Molecular and Evolutionary Properties of Non-Gene-Coding Regions in Bacteria Using Comparative Genome Bioinformatics

by

Tika Y. Sukarna

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

2010

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.

December 22, 2009	Dr. Weigang Qiu, Hunter College
Date	Chair of Examining Committee
January 28, 2010	Dr. Laurel A. Eckhardt
Date	Executive Officer

Dr. Michael E. Steiper, Hunter College

Dr. Benjamin Ortiz, Hunter College

Dr. Stephane Boissinot, Queens College

Dr. Dustin Brisson, University of Pennsylvania

Dr. John J. Dunn, Brookhaven National Laboratory (Reader)

Supervising Committee

THE CITY UNIVERSITY OF NEW YORK

Abstract

Molecular and Evolutionary Properties of Non-Gene-Coding Regions in Bacteria Using Comparative Genome Bioinformatics

by

Tika Y. Sukarna

Adviser: Professor Weigang Qiu

Although most non-gene-coding regions of the genome have long been thought of as nonfunctional, a growing body of literature now show that many sites act as important sources of phenotypic variation and complexity. In eukaryotes, this has been attributed to the sophisticated gene regulatory apparatus that includes cis-acting regulatory elements acting on multiple levels. In bacteria, this level of regulatory multiplicity is reduced, as is reflected by the lower percentage of intergenic segments in their genomes and the lower capacity for metabolic and catabolic activities. Most non-gene-coding intergenic portion of the genome of bacteria is thought of as functionally compact, mostly transcriptional and translational regulatory in nature, containing only limited number of infrastructural or regulatory RNAs. This study addresses the extent of this cis-regulatory organization on noncoding genomic regions in the Lyme bacteria *Borrelia burgdorferi* as apparent in their molecular and evolutionary properties and how they can influence and be applied to the bioinformatic predictions of cis-regulatory function. Several general molecular and evolutionary properties of a bacterial non-coding genome were identified. Overall, most non-gene-coding intergenic portion of *Borrelia* are constrained, functionally compact and degenerate. A phylogenetic footprinting approach for very closely related species (> 90% nucleotide sequence identity) was developed to test for specific sites of transcriptional

regulation, which was additionally tested using the *Escherichia coli* genome dataset. The method finds most constrained regions to coincide with several general properties of promoter binding, suggesting that constraint levels are differentiable even amongst these very closely related bacterial species, providing a way to measure for molecular function at a fine phylogenetic level through the understanding of the patterns of DNA sequence evolution.

Acknowledgments

This dissertation is dedicated to mamah and bapa who always believed in me. Many thanks to all family members for being there. My thanks to Dr. Weigang Qiu for his patient support, also to committee/examining members Dr. Benjamin Ortiz, Dr. Michael Steiper, Dr. Stephane Boissinot, Dr. Dustin Brisson and Dr. John J. Dunn without whom this work may not be possible. I thank also the Department of Biology at Hunter College and The Graduate Center of the City University of New York for giving me the opportunity to study and conduct my research these many years under their support. Thanks also to Qiu Evolutionary Bioinformatics Lab members past and present for their cheerful presence even if most of the time it was by remote access. Finally, to Roby and Toby for always being around.

Table of Contents

List of Tables

List of Figures

1. Introduction

Most non-gene-coding regions of the genome have in general been thought of as nonfunctional, but evidence now show that many sites act as important sources of genetic variation. King and Wilson (1975) argued that the low level of divergence (~1%) between human and chimpanzee protein sequences cannot account for such profound phenotypic differences observed between these species. They concluded that some differences must also come from changes in genomic regions controlling gene expression. Since then, studies show several classes of highly conserved non-coding DNA to be regulatory in nature in several eukaryotic species (Andolfatto 2005; Bejerano et. al. 2004; Boffelli, Nobrega and Rubin 2004; Dermitzakis and Clark 2002; Gaffney and Keightley 2006; McEwen et. al. 2006; Sandelin et. al. 2004; Woolfe et. al. 2005; Vavouri et. al. 2007). . These sites include short degenerate regulatory elements acting on multiple levels, from structural, to transcriptional, post-transcriptional and translational regulation (Taft et al. 2007).

In bacteria, this level of regulatory multiplicity is reduced, as is reflected by the lower percentage of non-gene-coding intergenic segments in their genomes and the lower capacity for metabolic and catabolic activities, particularly those relying on host functions. Most of the intergenic portion of the genome of bacteria is thought of as functionally compact, mostly transcriptional and translational regulatory in nature, containing only limited number of infrastructural or regulatory RNAs involved in post-transcriptional regulation (Gottesman 2005). Following the publication of their seminal paper on the *lac* operon of *E. coli,* Jacob and Monod (1961) described the possible role of mutations in operators during evolution. They argued that the function of every gene depends not only on what its product does but also on the

circumstances under which they are produced as controlled by non-gene-coding regulatory elements such as the operon, emphasizing the importance of non-coding DNA evolution on gene function.

This study addresses the functional organization of noncoding intergenic genomic regions in bacteria as maybe apparent in the molecular evolution of the genomes of Lyme bacteria *Borrelia burgdorferi* and how they can influence and be applied to the bioinformatic predictions of regulatory functional sites, namely for this study, probable promoter/transcription factor binding sites. A comparative genomics approach using multiple genomes of closely related *Borrelia* species (> 90% nucleotide sequence identity) is used to identify patterns of evolutionary constraint on intergenic regions. Such constraint predict functional role and was used subsequently to develop a phylogenetic footprinting method that attempts to correlate fine levels of nucleotide substitution rates in intergenic regions to possible promoter/transcription factor binding sites. A collection of three *Escherichia coli* strains was used to test the applicability of the method in a separate genome dataset. Thus, the aim of this study was to develop a comparative genomics bioinformatic strategy using available genome data to uncover possible functional regulatory sites in noncoding intergenic regions, in addition to determining the fundamental properties of the evolution of functional noncoding regions at a fine scale.

i. Experimental Methods to Identify Transcription Factor Binding Sites

In experimental methods to identify transcription factor binding sites, the core idea is to isolate DNA and DNA binding protein complexes as related to the biological process of interest, after which the task is to determine the bound DNA sequence in relation to its genomic site.

Initial methods to isolate bound DNA-protein complexes include footprinting methods, gel-shift assays and southwestern blotting (Bulyk 2003; Bowen et. al. 1980; Galas and Schmitz 1978; Garner and Revzin 1981). A number of high-throughput genomic level technologies have been developed to isolate DNA sites as they are bound by their cognate transcription factors. Genomic SELEX uses a genomic library to select for double stranded DNA fragments that bind with high affinity to a protein of interest *in vitro* (Shimida et. al. 2005; Shtatland et. al. 2000; Kim et. al. 2003). Another method takes advantage of DNA microarrays to decipher bound DNA sites by using transcription factors tethered to DNA adenine methyl transferase (DAM), whereby bound sites will exhibit DNA methylation signal(s) (Van Steensel et. al. 2001). Chromatin immunoprecipitation 'ChIP' is currently the most common method to identify genomic TFBSs *in vivo* (Lieb et. al. 2001; Ren et. al. 2000; Iyer et. al. 2001; Buck et. al. 2004; Farnham, P. J. 2009). Here, DNA-binding proteins are initially crosslinked to their target DNA by treating cells with formaldehyde. Crosslinked sites are immunoprecipitated with an antibody that is specific to the regulatory protein of interest. Bound DNA-protein are further analyzed, usually through PCR, microarray (ChIP-chip) analysis or various sequencing technologies, the most current of which is the 'ChIP-seq' which takes advantage of high-throughput-next-generation sequencing technology to analyze the DNA sequence bound (Farnham, P. J. 2009).

High-throughput methods such as the ChIP assays require a large number of cells as they represent the average level of binding of a factor at a site in the cell population. Unlike ChIP-chip assays, ChIP-seq allows the scanning of overlapping genomic fragments usually 27-50 nucleotides in length, and hence is regarded to provide better resolution mapping of transcription factor bound sites than their ChIP-chip predecessor (Farnham, P. J. 2009). As the genomic

sequence space for such potential sites are vast, bioinformatic methods that allow for a prior prediction of a subset of sites with high probability of true hits becomes essential.

ii. Computational Methods To Identify Transcription Factor Binding Sites

A number of computational approaches have been developed to predict for DNA regulatory binding sites. While still exhibiting a high false-positive rate, they provide a means to determine likely candidates for experimental work, increasing the likelihood of positive hits. For example, in detecting a possible regulator candidate for an arbitrary gene in *E. coli*, without any previous information, the candidates would be ~4500 genes, the total number of genes in the *E. coli* genome. Bioinformatic methods are able to reduce this to a manageable set of a dozen of potential regulators (Balleza et. al. 2009). These algorithms aim to deduce the regulatory elements by looking at regions of several (putatively) co-regulated genes from single or multiple genomes by searching for instances of overrepresented motifs (Blanchete and Tompa 2002).

Difficulties in the bioinformatic identification of regulatory sites from sequence data result from their short (6-30bp) and degenerate property. Such short sequence patterns can appear thousands of times in a genome through random chance alone (Hahn 2007). DNA regulatory binding sites maybe represented as a position weight matrix (PWM) or a consensus sequence to address this degeneracy (Stormo 2000; Mount 2001; Jones and Pevzner 2004). These representations are used to identify conservation from genome sequence comparisons where conventional alignment algorithms to identify conserved regions is not sufficient (Lawrence et al. 1993; Neuwald, Liu, and Lawrence 1995; Hughes et al. 2000; Mount 2001; Thompson, Rouchka, and Lawrence 2003; Jones and Pevzner 2004; Pavesi, Mauri, and Pesole

2004). PWM is essentially a probability scoring assignment for a list of known regulatory sites compared to unknown sites. PWM method is dependent on having already a list of known regulatory binding sites. By using this known list, a PWM is used to scan through unknown regions for a match. A match is when the scan retrieves a PWM probability score that is statistically significant. PWM is limited to known regulatory sites and does not look for new types of regulatory sites that have not been previously identified.

Phylogenetic footprinting (Gumucio, et. al. 1992; Duret and Bucher 1997) relies on the detection of high degrees of conservation across different species to finding functionally important sequences in the genome without the need for a list of "known sites". Function is determined based on a measure of evolutionary conservation amongst orthologous sites (Dermitzakis and Clark 2003; Zhang and Gerstein 2003). Figure 1 shows an example of an alignment whereby constrained orthologous regions coincide with known regulatory functional sites. A statistical model of the rates of evolution between conserved and non-conserved sequences maybe used to score a match (Blanchette and Tompa 2002). Phylogenetic type methods require knowledge of the evolutionary properties of the regions being investigated for the presence of functional regulatory sites. It is assumed that sites that are functional will influence the fitness of the organism. This fitness is quantifiable using statistical tests that are based on biological evolutionary properties of genomic regions being compared. These properties are based on the idea that only functional regions will influence the fitness of the organism, while non-functional sites will evolve essentially randomly. This study addresses the development of a phylogenetic footprinting method specifically designed to detect for subtle changes within bacterial intergenic sequence of > 90% in nucleotide sequence identity.

iii. Molecular Evolutionary Statistics

The power of comparative genomics to predict functional sites in noncoding genomic regions lies on the assumption that genome sequence sites maybe differentiated from one another based on how they affect the evolutionary fitness of the whole organism. When genome sequences are compared, sites may exhibit three types of evolutionary properties:

1. Neutrally Evolving Sites

Sites that confer no evolutionary fitness to the organism are considered as "neutral" or "neutrally evolving". Rates of nucleotide substitutions in these sites will be similar to other sites that are non-functional, aside from differences in substitution rate.

2. Constrained Sites (Sites Undergoing Negative or Purifying Selection)

Sites that are functional may exhibit less sequence variety compared to neutrally evolving sites. Nucleotide substitutions occurring here will usually be removed, or negatively selected for, as they are deleterious to the fitness of the organism. Rates of nucleotide substitutions at these sites are lower or more constrained than neutral sites. When orthologous sequences are compared, such sites will exhibit a high degree of conservation.

3. Adaptive Sites (Sites Undergoing Positive Selection)

Functional sites may exhibit greater sequence variety when compared to neutral sites as nucleotide substitutions occurring here confers a fitness advantage to the organism, and are positively selected for. Rates of nucleotide substitutions at these sites are higher than neutral rates, and thus are positively selected for. When orthologous sequences are compared, such sites will exhibit a high degree of sequence variation.

Using this categorization, sites maybe differentiated with one another by the rates of

nucleotide substitutions within a region when compared to neutrally evolving sites. This becomes the basis for the statistical evolutionary measurements used to identify functional regulatory sites through comparative genomics. This idea is based on the neutral theory of molecular evolution, as developed by Kimura (1968) and King and Jukes (1969), which has been used as a null model to detect different forms of selection (Wayne and Simonsen 1998; Nielsen 2001). By finding genomic regions in which selection has been acting, one can verify or identify regions in genomes that confer function. When sequences are compared, functional sites are those that show a level of sequence variation that significantly differs from neutrally evolving sites. Those that show higher sequence variation are deemed as adaptive or positively selected for, while those that show a lower level of sequence variation are thought of as constrained or negatively selected for.

In protein coding regions, the level of selection acting on a DNA sequence site maybe measured by comparing the rate of substitutions in non-synonymous sites (dN) to synonymous sites (dS), where synonymous sites are taken as the background neutral substitution rate (Kimura 1977; Nei and Gojobori 1986; Hendrick 2005). Hence, for neutral sites, $dN = dS$ (or $dN/dS = 1$), but if there is constraint then $dN < dS$ (or $dN/dS < 1$), while sites which are adaptive will have $dN > dS$ (or $dN/dS > 1$). Similarly, substitution rates on a particular noncoding intergenic site (dI) maybe compared to synonymous sites (dS) to detect selection whereby sites with $dI > dS$ are neutral, while sites with $dI < dS$ are constrained and sites with $dI > dS$ are positively selected for (Hahn 2007).

Initially, phylogenetically based methods relied on strict conservation between the sequences compared as a way to differentiate functional versus nonfunctional sites, as the

phylogenetic distance of the species compared are relatively distant enough that the few sites that share the same evolutionary history are usually constrained. But as more genome sequences from species of varying phylogenetic distance becomes available, conservation alone to identify sites of function in noncoding regions becomes insufficient, as selection may act to both constrain or increase the level of variation within genomes. Depending on the evolutionary distance between the species compared, functional noncoding sites will exhibit varying levels of evolutionary selection. Especially for very closely related species, functional sites maybe obscured, as most sites will not have diverged far enough to be differentiated from neutrally evolving sites. Alternately, if species used are too divergent from one another, functional sites that are absent in some of the species compared will not be detected (Hahn 2007). Thus, the inclusion of sensitive molecular evolutionary statistical measures for the prediction of functional sites becomes necessary. As functional noncoding sites (i.e. transcriptional binding sites etc.) are known to be short and degenerate, understanding the fundamental nature of noncoding evolution becomes essential and forms the basis of this study.

iv. Evolution of Noncoding Regions in Eukaryotes Versus Prokaryotes

Around 6-14% of bacterial genomes are non-protein-coding compared to 95% in eukaryotic genomes (Shabalina et. al. 2001; Chamary et. al. 2006). Most intergenics within bacterial genomes are thought to be "functional", not “junk”, and evolves mostly under constraint. In mammals only ~15% of noncoding regions are “functional”, while bacterial genomes are considered to be functionally compact. The percentage of noncoding regions in the genomes of mammals do not coincide with their complexity, while for prokaryotes, the larger percentage of

noncodings in their genomes are directly related to their complexity (Taft et. al. 2007). Studies have shown that purifying selection is the main mode for the evolution of most noncoding regions in prokaryotes, while several instances of adaptive selection have been observed in the evolution of eukaryotes. This maybe related to the different ways genomes are organized in the prokaryotic versus the eukaryotic system.

Noncoding regions in eukaryotes have been observed to exhibit both negative and positive selection in sites assumed to have regulatory properties (Keightley and Gaffney 2003; Gaffney and Keightley 2006; Halligan and Keightley 2006; Andolfatto 2005). Comparisons of DNA in rodents estimate that the genomic deleterious point mutation rate for non-coding DNA is similar to that for coding DNA (Keightley and Gaffney 2003). Further work revealed rat-mouse non-coding DNA to have at least three times as many selectively constrained (negative selection), functional non-coding sites as coding sites (Gaffney and Keightley 2006). The *Drosophila* genome is inferred to have more than three times as much functional non-coding DNA as protein-coding DNA (Halligan and Keightley 2006). Interestingly, a study of population level *Drosophila* X Chromosome variability provided evidence for adaptive evolution in several classes of non-coding DNA (Andolfatto 2005).

The explosion of “Evo-Devo” research has highlighted the importance of non-coding DNA variation on the evolution of form (Carroll 2005). Most interestingly, recent studies have unraveled highly conserved non-coding elements (CNE's) that are often more conserved than gene coding sites associated with enhancer activity controlling genes involved in development (Bejerano et. al. 2004; Boffelli, Nobrega and Rubin 2004; Dermitzakis and Clark 2002; McEwen et. al. 2006; Sandelin et. al. 2004; Woolfe et. al. 2005; Vavouri et. al. 2007).

Compared to eukaryotes, prokaryotes have significantly reduced non-coding DNA in their genomes. Through comparative analysis of non-coding regions from 39 species of bacteria, (Rogozin et. al. 2002) found that non-coding DNA evolves mainly through evolutionary constraint to minimize the amount of non-functional DNA. Using a quantitative model comparing genomes of *E. coli* and a series of gamma proteobacteria (Rajewsky et. al. 2002) also found that most non-coding regions in this collection of prokaryotic species are mostly evolutionarily constrained. Similarly, evidence of constraint has also been found at functionally related noncoding positions in 22 clades of bacteria (Molina and Nimwegen 2008).

In this study, a collection very closely related species (> 90% sequence identity) of Lyme bacteria *Borrelia burgdorferi* genomes was used to explore such questions using a comparative genome approach. The *Borrelia burgdorferi sensu lato* (s. l.) complex is a spiral-shaped gram-negative bacterium of the Spirochaeta family. Its genome is composed of a 910kbp linear chromosome and variable numbers of plasmids (up to 21 in number and 610kbp in total) (Casjens et al. 1995; Fraser et al. 1997; Wang et al. 1999; Casjens et. al. 2000; Lin et al. 2003; Glockner et al. 2004; Stewart et al. 2005). It is the agent of one of the most prevalent bacterial arthropod borne disease in North America and Europe. Of the 11 species within the *Borrelia burgdorferi s. l.* complex, three are known to be associated with Lyme disease: *B. burgdorferi sensu stricto* (s. s.), *B. garinii*, and *B. afzelii*. *Borrelia burgdorferi* goes through a complex transmission cycle of infection, requiring both colonization of its arthropod vector and its mammalian host: (i) migration from tick midgut to salivary glands during blood meal, (ii) entry and colonization of mammalian tissues, (iii) establish chronic/systemic infection, and (iv) uptake and colonization of non-infected tick (Fisher et al. 2005). The availability of genomes from

various species of this complex will provide a valuable resource to understand the ability of *Borrelia burgdorferi* to regulate such complex activities, especially within its relatively small genome compared to other prokaryotes. With additional genomic sequences for *Borrelia burgdorferi* strains JD1, N40, 297, DN127 and PKo (Sherwood et. al. unpublished data), and the European genospecies *B. garinii* (strain PBi) (Glockner et. al. 2004), it is a rich collection of dense genome data to identify patterns of non-coding genome evolution at a fine scale.

For this study, levels of evolutionary selection acting on different classes of genomic segments of the *Borrelia burgdorferi* species collection were measured to further investigate the extent of selective pressure within a model prokaryotic lineage and how they correlate to function at a genome wide scale. Most noncoding intergenic sites in *Borrelia* were shown to exhibit functional constraint, as seem to be typical of most prokaryotic species. Using a collection of within species and between species genome sequence data, the extent of this constraint as species diverged was determined at within and between population level. Measures of selection were conducted by examining differences at constraint levels between fourfold synonymous (F), nonsynonymous (N) and intergenic (I) sites, which were then related to their location in the genome and possible role in regulatory function.

From this, a phylogenetic footprinting method was developed to calculate differences in substitution rates between intergenic fragments > 90% in sequence identity, 6bp in length. The method gives a measurement of deviations from fourfold synonymous rates. Intergenic fragments are appended to its flanking orthologous upstream and downstream fourfold ORF rates to determine fluctuations of specific intergenic sites from neutral rates. The method was tested on a separate *Escherichia coli* genome data set, focusing on intergenics spanning 227 ORFs in

three strains.

2. Materials and Methods

Documentation of computational analysis is further elaborated at the Appendix.

i. Computer System

All final analysis and data storage was performed on a Linux (Ubuntu) machine (Lenovo x60s IBM Thinkpad) with 1.66 GHz Intel Core Duo Processor, 71G hard disk drive with the following open source software packages installed:

1. PostgreSQL database system to process and store sequence data.

2. PERL programming language including the CGI, GD and Bioperl modules, for manipulation of sequence strings and image making (Stajich 2002).

3. Standalone command line NCBI BLAST ('blastall') to match local alignments of protein and/ or DNA sequences against local database (Altschul et. al. 1997).

4. ClustalW multiple sequence alignment program (Thompson et al. 1994)

5. R Statistical Package (R Development Core Team 2008)

6. PAML package for phylogenetic analysis (Yang 2007)

7. PHYLIP package for inferring phylogenies (Felsenstein 1989)

8. The Linux distribution of the PATSER software (Patser-v3b) to scan DNA sequences with position-specific-scoring-matrix (PSSM).

ii. Genome Source

a. *Borrelia burgdorferi sensu lato* Dataset

The *Borrelia burgdorferi sensu lato* complex genome collection is as tabulated in Table 1 with the following species (strains): *Borrelia burgdorferi sensu stricto* (strains B31, JD1, N40,

297), *Borrelia garinii* (strain PBi), *Borrelia bissetti* (strain DN127) and *Borrelia afzelii* (strain Pko). Genome sequence for PBi was obtained from the NCBI database, while genome sequences for JD1, N40, 297, DN127 and PKo were obtained through personal communications (Sherwood et. al. unpublished data). The statistical power of the genome collection was determined through the statistical model developed by Eddy (2005). Since the genome data used is under ongoing assembly, analysis was done on two types of data set, one for the draft assembly and the other for the final complete form, as will be elaborated in the results section. A Bayesian inferred phylogenetic tree of the *Borrelia* species (fig. 2) used was done by using a CLUSTALW (Thompson et al. 1994) multiple alignment of outer surface protein A (ospA) and the MrBayes program (Ronquist and Huelsenbeck 2003).

b. *Escherichia coli* Dataset

Total genome data from strains K12, APEC and O157 of *Escherichia coli*, which approximates the phylogenetic distance of the *Borrelia* dataset, were downloaded from NCBI genome sequence database (NCBI reference sequence: NC_010473, NC_008563, NC_002655), where 227 ORFS were used for subsequent analysis.

iii. Bioinformatic Pipeline

In summary, genomes are first divided into orthologous intergenic, nonsynonymous (first and second codon position) and fourfold-synonymous sites. They were then aligned and separated into within and between species groups. The substitution rate for each aligned sequence was calculated using the PAML software (Yang 2007) using the output "tree length".

The aligned genome sequences were further divided into convergent, unidirectional and divergent intergenics (fig. 3) to calculate rates at near ORF boundary sites in relation of probable 5' ORF cis-regulatory sites. Substitution rates at particular intergenic sites were determined using phylogenetic footprinting of intergenic regions in relation to known 5' ORF cis-functional sites.

The bioinformatic pipeline consist of:

- Dividing genome collection into ORF and Intergenic sequences.
- Identifying orthologous ORF and Intergenic sequences
- Aligning orthologous ORF and intergenic sequences then separating the alignments into within and between species grouping.
- Dividing ORF sequences into nonsynonymous and fourfold synonymous sites.
- Dividing Intergenics further into convergent and divergent intergenics, also 5' cis-ORF-upstream sites.
- Calculating rates of each site in comparison to fourfold synonymous sites to measure differences in constraint.

Figure 4 and Figure 5 summarizes the bioinformatic pipeline to calculate substitution rates following division of genomes into intergenic and ORF regions.

iv. Create Orthologous ORF Database

Data of non-overlapping ≥250-bp ORFs have been previously annotated for all genomes (Qiu et. al. 2004). The annotation process included the alignment of genome assemblies using

Nucmer of the Mummer 3.0 software package (Kurtz et. al. 2004), and the identification of non-overlapping ORFs for each genome using the GLIMMER 2.0 software (Kurtz et. al. 2004).

