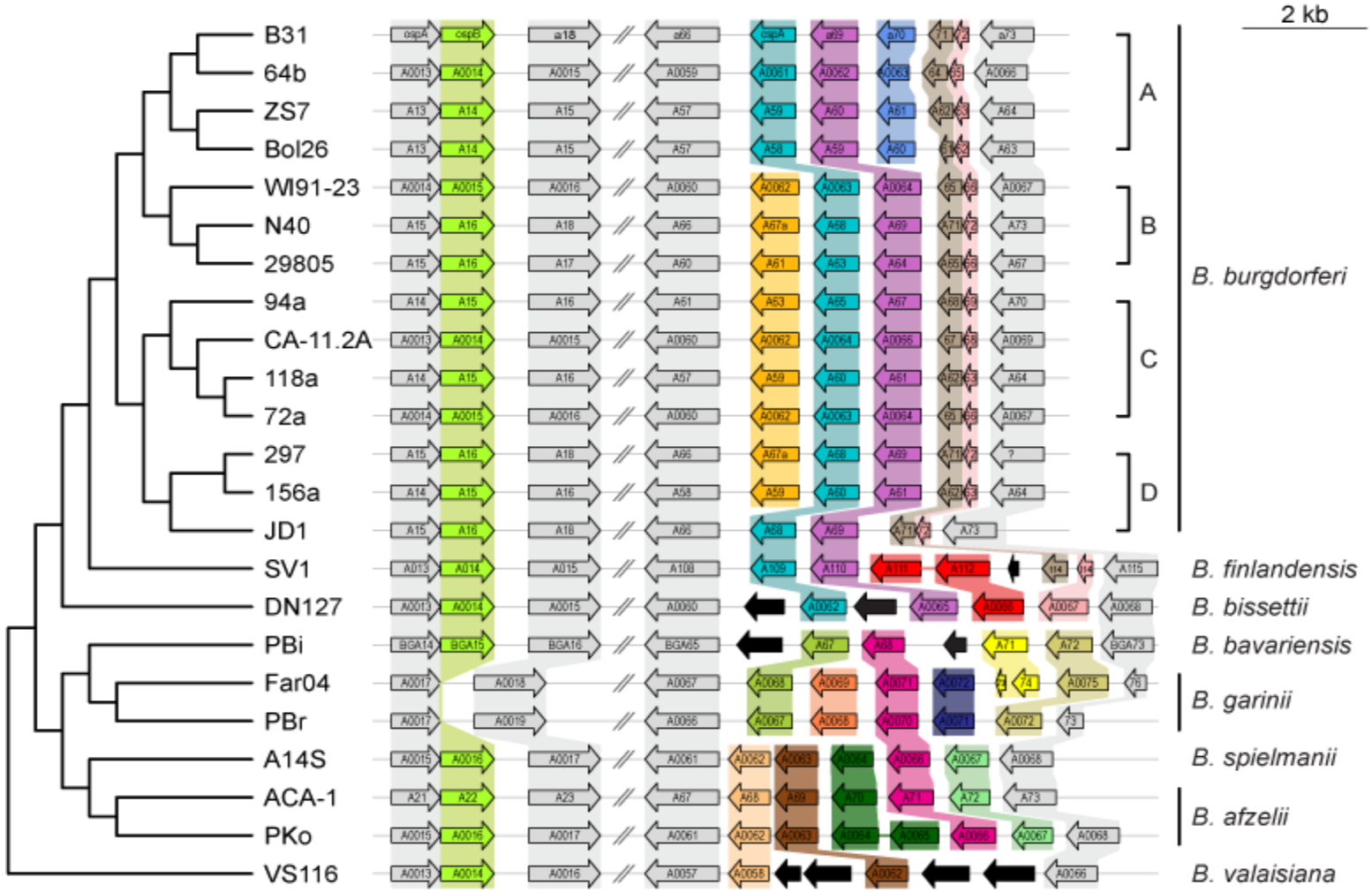


Figure 1-4: Parsimonious reconstruction revealed significant gene gains and losses on lp54. From right t o left: SPN based tree showing analyzed strains, ORFs included in the current study (arrows), Clade type (A,B,C,D), Genospecies, and the scale (2kb). Orthologous ORFs are represented by the same color. Orthologous ORFs present in all strains are colored grey. ORFs with no known orthologs are colored black. A "?" represent regions of genome where sequencing information is unavailable.

ORFs WITH ELEVATED *KA/KS* RATIOS WITHIN SPECIES

In pathogens, genes with high within-species variation are usually suggested to be involved in adaptive immunity evasion. To identify genes putatively involved in adaptive host immunity, we conducted within-species evolutionary analyses. Here, two lipoprotein genes associated with virulence, ospC and dbpA, show elevated levels of synonymous and nonsynonymous polymorphisms within *B. burgdorferi sensu stricto* (Figure 1-5). Interestingly, we observe an additional 3 genes which also display elevated levels of within-species polymorphisms, *b08* (a putative lipoprotein gene), *a07* (a putative *chpAI* gene) and *bbo258* (undecaprenol kinase). On cp26, two additional genes: *b18* and *b22* appear to have elevated levels of K_S. This observation however is actually an artifact due to their proximity to *ospC*.

ORFs UNDERGOING SPECIES-SPECIFIC ADAPTION

Between-species analysis of synonymous and nonsynonymous nucleotide substitution rates revealed a number of ORFs influenced by positive natural selection, which may be of functional importance. On the main chromosome, five ORFs; *bb0199* (putative membrane protein), *bb0441* (ribonuclease P protein), *bb0553* (unknown), *bb0576* (unknown) and *bb0831* (xylose regulator) contained codon sites with a K_A/K_S ratio significantly greater than 1 (Table 2). Lp54, contained nine ORFs; *dbpA*, a05 (encoding s1 antigen), *a15* (*ospA*), *a44* (unknown), *a50* (unknown), *a52* (putative outer member protein), *a57* (*P45-13*), a65 (encoding a putative lipoprotein), and *a73* (encoding putative P35 antigen) which possessed codon sites under positive selection. Notably, on cp26 only *ospC* was observed to be under significant positive selection.

ELEVATED ORFS ON CP26, LP54 AND MAIN CHROMOSOME

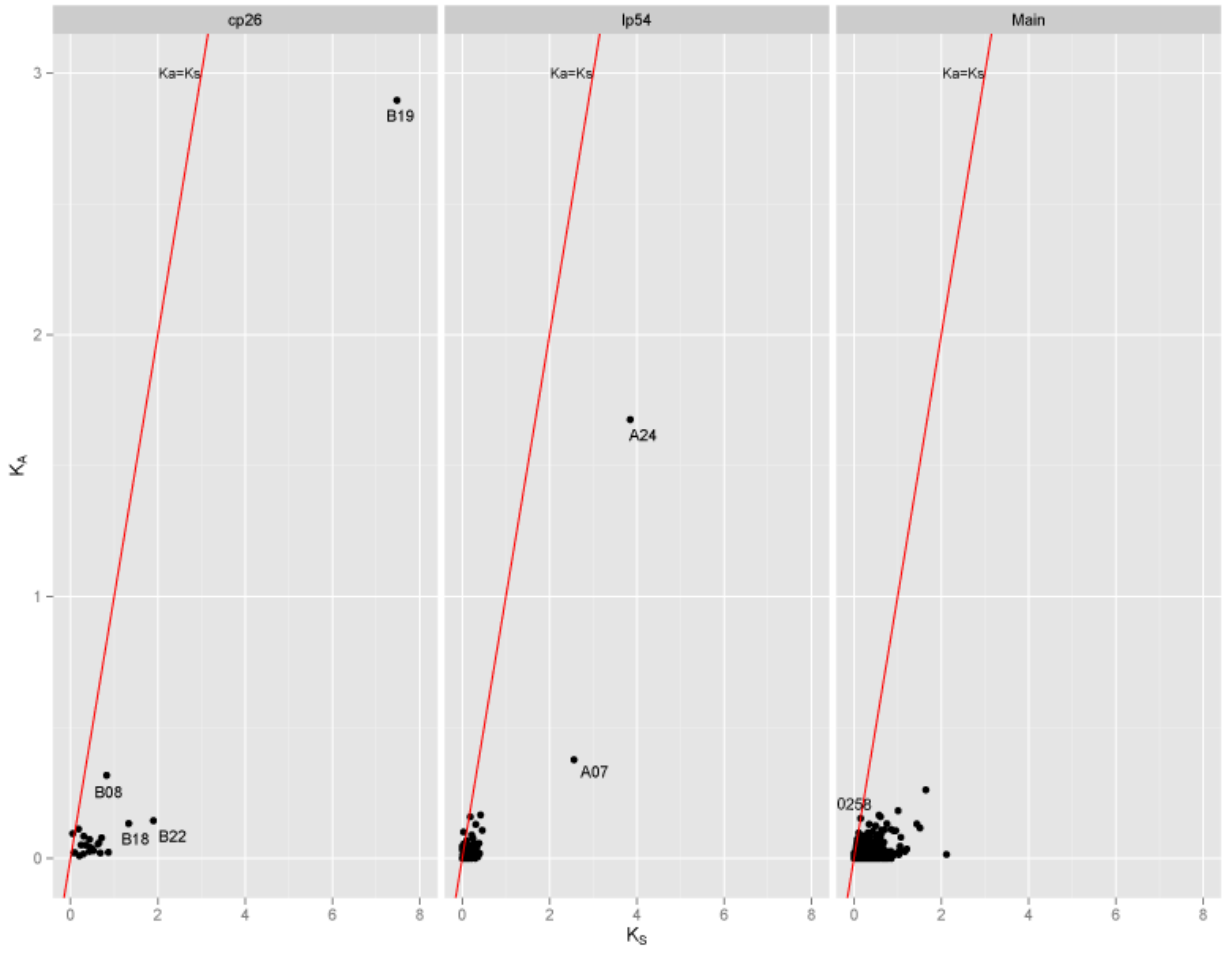


Figure 1-5: ORFs with elevated K_A/K_S ratio on cp26, lp54 and main chromosome. K_A rates are along the Y– axis while K_S rates are along the X-axis. Only ORFs with elevated K_A/K_S ratio are labeled. From each respective ORF: cp26 (*b08, b18, b19* and *b22*), lp54 (*a07* and *a24*) and Main Chromosome (*B0258*).

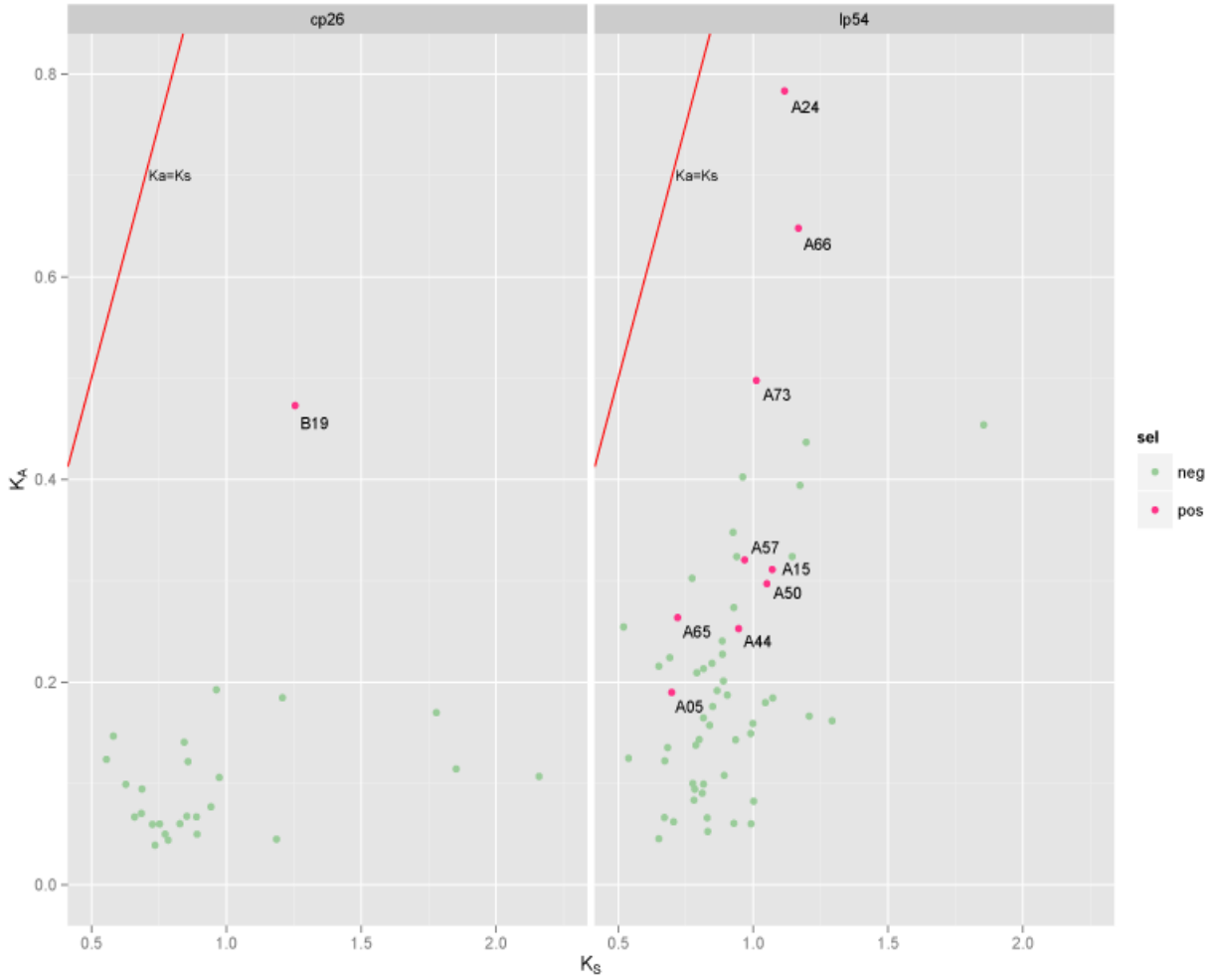


Figure 1-6: ORFs containing codon sites under positive selection. Non-synonymous rates (K_A) are depicted on the Y-axis. Synonymous rates (K_S) are shown on the X-axis. Positively selected ORFs are labeled and colored red. ORFs under purifying selection are colored green.

POSITIVELY SELECTED ORFs: All replicons

ORF ID	K_A	K_S	K_A/K_S	Delta	Annotation	Biological Function	Replicon
bbb19	0.721	1.9131	0.37688	15.1672	Outer surface protein C (OspC)	Antigen surface protein	cp26
bba05	0.2895	1.0638	0.27214	11.4922	S1 antigen	Immune evasion	lp54
bba15	0.4744	1.6319	0.2907	7.40942	Outer surface protein A (OspA)	Antigen surface protein	lp54
bba24	0.9801	1.3964	0.70188	11.1367	Decorin-binding protein A	Antigen surface protein	lp54
bba44	0.3622	1.3552	0.26727	9.81116	Conserved hypothetical protein	Unknown	lp54
bba52	0.6135	1.4651	0.41874	12.7388	Outer membrane protein	Membrane	lp54
bba57	0.4705	1.4204	0.33124	14.1548	P45-13	Surface lipoprotein	lp54
bba65	0.387	1.0562	0.36641	14.2217	Lipoprotein, putative	Unknown	lp54
bba66	0.9878	1.7798	0.55501	7.97363	Outer surface protein	Surface protein	lp54
bba73	0.7203	1.4637	0.49211	19.5278	Putative antigen P35	Antigen Immune Evasion	lp54
bb0199	0.68387	3.48352	0.19632	7.18018	Membrane protein, putative	Membrane	main
bb0441	0.54319	2.54677	0.21329	11.9761	Ribonuclease P protein component	Transcription	main
bb0553	0.18627	0.9454	0.19703	10.6221	Conserved hypothetical protein	Unknown	main
bb0576	0.22042	0.6759	0.32611	8.4397	conserved hypothetical protein	Unknown	main
bb0831	0.26776	3.32064	0.08064	7.72559	Xylose operon regulatory protein	Regulation	main

Table 2: ORFs under positive selection which contain codon sites with a K_A/K_S ratio significantly (P<= .001) greater than 1. The Delta value is the differences in log-likelihoods from positive verses nearly neutral models, and was used to conduct a likelihood ratio test at 2 degrees of freedom.

DISCUSSION

LINEAGE-SPECIFIC GENE GAINS AND LOSSES ON lp45

In this analysis, a well-supported phylogenetic tree (Mongodin et al. 2013) was used to reconstruct the history of gains and losses at five gene loci on the lp54 plasmid (Figure 1-4). In contrast to the conserved gene composition observed in cp26, the lp54 plasmid shows extensive evidence for gene gains and losses across *B. burgdorferi s.l.* strains. Specifically, we observed the loss of *ospB* in *B. garinii*; an avian-associated species (Vollmer et al. 2013), and high variability in gene composition within the PF54 region of lp54. The Clade-A specific *a70* orthologs in the PF54 array of *B. burgdorferi s.s.* is most likely a gene gain; considering that it is unlikely for its absence in other strains to be the result of multiple independent gene losses. This observation is particularly interesting given that all Clade-A isolates are RST1; a ribotype associated with disseminative Lyme borreliosis (Dykhuizen et al. 2008, Strle et al. 2011). Previous, studies to identify genes involved in disseminative Lyme borreliosis employed proteomic approaches, and the use of micro-arrays (Ojaimi et al. 2005, Nowalk, Gilmore, and Carroll 2006). Our current study; which adopts a comparative genomic approach, highlights *a70* as a possible candidate gene contributing towards the disseminative Lyme borreliosis phenotype in *B. burgdorferi s.s.* Other notable variations within the PF54 array include; the independent loss of *a67.5* in *B. burgdorferi s.s.* Clade A and JD1 strains, a truncated form of *B. bissettii a72* orthologs (verified via alignments) in *B. burgdorferi s.s.* and *B. finlandesis* strains, and limited ORF synteny among non-*sensu stricto* strains. Though many studies have implicated PFam54 genes in host resistance (Kenedy et al. 2009, Patton et al. 2013), it is not clear how the variations mention above are associated with host-species preference or human virulence.

Another interesting observation is the loss of *ospB*, a lipoprotein gene, in *B. garinii* (an avian specific strain). Here, Sequence alignments revealed that the loss of *ospB* was due to a deletion in the region encompassing *ospB,* and its upstream 9-base intergenic sequences, which left *a18* (a PF62 partition gene) and its upstream sequences intact. Despite the fact that *ospA* and *ospB* share a single promoter in *B. burgdorferi*, *ospA* is predominantly expressed in tick-adapted spirochetes while the expression of *ospB* is known to be higher in a mammalian environment (Liang, Yan, et al. 2004). Studies show that *ospB* is selectively expressed throughout murine infection and is targeted by the host complement-independent defense system (Liang, Caimano, et al. 2004, LaRocca et al. 2009). The observed absences of *ospB* from the genomes of *B. garinii*, is consistent with its apparent mammalian-specific functionality. Notably, studies involving *ospB* knockouts revealed that it is also necessary for *B. burgdorferi* adherence and persistence within *Ixodes* ticks, a function divergent from its role during vertebrate infection which is associated with immune escape (Neelakanta et al. 2007). Collectively, our findings along with the results of experimental studies mentioned above, suggest a probable role for *ospB* as an important virulent factor specific to, and required for *B. burgdorferi* maintenance in its mammalian-tick enzootic cycle.

