RESEARCH ARTICLE

Phylogenomic Identification of Regulatory Sequences in Bacteria: an Analysis of Statistical Power and an Application to *Borrelia burgdorferi Sensu Lato*

Che I. Martin,[a] Tika Y. Sukarna,[a]* Saymon Akther,[b] Girish Ramrattan,[b] Pedro Pagan,[b] Lia Di,[b] Emmanuel F. Mongodin,[c] Claire M. Fraser,[c] Steven E. Schutzer,[d] Benjamin J. Luft,[e] Sherwood R. Casjens,[f] Wei-Gang Qiu[a,b]

Department of Biology, The Graduate Center, City University of New York, New York, USA[a]; Department of Biological Sciences and Center for Translational and Basic Research, Hunter College, City University of New York, New York, USA[b]; Institute for Genome Sciences, University of Maryland BioPark, Baltimore, Maryland, USA[c]; Department of Medicine, New Jersey Medical School, Rutgers, the State University of New Jersey, Newark, New Jersey, USA[d]; Department of Medicine, Health Science Center, Stony Brook University, Stony Brook, New York, USA[e]; Department of Pathology, Division of Molecular Cell Biology and Immunology, University of Utah School of Medicine, Salt Lake City, Utah, USA[f]

* Present address: Tika Y. Sukarna, Akonlabs Research, Kebayoran Baru, Jakarta Selatan, Indonesia.

C.I.M. and T.Y.S. contributed equally to this article.

ABSTRACT **Phylogenomic footprinting is an approach for *ab initio* identification of genome-wide regulatory elements in bacterial species based on sequence conservation. The statistical power of the phylogenomic approach depends on the degree of sequence conservation, the length of regulatory elements, and the level of phylogenetic divergence among genomes. Building on an earlier model, we propose a binomial model that uses synonymous tree lengths as neutral expectations for determining the statistical significance of conserved intergenic spacer (IGS) sequences. Simulations show that the binomial model is robust to variations in the value of evolutionary parameters, including base frequencies and the transition-to-transversion ratio. We used the model to search for regulatory sequences in the Lyme disease species group (*Borrelia burgdorferi sensu lato*) using 23 genomes. The model indicates that the currently available set of *Borrelia* genomes would not yield regulatory sequences shorter than five bases, suggesting that genome sequences of additional *B. burgdorferi sensu lato* species are needed. Nevertheless, we show that previously known regulatory elements are indeed strongly conserved in sequence or structure across these *Borrelia* species. Further, we predict with sufficient confidence two new RpoS binding sites, 39 promoters, 19 transcription terminators, 28 noncoding RNAs, and four sets of coregulated genes. These putative *cis*- and *trans*-regulatory elements suggest novel, *Borrelia*-specific mechanisms regulating the transition between the tick and host environments, a key adaptation and virulence mechanism of *B. burgdorferi*. Alignments of IGS sequences are available on BorreliaBase.org, an online database of orthologous open reading frame (ORF) and IGS sequences in *Borrelia*.**

IMPORTANCE **While bacterial genomes contain mostly protein-coding genes, they also house DNA sequences regulating the expression of these genes. Gene regulatory sequences tend to be conserved during evolution. By sequencing and comparing related genomes, one can therefore identify regulatory sequences in bacteria based on sequence conservation. Here, we describe a statistical framework by which one may determine how many genomes need to be sequenced and at what level of evolutionary relatedness in order to achieve a high level of statistical significance. We applied the framework to *Borrelia burgdorferi*, the Lyme disease agent, and identified a large number of candidate regulatory sequences, many of which are known to be involved in regulating the phase transition between the tick vector and mammalian hosts.**

Received 7 January 2015 **Accepted** 10 March 2015 **Published** 14 April 2015

Citation Martin CI, Sukarna TY, Akther S, Ramrattan G, Pagan P, Di L, Mongodin EF, Fraser CM, Schutzer SE, Luft BJ, Casjens SR, Qiu W-G. 2015. Phylogenomic identification of regulatory sequences in bacteria: an analysis of statistical power and an application to *Borrelia burgdorferi sensu lato*. mBio 6(2):e00011-15. doi:10.1128/mBio.00011-15.

Editor Louis M. Weiss, Albert Einstein College of Medicine

Address correspondence to Wei-Gang Qiu, weigang@genectr.hunter.cuny.edu.

A major rationale for sequencing a large number of closely related genomes is to identify candidate gene-regulatory elements and networks based on the observation that functional elements tend to be conserved in DNA sequences between as well as within genomes (1–3). Such evolutionary approaches, which may be called phylogenomic footprinting (4), are relatively cost-effective and have been successfully used in revealing candidate regulatory elements in humans (5, 6), *Drosophila* species (7, 8), and yeasts (9, 10). The evolutionary approach is especially valuable for non-model bacterial species for which a method of experimental and genetic manipulations is limited or nonexistent (11).

Borrelia burgdorferi sensu lato, a non-model bacterial species group of Gram-negative spirochetes, consists of at least 18 named and putative species (12, 13). Several species of this complex are causative agents of Lyme disease, a tick-borne infectious disease that is increasing in prevalence throughout North America, Eu-

rope, and East Asia (13, 14). Three species, *Borrelia garinii*, *Borrelia afzelii*, and *B. burgdorferi sensu stricto*, cause the majority of Lyme disease worldwide. In North America, Lyme disease is predominantly caused by *B. burgdorferi sensu stricto*. At least 20 evolutionary lineages of *B. burgdorferi sensu stricto* exist in Europe and North America, some of which are more likely than others to cause disseminated Lyme disease in humans (15–17).

As an obligate parasite, *B. burgdorferi* must survive in two physiologically distinct environments between the tick and its vertebrate host for its maintenance in nature, and hence elaborative mechanisms for regulating levels of gene expression during such phase transitions have evolved (18–20). Over 100 genes (~10%) in the *B. burgdorferi* genome are differentially expressed during the transition between the tick and mammalian phases (21–23). RpoS (σ^s), an alternative sigma factor, appears to be a main transcriptional control mechanism regulating the tick-mammal transitions via the Rrp2-RpoN-RpoS gene regulatory pathway (19, 22, 23). For example, the Rrp2-RpoN-RpoS pathway is activated during tick feeding, leading to the upregulation of mammalian phase lipoprotein genes (e.g., *ospC* [encoding outer surface protein C] and *dbpAB* [encoding decorin-binding proteins A and B] operon) and the simultaneous downregulation of tick phase genes (e.g., *ospA* [encoding outer surface protein A]) (18, 19, 24–26). Five genes (*ospC*, *dbpA*, *oppA5*, *bba66*, and *bba07*) have been identified to contain a consensus RpoS-dependent promoter sequence (27). Additional gene-regulatory pathways important for *B. burgdorferi sensu lato* pathogenesis are beginning to be understood, such as post-transcriptional control with small RNAs, genes targeting the host complement systems, and genes responsible for its persistent infection in hosts (18, 19, 28). In spite of these new findings, the majority of downstream targets of key gene regulatory mechanisms, including the Rrp2-RpoN-RpoS pathway, remain to be identified (29).

Much of the knowledge about *Borrelia* gene regulation, e.g., the discovery of the RpoN-RpoS pathway, benefited from prior studies of homologous proteins in model organisms, such as *Escherichia coli* (19). Recently, we sequenced the genomes of 13 strains of *B. burgdorferi sensu stricto* and nine strains of other *B. burgdorferi sensu lato* species, bringing the total number of completed or draft *B. burgdorferi sensu lato* genomes to at least 24 (30, 31). These genomes make it possible to use phylogenomic footprinting for *ab initio* discovery of *Borrelia*-specific gene-regulatory elements and networks that may not exist in other bacterial groups. Previously, five putative noncoding RNAs (ncRNAs) have been identified based on a comparison of three genome sequences (32). Five additional candidate ncRNAs on lp54 and cp26, the two constitutive plasmids, have been identified using these *B. burgdorferi sensu lato* genomes (33). Here, we describe the results of a more comprehensive and systemic search for highly conserved putative regulatory genomic elements in the core *B. burgdorferi sensu lato* genome. In addition, we propose a statistical framework for guiding the search for candidate functional elements using phylogenomic footprinting in *Borrelia* or other bacterial groups.

RESULTS AND DISCUSSION

Genome sequences. (i) Genomes and orthologous ORFs. We and other groups have sequenced and released the genome sequences of 23 *B. burgdorferi sensu lato* strains isolated from North America and Europe encompassing eight *B. burgdorferi sensu lato* species (see Table S1 in the supplemental material). The present study is based on the genomic sequences of the three universally present replicons, including the cp26 and lp54 plasmids and the main chromosome. We have previously identified, by using automated homology searches and manual synteny analysis, 837 orthologous open reading frame (ORF) families, including 750 on the main chromosome, 26 on the cp26 plasmid, and 62 on the lp54 plasmid (30, 34).

(ii) Orthologous IGS families. After identifying consensus start codon positions for orthologous ORF families (see Materials and Methods), discarding short (<150-base) predicted ORFs, and filtering out short (<30-base) intergenic spacer (IGS) sequences and IGS sequences not present in seven or more sequenced *B. burgdorferi sensu lato* species, the final data set for all subsequent analysis consists of 17 orthologous IGS families on the cp26 plasmid, 26 orthologous IGS families on the lp54 plasmid, and 203 orthologous IGS families on the main chromosome (Table 1).

Power analysis. (i) Synonymous tree lengths. Synonymous tree lengths (T_S) of ORFs are the key parameter for determining the statistical significance of sequence variability of flanking IGSs (see Materials and Methods, equations 1 and 2). T_S and nonsynonymous tree lengths (T_N) were obtained for 23, 41, and 327 IGS-flanking ORF families on cp26, lp54, and the main chromosome, respectively. Median tree length values are listed in Table 1, since the tree lengths are not normally distributed and many outliers exist. The outliers include high T_S values at *0256* (*rpsU*, encoding ribosomal protein S21), at *b10*, *b12*, and *b13* (three plasmid-partitioning genes on cp26), and at *b19* (*ospC*) and high T_N values at *a24* (*dbpA*) and *b19* (*ospC*). An earlier study using between-species comparisons found a similar group of outliers (33). The smaller T_S values of chromosomal ORFs than those of plasmid-borne ORFs have more to do with higher effective recombination rates caused by diversifying natural selection on the plasmids than with the unsequenced chromosome of strain 297 (34). The T_N/T_S ratios indicate that ORFs on the main chromosome and cp26 are about twice as conserved as ORFs on lp54, consistent with a previous study based on pairwise comparisons between B31 and another strain (30).