ORF nucleotide sequences were obtained using the start and stop annotation of the genome(s) ORFs for one strand and then reverse complemented using PERL scripts. All ORF nucleotide sequence(s) were converted into 6-frame amino acid sequences, which were then filtered to exclude internal stop codons to obtain one inframe amino acid sequence for each ORF. B31 was set as the reference genome for the one-to-one orthologous ORF search. Using the formatdb option of the Blastall package (Altschul et. al. 1997), a database of the B31 amino acid sequences were constructed and subsequently used to query a set of amino acid sequence from one other genome from the collection using the blastp option of the Blastall package, with e-value set at 0.001. Top scoring matches (first match in the list of blast result) were selected to get a one-to-one match list, which were stored into the PostgreSQL database.

v. Create Orthologous Intergenic Database

A PERL script extracts a list of intergenics from nucleotide start and stop data of ORF(s) for all genomes then stored into the PSQL database. Data from top scoring orthologs ORFs were used to identify orthologous intergenics then filtered through the *Borrelia* Genome Synteny browser (fig. 6) and PERL scripts. Only intergenics > 25bp located between orthologous ORFs are chosen through the synteny browser to be further analyzed (fig. 6). The *Borrelia* Genome Synteny browser was constructed using the PERL GD package and the CGI handle to access PSQL database for both ORF and Intergenic data set. The browser allows for the visualization of ORF and intergenic sites corresponding to its position in the main or plasmid cp26

chromosome. Each ORF ortholog is aligned vertically and color coded to signify match. ORFs that appear to be non-syntenic are removed from the data. Following this filter, the database is adjusted accordingly.

vi. Alignments

a. Intergenic Sequence Alignment

ClustalW was used to obtain multiple alignments of the orthologous genome regions of within and between *Borrelia* species. A modified DNA matrix (matrix_IUB_intergenic) where matches score 1.0 and mismatches score 0 were used. The matrix includes ambiguity symbols (except X/N) with addition of symbol Q as ORF-Intergenic boundary, where matches of overlapping Q scores = 50 to 100 (fig. 7). This high matching score aligns Q boundaries as fixed anchors. To improve the anchoring of intergenic CLUSTALW alignments, 300bp ORF sequences were added flanking both the 3' and 5' end of intergenic regions, which were removed during subsequent analysis.

b. Coding Sequence Alignment

For ORFs, each nucleotide sequence were converted into 6-frame amino acid sequences using PERL scripts, which were then filtered to exclude internal stop codons to obtain one inframe amino acid sequence. Amino acid sequences for each ORF using ClustalW were aligned then converted back into nucleotide alignments. Nucleotide alignments were separated into within and between species alignments for further analysis.

vii. Calculation of Substitution Rates

Figure 4 and Figure 5 shows the bioinformatic strategy used for the calculation of substitution rates for intergenic (fig. 4) and ORF (fig. 5) sequences. Analysis of within and between species alignments was done separately. The PAML software was used to calculate substitution rates (tree length) for each CLUSTALW multiple alignments. To determine the effect of the molecular evolution models, each intergenic data set was subjected to the JC69 (Jukes and Cantor 1969), HKY85 (Hasegawa et. al. 1985) and REV (Rodriguez et. al. 1990; Yang et. al. 1994) models of the baseml program and compared to synonymous and nonsynonymous rates determined through the codeml program with free dN/dS ratios for each branch of PAML (fig. 4, fig. 5 and Table 2). The three models were determined as it represents a range that incorporates all the parameters for all available models. Ultimately, the baseml F81 model (Felsenstein 1981) with no clock or free rates each branch was chosen, for calculations of substitution rates for all data set. For calculation of first-second codon position (Nonsynonymous) and fourfold (Fourfold Synonymous) codon position, first-second and fourfold position sub-alignment from total ORF alignment was extracted using PERL scripts. This sub-alignment was used in PAML baseml runs to determine Nonsynonymous rates (dN_w and dN_b) and Fourfold Synonymous rates (dF_w and dF_b) (fig. 5). The Phylip program (Felsenstein 1989) was used to generate trees from CLUSTALW alignments, as input for all PAML runs. Nonsynonymous (dN_{wc} and dN_{bc}) and synonymous (dS_{wc} and dS_{wc}) substitution rates were determined using the codon substitution model of the codeml program in PAML (fig. 5). Rates from baseml runs were used to calculate nonsynonymous rates (dN) for subsequent analysis.

viii. Proportion of Constrained Versus Non-Constrained Intergenics

Counts of intergenics where dI/dF < 1 (constrained sites) and dI/dF > 1 (non-constrained sites) were determined for main chromosome substitution rate data. Randomized sampling (sample with replacement) of fourfold rates (dF) was generated using R statistical package sample (R Development Core Team 2008). The number of random fourfold sample equals the number of intergenic sample to be tested for both constrained and non-constrained sites. Significance of difference of constrained intergenics compared to non-constrained intergenic counts is calculated using the Chi-Square test, where expected proportion is the number of counts from fourfold rates random samples.

ix. Measuring Changes in Constraint for Within and Between Species

The mean upstream and downstream flanking fourfold ORF sites substitution rates (dF_{mean}) were used for the background neutral rates for each intergenic sites, while nonsynonymous sites uses its corresponding ORF fourfold rate (dF). Constraint is measured as the ratio of substitution rates for nonsynonymous (dN) and intergenic (dI) sites over neutral rates where K = dN/dF for nonsynonymous or K = dI/dF_{mean} for intergenics. K was measured for both within and between populations for each intergenic and nonsynonymous sites, where K_{within} is substitution rate ratio for within species and $K_{between}$ is rate ratio for between species. Only the main chromosome was used in this analysis.

Only K < 1 was used for the analysis as indicator of constraint. At a particular site, changes in constraint level is constant within to between species if $K_{within} = K_{between}$. If more sites show $K_{within} > K_{between}$ then constraint is increasing between species compared to within species,

and if more sites show $K_{within} < K_{between}$ then constraint is decreasing between species compared to within species.

To measure the significance of change in K within to between species, dI and dN rates was substituted to randomized fourfold rates (sampling with replacement) using the R statistical package to get K expected, where for intergenics $K_{expected} = dF_{random}/dF_{mean}$, and for ORFs, $K_{expected} = dF_{random}/dF$. Significance is calculated using the Chi-Square test of 2x2 table for counts for $K_{within} > K_{between}$ and $K_{within} < K_{between}$ using the R statistical package (simulating for P-value).

x. 2X2 Contingency Test

The number of substitutions at intergenic, nonsynonymous and fourfold sites were obtained by multiplying substitution rates (dI or dN or dF) with the nucleotide sequence length of the reference genome for each site. Analysis was done using the total number of substitution for the whole genome, which is taken as the sum of the number substitutions for each site. Significance is determined using the Chi Square test of the R statistical package.

xi. Comparing Rates Ratios at Sigma-54 Promoter Sites

Intergenic data were compiled and grouped according to their relation to published Sigma-54 promoter transcriptional regulation activity (Fisher et. al. 2005). In this published study, transcriptional activity between Sigma-54 mutant and wildtype strains of *Borrelia burgdorferi* B31 was determined using Microarray, RT-PCR and Quantitative PCR Analyses. Here, *B. burgdorferi* genome sequences were analyzed with the program seqscan by using a scoring matrix derived from 186 known sigma 54 dependent promoters.

For this study, intergenics were grouped according to how downstream genes (5' to 3') are differentially transcribed in Sigma-54 mutant versus wildtype *borrelia* strains according to such published data. The intergenic grouping is as follows:

1. Intergenics 5' upstream of genes with altered transcriptional activity.
2. Intergenics 5' upstream of genes with no change in transcriptional activity.
3. Intergenics 5' upstream of genes with altered transcriptional activity and computationally predicted Sigma-54 promoter binding sites.

Substitution rates for synonymous and nonsynonymous sites were calculated using the codeml program of the PAML software, using total ORF sequence alignment and the free dN/dS ratios for each branch. Substitution rates for intergenics were calculated using the baseml program of the PAML software and the HKY85 nucleotide substitution model for total intergenic region. For all sites, substitution rates were converted to rate ratios by dividing with median synonymous rates then plotted as a density distribution using the R statistical software package (fig. 8). Draft genome data of main chromosome was used for this analysis.

xii. Sigma-54 Patser Analysis

Using the PATSER software (Hertz and Stormo 1999), intergenic between species alignments in phylip format was computationally scanned for matches to a Sigma-54 promoter binding site position weight matrix (PWM) using A 1:4 A:T to G:C *borrelia* genome nucleotide sequence bias (fig. 9). Cut-off score were set as default. The number hits for each intergenic were tabulated into a PSQL database. The R statistical software was used to plot the number Sigma-54 hits in each intergenic regions compared to intergenic length.

xiii. Intergenic Phylogenetic Footprinting

A PERL script (windows-anal.pl) extracts 6bp to 25bp overlapping fragments from intergenic alignments, each fragment a 1bp step from the next (fig 10). Each 6bp/25bp intergenic alignment fragment is concatenated with its upstream and downstream fourfold ORF alignment sequences to generate a concatenated sequence > 25bp. This concatenated sequence is used to determine substitution rates using the PAML baseml program. Thus, constraint at intergenic sites is a measure of deviations from fourfold rates.

a. *Borrelia* Dataset

For comparisons of substitution rates at convergent and divergent intergenics, substitution rates were calculated for 1bp step, 25bp overlapping window fragments starting at -12bp to 37bp from ORF start/stop sites. The F81 model of the baseml PAML program was used to calculate rates.

Constraint at 5' ORF upstream intergenics was determined using a 1bp step, 6 bp overlapping intergenic fragment window analysis. Each 6bp intergenic fragment is concatenated with its upstream and downstream fourfold ORF sequences to generate a concatenated sequence > 25bp. This concatenated sequence is used to determine substitution rates using the PAML baseml program. Thus, constraint at intergenic sites is a measure of deviations from fourfold rates. To calculate significance for constraint, a sample of fourfold only fragments was generated as the distribution of random rates deviations. Additionally, 6bp fragments of upstream and downstream fourfold ORF sequences were used instead of 6bp intergenics to construct fourfold only concatenated sequence. The random distribution of whole genome

fourfold only rates is generated from a 1000 resampling (with replacement) of fourfold only concatenated sequence substitution rates using the R statistical package. P-values of 0.1 and 0.5 are Quantiles of 0.1 and 0.5 probabilities from this random distribution.

b. *E. coli* Dataset

For *Escherichia coli* dataset, substitution rates were calculated for total intergenic alignment from its upstream intergenic-ORF boundary to its downstream intergenic-ORF boundary. Only fourfold alignments > 25 were used to make a minimal of total 50bp fourfold flanking alignments. Standard errors were constructed by resampling (with replacement) a bin of 6bp overlapping rate measurements. Here, rates were resampled 100 times for each 6bp fragment using the R statistical package, generating a distribution of medians for each bin. Standard error (SE) is calculated as the square root of the variance of medians (using the function variance). For baseline neutral rate, instead of using a whole genome rate as was done for the *borrelia* dataset, local upstream and downstream fourfold rate were concatenated without intervening intergenic fragments and used as a measure of local fourfold rate. Constraint is defined as intergenic rates < neutral fourfold rates.

All results for rate calculations were stored in a PSQL database system. The Regulondb (http://regulondb.ccg.unam.mx/) database for *Escherichia coli K12* and the SwissRegulon (http://swissregulon.unibas.ch) database motevo prediction was used as the transcription factor binding site reference, while the Database of prOkaryotic OpeRons (DOOR http://csbl1.bmb.uga.edu/OperonDB/) was used as the operon reference. Required datasets from these databases were downloaded and stored in the local PSQL database system.

The level of AT/GC content and instances of palindromic patterns for each intergenic fragment were scanned using PERL scripts. The results are stored in a local PSQL database and compared to levels of substitution rates as determined previously, using PSQL queries and the R statistical Package. The number of ORFs at leading versus lagging strand and ribosomal genes were calculated based on the annotated ORFs as stored in a local PSQL database. Standard errors are based on a binomial proportion confidence interval calculation where $SE = \sqrt{p(1-p)/n}$, where p is the proportion and n is total sample number.

3. Results

i. Statistical Power of The Genome Collection

The average distance for the *borrelia* genome within and between species collection is Dw = 0.01 and Db = 0.1 respectively (fig. 2). Using Jukes-Cantor model of molecular evolution, the probability that two sites are identical for within species is:

$$Pw=\frac{1}{4}+\frac{3}{4}e^{-\frac{4}{3}Dw}=\frac{1}{4}+\frac{3}{4}e^{-\frac{4}{3}(0.01)}=0.99$$

While the probability that two sites are identical for between species is:

$$Pb=\frac{1}{4}+\frac{3}{4}e^{-\frac{4}{3}Db}=\frac{1}{4}+\frac{3}{4}e^{-\frac{4}{3}(0.1)}=0.91$$

The combined probability of using both the within and between species data set is:

$$P=Pw \times Pb=0.9$$

The probability that a sequence site being compared will exhibit c = 0 changes where:

L= length of site being observed (i.e. length of transcription factor binding sites).

N=number of genomes compared

maybe measured using the following False Positive (FP) probability using the above combined probability (with c = 0):

$$FP(c;N,P)=\binom{NL}{c}(1-P)^{c}(P)^{NL-c}=P^{NL}$$

(Eddy 2005).

The false positive (FP) probability that a sequence site will exhibit no change at N=7 (the total number of genomes being compared), length of site L=6 (minimal length of transcription factor binding sites), and P = 0.9 (probability that two sites are identical for the genomes compared) is FP=0.01. While the False Positive probability for just between species genome comparisons (N = 4, P = 0.91) is FP = 0.1, and for within species genome comparisons (N = 3, P = 0.99) is FP = 0.83.

ii. Distribution of Regulatory Binding Sites Are Length Dependent For ORF and Intergenic Sites

Using the PATSER software (Hertz and Stormo 1999) and the consensus sequence "Matrix" of Sigma-54 promoter binding sites (fig. 9), it was found that both ORF and Intergenic sites show linear distribution of the number of binding sites to increase in length, indicating that such sites are distributed randomly throughout the genome and are dependent on its length (the longer the length, the greater the number of sites observed) (fig. 11).

iii. Intergenic Substitution Rates Are Evolutionarily Constrained

Rates of substitutions at these within and between *Borrelia burgdorferi* species grouping were determined for noncoding intergenics (I), coding fourfold synonymous sites (F) and nonsynonymous sites (N). The F81 nucleotide substitution model was used, incorporating

unequal base frequencies GC:AT of 1:4 for all sites. Tests using a broad range of nucleotide substitution models produced slight differences, though overall differences amongst sites show the same trend (fig. 12). Ultimately, the PAML F81 model was selected for subsequent analysis as most conservative model with least error in calculating substitution rates compared to other models (data not shown). Synonymous and nonsynonymous substitution rates were determined through using the codon substitution model under the PAML codeml program to compare to rates obtained from the baseml program (fig. 13). Median rates when using different nucleotide substitution models seem to show no difference (fig. 12), though the variance for synonymous rates are greater using the nucleotide substitution model at fourfold sites (fig. 13). Pearson correlation of intergenic, fourfold synonymous and coding sites show that sites within each separate category seem to evolve independently, whereby rates of substitutions are uncorrelated at p-value > 0.1 (Table 3).

Overall, median intergenic rate (dI) is lower than median fourfold rate (dF) though still higher than median nonsynonymous rate (dN) (fig. 14) The lower median substitution rate observed in intergenic and nonsynonymous compared to neutral rate maybe due to either functional constraint or the result of a lower overall mutation rate when compared to neutral sites. To test for this, the distribution of dI and dN to dF rates for both within and between chromosome was compared. Unlike the main chromosome, the rates distribution of cp26 plasmid within species intergenics did not show significant difference to fourfold sites, perhaps due to its small sample size (n = 15). Wilcox test show $p < 0.001$ for main chromosome within and between species intergenics and cp26 plasmid between species intergenics, but $p = 0.37$ for cp26 plasmid within species intergenics. All nonsynonymous sites show rates distribution to be

significantly different from fourfold rates ($p < 0.001$). Additionally, the proportion of constrained intergenic sites are greater than those that are not, with significance $p < 0.0005$ when compared to fourfold rates (Table 4). Overall, this indicates that the distribution of intergenic and nonsynonymous rates are significantly different from fourfold substitution rates, thereby functionally constrained, with most constraint observed for nonsynonymous sites.

iv. Measuring Changes In Constraint and 2X2 Contingency Test For Selection Using Within and Between Species Data.

Substitution rates from within and between species were compared. Here, constraint is measured as the ratio of substitution rates for nonsynonymous and intergenic sites over neutral rates ($K = dN/dF$ and $K = dI/dF$ respectively), where ratios for within and between ratios was compared for constrained sites ($K < 1$) in *Borrelia* main chromosome. It was found that 301 nonsynonymous sites show $K_{within} < K_{between}$, while 414 sites show $K_{within} > K_{between}$ (Table 5 and Table 6). Thus, a greater proportion of nonsynonymous sites are more constrained for between species compared to within species. The change is significant compared to expected K distribution with Chi Square test $p = 0.005$. For K intergenics, 173 sites show $K_{within} < K_{between}$, while 95 sites show $K_{within} > K_{between}$, indicating that for intergenic sites, a greater proportion of sites are less constrained in between species compared to within species. The change is significant compared to expected K distribution at Chi-Square Test $p = 0.013$.

Hence, most intergenic sites show less constraint between species compared to within species, while nonsynonymous sites show the reverse trend, indicating that substitutions at intergenic sites are less deleterious than coding sites. Using a 2X2 Contingency Test that is

based on the McDonald-Kreitman (MK) test (McDonald and Kreitman 1991) for gene coding analysis, it was found that for intergenic sites, the ratio of intergenic to fourfold within species polymorphism is 309/1270=0.24 while the ratio of intergenic to fourfold between species differences is 8967/23658=0.38 (Table 7), suggesting that substitutions occurring at intergenic sites are not only less deleterious, but are positively selected for. Test of nonsynonymous sites show within nonsynonymous to fourfold ratio to be higher (302/1270=0.24) than between ratio (3891/23658=0.16) (Table 8), indicating that substitutions that occur here are deleterious are eliminated before they become fixed in different species. The difference in ratio for both intergenic and nonsynonymous data is statistically significant using a Chi Square test at $p < 0.0005$.

v. Rates Ratios Distribution at Putative Sigma-54 Promoter Sites

Intergenic data were compiled and grouped according to their relation to published Sigma-54 promoter transcriptional regulation activity. Figure 8 show density plot of rate ratios from nonsynonymous, synonymous and intergenic data. For all sites, substitution rates are converted to rate ratios by dividing with median synonymous rates then plotted as a density distribution using the R statistical software package. Using this published study, intergenics are grouped according to how downstream genes (5' to 3') are differentially transcribed in Sigma-54 mutant versus wildtype *borrelia* strains. Both nonsynonymous and intergenics show rate ratios to be lower than synonymous sites, indicating constraint. Nonsynonymous sites show significant shift of ratios distribution for any group, while intergenic sites show shift for sites related to altered transcriptional activity with computationally predicted sigma-54 sites.

vi. Intergenic Phylogenetic Footprinting Analysis

a. *Borrelia* Dataset

a.1. Rates of Substitutions Near Intergenic-ORF Boundaries

Substitution rates starting at 37 bp upstream of 5' divergent and 3' convergent sites were analyzed, assuming 5' sites to have a higher probability of cis-regulatory function compared to 3' sites and therefore more constrained (Cooper, 2004). Each 1bp step overlapping window is 25 bp in length. Substitution rates do not vary significantly at sites within the cis-ORF boundary (fig. 15). Yet, comparing overall divergent to convergent rates show that sites upstream of 5' divergent have a slightly lower median substitution rate when compared to 3' convergent (fig. 15 and fig. 16). Additionally, distribution of rates variance is wider for convergent sites and significantly different ($p = 0.007$ Two Sample Wilcoxon test).

a.2. Constraint Coincides With Promoters of The Alternative Sigma (Sigma-S - Sigma-54) Regulatory Cascade

Known promoter sites upstream of OspC and computationally predicted promoter upstream of BB0680 were used as reference promoter binding sites. Figure 17. show analysis of Sigma-S promoters for cp26 plasmid ORF BBB19 (fig. 17a and fig. 18a), an Outer Surface Protein C (OspC) of the cp26 plasmid, and main chromosome ORF BB0680 (fig. 17b and fig. 18b), a bacterial chemotaxis protein (mcp-4) (Caimano et. al. 2007). Sigma-S sites coincide

with highly constrained intergenic sites, at a significance threshold of $p < 0.1$ when compared to a distribution of randomized fourfold only rates. Constraint levels and visual analysis of sequence data from a collection of *borrelia* -35 sites suggest a 'corrected' position for BB0680 -35 site, that is different from the reference computationally predicted site (Caimano et. al. 2007). Sites with the same or lower constraint level ($p < 0.1$ and $p < 0.05$) with no known function maybe related to enhancer sites for the specific promoter recognition of Sigma-S. In this case, it maybe due to how ospC and mcp-4 Sigma-S promoters are differentially controlled by Sigma-54, wherein only ospC Sigma-S promoters is co-regulated by Sigma-54. Comparing the distribution of rates for intergenic sites to fourfold , it was found that rates distribution for intergenic sites encompassing these promoter regions to be significantly different from fourfold only rates (Wilcox test $p < 0.001$).

b. *Escherichia coli* Dataset

For this dataset, rather than just a portion of 5' ORF-intergenic boundary regions as was done for the *borrelia* dataset, the total expanse of intergenic sites were scanned for matches to reference promoter binding sites obtained from the *E. coli* K21 RegulonDB database. Figure 19 show example matches for a scan of Sigma-70 promoter binding sites in three intergenic regions. Substitution rates for all matching Sigma-70 sites (colored red) are well below their fourfold flanking 'neutral' rates (data not shown). Interestingly, promoters span the most linear portion of the graphs and are located within the lower bounds. Figure 20 shows a close up of the Sigma-70 promoter region of the NADH dehydrogenase gene, with -35 and -10 binding regions at a slightly higher rate than +1 transcriptional start site (RBS), which is a similar pattern as that

observed in *borrelia* OspC (BBB19) promoter. Overall, promoter site coincide with those showing lowest constraint within the span of the intergenic sequence.

As in *Borrelia*, cases of 'unknown' sites with similar constraint levels are found. To determine how such sites maybe related to possible 'unknown' regulatory sites versus instances of 'false positives', the percentage of AT/GC content and palindromic types for each fragment were compared to their corresponding substitution rate. About 77% (± 2.53%) of palindromic fragments are constrained (fig. 21). Similarly, about 32% (± 0.3%) of AT rich sites are constrained, compared to 16% (± 0.2%) of those that are not (the rest have equal proportion of AT to GC content) (fig. 22). Additionally, by classifying fragments according to their upstream and downstream ORFs orientation (fig. 23), > 53% of constrained sites are associated divergent (⟵⟶ or unidirectional (←←and →→ ORFS, indicating potential cis-transcriptional regulatory role. The highest percentage of constrained sites (85%± 1.34%) are located between unidirectional (←←or --) ORFs on the leading (-) strand of the *E. coli* genome, which maybe accounted for by the greater number of ORFs (~51%) as well as highly expressed ribosomal genes (~69%) on the leading strand (Rocha 2002).

4. Discussion

A comparative genomics approach using multiple genomes of closely related *borrelia* species was used to identify patterns of evolutionary constraint on non-gene-coding intergenic regions. The aim of this study was to develop a comparative genomics bioinformatic strategy using available genome data to predict for functional sites in intergenic regions, particularly in developing a novel phylogenetic footprinting approach using very closely related bacterial genome data (>90% nucleotide sequence identity) to identify cis-regulatory functional sites. A collection of 7 *borrelia* genome strains > 90% in nucleotide sequence identity were used to develop the strategy, while additional distinct 3 *E. coli* genome strains at approximately the same phylogenetic distance were used to corroborate the findings and identify general molecular patterns as it is related to sequence evolution and cis transcriptional regulation. As data from close species genome sequencing becomes increasingly available, especially for bacteria, the basic strategy as developed in this study will provide a way to take advantage of this influx of data, to uncover new biological properties, to test for new hypotheses.

The study has revealed several fundamental properties on the evolution of intergenic non-gene-coding regions in a collection of closely related bacterial species:

1. That most non-gene-coding intergenic portion are evolving under constraint, though not as strong as the gene coding portion of the genome.
2. That constraint is more relaxed for intergenics compared to coding where substitutions for total genome intergenic region is positively selected for, indicating that, as expected for such degenerate sites, substitutions at intergenic sites are less deleterious than at coding sites.

3. That a slightly higher level of constraint is observed near Intergenic-5'ORF boundaries, indicating some relation of constraint to 5'cis regulatory function.
4. That some constrained intergenic sites correspond well to promoter binding sites, while the observation of constraint at many more sites within the intergenic region indicate that intergenic sites are functionally compact.
5. Using a separate *E. coli* genome dataset, it was found that the most constrained sites not only coincided with several known promoter sites and potential cis-regulatory sites, but also show several general patterns for cis transcriptional regulatory sequence.

i. The Power of Comparative Genomics

Comparative genomics methods rely heavily on the level of evolutionary distance between the species compared (McCue et al. 2002; Eddy, S. R. 2005). Especially for very closely related species, as for the case of this study, functional sites maybe masked, as sites may not have diverged far enough to be differentiated from neutrally evolving sites. Alternately, if species used are too divergent from one another, functional sites that are absent in some of the species compared will not be detected (Hahn 2007). Thus, a measure of effective evolutionary distance becomes essential (Eddy 2005).