ORFs UNDER WITHIN-SPECIES BALANCING SELECTION

For our within-species analysis, in order to identify genes with a putative role in adaptive immunity evasion, we employed the software PAML to estimate synonymous (K_A) and non-synonymous (K_S) rates for orthologous ORF sets of *B. burgdorferi* core genome. Here, two lipoprotein genes associated with virulence, *ospC* and *dbpA*, were observed to have elevated levels of synonymous and nonsynonymous polymorphisms (high sequence diversity) within *B.*

burgdorferi s.s. Both these genes are highly expressed during the early stages of host invasion and have been shown to be under strong positive natural selection; driven by negative-frequency-dependent immune escape (Haven et al. 2011, Wang et al. 1999, Qiu et al. 2004). Additionally, independent studies have also implicated within-population host differentiation as a contributing factor to the high sequence diversity observed at these two loci within *B. burgdorferi s.s* (Brisson and Dykhuizen 2004).

Three other genes: *b08*, *a07*, and *bb0258* also displayed elevated levels of within-species polymorphisms, and by the same rational may be involved host differentiation, or subjected to immune driven within-species diversifying selection (Figure 1-5). It is possible however as an alternative explanation, that the higher sequence variation observed at these three loci could be attributed to slightly deleterious mutations in these ORFs segregating within *B. burgdorferi s.s.*

OspC and DbpA have been extensively studied as key virulent factors in *B. burgdorferi.* While the function of *b08* remain unknown, *bb0258* has been annotated to as a undecaprenol kinase (a stress response gene in other bacteria), and *a07*(a putative chpAI gene) has been shown to be required for infection of mammals via tick transmission (Xu et al. 2010). The findings of elevated synonymous and nonsynonymous polymorphisms in *b08*, *a07* and *bb0258,* coupled with the experimentally identified functional importance of *a07* and *bb0258*, suggest that these three ORFs may prove to be interesting targets for further studies.

ADAPTIVELY EVOLVING ORFS BETWEEN *Borrelia burgdorferi sensu lato*

To identify putative genes contributing to functional divergence within *Borrelia* core genome, we conducted a scan for the evidence of positive selection on between-species

orthologous ORF sets. Our between-species analyses revealed fifteen ORFs: 5 from the main chromosome, 1 from cp26, and 9 from lp54, containing codon sites with a K_A/K_S ratio significantly ($P \leq 0.001$) greater than 1 (Table 2). Sequence alignments revealed that most codon sites with a high K_A/K_S ratio show three or more amino-acid states between species. Since these ORFs show species-specific nonsynonymous variations, they are more, likely to be associated with adaptive divergence between species in *B. burgdorferi sensu lato.* Interestingly, while cp26 was observed to have higher levels of synonymous polymorphism and divergence due to strong balancing selection, ORFs on cp26 showed lower levels of nonsynonymous substitutions than those on lp54, suggesting stronger purifying selection on the cp26 genes in general.

Notably, genes identified as being influenced by positive selection spanned a variety of functional roles including: gene regulation, transcription, immune evasion, membrane proteins, and lipoproteins (Table 2). Studies have attributed functional importance to two of the five genes from the main chromosome identified as being influenced by positive selection. Specifically, *bb0441*; a putative transcription factor, was shown to display elevated levels of expression with an increase in temperature (Ojaimi et al. 2003) and *bb0831*, a putative regulatory protein, was predicted to be a xylose operon regulator (Schutzer et al. 2011a). The observation of positive selection in these two non-membrane genes of the main chromosome is plausible, since it has been reported that other pathogens such *Escherichia coli* and *Streptococcus* also poses genes under positive selection in similar functional categories. Such categories include: nutrient acquisition, gene regulation and transcription (Chen et al. 2006, Lefébure and Stanhope 2007). This suggests that such-like genes in bacterial pathogens are under similar selective pressures. The function of the three remaining main chromosome genes (hypothetical proteins)

remains unknown. However, given that they were considered to be under positive selection at such high significance, they should be considered interesting targets for future research.

Most of the ORFs reported to be under the influence of positive selection were located on lp54 (9 of 15 ORFs) (Table 2) and were described as either lipoproteins or antigenic proteins. The cp26 plasmid contained only one ORF, *ospC*, also defined as a lipoprotein, to be significantly under positive selection (Table 2)**.** Experimental data show that lipoproteins *ospA*, *ospC*, *dbpA*, *a52*, *a57*, *a66* and *a73* all play crucial roles in host-pathogen interactions (Patton et al. 2013, Yang et al. 2013, Kumar et al. 2011, Samuels 2011, Wywial et al. 2009). The role of the remaining three lipoproteins: *a15*, *a44* and *a50* have yet to be defined.

It should be noted that in our evolutionary analyses, one possible source of error is the overestimation of positive selection due to poorly aligned ORFs. This is particularly the case in our main chromosome dataset where the start codon synchronization was not completed. To ensure the accuracy of our results, the alignment of all ORFs identified as significant in our analyses were manually inspected. If any significant ORFs showed poor alignments (excessive number of gaps in the N-terminal or other regions), they were omitted from our reported results. Additionally, the fact that PAML ignores gap sites, and that most inconsistencies occurred at the highly conserved N-terminal of studied ORFs, the unsynchronized main chromosome data had negligible impact on our ability to detect significant elevated K_A/K_S level or positive selection.

CONCLUSION

In our current study, we employed comparative genomics to identify functionally important genes involved in host evasion within the core genome of *B. burgdorferi*. Through parsimony analysis we have identified a number of gene gains and losses on lp54. Here, two particularly interesting ORFs *a70*, a Clade A-specific gene gain and *a16*, a *B. garinii* specific gene loss, may be putatively implicated in disseminative Lyme borreliosis, and the mammalian-tick enzootic cycle respectively. Our evolutionary analyses revealed five genes putatively under immune-driven balancing selection and fifteen genes influenced by positive natural selection, putatively contributing to adaptive divergence. We conclude that the evolutionary significance of our presented results, provide new insights into genes with putative functional roles in *B. burgdorferi* persistence and virulence in hosts including humans. These findings may contribute to the experimental design of site-directed mutagenesis studies, and provide crucial information towards the identification of potential targets for vaccine development.

CHAPTER II –THE IDENTIFICATION OF *CIS* AND *TRANS*-ACTING REGULATORY ELEMENTS IN THE PLASMID COMPONENT OF *Borrelia burgdorferi sensu lato* CORE GENOME

INTRODUCTION

GENE REGULATION AND THE RPON-RPOS PATHWAY

The ability to regulate gene expression throughout the course of an infection is important for the survival of pathogens in their respective hosts. As an obligate parasite, *B. burgdorferi* must survive in two biologically distinct environments for its maintenance in nature, and hence employ tight regulation of gene functions during transition between ticks and its vertebrate host (Samuels 2011). Specifically, approximately 150 genes (~9%) of the *B. burgdorferi s.s.* B31 genome are differentially expressed during the transition between the tick and mammalian infection phases (Brooks et al. 2003). RpoS (σ^S), an alternative sigma factor, has been identified as a major regulator of *B. burgdorferi* virulence genes and has been shown to achieve this precise control via an elaborate RpoN (σ^{35})-RpoS gene regulatory pathway (Smith et al. 2007, Ouyang et al. 2012, Dunham-Ems et al. 2012). Under this pathway, RpoS synthesis is regulated at both the transcriptional and post-transcriptional levels. It should be noted that RpoS mRNA exists in two isoforms: (i) a RopN-dependent (temperature and pH sensitive) short transcript and (ii) an RpoD (σ^{70})-dependent (temperature sensitive) long transcript; regulated post-transcriptionally by ncRNAs. Studies show that as temperature increases and pH decreases (conditions which mimic tick to host transmission), virulent genes such as *ospC* and *dbpAB* are up-regulated, while tick-phase genes (e.g. *ospA*) are down regulated in an RpoS dependant manner (Strle et al. 2011).

Despite the fact that RpoS has been observed to alter the expression of many *B .burgdorferi* genes, to date only five genes; *ospC*, *dbpA*, *oppA5, a66* and *bba07*, have been identified to be directly transcribed by RpoS (Caimano et al. 2007). Notably, all four genes

mentioned above, were also observed to contain known RpoS-dependant promoter sequences TTGA(a/t)(t/a)$N_{10\text{-}11}$TG(g/a)(g/a)ATA(t/a)ATT in their upstream regions (Caimano et al. 2007). Our knowledge of the RopS-RpoN pathway's influence on pathogenicity in *Borrelia* is still in its infancy. Recent studies have shown that RopS negative mutants modified to constitutively express OspC, DbpA, and DbpB, were unable to infect murine host; suggesting that other downstream targets of the RpoN-RpoS regulatory pathway remain to be identified (Xu et al. 2012).

In addition to the RpoS-RpoN pathway, a numbers of regulatory element have been identified in the intergenic regions of the *B .burgdorferi* genome. These elements vary from cis-regulatory sequences to ncRNAs. One such example can be found in the non coding region upstream of *ospC*. Here, studies reveal a pair of inverted repeats which function as an operator required for the repression of *ospC* during vertebrate host infection (Xu, McShan, and Liang 2007). Note, although OspC is required initially to establish infection, it is a potent immunogen and must be repressed early during *B .burgdorferi* infection (Crother et al. 2004, Liang et al. 2002, Liang, Nelson, and Fikrig 2002). Another interesting regulator identified in *B. burgdorferi* is $DsrA_{Bb}$. $DsrA_{Bb}$ is an ncRNA which binds to, and regulates RpoS translation in a temperature-dependant manner (Samuels 2011). Clearly the identification of additional putative regulators will further our understanding in the modulation of gene expression in *B. burgdorferi*.

IDENTIFICATION OF FUNCTIONALLY IMPORTANT *cis*- and *trans*-acting ELEMENTS

Gene regulation is often associated with *cis*-acting sequences and *trans*-acting proteins that cooperatively affect the function of RNA polymerase (RNAP). A number of studies have identified functional *cis*- and *trans*-acting elements which were critical to the regulation of virulent genes in *B. burgdorferi* and other pathogens (Lin et al. 2009, Xu, McShan, and Liang 2010, Ouyang, Deka, and Norgard 2011, Dale et al. 2012).

Experimentally, the creation of transcription factor binding maps, and hence the identification of functionally important non-coding regions employs the use of chromatin immunoprecpitation (ChIP) (Farnham 2009). Here, DNA-binding proteins are cross-linked to the target DNA in vivo via formaldehyde. The DNA-protein complex is then immunoprecipitated using antibodies to binding proteins of interests. This retrieved DNA-protein complex is then further analyzed thru PCR, microarray (ChIP-chip) or sequencing technology capable of analyzing the bound DNA sequence (ChIP-seq) (Farnham 2009). In studies such as these, preliminary evolutionary analyses can prove to be a valuable tool towards the identification of putatively functional *cis,* and *trans*-acting elements for further experimental studies. Under this evolutionary framework, the approach of "phylogenetic shadowing" compares orthologous non-coding sequences among closely related species while "phylogenetic footprinting" refers to the comparison of more distantly related genomes. For instance, in *Drosophila melanogaster*, evolutionary analysis revealed non-coding but functionally important regulatory genomic sequences through inter-species comparisons (Andolfatto 2005). Similar analyses on the genomes of *Buchnera aphidicola*, revealed the presence of conserved, functional, intergenic spacers (IGS) (Degnan, Ochman, and Moran 2011). Previously, such studies on *B. burgdorferi* were restricted in their ability to detect significantly-conserved, putative *cis*- and *trans*-acting

elements due to the limited availability of whole genome sequences (Eddy 2005). However, with the newly available *Borrelia* genome data, the sequence scope for the identification of potential functional sites in this pathogen has greatly increased. In the current study, we employed comparative analyses towards the identification of putatively functional *cis,* and *trans*-acting elements within IGS regions of *B. burgdorferi* lp54 and cp26 plasmids. Specifically, we scanned IGS sequences for signals of purifying and positive natural selection based on intra- and inter-specific comparisons. Identification of conserved, putative cis-regulatory sequences, or putative trans-acting factors, will improve our understanding of the genomic basis for gene regulation and virulence in *B. burgdorferi s.l.*

MATERIALS AND METHODS

GENERATION OF IGS DATASET

To identify orthologous IGS sequences, start and end positions of orthologous ORFs were used as a reference. This approach is based on the assumption that IGS sequences are orthologous if they are flanked by orthologous ORFs. Previous synchronization of start-codon positions within orthologous ORF sets minimized the erroneous mixing of true IGS, and sequences which may be a part of neighboring ORFs. Orthologous IGS sequences and respective 25-codon sequences from its two flanking codon regions were extracted using a customized Perl script based on BioPerl (Stajich et al. 2002). Extracted IGS sequences were aligned with MUSCLE (Edgar 2004), while their flanking ORF sequences were aligned according to the MUSCLE alignment of translated protein sequences. IGS loci were categorized into three types based on the relative transcription directions of its flanking ORFs: i) divergent, IGS was located

at the 5' end of both flanking ORFs; ii) tandem, IGS at the 5' of one of the two flanking ORFs; and iii) convergent, IGS at the 3' end of both flanking ORFs (Appendix II).

TESTS OF SEQUENCE CONSERVATION IN IGS

To identify evolutionarily conserved IGS, we conducted comparative intra- and inter-specific (within- and between species) analyses for evidence of evolutionary constraint. The dataset for our within-species analysis consists of fourteen *B. burgdorferi s.s.* genomes, whereas our between-species analysis included one genome from each of eight *B. burgdorferi s.l.* species (Figure 1-1).

For our intra-specific analysis, nucleotide substitution rate among orthologous IGS sequences (K_I), were compared to the synonymous substitution rate (K_S) of their respective flanking orthologous ORFs segments (25-codons long). Maximum-likelihood estimated of these nucleotide substitution rates were obtained using the $PAML_{v4.4}$ software package (Yang 2007). Specifically, the K_I was obtained using the BASEML program with a neighbor-joining tree of the IGS sequences as the input tree and the following key parameters: cleandata=1 (removing gaps), runmode=0 (user-defined tree), model=4 (HKY substitution model), clock=0 (no clock), fix_kappa=0 (estimating κ, the transition to transversion ratio), fix_alpha=0 (estimating α, the rate heterogeneity parameter), and fix_length=1 (tree branch lengths as initials). The K_S was obtained using the CODEML program with the SNP-based tree (Mongodin et al. 2013. in review) as the input tree and the key parameters the same as in the BASEML runs (see above) except the following: model=0 (one rate for all branches) and NSSites=0 (a single d_N/d_S ratio for all branches).

For our inter-specific analysis, the SNP tree was used in both the BASEML and CODEML runs in order to obtain maximum-likelihood estimates of substitutions rates. Here, nucleotide substitutions rates (K_I) from orthologous IGS, and the synonymous substitution rates (K_S) from their flanking ORFs were used to compute their (K_I/K_S) ratio. This ratio is then used as a measure of evolutionary constraint for IGS sequences; with $K_I/K_S < 1$ indicating sequence conservation and purifying selection, and $K_I/K_S = 1$ as the neutral expectation. Conservation of an IGS locus was also measured by the proportion of gapped sites in an IGS alignment. Here, conserved IGS which contained few gaps (<20%) and (K_I/K_S<0.5) were considered most significant. To further verify conservation, we conducted a two dimensional (polymorphism *vs.* divergence and neutral (π_S) *vs.* π_a differences) MK analysis (Eyre-Walker 2006). Here if an intergenic region is observed to have a higher proportion of within species polymorphism versus between species divergence it is considered to be conserved (Appendix II).

COMPUTATIONAL PREDICTION OF REGULATORY IGS SEQUENCES

PROMOTERS AND ncRNAs

To identify putatively functional elements within non-coding regions of the *Borrelia* genome, we tested for the presence of promoter and, non-coding RNA (ncRNA), in significantly conserved IGS. PromPredict$_{v1.0}$, which detects differences in free energy between promoter and non-promoter regions in bacterial genomes (Rangannan and Bansal 2007), was used to predict promoter sequences our within orthologous IGS sequences Promoters were only reported if they were identified in all sequences. Here we considered known promoter regions (e.g. *ospC* and *dbpA*) as our positive control and convergent IGS segments within consecutive ORFs as negative controls. The identification of putatively functional non-coding RNAs within individual conserved ISG alignments, was accomplished using the software RNAz$_{v2.1}$ (Gruber et al. 2010).

Here, we only reported the presence of an element if it was present in all strains. Our results were further filtered for putative ncRNA alignments which were < = 10% gapped. A consensus secondary structure of predicted ncRNA elements was visualized using RNAalifold (Bernhart et al. 2008).

IDENTIFICATION OF POTENTIAL RPOS-BINDING SITES

A Customized search algorithm which incorporated the previously published B. *burgdorferi B31* RpoS-binding consensus sequence (Caimano et al. 2007) was used to initially identify RpoS-binding sequences in IGS alignments (AppendixV). Identified RpoS sequences were then used as a query in a BLASTN search (megablast, - no dust options) on their respective IGS alignments in order to identify their between species orthologs. To obtain a new inter-specific consensus sequence, our paralogous predicted RpoS-binding sequences (one representative strain per species from each IGS locus) were aligned with MUSCLE (Edgar 2004) and used as input into the web-based application WebLogo (Crooks et al. 2004) for the generation of an RpoS-binding site sequence logo.

FUNCTIONAL ELEMENTS WITHIN PERFECTLY CONSERVED INTERGENIC BLOCKS (PCIBS)

Aware that conserved IGS sequences are likely to contain functional elements other than those listed above, we identified and extracted perfectly conserved sequence blocks from our IGS alignments. We extracted all un-gapped segments longer than 6 nucleotides that were perfectly conserved among all sequenced genomes using a customized PERL script. The statistical significance of observed PCIBs was estimated by randomly shuffling each IGS alignments ten times and comparing the length distributions of PCIBs between observed and shuffled alignments. The resultant perfectly conserved IGS blocks (PCIBs) represent the most

significant conserved segments of IGS sequences, and potentially contain a number of uncharacterized functional elements. To further explore this possibility, extracted PCIBs and entire IGS segments were analyzed with TransTermHP$_{v1.0}$; which predicts Rho-independent transcription terminators via scanning for hairpin loops followed by a thymine-rich segment (Kingsford, Ayanbule, and Salzberg 2007). The results of our TransTermHP analysis were the confirmed using another prediction software, ARNold (Naville et al. 2011). Note, both Arnold and TransTermHP attain high specificities at 95.3% and 97.5%, respectively, this equals to an average of 47 and 25 false positives hits per 1,000 intergenic regions of size 115 nt. If transcription terminators were identified by both algorithms, a BLASTN search was done (using the megablast options mentioned above) in order to confirm the presence of orthologous sequences within respective IGS or PCIBs. As an additional, analysis for the identification of putatively functional elements, BLASTN was used in an all against all blast to determine PCIBs which were similar to each. For this analysis, the following parameters were used: -F 0 (no low-complexity sequence filtering), -e 1e-5 (an expect value cutoff 10-5), -w 5 (word size 6) (Camacho et al. 2009).