(ii) Plasmid-borne elements are more easily resolved. Using the median T_S value of 1.5 (Table 1) as the expected number of neutral substitutions per site for IGSs on the main chromosome during the evolutionary diversification among the 22 *Borrelia* genomes, levels of statistical significance of an IGS segment (with a length [L] of 5, 10, 15, or 20 bases) showing $n = 0$ to 10 substitutions are plotted according to equation 1 in Materials and Methods (Fig. 1A). These results show that one may not expect comparative analysis of these genomes to reveal functional chromosomal IGS elements shorter than five bases (Fig. 1, the "L=5" line). Regulatory sequences with a length of 10 bases having 0 to 3 variable sites would be marginally significant (Fig. 1, the "L=10" line). With a higher median T_S value of 1.85 for sequences on the plasmids, shorter and more variable regulatory sequences could be significantly detected using these *Borrelia* genomes (Fig. 1B).

(iii) A need for divergent genomes. We simulated the evolution of plasmid-borne IGS sequences from the 23 genomes under neutral conditions using EVOLVER (35). A tree (Fig. 1C) was inferred, and synonymous subtree lengths (Fig. 1D, gray vertical lines) were obtained at various levels of phylogenetic divergence. The analysis shows that even long (L = 20 bp) functional IGS elements would not be resolvable if using only the genomes of *B.*

TABLE 1 Orthologous ORFs and IGSs

Characteristic	Value		
	Main chromosome	lp54	cp26
No. of orthologous ORF families	750	62	26
Synonymous tree length (T_S)[a]	1.4965	1.8959	1.7931
Nonsynonymous tree length (T_N)[a]	0.1092	0.3684	0.1925
Ratio (T_N/T_S)	0.07297	0.1944	0.1074
No. of orthologous IGS families[b]	203	27	17
No. of convergent IGSs (no. conserved[c]; %)	31 (6[d]; 19.4)	3 (0; 0)	3 (0; 0)
No. of tandem IGSs (no. conserved[c]; %)	109 (41[e]; 37.6)	19 (8[g]; 42.1)	8 (0; 0)
No. of divergent IGSs (conserved[c]; %)	63 (40[f]; 63.5)	5 (0; 0)	(3[h]; 50)

[a] Median values obtained by PAML (35) among 23, 41, and 327 orthologous ORF families on cp26, lp54, and the main chromosome, respectively. The total number of sequences in individual ORF families is 22 for those on the main chromosome and 23 for those on lp54 and cp26.

[b] Includes only IGSs with an alignment length of 30 bases or more.

[c] With nucleotide substitution rates obtained by Rates4site (63) significantly lower ($P < 0.001$ by t test) than those of flanking third-codon sites.

[d] *bb0004-bb0005*, *bb0364-bb0365*, *bb0459-bb0460*, *bb0536-bb0537*, *bb0688-bb0689*, *bb0758-bb0759.*

[e] *bb0034-bb0035*, *bb0057-bb0058*, *bb0089-bb0090*, *bb0146-bb0147*, *bb0163-bb0164*, *bb0172-bb0173*, *bb0208-bb0209*, *bb0219-bb0220*, *bb0247-bb0248*, *bb0250-bb0251*, *bb0255-bb0256*, *bb0278-bb0279*, *bb0328-bb0329*, *bb0339-bb0340*, *bb0347-bb0348*, *bb0380-bb0381*, *bb0381-bb0382*, *bb0389-bb0390*, *bb0390-bb0391*, *bb0430-bb0431*, *bb0434-bb0435*, *bb0539-bb0540*, *bb0542-bb0543*, *bb0567-bb0568*, *bb0584-bb0585*, *bb0603-bb0604*, *bb0608-bb0610*, *bb0642-bb0643*, *bb0647-bb0648*, *bb0671-bb0672*, *bb0679-bb0680*, *bb0693-bb0694*, *bb0715-bb0716*, *bb0726-bb0727*, *bb0744-bb0745*, *bb0755-bb0756*, *bb0770-bb0771*, *bb0773-bb0774*, *bb0776-bb0777*, *bb0808-bb0809*, *bb0830-bb0831.*

[f] *bb0007-bb0008*, *bb0023-bb0024*, *bb0045-bb0046*, *bb0100-bb0101*, *bb0133-bb0134*, *bb0135-bb0136*, *bb0154-bb0155*, *bb0190-bb0192*, *bb0201-bb0202*, *bb0214-bb0215*, *bb0226-bb0227*, *bb0236-bb0237*, *bb0253-bb0254*, *bb0313-bb0314*, *bb0336-bb0337*, *bb0346-bb0347*, *bb0365-bb0366*, *bb0373-bb0374*, *bb0400-bb0401*, *bb0436-bb0437*, *bb0454-bb0455*, *bb0457-bb0458*, *bb0460-bb0461*, *bb0507-bb0508*, *bb0560-bb0561*, *bb0571-bb0572*, *bb0596-bb0597*, *bb0598-bb0599*, *bb0620-bb0621*, *bb0623-bb0624*, *bb0629-bb0630*, *bb0655-bb0656*, *bb0706-bb0707*, *bb0723-bb0724*, *bb0734-bb0735*, *bb0748-bb0749*, *bb0760-bb0761*, *bb0812-bb0814*, *bb0828-bb0829*, *bb0835-bb0836.*

[g] *bba14-bba15*, *bba16-bba18*, *bba21-bba23*, *bba24-bba25*, *bba39-bba40*, *bba51-bba52*, *bba64-bba65*, *bba65-bba66.*

[h] *bbb08-bbb09*, *bbb25-bbb26*, *bbb27-bbb28.*

burgdorferi sensu stricto and its closest relative, SV1 (Fig. 1D, "SV1" line). The plot further suggests that elements shorter than 5 bases would not be resolved at a false discovery rate smaller than a *P* value of 0.001 even with additional genomes. Nevertheless, sequencing more genomes from phylogenetically distinct *B. burgdorferi sensu lato* lineages would be the most cost-effective for the identification of candidate functional elements using phylogenomics (Fig. 1C and D). In North and South America, divergent *B. burgdorferi sensu lato* species not represented by the present genome data set include *Borrelia carolinensis*, *Borrelia kurtenbachii*, *Borrelia californiensis*, an unnamed species ("geno-species 2"), *Borrelia americana*, *Borrelia andersonii*, and *Borrelia chilensis* (13, 36). In Eurasia, *B. burgdorferi sensu lato* species highly divergent from those in the present study include *Borrelia sinica*, *Borrelia yangtze*, *Borrelia tanukii*, *Borrelia japonica*, *Borrelia lusitaniae*, and *Borrelia turdi* (13, 36). Comparison among distantly related genomes, however, introduces its own problems, such as an inability to identify species-specific regulatory elements since such elements evolved recently and are not conserved across all genomes (37).

***Borrelia* IGSs are enriched in conserved elements. (i) Conserved IGSs.** Conserved IGSs were identified as those with significantly low (by t tests at $P < 0.001$) nucleotide substitution rates relative to the rates at flanking third-base sites. At each IGS locus, substitution rates of IGS and ORF sites were coestimated with a concatenated alignment using Rates4site (38). For example, the *b08-b09* IGS contains a higher proportion of slowly evolving sites than its flanking third-base sites (Fig. 2B). Among the three directional types of IGSs, divergent IGSs tend to contain a large number of conserved sequences (63.5% for chromosomal IGSs) while convergent IGSs have relatively few conserved sequences (19.4% for chromosomal IGSs) (Table 1). This observation is consistent with the expectation that divergent and tandem IGSs are more likely than convergent IGSs to house *cis*-regulatory sequences. Among the three replicons, IGSs and ORFs on the main chromosome (Fig. 2A, right-most panels) contain a higher proportion of low-rate sites and are therefore more conserved than IGSs and ORFs on the plasmids (Fig. 2A, left-most and middle panels). Relatively low evolutionary rates on the main chromosome are expected, since it has lower effective recombination rates than the plasmids (34).

(ii) PCIBs. We identified a total of 935 and 276 perfectly conserved intergenic blocks (PCIBs) with a minimal length of six nucleotides on the main chromosome and the plasmids, respectively. These PCIBs occur within 125 nucleotides upstream or downstream of an ORF. The total lengths of these ORF-flanking PCIBs on the main chromosome and the two plasmids are, respectively, 26,417 and 10,199 bases, or 43.3% and 29.8% of the selected IGS sequences on the B31 genome. The comparable numbers from randomly permuted IGS alignments are 27.4% and 10.6%, respectively, indicating that *B. burgdorferi* IGSs are about 1.5 times and 3.0 times as enriched in conserved sequence blocks as expected by chance on the main chromosome and the plasmids, respectively. While the observed PCIBs outnumber those in shuffled alignments in nearly every length category, those on the main chromosome ($P = 8.1e{-}12$ by a one-tailed t test; Fig. 3A) are not as significant as those on the plasmids ($P = 7.3e{-}15$; Fig. 3B). Such deficiency in enrichment or lack of significance of conserved sequences on the main chromosome, however, does not necessarily imply that chromosomal IGSs harbor a proportionally smaller number of regulatory sequences. Rather, these deficiencies reflect a relative lack of statistical power for distinguishing functional IGS elements from neutrally evolving sequences on the main chromosome, which has a lower level of overall phylogenetic divergence than the plasmids (Fig. 1, Table 1). Sequencing the genomes of additional *B. burgdorferi* species is therefore expected to increase the resolving power of phylogenomics toward revealing shorter