The *borrelia* genome collection is very closely related to one another, with sequence identity > 90% for all the seven species used (fig. 2) and around ~7% of genomic regions are noncoding intergenics. It has been assumed that data from such a closely related sequence collection will not be able to provide enough information to discriminate functional fitness in noncodings, such as those short sequence binding sites for cis transcriptional regulation (Cliften

et. al. 2001; McCue et. al. 2002). The problem for such very closely related genome collection is the ability to differentiate sites that are conserved due to actual evolutionary selection for function as opposed to chance.

Using the short sequence pattern search application PATSER (Hertz and Stormo 1999), a search of a representative Sigma-54 matrix (fig. 9) show that such sites are distributed randomly throughout the genomes that seem to be strongly dependent on sequence length. This is evident both in coding (ORFs) and intergenic sequences (fig. 11). Thus, a search based on the occurrence sequence conservation alone was not able to distinguish conserved sites from random hits in a collection of genome sequence > 90% in nucleotide identity. Comparing the number of matching Sigma-54 sites to positions relative to 5'ORFS did not show any significant bias of pattern hit to genome location (data not shown).

This pattern search method is dependent on the matrix (PWM) used to search for such sites, which in this case, were based on a collection of bacterial transcriptional sites not specific to *borrelia*. As functional sites may vary amongst species, the matrix may become inappropriate for species specific sites as they maybe obscured by having additional species sequence represented within the matrix. A *borrelia* scoring matrix for cis-regulatory promoter binding sites such as the Sigma-54 was not applicable due to lack of data for the *borrelia* sequence, and thus are not statistically robust. Significance scoring in this approach is based on an information content model of statistical inference (Pavesi, Mauri, and Pesole 2004)), which does not include molecular evolutionary assumptions.

The purpose of this study is to determine if by incorporating some fundamental properties of sequence evolution, a better discriminating power could be achieved. That is, developing a

phylogenetic footprinting strategy that is aimed at close species data, in this case, > 90% nucleotide sequence identity. What needs to be determined is whether such a closely related species dataset will be able to provide the discriminating power needed, or are they are indeed "uninformative" for phylogenetic footprinting type of analysis.

First, a test on the theoretical probability of the information value of the *borrelia* sequence collection was done. The test gives a measure of the significance of conserved sites as they are related to functional fitness. Are sequence conservation observed due to just background neutral substitution rates or to actual evolutionary selection? In other words, what is the probability of finding false positives (FP) using the genome collection available in order to find "true" functional constraint? For this purpose, the Jukes-Cantor model of molecular evolution was used as the model for neutral substitution rate.

Using a simplified version of Sean Eddy's mathematical model (Eddy 2005), it was found that a site of at least 6bp in length will result in a false probability of FP = 0.01 if all 7 genomes are compared. In other words, when sequences are compared, there is a 1/100 chance that sites observed to be under constraint are in fact evolving neutrally and bear no fitness. If only within species genomes are compared, FP = 0.83, while for between species comparisons, FP = 0.1.

In general, using this model, it was found that if all 7 of the *borrelia* genome sequences were used, functional constraint may be observed at significance level of P = 0.01. Thus, the *Borrelia* genome sequence as a total collection is theoretically able to differentiate constraint from neutral evolution 99% of the time at a minimal nucleotide sequence length of 6bp. Comparisons of within species sequences only will lead to a high number of false positives,

while between species comparisons is weakly significant. The next step is to determine how this applies to an actual measure of functional constraint in the sequence dataset. To do this, an understanding of the background evolutionary properties of these non-gene-coding intergenic sites becomes important.

ii. Molecular Evolutionary Properties of *Borrelia burgdorferi* Noncoding Intergenic Regions

a. Justification for Within Species PAML Analysis

Seven *Borrelia* genomes were divided into within and between species phylogenetic group based on OspA (Outer Surface Protein A) alignments (fig. 2, Table 1), a Lyme disease related loci that has been extensively used to serotype *Borrelia* isolates (Lebech et. al. 1994; Wang et. al. 1999). This phylogenetic grouping is supported by recent studies using MLST (Multiple Locus Sequence Typing) (Hoen, et. al. 2009; Margos, et. al. 2008; Qiu, et. al. 2008). *Borrelia afzelii* and *Borrelia garinii* occurs in Eurasia (Comstedt et. al. 2009; Krupka et. al. 2009) while *Borrelia burdorferi sensu stricto* has been found in both Europe and North America and is the predominant strain in North America that cause Lyme Disease (Hoen, et. al. 2009; Margos, et. al. 2008; Qiu et. al. 2008). *Borrelia burgdorferi sensu stricto* is composed of stable clonal groups, with only occasional recombinant phenotypes caused by plasmid exchanges (Qiu et. al. 2004). It is suggested that North American *Borrelia burgdorferi sensu stricto* clonal groups originated from Europe with a prehistoric population size expansion and east-to-west radiation of descendant clones from founding sequence types in the Northeast (Hoen et. al. 2008). Comparisons of North American *Borrelia burgdorferi sensu stricto* strains using MLST (Multi Locus Sequence Type) analysis of housekeeping loci indicate a stable phylogeny with

restricted geographic distribution for each clonal type, suggesting that they once belonged to an admixed population but are now isolated (Hoen, et. al. 2009; Margos, et. al. 2008; Qiu, et. al. 2008). Clonal types has also been associated to differences in host specificity and human pathogenesis in Lyme Disease (Brisson and Dykhuizen 2004). These studies support the use of the designated within and between species phylogenetic grouping to determine variations in substitution rates at a population genetic level using PAML.

b. Constraint of Intergenic Regions

Genome sequence sites maybe differentiated from one another based on how they affect the evolutionary fitness of the whole organism. Differences in evolutionary fitness are determined through nucleotide sequence variation as measured by the rates of nucleotide substitutions amongst orthologous sites being compared. Orthologous sites are those with shared evolutionary history through a recent common ancestor, and have not undergone duplications. Differences in evolutionary fitness are measured as deviations from neutrally evolving sites.

Neutrally evolving sites are assumed to bear no fitness and therefore not functionally selected for. For this study, fourfold-synonymous sites were chosen as the background neutral substitution rate (dF), which are sites where any of the four base changes at third position codon does not to lead to amino acid change. Sites with higher or lower substitution rates than dF are considered to be functionally selected for. Sites with rates that are higher than dF are considered to exhibit selective advantage and substitutions occurring at such sites are positively selected for, whereas sites with rates that are lower than dF are considered to be selectively constrained and substitutions occurring at such sites are deleterious and negatively selected for or removed.

Comparisons of substitution rates for *borrelia* genome sequence data show an overall trend that most noncoding intergenic regions have lower nucleotide substitution rates (dI) than fourfold synonymous rates (dF), though still higher than non-synonymous rates (dN). Overall, *Borrelia* noncoding intergenic regions show evolutionary constraint for both within and between species population. Evolutionary constraint is measured by comparing intergenic rates (dI) and nonsynonymous rates (dF) to fourfold synonymous rates (dF), and constraint occurs if dI < dF and dN < dF. Both within and between species show median intergenic rates (dI) and nonsynonymous rates (dN) to be lower than fourfold sites, though dI is still higher than dN (fig. 14), thus, intergenics seem to be constrained for most sites, though at a lesser extent compared to noncoding sites.

Wilcox tests of intergenic, fourfold synonymous and nonsynonymous substitution rates distribution indicates that dI and dN rates distribution are significantly different from fourfold substitution rates, suggesting constraint. Most intergenic sites show constraint when compared to randomized data using fourfold sites. Additionally, the proportion of constrained intergenics are greater than those that are not. Thus, constraint even within these closely related species (> 90% sequence identity) are indicative of selection rather than insufficient time to evolve.

These results confirm previous studies of bacterial noncoding evolution where constraint (purifying selection) seems to be the main mode of evolution for such sites (Rogozin et al. 2002, Rajewsky et al. 2002, Molina et al. 2008). The results maybe affected by selection occurring at synonymous sites due to the preferred use of codons for highly expressed genes (Hirsh et al. 2005, Akashi 2001). *Borrelia* shows codon usage bias between leading and lagging strands of replication, resulting in a greater number of genes located at leading strands, where

transcriptional selection appears to be responsible for such enrichment (McInerney 1998). Rates of substitutions at these synonymous sites will be lower than neutral expectation, hence such bias will increase the number sites exhibiting positive selection, and decrease the number of those constrained. The analysis show that constraint levels for a large proportion of intergenics are lower than at fourfold sites, for within and between species, thus codon bias seem to produce only a negligible impact on our ability detect constraint. Positive selection has been observed occurring on synonymous sites in mammalian genes (Alissa et al. 2007), exhibiting higher rates of substitution in neutral sites. This effect will increase the number of intergenic and nonsynonymous sites seemingly under constraint. Adaptive evolution of synonymous sites has yet to be found in prokaryotes and therefore not corrected for in this study.

To test for bias in background substitution rate for each genome category, substitution rates were compared between intergenic, nonsynonymous and fourfold regions (Table 3). Pearson correlation show that substitution rates between the categories are uncorrelated, meaning that variation between sites and the observation of such constraint are not due bias from background substitution rates, but most probably as evidence for constraint.

Observed constraint may also be affected by difference in AT and CG content between intergenic, ORF coding sites. Overall, the *borrelia* genome is GC poor with approximate GC:AT composition of 1:4 for all sites, showing only negligible difference in composition between intergenic versus ORF coding sites. Additionally, orthology assignment did not take into account genome rearrangements as they occur based on total phylogeny, but are based on a 1:1 comparison to the B31 *Borrelia burgdorferi* strain as reference genome. Some sites may not be truly orthologous though most main chromosomal regions are syntenic, indicating seemingly

few duplication events.

These observations indicate that intergenics contain functional sites with different underlying evolutionary properties when compared to coding sites. That intergenic sites, as expected, seem to be more degenerate, and are more open to 'changes' compared to coding sites, or are less deleterious. In order to further explore for this idea a comparison of rates between different *borrelia* populations was conducted.

c. Using Population Genetic Inference, Substitutions at Intergenic Sites Are Less Deleterious Than Coding Sites And Are Positively Selected For

If sites were neutral, the substitution rates at within species populations will be the same as rates at between species populations after taking into account changes in neutral rates amongst different populations. To take into account this change in neutral rates, we can calculate a rate ratio for within (K_{within}) and between ($K_{between}$) species populations where K = site rate/neutral rate. The site maybe within an intergenic region with rate dI, or a coding nonsynonymous region with rate dN. The fourfold synonymous rate (dF) was used as the neutral rate. If all $K < 1$ and hence all sites are under constraint, we can say that constraint increases if more sites show $K_{within} > K_{between}$, and constraint decreases if more sites show $K_{within} < K_{between}$.

By calculating the number of sites where $K_{within} > K_{between}$ and $K_{within} < K_{between}$, it was found that intergenics show a greater number of sites with decreased constraint at between species compared to within species populations, while coding sites show a reverse trend, indicating that nucleotide substitutions occurring at intergenics seem to be less deleterious as species diverge. This difference is significant when compared to expected rate due to neutral rate changes at p =

0.013.

Though most intergenic sites are constrained when compared to fourfold synonymous sites, using a 2x2 Contingency Test, it was found that substitutions at total intergenic sites are not only less deleterious, but are also positively selected for. The test is similar to the McDonald-Kreitman (MK) test (Mcdonald and Kreitman 1991) for gene coding analysis, but as applied to intergenic non-coding sites. In the MK test, it is assumed that when observed variation is neutral, then the rate of substitution between species and the amount of variation within species is both a function of the substitution rate. Thus, under neutrality, the ratio of nonsynonymous to synonymous (N/S) differences between species should be the same as the ratio of nonsynonymous to synonymous (N/S) polymorphisms within species. In effect, N/S between > N/S within species indicates positive selection while N/S between < N/S within species indicates negative purifying selection. For the intergenic 2X2 Contingency Test, nonsynonymous sites (N) is replaced by counts for intergenic sites (I).

For the borrelia genome data, intergenic sites (I) were compared to fourfold sites (F) by calculating the number of substitutions occurring at total intergenic sites (I) to the number of substitutions occurring at total fourfold synonymous sites (F). Similarly, nonsynonymous sites (N) were compared to fourfold sites (F) by calculating the number of substitutions occurring at total nonsynonymous sites (N) to total fourfold sites (F). Intergenics show I/F between > IF within (0.38 > 0.24) while nonsynonymous show N/F between < N/F (0.16 < 0.24), indicating that while most intergenic sites are constrained, when substitutions occur, they are not only less deleterious than coding sites, but also are positively selected for.

This property may be related to the different functional significance of nucleotide

changes occurring at these different regions of the genome. As most intergenic sites are considered to be regulatory in function, the observation that constraint is relaxed as species diverge and positive selection drives changes indicate that for these *Borrelia* species differences maybe more dependent on changes at regulatory as opposed to gene coding sites.

In relation to the Lyme Disease phenotype, uncovering the molecular basis of the different Lyme pathogenicities observed between species of *borrelia* will increasingly depend on how such phenotypes are influenced by changes of gene regulatory sequences within the non-gene-coding portion of the genome. As such, non-gene-coding sequence variation may play a larger role than previously thought, in the virulence and clinical manifestations of Lyme. This includes transcriptional, post-transcriptional and translational regulation as determined by elements within the intergenic portion of the bacterial chromosome.

In effect, it becomes interesting to determine the extent of how constraint is maintained at different regions of the intergenic regions, and how this is related to some regulatory function. For this study, the focus will be on how possible transcriptional cis-regulatory function is evolutionarily maintained.

d. Using Published Expression Data, Intergenic Sites Are Less Deleterious Than Coding Sites

When sites are grouped according to their relation to sigma-54 related transcriptional activity, nonsynonymous sites show no significant shift of ratios distribution for any group (fig. 8). On the other hand intergenic sites show a shift for sites related to altered transcriptional activity with computationally predicted Sigma-54 sites only, while total sites related to altered

transcriptional activity seem to show no difference from total sites. Intergenic sequence sites seem to be more variable for that subset harboring Sigma-54 putative promoter sites, indicating that such sites are less deleterious than the bulk of intergenics. For the case of *Borrelia*'s relation to the Lyme phenotype, perhaps as Sigma-54 sites are involved in the regulation of key virulence factors, constraint within these regions are more relaxed which allows for the evolution of new virulence phenotypes; this observation maybe indication that varying levels of constraint seem to be related to some differences in function.

Other than cis transcriptional regulation, changes in constraint at intergenic sites may also be influential in other forms of sequence dependent regulatory patterns. Studies have demonstrated a key role of intergenic small RNA's in bacterial response to stress and the regulation of factors important for virulence, whose role is to modulate translational activity through changing the stability of mRNA post-transcriptionally (Gottesman 2005). Recently it was observed that intergenic regions of *Borrelia* carry evolutionarily stable RNA secondary structure motifs in a form of repeat elements, some associated with protein genes of large sequence variability. It is thought that this RNA motif conservation allows a large variability of amino acid sequence, perhaps to create new virulence factors (Delihas 2009).

In *Borrelia*, it is now well established that efficient infection of *borrelia* to ticks or mammalian hosts involves the expression of virulence determinant proteins, mostly in plasmid sequences. The high sequence variability of these proteins is thought to be the main determinants of Lyme infection. For example, the outer surface lipoprotein C (OspC) is a highly polymorphic single-copy gene of 22 major ospC groups worldwide within the *Borrelia burgdorferi sensu stricto* lineage (Brisson and Dykhuizen, 2004). Its expression is heightened during tick feeding,

allowing *borrelia* to migrate from the midgut of the tick to the salivary glands during feeding. It will be interesting to determine how such clinical manifestations maybe related to the variability of constraint as observed in between bacterial intergenic sequences.

iii. Phylogenetic Footprinting for Bacteria

Available phylogenetic based bioinformatic methods to predict for regulatory sites within noncoding regions are usually optimized to assume the greater sequence heterogeneity of noncoding regions in eukaryotes, even for within species comparisons (Boffelli et. al. 2003; Boffelli et. al. 2004; Shringarpure et. al. 2008). It has been assumed that comparisons of species with sequence identity > 70% will not be able to distinguish sites of true conservation to from background neutral substitution rate as species have not evolved far away enough from each other to be able to resolve this difference (Cliften et. al. 2001; McCue et. al. 2002). Yet, this study suggest that constrained sites are differentiable from fourfold synonymous neutral rates for both within and between species comparisons, even amongst species > 90% nucleotide sequence identity.

Additionally, there is an indication that substitutions occurring at coding versus noncoding regions exhibit selective properties distinct from each other and from background neutral substitution rate. As a result, the inability to distinguish sites maybe the result of the compact nature of bacterial genomes, that is, most sites will show functional constraint. The task is for this kind of dataset therefore is to distinguish the different units of function in one sequence span (i.e. different binding site regions). To do this, an understanding the variability of sequence evolution patterns across the total intergenic space is needed.

Using this assumption, a phylogenetic footprinting strategy was developed that gives a measure for this pattern of fine sequence evolution that is distinct from background neutral rate (fig. 10). In this case, it measures distinct evolutionary patterns per short sequence fragments across noncoding intergenics.

Traditionally, phylogenetic footprinting methods rely only on the alignment of the sequence of interest to measure differences in substitution rates between sites (fig. 1). In the current method, the intergenic fragment of interest is appended to their local orthologous synonymous fourfold rates. The result is a measure of deviations from local neutral substitution across sites as reference for functional constraint (fig. 10). The method is dependent on the assumed model of nucleotide sequence evolution as it is compared to fourfold rates. First, bias at fourfold rates used for the neutral reference may mask actual deviations from 'true' neutral rates. The method is also dependent on the choice of the statistical significance level for constraint. Significance was determined either through a distribution of whole genome fourfold substitution rate in *borrelia*, or a distribution of rates as observed locally across one intergenic space in *E. coli*. Further improvements to the method will need to correct for such biases. Overall it provides an initial approach to determine a measure for fine sequence evolution at for short sequence fragments.

The method was used to measure for constraint at 5' ORF-Intergenic boundaries and known promoter sites in *borrelia*. Using a separate *E. coli* data set, the method was further tested on a list of known promoter binding sites in *E. coli*, and provided a means to further characterize the molecular evolutionary properties of noncoding intergenics.

a. Constraint At Intergenic Sites Is Related To Possible Cis-Upstream Gene Regulatory Activity in Borrelia

Substitution rates do not vary significantly at different locations from cis-ORF boundary (fig. 15). Yet, comparing median rates from 5' divergent versus 3' convergent intergenics show greater variance for 5' divergent intergenics. Additionally, 5' divergent intergenics show slightly lower median substitution rate than 3' convergent intergenics (fig. 16). The finding that higher constraint coincide with 5' ORF sites are supported by previous study using 22 clades of bacteria (Molina and Nimwegen, 2008), also in studies using rodents (Keightley and Gaffney 2003), indicating such properties to be fairly universal for both prokaryotic and eukaryotic species. As they are located cis-upstream of 5' ORF regions, most constraint observed within intergenic regions maybe dominated by some form of cis-transcriptional regulation.

b. Constraint at 6bp Intergenic Sites Coincide with Known Promoter Binding Sites In *Borrelia*

To evaluate the extent of constraint at particular cis-regulatory sites, constraint was measured at a 6bp window intergenic regions harboring Sigma-S dependent promoters. Sigma-S is involved in stress response in *Borrelia burgdorferi* and is known to activate a subset of virulence determinant related genes in Lyme infection (Caimano et. al. 2004, Eggers et. al. 2004). A subset of Sigma-S promoters are co-regulated by Sigma-54 promoters within the Sigma-S -Sigma-54 cascade, whose genes being regulated are expressed specifically during mammalian infection of *borrelia* in Lyme Disease (Fisher et. al. 2005, Hubner et. al. 2001, Smith et. al. 2007, Ouyang et. al. 2008).

Known promoter sites upstream of BBB19 (OspC) and hypothetical promoter upstream of BB0680 were used as reference promoter binding sites. It was found that -10 and -35 Sigma-S sites coincide with highly constrained intergenic sites, at a $p < 0.1$ significance threshold. There are sites with the same or lower constraint level ($p < 0.1$ and $p < 0.05$) with no known function perhaps related to enhancer sites for the specific promoter recognition of Sigma-S. In this case, ospC and mcp-4 Sigma-S promoters are differentially controlled by Sigma-54, wherein only ospC Sigma-S promoters is co-regulated by Sigma-54. Interestingly, the window analysis suggest a correction for the -35 binding site from the reference computationally predicted site, which is more similar to the -35 sequence of known BBB19 promoter (fig. 17b and fig. 18b).

A number functionally unknown sites are constrained at a similar significance level to Sigma-S promoter binding sites ($p < 0.1$ and $p < 0.05$), which maybe a reflection of the functionally compact nature of noncoding intergenic sites in *borrelia*, rather than insufficient time for sites to evolve. To test for this, distribution of rates for intergenic sites to fourfold only rates were compared in order to determine if the distribution of rate substitutions at intergenic sites is the same as neutral rates. It was found that distribution for intergenic sites encompassing these promoter regions to be significantly different from fourfold only rates (Wilcox test $p < 0.001$). This supports the idea that most sites within these promoter intergenic regions are not evolving under neutral rates and that intergenic sites are functionally compact and functionally constrained. These findings are further explored by an analysis of a distinct dataset of *E. coli* genomes.

c. Test Case: Using *Escherichia coli* Genome Dataset

Genomes from *E. coli* strains K21, O157, APEC were used to test the extent of the phylogenetic footprinting method as it applies to a distinct dataset of similar phylogenetic distance to *borrelia*. The advantage to this dataset is the availability of associated functional studies, such as was found in the Regulondb, SwissRegulon and DOOR database for transcription factor binding and operon site respectively.

The phylogenetic footprinting method across total intergenic regions between 227 ORFs show predicted/known promoter binding to be associated with the most constrained sites within the intergenic regions (at $p \sim 0.1$) and maybe differentiated from neighboring sites across the intergenic sequence span. Similar to the *borrelia* dataset, many more observed constrained sites do not correspond to any known function. Thus, constrained sites, those that show substitution rate level lower than fourfold substitution rate ($dI < dF$), were analyzed for several known patterns of regulatory binding regions.

Transcription binding sites that bind helix-turn-helix pattern of transcription factors are pallindromic, and the classic promoter is known to have higher AT content to accommodate reduced strand separation energies during transcription initiation and unwinding (DeHaseth and Helmann 1995). A large percentage of pallindromic fragments are constrained, and a slightly higher percentage of constrained fragments are AT rich, thus most of constrained sites show indication of some general patterns of transcription factor binding site, thereby suggesting that most constrained sites are not instances of 'false positives', but is indication of functional constraint.

Interestingly, the highest percent of constrained sites are located within the leading

replicative strand of the *E. coli* genome. As in the *borrelia* genome, this strand asymmetry of constraint maybe accounted by transcriptional selection occurring at synonymous sites due to the preferred use of codons for highly expressed genes, and/or the result of mutational bias as known to occur on *E. coli* lagging strand synthesis (Maliszewska-Tkaczyk et. al. 2000; Veaute et. al. 1993). The observation that slightly greater number of ORFs, including highly expressed ribosomal genes, and annotated promoters/operon are coded in the leading strand suggests transcriptional selection to be a greater underlying factor for such constraint bias on the leading strand.

Codon bias may mask the true level of constraint at intergenic sequence, wherein a higher level synonymous rate will increase the number of constraint observed while lower synonymous rates will decrease it. Yet, for this phylogenetic footprinting framework, such bias will not affect the measure of constraint and comparisons thereof between sites. This occurs as constraint is taken as deviations from a common fourfold synonymous rate of flanking ORFs, and predictions of functional constraint is based on the substitution rate variation of neighboring sites across one intergenic sequence span, not across the genome. However, uncommonly high levels synonymous rates will lower the scale of differences observable across sites; alternatively, an uncommonly low level of synonymous rates may exaggerate the differences between sites such that they are not differentiable from random, these caveats remains to be tested.

5. Conclusion

Comparisons of genomes of the *Borrelia burgdorferi sensu lato* species, show that most noncoding intergenic regions are evolutionarily constrained. Substitutions occurring at intergenic sites are less deleterious when compared to nonsynonymous sites and that constraint is observed to be slightly higher upstream of 5' ORF sites compared to 3' ORF sites. This is taken as evidence of possible relation of constraint to cis-regulatory function.