RESULTS

OVERLAPPING ORFS

Inter-specific comparative analysis revealed a number of overlapping reading frames conserved across species on these two *B. burgdorferi s.l.* plasmids. Of the 26 possible IGS loci on cp26, three were absent due to overlapping ORFs. These IGS included: *b01-b02*, *b11-b12* and *b24-b25*. On the lp54 plasmid, the following overlapping ORFs were identified: *a08-a09*, *a12-*

a13-a14, *a19-a20*, *a42-a43-a44*, *a46-a47*, *a48-a49*, and *a55-a56*. Interestingly, all overlapping ORFs on lp54 are oriented in the same 5` to 3` direction with respect to transcription.

IGS MORE CONSERVED THAN FLANKING SYNONYMOUS SITES

Both our intra-and inter-species analyses revealed on average IGSs nucleotide substitution rates (K_I) were less than the rates of synonymous substitutions in their flanking ORFs (K_S) (Figure 2-1). These results suggested that for the most part our IGS sequences were under purifying selection with a $K_I/K_S<1$. A between- vs. within-species K_I/K_S plot further bolstered these findings revealing most IGS segments had a K_I/K_S ratio of < 0.5 for both inter- and intra-specific analyses (Figure 2-2). Notably, our between- vs. within-species plot revealed one IGS; *a33-a34*, with an elevated K_I/K_S ratio. This observation however is considered to be an artifact caused by the high percent of gapped sites (72.38%) within *a33-a34*. Our intra-specific analysis (within-species) analysis was less informative showing relatively high levels of conservation in all IGS, hence the results of our inter-specific analysis were used to determine significance of conservation in IGS segments. The filtered results of our inter-specific analysis revealed five IGSs from cp26, and eight from lp54 which were significantly conserved (K_I/K_S <= 0.5 and < 20% gap sites) (Table 3). To further identify regulatory sequences within IGSs segments, 82 and 300 perfectly conserved intergenic sequence blocks (PCIBs) with a minimal length of 6 nucleotides were identified on cp26 and lp54, respectively. Simulated PCIBs from ten sets of shuffled alignments revealed that the majority of our identified 6-, 7-, and 9-mer PCIBs are likely to be random and hence are false positives. The false-positive rates of observed PCIBs in 8-mer and 10- through 24-mers are approximately 50% (Figure 2-3). In total, we identified 286 significant (length >10) PCIBs on cp26 and lp54.

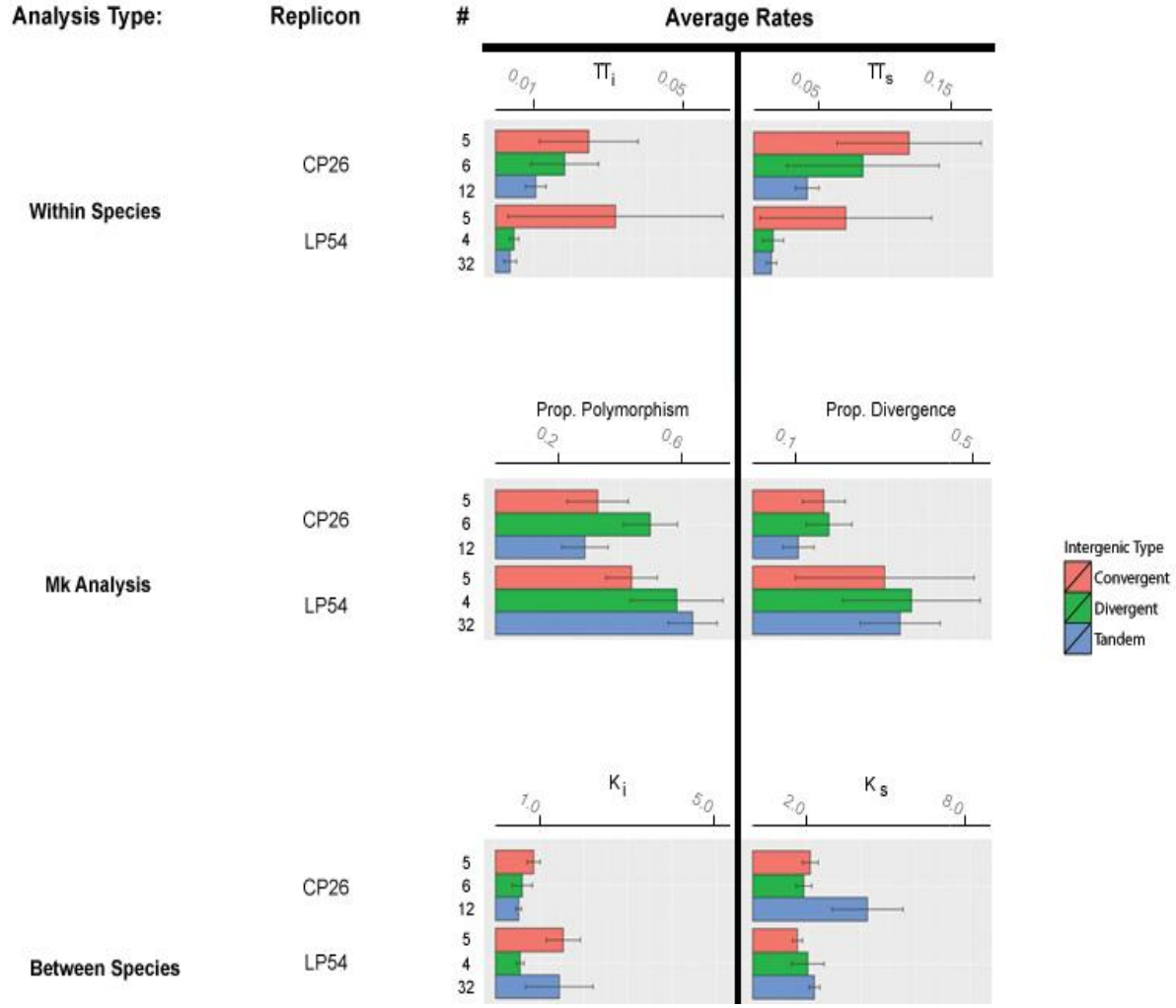


Figure 2-1: Summary plots illustrating overall purifying selection on both cp26 and lp54. Analysis type, replicon source and number of each intergenic segment (IGS) type analyzed are illustrated on the left. Convergent, divergent and tandem IGS are colored red green and blue respectively. Within species plot compares average pairwise difference π_i of each IGS type to their flanking synonymous rates π_s. The MK Analysis compares the average within species polymorphism to average between species divergence. Our between species analysis compared the average IGS non-synonymous rates to average the synonymous rates of its flanking ORFs.

MOST IGS SEGMENTS ARE CONSERVED WITH K_I/K_S RATIO < 0.5

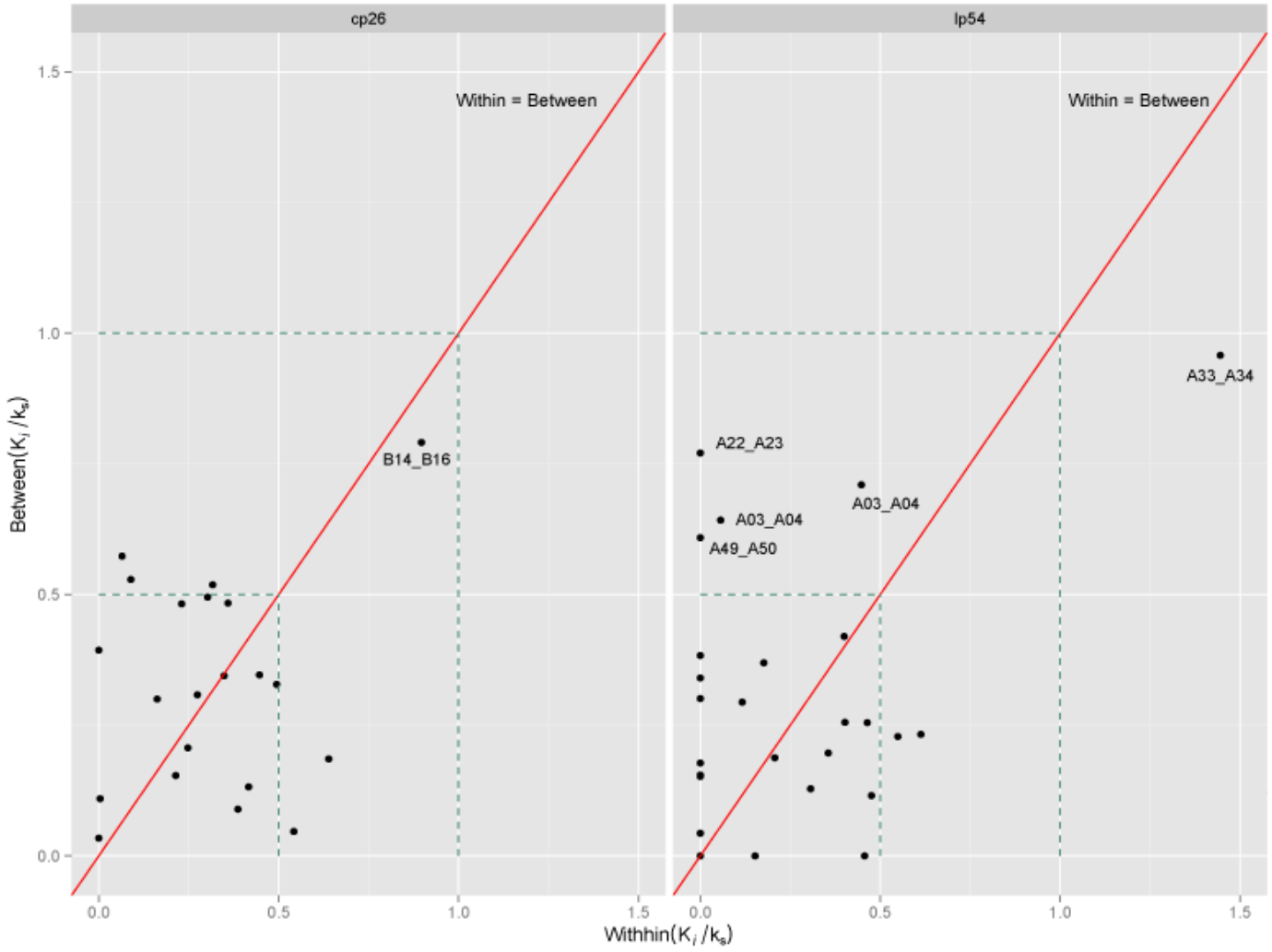


Figure 2-2: Within vs. Between K_I/K_S plot for cp26 and lp54. Within ratios are plotted on the X-axis, while between ratios are plotted on the Y-axis. Dashed lines (green) illustrate areas of the plot where K_I/K_S rates for within and between analysis is within the range of 0.5 and 1.0 respectively. IGS segments with rates > 0.5 are labeled with their respective names. Note the IGS *a33_a34* has a rate > 1.0, but is not considered to be a significant result due to the high % of gap observed in the alignment.

PUTATIVE FUNCTIONAL ELEMENTS IDENTIFIED WITHIN SIGNIFICANTLY CONSERVED IGS SEGMENTS

IGS	Alignment length	gap %	*Ka/Ks*	Identified Putative Functional Elements			Within PCIBs
				ncRNA	terminator	promoters	
BBB08-BBB09	166	12.0482	0.1298			◆	promoter -35
BBB12-BBB13	75	5.3333	0.0659	◆			
BBB13-BBB14	286	15.035	0.2254	◆	◆		terminator
BBB28-BBB29	357	19.0476	0.153			◆	promoter -35
BBB29-BBB01	97	5.1546	0.5187	◆	◆		terminator ncRNA
BBA07-BBA08	183	4.92	0.3451			◆	
BBA16-BBA18	609	10.34	0.1458	◆	◆ **	◆	
BBA25-BBA30	957	19.75	0.1702	◆	◆ **	◆	
BBA32-BBA33	264	13.64	0.082			◆	
BBA37-BBA38	372	7.26	0.2079	◆	◆ **	◆	terminator
BBA57-BBA59	517	18.96	0.1566	◆	◆**	◆	
BBA61-BBA62	255	2.75	0.3559			◆	
BBA65-BBA66	184	4.89	0.1369	◆			

Table 3: Identified putative ncRNA, transcription terminators and promoters within specific IGS segments are indicated via a blue, red, and green diamonds respectively. ** indicate transcription terminators identified by both ARNold and TranstermHP. Elements identified within absolutely conserved blocks (PCIBs) were listed in the rightmost column of their respective IGS segments.

DETERMINATION OF SIGNIFICANT LENGTH FOR IDENTIFIED PCIBs

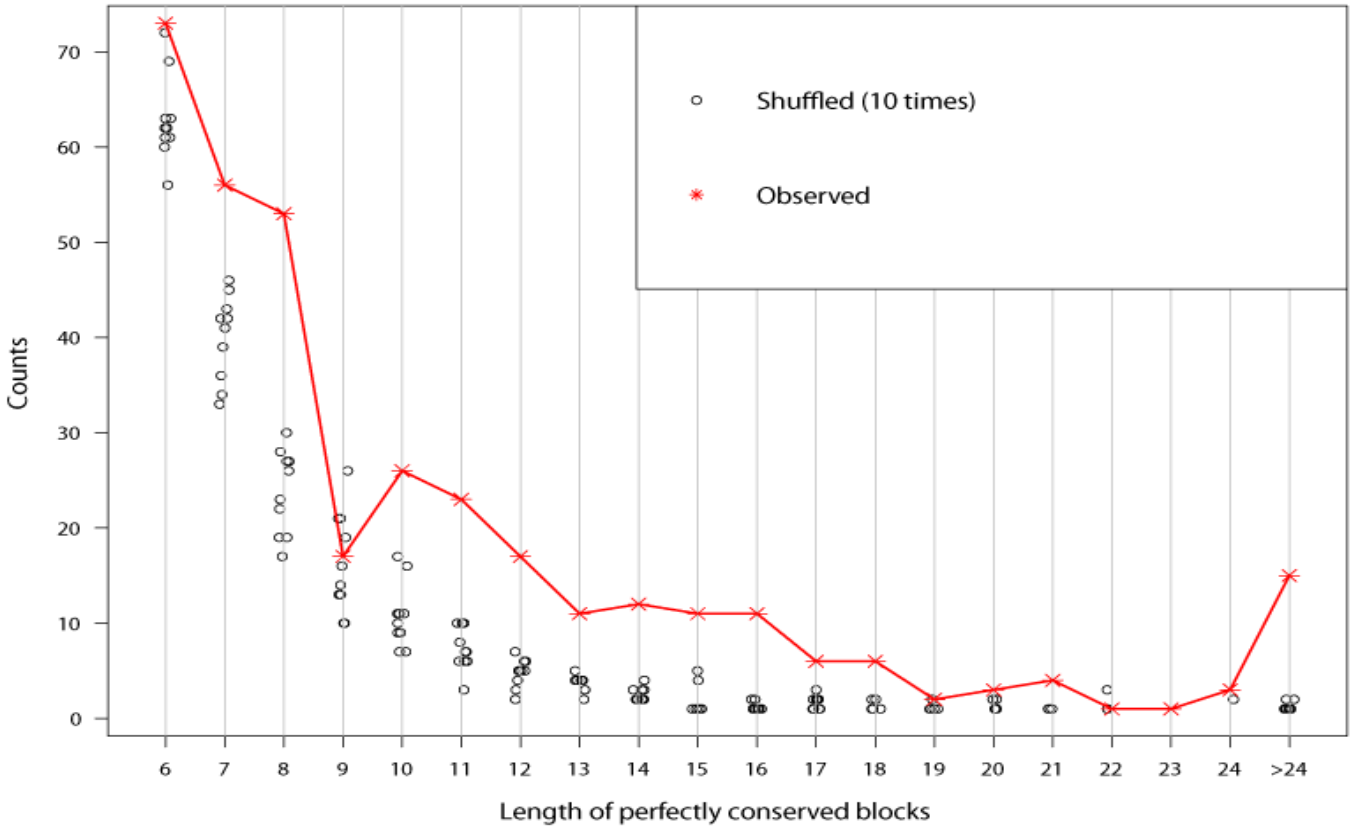


Figure 2-3: Alignments of PCIBs ranging in lengths from 6 to >24 nucleotides were randomly shuffled 10 times. The resultant frequencies for different lengths of shuffled PCIBS were plotted against the observed frequency for their respective original un-shuffled PCIB. Counts are plotted on the Y-axis while the lengths of PCIBs are plotted on the X-axis. Shuffled frequencies are represented by black circles, while frequencies of the original observed lengths are represented by red asterisk.

PREDICTED REGULATORY IGS ELEMENTS

PROMOTERS, ncRNAs AND TRANSCRIPTION TERMINATORS

We constructed computational PERL-based pipelines that employed the use of propriety software packages to perform free energy ($PromPredict_{v1.0}$, $RNAz_{v2.1}$) and structural-homology ($TransTermHP_{v1.0}$) analyses on our IGS segments. In cp26, we found five significantly conserved IGS: *BBB08-BBB09*, *BBB12-BBB13*, *BBB13-BB14,* *BBB28-BBB29*, and *BBB29-BBB01* containing putative elements including: two promoters, three ncRNAs, and two transcription terminator (Table 3). Lp54 contained eight significantly conserved IGS: *BBA07-BBA08*, *BBA16-BBA18*, *BBA25-BBA30*, *BBA32-BBA33,* *BBA37-BBA38*, *BBA57-BBA59,* *BBA61-BBA62* and *BBA65-BBA66* which collectively contained five ncRNAs, four transcription terminators, and seven promoters (Table 3). Overall, we identified twenty-one putative ncRNAs, sixteen putative transcription terminators, and sixteen putative promoters within our analyzed IGS segments (data not shown**)**. Of all the identified transcription terminators, six; 1 from cp 26 and 5 from lp54, were predicted by both TranstermHP and ARNold. Notably, four PCIBs; 2 from cp26 and 2 from lp54 contained predicted transcription terminators (Table 3). Our all against all BLAST identified eight groups of PCIBs which shared significantly matched sequences (Figure 2-4). Additionally, we identified six long (> 14 nucleotides) conserved palindromes which contained inverted repeats, and were further confirmed to be the site of predicted ncRNAs in their respective IGS segments. Interestingly, three of these long conserved palindromes: *A16-A18*, *B12-B13*, and *B29-B01*, were additional located within significantly conserved IGS segments (Table 3).