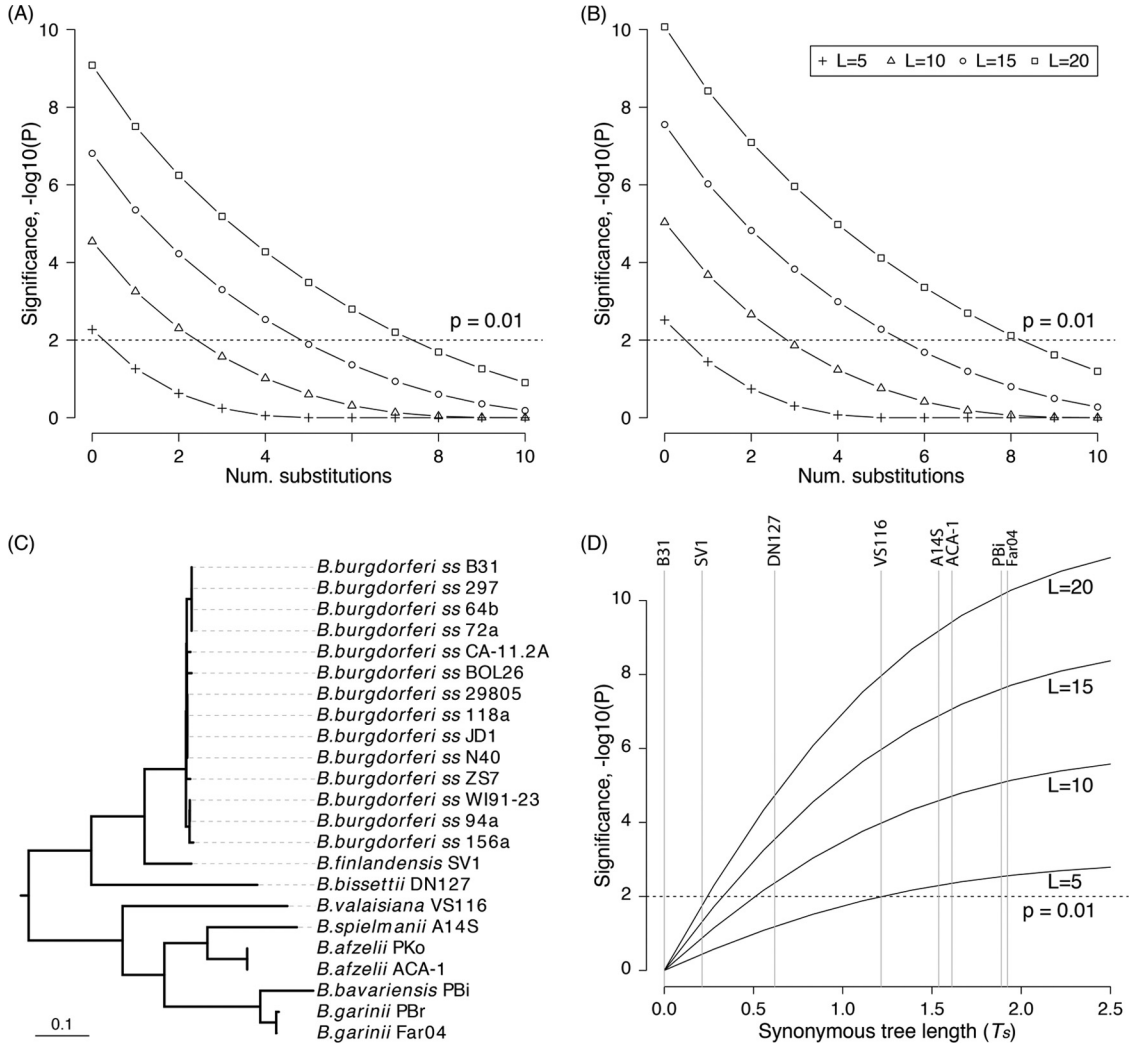

FIG 1 Statistical power of phylogenomic footprinting. (A and B) Each data point represents the probability (*y* axis, in −log10) of an *L*-mer IGS segment having *n* substitutions (*x* axis) after evolving with an expected neutral distance of T_0. These probabilities were calculated according to equation 1 in Materials and Methods and obtained using the R function *pbinom* (58). (A) Probabilities for IGSs on the main chromosome, with the neutral distance T_0 approximated by $T_S = 1.5$ substitutions/site (Table 2); (B) probabilities for IGSs on the plasmids, with the neutral distance T_0 approximated by $T_S = 1.85$ substitutions/site (Table 2). These two plots show that the statistical power of identifying regulatory elements using phylogenomic footprinting increases with the length of the element (*L*), the degree of its sequence conservation (*n*), and the total neutral divergence among the genomes (T_S). (C) Phylogenetic tree of neutrally evolved IGS sequences (each 10,199 bp long) simulated by Evolver (35) with parameters taken from a typical plasmid-borne gene (*a39*, with $T_S = 1.85$, %GC = 21.3%, and a transition-to-transversion ratio of 3.66). (D) Sensitivity of statistical power (*y* axis, calculated by equation 2 in Materials and Methods) to phylogenetic diversity (*x* axis, measured by T_S). Vertical gray lines indicate subtree distances from B31 up to a labeled strain.

and more reliable functional IGS elements on the main chromosome as well as on the plasmids.

(iii) The binomial model is robust to substitution models. PCIB counts obtained from simulated IGS sequences using genes (*0457* and *a39*) having median T_S values are not significantly different from the counts obtained by equation 2 in Materials and Methods ($P = 0.1017$ and $P = 0.0924$, respectively, by paired *t* tests). The close match between the counts from realistically simulated sequences and counts from the simplest sequence evolution model indicate that the analytical model is robust to variations in the value of evolution parameters, such as unequal base frequencies, bias in transitions to transversions, and rate heterogeneities among sites. This result is consistent with the original binomial model, which similarly was shown through simulations to be robust to models of base substitutions (2). Counts from the analytical model and the simulations, however, deviate greatly from permutation-based counts (Fig. 3A and B). This large discrepancy is likely due to the fact that individual IGS sequences vary greatly in T_S, while the simulation and analytical results are based on a single T_S value, considering that the analytical model is sensitive to the T_S value (Fig. 1D).

RpoS-dependent promoter regions are conserved. (i) *cis*-regulatory sequences of *ospC*. On the cp26 plasmid, *ospC* is directly regulated by RpoS through its binding to a *cis*-acting promoter sequence (39–41). The −35/−10 promoter sequence of *ospC* is indeed highly conserved and contains PCIBs among the *B. burgdorferi sensu lato* species. Notably, the functionally critical C and T at −15 and −14, respectively, are constant among the ge-

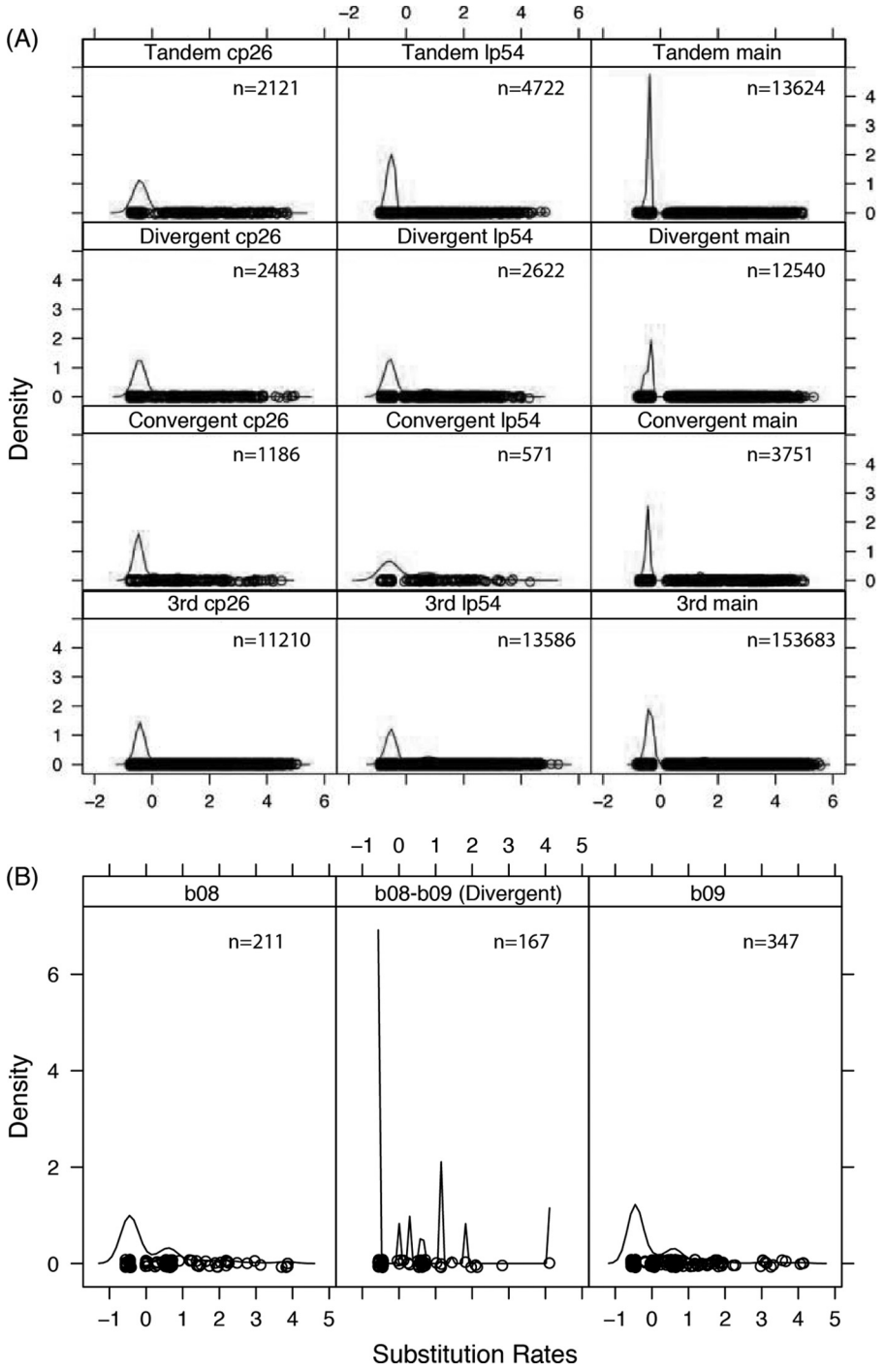

FIG 2 Frequency distributions of base substitution rates. (A) Normalized base substitution rates (*x* axis), obtained by using concatenated IGS-ORF alignments and calculated by Rates4site (38), are distributed similarly among the three types of IGSs (top three rows) and the third-base sites (bottom row). Chromosomal sequences (right column) are more conserved than plasmid-borne sequences (left and middle columns). (B) Substitution rates of a conserved divergent IGS (middle panel) consist of a significantly higher ($P = 5.6e{-}06$, by a Wilcoxon rank sum test) density of low-rate sites than its flanking third-base sites (left and right panels).

nomes (Fig. 4B). Our comparative analysis thus supports the functional importance of the RpoS recognition sequence. Further upstream of the RpoS recognition site, two sets of inverted repeats (IRs) function as operators for post-invasion repression of *ospC* (42–45). These IRs were not necessary for *ospC* induction in *trans*-complementation experiments but may be required for *cis* induction of *ospC* (19, 39, 41, 46). Additionally, RNA structural analysis using RNAz showed that IRs in *ospC* promoters of all *B. burgdorferi sensu lato* species form stable secondary structures, although their sequences are not conserved between the species (Fig. 4B).