Additionally, intergenic regions in *borrelia* seem to be functionally compact. That constraint even within these closely related species (> 90% sequence identity) are indicative of selection rather than insufficient time to evolve. It has been shown recently that transcription factor (TF) binding sites are more abundant in bacteria with smaller genomes (Molina and Nimwegen 2008). The number of TF binding site are inversely related to size of genome, whereas the number of TF 's increases quadratically with genome size. *Borrelia* has one of the smallest genomes for a prokaryotic species, hence its need to accommodate a larger number TF binding sites at a small intergenic space. The observation of promoter sites located within ORFs in prokaryotic systems maybe a consequence of this phenomena (Kawano et al. 2005). Overall, these observations provide additional support for the idea that noncoding intergenic sites in the *borrelia* genome are functionally compact.

Analysis of two alternative sigma promoter sites regulating ospC and mcp-4 genes in cp26 plasmid and main chromosome of *borrelia* respectively show that -10 and -35 sites coincide with sites that show most constraint. Several other sites within the intergenic region also show similar levels of constraint that we take as unknown functional regions that may or may not be related to such sigma promoter regulatory function. By using an additional set of *E.*

coli genomes, it was found that not only constrained sites similarly coincide with known binding regions, most constrained sites correspond well to several known properties of transcriptional factor binding sites.

The prospect of additional genome sequence information from various species of *borrelia*, as is for many other bacterial species, should provide the means to increase the utility of comparative genome methods in the identification noncoding regulatory functional sites important in developing species specific fitness variability, such as virulence determinants in *borrelia* in relation to Lyme infection. Non-gene-coding sequence variation may play a larger role in the determination of virulence and clinical manifestations. This includes transcriptional, post-transcriptional and translational regulation as determined by elements within the intergenic portion of the bacterial chromosome.

In summary, a phylogenetic footprinting approach for very closely related species (> 90% nucleotide sequence identity) was developed to test for specific sites of transcriptional regulation, which was additionally tested using the *Escherichia Coli* genome dataset, providing way to measure for molecular function at a fine phylogenetic level through the understanding of the patterns DNA sequence evolution. The challenge is to design increasingly sensitive methods that would allow the harvesting of functional information (i.e. specific TF binding sites) through using comparative genome information as a function of its phylogenetic information (Eddy, S.R. 2005). For small bacterial genomes, such as *borrelia*, methods to measure evolutionary changes within closer phylogenetic range becomes important as individual functional sites maybe masked due its compact nature, especially for the identification those species specific sites that maybe lost at greater phylogenetic distance.

FIG. 1. Phylogenetic Footprinting Example. " CLUSTALW alignment of SNR39 sequences (encoding a snoRNA) from eight different Saccharomyces species. Box C and Box D are known functional elements; the guide" is the sequence complementary to rRNA sequence adjacent to the methylation site. The structure of the RPL7A transcript including the intronic SNR39 gene is shown above the sequence alignment (Cliften et al 2001).

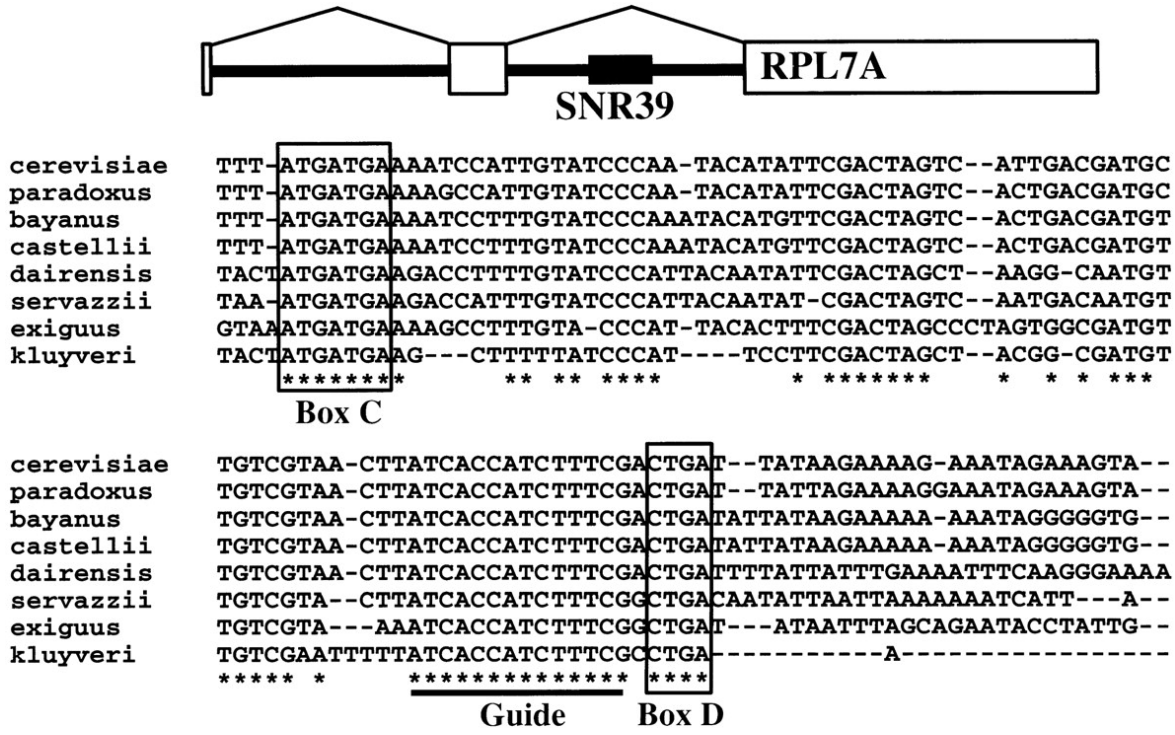


FIG. 2. *Borrelia*'s outer surface protein A (ospA) MrBayes tree where each branch represents the species with strain name. There are seven strains used in this study, four within species group (*Borrelia burgdorferi* strains 297, B31, JD1, N40) and four between species group (*Borrelia burgdorferi, Borrelia afzelii, Borrelia garinii and Borrelia bissetti*). Within species group has on average 99% sequence identity while between species group has on average 91% sequence identity. Red numbers indicate bootstrap values of branching points as estimated using the MrBayes program.

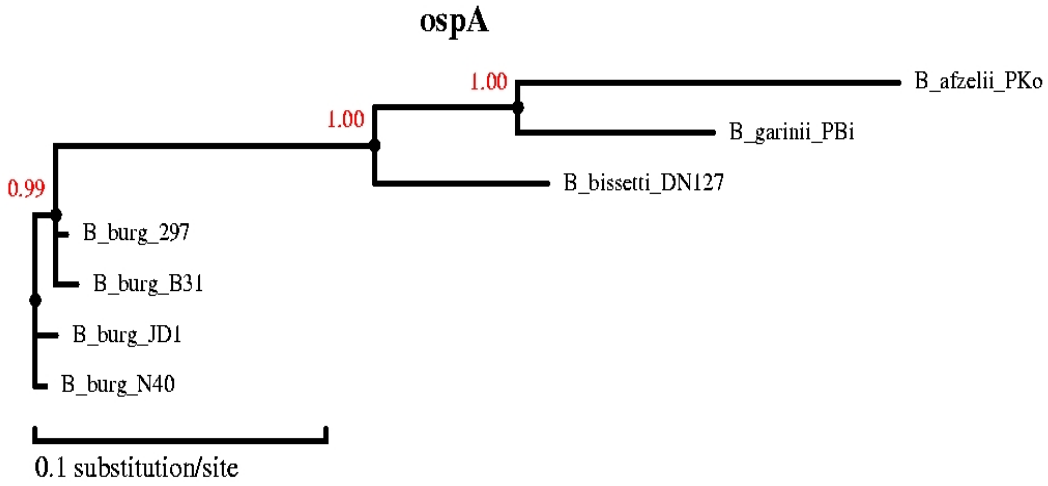
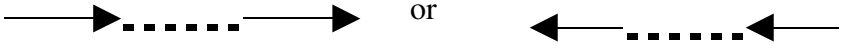


FIG. 3. Types of Intergenics (Intergenic regions shown as dashed line, ORF shown as arrows pointing from 5' to 3').

i. Convergent

ii. Unidirectional



iii. Divergent

FIG. 4. Intergenic data analysis bioinformatic pipeline to calculate substitution rates. Intergenics 25-500bp in length are aligned using the CLUSTALW software. Each intergenic alignment includes 300bp upstream and downstream flanking sequences from ORF region. Following alignment, each intergenic data set are separated into within and between species alignments, which includes the removal of 300bp flanking regions. Rates of substitutions between alignments are determined using the BASEML program of the PAML software, which includes phylogenetic tree data obtained using PHYLIP. Substitution rates are: dI_w for within species intergenics and dI_b for between species intergenics.

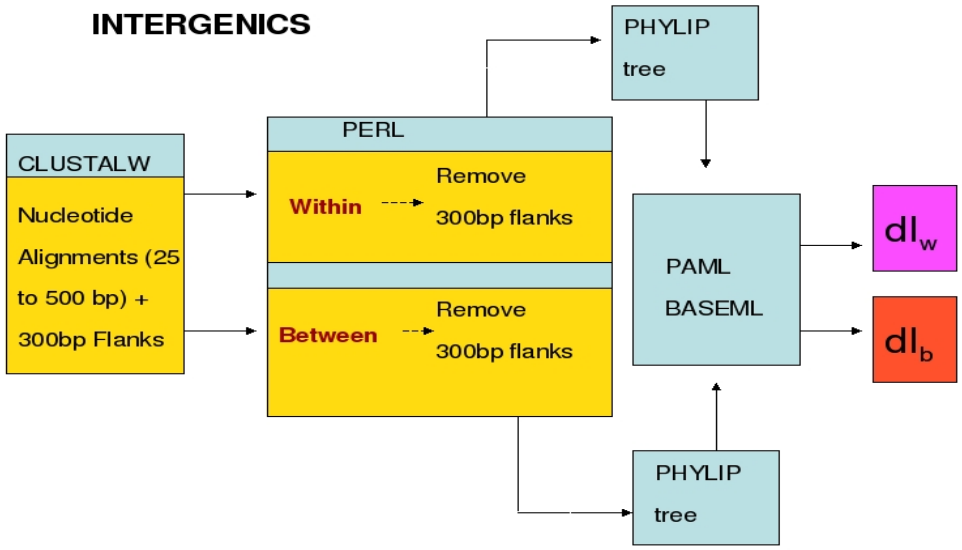


FIG. 5. ORF data bioinformatic pipeline to calculate substitution rates. Following translation of ORF nucleotide sequences, amino acid sequences are aligned using CLUSTALW software then converted back to nucleotide sequences. Alignments are then separated into within and between species into total codon, fourfold third, and coding sites. Rates of substitutions for fourfold and coding positions are determined using the BASEML program of the PAML software, while the CODEML program of the PAML software determines nonsynonymous and synonymous rates from total codon. PAML runs include phylogenetic tree data obtained using PHYLIP. Substitution rates from baseml program are: dN_w for within species nonsynonymous , dN_b for between species nonsynonymous, dF_w for within species fourfold synonymous and dF_b for between species nonsynonymous. Substitution rates from codeml program are: dN_{wc} for within species nonsynonymous, dN_{bc} from between species nonsynonymous, dS_{wc} for within species synonymous and dS_{bc} for between species synonymous.

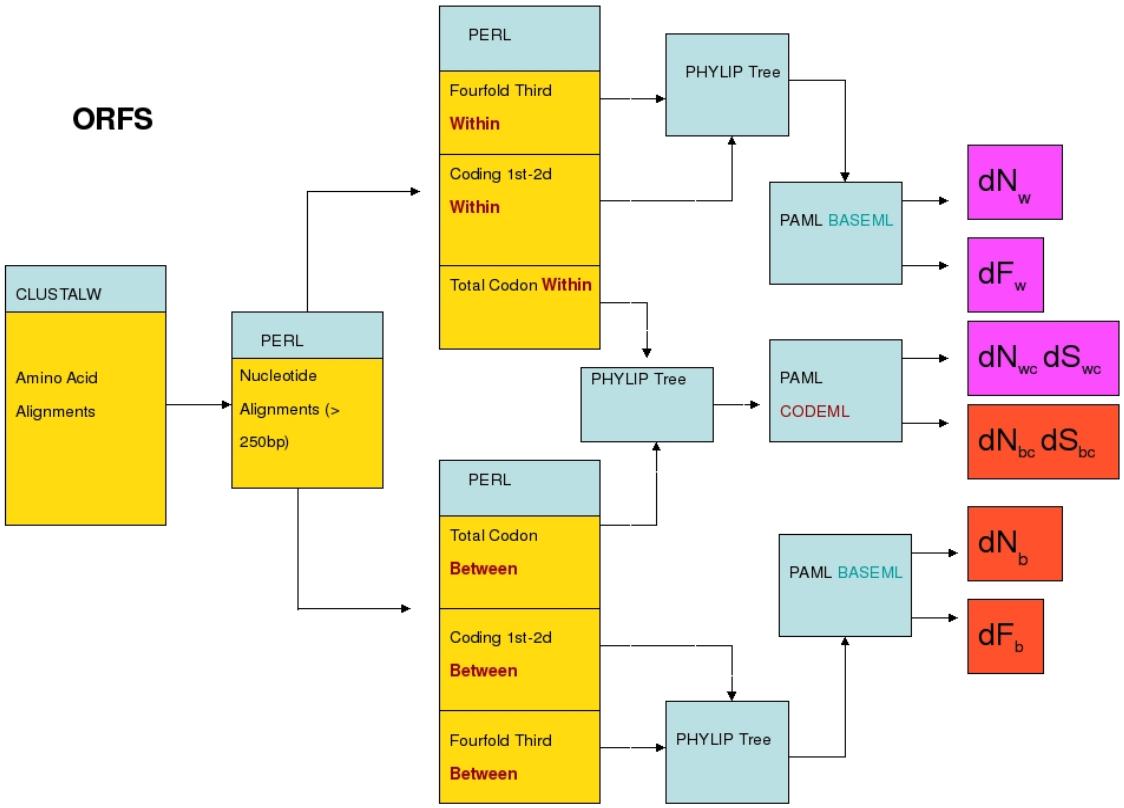


FIG. 6. *Borrelia* synteny browser, used to visualize syntenic regions in plasmids and main chromosome data to determine orthologous set. Top scoring blastp matches are color coded uniformly and those that are seen as syntenic are selected for further analysis.

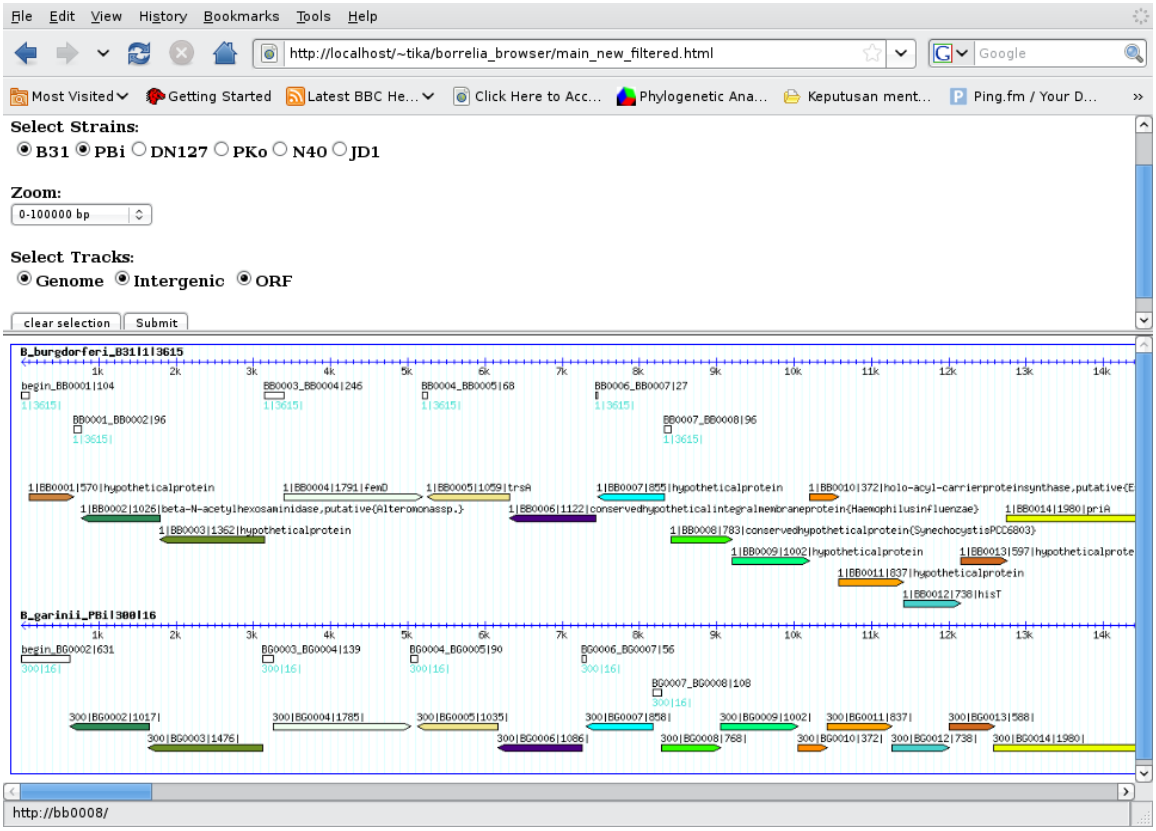


FIG. 7. Intergenic DNA Matrix

	A	B	C	D	G	H	K	M	R	S	T	U	V	W	Y	Q	
A	1	0	0	1	0	1	0	1	1	0	0	0	1	1	0	0	
B	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	
C	0	1	1	0	0	1	0	1	0	1	0	0	1	0	1	0	
D	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0	
G	0	1	0	1	1	0	1	0	1	1	0	0	1	0	0	0	
H	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	
K	0	1	0	1	1	1	1	0	1	1	1	1	1	1	1	0	
M	1	1	1	1	0	1	0	1	1	1	0	0	1	1	1	0	
R	1	1	0	1	1	1	1	1	1	1	0	0	1	1	0	0	
S	0	1	1	1	1	1	1	1	1	1	0	0	1	0	1	0	
T	0	1	0	1	0	1	1	0	0	0	1	1	0	1	1	0	
U	0	1	0	1	0	1	1	0	0	0	1	1	0	1	1	0	
V	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	0	
W	1	1	0	1	0	1	1	1	1	0	1	1	1	1	1	0	
Y	0	1	1	1	0	1	1	1	0	1	1	1	1	1	1	0	
Q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	50 or 100	

FIG. 8. Rate Ratios Distribution of intergenic (black), nonsynonymous (blue) and synonymous (red) sites categorized according to published Sigma-54 promoter transcriptional regulation activity. Solid lines are rate ratios for total sample, dashed lines are those with altered transcriptional activity, while dotted lines are those with altered transcriptional activity plus presence of predicted Sigma-54 promoter binding site.

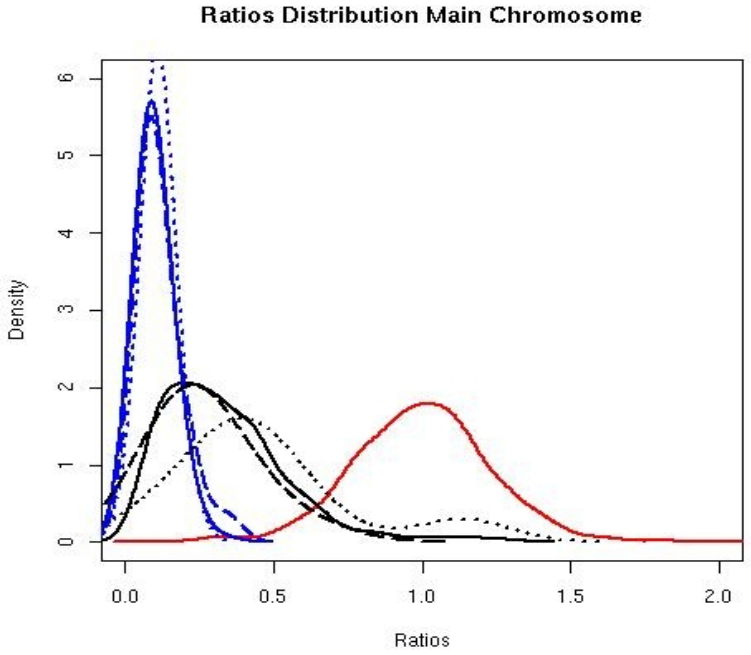


FIG. 9. Weight Matrix Sigma-54

Sigma 54 Matrix:
Weight matrix based on 186 characterized -24/-12-type promoters (Dombrecht et. al. 2002)

	T	G	G*-24	C	A	C	G	N	N	N	N	T	T	G	C*	W
A	12	2	0	12	139	11	55	51	46	44	38	13	4	1	9	76
C	14	0	0	147	23	122	17	48	64	42	62	22	18	2	173	5
G	10	184	186	6	18	10	103	69	36	35	43	15	10	181	1	17
T	150	0	0	21	6	43	11	18	40	65	43	136	154	2	3	88

FIG. 10. Intergenic Phylogenetic Footprinting. An alignment of intergenic sequence flanked by its upstream and downstream ORF. In this method, fragments of overlapping intergenic fragments are concatenated by their upstream and downstream fourfold synonymous sites, which are then analyzed by the PAML software to calculate substitution rate deviations from fourfold rates.

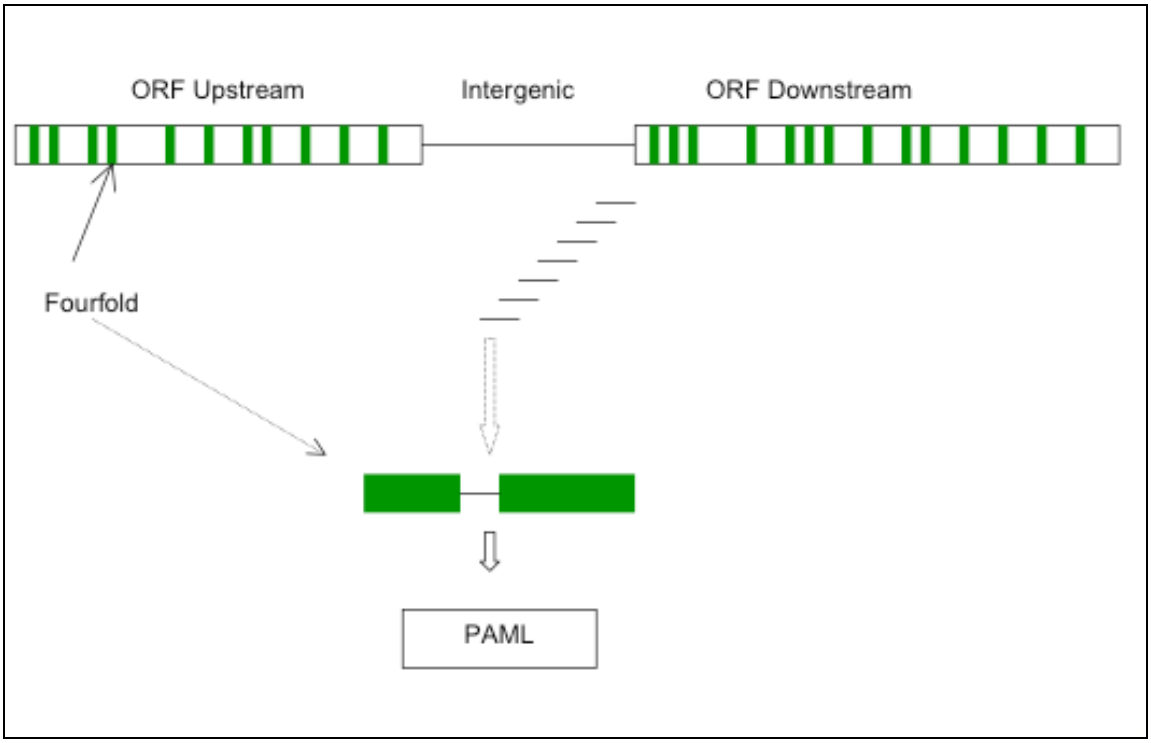
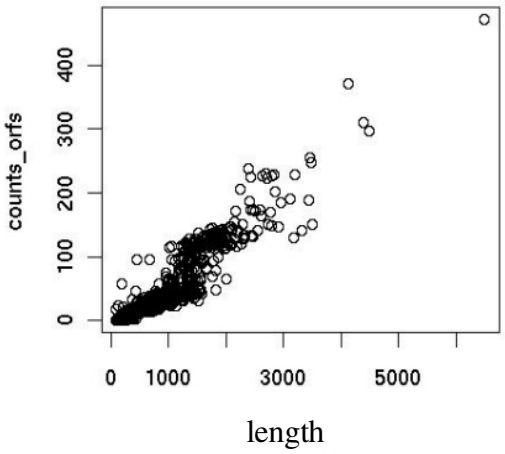


FIG. 11. Sigma-54 frequency per length in ORFs (A) versus Intergenics (B).