FIGURE 2-4: PREDICTED REGULATORY ELEMENTS AND AGAINST ALL BLAST RESULTS

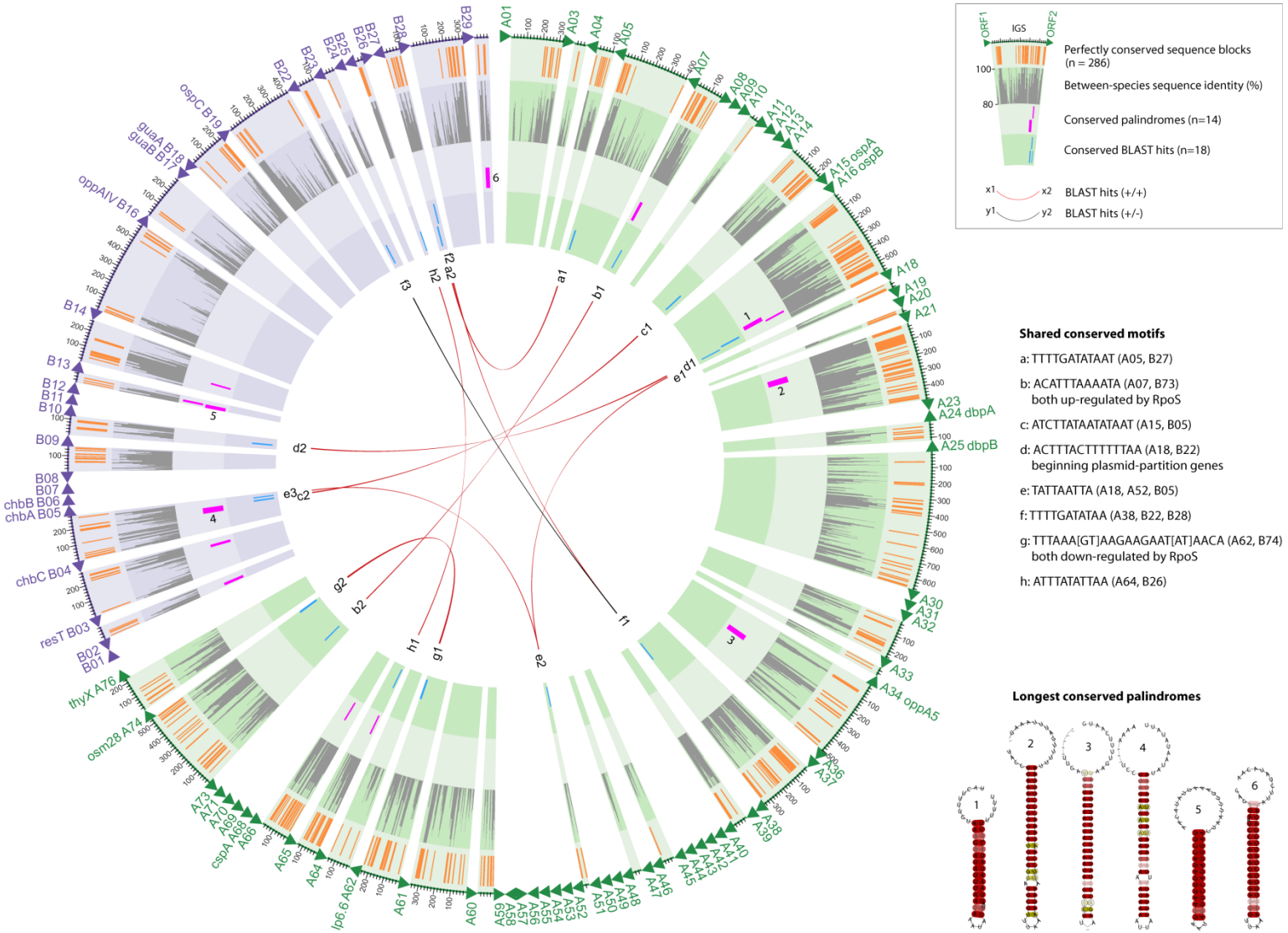


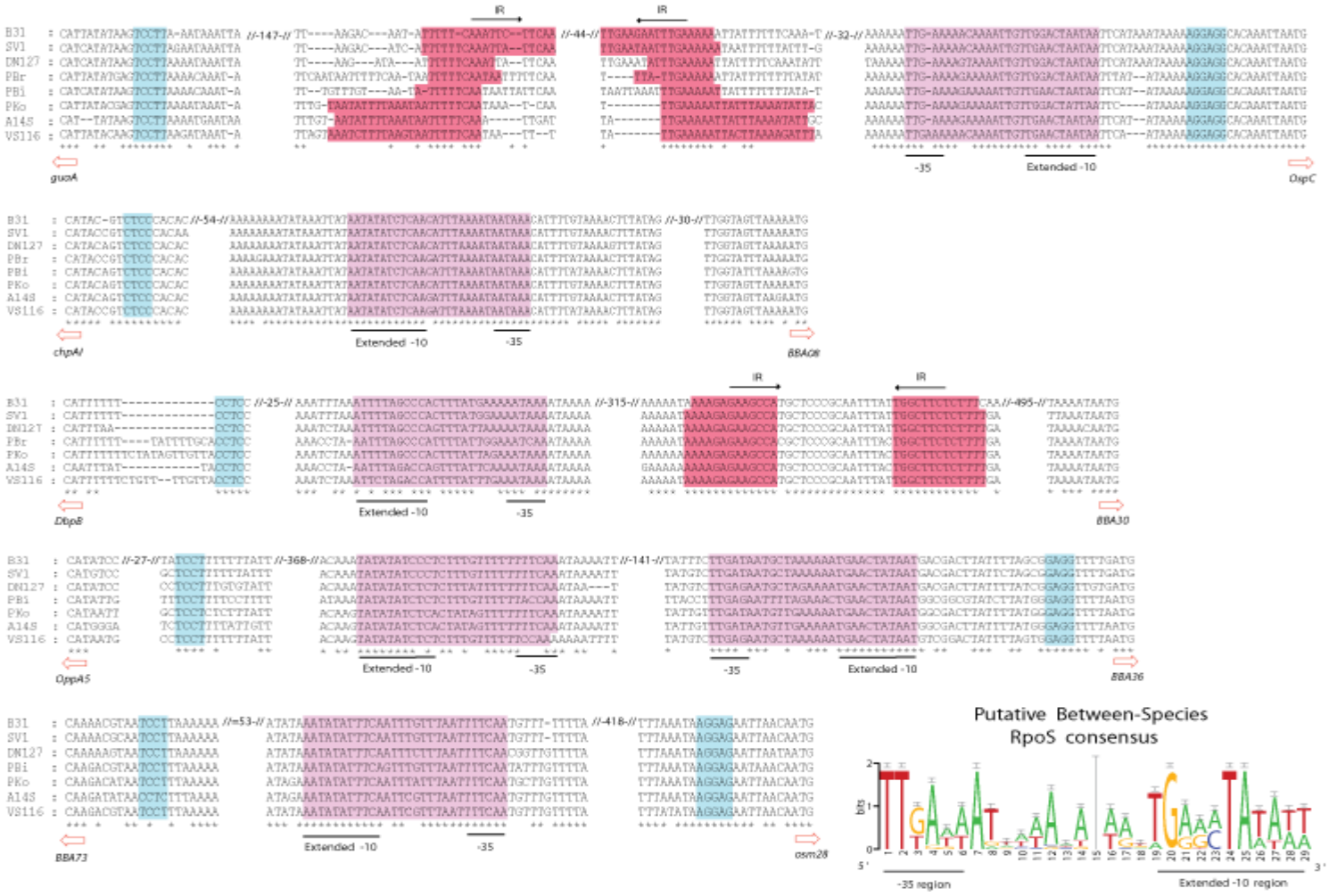
Figure 2-4: Identified putative functional elements within cp26 and lp54 plasmids. Cp26 is represented by purple, while lp54 is depicted by the color green. Moving inwards from the outermost circle, we illustrate PCIBs (orange), between-species % identity, conserved palindromes, and conserved BLAST hits. Directionality of BLAST hits are represented by red and grey for (+/+) and (+/-) hits respectively. Predicted, stable ncRNA structures for longest conserved palindromes (>14 nucleotides) are depicted in the lower right.

RpoS BINDING SITES AND BETWEEN SPECIES CONSENSUS SEQUENCE

Putative binding sites were identified in our cp26, and lp54 IGS segments using an algorithm which incorporated the previously published *B. burgdorferi B31* RpoS-binding sequence. We detected a total of six RpoS-binding, cis-regulatory sequences in five IGS orthologous segments: *BBB18-BBB19*, *BBA07-BBA08*, *BBA25-BBA30*, *BBA34-BBA36*, and *BBA73-BBA74* (Figure 2-5). Interestingly, these results included two previously unreported binding sites within the intergenic region upstream of *BBA34* and *BBA73* respectively. Paralogous and orthologous comparisons of our identified putative binding sites revealed significant differences from the previously published consensus sequence. Using the paralogous alignments of our predicted sites as our input into the web-based application, WebLogo, we predicted a new putative between-species RpoS consensus sequence (Figure 2-5).

Figure 2-5: Six putative RpoS binding sites were identified within five IGS of cp26 and lp54 plasmids. Between species alignment contain one representative sequence per genospecies where available. Putative RopS binding sites are shaded purple with -10and -35 regions labeled. Ribosome binding sites are shaded blue. Conserved regulatory inverted repeats are colored red with their directionally denoted by black arrows (above). Start codons are underscored by red arrows. Gene names are also located below their respective start codons. Newly proposed between-species consensus illustrating directionality and -10 and -35 regions is depicted as a Web-logo plot (right).

IDENTIFIED PUTATIVE RpoS BINDING SITES FIGURE 2-5



DISCUSSION

CONSERVED PLASMIDS COMPONENT OF CORE GENOME CONTAIN FUNCTIONAL ELEMENTS WITHIN SIGNIFICANTLY CONSERVED IGS AND PCIBs

Many different computational-based comparative methods have been developed for the identification of putatively functional elements, and the determination of natural selective forces within non-coding genomic regions. Such analyses have been previously used to successfully identify ncRNA in *Buchnera* (Degnan, Ochman, and Moran 2011), promoters in *Escherichia coli* and *Mycobacterium tuberculosis* (Rangannan and Bansal 2010), transcription terminators in *Bacillus subtilis* (Kingsford, Ayanbule, and Salzberg 2007, Naville et al. 2011) and putatively functional, adaptive non-coding elements in *Drosophila melanogaster* (Andolfatto 2005). In the current study, we employed the use of evolutionary, free-energy and structural analyses, in order to identify signatures of conservation, and putatively functional sites within the intergenic regions of the *B. burgdorferi* genome. For the most part, we observed that the intergenic regions of cp26 and lp54 plasmids were predominantly under purifying selection (conserved). In fact, only one IGS, *BBA24-BBA25* was detected to be under adaptive evolution between species. This single occurrence of adaptive evolution however, is considered to be a probable artifact due the fact that *BBA24-BBA25* is a relatively short IGS (~120 nucleotides), with a highly adaptive flanking ORF *BBA24*. Both plasmids contained significantly conserved IGS and PCIBs which contained predicted functional elements (Table 3). The most conserved segments of our intergenic regions; PCIBs, were observed to primarily contain predicted transcription terminators, and in a few cases, either the -10 or -35 regions of known and putative promoters. This observation suggested that elements such as ncRNAs and promoters may be structurally conserved, but not perfectly conserved in primary sequences among species; a telltale sign of possible differences in gene regulation mechanisms. Within our PCIBs, we observed two

interesting results: six conserved inverted repeats; which were predicted to form stable ncRNA and highly conserved sequences shared among different PCIB segments. We postulate that these shared sequences between different PCIBs may be putative regulatory elements, and that the genes in their downstream regions probably share a common regulatory pathway. The additional observation of conserved inverted repeats across different species was very intriguing given that these types of complementary sequences have been shown to be present, and required, for the regulation of two known virulent genes; *ospC* and *dbpA* in *Borrelia*. Even more interesting is the fact that these inverted repeats were predicted to form structurally stable ncRNAs. In *Borrelia* ncRNAs are known to be involved post-transcriptional regulation. Although the precise function of the identified inverted repeats is unknown, the fact that they are predicted to form stable ncRNA structures suggests that they may be functionally-important regulatory elements of gene expression in the *B. burgdorferi s.l.* complex.

TWO NEW PUTATIVE GENES UNDER DIRECT RPOS REGULATION

RpoS is known to be a master regulator of transcription for virulent genes in *B. burgdorferi.* Despite this obvious crucial role, to date only a total of five genes from cp26 and lp54 collectively have been reported to be directly transcribed this alternative sigma factor. These genes include: *ospC* and *dbpA,* which encode virulent lipoproteins; *chpAI* and *A66*, which are expressed to facilitate efficient transmission of *B. burgdorferi* to mammalian hosts; and *oppA5,* a gene involved host-specific adaptation of *B. burgdorferi* in vertebrate hosts. Our current analyses revealed two additional genes with putative RpoS binding sites in their upstream regions *A36*, and *A73*. Notably, these newly identified putative *cis*-regulatory sequences are relatively close in proximity to ribosome binding sites; an observation which further supports

their probable role as true RpoS binding sequences. Experimental studies have revealed *A36* to be an outer surface protein which localizes to the surface of *B. burgdorferi* during mammalian host infection, while *A73* has been shown highly up-regulated in temperature and pH conditions which simulate environmental changes during tick to host transmission (Brooks et al. 2006, Gilmore et al. 2008). Additionally, *A73* has been previously identified as a potential vaccine candidate due to its observed antigenicity and conservation among other *B. burgdorferi s.l.* species (Poljak et al. 2012, Wywial et al. 2009). To further bolster our results, studies show that both *A36* and *A73* were highly up-regulated in the presence of RpoS(Caimano et al. 2007).

Notably, none of our identified PCIBs contained an entire RpoS-binding sequence. We did observe however, that the -10 or -35 regions of these cis-regulatory sequences were contained within some PCIBs. The lack of primary sequence conservation observed both within orthologs and between paralogs comes as no surprise given the fact that in smaller genomes, binding specificity at promoter regions tends to be more relaxed than in larger genomes. In accordance with the preceding theory, we noted some variability in the -10 and -35 regions of the previously published RpoS consensus sequence; despite the fact that it was based on observations from a single *B. burgdorferi* strain *(B31*). Our study was base on twenty-four different strains of *B. burgdorferi s.l.* spanning eight different species. A consensus sequence of RpoS-binding sequences paralogs from our inter-specific data show significant difference at five nucleotide positions to the previously published consensus sequence. This newly proposed between-species consensus sequence would prove to be a valuable tool towards the identification, exploration and verification of RpoS-binding *cis*-regulatory sequences in *B. burgdorferi*.

CONCLUSION

Collectively, we have reported a number of identified putatively functional elements present in the IGS of *B. burgdorferi*. The most significant of these elements are those considered to be conserved across all species either structurally or in their primary sequence. Further investigation into such identified elements including ncRNA, predicted promoters, and putative transcription terminators could provide valuable insights into the mechanisms behind genome-based gene regulation in *B. burgdorferi*.

CHAPTER III STRUCTURAL AND PHYLOGENETIC ANALYSES OF OspA AND OspB

INTRODUCTION

OspA AND OspB

A primary strategy employed by *B. burgdorferi* to evade host immune systems and to maintain its persistence in nature is the differential expression of its outer surface proteins. Two lipoproteins located on the lp54 plasmid: OspA and OspB, were shown to be selectively expressed under a single sigma 70 (σ^{70}) dependent *ospAB* operon, and are known to play a crucial role in protecting *Borrelia* as it navigates between the biologically distinct environments of its vector and mammalian host. Specifically, OspA and OspB were observed to be the most prominent surface molecules present on *Borrelia burgdorferi* in engorged and unfed ticks (Xu et al. 2010). Notably, both are down regulated as the *Borrelia* transverses from the tick to its mammalian host. This precise regulation of OspA/B expression was observed to be accomplished by *cis*-regulatory elements flanking the *ospAB* promoter (Xu, McShan, and Liang 2010).

OspA has been shown to bind the tick receptor TROSPA and has also been implicated as an antibody-shielding agent during blood meals on immune hosts (Battisti et al. 2008, Pal et al. 2004). Interestingly, other studies suggest that OspA also primes and activates human neutrophils (Hartiala et al. 2008). OspB, an apparent paralog of OspA, has also been shown to be an adhesive agent required for *Borrelia* adhesion to the tick gut, and survival within the vector (Neelakanta et al. 2007). Additionally, studies reveal that OspB also has the ability to inhibit phagocytosis and oxidative burst of human neutrophils(Hartiala et al. 2008).

Both OspA and OspB may be selectively expressed in the late stages of human Lyme disease. Structural studies have revealed key protective B-cell epitopes in the C-terminal domain of OspA which bind to the murine monoclonal antibody (LA-2), and an analogous region in the C-terminal OspB which binds to fragment antigen binding (Fab) H6831(Becker et al. 2005, Ding et al. 2000). Notably, full-length crystalline structures are only available for OspA. The OspB results mentioned above were obtained via the crystal structure of a C-terminal fragment of OspB spanning residues 152-296. Despite the fact that numerous studies have highlighted important roles for both OspA and OspB, the structural and genomic basis for their observed functionality remains unclear. In this current study, we employ structural and evolutionary analyses in order to gain insights into the functional details of OspA and OspB.

MATERIALS AND METHODS

DATA ASSEMBLY

Orthologous inter-specific ORFs sets for OspA and OspB were obtained from our online *Borrelia* ORF database http://borreliagenome.org/orth_get. These contained sequences from approximately eight strains of *Borrelia*, representing eight different *B. burgdorferi s.l.* species including: *B. burgdorferi s.s.*, *B. finlandensis, B. bissettii*, *B. afzelii*, *B. spielmanii*, *B. garinii*, *B. bavariensis* and *B. valaisiana*. The strains retrieved and used in this study included: B31, SV1, DN127, PKo, A14S, PBr, PBi, and VS116 respectively.

STRUCTURAL MODELING OF OspB AND COMPARATIVE ANALYSIS TO OspA

To evaluate putative structural and hence functional differences between *B. burgdorferi s.s* B31 OspA and OspB lipoproteins, a tertiary structure model of OspB was constructed for our

B. burgdorferi s.s B31 strain. The crystal structure of OspA (PDB: 1OSP) which shares 56% amino acid sequence identity with OspB (Figure S3), served as a structural template for the homology modeling of *B. burgdorferi s.s* B31 OspB. Template based modeling was performed using the I-TASSER software (Zhang 2008) which combines ab intio prediction with template based modeling.

Model assessment was done using two different software packages: Verify 3D (Luethy et al., 1992) and ProSA (Wildenstein, M. 2007). Specifically, Veriy3D analyzes the compatibility of the predicted structure's atomic model (three-dimensional profile computed from the structure atomic coordinates) with its own amino acid sequence. Here, atomic models computed from accurate protein structures match their own sequences with high scores, while incorrectly modeled regions score poorly. In Verify3D, scores range between -10 and 10. Incorrectly modeled regions will carry negative scores (Luethy and al 1992). ProSA evaluates the energy of a structure using atomic coordinates, to produce an energy plot of its residual energies. Generally, on the energy plot, positive energies correspond to poorly modeled regions. Loop refinement for our predicted model was conducted using the stand alone MODELLER (Eswar et al. 2007) software.

The generated model was visualized and structurally aligned to OspA using UCSF Chimera v1.6.1; a visualization system for exploratory research and analysis (Pettersen et al. 2004)..