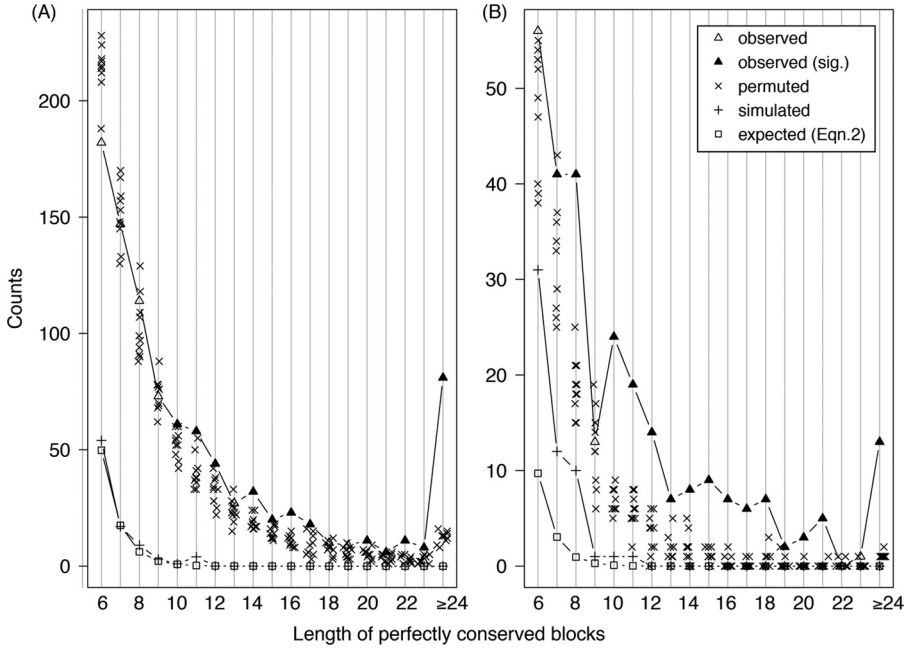

FIG 3 Observed and predicted counts of perfectly conserved intergenic blocks (PCIBs) on the chromosome (A) and plasmids (B) (note the scale difference of the *y* axis). A PCIB has no nucleotide variations or alignment gaps. The minimum length of a PCIB is six nucleotides. "Observed," length distribution of 935 PCIBs on the main chromosome and 276 PCIBs on the plasmids; "permuted," counts of *L*-mer PCIBs from 10 rounds of permutations of original IGS alignments; "simulated," PCIB counts from simulated sequences using Evolver (35); "expected," PCIB counts obtained by equation 2 in Materials and Methods. Solid triangles represent *L*-mers having significantly higher counts than permuted counts ($P < 0.001$ by one-tailed *t* tests).

This finding corroborates the suggestion in an earlier study which highlighted the functional significance of these IRs' secondary structures (46).

(ii) Newly identified putative RpoS-dependent promoters. Ten genes in B31 have been identified through a combination of genetic manipulations and quantitative PCR as being absolutely dependent on RpoS for their expression, including *ospC* on cp26 and *bba07*, *bba25* (*dbpB*)-*bba24* (*dbpA*), *bba34* (*oppA5*), and *bba66* on lp54 (27). The putative RpoS-dependent promoter regions consisting of −35 and −10 promoter sequences in the upstream of *ospC*, *a07*, *a25*, and *a34* are indeed highly conserved across the *B. burgdorferi sensu lato* species (Fig. 4A to D). In addition, the two inverted repeats upstream of *a25* (*dbpB*) are perfectly conserved across these species (Fig. 4C). Using a customized RpoS motif-searching script (see Materials and Methods), we identified two additional putative RpoS-dependent promoters upstream of *a36* and *a73* (Fig. 4D and E), both of which encode lipoproteins that are highly upregulated in the presence of RpoS (27). All these putative RpoS recognition sites are highly conserved among the orthologous IGS sequences but vary considerably among the coregulated genes, suggesting differential binding affinity to RpoS. The WebLogo analysis showed significant nucleotide conservation at the −10 and −35 sites, while the intervening region between these two sites varies in sequence as well as in length (Fig. 4F).

Putative noncoding RNAs and coregulated genes. Gene regulation is often associated with *cis*-acting sequences and *trans*-acting proteins that cooperatively affect the function of RNA polymerase. A number of studies have identified functional *cis*- and *trans*-acting elements that are critical to the regulation of virulent genes in *B. burgdorferi* and other pathogens (19, 41–43, 46). Although PCIBs include RpoS recognition and other known *cis*-regulatory sequences, regulatory sequences are not necessarily perfectly conserved among the species and even less so among the coregulated genes. For example, the RpoS recognition sequences varied considerably among coregulated genes (Fig. 4).

To further identify putative regulatory sequences and coregulated genes, we performed a stand-alone NCBI-BLAST (47) search for statistically significant matches among the IGS sequences on the *B. burgdorferi* B31 core genome. Close to 1,610 matches were identified using an *E* value cutoff of 0.01. Under the assumption that regulatory sequences are highly conserved among orthologs, we retained only BLAST hits with an average between-species sequence identity of 90% or more for both the query and subject sequences and matches occurring within 125 nucleotides of flanking genes. We also removed BLAST matches with query or subject sequences located in regions with more than 10% gapped alignment sites. A total of 393 unique BLAST matches remained after these conservation-based filtering procedures, which include 40 self-matching palindromic sequences and 353 other sequences. These 393 BLAST hits are likely to be regulatory sequences, because they are not only highly conserved between species but also either self-matching palindromes or similarly oriented with respect to their downstream ORFs.

(i) Coregulated genes. Table S2 in the supplemental material lists four examples of putative coregulatory gene sets that are supported by multiple lines of evidence, including (i) cross-species sequence conservation of the shared IGS elements, (ii) being approximately equally distant from the downstream genes, and (iii) similar biological functions of downstream genes. The 5′ ACATT TAAAATA 3′ motif shared between *a07* and *a73* may contribute to their RpoS-mediated upregulation (27). A shared 15-base 5′

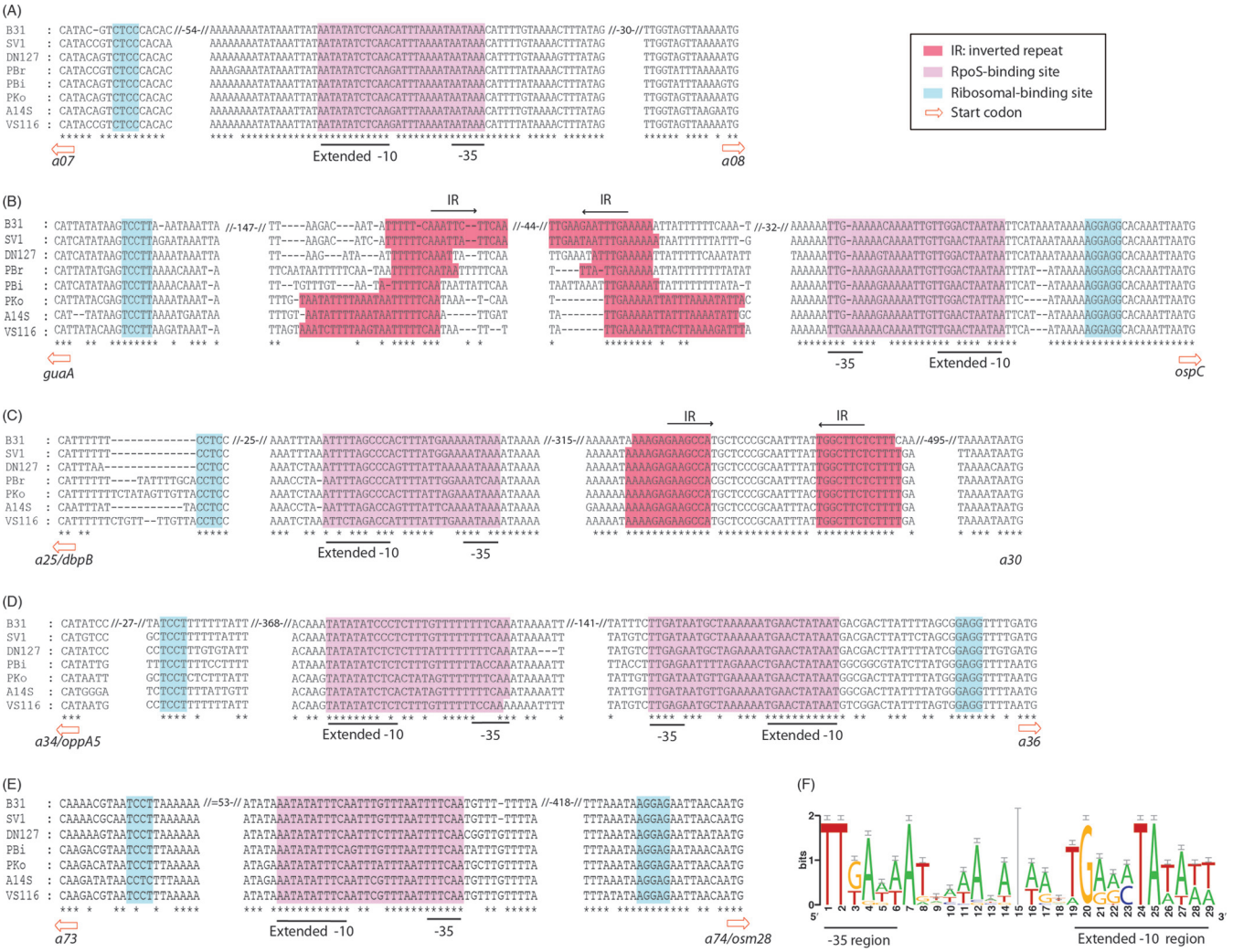

FIG 4 Conserved RpoS-dependent promoter sequences. RpoS recognition sites, ribosome-binding sites, and inverted repeats are conserved among *B. burgdorferi sensu lato* species in the promoters of six genes known to be upregulated during transmission from the tick to mammalian hosts. A new putative RpoS consensus sequence is derived (inset). The inverted repeats upstream of *ospC* are conserved in secondary structure but not in primary sequence.