(A) ORFS



(B) Intergenics

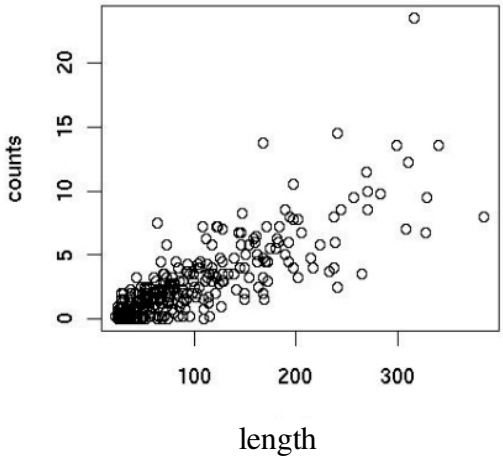


FIG. 12. Comparison of nucleotide substitution rates using the JC69 (I0), HKY85 (I4) and REV (I7) model as range of increasingly complex models, from baseml nucleotide substitution program of the PAML software using the same intergenic data set (between species genome data).

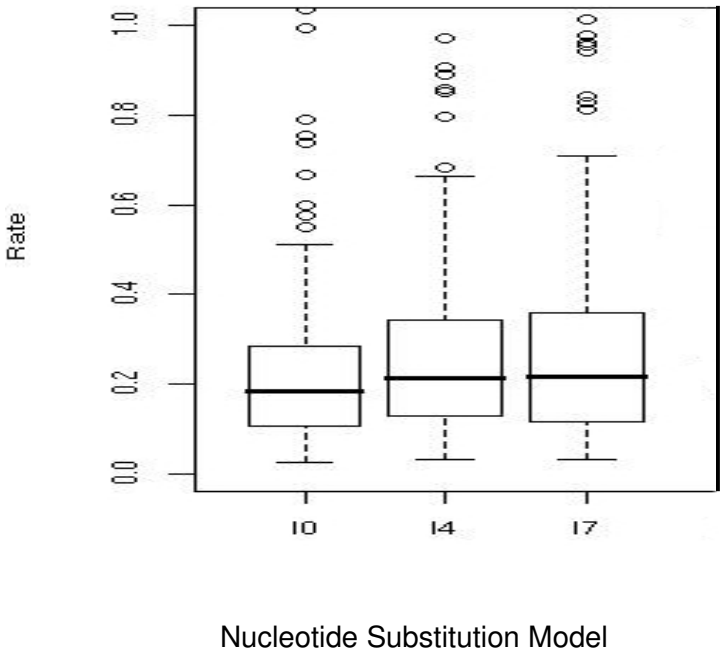


FIG. 13. Comparing nucleotide substitution rates using baseml and codeml program. Rate categories using the baseml program are dF (fourfold rates), dI (intergenic rates) and dN (nonsynonymous rates). Rate categories using the codeml program are dSc (synonymous rates) and dNc (nonsynonymous rates).

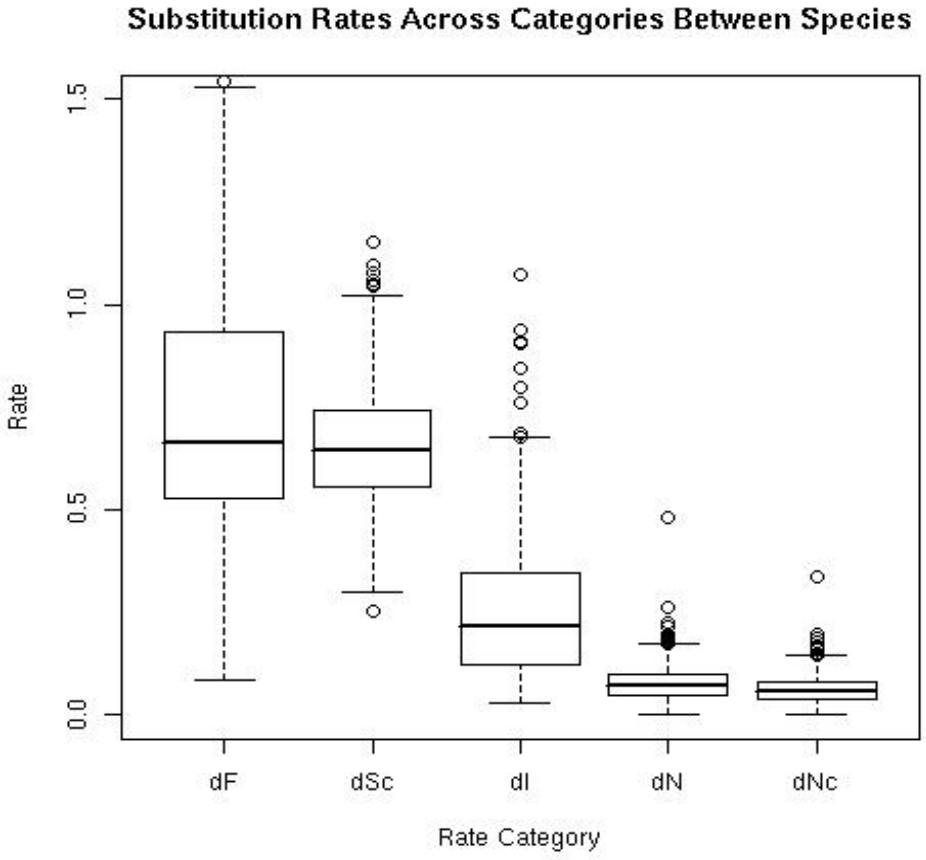
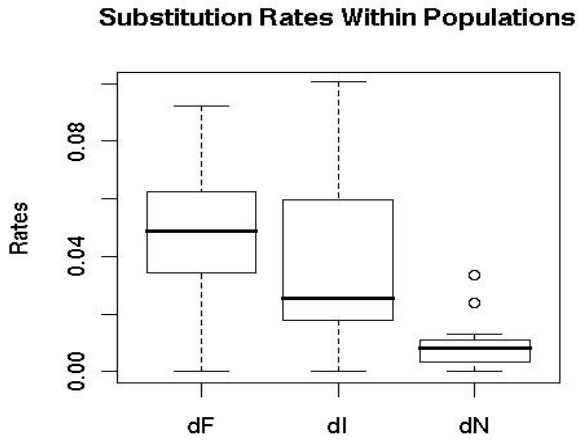
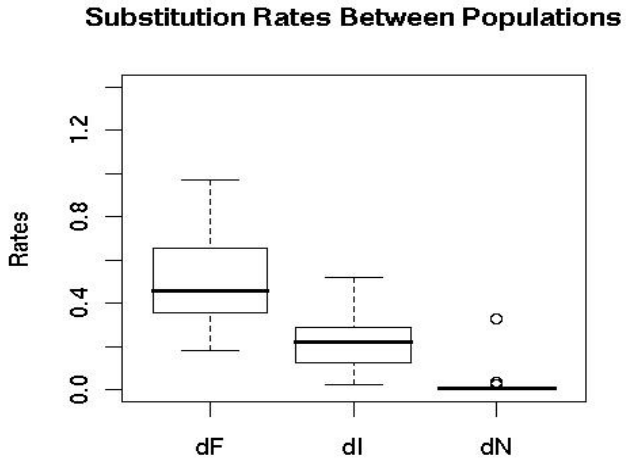


FIG. 14. Substitution rates differences between categories. Rates are designated as the following: dF for fourfold synonymous rate, dN for nonsynonymous rate and dI for intergenic rate. Main Chromosome Substitution Rates in: (A) within species plot, n=811 for dF, n=812 for dN, n=316 from dI and (B) between species plot n=716 for dF, n=716 for dN, n=334 for dI. Plasmid cp26 Substitution Rates in: (C) within species plot n=25 for dF, n=25 for dN, n=15 for dI and (D) between species plot, n= 22 for dF, n= 22 for dN, n= 15 from dI.

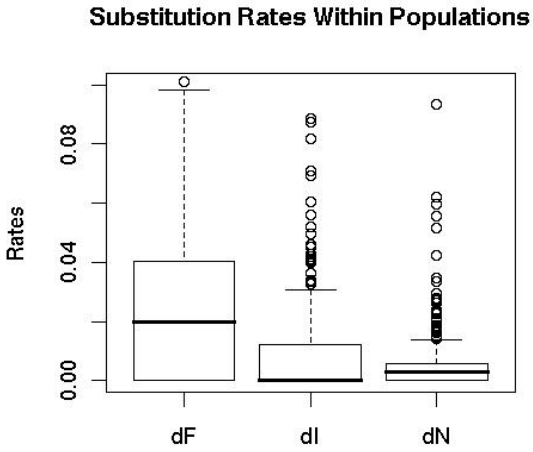
A.



B.



C.


Substitution Rates Within Populations
Rates
0.00
0.04
0.08
dF
dI
dN

D.

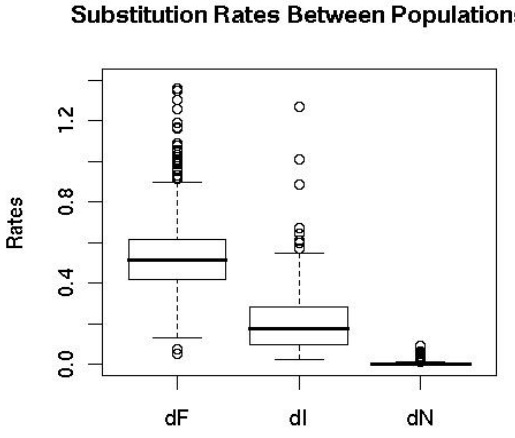
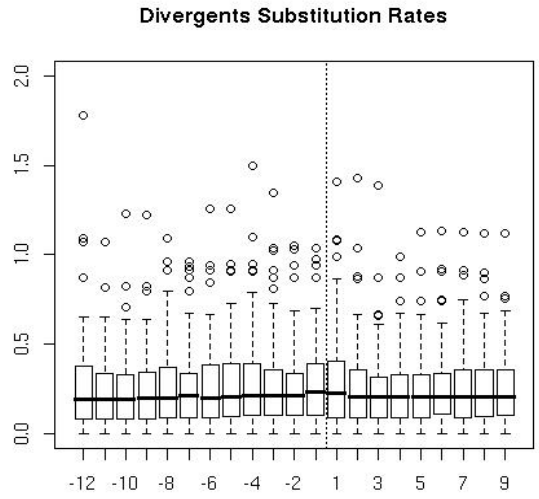

Substitution Rates Between Populations
Rates
0.0
0.4
0.8
1.2
dF
dI
dN

FIG. 15. Rates of substitutions per 25 bp fragment for (A) divergent compared to (B) convergent sites in main chromosome. Rates are calculated as # of substitution per site divided by downstream ORF fourfold rates. Each boxplot is range of substitution rates taken from intergenic sites for each fragment. Dotted line indicates intergenic-ORF boundary.

A.



B.

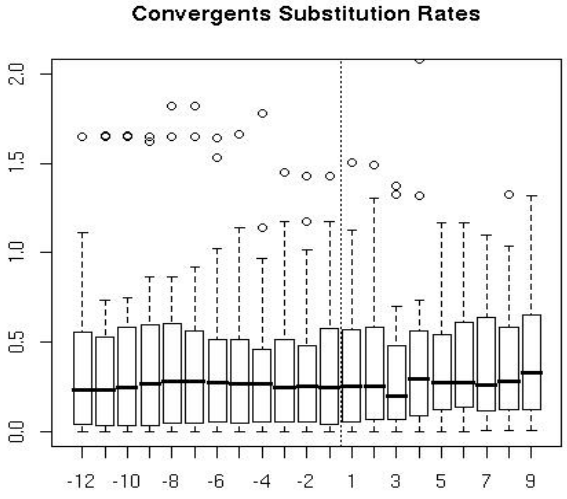
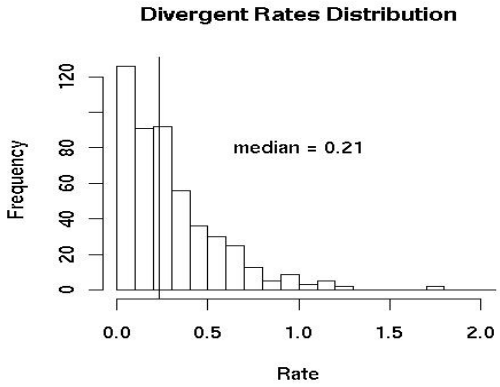


FIG. 16. Distribution of (A). Divergent vs. (B) Convergent different at P=0.007 (Two Sample Wilcoxon test or Mann-Whitney test). N = 500.

A.



B.

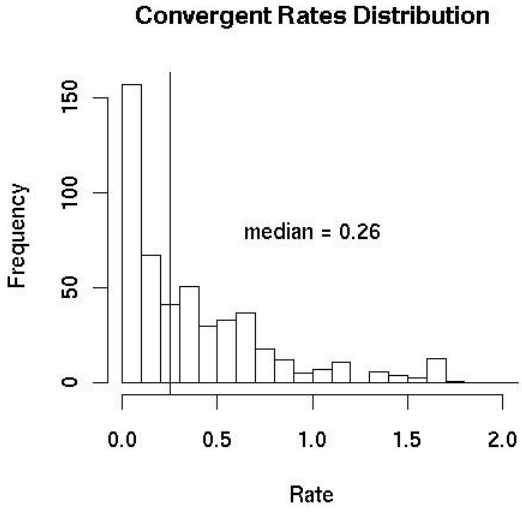
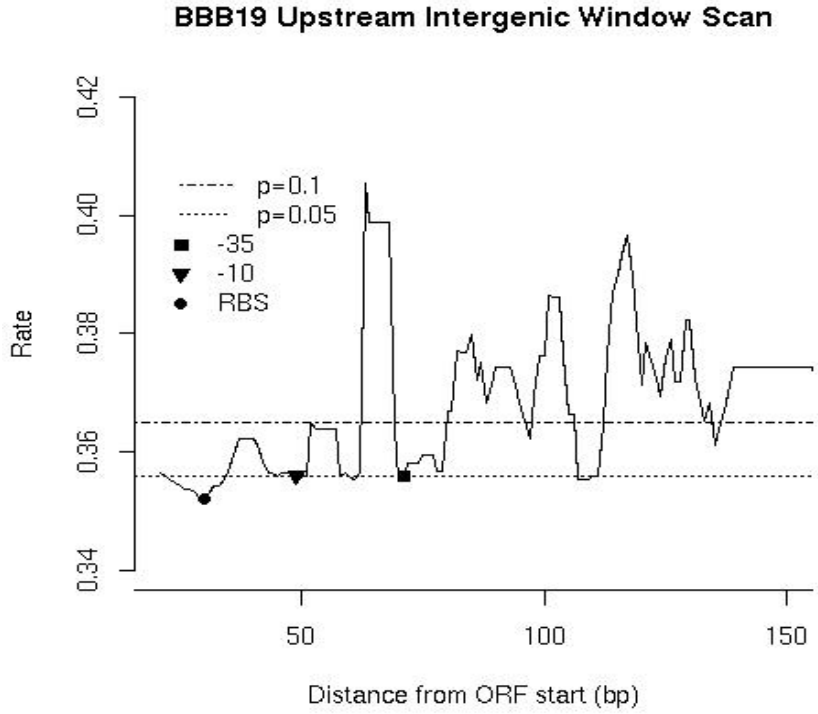


FIG. 17. Phylogenetic Footprinting of Sigma-S promoter binding sites in intergenics upstream of (A) cp26 BBB19 (ospC) and (B) Main chromosome BB0680 (mcp-4 methyl accepting chemotaxis protein) ORFs.

A.



B.

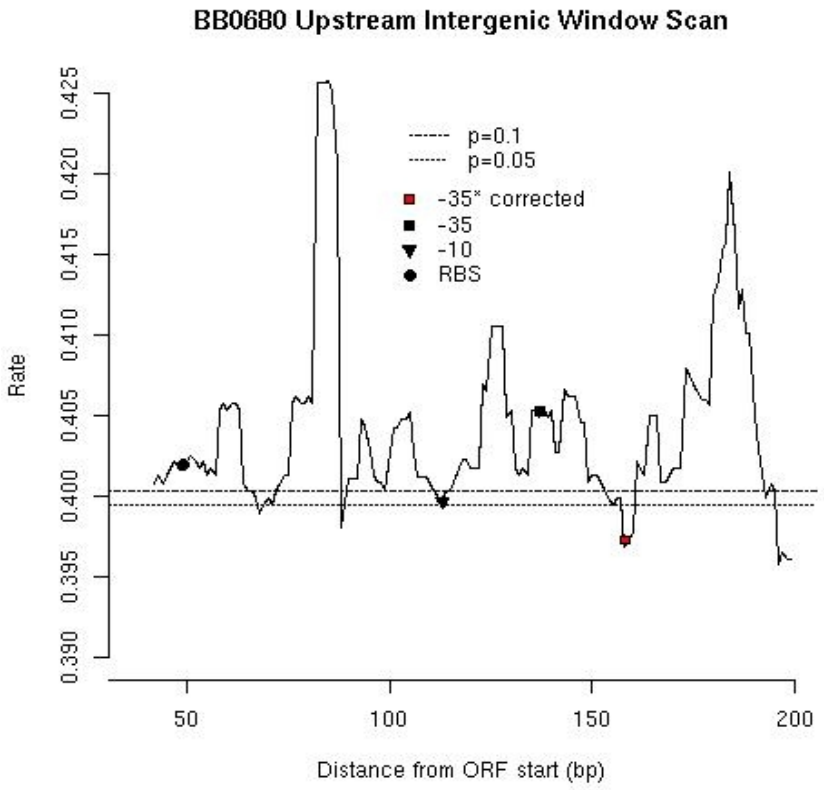

BB0680 Upstream Intergenic Window Scan
p=0.1
p=0.05
-35* corrected
-35
-10
RBS
Rate
0.390
0.395
0.400
0.405
0.410
0.415
0.420
0.425
50
100
150
200
Distance from ORF start (bp)

FIG. 18. Sigma-S-dependent Bb promoters for (A) cp26 BBB19 (ospC) and (B) Main Chromosome BB0680 (mcp-4 methyl accepting chemotaxis protein

A.

```
CLUSTAL W(1.81) multiple sequence alignment

Borrelia burgdorferi JD1 TATQTTAATTTTAGCATATTTGGCTTTGCTTATGTCGATTTTAAAATCAAATTAA-----
Borrelia burgdorferi N40 CATQTTAATTTTAACATATTTAGCTTTGCTTATGCTGATTTTAAAATCAAATTAA-----
Borrelia burgdorferi 297 CATQTTAATTTTAGCATATTTGGCTTTGCTTATGTCGATTTTAAAATCAAATTAA-----
Borrelia burgdorferi B31 CATQTTAATTTTAGCATATTTGGCTTTGCTTATGTCGATTTTAAAATCAAATTAA-----
Borrelia bissetti DN127  CACQCTAATTTTAGCATATTTAGCCTTGCTTATGTCGATTTTAAAATTAAATTAA-----
Borrelia garinii PBi     CATQTTAATTCTAGCATATTTAGCTTTGCTTCTGTCAATCTTAAAATCAATTAAGATAGT
                          *    ***** **.*******.** ******.**  .** ******* **:*:*.

Borrelia burgdorferi JD1 -----TACAATATTTTTCAA--ATTCTTCAATATTTAT----------------------
Borrelia burgdorferi N40 -----TACAATATTTTTCAA--ATTCTTCAATATTTAT----------------------
Borrelia burgdorferi 297 -----GACAATATTTTTCAA--ATTCTTCAATATCTTG---------------------
Borrelia burgdorferi B31 -----GACAATATTTTTCAA--ATTCTTCAATATTTAT----------------------
Borrelia bissetti DN127  -----GATAAT-TTTTTCAA--ATTATTCAATAGTTATG--------ATTGAATAATTTT
Borrelia garinii PBi     TTGTTTGTAATATTTTTCAATAATTATTCAATAATTATTCAATAATTATTCAATAATTAT
                               . *** ********  ***.*******  *:

Borrelia burgdorferi JD1 TCAAGA--TATTGA-----AGAATTTGAAAAAATTATTTT---------TTCAAATAAAA
Borrelia burgdorferi N40 TCAATA--TATTGA-----AGAATTTGAAAAAATAATTTT---------TTCAAATAAAA
Borrelia burgdorferi 297 AATAAA--TATTGA-----AGAATTTGAAAAAAT-ATTTT---------TTCAAATAAAA
Borrelia burgdorferi B31 TCAAGA--TATTGA-----AGAATTTGAAAAAATTATTTT---------TTCAAATAAAA
Borrelia bissetti DN127  TCAATA--AATTGACTTGAAATATTTGAAAAATTATTTTTCAAATATTTTTCAAATAAAA
Borrelia garinii PBi     TCAATAATTATTCAATAATTAAATTTGAAAAATTATTTTT---------TTTATATAAAA
                         :.:* *  :*** *     :.:**********:* :****         ** *:******
                           -35                -10              RBS
Borrelia burgdorferi JD1 AATTGAAAAACAAAATTGTTGGACTAATAATTCATAAATAAAAAGGAGGCACAAATTAQA
Borrelia burgdorferi N40 AATTGAAAAACAAAATTGTTGGACTAATAATTCATAAATAAAAAGGAGGCACAAATTAQA
Borrelia burgdorferi 297 AATTGAAAAACAAAATTGTTGGACTAATAATTCATAAATAAAAAGGAGGCACAAATTAQA
Borrelia burgdorferi B31 AATTGAAAAACAAAATTGTTGGACTAATAATTCATAAATAAAAAGGAGGCACAAATTAQA
Borrelia bissetti DN127  A-TTGAAAAGTAAAATTGTTGGACTAATAATTCATAAATAAAAAGGAGGCACAAATTAQA
Borrelia garinii PBi     AATTGAAAAGAAAAATTGTTGAACTAATAATTCATA--TAAAAAGGAGGCACAAATTAQA
                         * *******. **********.**************  ********************  *


Borrelia burgdorferi JD1 TG
Borrelia burgdorferi N40 TG
Borrelia burgdorferi 297 TG
Borrelia burgdorferi B31 TG
Borrelia bissetti DN127  TG
Borrelia garinii PBi     TG
                         **
```

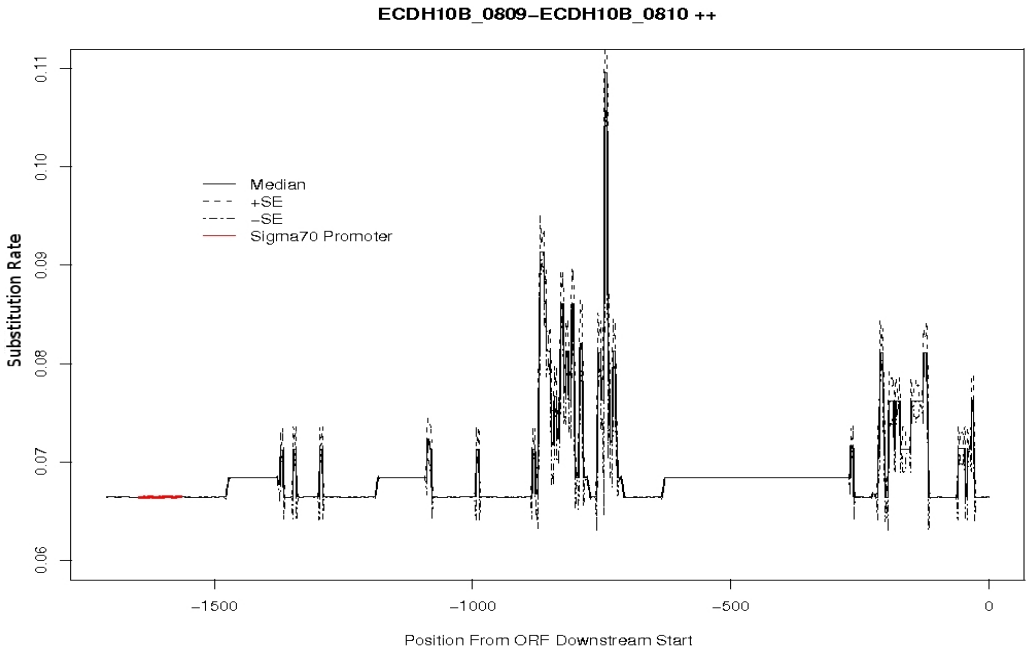
B.

```
CLUSTAL W(1.81) multiple sequence alignment

                                                                    -35*
Borrelia burgdorferi JD1 ---QTAGCAGTTAATAATAATAGTTTATAATTTGTTTTTCATGTTGTGAAAATTCTTTTA
Borrelia burgdorferi N40 ---QTAGCAGTTAATAATAATAGTTTATAATTTGTTTTTCATGTTGTGAAAATTCTTTTA
Borrelia burgdorferi B31 ---QTAGCAGTTAATAATAATAGTTTATAATTTGTTTTTCATGTTGTGAAAATTCTTTTA
Borrelia afzelii 400     ---QTAGTAGTTAATA---ATGGTCTATAATTTGTTTTTCATATTTTGAAAATTCTTTTA
Borrelia garinii 300     ---QTAGCAGTTAATAT--GTAGCTTATAACTTGTTTTTCATATTGTGAAAATTCTTTTA
                            *    ********    * *  ***** *********** ** **************
                              -35                      -10
Borrelia burgdorferi JD1 TT-TTTTTGAGAGTTTTTGATTTTTTGCTGTAAATTTTTTTTTGATACCCTTT--GGTCT
Borrelia burgdorferi N40 TT-TTTTTGAGAGTTTTTGATTTTTTGCTGTAAATTTTTTTTTGATACCCTTT--GGTCT
Borrelia burgdorferi B31 TT-TTTTTGAGAGTTTTTGATTTTTTGCTGTAAATTTTTTTTTGATACCCTTT--GGTCT
Borrelia afzelii 400     AT-TTTTTAAGAGTTTTTAAATTTTTGCTGTAAATTTTTTTTTGATACCCTTTTTGATCT
Borrelia garinii 300     ATATTTTTAAGAGTTTTTAATTTTTTGCTGTAAATTTTTTTT-AATACCCTTTT-GGTCT
                          * ***** ********* * *********************  *********  * ***

                                                          RBS
Borrelia burgdorferi JD1 TGAGTTTATTTGATTAAATAAGTAGGTGATTTGTGAGGTAGTTTATTQATG
Borrelia burgdorferi N40 TGAGTTTATTTGATTAAATAAGTAGGTGATTTGTGAGGTAGTTTATTQATG
Borrelia burgdorferi B31 TGAGTTTATTTGATTAAATAAGTAGGTGATTTGTGAGGTAGTTTATTQATG
Borrelia afzelii 400     TTAGTTTATTTGATTAAATAAGTAGATGATTTGTGAGGTAGTTTATTQATG
Borrelia garinii 300     TAAGTTTGTTTGATTAAATAAGTAGATGATTTGTGAGGTAGTTTATTQATG
                         * ***** ***************** ************************
```