Generation of electrostatic potential maps for OspA and OspB were accomplished using the software Delphi (Honig and Nicholls 1995) and visualized using PyMOL Molecular Graphics System, (Version 1.5.0.1 Schrödinger, LLC) in order to compare possible differences

in electrostatic profiles. Hydrophobicity surface maps for the comparison of our modeled OspB structure to OspA were generated using the software UCSF Chimera v1.6.1 (Pettersen et al. 2004)

Surface topology analysis for the identification of putative pockets was conducted using two different web-based algorithms: Pocket finder (Laurie and Jackson 2005) and CASTp (Dundas et al. 2006). Only pockets identified by both algorithms were reported. The identification of possible hydrogen bonds between residues within OspA and OspB structures respectively, was done using the software UCSF Chimera v1.6.1 (Pettersen et al. 2004)

EVOLUTIONARY ANALYSES

ANALYSIS FOR FIXED DIFFERENCES BETWEEN OspA AND OspB

In order to identify fixed differences (amino-acid substitutions) between OspA and OspB which may contribute to their divergent functionality, we conducted a branch-site test of positive selection using the CODEML program of the PAML software package (Yang 2007). Here, we employed a neighbor-joining tree which was generated using the software CLUSTALW (Larkin et al. 2007). The phylogenetic tree consisted of OspA and OspB inter-specific orthologs, with OspA as our foreground branch. Key PAML "branch-site model" parameters included, "runmode = 0" for user-defined tree, "model = 2" for 2 or more d_N/d_S ratios for branches; "NSsites = 2" for positively selective sites. Only positively selected sites with a probability >95% under the PAML Naive Empirical Bayes (NEB) analysis and which also represented fixed differences were reported.

ANALYSIS OF SIGNIFICANTLY INCONSISTENT AMINO ACID STATES IN INTER-SPECIFIC ALIGNMENTS OF OspA AND OspB

Our previous evolutionary analyses on inter-specific orthologous ORF sets revealed OspA, but not OspB contained amino-acid sites under significant positive selection. This observation was intriguing given that OspB is observed to have higher variability in its between-species alignment when compared to OspA. To further investigate this interesting result we conducted parsimony analyses on all sites in both OspA and OspB inter-specific alignments with a probability of $d_N/d_S > 1$ under PAML Naive Empirical Bayes (NEB) analysis.

RESULTS

STRUCTURAL COMPARISON OF OspA AND OspB TERTIARY MODELS

To further characterize OspB, we constructed its tertiary structure via template-based modeling and then assessed its similarity to the well-defined crystalline structure of OspA.

Template-based modeling yielded a tertiary structure similar to OspA having a repetitive antiparallel β topology. Here, we observe a non-globular domain of "freestanding" sheet connecting globular N- and C-terminal domains (Figure 3-1 [A]). Our study also revealed that OspA and OspB shared similar electrostatic characteristics. Specifically, we observed OspA and OspB to be relatively basic in their non-globular central domain, with both lipoproteins containing a hydrophobic cavity in a positively charged cleft at the C-terminal domain; a potential binding site for an unknown ligand (Figure 3-1 [B]). Notably, we did observed one key difference where The N-terminus of OspB is relatively more basic than that of OspA (Figure 3-1 [B]). A hydrophobicity surface analysis revealed that OspA and OspB had strikingly differently hydrophobic profiles. Here, OspA was observed to be more hydrophobic along the non-globular

domain between the N- and C-terminals. Additionally, the N-terminal of OspA was relatively more hydrophobic than the N-terminal of OspB. Verify3D and Prosa verification revealed that our predicted OspB model was of good quality. Structural alignment revealed a global root-mean-square deviation (RMSD) of 1.81 angstroms (Figure 3-1 [C]). A number of sites were also observed to be absolutely conserved across both OspA and OspB paralogs including a C-terminal Tryptophan identified to be an important residue among OspA antibody binding epitopes (Figure 3-1 [A]).

Topography analysis of OspB revealed two distinct pockets directly opposite to each other. The first pocket ("A") included residues: Arg-162, Thr-166, Thr-167, Leu-168, Glu-184, Thr-185, Lue-186, Phe-260 and Glu-264, while the second identified pocket ("B") contained residues: Lys-158, Leu-160, Try-170, Asn-178, Leu-196, Val-197, Gly-198, Thr-201, Ile-216, and Glu-217 (Figure 3-3). It should be noted that both OspA and OspB contained these putative clefts.

EVOLUTIONARY ANALYSES

Our previous evolutionary analysis on individual orthologous sets revealed that OspA, and not OspB was under significant positive selection. Parsimony analysis showed that in OspA, the identified positively selected site was phylogenetically inconsistent, while sites identified as significantly variable in OspB were phylogenetically consistent (Figure 3-2).

Branch-site PAML analysis of our OspA/B paralogous identified fourteen fixed differences; twelve of which were significantly (>99%) positively selected for under PAML Naive Empirical Bayes (NEB) analysis (Figure 3-2). Identified fixed differences were assigned into three categories: type I, II, and III. Type I represent fixed substitutions where amino-acids

were consistent within, but not between paralogs. Type II represent fixed differences where amino-acids were conserved in OspA but variable in OspB, while type III represented fixed differences where amino-acids were conserved in OspB but variable in OspA (Figure 3-2 and 3-3). We also observed from our structure based alignment, that the OspB residues Arg-162, Thr-166 and Glu-184; previously identified as a putative serine protease catalytic triad (Becker et al. 2005), were also conserved in OspA. Notably, this conserved catalytic triad was also found to be within close proximity to residues representing fixed amino-acid differenced between OspA and OspB in our tertiary structures. Here, our analysis showed that the OspB reside Asp-263; a type III fixed difference between OspA and OspB may form an H-bond (hydrogen bond) with Thr-166; a member of the putative catalytic triad (Figure 3-4). No such analogous bond was observed in the OspA structure 1OSP (not shown).

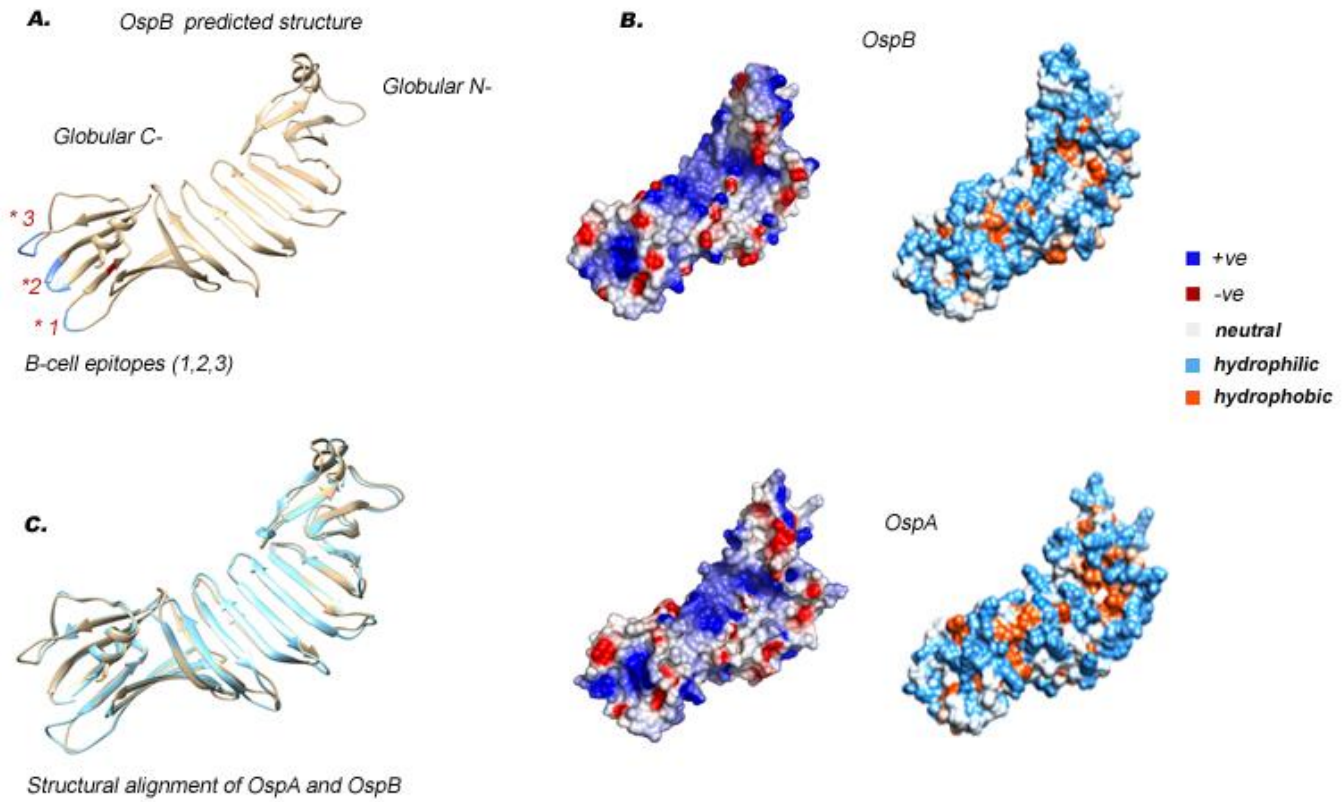


Figure 3-1: All analyses are depicted in the same orientation**.** (A); Predicted tertiary structure of OspB showing N-and C-terminal globular domains as well as B-cell epitopes. (B); Comparison of electrostatic and hydrophobic profiles of OspB (top) and OspA (bottom). These lipoproteins show notable differences in their electrostatic and hydrophobic profiles in their N and C-terminal domains, as well as in the central region proximal to the C-terminus. (C); Structural alignment between OspA and OspB reveals a global RMSD of 1.81, and suggests similar topology. A critical and absolutely conserved (in OspA and OspB) Tryptophan residue in the C-terminal domain that is involved in antibody binding, is shaded red.

CONSERVATION AND FIXED DIFFERENCES BETWEEN OspA (blue branches) AND OspB (red branches)

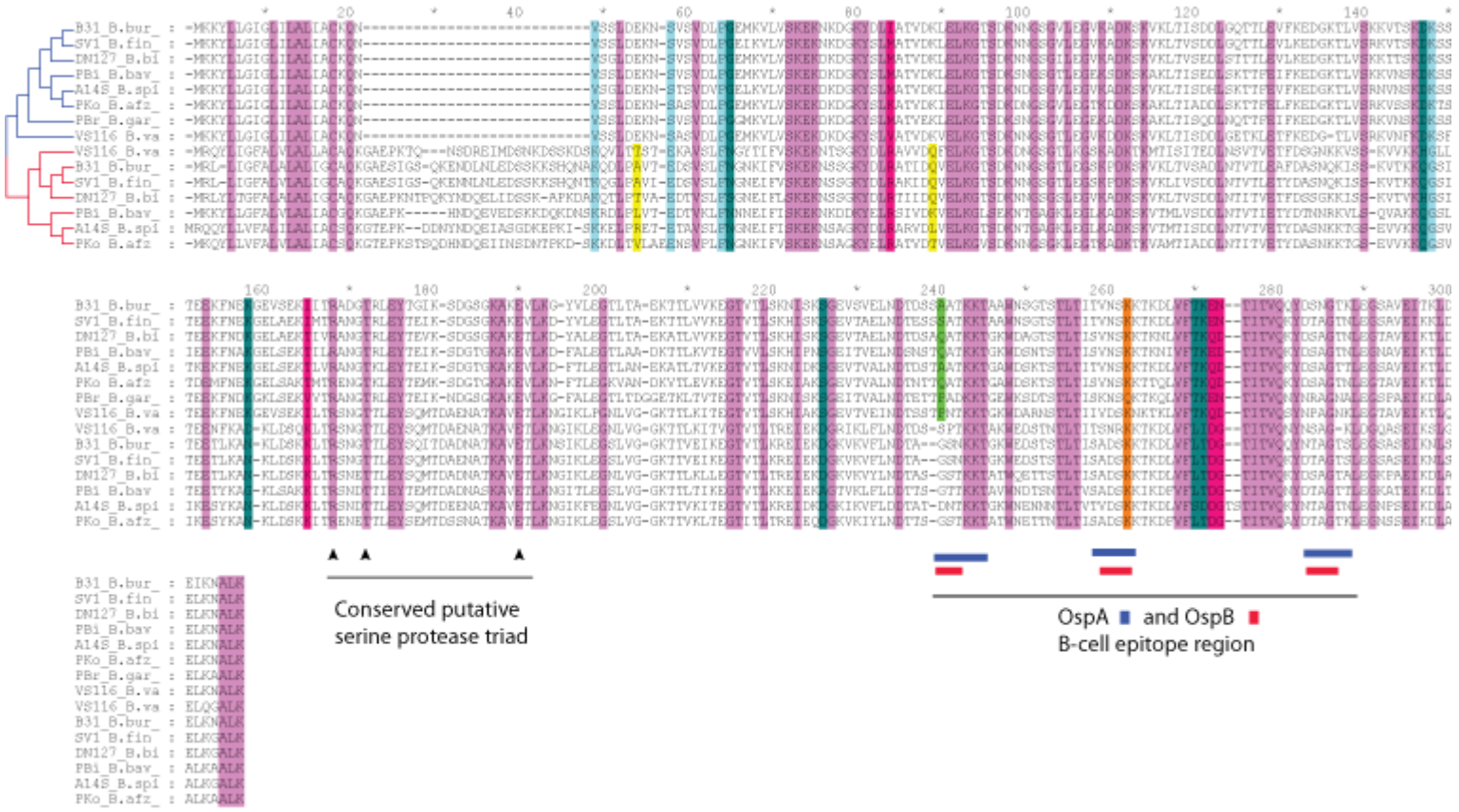


Figure 3-2: MUCSLE generated sequence alignment of OspA (blue branches on cladogram) and OspB (red branches on cladogram). Purple shading indicates sites absolutely conserved between the two lipoproteins. Three categories of fixed differences were identified between OspA and OspB. Blue shading represent fixed substitutions where amino-acids were consistent within but not between paralogs. Sites shaded in teal represent fixed differences where amino-acids are conserved in OspA but variable in OspB. Sites shaded in rose represent fixed differences where amino-acids are conserved in OspB but variable in OspA. Yellow shading depicts highly variable sites in OspB. These sites are phylogenetically consistent and therefore did not register as being under positive selection in our PAML analysis. Green shading indicates phylogenetically inconsistent and hence positively selected sites within the OspA orthologous set. Regions of B-cell epitope are indicated by red bars. The previously indentified putative serine protease catalytic triad, Thr-166, Arg-162 and Glu-184, are indicated by black arrows. Lys- 253 a proposed antibody "hotspot" is shaded in orange. Notably this Lys is conserved in all species except *B. garinii.*

PUTATIVE FUNCTIONALLY IMPORTANT SITES ON OspB

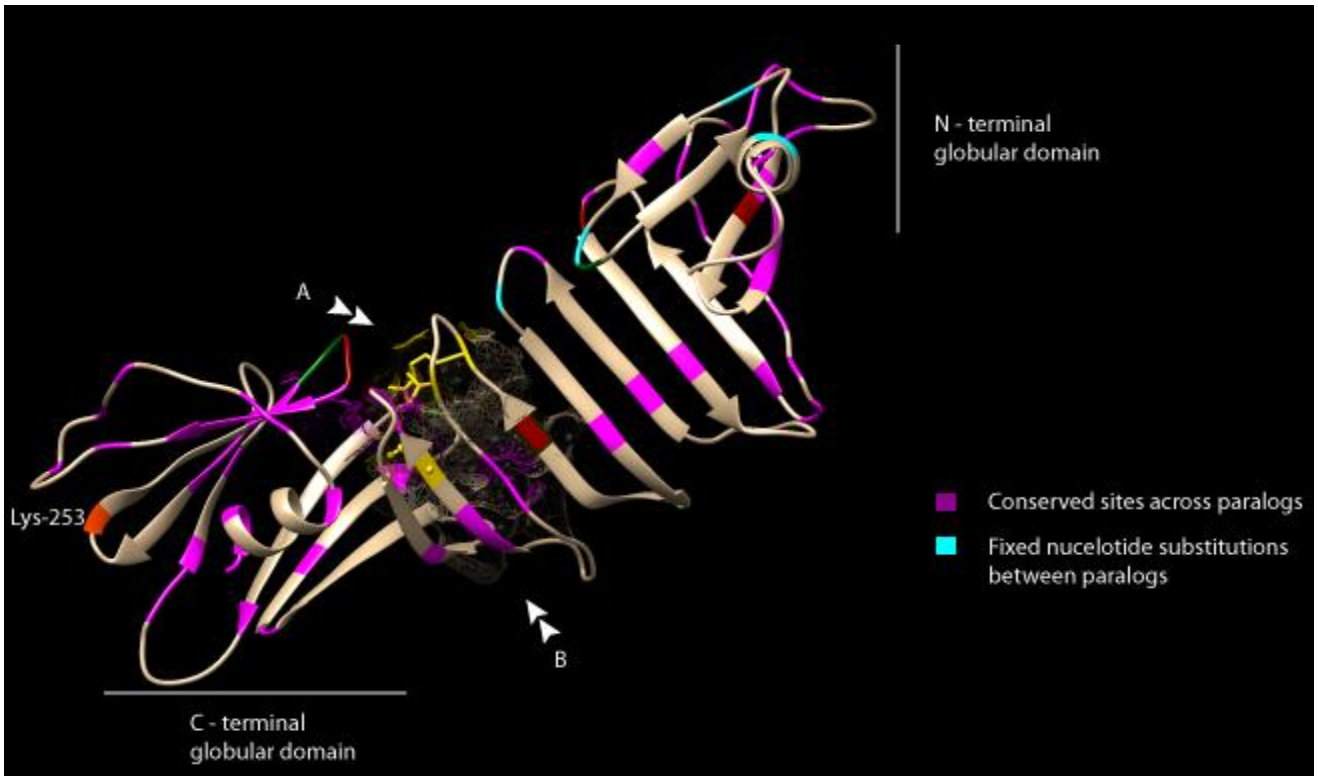


Figure 3-3: Predicted structure of OspB illustrating putatively functional sites. N- and C-terminal domains are labeled. Absolutely conserved sites between OspA and OspB are illustrated in the color purple. Fixed substitutions where amino-acids were consistent within but not between paralogs are shaded light blue (type I). Fixed differences where amino-acids are conserved in OspA but variable in OspB are shaded green (type II). Sites shaded in red represent fixed differences where amino-acids are conserved in OspB but variable in OspA (type III). The proposed antibody "hotspot" is shaded in orange. Arrows A and B represents two predicted binding pockets (mesh area). Notably, the predicted binding pocket "A" was observed to be adjacent to sites identified as fixed differences between the two lipoproteins. These may represent sites contributing to the observed functional differences between OspA and OspB. The putative serine protease triad conserved in both OspB and OspA are shaded in gold. Interestingly, all three residues (Arg-162, Thr-166 and Glu-184) are located adjacent to the predicted binding pocket "A "of OspB.