ATCTTATAATATAAT 3′ motif upstream of *a15* (*ospA*) and *b05* (*chbA*) hints at possible coregulation of these genes. The 21-base-long, G-rich motif shared between *a62* (*lp6.6*) and *a74* (*osm28*) may contribute to their RpoS-mediated repression in mammals besides the conserved T-rich motif (27). It is equally interesting to note that the same 13-base sequence, 5′ ACTTTACTTTTTT 3′, resides at similar locations in the upstream of *a18* and *b10*, the first of three plasmid-partitioning genes on lp54 and cp26, respectively.

(ii) Noncoding RNAs. Of the 40 identified palindromic sequences, 28 were significantly predicted (RNA class probability of 0.9 or greater) by RNAz to form stable ncRNA secondary structures (see Table S3 in the supplemental material). The biological significance of these putative ncRNAs inferred using the Infernal software (48) included clustered regularly interspaced short palindromic repeat (CRISPR)-RNA direct repeat elements, a novel finding in *B. burgdorferi*. The functional significance of the longest putative ncRNAs is further highlighted by the compensatory changes within the stem regions and the high variability within the loop regions (Fig. 5). Note that the *0577-0578* IGS overlaps with $DsrA_{Bb}$, a small ncRNA that is responsible for temperature-dependent production of RpoS (49). It is not identified by our methods due to an overlap of 29 nucleotides with the 3′-end sequence of *0577*. Nevertheless, the alignment of *0577-0578* IGS sequences (not shown) reveals a high degree of sequence conservation across the eight species in the region that binds to the *rpoS* transcript, except that a TTAAA tandem repeat (previously known as the TAAAT repeat [50]) varies with 3 to 7 copies among the *B. burgdorferi sensu stricto* and *Borrelia finlandensis* strains. This genetic variation appears to be evolutionarily recent and specific to the aforementioned lineages, since there is no length variation in the same region among other species.

(iii) Promoters and terminators. The IGS sequences were tested for the presence of promoters and transcription terminators (see Materials and Methods). Overall, 57 such putative elements within 125 nucleotides of the flanking ORFs were identified, including 39 predicted promoters (including the well-studied *ospC* promoter) and 19 putative transcription terminators

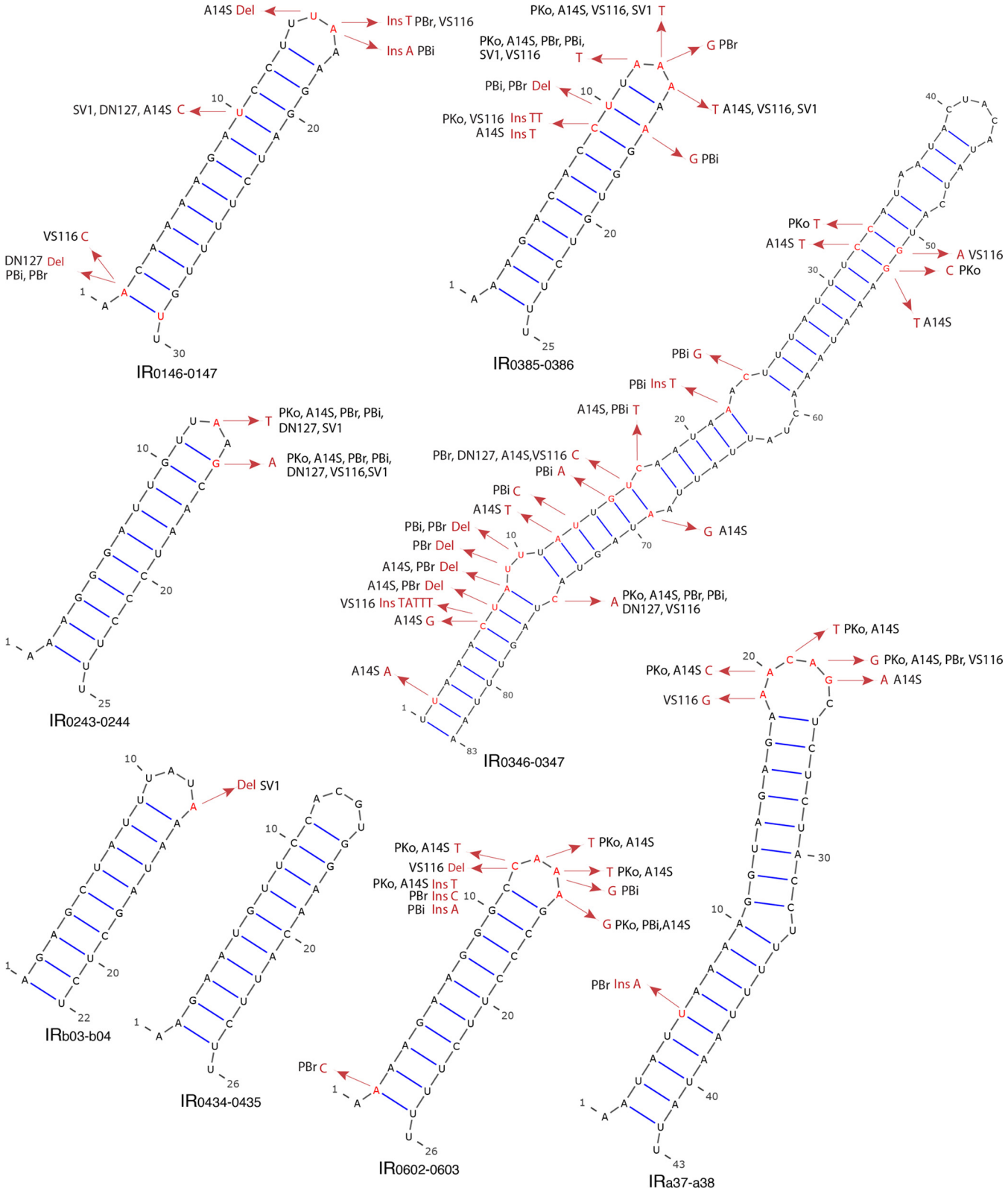

FIG 5 Predicted secondary structures of highly conserved putative ncRNAs. Structures of these eight longest inverted repeats (IRs) were predicted using RNAz (61) and plotted with B31 sequences using Varna (62). Arrows point to variations in the indicated strains. The Rfam accessions and annotations based on searches using Infernal (48) are as follows: $IR_{0146\text{-}0147}$-RF00082, small RNA G (SraG); $IR_{0243\text{-}0244}$-RF02152, long noncoding RNA (MINT_2); $IR_{0434\text{-}0435}$-RF00074, pre-miRNA (mir-29); $IR_{0346\text{-}0347}$-RF01350, CRISPR direct repeat element (CRISPR-DR41); $IR_{0385\text{-}0386}$-RF01379, CRISPR direct repeat element (CRISPR-DR66); $IR_{0602\text{-}0603}$-RF02066, bacterial small RNAs (STnc320); $IR_{a37\text{-}a38}$, RF02058-bacterial small RNAs (STnc400); and $IR_{b03\text{-}b04}$, RF00741-pre-miRNA (mir-378). Structures of another six long conserved IRs in *Borrelia* ($IR_{b04\text{-}b05}$, $IR_{b12\text{-}b13}$, $IR_{b29\text{-}b01}$, $IR_{a16\text{-}a18}$, $IR_{a21\text{-}a23}$, and $IR_{a34\text{-}a36}$) have been published earlier (32, 33).

TABLE 2 Predicted regulatory elements

IGS[a]	Orientation	ncRNA[b]	Promotor[c]	Terminator[d]
a01-a03	Tandem			+
a03-a04	Convergent			+
a05-a07	Tandem		+	+
a07-a08	Divergent		+	
a14-a15	Tandem		+	
a16-a18	Tandem	+	+	+
a21-a23	Tandem	+		
a25-a30	Divergent		+	
a34-a36	Divergent	+	+	
a37-a38	Tandem	+		+
a61-a62	Divergent		+	
a62-a64	Convergent		+	
a64-a65	Tandem			+
a73-a74	Divergent		+	
b03-b04	Tandem	+	+	
b04-b05	Divergent	+	+	+
B12-B13	Tandem	+		
B13-B14	Convergent			+
b16-b17	Convergent	+		+
B18-b19	Divergent	+	+	+
b19-b22	Convergent			++(2)
b28-b29	Tandem	+	+	
b29-b01	Tandem	+		+
0089-0090	Tandem	+		
0100-0101	Divergent		+	
0103-0104	Tandem		+	
0135-0136	Divergent		+	
0146-0147	Tandem	+		
0166-0167	Divergent		+	
0190-0192	Divergent		+	
0195-0196	Divergent		+	
0214-0215	Divergent		+	
0236-0237	Divergent		+	
0239-0240	Divergent		+	
0243-0244	Convergent	+		
0247-0248	Tandem	+		
0253-0254	Divergent		+	
0327-0328	Divergent	+		
0346-0347	Divergent	+	+	
0364-0365	Convergent	+		
0384-0385	Tandem		+	
0385-0386	Tandem	+		
0408-0409	Divergent	+		
0421-0422	Tandem	+	+	
0434-0435	Tandem	+		
0436-0437	Divergent	+	+	
0437-0438	Tandem		+	
0460-0461	Divergent	+		+
0472-0473	Tandem			+
0536-0537	Convergent			+
0543-0544	Tandem		+	
0551-0552	Divergent		+	
0571-0572	Divergent		+	
0574-0575	Tandem		+	
0577-0578[e]	Tandem		+	
0596-0597	Divergent		+	
0602-0603	Convergent	+		
0603-0604	Tandem	+	+	+
0608-0610	Tandem			+
0620-0621	Divergent		+	
0676-0677	Divergent		+	
0723-0724	Divergent	+		
0744-0745	Tandem			+
0772-0773	Tandem		+	
0775-0776	Divergent		+	
0828-0829	Divergent	+		

TABLE 2 (Continued)

[a] Including IGSs on chromosome, lp54, and cp26 that are ≥30 nucleotides and present in at least seven of the eight *B. burgdorferi sensu lato* species.
[b] Presence (+; $n = 28$) of a conserved RNA structure predicted by RNAz (61). Sequences are available in Table S3 in the supplemental material.
[c] Presence (+; $n = 39$) of a conserved promoter predicted by PromPredict (51). Sequences are available in Table S4 in the supplemental material.
[d] Presence (+; $n = 19$) of a conserved transcription terminator predicted by TransTermHP (52). Sequences are available in Table S5 in the supplemental material.
[e] *0577-0578* contains $DsrA_{Bb}$, a small ncRNA that regulates *rpoS* expression. It is not identified here due to an overlap with the 3′ end of *0577* (49).