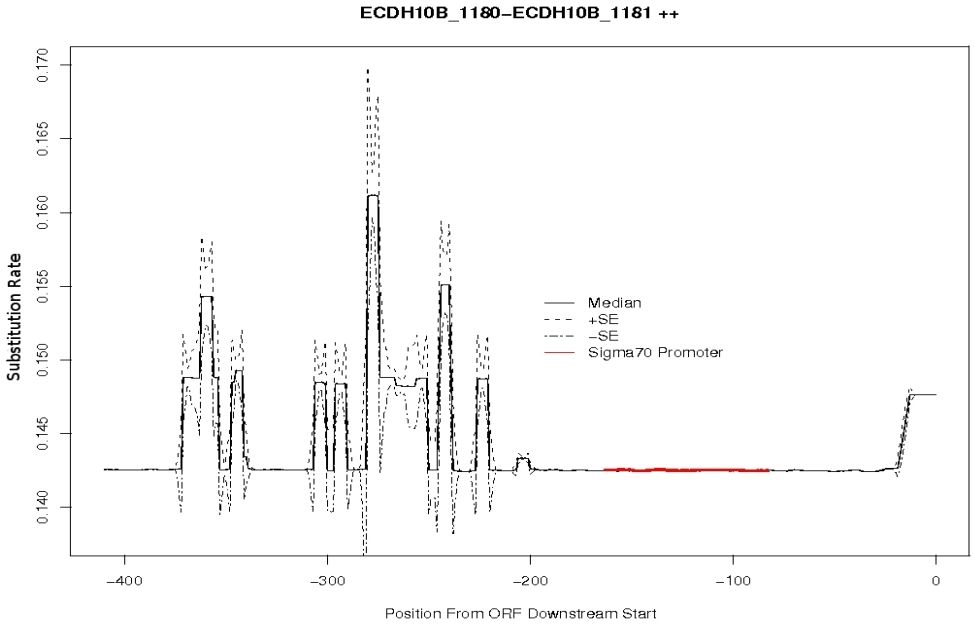
Note.— *corrected -35 site.

FIG. 19. Example Intergenic Phylogenetic Footprinting in (A) ECDH10B_0810 (B) ECDH10B_1181 (NADH Dehydrogenase) and (C) ECDH10B_0393 *E. Coli* Sigma-70 Promoters.

A.



B.


ECDH10B_1180–ECDH10B_1181 ++
Substitution Rate
0.140
0.145
0.150
0.155
0.160
0.165
0.170
Median
+SE
–SE
Sigma70 Promoter
–400
–300
–200
–100
0
Position From ORF Downstream Start

C.

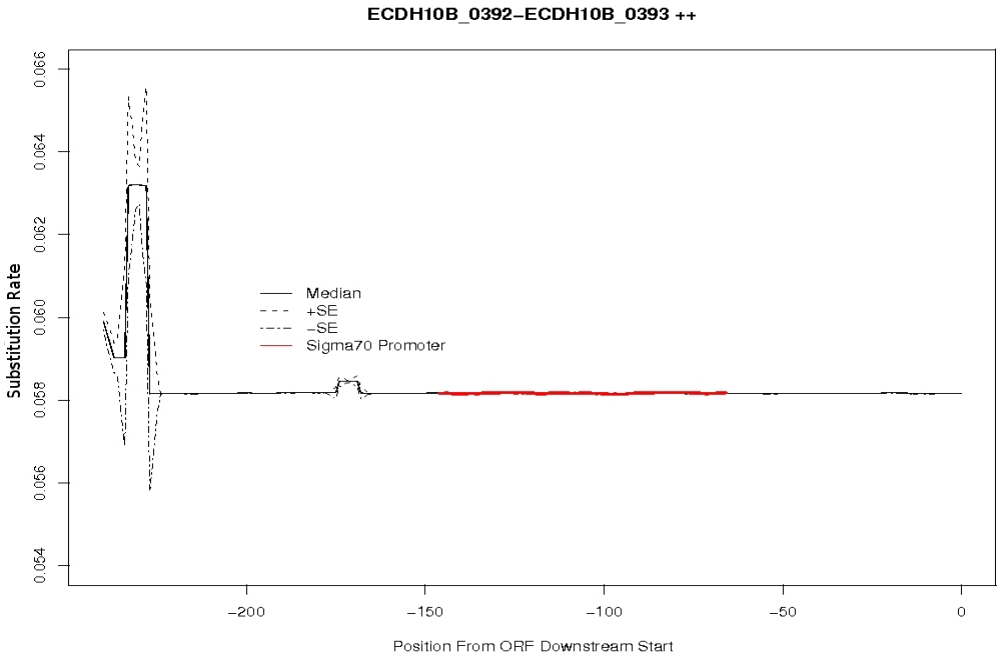
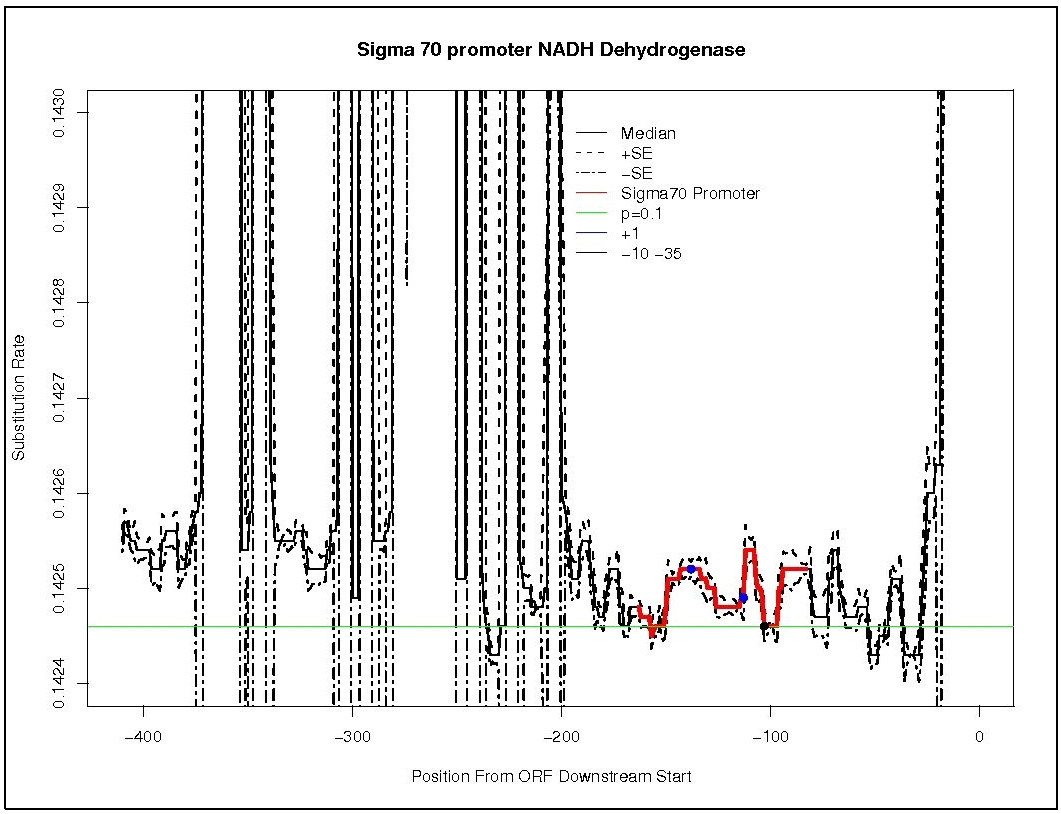

ECDH10B_0392–ECDH10B_0393 ++
Substitution Rate
0.054
0.056
0.058
0.060
0.062
0.064
0.066
Median
+SE
–SE
Sigma70 Promoter
–200
–150
–100
–50
0
Position From ORF Downstream Start

FIG. 20. Closeup of ECDH10B_1180 (NADH Dehydrogenase) promoter phylogenetic footprinting result.



-35 -10 +1

aacttgattaacatcaattttggtat**t**gaccaatgcaccattcatgttatt**c**tcaatagcg**A**agaacattttcattgctgta

FIG. 21. Percentage of of pallindromes that are constrained compared to those that are not constrained.

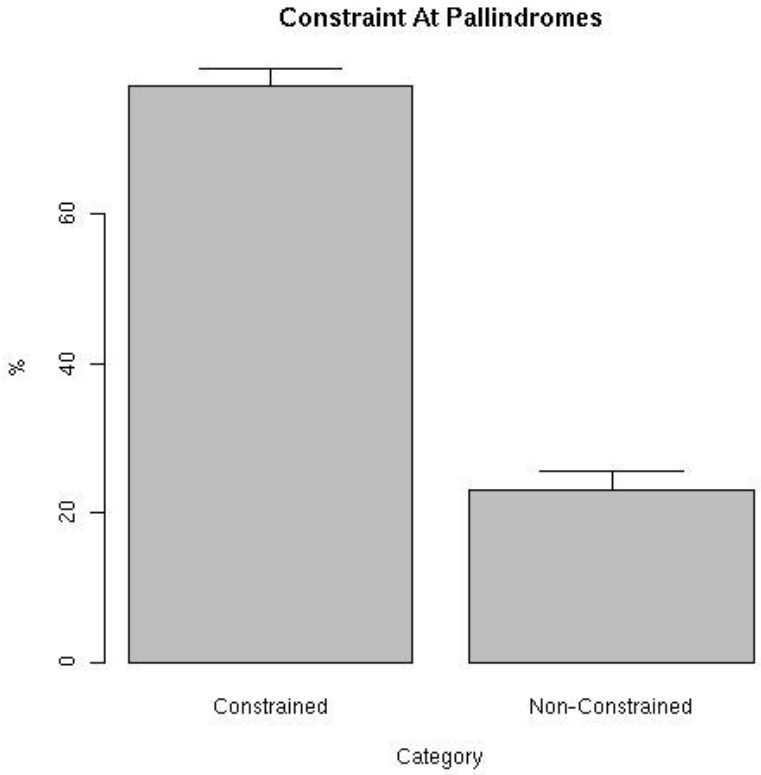


FIG. 22. Percentage of Constraint at AT Rich vs AT poor sites.

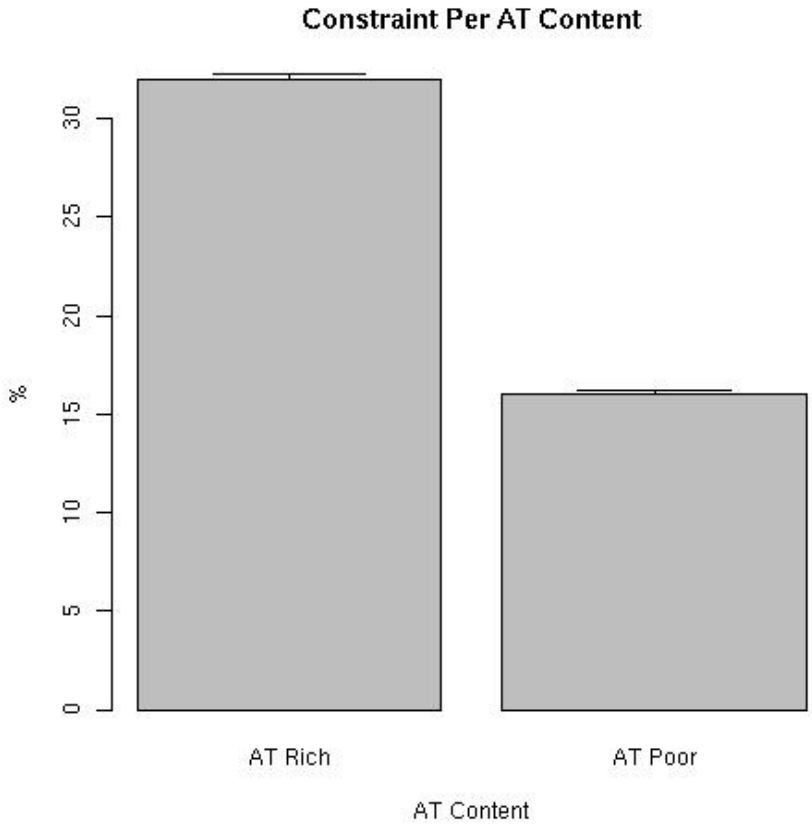


FIG. 23. Percent of constrained sites at divergent (←→-+), convergent (→←+-) and unidirectional (←←-- or →→++) integenic type, where each intergenic type correspond to a particular genome locality oriented according to upstream and downstream flanking ORF direction.

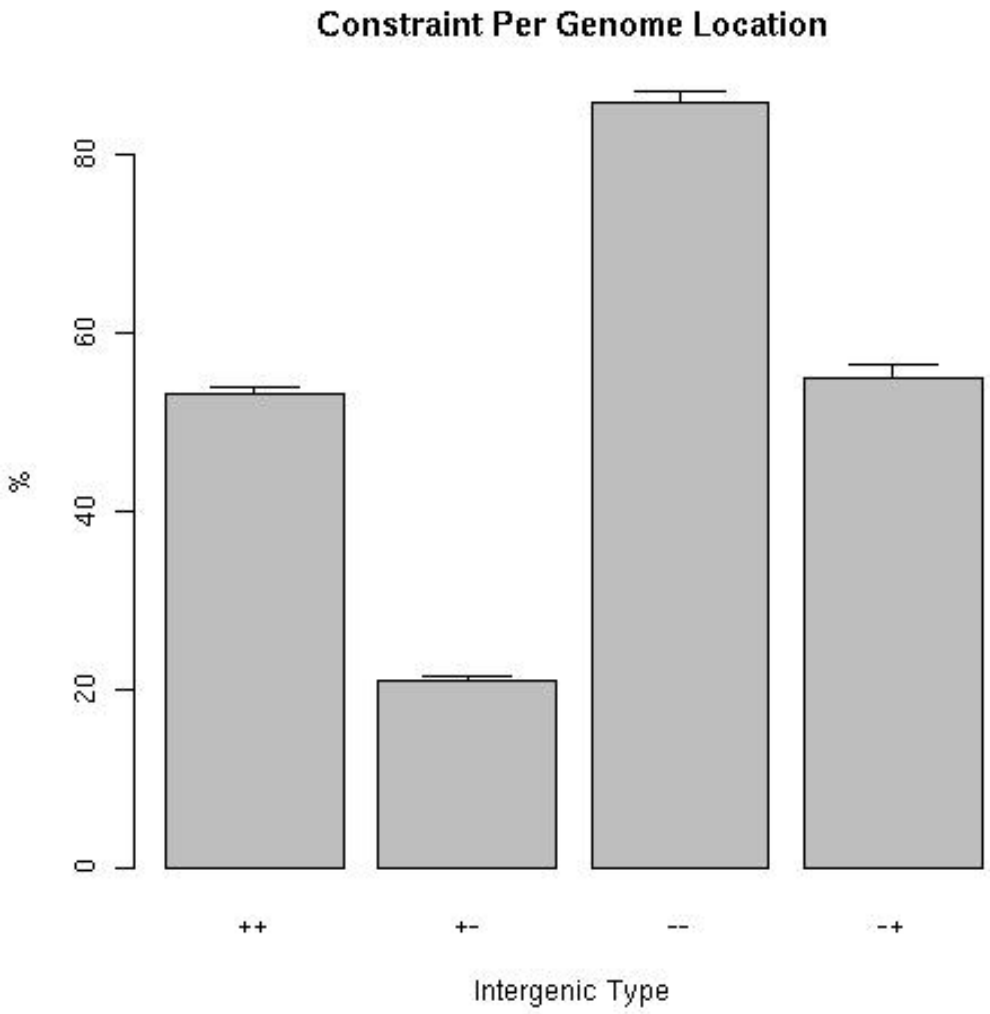


Table 1
***Borrelia* Genome Collection**

Molecule	Species	Strain
Main Chromosome Within Species Collection	*Borrelia burgdorferi sensu stricto*	B31, JD1, N40
Main Chromosome Between Species Collection	*Borrelia burgdorferi sensu stricto*	B31
	Borrelia garinii	PBi
	Borrelia afzelii	PKo
	Borrelia bissetti	DN127
Plasmid cp26 Within Species Collection	*Borrelia burgdorferi sensu stricto*	B31, JD1, N40, 297
Plasmid cp26 Between Species Collection	*Borrelia burgdorferi sensu stricto*	B31
	Borrelia garnii	PBi
	Borrelia bissetti	DN127

Note.—Total genome size for main chromosome ~ 910kbp, for plasmid cp26 ~ 26kbp.

Table 2

Molecular Evolution Models Used in This Study Using the PAML Software

Model	Type (PAML program)	Base Frequencies	Substitution Rates
JC69 (Jukes-Cantor)	Nucleotide Substitution Model (baseml)	Equal	All substitutions equally likely
F81 (Felsenstein)	Nucleotide Substitution Model (baseml)	Unequal	All substitutions equally likely
HKY85 (Hasegawa et. al.)	Nucleotide Substitution Model (baseml)	Unequal base frequencies	Transversions and transitions have different substitution rates
REV (General Reversible)	Nucleotide Substitution Model (baseml)	Unequal base frequencies	All six pairs of substitutions have different rates
2 or more dN/dS ratios for branches	Codon Substitution Model (ORF only)	-	Free dN/dS ratios for each branch

Table 3

Pearson Correlation of substitution rates of intergenic (dI), fourfold syononymous (dF) and nonsyononymous (dN).

	dF	dN	dI*
dF	-	**cor = 0.13** **p = 0.24**	**cor = -0.01** **p = 0.90**
dN	-	-	**cor = -0.02** **p = 0.89**
dI	-	-	-

Note. —dI = (dI_u + dI_d)/2, where dI_u is upstream intergenic and dI_d is downstream intergenic from ORF analyzed.

Table 4

Proportion of Constrained Intergenics (Chi-square test p-value = 0.0004998)

	dI/dF > 1	dI/dF < 1
Intergenics	43 out of 268 (19%)	218 out of 268 (81%)
Randomized Fourfold	124 out of 268 (46%)	144 out of 268 (60%)

Note. —Intergenic data set n = 268. Fourfold data set adjusted through random sampling.

Table 5

Intergenic Within and Between Species Rates Ratios (K)

P=0.016 (Chisq.test)	Kwithin > Kbetween	Kwithin < Kbetween
Intergenics	95	**173**
Expected	124	144

Note. —Expected from simulated fourfold only rates

Table 6

Nonsynonymous Within and Between Species Rates Ratios (K)

P=0.004 (Chisq.test)	Kwithin > Kbetween	Kwithin < Kbetween
Nonsynonymous	**414**	301
Expected	358	357

Note. —Expected from simulated fourfold only rates

Table 7
2X2 Contingency Test Intergenics

P<0.0005 (Chisq.test)	Within	Between
I	309	8967
F	1270	23658
I/F	0.24	0.38

Table 8
2X2 Contingency Test Nonsynonymous

P<0.0005 (Chisq.test)	Within	Between
N	302	3891
F	1270	23658
N/F	0.24	0.16

APPENDIX

PERL Scripts stored in: intergenic_analysis_clean.tar.gz
(includeds additional accesory files for PAML and PHYLIP runs
(ie. baseml.ctl, matrix_IUB_intergenic, input etc.)

Databases stored in: intergnic.out.gz, bb2.out.gz

1. Inputs
filename

molecule:
E. coli: main
Borrelia: main, lp54, cp26

strain:
E. coli:K12, O157, APEC
Borrelia: JD1, N40, PKo, PBi, B31

table-id :
ecoli, darwin_new, bb2 etc.

database: intergenic, bb2

reference-strain:
Borrelia: B31
E. coli: K12

query-strain:
Borrelia: JD1,N40
E. coli: O157, APEC

2. Tables

2.1.*orf-table*
E. coli: orf_ecoli
Borrelia: orf_darwin_new

Column	Type	Modifiers
genome_id	text	
mol_id	text	
start	integer	
stop	integer	

orf_id	text
strand	text
annotation	text
exclude	boolean

2.2. *contig-table*
E. coli: contig_darwin_new
Borrelia: contig_ecoli

Column	Type	Modifiers
genome_id	smallint	
mol_id	smallint	
mol_name	text	
seq	character varying(9000000)	

2.3. *blast-data-table*
E. coli: blastp_ecoli
Borrelia: blastp_linkagejr_new

Column	Type	Modifiers
q_genome_id	text	
q_mol_id	text	
q_mol_name	text	
q_orf_start	integer	
q_orf_end	integer	
q_orf_id	text	
q_strand	text	
s_genome_id	text	
s_mol_id	text	
s_mol_name	text	
s_orf_start	integer	
s_orf_end	integer	
s_orf_id	text	
s_strand	text	
percent_id	real	
align_len	integer	
mismatches	integer	
gaps	integer	
q_start	integer	
q_end	integer	

s_start	integer
s_end	integer
e_value	double precision
bit_score	real

2.4. *genome-data-table*
E. coli: genome_data_ecoli
Borrelia: genome_data_new, genome_data_bb2

Column	Type	Modifiers
molecule	text	
species	text	
strain	text	
genome_id	text	
mol_id	text	
length	integer	
push_contig	integer	
browser_order	integer	
flip	text	

2.5. *intergenic-table*
E. coli: intergenic_ecoli
Borrelia: intergenic_new, intergenic_bb2

2.6. *orth-contig-bb2-table*

2.7. *orth-orf-26spp-view-bb2-table*

Column	Type	Modifiers
genome_id	smallint	
mol_id	smallint	
description	text	
start	integer	
length	integer	
orf_upstrm	text	
orf_upstrm_strand	text	
orf_downstrm	text	
orf_downstrm_strand	text	
seq	character varying(2000000)	

2.8. *intergenic-table_filtered*

2.9. *top-score-blastp-table*
E. coli: top_score_blastp_ecoli
Borrelia: top_score_blastp_linkagejr_new

Column	Type	Modifiers
orth_id	text	
hit_genome_id	smallint	
hit_orf	text	
hit_mol_id	smallint	
query_genome_id	smallint	
query_orf	text	
query_mol_id	smallint	
e_value	double precision	

2.10. *browser-orf-table*
E. coli: browser_orf_ecoli and browser_orf_new_ecoli
Borrelia: browser_orf_new, browser_orf_bb2

Column	Type	Modifiers
genome_id	smallint	
mol_id	smallint	
orf_id	text	
start	integer	
stop	integer	
strand	text	
color	text	
orf_length	integer	
contig_length	integer	
exclude	boolean	
annotation	text	
orth_id	text	

2.11. *browser-intergenic-table*
E. coli: browser_intergenic_ecoli
Borrelia: browser_intergenic_new,
browser_intergenic_new_filtered, browser_intergenic_bb2

Column	Type	Modifiers
orf_upstrm	text	
orf_downstrm	text	
genome_id	smallint	
mol_id	smallint	
start	integer	
stop	integer	
length	integer	
exclude	boolean	
orf_upstrm_strand	text	
orf_downstrm_strand	text	
seq_300bp_included	character varying(5000000)	

2.12. *window-anal-table*
E. coli : Window_anal_ecoli

Column	Type	Modifiers
orf_id	text	
num	text	
start	integer	
stop	integer	
strand	text	
tree_length	real	