OspB PUTATIVE SERINE PROTEASE

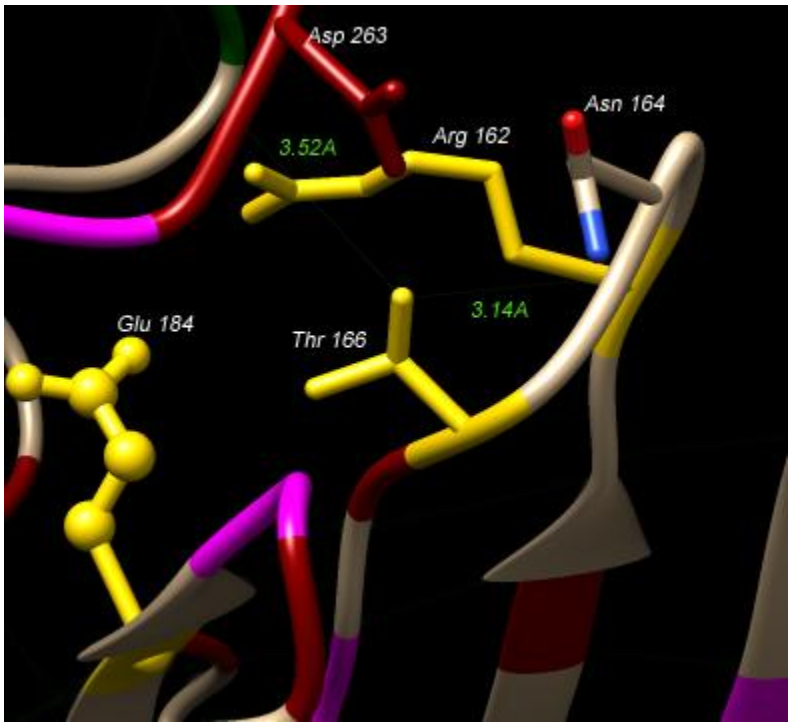


Figure 3-4: Previously identified putative catalytic triad of OspB. Residues of the putative catalytic triad are observed to be within close proximity of each other in the tertiary structure of OspB (gold). Asn-164; a possible influential residue was identified in a recent structural study to form an H-bond with Thr-166 (labeled). Possible H-bonds and relative distances are represented in light green. A total of seven functional differences spanning three categories, type I (light blue), type II (green) and type III (dark-red) are observed within the vicinity of the putative triad residues.

DISCUSSION

CHARACTERIZATION OF OspA AND OspB

To gain insight into the structural and genomic basis for the observed functional differences between OspA and OspB, we constructed a predicted model of full-length OspB, and conducted evolutionary and structural analyses against it paralog, OspA.

The tertiary structure of OspB revealed it to be structurally similar to OspA, consistent with the evolutionary evidence of these two lipoproteins being paralogs. It should be noted however that OspB has a 23 amino-acid insertion at its N-terminal region resulting in a predicted short helix (Figure 3-1 [A]). This observed structural difference between OspA and OspB N-termini is probably associated with distinct membrane-associated functions. In addition to their similar tertiary structure, OspA and OspB were observed to be similar in their electrostatic profiles and primarily positively charged (basic) along the central domain, suggesting that they could be attracted to negatively charged membranes through non-specific electrostatic interactions. Notably, although the central non-globular regions of both lipoproteins were primarily basic, OspB was slightly more basic in its central domain and proximal N-terminal regions (Figure 3-1 [B]). This observation suggests that both OspA and OspB may differentially interact with polar component or ligands of the epithelial membrane in the tick gut along this region. Additionally, our hydropathy analysis supported this premise revealing that the hydrophobicity profiles of OspA and OspB were strikingly different specifically in the N-terminal region, C-terminal region, and the central domain proximal to the C-terminus. The C-terminal barrel domains of both lipoproteins have been shown to interact with specific antibodies within particular looped regions (Figure 3-1 [A]) (Ding et al. 2000, Becker et al. 2005) and are observed to feature, pronounced, predominantly basic clefts with adjacent patches of hydrophobic residues. This observation suggests that these proteins recognize or interact with similar or closely related targets via their C-terminal domains.

Evolutionary analyses revealed three categories of fixed amino-acid substitutions between OspA and OspB: type I; fixed substitutions where amino-acids were consistent within but not between paralogs; type II, fixed differences where amino-acids are conserved in OspA

but variable in OspB; and type III, fixed differences where amino-acids are conserved in OspB but variable in OspA. Notably, all functional differences: type I, II and type III were confined to the region proximal to the C-terminus (within the vicinity of a putative ligand binding pocket labeled "A") (Figure 3.3), and the N-terminal domain. This observation is not only consistent with the observed difference in hydrophobic profiles, but also suggest that both these regions may significantly contribute to OspB specific functionality especially at type I and III sites where amino-acid changes from OspA are 100% conserved within OspB.

Despite the fact that OspB sequences (71.1% sequence identity) were more variable than OspA sequences (80.1% identity) in cross-species comparisons, and that the *ospB* gene showed a higher nonsynonymous-to-synonymous ratio of nucleotide substitution rates (K_A/K_S =0.34) than *ospA* (K_A/K_S =0.29), no significantly positively selected sites were identified in OspB. Interestingly, our evolutionary analyses did identify a significantly positive selected site in OspA which was also observed to be within the one of the looped region mentioned above as an identified putative antibody epitope (Ding et al. 2000) (Figure 3-2). Parsimony analyses revealed that although OspB contained a number of variable sites, the observed variability was phylogenetically consistent explaining the absence of significant positive selection (Figure 3-2). Our analyses also reveal a number of sites absolutely conserved across both paralogs. The fact that these sites are conserved in all species across both paralogous lipoproteins suggests their putative functional importance.

Truncated C-terminal forms of OspB are commonly observed both in *vivo* and in *vitro* (Becker et al. 2005). Structural studies on these truncated variants have led to the suggestion that OspB may have an intrinsic autoproteolysis characteristic; given that it contain three residues: Thr-166, Arg-162 and Glu-184, which resemble the catalytic triad of the serine proteases

(Becker et al. 2005). The catalytic action serine proteases usually involve the interaction of a general base (His), an acid (Asp), and a nucleophile (Ser). Notably, two residues of OspB putative catalytic triad; Thr and Glu are also commonly found in some members of the proteases family as a nucleophile and acid respectively. Additionally, studies show that Arg can function as a catalytic base in some cleavage enzymes (Polgár 2005). Cross-species sequence comparison of OspA and OspB showed that these three residues were absolutely conserved between these two lipoproteins; across all species (Figure 3-2). This observation suggests that the three residues mentioned above are not definitively responsible for the observed instability of OspB relative to OspA. Notably, we did identify a putative ligand binging pocket proximal to the C-terminal barrel region of OspB (Figure 3-3 pocket "A") which corresponds to an analogous identified putative binding site in OspA (Pawley, Koide, and Nicholson 2002). The above mentioned conserved putative serine protease residues: Thr-166, Arg-162 and Glu-184, were observed to be within the vicinity of this predicted OspB cleft. Interestingly, seven of the fourteen identified fixed differences were located within the region of this putative binding pocket (Figure 3-3 and 3-4). In Figure 3-4, the predicted tertiary structure of OspB showed, that residues of the putative catalytic triad are bought into close proximity to each other and another suggested influential residue; Asn-164 which may form an H-bond with Thr-166, and possibly aid in the formation of an oxyanion hole (Becker et al. 2005). Furthermore, our results show that Asp-263; a conserved residue in OspB, and a type III fixed difference between OspA and OspB may also form an H-bond with Thr-166, an observation not seen in the analogous (fixed difference) Glu-240 OspA residue. Collectively, these observations may explain the diverse functionality and the relative stability of OspA verses OspB.

Another interesting observation is that in OspB the residue Lys-253; cited to be a potential hot-spot for antibody recognition, is conserved in both lipoproteins across all genospecies except *B. garinii*; the species lacking the *ospB* gene.

Previous studies revealed that the functional binding domains of OspA are located within its central and COOH-terminus regions (Pal et al. 2000). It is then reasonable to assume that the functional domains of OspB are likely situated within the same structural domains; given their similar tertiary confirmation. Our results further support this notion given that OspA and OspB are functionally distinct, and that most of our observed differences in electrostatic, hydrophobic and evolutionary profiles between OspA and OspB are localized to the same C-terminal and central domain regions.

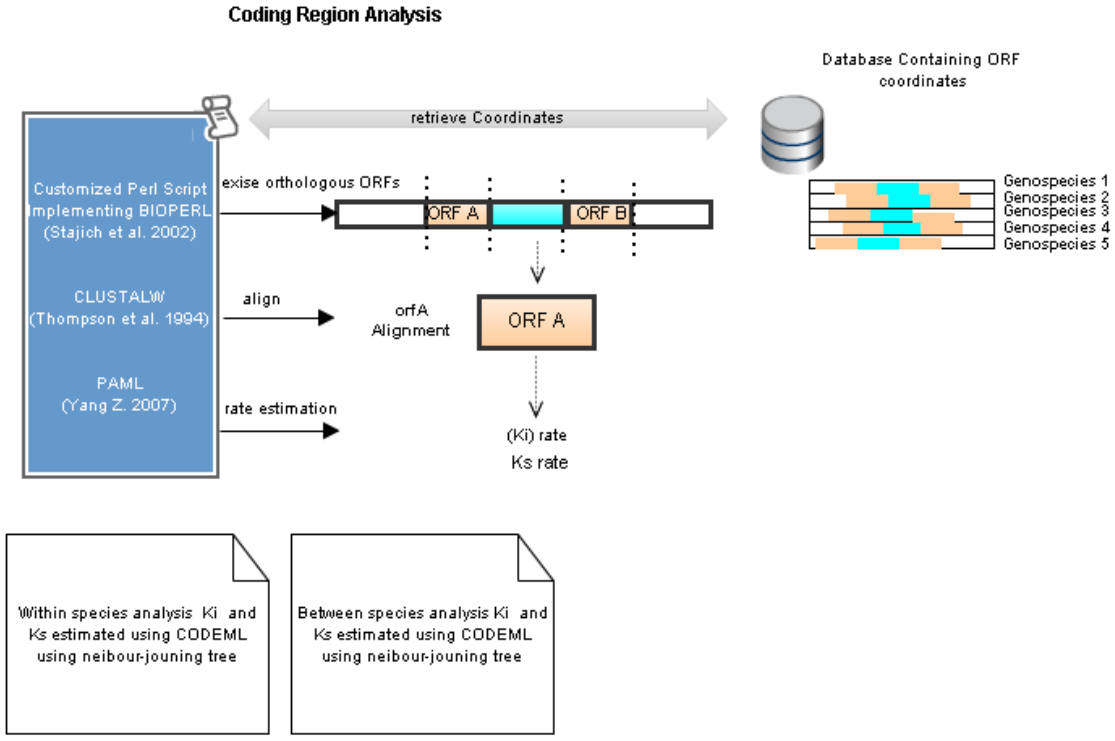
CONCLUSION

Despite the fact that numerous studies have demonstrated that OspA and OspB are multifunctional lipoproteins crucial to *Borrelia* as it navigates its enzootic cycle, the genomic and structural basis for their differential functionally is still unknown. Here, we have conducted structural and evolutionarily analysis, the results of which may shed some insight into the mechanisms behind the different functions of these two *B. burgdorferi* lipoprotein paralogs. To summarize, we have constructed a full-length predicted model of OspB and showed that it is structurally similar to OspA with a RMSD of 1.81. We showed that although there are a number of absolutely conserved sites across the central domain, and the N- and C-terminal regions, the observed fixed difference between these two lipoproteins were primarily concentrated in the region proximal to the C-terminal barrel domain and the N-terminal globular domain. The former observation is consistent with studies showing that the central and COOH-terminus regions as primary functional domains of OspA. Additionally the electrostatic and hydrophobic profiles of

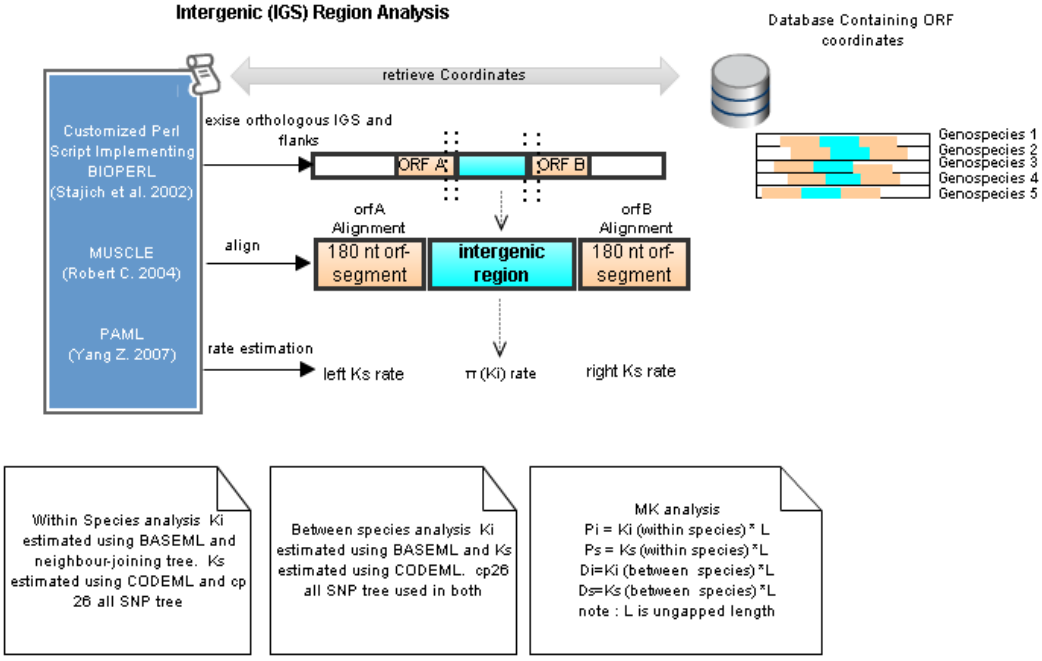
OspA and OspB also show marked variability along the same regions. Furthermore, we identified a putative binding pocket proximal to C-terminal globular domain of OspB. Additionally, residues within the region of our putative cleft represented fixed nucleotides substitutions between OspA and OspB paralogs. These sites may be functionally important and may contribute towards the observed functional difference between these two lipoproteins. Collectively our results provide some insight into the residues and structural regions within OspB that may be of functional significance.

APPENDIX

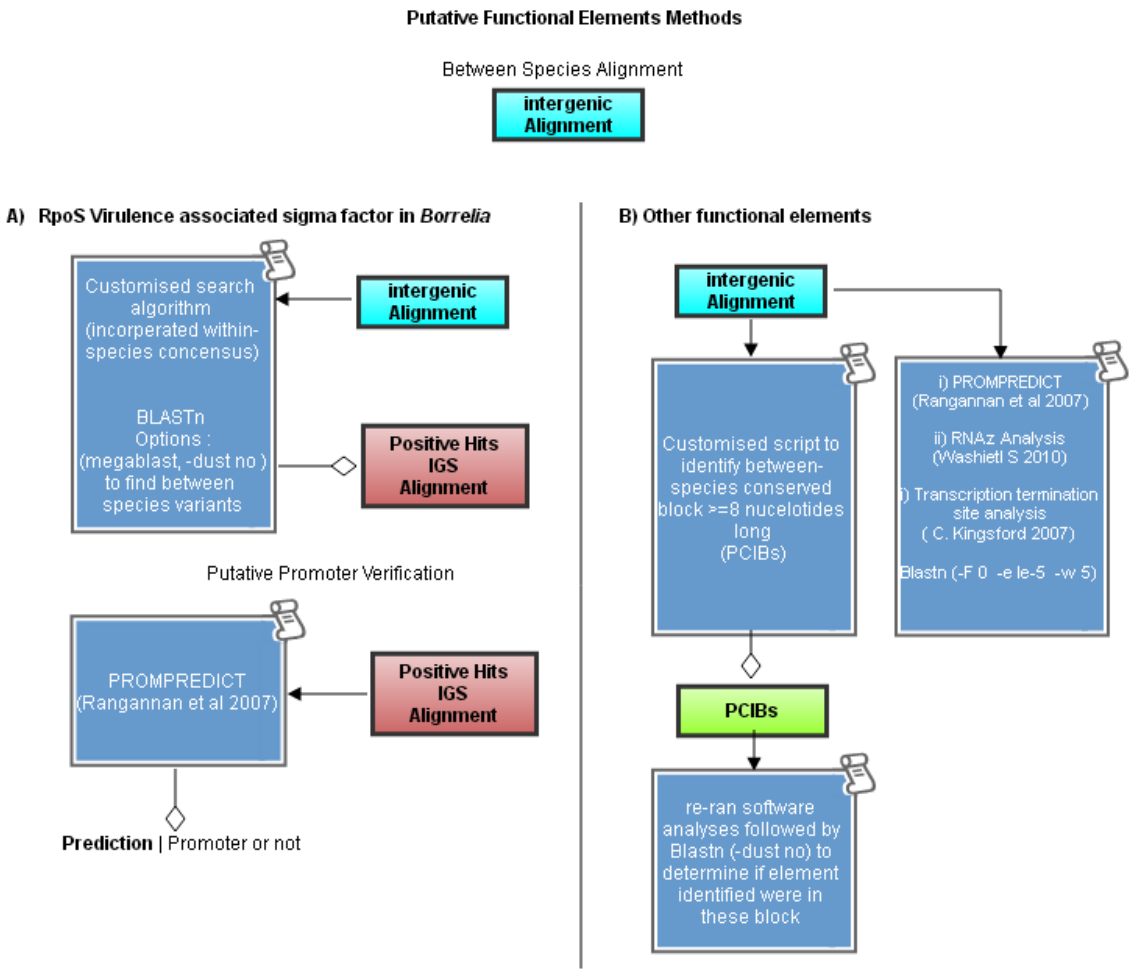
APPENDIX I: CODING REGION ANALYSIS LOGIC MAP



