Maximum antigen diversification in a lyme bacterial population and evolutionary strategies to overcome pathogen diversity

Lia Di1,2,1*, Saymon Akther2,7, Edgaras Bezrucenkovas1,1, Larisa Ivanova5,6, Brian Sulikow1,1, Bing Wu1, Saad Mneimneh2,4, Maria Gomes-Solecki3 and Wei-Gang Qiu1,2,5,6

© The Author(s), under exclusive licence to International Society for Microbial Ecology 2021

INTRODUCTION

Antigen diversification driven by host–pathogen co-evolution

Negative-frequency-dependent selection (NFDS) is an evolutionary mechanism that favours rare phenotypes over common ones, promoting biological novelty [1–3]. Driven by NFDS, antigenic variation is a molecular strategy widely shared among viral, bacterial, and eukaryotic pathogens to evade host immune defense [4–6]. Consequently, the power of NFDS in driving pathogen diversity becomes a fundamental challenge for developing broadly effective diagnostics and vaccines against fast-evolving microbial pathogens [7–9]. Although bacterial pathogens do not evolve as rapidly as viral pathogens, development of broadly effective diagnostics and vaccines is nonetheless hampered by a large number of cell-surface antigens as well as by the vast allelic diversity segregating at individual antigen loci in natural bacterial populations [5].

Here we hypothesize that, besides the structural and functional constraints to the relentless and seemingly random sequence diversification of microbial surface antigens, evolutionary rules govern antigen diversification as well. Specifically, we propose and test the hypothesis of maximum antigen diversification (MAD) that co-existing antigen variants in a microbial population are obligatorily distinct from one other in antigenicity. The MAD hypothesis is a corollary of the strain theory of pathogen–host co-evolution, which posits that host immunity drives pathogen populations into distinct genotypes (“strains”) separated from one another by large genetic distances [3, 6]. Under the strain model, co-existing pathogen strains occupy high-fitness peaks on an antigenic landscape shaped by host immunity where any off-peak antigen variants (e.g., recombinants) are at a selective disadvantage and would be eliminated by the host immunity [3]. This evolutionary rule may be exploited to tip the balance of the host–pathogen co-evolution for the benefit of the host. For example, the precarious coexistence of pathogen strains could be destabilized and the pathogen population be eliminated if the host immunological landscape is remolded by, e.g., an introduction of novel antigen variants as vaccines.

Antigenic variations in Lyme disease pathogens

For over three decades, Lyme disease has been the most prevalent vector-borne disease in the United States and Europe [10]. It is...
caused by spirochetes of the *Borrelia burgdorferi sensu lato* (*Bbsl* hereafter) species complex, also known as a new bacterial genus *Borrelliella* [11, 12]. A single species, *B. burgdorferi sensu stricto* (*B. burgdorferi* hereafter), transmitted by *Ixodes scapularis* ticks in the Northeast and Midwest and *I. pacificus* in the West, causes the majority of Lyme disease cases in the US. Genes encoding cell-surface lipoproteins are overrepresented in the ~1.5 Mbp genome of *B. burgdorferi*, totaling 4.9% of the chromosomal genes and 14.5% of the plasmid-encoded genes, in contrast to ~2.0% lipoprotein-encoding genes in other bacterial pathogens such as *Helicobacter pylori* and *Treponema pallidum* [13]. Genome comparisons further revealed that lipoprotein-encoding genes are the most variable loci within the genome, consistent with their roles in evading vector and host immunity [14]. Specifically, *B. burgdorferi* modulates cell-surface lipoprotein composition when migrating between the tick vector and the mammalian host. For example, the expression of the outer surface protein A (*OspA*) is downregulated within a mammalian host while the expression of the outer surface protein C (*OspC*) is upregulated during host invasion and, subsequently, the spirochete cells generate and express variant antigens at the *vls* (variable membrane protein-like sequences) locus to enable persistent infection of the host [15–17].

Driven by diverse forms of natural selection and with distinct cellular functions, *Bbsl* surface antigens differ in the rate and pattern of sequence evolution [14, 18–20]. For example, DNA sequences encoding *OspA* vary little among strains of the same *Bbsl* species while differing greatly among the *Bbsl* species [21]. Genes encoding *OspC* display high sequence variability within as well as between the *Bbsl* species as a result of diversifying selection [22–25]. The silent cassettes at the *vls* locus vary greatly between-species, within-species, as well as during the course of infection as a result of host adaptive immunity [17, 26–30].