2.13. *orth-orf-table*
Borrelia: orth_orf
E. coli: orth_orf_ecoli

2.14. Schema for E. coli Base Composition and Promoter Analysis

i. ecoli_gc

Column	Type	Modifiers
orf_upstrm	text	
orf_downstrm	text	
num	integer	

tree_length	real
genome_id	integer
start	integer
stop	integer
seq	text
gc_percent	real

ii. ecoli_operon_door

Column	Type	Modifiers
operon_id	integer	
gi	integer	
orf_id	text	
start	integer	
stop	integer	
strand	text	
length	integer	
cog_number	text	
product	text	

iii. ecoli_pallindrome

Column	Type	Modifiers
orf_upstrm	text	
orf_downstrm	text	
num	integer	
genome_id	integer	
start	integer	
stop	integer	
seq	text	
tree_length	real	

iv. ecoli_sigma70_regulondb

Column	Type	Modifiers
promoter_identifier	text	
promoter_name	text	
dna_strand	text	
transc_start	integer	
sigma_factor	text	
promoter_seq	text	
evidence	text	

v. ecoli_sigmatotal_regulondb

Column	Type	Modifiers
promoter_identifier	text	
promoter_name	text	
dna_strand	text	
transc_start	integer	
sigma_factor	text	
promoter_seq	text	
evidence	text	

vi. tf_binding_site_ecoli

Column	Type	Modifiers
tf_bs_left	integer	
tf_bs_right	integer	
tf_bs_name	text	
strand	text	
tf_name	text	
binding_site	text	

vii. tf_binding_site_ecoli_swissregulon (or tf_binding_site_ecoli_swissregulon2)

Column	Type	Modifiers
start	integer	
stop	integer	
strand	text	
tf	text	

viii. window_anal_ecoli

Column	Type	Modifiers
orf_upstrm	text	
orf_downstrm	text	
num	integer	
start	integer	
stop	integer	
orf_upstrm_strand	text	
orf_downstrm_strand	text	

tree_length	real	

ix. window_anal_ecoli_annotated

Column	Type	Modifiers
orf_upstrm	text	
orf_downstrm	text	
orf_upstrm_strand	text	
orf_downstrm_strand	text	
start	integer	
stop	integer	
num	integer	
fourfold_rate	real	
int_rate	real	
tf_start	integer	
tf_stop	integer	

x. window_anal_ecoli_annotated_neut

Column	Type	Modifiers
orf_upstrm	text	
orf_downstrm	text	
orf_upstrm_strand	text	
orf_downstrm_strand	text	
start	integer	
stop	integer	
num	integer	
fourfold_rate	real	
int_rate	real	
tf_start	integer	
tf_stop	integer	

xi. window_anal_ecoli_annotated_regulondb

Column	Type	Modifiers
orf_upstrm	text	
orf_downstrm	text	
orf_upstrm_strand	text	
orf_downstrm_strand	text	

start	integer	
stop	integer	
num	integer	
fourfold_rate	real	
int_rate	real	
tf_start	integer	
tf_stop	integer	

xii. window_anal_ecoli_neutral

Column	Type	Modifiers
orf_upstrm	text	
orf_downstrm	text	
num	integer	
start	integer	
stop	integer	
orf_upstrm_strand	text	
orf_downstrm_strand	text	
tree_length	real	

xiii. window_anal_ecoli_neutral_flankonly

Column	Type	Modifiers
orf_upstrm	text	
orf_downstrm	text	
num	integer	
start	integer	
stop	integer	
orf_upstrm_strand	text	
orf_downstrm_strand	text	
tree_length	real	

xiv. blastp_ecoli_promoter

Column	Type	Modifiers
promoter_identifier	text	
genome_ref	text	
prom_start	integer	

```
prom_stop           | integer           |
genome_start        | integer           |
genome_stop         | integer           |
e_value             | double precision  |
```

3. Creating Orthologous ORFS and Intergenic Database:

3.1. ORF extraction
Get genome data and put into *orf-table* and *contig-table* (shell scripting, manual). Make *genome-data-table*.

get_orf_sequence.pl *genom-id mol-id orf-table contig-table* > *strain*.fas

reverse_complement.pl *strain*.fas *strain*_fas_rev

translate_filter_outofframe5.pl *strain*.fas

translate_filter_outofframe5.pl *strain*.fas_rev

cat *strain*.fas.aa *strain*_fas_total.aa

cat *strain*.fas_rev.aa > *strain*_fas_total.aa

Create *reference-database*:
formatdb -i *strain*_fas_total.aa -p T -n *reference-database*

BLAST all strains:

blastall -p blastp -d *reference-strain* -i *query-strain*_fas_total.aa -e 0.001 -m 8 -o *reference-strain_query-strain*.blastout_8

cat *reference-strain_query-strain*.blastout_8 | grep "\w" | sed 's/\t/|/g' | sed 's/_r//g' > *reference-strain_query-strain*.blastout_8_todb

cat *blastout_8_todb > blastout_8_todb_total

psql intergenic

\copy blastp_*table-id* from 'blastout_8_todb_total' using delimiters '|'

\q
get_top_score_blastp5_new.pl *reference_genome_id genome_id mol_id table_id*

mv top_score_blastp_tmp top_score_blastp_<table_id>

3.2. Intergenic Extraction

i. Get intergenic (as bulk), put into intergenic table:

Run script in directory *dir*

Files needed:

-orf fas files: *filename*.fas
-input tables: *genome-data-table*, *contig-table*
-output put into table: *intergenic-table*

intergenic3_new_bulk.pl molecule

*strain*1_*strain2_strain3*... *dir*

psql intergenic

\copy *intergenic-table_table-id* from 'to_db' using delimiters '|'

update *orf-table* set exclude = 't' where stop - start < 250 (not necessary for E. coli dataset).

Require Input Tables:
top-score-blastp-table
intergenic-table
orf-table
genome-data-table

Output(s) put into tables:
browser-orf-table
browser-intergenic-table

Outfile: orf_outfile, intergenic_outfile modified for database prep as orf_outfile_edit, intergenic_outfile_edit

orthologous_total_table_batch2_new.pl *molecule table-id orf-table*

or for bb2 data: orthologous_match_darwin.pl *molecule table-id orth-orf-table orth-con-id*

psql intergenic

\copy *browser-orf-table* from 'orf_outfile_edit' using delimiters '|'

\copy *browser-intergenic-table* from 'intergenic_outfile_edit' using delimiters '|'

update *browser-intergenic-table* set exclude = 't' where length < 25;

update *browser-orf-table* set exclude = 'f' where length > 250;

#add same color to orthologous orfs in *browser-orf-table*.

 browser_orf_color_auto_new.pl *table_id*

ii. Local intergenic isolation based on ortholous id

Require Tables:
orf_table
orth_table
contig_table
gene_order table

extract-intergenic.pl

3.3. Visualization

create_browser_perlgd.pl *table-id molecule browser-name*

Outfiles:
filename.html
*filename*_stdin.cgi

mv *filename*.html to ~/public_html/borrelia_browser.html

mv *filename*_stdin.cgi to ~/public_html/cgi-bin/

```
chmod ugo+x filename_stdin.cgi

additional files (to make browser):
mv ../intergenic_analysis_clean/filename.cgi  ~/public_html/cgi-
bin/

browser store at ../public_html/borrelia_browser/filename.html

#Filter data

intergenic3_filter.pl molecule strain1_strain2_... database
contig_table

cat *to_db | sed 's/ //g' | grep '^\w' > to_db_final
psql intergenic

#create table intergenic-table_filtered

\copy intergenic-table_filtered from 'to_db_final' using
delimiters '|'

#choose orthologous contig id (orth-con-id) according to orth-
orf-table
orthologous_match_darwin_filter.pl molecule database
browser_table orth-con-id

Outfiles:
intergenic_outfile, orf_outfile modified for database as
intergenic_outfile_edit, orf_outfile_edit

psql intergenic

delete from browser-intergenic-table ;
delete from browser-orf-table;

\copy browser-intergenic-table from 'intergenic_outfile_edit'
using delimiters '|'
\copy browser-orf-table from 'orf_outfile_edit' using delimiters
'|'

update <browser intergenic table> set exclude = 't' where length
< 25;
update <browser_orf_table> set exclude = 'f' where orf_length >
250;
```

\q

browser_orf_color_auto.pl *table-id*

4. Alignments

4.1. Intergenics Alignment

Require Table:
match_orf_orth_new
match_orf_orth_new.pl

i. Get intergenics:

TOTAL GENOME extraction:
Borrelia:

\copy match_orf_orth_new from 'outfile' using delimiters '|'

align_intergenics_new.pl *genome_id mol_id*

#get matrix_IUB (used different matrix for window analysis e_coli where Q=100 not 50).

splice_alignment_23.pl

BULK from list of Orthologous ORF ids (*orth-id*) from *orth-orf-table*
E.coli:

~/scripts/intergenic_analysis_clean/bulk_perl.pl -l orth_orf_list_edit -s '~/scripts/intergenic_analysis_clean/extract-intergenic-ecoli.pl -c 1 -i R -d intergenic -a'

ls *intergenic_seq.aln > list

remove_Q_flank_from_intergenic.pl list
use *fas_aln file for window analysis

INDIVIDUAL
Borrelia :

extract-intergenic.pl -o *orth-id* -c 1 -i R -d *database* -a

E. coli:

extract-intergenic-ecoli.pl -o *orth-id* -c 1 -i R -d *database* -a

4.2. ORFs Alignment

Borrelia:
get_seq_from_browser_orf_2.pl *molecule*
number_of_genomes_within_species
number_of_genomes_between_species

input file for phylip run

ls *nt_aln > list

bulk_phylip_3_orfs_new.pl list

ls *phy > list

#change interleaved phylip format
to_interleaved_3.pl list

ls *outtree > list

reformat_phylip_4.pl list n

E. coli:
get_seq_from_browser_orf_2_edit.pl *molecule* *strand*

or

get_seq_from_browser_orf_2_edit2.pl *molecule*

ls *nt.aln >list

*need file 'input'

bulk_phylip_3_orfs_new_edit.pl list

```
ls *phy > list

to_interleaved_3.pl list

ls *outtree > list

reformat_phylip_4.pl list n
```

5. PAML Analysis

5.1. Nonsyn

```
cp *phyI
cp *ph
ls phyI > list
batch_get_nonsyn_pos.pl

ls *final .phy > list
edit_id.pl list
ls *final.phy.edit >list

#need to make empty file called "tree"

bulk_paml.pl list
```

5.2. ORFS FOURFOLD

Borrelia:

```
cp *within.nt.aln
cp *within.aa.aln
cp  *ph .
ls *within.nt.aln > list
get_fourfold_codons_2_thirdpos.pl list
number_of_genomes_aligned(ie.3 or 4 etc) > out

ls *phy > list
get_codons.pl
to_interleaved_3.pl list
#need baseml.ctl file
ls *phyI > list
bulk_paml.pl list
get_mlb_data_orfs.pl list > fourfold_mlb_data
```

E. coli:

```
cp ../*nt.aln .
cp ../*aa.aln .
cp ../phylip/*ph .

ls *aa.aln > list
get_fourfold_codons_2_thirdpos_edit.pl list 3

ls *phy > list
text_edit_phylip/to_interleaved_3.pl list

get baseml.ctl

ls *fourfold.phy > list

phylip_to_fasta_universal.pl list

#use *fas_aln file for window analysis

5.3.  Intergenic Constraint Within/Between

#need files:
*B*flankboth_aln from intergenic alignments.

rm *__* rm _*

ls *flankboth_aln > list

#need file: input for phylip run

remove_Q_from_nonB31_within_between.pl list within

ls *phy_intergen*  > list

to_interleaved.pl list

ls *outtree > list

reformat_phylip_3.pl list n

ls *phyI > list

edit_id.pl list
```

```
#need baseml.ctl file

ls *phyI.edit > list

bulk_paml.pl list

get_mlb_data.pl list > int_baseml_data_within

cat int_baseml_data_within | grep "^.*|.*|.*|.*|.*|.*$" >
int_baseml_data_within_edit
```

Get fourfold, noncoding data from orfs.

5.4 2x2 Contingency Test

```
psql intergenic

select int_baseml_data_within_2.tree_length,
browser_intergenic_new_filtered.length from
int_baseml_data_within_2, browser_intergenic_new_filtered where
int_baseml_data_within_2.orth_downstrm =
browser_intergenic_new_filtered.orf_downstrm and
browser_intergenic_new_filtered.genome_id = '1' and
browser_intergenic_new_filtered.mol_id = '3615' and
browser_intergenic_new_filtered.exclude = 'f' \o
mk_int_within_browser

select int_baseml_data_between_2.tree_length,
browser_intergenic_new_filtered.length from
int_baseml_data_between_2, browser_intergenic_new_filtered where
int_baseml_data_between_2.orth_downstrm =
browser_intergenic_new_filtered.orf_downstrm and
browser_intergenic_new_filtered.genome_id = '1' and
browser_intergenic_new_filtered.mol_id = '3615'and
browser_intergenic_new_filtered.exclude = 'f' \o
mk_int_between_browser

select orfs_fourfold_within.rates,
browser_intergenic_new_filtered.length from
orfs_fourfold_within, browser_intergenic_new_filtered where
orfs_fourfold_within.orf_id =
browser_intergenic_new_filtered.orf_downstrm and
browser_intergenic_new_filtered.genome_id = '1' and
browser_intergenic_new_filtered.mol_id = '3615' and
```

```
browser_intergenic_new_filtered.exclude = 'f' \o
mk_fourfold_within_browser

select orfs_fourfold_between.rates,
browser_intergenic_new_filtered.length from
orfs_fourfold_between, browser_intergenic_new_filtered where
orfs_fourfold_between.orf_id =
browser_intergenic_new_filtered.orf_downstrm and
browser_intergenic_new_filtered.genome_id = '1' and
browser_intergenic_new_filtered.mol_id = '3615'and
browser_intergenic_new_filtered.exclude = 'f' \o
mk_fourfold_between_browser


select orfs_nonsyn_within.rates,
browser_intergenic_new_filtered.length from orfs_nonsyn_within,
browser_intergenic_new_filtered where orfs_nonsyn_within.orf_id
= browser_intergenic_new_filtered.orf_downstrm and
browser_intergenic_new_filtered.genome_id = '1' and
browser_intergenic_new_filtered.mol_id = '3615'and
browser_intergenic_new_filtered.exclude = 'f' \o
mk_nonsyn_within_browser

select orfs_nonsyn_between.rates,
browser_intergenic_new_filtered.length from orfs_nonsyn_between,
browser_intergenic_new_filtered where orfs_nonsyn_between.orf_id
= browser_intergenic_new_filtered.orf_downstrm and
browser_intergenic_new_filtered.genome_id = '1' and
browser_intergenic_new_filtered.mol_id = '3615' and
browser_intergenic_new_filtered.exclude = 'f'  \o
mk_nonsyn_between_browser


tcat mk_fourfold_between_browser | sed 's/ //g' | sed 's/|/ /g'
> mk_fourfold_between_browser_edit

cp mk_fourfold_between_browser_edit input
get_mk_counts_browser.pl input

etc.

R

> x<-c(309,8967,1270,23658)
> y<-matrix(data=x,nrow=2,ncol=2,byrow=TRUE,dimnames=NULL)
```

```
> chisq.test(y,simulate.p.value=TRUE)

     Pearson's Chi-squared test with simulated p-value (based on
2000
     replicates)

data:  y
X-squared = 47.7465, df = NA, p-value = 0.0004998

> y
     [,1]  [,2]
[1,]  309  8967
[2,] 1270 23658

etc.
```

6. Window analysis/Phylogenetic Footprinting

Borrelia:

```
need baseml.ctl
need *ph tree files

window_anal_group_fourfold_flank.pl upstrm

ls *phyI > list_phyI

bulk_paml_window_indv_fourfold_flank.pl list_phyI

ls *mlb > list_mlb

get_mlb_data_window_2_indv.pl list_mlb > to_db

psql intergenic
delete from int_baseml_data_window_indv_fourfold
\copy int_baseml_data_window_indv_fourfold from 'int_to_db'
using delimiters '|'

select nt_num, rates from int_baseml_data_window_indv_fourfold
order by nt_num \o int_data_out

cat int_data_out |sed 's/ //g'| sed '/|/ /g' > int_data_out_edit
```

```
Neutral/Fourfold only:
bulk_paml_window_indv_fourfold_flank.pl fourfold-only-phylip-
alignments
(use outfile from intergen analysis above, remove intergenic,
run script).

plot_window_borrelia.R * need to make this file see p. 10


E. coli:

#need files: from intergenic analysis and fourfold analysis
(*fas_aln)
#need baseml.ctl file
#need phylip 'input' file

#fragment-length = 6 (for neutral and intergenic) or 3 (for
coding)
#intergenic-filename and fourfold-filename in format: upstream-
orf-id.filename and orf-id.filename
#need files: 'fourfold_list' and 'intergenic_list' to list orf-
id and upstream-orf-id
window-analysis.pl  -f fragment-length -i intergenic-filename -f
fourfold-filename
#example:

#intergenic analysis
window-analysis.pl -f 6 -i intergen_only.fas_aln -n
main.fourfold.phy.fas_aln

#neutral analysis:
window-analysis.pl -f 6 -i main.fourfold.phy.fas_aln -n
main.fourfold.phy.fas_aln
or

#only flanks

~/scripts/intergenic_analysis_clean/window-analysis-flankonly.pl
-f 0 -i main.fourfold.phy.fas_aln -n main.fourfold.phy.fas_aln

#coding analysis:
window-analysis.pl -f 6 -i main.nt.aln -n
main.fourfold.phy.fas_aln
```

7. Base Composition Analysis

7.1 Pallindrome Analysis

```
#get Pallindrome data
base_composition_anal.pl -f 6 -t window_anal_ecoli -i
intergen_only.fas_aln > out_palindrome

cat out_pallindrome | sed 's/ /|/g' | sed 's/||/|/g' | sed
's/_E/|E/g' | cut -f1-7,10 -d '|' > out_pallindrome_todb

 \copy ecoli_pallindrome from 'out_pallindrome_todb' using
delimiters '|'

constrained pallindrome = 212

select count(*) from ecoli_pallindrome,
window_anal_ecoli_annotated where ecoli_pallindrome.tree_length
< window_anal_ecoli_annotated.fourfold_rate and
ecoli_pallindrome.orf_upstrm =
window_anal_ecoli_annotated.orf_upstrm and ecoli_pallindrome.num
= window_anal_ecoli_annotated.num;


total pallindrome = 276

=77% pallindromes are constrained SE 2.53%

100*((0.77*(1-0.77)/276)^0.5)

[1] 2.533114

R
> se=c(2.53,2.53)

> xbar=c(77,23)

jpeg(filename="constraint_prop_at_pallindromes.jpeg",
type="quartz")

>
barplot2(xbar,plot.ci=TRUE,ci.u=xbar+se,ci.l=xbar,names=c("Const
rained","Non-Constrained"),main="Proportion of Pallindromes at
```

```
Constraint vs. Non-Constrained
Sites",ylab="%",xlab="Category")
```

7.2. Window analysis/Phylogenetic Footprinting

```
misc/intergen_window_anal.pl  upstrm.fourfold.fas
downstrm.fourfold.fas intergen.fas 6

get tree for window anal

get 'infile' for phylip run

misc/fourfold_intergen_fourfold_totalintergen.pl

ls *outtree > list

text_edit_tree/reformat_phylip_4.pl list n

#regulon analysis:

misc/select_treelength_bins.pl 0.01

#graphing compare to fourfold_flank only:

 graph_window_anal_int_vs_neut_flankonly.pl
 > out_int
#get tf data  from out_int put into file out_int_edit_tf

#get nontf data from out_int put into file out_int_edit_nontf

psql intergenic

\copy window_anal_ecoli_annotated from 'out_int_edit_tf' using
delimiters '|'

\copy window_anal_ecoli_annotated from 'out_int_edit_nontf'
using delimiters '|'

#select from above to db variations in rate
 (see book 6)

psql intergenic
select fourfold_rate from window_anal_ecoli_annotated \o
fourfold_flank_rate
```

```
#edit file > fourfold_flank_rate_edit

7.3. Promoter Analysis

put all sigma data into table : ecoli_sigmatotal_regulondb

#blastp_ecoli_promoter table  == matches sigma70 promoter sites
from regulondb to DH10B

select blastp_ecoli_promoter.promoter_identifier,
window_anal_ecoli_annotated_regulondb.orf_upstrm,
window_anal_ecoli_annotated_regulondb.orf_downstrm,window_anal_e
coli_annotated_regulondb.stop,
window_anal_ecoli_annotated_regulondb.int_rate from
window_anal_ecoli_annotated_regulondb, blastp_ecoli_promoter
where window_anal_ecoli_annotated_regulondb.start =
blastp_ecoli_promoter.genome_start \o DH10B_prom_approximate

Example for ECDH10B_1180:

select
orf_upstrm,orf_downstrm,orf_upstrm_strand,orf_downstrm_strand,nu
m,int_rate, fourfold_rate from
window_anal_ecoli_annotated_regulondb where orf_upstrm =
'ECDH10B_0959' order by num \o ECDH10B_1180_int_rate

 more ECDH10B_1180_int_rate | sed 's/orf.*//g' | sed 's/(.*//g'
| sed 's/ //g' | cut -f 5-7 -d '|' | sed 's/|/ /g' | sed 's/+*//
g' | sed 's/-*//g' > ECDH10B_1180_int_rate_edit

cp ECDH10B_1180_int_rate_edit > rate

cp *int_rate_edit > rate

plot_ECD10B_*_boot.R (in prom_plot_ecoli.tar)

#get total low gc's that are pallindromic
intergenic=# select count(*) from ecoli_gc_pallindrome where
gc_percent < 50;

 count
-------
378 = 77% (from total 499)
```

```
SE=100*(.77*.23/499)^.5=1.9%


#compare to neutral with fourfold fragments as intergenics:

~/scripts/intergenic_analysis_clean/graph_window_anal_neut_tf.pl
> out_int_neut

\copy window_anal_ecoli_annotated_neut from 'out_int_neut_edit'
using delimiters '|'


7.4 Analysis at Different "Intergenic Types"

misc/strandness_anal.pl

#strandness with binomial SE:

++: 53.2% SE 0.64%
100*(.532*(1-.532)/6058)^.5

+-:20.8% SE 0.55%
 100*(.208*(1-.208)/5445)^.5

--:85.8% SE 1.34%
 100*(.858*(1-.858)/677)^.5

-+:54.8% SE 1.53%
100*(.548*(1-.548)/1055)^.5


R
>library(gplots)
se=c(0.64,0.55,1.34,1.53)

xbar=c(53.2,20.8,85.8,54.8)
 jpeg(filename="constraint_per_genome_location",type="quartz")

> barplot2(xbar,plot.ci=TRUE,ci.u=xbar+se,ci.l=xbar,names=c("+
+","+-","--","-+"),main="Constraint Per Genome
Location",ylab="%",xlab="Upstream and Downstream Orthologous ORF
Strand")


7.5 AT/GC Content Analaysis (see 6.1)
```

```
#get AT/GC data
base_composition_anal_getfrag.pl -f 6 -t window_anal_ecoli -i
intergen_only.fas_aln > out_frag
(see 6.1, do like pallindrome)

constrained AT rich:
intergenic=#  select count(*) from ecoli_gc,
window_anal_ecoli_annotated_regulondb where ecoli_gc.orf_upstrm
= window_anal_ecoli_annotated_regulondb.orf_upstrm and
ecoli_gc.num = window_anal_ecoli_annotated_regulondb.num and
ecoli_gc.tree_length <
window_anal_ecoli_annotated_regulondb.fourfold_rate and
ecoli_gc.gc_percent < 50;

 count

-------

  7873

(1 row)

not constrained AT rich:
intergenic=#  select count(*) from ecoli_gc,
window_anal_ecoli_annotated_regulondb where ecoli_gc.orf_upstrm
= window_anal_ecoli_annotated_regulondb.orf_upstrm and
ecoli_gc.num = window_anal_ecoli_annotated_regulondb.num and
ecoli_gc.tree_length >
window_anal_ecoli_annotated_regulondb.fourfold_rate and
ecoli_gc.gc_percent < 50;

 count

-------

  4012


total at rich:
intergenic=#  select count(*) from ecoli_gc,
window_anal_ecoli_annotated_regulondb where ecoli_gc.orf_upstrm
= window_anal_ecoli_annotated_regulondb.orf_upstrm and
ecoli_gc.num = window_anal_ecoli_annotated_regulondb.num and
ecoli_gc.gc_percent < 50;
```

```
-------

 24605


Constrained and < 50% GC: 32%
SE=sqrt .32*.68/24605=0.002973841 = .30%

Not Constrained and < 50% GC: 16%
SE=100*(.16*.84/24605)^.5= 0.2337157
%

constrained AT  rich=32%
nonconsrained AT rich=16%
```

Bibliography

Akashi H. 2001. Gene expression and molecular evolution. *Curr Opin Genet Dev*. 11: 660–666.