(see Tables S4 and S5 in the supplemental material). Note that these predicted IGS terminators do not include an intragenic terminator that is a part of the *bmpB* (*bb_0382*) coding sequence (50). The alignment of *bmpB* sequences (not shown) displays an absence of nucleotide substitutions in the terminator region across all eight species, except at two opposite positions of the stem region. These two sites show an A-T pairing in six species, two compensatory changes resulting in a G-C pairing in *B. burgdorferi sensu stricto*, and one substitution resulting in a G-T mismatch in *Borrelia bissettii* DN127. All variations at these two sites are synonymous. Strong sequence conservation and compensatory substitutions support the functional importance of this and other terminators as a mechanism for regulating differential expression of cotranscribed genes in *Borrelia* (50).

Concluding remarks. The present study is the first systematic search of gene regulatory elements in *B. burgdorferi* using a large number of genomes. Previous efforts were either based on a limited number of genomes (32) or using plasmid sequences only (33). The phylogenomic search identified a large number of candidate *cis*-regulatory (Table 2, Fig. 4) and *trans*-regulatory (Fig. 5) elements that are highly conserved among *B. burgdorferi sensu lato* species. We caution, however, that regulatory elements may not be conserved in primary sequences. For example, the RpoS-binding sites span across a variable region (Fig. 4F). The inverted repeats upstream of *ospC* are conserved in the secondary structure but not in the primary sequence (Fig. 4B). Computational approaches not based on sequence conservation, such as PromPredict (51) and TransTermHP (52), are therefore valuable complementary tools for predicting regulatory elements. To aid future computational and experimental characterization of the genome-wide regulatory network in *B. burgdorferi sensu lato*, we released all IGS alignments on BorreliaBase.org, a publicly accessible online database of orthologous ORFs and IGSs in *Borrelia* (53). The website will be periodically updated to include newly released *Borrelia* genomes.

The statistical approach we developed here based on an earlier model (2) suggests that genome sequences from additional *B. burgdorferi sensu lato* species are needed to identify IGS elements shorter than five bases and to further reduce false discovery rates, especially for those on the main chromosome (Fig. 1 and 3). The GERP++ tool, which similarly estimates statistical significance using empirically calculated neutral substitution rates, identifies putatively functional elements conserved among vertebrates in a more automated fashion (5). In the future, one may consider adapting the GERP++ approach to identify functional IGS elements in bacterial genomes by using synonymous tree lengths as neutral expectations. For now, we expect the proposed statistical framework to be helpful for estimating the false discovery rates of conserved IGS sequences as well as for determining the number of

genomes (and at what phylogenetic levels) necessary for achieving a certain level of statistical significance in *Borrelia* and other bacterial species.

MATERIALS AND METHODS

Identification of orthologous IGSs. (i) Consensus start positions. We used these orthologous ORFs as anchors for identifying orthologous IGS sequences based on the assumption that IGS sequences are orthologous if they are flanked by orthologous ORFs (11). A major problem in IGS identification is the inconsistent start codon positions among the orthologous ORFs, each of which had been predicted independently by the program Glimmer3 (54). In fact, one important rationale for sequencing multiple genomes of a single species or species group is to improve the prediction of genes and their start codon positions (55). To minimize the erroneous mixing of true IGSs and sequences that may in fact be a part of ORFs, we identified a consensus start codon position for each orthologous ORF family based on the majority of predicted start codon positions among (but not within) *B. burgdorferi sensu lato* species. After the identification of a consensus start codon position for each orthologous ORF family, we used a customized Perl script based on BioPerl (56) to extract orthologous IGSs. IGS sequences were aligned directly with MUSCLE (57), while flanking ORF sequences were aligned according to the MUSCLE alignment of translated protein sequences. IGS loci were categorized into three types based on their orientation relative to the transcription directions of its two flanking ORFs: a "divergent" IGS is located at the 5′ ends of both flanking ORFs, a "tandem" IGS at the 5′ end of one of the two flanking ORFs and the 3′ end of another flanking ORF, and a "convergent" IGS at the 3′ ends of both flanking ORFs.

(ii) Filtering by length. In identifying reliable IGSs, we used only long (≥150-base) orthologous ORFs and those that are present in at least seven of the eight genome-sequenced *B. burgdorferi sensu lato* species. ORFs with limited phylogenetic presence are likely to be erroneously predicted, and their flanking IGSs were excluded from analysis. We further excluded IGS loci with an alignment length of 30 or fewer bases. Short IGSs are likely to be between genes that are cotranscribed (e.g., part of an operon) and thus lacking regulatory elements.

Identification of conserved IGS elements. (i) Substitution rates. We identified evolutionarily conserved IGSs by coestimating per-site nucleotide substitution rates for an IGS with its two flanking ORFs. At each IGS locus, alignments of the IGS sequences and two flanking ORF sequences were concatenated using a customized Perl script. Nucleotide substitution rates were subsequently estimated using Rates4site with the HKY model and 16 discrete categories (38). Customized Perl scripts were then used to extract per-site substitution rates at the IGS as well as at the first, second, and third codon positions of the flanking ORFs. Conserved IGS sequences were identified as those having significantly lower substitution rates than the flanking third codon positions by *t* tests or Wilcoxon rank sum tests (nonparametric equivalent of *t* test) in an R statistical environment (58).

(ii) Power analysis. The statistical power of detecting conserved noncoding sequences using phylogenetic footprinting increases with the length of conserved elements, the number of genomes, and the evolutionary distance. In a hypothetical, simplified case of a sequence with a length of L nucleotides evolving under the Jukes-Cantor model and using a group of N equally related genomes, each of which deviates from an ancestral sequence by a distance of D substitutions, the probability of false positives (FP; i.e., selectively neutral elements misidentified as conserved sequences) is given by a cumulative binomial function, $\mathrm{FP} = P(\leq C) = \sum_{k=0}^{C}\left[\binom{NL}{k}\left(\frac{3}{4}-\frac{3}{4}e^{-\frac{4D}{3}}\right)^{k}\left(\frac{1}{4}+\frac{3}{4}e^{-\frac{4D}{3}}\right)^{NL-k}\right]$, where k is the number of base substitutions and C is the threshold number of base changes below which a sequence is considered conserved (2). For genomes related not by a star phylogeny, one may consider a Poisson model in which the probability that a nucleotide remains identical after evolving with an expected number of substitutions given by the neutral tree length T_0 is $P(k=0) = \frac{1}{4}+\frac{3}{4}e^{-\frac{4T_0}{3}}$. Using the synonymous tree length (T_S) of the flanking ORFs to approximate the neutral tree length (T_0), the statistical significance of deviation from the neutral expectation of an IGS displaying n substitutions is given by

$$P(\leq n) = \sum_{k=0}^{n}\left[\binom{L}{K}\left(\frac{3}{4}-\frac{3}{4}e^{-\frac{4T_s}{3}}\right)^{k}\left(\frac{1}{4}+\frac{3}{4}e^{-\frac{4T_s}{3}}\right)^{L-k}\right] \quad (1)$$

The number n can be estimated either by the number of variable sites or by the total tree length of the IGS itself. In the special case of $n = 0$ (i.e., an absence of base substitution) in such an L-mer sequence,

$$FP = P(n=0) = \left(\frac{1}{4}+\frac{3}{4}e^{-\frac{4T_s}{3}}\right)^{L} \quad (2)$$

This FP discovery rate decreases with increasing L, thus defining the minimum length of functional L-mer conserved elements that can be identified given a set of genomes with a total phylogenetic diversity measured by T_S. We used the CODEML program of the PAML (version 4.8) package (35) to obtain the T_S and nonsynonymous tree lengths (T_N) for each flanking ORF family. The R *pbinom* function was used to obtain the cumulative binomial probabilities (58).

(iii) Validation by simulations. We tested the validity of the T_S-based binomial model with simulated sequences generated by the Evolver program of the PAML (version 4.8) package (35). Option 5 of Evolver simulates the evolution of noncoding nucleotide sequences given user-specified phylogeny, length of sequences, total tree length, and a nucleotide substitution model. For the phylogeny and total tree length, we used those estimated by the CODEML program of PAML for a typical ORF (with a T_S close to the median) on the main chromosome or plasmid. For the nucleotide substitution model, we used an HKY model with base frequencies, transition-to-transversion ratio (κ), and rate heterogeneity parameters (α and γ) estimated by CODEML for the same ORF. Counts of perfectly conserved L-mer sequences were compared with the analytically predicted counts (equation 2) as a test of the validity as well as the robustness of the analytical model, which is based on the Juke-Cantor model, the simplest of nucleotide substitution models.

(iv) Permutation test. Statistical significance of conserved IGS sequences was also empirically estimated by permuting IGS alignments (11). We used customized Perl scripts to permute each IGS alignment 10 times and extracted all ungapped segments 6 bases or longer that were perfectly conserved among all sequenced genomes. The numbers of occurrences of L-mer perfectly conserved IGS blocks (PCIBs) were then compared with the observed numbers with one-tailed *t* tests. L-mers that are significantly more numerous than permuted counts have relatively low false-positive discovery rates.

Prediction of regulatory IGS sequences. (i) Ribosome-binding sites, promoters, and intrinsic terminators. To identify putative functional elements contributing to the evolutionary conservation of IGSs, we tested for the presence of ribosome-binding sites (RBS), promoters, intrinsic transcription terminators, noncoding RNAs (ncRNAs), and RpoS recognition sites. Only elements discovered within 125 nucleotides from each flanking ORF and present in all studied strains were reported. This allowed us to filter out recently pseudogenized sequences, which tend to be conserved and closer to the center of a long IGS. The RBS profile specific for *B. burgdorferi sensu lato* was identified using the RBSFinder algorithm with the 16S rRNA sequence of *B. burgdorferi* B31 as the reference and the 5′ upstream sequences of 26 ORF sequences on the cp26 plasmid of *B. burgdorferi* B31 as sample sequences (59). PromPredict (version 1.0), which detects differences in free energy between promoter and nonpromoter regions in bacterial genomes, was used to predict promoter sequences in individual IGS sequences (51). Promoters were reported only if they were identified in all orthologous IGS sequences. We considered known promoter regions (e.g., *ospC* and *dbpA*) as our positive controls and convergent IGS segments within consecutive ORFs as negative controls. TransTermHP (version 1.0), which identifies the pattern of a hair-

pin loop followed by a thymine-rich segment, was employed to predict Rho-independent transcription terminators (52). Terminators were reported only if they were identified in all orthologous IGS sequences.