Among the large repertoire of genes encoding cell-surface lipoproteins in *Bbsl*, *ospC* plays an outsized role in evading host immunity and establishing infection. First, *ospC* is required for the initial invasion into the host, suggesting its role in defense against host innate rather than adaptive immunity [31, 32]. Host cellular and molecular targets of *OspC* remain to be identified, although it has been shown that *ospC* expression was associated with the anti-phagocytosis and plasminogen-binding activities of the spirochete [33, 34].

Second, *OspC* is the immunodominant and serotype-determinant antigen of *Bbsl* strains [24, 35]. Experimental immunization of mice with recombinant *OspC* variants elicited strain-specific protective immunity against strains expressing homologous but not heterologous *OspC* variants [36–39]. Further, experimental immunization of mice using whole sera from infected mice showed that polyclonal antibodies binding *OspC* were the main components of strain-specific immunity [39]. Field-based studies further supported NFDS acting on the *ospC* locus being the main evolutionary mechanism maintaining genomic diversity in natural *B. burgdorferi* populations [22, 40–42].

Third, sequence variations at *ospC* are in nearly complete linkage disequilibrium with genomic lineages in North America, suggesting that the within-population *B. burgdorferi* strain diversification is driven by *ospC* variability [40, 43]. Simulations based on principles of population genetics showed that the nearly one-to-one correspondence between the major *ospC* alleles and the co-existing *B. burgdorferi* lineages was consistent with a history of within-population genome diversification driven by NFDS targeting the *ospC* locus [40]. Additional evidence supporting the *ospC*-driven diversification of *B. burgdorferi* strains includes the high recombination rate at *ospC* and the uniform distributions of *ospC* alleles [23, 42]. While it remains a possibility that sequence variation at *ospC* is associated with host diversity in this generalist parasite [25], the “multiple-niche” hypothesis appears to be inconsistent with results of field studies of *B. burgdorferi* populations in North America and *B. afzelii* populations in Europe as well as with results of experimental investigations [44, 45].

**Quest for broadly cross-reactive OspC molecules**

Immuno-dominance of *OspC* makes it a valuable target for anti-Lyme diagnostics and vaccines, yet its clinical potentials are limited by its sequence hyper-variability. Thus far, strategies to overcome *OspC* diversity included identification of conserved epitopes or variable epitopes distinct among natural variants [46–48]. However, conserved sequences and domains on the *OspC* molecule were ineffective targets of vaccination [46, 49]. In an high-throughput investigation, key *OspC* epitopes were mapped to the hypervariable C-terminal region with the use of protein arrays and sera from mice and humans [47]. A minimum set of *OspC* variants had been identified as candidates of broadly effective diagnostics on the basis of quantifying the antigenic breadth of *OspC* variants with the use of sera from immunized mice as well as sera from naturally infected hosts [48]. A concatenation of eight *OspC* epitopes associated with distinct natural variants were the base of a broadly immunogenic vaccine for canine use [46, 50].

The MAD hypothesis suggests an alternative and novel strategy to overcome the limitation of *OspC* sequence diversity to the development of *OspC*-based diagnostics and vaccines. First, on the basis of frequency distributions of antigen variants in nature and experimental immunization of mice, here we tested MAD among the 16 *OspC* variants co-existing in natural populations of the Lyme disease pathogens in the Northeast United States [40, 51]. Second, we used evolutionary algorithms to design analogs of natural *OspC* molecules with minimized sequence differences to the 16 natural variants. We cloned and purified these synthetic *OspC* molecules and tested their antigenic breadths using sera from artificially and naturally infected hosts. Third, we explored molecular mechanisms underlying the broad antigenicity of evolutionary antigens with a mathematical model and computer simulations. One of our evolution-based designs—the consensus algorithm—was similar to the COBRA approach used to design broadly reactive vaccines against the influenza virus [8]. The root algorithm—another evolution-informed algorithm implemented in the present study—has been used to design vaccine candidates against diverse strains of the human immunodeficiency virus type 1 (HIV-1) [52]. Critically, these evolution-based strategies are automated and able to generate synthetics analogs that assume the structure, function, and epitope configurations similar to those of the native antigens while displaying broader antigenicity.