APPENDIX II: IGS ANALYSIS LOGIC MAP



APPENDIX III: IDENTIFICATION OF RPOS AND PUTATIVE FUNCTIONAL ELEMENTS LOGIC MAP



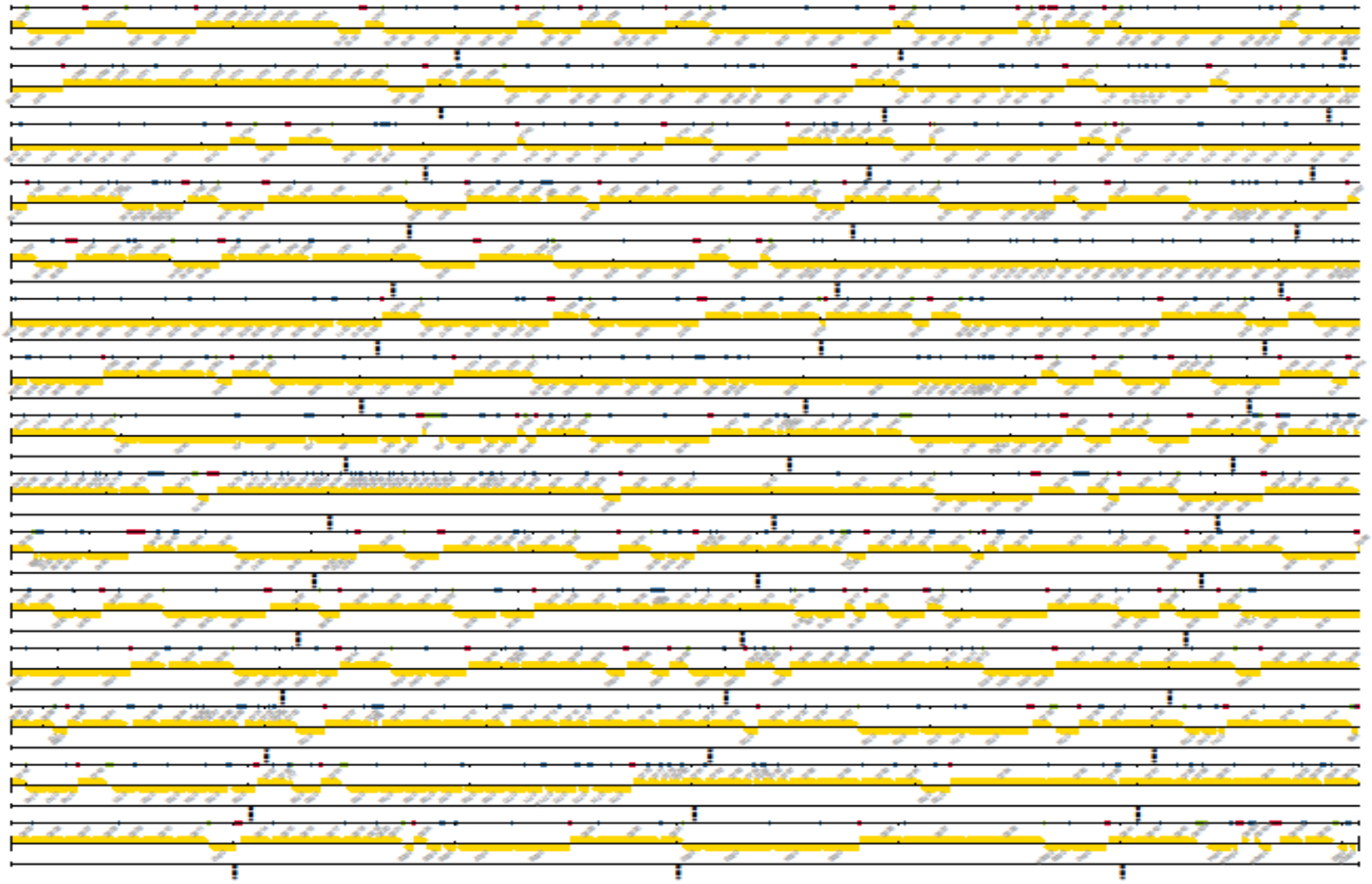


Figure S1: Main Chromosome ORFs used in the current study

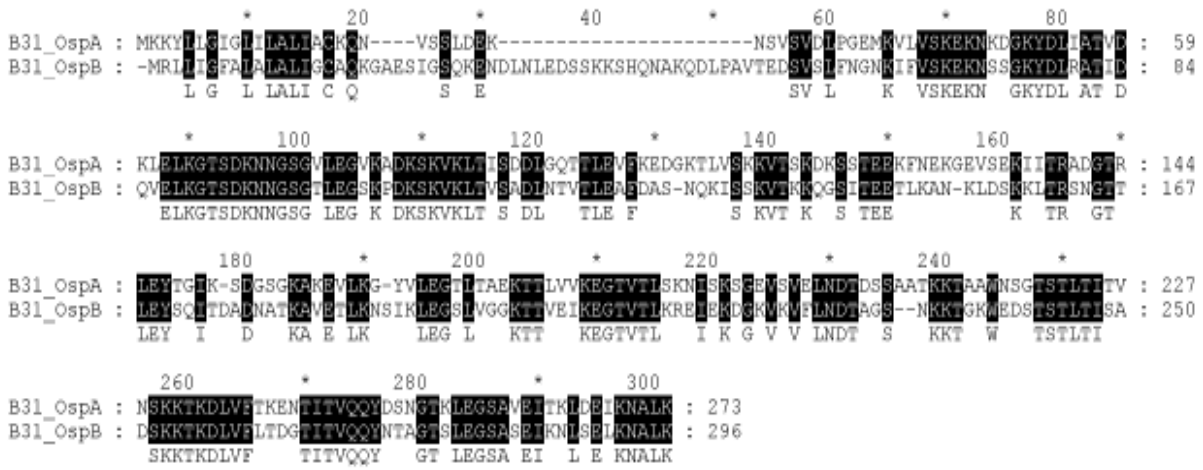


Figure S1: CLUSTALW generated amino-acid sequence alignment of B31 *B. burgdorferi s.s.* OspA and OspB paralogous proteins showing 56% sequence identity. Conserved amino-acid sites are shaded in black.

REFERENCES

Andolfatto, P. 2005. "Adaptive evolution of non-coding DNA in Drosophila." *Nature* no. 437 (7062):1149-52. doi: nature04107 [pii]

10.1038/nature04107.

Bacon, R. M., K. J. Kugeler, P. S. Mead, and Centers for Disease Control and Prevention (CDC). 2008. "Surveillance for Lyme disease--United States, 1992-2006." M*MWR Surveill Summ* no. 57 (10):1-9.

Battisti, J. M., J. L. Bono, P. A. Rosa, M. E. Schrumpf, T. G. Schwan, and P. F. Policastro. 2008. "Outer surface protein A protects Lyme disease spirochetes from acquired host immunity in the tick vector." I*nfect Immun* no. 76 (11):5228-37. doi: 10.1128/IAI.00410-08.

Becker, M., J. Bunikis, B. D. Lade, J. J. Dunn, A. G. Barbour, and C. L. Lawson. 2005. "Structural investigation of Borrelia burgdorferi OspB, a bactericidal Fab target." J *Biol Chem* no. 280 (17):17363-70. doi: 10.1074/jbc.M412842200.

Bernhart, S. H., I. L. Hofacker, S. Will, A. R. Gruber, and P. F. Stadler. 2008. "RNAalifold: improved consensus structure prediction for RNA alignments." B*MC Bioinformatics* no. 9:474. doi: 10.1186/1471-2105-9-474.

Biesiada, G., J. Czepiel, M. R. Leśniak, A. Garlicki, and T. Mach. 2012. "Lyme disease: review." A*rch Med Sci* no. 8 (6):978-82. doi: 10.5114/aoms.2012.30948.

Brisson, D., N. Baxamusa, I. Schwartz, and G. P. Wormser. 2011. "Biodiversity of Borrelia burgdorferi strains in tissues of Lyme disease patients." P*LoS One* no. 6 (8):e22926. doi: 10.1371/journal.pone.0022926.

Brisson, D., D. Drecktrah, C. H. Eggers, and D. S. Samuels. 2012. "Genetics of Borrelia burgdorferi." A*nnu Rev Genet* no. 46:515-36. doi: 10.1146/annurev-genet-011112-112140.

Brisson, D., and D. E. Dykhuizen. 2004. "ospC diversity in Borrelia burgdorferi: different hosts are different niches." G*enetics* no. 168 (2):713-22.

Brooks, C. S., P. S. Hefty, S. E. Jolliff, and D. R. Akins. 2003. "Global analysis of Borrelia burgdorferi genes regulated by mammalian host-specific signals." I*nfect Immun* no. 71 (6):3371-83.

Brooks, C. S., S. R. Vuppala, A. M. Jett, and D. R. Akins. 2006. "Identification of Borrelia burgdorferi outer surface proteins." I*nfect Immun* no. 74 (1):296-304. doi: 10.1128/IAI.74.1.296-304.2006.

Brooks, C. S., S. R. Vuppala, A. M. Jett, A. Alitalo, S. Meri, and D. R. Akins. 2005. "Complement regulator-acquiring surface protein 1 imparts resistance to human serum in Borrelia burgdorferi." J *Immunol* no. 175 (5):3299-308.

Caimano, M. J., R. Iyer, C. H. Eggers, C. Gonzalez, E. A. Morton, M. A. Gilbert, I. Schwartz, and J. D. Radolf. 2007. "Analysis of the RpoS regulon in Borrelia burgdorferi in response to mammalian host signals provides insight into RpoS function during the enzootic cycle." M*ol Microbiol* no. 65 (5):1193-217. doi: 10.1111/j.1365-2958.2007.05860.x.

Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden. 2009. "BLAST+: architecture and applications." B*MC Bioinformatics* no. 10:421. doi: 10.1186/1471-2105-10-421.

Casjens, S. 2000. "Borrelia genomes in the year 2000." J *Mol Microbiol Biotechnol* no. 2 (4):401-10.

Casjens, S., N. Palmer, R. van Vugt, W. M. Huang, B. Stevenson, P. Rosa, R. Lathigra, G. Sutton, J. Peterson, R. J. Dodson, D. Haft, E. Hickey, M. Gwinn, O. White, and C. M. Fraser. 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." M*ol Microbiol* no. 35 (3):490-516. doi: mmi1698 [pii].

Casjens, S. R., C. M. Fraser-Liggett, E. F. Mongodin, W. G. Qiu, J. J. Dunn, B. J. Luft, and S. E. Schutzer. 2011. "Whole genome sequence of an unusual Borrelia burgdorferi sensu lato isolate." J *Bacteriol* no. 193 (6):1489-90. doi: JB.01521-10 [pii]

10.1128/JB.01521-10.

Casjens, S. R., E. F. Mongodin, W. G. Qiu, J. J. Dunn, B. J. Luft, C. M. Fraser-Liggett, and S. E. Schutzer. 2011. "Whole-genome sequences of two Borrelia afzelii and two Borrelia garinii Lyme disease agent isolates." J *Bacteriol n*o. 193 (24):6995-6. doi: 10.1128/JB.05951-11.

Chen, S. L., C. S. Hung, J. Xu, C. S. Reigstad, V. Magrini, A. Sabo, D. Blasiar, T. Bieri, R. R. Meyer, P. Ozersky, J. R. Armstrong, R. S. Fulton, J. P. Latreille, J. Spieth, T. M. Hooton, E. R. Mardis, S. J. Hultgren, and J. I. Gordon. 2006. "Identification of genes subject to positive selection in uropathogenic strains of Escherichia coli: a comparative genomics approach." Pr*oc Natl Acad Sci U S A n*o. 103 (15):5977-82. doi: 0600938103 [pii]

10.1073/pnas.0600938103.

Crooks, G. E., G. Hon, J. M. Chandonia, and S. E. Brenner. 2004. "WebLogo: a sequence logo generator." Gen*ome Res no*. 14 (6):1188-90. doi: 10.1101/gr.849004.

Crother, T. R., C. I. Champion, J. P. Whitelegge, R. Aguilera, X. Y. Wu, D. R. Blanco, J. N. Miller, and M. A. Lovett. 2004. "Temporal analysis of the antigenic composition of Borrelia burgdorferi during infection in rabbit skin." In*fect Immun no*. 72 (9):5063-72. doi: 10.1128/IAI.72.9.5063-5072.2004.

Dale, J. L., M. J. Raynor, P. Dwivedi, and T. M. Koehler. 2012. "cis-Acting elements that control expression of the master virulence regulatory gene atxA in Bacillus anthracis." J B*acteriol no*. 194 (15):4069-79. doi: 10.1128/JB.00776-12.

Degnan, P. H., H. Ochman, and N. A. Moran. 2011. "Sequence conservation and functional constraint on intergenic spacers in reduced genomes of the obligate symbiont Buchnera." PLo*S Genet no*. 7 (9):e1002252. doi: 10.1371/journal.pgen.1002252.

Delcher, A. L., K. A. Bratke, E. C. Powers, and S. L. Salzberg. 2007. "Identifying bacterial genes and endosymbiont DNA with Glimmer." Bio*informatics no*. 23 (6):673-9. doi: 10.1093/bioinformatics/btm009.

Ding, W., X. Huang, X. Yang, J. J. Dunn, B. J. Luft, S. Koide, and C. L. Lawson. 2000. "Structural identification of a key protective B-cell epitope in Lyme disease antigen OspA." J M*ol Biol no*. 302 (5):1153-64. doi: 10.1006/jmbi.2000.4119.

Dundas, J., Z. Ouyang, J. Tseng, A. Binkowski, Y. Turpaz, and J. Liang. 2006. "CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues." Nuc*leic Acids Res no*. 34 (Web Server issue):W116-8. doi: 10.1093/nar/gkl282.

Dunham-Ems, S. M., M. J. Caimano, C. H. Eggers, and J. D. Radolf. 2012. "Borrelia burgdorferi requires the alternative sigma factor RpoS for dissemination within the vector during tick-to-mammal transmission." PLo*S Pathog no*. 8 (2):e1002532. doi: 10.1371/journal.ppat.1002532.

Dykhuizen, D. E., D. Brisson, S. Sandigursky, G. P. Wormser, J. Nowakowski, R. B. Nadelman, and I. Schwartz. 2008. "The propensity of different Borrelia burgdorferi sensu stricto genotypes to cause disseminated infections in humans." Am *J Trop Med Hyg no*. 78 (5):806-10. doi: 78/5/806 [pii].

Díaz, M., E. Ferreras, R. Moreno, A. Yepes, J. Berenguer, and R. Santamaría. 2008. "High-level overproduction of Thermus enzymes in Streptomyces lividans." App*l Microbiol Biotechnol no*. 79 (6):1001-8. doi: 10.1007/s00253-008-1495-1.

Eddy, S. R. 2005. "A model of the statistical power of comparative genome sequence analysis." PLo*S Biol no*. 3 (1):e10. doi: 10.1371/journal.pbio.0030010.

Edgar, R. C. 2004. "MUSCLE: multiple sequence alignment with high accuracy and high throughput." Nuc*leic Acids Res no*. 32 (5):1792-7. doi: 10.1093/nar/gkh340.

Ellegren, H. 2008. "Comparative genomics and the study of evolution by natural selection." Mol *Ecol no*. 17 (21):4586-96. doi: MEC3954 [pii]

10.1111/j.1365-294X.2008.03954.x.

Enright, A. J., S. Van Dongen, and C. A. Ouzounis. 2002. "An efficient algorithm for large-scale detection of protein families." Nucl*eic Acids Res no.* 30 (7):1575-84.

Eswar, N., B. Webb, M. A. Marti-Renom, M. S. Madhusudhan, D. Eramian, M. Y. Shen, U. Pieper, and A. Sali. 2007. "Comparative protein structure modeling using MODELLER." Curr *Protoc Protein Sci no.* Chapter 2:Unit 2.9. doi: 10.1002/0471140864.ps0209s50.

Eyre-Walker, A. 2006. "The genomic rate of adaptive evolution." Tren*ds Ecol Evol no.* 21 (10):569-75. doi: S0169-5347(06)00203-5 [pii]

10.1016/j.tree.2006.06.015.

Farnham, P. J. 2009. "Insights from genomic profiling of transcription factors." Nat R*ev Genet no.* 10 (9):605-16. doi: nrg2636 [pii]

10.1038/nrg2636.

Fingerle, V., G. Goettner, L. Gern, B. Wilske, and U. Schulte-Spechtel. 2007. "Complementation of a Borrelia afzelii OspC mutant highlights the crucial role of OspC for dissemination of Borrelia afzelii in Ixodes ricinus." Int J *Med Microbiol no. 2*97 (2):97-107. doi: 10.1016/j.ijmm.2006.11.003.

Fingerle, V., S. Rauser, B. Hammer, O. Kahl, C. Heimerl, U. Schulte-Spechtel, L. Gern, and B. Wilske. 2002. "Dynamics of dissemination and outer surface protein expression of different European Borrelia burgdorferi sensu lato strains in artificially infected Ixodes ricinus nymphs." J Clin *Microbiol no. 4*0 (4):1456-63.

Fraser, C. M., S. Casjens, W. M. Huang, G. G. Sutton, R. Clayton, R. Lathigra, O. White, K. A. Ketchum, R. Dodson, E. K. Hickey, M. Gwinn, B. Dougherty, J. F. Tomb, R. D. Fleischmann, D. Richardson, J. Peterson, A. R. Kerlavage, J. Quackenbush, S. Salzberg, M. Hanson, R. van Vugt, N. Palmer, M. D. Adams, J. Gocayne, J. Weidman, T. Utterback, L. Watthey, L. McDonald, P. Artiach, C. Bowman, S. Garland, C. Fuji, M. D. Cotton, K. Horst, K. Roberts, B. Hatch, H. O. Smith, and J. C. Venter. 1997. "Genomic sequence of a Lyme disease spirochaete, Borrelia burgdorferi." Nature *no. 3*90 (6660):580-6. doi: 10.1038/37551.

Gilmore, R. D., R. R. Howison, V. L. Schmit, and J. A. Carroll. 2008. "Borrelia burgdorferi expression of the bba64, bba65, bba66, and bba73 genes in tissues during persistent infection in mice." Microb *Pathog no. 4*5 (5-6):355-60. doi: 10.1016/j.micpath.2008.08.006.

Glöckner, G., R. Lehmann, A. Romualdi, S. Pradella, U. Schulte-Spechtel, M. Schilhabel, B. Wilske, J. Sühnel, and M. Platzer. 2004. "Comparative analysis of the Borrelia garinii genome." Nucle*ic Acids Res no. 3*2 (20):6038-46. doi: 32/20/6038 [pii]

10.1093/nar/gkh953.

Glöckner, G., U. Schulte-Spechtel, M. Schilhabel, M. Felder, J. Sühnel, B. Wilske, and M. Platzer. 2006. "Comparative genome analysis: selection pressure on the Borrelia vls cassettes is essential for infectivity." BMC Gen*omics no. 7:*211. doi: 1471-2164-7-211 [pii]

10.1186/1471-2164-7-211.

Grimm, D., K. Tilly, R. Byram, P. E. Stewart, J. G. Krum, D. M. Bueschel, T. G. Schwan, P. F. Policastro, A. F. Elias, and P. A. Rosa. 2004. "Outer-surface protein C of the Lyme disease spirochete: a protein induced in ticks for infection of mammals." Proc Nat*l Acad Sci U S A no. 101* (9):3142-7. doi: 10.1073/pnas.0306845101.

Gruber, A. R., S. Findeiß, S. Washietl, I. L. Hofacker, and P. F. Stadler. 2010. "RNAz 2.0: improved noncoding RNA detection." Pac Symp *Biocomput:69-79.*

Halperin, J. J. 2012. "Lyme disease: a multisystem infection that affects the nervous system." Continuu*m (Minneap Minn) no. 18* (6 Infectious Disease):1338-50. doi: 10.1212/01.CON.0000423850.24900.3a.

Hartiala, P., J. Hytönen, J. Suhonen, O. Leppäranta, H. Tuominen-Gustafsson, and M. K. Viljanen. 2008. "Borrelia burgdorferi inhibits human neutrophil functions." Microbes *Infect no. 10* (1):60-8. doi: 10.1016/j.micinf.2007.10.004.

Haven, J., L. C. Vargas, E. F. Mongodin, V. Xue, Y. Hernandez, P. Pagan, C. M. Fraser-Liggett, S. E. Schutzer, B. J. Luft, S. R. Casjens, and W. G. Qiu. 2011. "Pervasive recombination and sympatric genome diversification driven by frequency-dependent selection in Borrelia burgdorferi, the Lyme disease bacterium." Genetics *no. 189* (3):951-66. doi: 10.1534/genetics.111.130773.