Alkema WB, Lenhard B, WassermanWW. 2004. Regulog analysis: detection of conserved regulatory networks across bacteria: application to *Staphylococcus aureus*. *Genome Res.* 14:1362-1373.

Altschul SF, Madden TF, Schaffer AA, Zhang J, Zheng Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25:3389-3402.

Andolfatto P. 2005. Adaptive evolution of non-coding dna in *Drosophila*. *Nature.* 437:1149-1152.

Babb K, McAlister JD, Miller JC, Stevenson B. 2004. Molecular characterization of *Borrelia burgdorferi* erp promoter/operator elements. *J Bacteriol.* 186:2745-2756.

Balleza E, Bojorquez-Lopez LN, Martinez-Antonio A, Resendis-Antonio O, Lozada-Chavez I, Balderas-Martinez YI, Encarnacion S, Collado-Vides J. 2008. Regulation by transcription factors in bacteria: beyond description. *FEMS Microbiol Rev.* 33:133-151.

Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D. 2004. Ultraconserved elements in the human genome. *Science.* 304:1321-1325.

Bird CP, Stranger BE, Dermitzakis ET. 2006. Functional variation and evolution of non-coding DNA. *Curr Opin Genet Dev.* 16:559-64.

Blanchette M, Tompa M. 2002. Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res.* 12:739-748.

Boffelli D, McAuliffe J, Ovcharenko D, Lewis KD, Ovcharenko I, Pachter L, Rubin EM. 2003. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science*. 299:1391–1394.

Boffelli D, Nobrega MA, Rubin, EM. 2004. Comparative genomics at the vertebrate extremes. *Nat Rev Genet.* 5:456-465.

Bowen B, Steinberg J, Laemmli UK, Weintraub H. 1980. The detection of DNA-binding proteins by protein blotting. *Nucleic Acids Res.* 8:1–20.

Brisson D, Dykhuizen DE. 2004. OspC diversity in *Borrelia burgdorferi*: different hosts are different niches. *Genetics.* 168:713-722.

Brooks CS, PS Hefty, SE Jolliff, Akins DR. 2003. Global analysis of *Borrelia burgdorferi* genes regulated by mammalian host-specific signals. *Infect Immun.* 71:3371-3383.

Buck MJ, Lieb JD. 2004. ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments, *Genomics.* 83:349-360.

Bulyk ML. 2003. Computational prediction of transcription-factor binding site locations. *Genome Biol.* 5:201

Caimano MJ, Eggers CH, Hazlett KRO, Radolf JD. 2004. RpoS Is Not Central to the General

Stress Response in *Borrelia burgdorferi* but Does Control Expression of One or More Essential Virulence Determinants. *Infect Immun*. 72:6433-6445

Caimano MJ, Iyer R, Eggers CH, Gonzales C, Morton EA, Gilbert MA., Schwartz I, Radolf JD. 2007. Analysis of the RpoS regulon in *Borrelia burgdorferi* in response to mammalian host signals provides insight into RpoS function during the enzootic cycle. *Mol Microbiol*. 65:1193-1217.

Carroll SB. 2005. Evolution at two levels: on genes and form. *PLoS Biol*. 3:e245.

Casjens S, Palmer N, Van Vugt R. et. al. (16 co-authors). 2000. A bacterial genome in flux: the twelve linear and nine circular extrachromosomal DNAs in an infectious isolate of the Lyme disease spirochete *Borrelia burgdorferi*. *Mol Microbiol*. 35:490-516.

Casjens SM, Delange M, Ley HL, Rosa P, Huang WM. 1995. Linear chromosomes of Lyme disease agent spirochetes: genetic diversity and conservation of gene order. *J Bacteriol.* 177:2769-2780.

Chamary JV, Parmley JL, Hurst LD. 2006. Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat Rev Genet*. 7:98-108.s.

Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, Majors J, Waterston R, Cohen BA, Johnston M. 2003. Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science.* 301:71-76.

Cliften PF, Hillier LW, Fulton L, Graves, T, Miner T, Gish W, Waterston RH, Johnston M. 2001. Surveying *Saccharomyces* genomes to identify functional elements by comparative DNA sequence analysis. *Genome Res.* 11:1175-1186.

Comstedt P, Asokline L, Eliasson I, Olsen B, Wallensten A, Bunikis J, Bergstrom S. 2009. Complex population structure of Lyme borreliosis group spirochete *Borrelia garinii* in subarctic Eurasia. *Plos One*. 4:e5841.

Davies SR, Chang L, Patra D, Xing X, Posey K, Hecht J, Stormo G, Sandell LJ. 2007. Computational identification and functional validation of regulatory motifs in cartilage-expressed genes. *Genome Res.* 17:1438-1447.

DeHaseth PL, Helmann JD. 1995. Open complex formation by *Escherichia coli* RNA polymerase: the mechanism of polymerase-induced strand separation of double helical DNA. *Mol Microbiol*. 16:817-24.

Delihas N. 2009. Intergenic regions of *Borrelia* plasmids contain phylogenetically conserved RNA secondary structure motifs. *BMC Genomics.* 10:101.

Dermitzakis ET, Clark AG. 2002. Evolution of Transcription Factor Binding Sites in Mammalian Gene Regulatory Regions: Conservation and Turnover. *Mol Biol Evol.* 19:1114-1121.

Dobrikova EY, Bugrysheva J, Cabello FC. 2001. Two independent transcriptional units control the complex and simultaneous expression of the bmp paralogous chromosomal gene family in *Borrelia burgdorferi*. *Mol Microbiol*. 39:370-378.

Drake JA, Bird C, Nemesh J, Thomas DJ. et. al. (11 co-authors). 2005. Conserved noncoding

sequences are selectively constrained and not mutation cold spots. *Nat Genet.* 38:223-227.

Duret L, Bucher P. 1997. Searching for regulatory elements in human noncoding sequences. *Curr Opin Struct Biol*. 7:399-406.

Eddy SE. 2005. A model of the statistical power of comparative genome sequence analysis. *PLoS Biol.* 3:e10.

Eggers CH, Caimano MJ, Radolf JD. 2004. Analysis of Promoter Elements Involved in the Transcriptional Initiation of RpoS-Dependent *Borrelia burgdorferi* Genes. *J Bacteriol.* 186:7390-7402.

Eisen MB, Spellman PT, Brown PO, Botstein D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*. 95:14863-14868.

Farnham PJ. 2009. Insights from genomic profiling of transcription factors. *Nat Rev Genet.* 10:605-616.

Felsenstein J. 1989. PHYLIP -- Phylogeny Inference Package (Version 3.2). Cladistics 5:164-166.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol*. 17:368-376.

Fisher MA, Grimm D, Henion AK, Elias AF, Stewart PE, Rosa P, Gherardini FC. 2005. Borrelia burgdorferi sigma54 is required for mammalian infection and vector transmission but not for tick colonization. *Proc Natl Acad Sci U S A.* 102:5162-5167.

Fraser CM, Casjens S, Huang, WM. et. al. (37 co-authors) 1997. Genomic sequence of a Lyme disease spirochaete *Borrelia burgdorferi*. *Nature*. 390:580-586.

Gaffney DJ, Keightley PD. 2006. Genomic selective constraints in murid noncoding dna. *PLoS Genet*. 2:e204.

Galas D, Schmitz A. 1978. DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Res.* 5:3157-70.

Garner MM, Revzin A. 1981. A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the *Escherichia coli* lactose operon regulatory system. *Nucleic Acids Res.* 9:3047–3060.

Ge Y, Old IG, Saint Girons I, Charon NW. 1997. Molecular characterization of a large *Borrelia burgdorferi* motility operon which is initiated by a consensus sigma70 promoter. *J Bacteriol.* 179:2289-2299.

Glockner G, Lehmann R, Romualdi A, Pradella S, Schulte-Spechtel U, Schilhabel M, Wilske B, Suhnel J, Platzer M. 2004. Comparative analysis of the *Borrelia garinii* genome. *Nucleic Acids Res.* 32:6038-6046.

Gottesman S. 2005. Micros for microbes: non-coding regulatory RNAs in bacteria. *Trends Genet*. 21:399–404.

Grimm D, Elias AF, Tilly K, Rosa PA. 2003. Plasmid stability during in vitro propagation of *Borrelia burgdorferi* assessed at a clonal level. *Infect Immun.* 71:3138-3145.

Gumucio DL, Heilstedt-Williamson H, Gray TA, Tarle SA, Shelton DA, Tagle DA, Slightom JL, Goodman M, Collins FS. 1992. Phylogenetic footprinting reveals a nuclear protein which binds to silencer sequences in the human gamma and epsilon globin genes. *Mol Cell Biol.* 12:4919-4929.

Hahn M. 2007. Detecting natural selection on cis-regulatory dna. *Genetica*. 129:7-18.

Halligan DL, Keightley, PD. 2006. Ubiquitous selective constraints in the *Drosophila* genome revealed by a genome-wide interspecies comparison. *Genome Res.* 16:875-84

Hasegawa M, Kishino H, Yano TA. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol.* 22:160-174.

Hendrick PW. 2005. Genetics of Populations. London: Jones and Bartlett Publishers.

Hertz GZ, Stormo GD. 1999. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics.* 15:563-77.

Hirsh AE, Fraser HB, Wall DP. 2005. Adjusting for selection on synonymous sites in estimates of evolutionary distance. *Mol Biol Evol.* 22:174–177

Hoen AG, Margos G, Bent SJ, Diuk-Wasser MA, Barbour A, Kurtenbach K, Fish D. 2009. Phylogeography of *Borrelia burgdorferi* in the eastern United States reflects multiple independent Lyme disease emergence events. *Proc Natl Acad Sci U S A.* 106: 15013-15018.

Hubner A, Yang X, Nolen DM, Popova TG, Cabello FC, Norgard MV. 2001. Expression of *Borrelia burgdorferi* OspC and DbpA is controlled by a RpoN-RpoS regulatory pathway. *Proc Natl Acad Sci U S A*. 98:12724-12729.

Hudson RR, Kreitman M, Aguade M. 1987. A test of neutral molecular evolution based on nucleotide data. *Genetics*. 116:153-9.

Hughes JD, Estep PW, Tavazoie S, Church GM. 2000. Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J Mol Biol*. 296:1205-14.

Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown, PO. 2001. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature.* 409:533-538.

Jacob F, Monod J. 1961. Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol.* 3:318-356.

Jones, NC, Pevzner PA. 2004. An introduction to bioinformatics algorithms. Cambridge: MIT Press.

Jukes TH, Cantor CR. 1969. Evolution of protein molecules. In Mammalian Protein Metabolism. New York: Academic Press.

Kawano M, Storz G, Rao BS, Rosner JL, Martin RG. 2005. Detection of low-level promoter activity within open reading frame sequences of *Escherichia coli*. *Nucleic Acids Res*. 33:6268-6276.

Keightley PD, Gaffney DJ. 2003. Functional constraints and frequency of deleterious mutations in noncoding dna of rodents. *Proc Natl Acad Sci U S A*. 100:13402-13406.

Kim K, Shi H, Lee DK, Lis JT. 2003. Specific SR protein-dependent splicing substrates identified through genomic SELEX. *Nucleic Acids Res*. 31:1955–1961.

Kimura M. 1968. Evolutionary rate at the molecular level. *Nature*. 217:624–626.

Kimura M. 1977. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature*. 267:275-276.

Kimura M. 1983. The Neutral Theory of Molecular Evolution. Cambridge, UK: Cambridge University Press.

King JL, Jukes TH. 1969. Non-Darwinian evolution. *Science.* 164:788-798

King MC, Wilson AC. 1975. Evolution at two levels in Humans and Chimpanzees. *Science*. 188:107-116.

Krupka I, Knauer J, Lorentzen L, O'Connor TP, Saucier J, Straubinger RK. 2009. Borrelia burgdorferi sensu lato species in Europe induce diverse immune responses agaisnt C6 peptides in infected mice. *Clin Vaccine Immunol.* 16:1546-1562.

Kurtz S, Phillipy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open software for comparing large genomes. *Genome Biol.* 5:R12.

Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC. 1993. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science.* 262:208-14.

Lebech AM, Hansen K, Wilske B, Theisen M. 1994. Taxonomic classification of 29 *Borrelia burgdorferi* strains isolated from patients with Lyme borreliosis: a comparison of five different phenotypic and genotypic typing schemes. *Med Microbiol Immunol.* 18: 325-341.

Lenhard B, Sandelin A, Mendoza L, Engström P, Jareborg N, Wasserman WW. 2003 Identification of conserved regulatory elements by comparative genome analysis. *J Biol*. 2:13.

Li H, Wang W. 2003. Dissecting the transcription networks of a cell using computational genomics. *Curr Opin Genet Dev*. 13:611-616.

Liang H, Lin YS, Li WH. 2008. Fast evolution of core promoters in primate genomes. *Mol Biol Evol*. 25:1239-1244.

Lieb JD, Liu X, Botstein D, Brown PO. 2001. Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. *Nat Genet*. 28:327-334

Lin TJ, Oliver HJ, Gao L. 2003. Comparative analysis of Borrelia isolates from southeastern USA based on randomly amplified polymorphic DNA fingerprint and 16S ribosomal gene sequence analyses. *FEMS Microbiol Lett*. 228:249-257.

Ludwig MZ, Palsson A, Alekseeva E, Bergman CM, Nathan J, Kreitman M. 2005. Functional evolution of a cis-regulatory module. *Plos Biol*. 3:588-598.

Maliszewska-Tkaczyk M, Jonczyk P, Bialoskorska M, Schaaper RM, Fijalkowska IJ. 2000. SOS mutator activity: unequal mutagenesis on leading and lagging strands. *Proc Natl Acad Sci U S A*. 97:12678-12783.

Margolis ND, Hogan W, Cieplak Jr, Schwan TG, Rosa PA. 1994. Homology between Borrelia burgdorferi OspC and members of the family of *Borrelia hermsii* variable major proteins. *Gene.*

143:105-10.

Margos G, Gatewood AG, Aanensen DM. et. al. (17 co-authors). 2008. MLST of housekeeping genes captures geographic population structure and suggests a European origin of Borrelia burgdorferi. *Proc Natl Acad Sci U S A* . 105:8730-8735.

Massie CE, Mills IG. 2008. ChIPping away at gene regulation. *EMBO reports.* 9:337-343.

McCue L, Thompson W, Carmack C, Ryan MP, Liu JS, Derbyshire V, Lawrence CE. 2001. Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nucleic Acids Res*. 29:774-782.

McCue LA, Thompson W, Carmack CS, Lawrence CE. 2002. Factors influencing the identification of transcription factor binding sites by cross-species comparison. *Genome Res.* 12:1523-1532.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature.* 351:652-654.

McEwen GKK, Woolfe A, Goode D, Vavouri T, Callaway H, Elgar G. 2006. Ancient duplicated conserved noncoding elements in vertebrates: A genomic and functional analysis. *Genome Res*. 16:451-65

Mcguire AM, Church GM. 2000. Predicting regulons and their cis-regulatory motifs by comparative genomics. *Nucleic Acids Res*. 28:4523-4530.

McInerney JO. 1998. Replicational and transcriptional selection on codon usage in *Borrelia burgdorferi*. *Proc Natl Acad U S A*. 95:10698-703.

Medina M. 2005. Colloquium Paper: Systematics and the Origin of Species: Genomes, phylogeny, and evolutionary systems biology. *Proc Natl Acad Sci U S A.* 102:6630-35.

Mirkin EV, Roa DC, Nudler E, Mirkin SM. 2006. Transcription regulatory elements are punctuation marks for DNA replication. *Proc Natl Acad Sci U S A.* 103:7276-7281.

Molina N, van Nimwegen E. 2008. Universal patterns of purifying selection at noncoding positions in bacteria. *Genome Res*. 18:148-160

Mount DW. 2001. Bioinformatics Sequence and Genome Analysis. New York: Cold Spring Harbor.

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol.* 3:418-426.

Neuwald AF, Liu JS, Lawrence CE. 1995. Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Sci*. 4:1618-32.

Nielsen R. 2001. Statistical tests of selective neutrality in the age of genomics. *Heredity.* 86:641-7.

Ojaimi C, Brooks C, Casjens S. et al. (15 co-authors). 2003. Profiling of temperature-induced changes in Borrelia burgdorferi gene expression by using whole genome arrays. *Infect Immun.* 71:1689-1705.

Ouyang Z, Blevins JS, Norgard MV. 2008. Transcriptional interplay among the regulators Rrp2,

RpoN and RpoS in *Borrelia burgdorferi*. *Microbiology*. 154:2641-2658.

Ouyang Z, He M, Oman T, Frank Yang X, Norgard MV. 2009. (BmtA), is required for virulence by the Lyme disease spirochete, *Borrelia burgdorferi*. *Proc Natl Acad U S A*. 106:3449-3454.

Pavesi G, Mauri G, Pesole G. 2004. In silico representation and discovery of transcription factor binding sites. *Brief Bioinform.* 5:217-236.

Prud'homme B, Gompel N, Carroll SB. 2007. Emerging principles of regulatory evolution. *Proc Natl Acad Sci U S A*. 104 Suppl 1:8605-8612.

Qin ZS, McCue LA, Thompson W, Mayerhofer L, Lawrence CE, Liu JS. 2003. Identification of co-regulated genes through Bayesian clustering of predicted regulatory binding sites. *Nat Biotechnol.* 21:435-439.

Qiu WG, Bruno JF, McCaig WD, Xu Y, Livey I, Schriefer ME, Luft BJ. 2008. Wide distribution of a high-virulence *Borrelia burgdorferi* clone in Europe and North America. *Emerg Infect Dis.* 14:1097-1104.

Qiu WG, Schutzer SE, Bruno JF, Attie O, Xu Y, Dunn JJ, Fraser CM, Casjens SR, Luft BJ. 2004. Genetic exchange and plasmid transfers in *Borrelia burgdorferi sensu stricto* revealed by three-way genome comparisons and multilocus sequence typing. *Proc Natl Acad Sci U S A*. 101:14150-14155.

R Development Core Team. 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Rajewsky N, Socci ND, Zapotocky M, Siggia ED. 2002. The evolution of dna regulatory regions for proteo-gamma bacteria by interspecies comparisons. *Genome Res*. 12:298-308.

Ramamoorthy R, McClain NA, Gautam A, Scholl-Meeker D. 2005. Expression of the bmpB Gene of *Borrelia burgdorferi* is modulated by two distinct transcription termination events. *J Bacteriol*. 187:2592-2600.

Ray P, Shringarpure S, Kolar M, Xing EP. 2008. CSMET: Comparative Genomic Motif Detection via Multi-Resolution Phylogenetic Shadowing. *PLoS Comput Biol.* 4:e1000090.

Ren B, Robert F, Wyrick JJ. et. al. (14 co-authors). 2000. Genome-wide location and function of DNA binding proteins, *Science.* 290:2306-2309.

Resch AM, Carmel L, Mariño-Ramírez L, Aleksey OY, Shabalina SA, Rogozin IB, Koonin EV. 2007. Widespread Positive Selection in Synonymous Sites of Mammalian Genes. *Mol Biol Evol*. 24:1821-1831.

Revel AT, Talaat AM, Norgard MV. 2002. DNA microarray analysis of differential gene expression in Borrelia burgdorferi, the Lyme disease spirochete. *Proc Natl Acad Sci U S A*. 99:1562-1567.

Rocha EPC. 2002. Is there a role for replication fork asymetry in the distribution of genes in bacterial genomes? *Trends in Microbiology.* 10:393-395.

Roderick DM, Holmes EC. 1998. Molecular Evolution A Phylogenetic Approach. Oxford:

Blackwell Science Ltd.

Rodriguez E, Oliver JL, Marin A, Medina JR. 1990. The general stochastic model of nucleotide substitution. *J Theor Biol.* 142:485-501.

Rogozin IB, Makarova KS, Natale DA, Spiridonov AN, Tatusov RL, Wolf YI, Yin J, Koonin EV. 2002. Congruent evolution of different classes of non-coding dna in prokaryotic genomes. *Nucleic Acids Res.* 30:4264-4271.

Rosa PA, Tilly K, Stewart PE. 2005. The burgeoning molecular genetics of the Lyme disease spirochaete. *Nat Rev Microbiol.* 3:129-143.

Salzberg S, Delcher A, Kasif S, White O. 1998. Microbial gene identification using interpolated Markov models. *Nucleic Acids Research* 26:544-548.

Sandelin A, Bailey P, Bruce S, Engström PG, Klos JM, Wasserman WW, Ericson J, Lenhard B. 2004. Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. *BMC Genomics*. 5:99.

Shabalina SA, Ogurtsov AY, Kondrashov VA, Kondrashov AS. 2001. Selective constraint in intergenic regions of human and mouse genomes. *Trends Genet*. 17:373-376.

Shimida T, Fujita N, Maeda M, Ishihama A. 2005. Systematic search for the Cra-binding promoters using genomic SELEX system. *Genes Cells*. 10:907–918.

Shtatland T, Gill SC, Javornik BE, Johanson HE, Singer BS, Uhlenbeck OC, Zichi DA, Gold L. 2000. Interactions of *Escherichia coli* RNA with bacteriophage MS2 coat proteins: genomic SELEX. *Nucleic Acids Res*. 28:e93.

Siepel A, Bejerano G, Pedersen JS. et. al. (16 co-authors). 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*. 15:1034-1050.

Smith AHH, Blevins JSS, Bachlani GNN, Yang XFF, Norgard MVV. 2007. Evidence that rpos (sigmas) in *Borrelia burgdorferi* is controlled directly by rpon (sigma54/sigman). *J Bacteriol*. 189(5):2139-44

Stajich JE, Block D, Boulez K. et. al. (21 co-authors). 2002. The Bioperl toolkit: Perl modules for the life sciences. *Genome Res*. 12:1611-1618.

Steensel BV, Henikoff S. 2000. Identification of in vivo DNA targets of chromatin proteins using tethered Dam methyltransferase. *Nat Biotechnol.* 18:424 – 428.

Stewart PE, Byram R, Grimm D, Tilly K, Rosa PA. 2005. The plasmids of *Borrelia burgdorferi*: essential genetic elements of a pathogen. *Plasmid*. 53:1-13.

Stormo GD. 2000. DNA binding sites: representation and discovery. *Bioinformatics* 16:16-23.

Taft RJ, Pheasant M, Mattick JS. 2007. The relationship between non-protein-coding DNA and eukaryotic complexity. *BioEssays.* 29:288 -299.

Thompson JD, Higgins DG, Gibson TJ. 1994. *Nucleic Acids Res.* 11:4673-4680.

Thompson W, Rouchka EC, Lawrence CE. 2003. Gibbs Recursive Sampler: finding transcription factor binding sites. *Nucleic Acids Res*. 31:3580-5.

Tilly K, Krum JG, Bestor A, Jewett MW. et. al. (11 co-authors). 2006. Borrelia burgdorferi OspC protein required exclusively in a crucial early stage of mammalian infection. *Infect Immun*. 74:3554-3564.

Van Steensel B, Delrow J, Henikoff S. 2001. Chromatin profiling using targeted DNA adenine methyltransferase. *Nat Genet*. 27:304-308.

Vavouri T, Walter K, Gilks WR, Lehner B, Elgar G. 2007. Parallel evolution of conserved noncoding elements that target a common set of developmental regulatory genes from worms to humans. *Genome Biol*. 8:R15+.

Veaute X, Fuchs RP. 1993. Greater susceptibility to mutations in lagging strand of DNA replication in *Escherichia coli* than in leading strand. *Science.* 261:598-600.

Wagner R. 2000. Transcription Regulation in Prokaryotes. New York: Oxford University Press.

Wang G, van Dam AP, Schwartz I, Dankert J. 1999. Molecular typing of *Borrelia burgdorferi sensu lato*: taxonomic, epidemiological, and clinical implications. *Clin Microbiol Rev*. 1999:4.

Wang G, van Dam AP, Schwartz I, Dankert J. 1999. Molecular Typing of *Borrelia burgdorferi Sensu Lato*: Taxonomic, Epidemiological, and Clinical Implications. *Clin Microbiol.* 12: 633-653.

Wayne ML, Simonsen KL. 1998. Statistical tests of neutrality in the age of weak selection. *Trends Ecol Evol*. 13:236-240.

Woolfe A, Goodson M, Goode DK. et. al. (16 co-authors). 2005. Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol*. 3:e7

Wray GA. The evolutionary significance of cis-regulatory mutations. *Nat Rev Genet*. 8:206-216.

Yang XF, Lybecker MC, Pal U, Alani SM, Blevins J, Revel AT, Samuels DS, Norgard MV. 2005. Analysis of the ospC regulatory element controlled by the RpoN-RpoS regulatory pathway in *Borrelia burgdorferi*. *J Bacteriol*. 187(14):4822-9.

Yang Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol Biol Evol* 24: 1586-1591

Zhang Z, Gerstein M. 2003. Of mice and men: phylogenetic footprinting aids the discovery of regulatory elements. *J Biol.* 2:11