**MATERIALS AND METHODS**

**Co-occurrences of OspC variants in field-collected *Ixodes* ticks and a test of *OspC* specificity**

We tested the immunological distinction of *B. burgdorferi* strains based on their co-occurrences in individual *Ixodes scapularis* ticks. Deep high-throughput sequencing of individual nymphal ticks, which had fed a single blood meal as larvae, revealed that multiple *B. burgdorferi* strains within a single tick were caused mainly by the mixed infection of the host rather than by a history of feeding on multiple hosts [23, 53]. As such, we hypothesized that immunologically distinct strains tended to co-infect a single tick while immunologically similar strains tended to be found in different ticks. We tested the hypothesis with a previously published data set that recorded the presence and absence of 20 Lyme pathogen strains within *n = 119 I. scapularis* ticks collected from New York State during 2015 and 2016 [54] (Supplementary Information Data S1). While the data set consisted of mostly adult ticks with only 25 nymphal ticks, there were no significant differences in the level of *B. burgdorferi* strain diversity carried by single ticks among the nymphal, adult male, and adult female ticks [23]. In other words, the individual infected nymphal ticks carried *B. burgdorferi* strains as diverse as the adult ticks despite an additional blood meal the adult ticks have taken. In a separate study using high-throughput sequencing of infected nymphal ticks from the same region, the authors...
Evolutionary algorithms for designing broadly reactive OspC antigens. The centroid algorithm used a genetic algorithm to search for antigen sequences with minimized differences to the natural OspC variants ("centroids"). The root algorithm inferred an antigen sequence representing the mid-point root of a phylogenetic tree of the natural OspC variants ("root"). The consensus algorithm generated an antigen sequence consisting of majority amino-acid residues at individual positions of an alignment of the natural OspC sequences ("consense"). These evolutionary algorithms produced OspC analogs with approximately half the sequence differences to the natural variants than the distances among the natural variants themselves.

Gene synthesis, protein overexpression, and protein purification
DNA sequences encoding the natural and synthetic OspC variants were codon-optimized, synthesized, and cloned into the pET24 plasmid vector, which was then used to transform the Escherichia coli BL21 cells. All DNA work was performed by a commercial service (GenelmMune Biotechnology Corp., Rockville, MD, USA). We designed the OspC constructs by excluding the first 18 residues encompassing the signal peptide and by adding a 10 × Histidine-tag on the N-terminus. These modifications were necessary for overexpression of OspC proteins in E.coli and to facilitate OspC purification.

For each E. coli strain containing a cloned ospC gene, a single colony was selected to inoculate 4 ml Luria-Bertani (LB) broth (Thermo Fisher Scientific, Waltham, MA, USA) containing vector-specific selective antibiotics (25 μg/ml for Kanamycin or 50 μg/ml for Ampicillin). The seeded culture was incubated overnight at 37 ℃ with vigorous shaking (250 rpm). A portion of the overnight culture was transferred into 50 ml fresh pre-warmed LB broth containing 0.4% glucose and the selective antibiotics. The culture was incubated at 37 ℃ with vigorous shaking until reaching exponential growth indicated by an OΔD600 of ~0.8 as measured by the NanoDrop Spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA). Expression of the cloned ospC was induced by adding isopropyl β-d-1-thiogalactopyranoside (IPTG) to a final concentration of 0.25–0.5 mM and by incubation overnight at 25 ℃. Cells were collected by refrigerated centrifugation at 4 ℃ and 7200 rpm for 15 min, resuspended in a lysis buffer containing 0.2 mg/ml lysozyme, 20 mM Tris-HCl (pH 8.0), 250 mM NaCl, and 1 mM dithiothreitol (DTT). After incubation for 1 h at 4 ℃, cells were further lysed by sonication until the solution become translucent. The lysate was centrifuged in refrigeration at 12,000 rpm for 20 min and the supernatant was withdrawn.

have similarly concluded that reservoir hosts were commonly infected by diverse B. burgdorferi strains [53]. Here, we quantified the over- or under-abundance of a pair of strains (i and j) as the fold change of the observed over the expected counts: $F(C) = \log_{10}(\frac{A_{ij}}{A_{ij}^{*}})$. Statistical significance of the relative abundance was obtained from a null distribution generated by permuting the occurrences of a pair of strains among infected ticks 1000 times while keeping the total occurrence constant. Significantly over- or underrepresented pairs of B. burgdorferi strains co-infecting single I. scapularis ticks would suggest an absence or presence, respectively, of immune cross-protection for the strain pairs.