He, M., T. Oman, H. Xu, J. Blevins, M. V. Norgard, and X. F. Yang. 2008. "Abrogation of ospAB constitutively activates the Rrp2-RpoN-RpoS pathway (sigmaN-sigmaS cascade) in Borrelia burgdorferi." Mol Micr*obiol no. 70* (6):1453-64. doi: 10.1111/j.1365-2958.2008.06491.x.

Hefty, P. S., S. E. Jolliff, M. J. Caimano, S. K. Wikel, J. D. Radolf, and D. R. Akins. 2001. "Regulation of OspE-related, OspF-related, and Elp lipoproteins of Borrelia burgdorferi strain 297 by mammalian host-specific signals." Infect I*mmun no. 69* (6):3618-27. doi: 10.1128/IAI.69.6.3618-3627.2001.

Honig, B., and A. Nicholls. 1995. "Classical electrostatics in biology and chemistry." Science *no. 268* (5214):1144-9.

Kenedy, M. R., and D. R. Akins. 2011. "The OspE-related proteins inhibit complement deposition and enhance serum resistance of Borrelia burgdorferi, the lyme disease spirochete." Infect I*mmun no. 79* (4):1451-7. doi: 10.1128/IAI.01274-10.

Kenedy, M. R., S. R. Vuppala, C. Siegel, P. Kraiczy, and D. R. Akins. 2009. "CspA-mediated binding of human factor H inhibits complement deposition and confers serum resistance in Borrelia burgdorferi." Infect I*mmun no. 77* (7):2773-82. doi: 10.1128/IAI.00318-09.

Kingsford, C. L., K. Ayanbule, and S. L. Salzberg. 2007. "Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake." Genome B*iol no. 8 (*2):R22. doi: 10.1186/gb-2007-8-2-r22.

Kumar, M., S. Kaur, T. Kariu, X. Yang, I. Bossis, J. F. Anderson, and U. Pal. 2011. "Borrelia burgdorferi BBA52 is a potential target for transmission blocking Lyme disease vaccine." Vaccine *no. 29* (48):9012-9. doi: 10.1016/j.vaccine.2011.09.035.

Lagal, V., D. Portnoï, G. Faure, D. Postic, and G. Baranton. 2006. "Borrelia burgdorferi sensu stricto invasiveness is correlated with OspC-plasminogen affinity." Microbes *Infect no. 8 (*3):645-52. doi: 10.1016/j.micinf.2005.08.017.

Larkin, M. A., G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson, and D. G. Higgins. 2007. "Clustal W and Clustal X version 2.0." Bioinfor*matics no. 23* (21):2947-8. doi: 10.1093/bioinformatics/btm404.

LaRocca, T. J., D. J. Holthausen, C. Hsieh, C. Renken, C. A. Mannella, and J. L. Benach. 2009. "The bactericidal effect of a complement-independent antibody is osmolytic and specific to Borrelia." Proc Nat*l Acad Sci U S A no. 106* (26):10752-7. doi: 10.1073/pnas.0901858106.

Laurie, A. T., and R. M. Jackson. 2005. "Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites." Bioinfor*matics no. 21* (9):1908-16. doi: 10.1093/bioinformatics/bti315.

Lefébure, T., and M. J. Stanhope. 2007. "Evolution of the core and pan-genome of Streptococcus: positive selection, recombination, and genome composition." Genome B*iol no. 8 (*5):R71. doi: 10.1186/gb-2007-8-5-r71.

Liang, F. T., M. J. Caimano, J. D. Radolf, and E. Fikrig. 2004. "Borrelia burgdorferi outer surface protein (osp) B expression independent of ospA." Microb P*athog no. 37* (1):35-40. doi: 10.1016/j.micpath.2004.02.007.

Liang, F. T., M. B. Jacobs, L. C. Bowers, and M. T. Philipp. 2002. "An immune evasion mechanism for spirochetal persistence in Lyme borreliosis." J Exp Me*d no. 195* (4):415-22.

Liang, F. T., F. K. Nelson, and E. Fikrig. 2002. "Molecular adaptation of Borrelia burgdorferi in the murine host." J Exp Me*d no. 196* (2):275-80.

Liang, F. T., J. Yan, M. L. Mbow, S. L. Sviat, R. D. Gilmore, M. Mamula, and E. Fikrig. 2004. "Borrelia burgdorferi changes its surface antigenic expression in response to host immune responses." Infect I*mmun no. 72* (10):5759-67. doi: 10.1128/IAI.72.10.5759-5767.2004.

Lin, T., L. Gao, D. G. Edmondson, M. B. Jacobs, M. T. Philipp, and S. J. Norris. 2009. "Central role of the Holliday junction helicase RuvAB in vlsE recombination and infectivity of Borrelia burgdorferi." PLoS Pat*hog no. 5 (*12):e1000679. doi: 10.1371/journal.ppat.1000679.

Luethy, and et al. 1992. Assessme**nt of protein models with three-dimensional profiles.**

Luikart, G., P. R. England, D. Tallmon, S. Jordan, and P. Taberlet. 2003. "The power and promise of population genomics: from genotyping to genome typing." Nat Rev *Genet no. 4 (*12):981-94. doi: nrg1226 [pii]

10.1038/nrg1226.

Margos, G., S. A. Vollmer, N. H. Ogden, and D. Fish. 2011. "Population genetics, taxonomy, phylogeny and evolution of Borrelia burgdorferi sensu lato." Infect Ge*net Evol no. 11 (*7):1545-63. doi: 10.1016/j.meegid.2011.07.022.

Mongodin, Emmanuel, Casjens. Sherwood, John Bruno, Xu Yun, Drabek Elliott Franco, Riley David, Cantarel Brandi, Pagan Pedro, Hernandez Yozen, Vargas Levy, Dunn John, Schutzer Steven, M. Fraser Claire, Qiu Wei-Gang, and Luft Benjamin. 2013. Inter- an**d Intra-Specific Pan-Genomes of Borrelia *burgdorferi sensu lato: Genome* Stability and Adaptive Radiation. BMC Gen**omics, In revision.

Mora, M., C. Donati, D. Medini, A. Covacci, and R. Rappuoli. 2006. "Microbial genomes and vaccine design: refinements to the classical reverse vaccinology approach." Curr Opin *Microbiol no. 9 (*5):532-6. doi: S1369-5274(06)00122-6 [pii]

10.1016/j.mib.2006.07.003.

Naville, M., A. Ghuillot-Gaudeffroy, A. Marchais, and D. Gautheret. 2011. "ARNold: a web tool for the prediction of Rho-independent transcription terminators." RNA Biol n*o. 8 (1)*:11-3.

Neelakanta, G., X. Li, U. Pal, X. Liu, D. S. Beck, K. DePonte, D. Fish, F. S. Kantor, and E. Fikrig. 2007. "Outer surface protein B is critical for Borrelia burgdorferi adherence and survival within Ixodes ticks." PLoS Patho*g no. 3 (3)*:e33. doi: 10.1371/journal.ppat.0030033.

Nowalk, A. J., R. D. Gilmore, and J. A. Carroll. 2006. "Serologic proteome analysis of Borrelia burgdorferi membrane-associated proteins." Infect Imm*un no. 74 (7)*:3864-73. doi: 10.1128/IAI.00189-06.

Ojaimi, C., C. Brooks, S. Casjens, P. Rosa, A. Elias, A. Barbour, A. Jasinskas, J. Benach, L. Katona, J. Radolf, M. Caimano, J. Skare, K. Swingle, D. Akins, and I. Schwartz. 2003. "Profiling of temperature-induced changes in Borrelia burgdorferi gene expression by using whole genome arrays." Infect Imm*un no. 71 (4*):1689-705.

Ojaimi, C., V. Mulay, D. Liveris, R. Iyer, and I. Schwartz. 2005. "Comparative transcriptional profiling of Borrelia burgdorferi clinical isolates differing in capacities for hematogenous dissemination." Infect Imm*un no. 73 (1*0):6791-802. doi: 10.1128/IAI.73.10.6791-6802.2005.

Ouyang, Z., R. K. Deka, and M. V. Norgard. 2011. "BosR (BB0647) controls the RpoN-RpoS regulatory pathway and virulence expression in Borrelia burgdorferi by a novel DNA-binding mechanism." PLoS Patho*g no. 7 (2)*:e1001272. doi: 10.1371/journal.ppat.1001272.

Ouyang, Z., S. Narasimhan, G. Neelakanta, M. Kumar, U. Pal, E. Fikrig, and M. V. Norgard. 2012. "Activation of the RpoN-RpoS regulatory pathway during the enzootic life cycle of Borrelia burgdorferi." BMC Microb*iol no. 12:44*. doi: 10.1186/1471-2180-12-44.

Pal, U., A. M. de Silva, R. R. Montgomery, D. Fish, J. Anguita, J. F. Anderson, Y. Lobet, and E. Fikrig. 2000. "Attachment of Borrelia burgdorferi within Ixodes scapularis mediated by outer surface protein A." J Clin Inv*est no. 106 (*4):561-9. doi: 10.1172/JCI9427.

Pal, U., X. Li, T. Wang, R. R. Montgomery, N. Ramamoorthi, A. M. Desilva, F. Bao, X. Yang, M. Pypaert, D. Pradhan, F. S. Kantor, S. Telford, J. F. Anderson, and E. Fikrig. 2004. "TROSPA, an Ixodes scapularis receptor for Borrelia burgdorferi." Cell no. 1*19 (*4):457-68. doi: 10.1016/j.cell.2004.10.027.

Patton, T. G., K. S. Brandt, C. Nolder, D. R. Clifton, J. A. Carroll, and R. D. Gilmore. 2013. "Borrelia burgdorferi bba66 gene inactivation results in attenuated mouse infection by tick transmission." Infect Imm*un. doi: 10.*1128/IAI.00140-13.

Pawley, N. H., S. Koide, and L. K. Nicholson. 2002. "Backbone dynamics and thermodynamics of Borrelia outer surface protein A." J Mol Biol *no. 324 (*5):991-1002.

Pettersen, E. F., T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin. 2004. "UCSF Chimera--a visualization system for exploratory research and analysis." J Comput C*hem no. 25 (1*3):1605-12. doi: 10.1002/jcc.20084.

Polgár, L. 2005. "The catalytic triad of serine peptidases." Cell Mol L*ife Sci no. 62 (1*9-20):2161-72. doi: 10.1007/s00018-005-5160-x.

Poljak, A., P. Comstedt, M. Hanner, W. Schüler, A. Meinke, B. Wizel, and U. Lundberg. 2012. "Identification and characterization of Borrelia antigens as potential vaccine candidates against Lyme borreliosis." Vaccine no. *30 (2*9):4398-406. doi: 10.1016/j.vaccine.2011.10.073.

Qiu, W. G., S. E. Schutzer, J. F. Bruno, O. Attie, Y. Xu, J. J. Dunn, C. M. Fraser, S. R. Casjens, and B. J. Luft. 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 no. 101 (*39):14150-5. doi: 0402745101 [pii]

10.1073/pnas.0402745101.

Radolf, J. D., M. J. Caimano, B. Stevenson, and L. T. Hu. 2012. "Of ticks, mice and men: understanding the dual-host lifestyle of Lyme disease spirochaetes." Nat Rev Mic*robiol no. 10 (2)*:87-99. doi: 10.1038/nrmicro2714.

Ramamoorthi, N., S. Narasimhan, U. Pal, F. Bao, X. F. Yang, D. Fish, J. Anguita, M. V. Norgard, F. S. Kantor, J. F. Anderson, R. A. Koski, and E. Fikrig. 2005. "The Lyme disease agent exploits a tick protein to infect the mammalian host." Nature no. *436 (7*050):573-7. doi: 10.1038/nature03812.

Rangannan, V., and M. Bansal. 2007. "Identification and annotation of promoter regions in microbial genome sequences on the basis of DNA stability." J Biosci no. *32 (5)*:851-62.

Rangannan, V., and M. Bansal. 2010. "High-quality annotation of promoter regions for 913 bacterial genomes." Bioinformat*ics no. 26 (24*):3043-50. doi: 10.1093/bioinformatics/btq577.

Rogovskyy, A. S., and T. Bankhead. 2013. "Variable VlsE is critical for host reinfection by the Lyme disease spirochete." PLoS One no. *8 (4):*e61226. doi: 10.1371/journal.pone.0061226.

Samuels, D. S. 2011. "Gene regulation in Borrelia burgdorferi." Annu Rev Mi*crobiol no. 65:479-*99. doi: 10.1146/annurev.micro.112408.134040.

Schutzer, S. E., C. M. Fraser-Liggett, S. R. Casjens, W. G. Qiu, J. J. Dunn, E. F. Mongodin, and B. J. Luft. 2011a. "Whole-genome sequences of thirteen isolates of Borrelia burgdorferi." J Bacteriol *no. 193 (4*):1018-20. doi: 10.1128/JB.01158-10.

Schutzer, S. E., C. M. Fraser-Liggett, S. R. Casjens, W. G. Qiu, J. J. Dunn, E. F. Mongodin, and B. J. Luft. 2011b. "Whole-genome sequences of thirteen isolates of Borrelia burgdorferi." J Bacteriol *no. 193 (4*):1018-20. doi: JB.01158-10 [pii]

10.1128/JB.01158-10.

Schutzer, S. E., C. M. Fraser-Liggett, W. G. Qiu, P. Kraiczy, E. F. Mongodin, J. J. Dunn, B. J. Luft, and S. R. Casjens. 2012. "Whole-genome sequences of Borrelia bissettii, Borrelia valaisiana, and Borrelia spielmanii." J Bacteriol *no. 194 (2)*:545-6. doi: 10.1128/JB.06263-11.

Shi, Y., Q. Xu, K. McShan, and F. T. Liang. 2008. "Both decorin-binding proteins A and B are critical for the overall virulence of Borrelia burgdorferi." Infect Immun *no. 76 (3):*1239-46. doi: 10.1128/IAI.00897-07.

Smith, A. H., J. S. Blevins, G. N. Bachlani, X. F. Yang, and M. V. Norgard. 2007. "Evidence that RpoS (sigmaS) in Borrelia burgdorferi is controlled directly by RpoN (sigma54/sigmaN)." J Bacteriol *no. 189 (5)*:2139-44. doi: 10.1128/JB.01653-06.

Stanek, G., G. P. Wormser, J. Gray, and F. Strle. 2012. "Lyme borreliosis." Lancet no. 3*79 (98*14):461-73. doi: 10.1016/S0140-6736(11)60103-7.

Steere, A. C., and L. Glickstein. 2004. "Elucidation of Lyme arthritis." Nat Rev Immu*nol no. 4 (2):1*43-52. doi: 10.1038/nri1267.

Storz, J. F. 2005. "Using genome scans of DNA polymorphism to infer adaptive population divergence." Mol Ecol no. *14 (3):*671-88. doi: MEC2437 [pii]

10.1111/j.1365-294X.2005.02437.x.

Strle, K., K. L. Jones, E. E. Drouin, X. Li, and A. C. Steere. 2011. "Borrelia burgdorferi RST1 (OspC type A) genotype is associated with greater inflammation and more severe Lyme disease." Am J Pathol no*. 178 (6):*2726-39. doi: 10.1016/j.ajpath.2011.02.018.

van Belkum, A., M. Struelens, A. de Visser, H. Verbrugh, and M. Tibayrenc. 2001. "Role of genomic typing in taxonomy, evolutionary genetics, and microbial epidemiology." Clin Microbio*l Rev no. 14 (3):5*47-60. doi: 10.1128/CMR.14.3.547-560.2001.

Vollmer, S. A., E. J. Feil, C. Y. Chu, S. L. Raper, W. C. Cao, K. Kurtenbach, and G. Margos. 2013. "Spatial spread and demographic expansion of Lyme borreliosis spirochaetes in Eurasia." Infect Genet *Evol no. 14:147-5*5. doi: 10.1016/j.meegid.2012.11.014.

Wang, I. N., D. E. Dykhuizen, W. Qiu, J. J. Dunn, E. M. Bosler, and B. J. Luft. 1999. "Genetic diversity of ospC in a local population of Borrelia burgdorferi sensu stricto." Genetics no. *151 (1):*15-30.

Weening, E. H., N. Parveen, J. P. Trzeciakowski, J. M. Leong, M. Höök, and J. T. Skare. 2008. "Borrelia burgdorferi lacking DbpBA exhibits an early survival defect during experimental infection." Infect Immun *no. 76 (12):*5694-705. doi: 10.1128/IAI.00690-08.

Wywial, E., J. Haven, S. R. Casjens, Y. A. Hernandez, S. Singh, E. F. Mongodin, C. M. Fraser-Liggett, B. J. Luft, S. E. Schutzer, and W. G. Qiu. 2009. "Fast, adaptive evolution at a bacterial host-resistance locus: the PFam54 gene array in Borrelia burgdorferi." Gene no. 445 *(1-2)*:26-37. doi: S0378-1119(09)00325-4 [pii]

10.1016/j.gene.2009.05.017.

Xu, H., M. He, J. J. He, and X. F. Yang. 2010. "Role of the surface lipoprotein BBA07 in the enzootic cycle of Borrelia burgdorferi." Infect Immun no*. 78 (7):2*910-8. doi: 10.1128/IAI.00372-10.

Xu, Q., K. McShan, and F. T. Liang. 2007. "Identification of an ospC operator critical for immune evasion of Borrelia burgdorferi." Mol Microbiol *no. 64 (1):22*0-31. doi: 10.1111/j.1365-2958.2007.05636.x.

Xu, Q., K. McShan, and F. T. Liang. 2008. "Essential protective role attributed to the surface lipoproteins of Borrelia burgdorferi against innate defences." Mol Microbiol *no. 69 (1):15*-29. doi: 10.1111/j.1365-2958.2008.06264.x.

Xu, Q., K. McShan, and F. T. Liang. 2010. "Two regulatory elements required for enhancing ospA expression in Borrelia burgdorferi grown in vitro but repressing its expression during mammalian infection." Microbiology no*. 156 (Pt 7*):2194-204. doi: 10.1099/mic.0.036608-0.

Xu, Q., Y. Shi, P. Dadhwal, and F. T. Liang. 2012. "RpoS regulates essential virulence factors remaining to be identified in Borrelia burgdorferi." PLoS One no. 7 *(12):e5*3212. doi: 10.1371/journal.pone.0053212.

Yang, X., J. Qin, K. Promnares, T. Kariu, J. F. Anderson, and U. Pal. 2013. "Novel microbial virulence factor triggers murine lyme arthritis." J Infect Dis no*. 207 (6):9*07-18. doi: 10.1093/infdis/jis930.

Yang, Z. 2007. "PAML 4: phylogenetic analysis by maximum likelihood." Mol Biol Evol *no. 24 (8):15*86-91. doi: msm088 [pii]

10.1093/molbev/msm088.

Zhang, Y. 2008. "I-TASSER server for protein 3D structure prediction." BMC Bioinformat*ics no. 9:40. doi:* 1471-2105-9-40 [pii]

10.1186/1471-2105-9-40.