(ii) RpoS recognition sites. To identify potential RpoS recognition sites, we first used previously published RpoS binding sequences in *B. burgdorferi* B31 (27) as the query sequences to search among orthologous IGS sequences using NCBI-BLASTN (with the megablast and 30 no-dust options) (47). We obtained a new consensus sequence by aligning predicted RpoS sequences (one representative strain per species for each IGS locus) with MUSCLE (57). The consensus sequence of these predicted RpoS recognition sites was obtained and visualized using WebLogo (version 2.8.2) (60).

(iii) Noncoding RNAs and coregulated genes. Based on the B31 IGS sequences, we used NCBI-BLASTN to identify similar segments with the following parameters: *-task* "blastn-short," -dust *0* (no sequence filtering), *-evalue 1e−5* (an expect value of 10^{-5}), -word_*size 5* (word size 5) (47). The BLASTN protocol identified IGS segments that are either similar to each other or self-similar palindromes. We retained only IGS segments that are highly conserved, showing a 90% or higher average sequence identity between the eight *B. burgdorferi sensu lato* species and 10% or less gapped alignment sites. Conserved palindromes were retained as putative ncRNAs (32, 33). RNAz (version 2.1) was used to identify the conserved secondary structure of the putative ncRNA (61). The consensus secondary structure of the predicted ncRNA elements was plotted using Varna (62). Infernal (version 1.1), which implements covariance models to search DNA sequence databases for similar RNA structures and sequences, was used to infer functions of putative ncRNAs (48). Genes sharing a conserved IGS segment were considered potential members of a coregulated network (1, 3).

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at http://mbio.asm.org/lookup/suppl/doi:10.1128/mBio.00011-15/-/DCSupplemental.

Table S1, DOCX file, 0.2 MB.
Table S2, DOCX file, 0.1 MB.
Table S3, DOCX file, 0.1 MB.
Table S4, DOCX file, 0.2 MB.
Table S5, DOCX file, 0.1 MB.

ACKNOWLEDGMENTS

This work was supported by Public Health Service grants AI107955 (to W.-G.Q.), AI37256 (B.J.L.), AI49003 (S.R.C.), and AI30071 (C.M.F. and S.E.S.) from the National Institute of Allergy and Infectious Diseases (NIAID) and grant MD007599 (Hunter College) from the National Institute on Minority Health and Health Disparities (NIMHD) of the National Institutes of Health (NIH).

The content of the manuscript is solely the responsibility of the authors and does not necessarily represent the official views of NIAID, NIMHD, or NIH.

REFERENCES

1. **Brohée S, Janky R, Abdel-Sater F, Vanderstocken G, André B, van Helden J.** 2011. Unraveling networks of co-regulated genes on the sole basis of genome sequences. Nucleic Acids Res **39:**6340–6358. http://dx.doi.org/10.1093/nar/gkr264.
2. **Eddy SR.** 2005. A model of the statistical power of comparative genome sequence analysis. PLoS Biol **3:**e10. http://dx.doi.org/10.1371/journal.pbio.0030010.
3. **Su J, Teichmann SA, Down TA.** 2010. Assessing computational methods of *cis*-regulatory module prediction. PLoS Comput Biol **6:**e1001020. http://dx.doi.org/10.1371/journal.pcbi.1001020.
4. **Kumar S, Filipski AJ, Battistuzzi FU, Kosakovsky Pond SL, Tamura K.** 2012. Statistics and truth in phylogenomics. Mol Biol Evol **29:**457–472. http://dx.doi.org/10.1093/molbev/msr202.
5. **Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S.** 2010. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol **6:**e1001025. http://dx.doi.org/10.1371/journal.pcbi.1001025.
6. **Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P, Ernst J, Jordan G, Mauceli E, Ward LD, Lowe CB, Holloway AK, Clamp M, Gnerre S, Alföldi J, Beal K, Chang J, Clawson H, Cuff J, Di Palma F, Fitzgerald S, Flicek P, Guttman M, Hubisz MJ, Jaffe DB, Jungreis I, Kent WJ, Kostka D, Lara M, Martins AL, Massingham T, Moltke I, Raney BJ, Rasmussen MD, Robinson J, Stark A, Vilella AJ, Wen J, Xie X, Zody MC, Broad Institute Sequencing Platform and Whole Genome Assembly Team, Worley KC, Kovar CL, Muzny DM, Gibbs RA, Baylor College of Medicine Human Genome Sequencing Center Sequencing Team, Warren WC, Mardis ER, Weinstock GM, Wilson RK, Genome Institute at Washington University, Birney E, Margulies EH, Herrero J, Green ED, Haussler D, Siepel A, Goldman N, Pollard KS, Pedersen JS, Lander ES, Kellis M.** 2011. A high-resolution map of human evolutionary constraint using 29 mammals. Nature **478:**476–482. http://dx.doi.org/10.1038/nature10530.
7. **Andolfatto P.** 2005. Adaptive evolution of non-coding DNA in *Drosophila*. Nature **437:**1149–1152. http://dx.doi.org/10.1038/nature04107.
8. **Stark A, Lin MF, Kheradpour P, Pedersen JS, Parts L, Carlson JW, Crosby MA, Rasmussen MD, Roy S, Deoras AN, Ruby JG, Brennecke J, Harvard FlyBase Curators, Berkeley Drosophila Genome Project, Hodges E, Hinrichs AS, Caspi A, Paten B, Park SW, Han MV, Maeder ML, Polansky BJ, Robson BE, Aerts S, van Helden J, Hassan B, Gilbert DG, Eastman DA, Rice M, Weir M, Hahn MW, Park Y, Dewey CN, Pachter L, Kent WJ, Haussler D, Lai EC, Bartel DP, Hannon GJ, Kaufman TC, Eisen MB, Clark AG, Smith D, Celniker SE, Gelbart WM, Kellis M.** 2007. Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. Nature **450:**219–232. http://dx.doi.org/10.1038/nature06340.
9. **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. http://dx.doi.org/10.1126/science.1084337.
10. **Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES.** 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature **423:**241–254. http://dx.doi.org/10.1038/nature01644.
11. **Degnan PH, Ochman H, Moran NA.** 2011. Sequence conservation and functional constraint on intergenic spacers in reduced genomes of the obligate symbiont *Buchnera*. PLoS Genet **7:**e1002252. http://dx.doi.org/10.1371/journal.pgen.1002252.
12. **Kurtenbach K, Hanincová K, Tsao JI, Margos G, Fish D, Ogden NH.** 2006. Fundamental processes in the evolutionary ecology of Lyme borreliosis. Nat Rev Microbiol **4:**660–669. http://dx.doi.org/10.1038/nrmicro1475.
13. **Margos G, Vollmer SA, Ogden NH, Fish D.** 2011. Population genetics, taxonomy, phylogeny and evolution of *Borrelia burgdorferi sensu lato*. Infect Genet Evol **11:**1545–1563. http://dx.doi.org/10.1016/j.meegid.2011.07.022.
14. **Bacon RM, Kugeler KJ, Mead PS, Centers for Disease Control and Prevention (CDC).** 2008. Surveillance for Lyme disease—United States, 1992–2006. MMWR Surveill Summ **57:**1–9.
15. **Dykhuizen DE, Brisson D, Sandigursky S, Wormser GP, Nowakowski J, Nadelman RB, Schwartz I.** 2008. The propensity of different *Borrelia burgdorferi sensu stricto* genotypes to cause disseminated infections in humans. Am J Trop Med Hyg **78:**806–810.
16. **Wang I-N, Dykhuizen DE, Qiu W, Dunn JJ, Bosler EM, Luft BJ.** 1999. Genetic diversity of *ospC* in a local population of *Borrelia burgdorferi sensu stricto*. Genetics **151:**15–30.
17. **Wormser GP, Brisson D, Liveris D, Hanincová K, Sandigursky S, Nowakowski J, Nadelman RB, Ludin S, Schwartz I.** 2008. *Borrelia burgdorferi* genotype predicts the capacity for hematogenous dissemination during early Lyme disease. J Infect Dis **198:**1358–1364. http://dx.doi.org/10.1086/592279.
18. **Radolf JD, Caimano MJ, Stevenson B, Hu LT.** 2012. Of ticks, mice and men: understanding the dual-host lifestyle of Lyme disease spirochaetes. Nat Rev Microbiol **10:**87–99. http://dx.doi.org/10.1038/nrmicro2714.
19. **Samuels DS.** 2011. Gene regulation in *Borrelia burgdorferi*. Annu Rev Microbiol **65:**479–499. http://dx.doi.org/10.1146/annurev.micro.112408.134040.
20. **Iyer R, Caimano MJ, Luthra A, Axline D, Corona A, Iacobas DA, Radolf JD, Schwartz I.** 2015. Stage-specific global alterations in the transcriptomes of Lyme disease spirochetes during tick feeding and following