Approximately 20 OspC variants commonly coexist in B. burgdorferi populations in the Northeast US [22, 23, 48]. We aligned the protein sequences of 16 common natural OspC variants (named as “A” through “N”, “T”, and “U) with the program muscle [55] (Supplementary Information Text S1). Pairwise sequence differences among the OspC variants were calculated with the alignment utility bioalign from the BpWrapper software suite, which was based on BioPerl [56, 57].

Evolutionary algorithms for designing broadly reactive synthetic OspC
Protein sequences analogous to natural OspC variants were optimized for broad reactivity using three evolutionary algorithms (Fig. 1). By generating OspC analogs close to the root of a molecular phylogeny with natural OspC variants, these evolutionary algorithms aimed to reduce the sequence distance between an evolutionary analog with the natural variants to be approximately half of the difference among the natural variants themselves. The initial input for all three algorithms was the aligned amino-acid sequences of 16 OspC variants (Supplementary Information Text S1). First, we inferred the hypothetical ancestral sequence at the midpoint root of the phylogeny of the natural OspC variants with RAxML [58] (the "Root Algorithm"). Second, we obtained a consensus sequence consisting of 20% majority residues at aligned sequence positions of the natural OspC variants with the consensus method implemented in the Bio: SimpleAlign module of the BioPerl library [57] (the "Consensus Algorithm"). Third, we used a genetic algorithm to generate sequences with minimal distances to natural OspC variants (the "Centroid Algorithm"). Briefly, we extracted amino acids at variable positions from 16 aligned natural OspC sequences. An initial seed population of random antigen sequences (N = 10) were generated through sampling the unique amino acids present at each variable site with uniform probabilities. For each randomly generated sequence i, we calculated its differences ($d_{ij}$) through "U") to the 16 natural variants. We defined the fitness of this sequence as the maximum value among its differences to all 16 natural sequences: $w = \max(d_{ij})$. This fitness measured its overall sequence similarity to the natural variants—the lower the $w$, the higher its overall similarity to the natural variants. The top ten most similar antigen sequences ("elites") in each generation were retained and others were discarded. Each elite sequence was then allowed to "reproduce" 10 times with mutations at randomly selected ten variable sites, resulting in ten mutated "gametes". The above process was repeated (e.g., for 5000 generations) to progressively lower the $w$ values, after which the final output included ten elite centroids that were the most similar to all 16 natural OspC variants. The centroid algorithm was implemented with the BioPerl library in Perl [57] and the DEAP package in Python [59]. The top four optimized centroid sequences were cloned, overexpressed, and purified for immunological assays of antigenic breadth.
The recombinant proteins were purified using nickel sepharose beads (Ni-NTA, Thermo Fisher Scientific, Waltham, MA, USA). The lysate supernatant from the 50 ml culture was mixed with 300 µl Ni-NTA beads and incubated overnight at 4 °C in the lysis buffer supplemented with 5 mM imidazole. The lysate-bead mixture was then loaded into a chromatography column and washed with 12 times the bed volume of the lysis buffer containing 100 mM imidazole. The elution was diazoyed to remove imidazole in phosphate-buffered saline (PBS, pH 7.4) containing 1 mM DTT and 20% glycerol.

The amount and purity of recombinant proteins were examined using the sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) containing 15% gel. The purified protein was identified and quantified by Coomassie Blue and stained in 45% methanol and 10% acetic acid. Concentration of the final purified protein solution was quantified using the Pierce Bradford Protein Assay Kit (Thermo Fisher Scientific, Waltham, MA, USA).

Sera from naturally infected hosts and immunized mice

The majority of human serum samples were provided by the Centers for Disease Control and Prevention (CDC) (Table 1). The human sera originated from patients diagnosed with early to late Lyme disease or from healthy individuals in endemic and non-endemic regions in the USA [61]. The CDC sera panel was previously screened using the standard two-tiered testing (STTT) for the presence of antibodies against B. burgdorferi, including IgM, IgG, or both antibodies against OspC (the 23 kD band) [62]. The CDC sera panel was custom compiled for the present study. Ten serum samples from Lyme disease patients were originally collected by the Stony Brook University Health Science Center, NY, USA. Ten serum samples were obtained from the natural reservoir of B. burgdorferi, the white-footed mouse (Peromyscus leucopus) from Milbrook, NY, USA. The latter human and mouse sera were screened for exposure to B. burgdorferi using the C6 ELISA (Immunetics, Boston, MA, USA).

Sixteen recombinant OspC were previously cloned from B. burgdorferi strains into the pETPlc plasmid [48]. The proteins were expressed in E. coli BL21 (DE3) pLysS and purified under native conditions by ion exchange chromatography using Q-Sepharose Fast Flow (GE Healthcare, Sweden) as described previously [63]. C3H/HeJ mice (Mus musculus) and white-footed mice (P. leucopus) were immunized with 10–20 µg of each of the 16 individual purified natural recombinant OspC proteins. Briefly, mice received a dose of recombinant protein on Day 1 and Day 4, and on Day 28 they were euthanized and blood collected by heart puncture. Animal experimentation followed the protocols approved by the Animal Care and Use Committee of University of Tennessee Health Science Center.

Immunological assays

MiniSlot assays of OspC variant-specific sera were performed using a MiniSlot/MiniBlotter 45 system (Immunetics, Boston, MA, USA). Briefly, a pyrography membrane with 225 mm Immunoblot membrane was placed into the MiniSlot and 25 µg of each purified protein was loaded individually into its parallel channels. The proteins were immobilized onto the PVDF membrane after the excess solution was removed by vacuum aspiration, resulting in a deposit of horizontal parallel stripes of antigens. The membrane was released from the MiniSlot and blocked in 10% skim milk (Difco, Sparks, MD, USA) for 2 h at room temperature. After blocking the membrane was rotated by 90 degrees and placed in the MiniBlotter 45. Diluted mouse serum (1:100 to 1:1000 in 3% of milk in TBS buffer with 0.05% Tween 20, 150 µl) was deposited in the individual vertical lanes of the MiniBlotter and was incubated for 1 h at room temperature. The membrane was washed three times with TBS containing 0.5% Tween 20 and was incubated with goat anti-mouse IgG conjugated with alkaline phosphatase (1:2000) (Kinkelaar & Perry Laboratories [KPL], Gaithersburg, MD, USA) for 1 h at room temperature. The BCIP/NBT Phosphatase Substrate (KPL) was used to visualize the signal. Serum of non-immunized mice and the bovine serum albumin were used as the negative serum control and the non-OspC antigen control, respectively. Binding intensity values were digitalized with Imagej [64].

Statistical analysis of OspC cross-reactivity

We tested antigenic specificity of natural OspC variants by performing a re-analysis of a data set generated from a previously published study [48]. The data set consisted of replicated ELISA readings of the reactivity between the 16 recombinant OspC variants with the antisera (n = 15) of C3H/HeJ mice (Mus musculus) artificially immunized with the purified OspC variants (Supplementary Information Data S2). Sera from uninfected mice were used as the negative control.

To quantify the antigenicity of OspC variants with variant-specific mouse sera, we first transformed the raw OD450 readings and the digitalized antigenic binding intensities into normalized z-scores: 

\[ z = \frac{x - \text{mean}(x)}{\text{SD}(x)} \]

where \( x \) and \( y \) are the binding value (OD450 reading or binding intensity) of the recombinant OspC variant with the variant-specific sera, and \( \text{mean}(x) \) and \( \text{SD}(x) \), respectively, the mean and the standard deviation of OD450 readings of all recombinant OspC variants with respect to the serum. The rescaling was necessary to account for the systematic differences among the variant-specific sera in non-specific, background bindings due to, e.g., the varying amount of total antibodies in a particular mouse serum. For example, the OD450 readings of variants—regardless homologous or heterologous antigens—with the variant K-specific serum were consistently higher than other sera (OD450 > 1.0) in the original study [48] (Fig. 1 therein). Normalization with the z-scores made it possible to compare the reactivity of an OspC variant across the serum samples by reducing serum-specific background noise. Furthermore, a score of \( z < -2.0 \) or \( z > 2.0 \) indicated a statistically significant (with \( p = 0.05 \)) deviation from the mean reactivity (\( z = 0 \)) of an OspC variant with the variant-specific sera. Normalization of ELISA readings with z-scores had been used to reduce serum-specific background noises in a clinical diagnostic test [65].