mammalian host adaptation. Mol Microbiol **95:**509–538. http://dx.doi.org/10.1111/mmi.12882.
21. **Brooks CS, Hefty PS, Jolliff SE, Akins DR.** 2003. Global analysis of *Borrelia burgdorferi* genes regulated by mammalian host-specific signals. Infect Immun **71:**3371–3383. http://dx.doi.org/10.1128/IAI.71.6.3371-3383.2003.
22. **Fisher MA, Grimm D, Henion AK, Elias AF, Stewart PE, Rosa PA, Gherardini FC.** 2005. *Borrelia burgdorferi* σ54 is required for mammalian infection and vector transmission but not for tick colonization. Proc Natl Acad Sci U S A **102:**5162–5167. http://dx.doi.org/10.1073/pnas.0408536102.
23. **Ouyang Z, Blevins JS, Norgard MV.** 2008. Transcriptional interplay among the regulators Rrp2, RpoN and RpoS in *Borrelia burgdorferi*. Microbiology **154:**2641–2658. http://dx.doi.org/10.1099/mic.0.2008/019992-0.
24. **Dunham-Ems SM, Caimano MJ, Eggers CH, Radolf JD.** 2012. *Borrelia burgdorferi* requires the alternative sigma factor *rpoS* for dissemination within the vector during tick-to-mammal transmission. PLoS Pathog **8:**e1002532. http://dx.doi.org/10.1371/journal.ppat.1002532.
25. **Ouyang Z, Narasimhan S, Neelakanta G, Kumar M, Pal U, Fikrig E, Norgard MV.** 2012. Activation of the RpoN-RpoS regulatory pathway during the enzootic life cycle of *Borrelia burgdorferi*. BMC Microbiol **12:**44. http://dx.doi.org/10.1186/1471-2180-12-44.
26. **Tilly K, Bestor A, Rosa PA.** 2013. Lipoprotein succession in *Borrelia burgdorferi*: similar but distinct roles for OspC and VlsE at different stages of mammalian infection. Mol Microbiol **89:**216–227. http://dx.doi.org/10.1111/mmi.12271.
27. **Caimano MJ, Iyer R, Eggers CH, Gonzalez 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. http://dx.doi.org/10.1111/j.1365-2958.2007.05860.x.
28. **Kraiczy P, Stevenson B.** 2013. Complement regulator-acquiring surface proteins of *Borrelia burgdorferi*: structure, function and regulation of gene expression. Ticks Tick-Borne Dis **4:**26–34. http://dx.doi.org/10.1016/j.ttbdis.2012.10.039.
29. **Xu Q, Shi Y, Dadhwal P, Liang FT.** 2012. RpoS regulates essential virulence factors remaining to be identified in *Borrelia burgdorferi*. PLoS One **7:**e53212. http://dx.doi.org/10.1371/journal.pone.0053212.
30. **Mongodin EF, Casjens SR, Bruno JF, Xu Y, Drabek EF, Riley DR, Cantarel BL, Pagan PE, Hernandez YA, Vargas LC, Dunn JJ, Schutzer SE, Fraser CM, Qiu W-G, Luft BJ.** 2013. Inter- and intra-specific pan-genomes of *Borrelia burgdorferi sensu lato*: genome stability and adaptive radiation. BMC Genomics **14:**693. http://dx.doi.org/10.1186/1471-2164-14-693.
31. **Norris SJ, Lin T.** 2011. Out of the woods: the remarkable genomes of the genus *Borrelia*. J Bacteriol **193:**6812–6814. http://dx.doi.org/10.1128/JB.06317-11.
32. **Delihas N.** 2009. Intergenic regions of *Borrelia* plasmids contain phylogenetically conserved RNA secondary structure motifs. BMC Genomics **10:**101. http://dx.doi.org/10.1186/1471-2164-10-101.
33. **Qiu W-G, Martin CL.** 2014. Evolutionary genomics of *Borrelia burgdorferi sensu lato*: findings, hypotheses, and the rise of hybrids. Infect Genet Evol **27:**576–593. http://dx.doi.org/10.1016/j.meegid.2014.03.025.
34. **Haven J, Vargas LC, Mongodin EF, Xue V, Hernandez Y, Pagan P, Fraser-Liggett CM, Schutzer SE, Luft BJ, Casjens SR, Qiu W-G.** 2011. Pervasive recombination and sympatric genome diversification driven by frequency-dependent selection in *Borrelia burgdorferi*, the Lyme disease bacterium. Genetics **189:**951–966. http://dx.doi.org/10.1534/genetics.111.130773.
35. **Yang Z.** 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol **24:**1586–1591. http://dx.doi.org/10.1093/molbev/msm088.
36. **Ivanova LB, Tomova A, González-Acuña D, Murúa R, Moreno CX, Hernández C, Cabello J, Cabello C, Daniels TJ, Godfrey HP, Cabello FC.** 2014. *Borrelia chilensis*, a new member of the *Borrelia burgdorferi sensu lato* complex that extends the range of this genospecies in the southern hemisphere. Environ Microbiol **16:**1069–1080. http://dx.doi.org/10.1111/1462-2920.12310.
37. **Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, Ward LD, Birney E, Crawford GE, Dekker J, Dunham I, Elnitski LL, Farnham PJ, Feingold EA, Gerstein M, Giddings MC, Gilbert DM, Gingeras TR, Green ED, Guigo R, Hubbard T, Kent J, Lieb JD, Myers RM, Pazin MJ, Ren B, Stamatoyannopoulos JA, Weng Z, White KP, Hardison RC.** 2014. Defining functional DNA elements in the human genome. Proc Natl Acad Sci U S A **111:**6131–6138. http://dx.doi.org/10.1073/pnas.1318948111.
38. **Mayrose I, Graur D, Ben-Tal N, Pupko T.** 2004. Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. Mol Biol Evol **21:**1781–1791. http://dx.doi.org/10.1093/molbev/msh194.
39. **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. http://dx.doi.org/10.1128/JB.186.21.7390-7402.2004.
40. **Marconi RT, Samuels DS, Garon CF.** 1993. Transcriptional analyses and mapping of the *ospC* gene in Lyme disease spirochetes. J Bacteriol **175:**926–932.
41. **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:**4822–4829. http://dx.doi.org/10.1128/JB.187.14.4822-4829.2005.
42. **Xu Q, McShan K, Liang FT.** 2008. Verification and dissection of the *ospC* operator by using *flaB* promoter as a reporter in *Borrelia burgdorferi*. Microb Pathog **45:**70–78. http://dx.doi.org/10.1016/j.micpath.2008.03.002.
43. **Xu Q, McShan K, Liang FT.** 2007. Identification of an *ospC* operator critical for immune evasion of *Borrelia burgdorferi*. Mol Microbiol **64:**220–231. http://dx.doi.org/10.1111/j.1365-2958.2007.05636.x.
44. **Margolis N, Hogan D, Cieplak W, Schwan TG, Rosa PA.** 1994. Homology between *Borrelia burgdorferi* OspC and members of the family of *Borrelia hermsii* variable major proteins. Gene **143:**105–110. http://dx.doi.org/10.1016/0378-1119(94)90613-0.
45. **Tilly K, Casjens S, Stevenson B, Bono JL, Samuels DS, Hogan D, Rosa P.** 1997. The *Borrelia burgdorferi* circular plasmid cp26: conservation of plasmid structure and targeted inactivation of the *ospC* gene. Mol Microbiol **25:**361–373. http://dx.doi.org/10.1046/j.1365-2958.1997.4711838.x.
46. **Drecktrah D, Hall LS, Hoon-Hanks LL, Samuels DS.** 2013. An inverted repeat in the *ospC* operator is required for induction in *Borrelia burgdorferi*. PLoS One **8:**e68799. http://dx.doi.org/10.1371/journal.pone.0068799.
47. **Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL.** 2009. BLAST+: architecture and applications. BMC Bioinformatics **10:**421. http://dx.doi.org/10.1186/1471-2105-10-421.
48. **Nawrocki EP, Eddy SR.** 2013. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics **29:**2933–2935. http://dx.doi.org/10.1093/bioinformatics/btt509.
49. **Lybecker MC, Samuels DS.** 2007. Temperature-induced regulation of RpoS by a small RNA in *Borrelia burgdorferi*. Mol Microbiol **64:**1075–1089. http://dx.doi.org/10.1111/j.1365-2958.2007.05716.x.
50. **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. http://dx.doi.org/10.1128/JB.187.8.2592-2600.2005.
51. **Rangannan V, Bansal M.** 2010. High-quality annotation of promoter regions for 913 bacterial genomes. Bioinformatics **26:**3043–3050. http://dx.doi.org/10.1093/bioinformatics/btq577.
52. **Kingsford CL, Ayanbule K, Salzberg SL.** 2007. Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. Genome Biol **8:**R22. http://dx.doi.org/10.1186/gb-2007-8-2-r22.
53. **Di L, Pagan PE, Packer D, Martin CL, Akther S, Ramrattan G, Mongodin EF, Fraser CM, Schutzer SE, Luft BJ, Casjens SR, Qiu W-G.** 2014. *BorreliaBase*: a phylogeny-centered browser of *Borrelia* genomes. BMC Bioinformatics **15:**233. http://dx.doi.org/10.1186/1471-2105-15-233.
54. **Delcher AL, Bratke KA, Powers EC, Salzberg SL.** 2007. Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics **23:**673–679. http://dx.doi.org/10.1093/bioinformatics/btm009.
55. **Wall ME, Raghavan S, Cohn JD, Dunbar J.** 2011. Genome majority vote improves gene predictions. PLoS Comput Biol **7:**e1002284. http://dx.doi.org/10.1371/journal.pcbi.1002284.
56. **Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H, Lehväslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka E, Wilkinson MD, Birney E.** 2002. The BioPerl toolkit: Perl modules for the life sciences. Genome Res **12:**1611–1618. http://dx.doi.org/10.1101/gr.361602.
57. **Edgar RC.** 2004. MUSCLE: multiple sequence alignment with high accu-

racy and high throughput. Nucleic Acids Res **32:**1792–1797. http://dx.doi.org/10.1093/nar/gkh340.
58. **R Core Team**. 2013. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Copenhagen, Denmark.
59. **Suzek BE, Ermolaeva MD, Schreiber M, Salzberg SL**. 2001. A probabilistic method for identifying start codons in bacterial genomes. Bioinformatics **17:**1123–1130. http://dx.doi.org/10.1093/bioinformatics/17.12.1123.
60. **Crooks GE, Hon G, Chandonia J-M, Brenner SE**. 2004. WebLogo: a sequence logo generator. Genome Res **14:**1188–1190. http://dx.doi.org/10.1101/gr.849004.
61. **Gruber AR, Findeiß S, Washietl S, Hofacker IL, Stadler PF**. 2010. RNAz 2. 0: improved noncoding RNA detection. Pac Symp Biocomput **2010:** 69–79. http://dx.doi.org/10.1142/9789814295291_0009.
62. **Darty K, Denise A, Ponty Y**. 2009. VARNA: interactive drawing and editing of the RNA secondary structure. Bioinformatics **25:**1974–1975. http://dx.doi.org/10.1093/bioinformatics/btp250.
63. **Pupko T, Bell RE, Mayrose I, Glaser F, Ben-Tal N**. 2002. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. Bioinformatics **18**(Suppl 1):S71–S77. http://dx.doi.org/10.1093/bioinformatics/18.suppl_1.S71.