To quantify the antigenicity of OspC variants with the naturally infected human and mouse sera, we first displayed the raw OD450 readings with bar graphs. To render the binding values comparable among the serum samples, we then transformed the OD450 values with respect to sera into z-scores as described above. The median value for a given antigen variant was a measure of the antigenic breadth of the variant-specific sera. Normalization of binding intensities were visualized with heatmaps, which further identified hierarchical clusters of OspC variants and sera according to the similarities in binding levels. The R package heatmap was used to generate the heatmaps. To summarize the antigenic breadth of an OspC variant across the serum samples, we designed a novel measure of total antigen reactivity. The antigenic reactivity characteristic (ARC) of an OspC variant was defined as a curve of cumulative binding values (with scaled OD450 readings) over the cumulative number of serum samples. A highly specific antigen variant would show as a low-lying ARC curve because of the consistently low (\( z < 0 \)) bindings with sera. As such, the ARC of a variant with the serum would show as a low-lying ARC curve because of the consistently low (\( z > 0 \)) bindings with sera. In contrast, a variant with low serum reactivity but generating all negative scores would show as an elevated ARC curve. To quantify the antigenicity of OspC variants with variant-specific sera, a score of \( z < -2.0 \) or \( z > 2.0 \) indicated a statistically significant (with \( p = 0.05 \)) deviation from the mean reactivity (\( z = 0 \)) of an OspC variant with the variant-specific sera. Normalization of ELISA readings with z-scores had been used to reduce serum-specific background noises in a clinical diagnostic test [65].

The ISME Journal (2022) 16:447 – 464

SPRINGER NATURE

L. Di et al.
Table 1. Serum samples.

| Label | Host   | Source | Description                  | STTT Interpretation<sup>b</sup> | C6 ELISA<sup>c</sup> |
|-------|--------|--------|------------------------------|----------------------------------|----------------------|
|       |        | EIA    | IgM 23 kDa band              | IgG 23 kDa band                   |                      |
| S01   | Human  | CDC    | EM<sup>*</sup> convalescence | +                                | +                    | NA<sup>d</sup> |
| S03   | Human  | CDC    | EM convalescent              | +                                | +                    | NA         |
| S04   | Human  | CDC    | Non-endemic control          | −                                | −                    | NA         |
| S05   | Human  | CDC    | Non-endemic control          | −                                | −                    | NA         |
| S10   | Human  | CDC    | Neurological Lyme            | +                                | +                    | NA         |
| S11   | Human  | CDC    | EM convalescent              | +                                | +                    | NA         |
| S14   | Human  | CDC    | Fibromyalgia (control)       | −                                | −                    | NA         |
| S16   | Human  | CDC    | Severe periodontitis (control) | −                                | −                    | NA         |
| S18   | Human  | CDC    | EM convalescent              | +                                | +                    | NA         |
| S21   | Human  | CDC    | Endemic control              | −                                | +                    | NA         |
| S22   | Human  | CDC    | Neurological Lyme            | +                                | +                    | NA         |
| S30   | Human  | CDC    | EM acute                     | −                                | +                    | NA         |
| T01   | Human  | CDC    | EM                           | +                                | +                    | NA         |
| T03   | Human  | CDC    | Lyme arthritis               | +                                | +                    | NA         |
| T04   | Human  | CDC    | EM                           | Equ<sup>o</sup>                  | −                    | NA         |
| T05   | Human  | CDC    | Lyme arthritis               | +                                | +                    | NA         |
| T06   | Human  | CDC    | EM                           | +                                | +                    | NA         |
| T07   | Human  | CDC    | EM                           | Equ                              | +                    | NA         |
| T08   | Human  | CDC    | EM                           | +                                | +                    | NA         |
| T09   | Human  | CDC    | EM                           | −                                | −                    | NA         |
| T10   | Human  | CDC    | EM                           | −                                | −                    | NA         |
| T11   | Human  | CDC    | EM                           | +                                | +                    | NA         |
| T12   | Human  | CDC    | Lyme arthritis               | +                                | −                    | NA         |
| T13   | Human  | CDC    | Neurological Lyme            | +                                | +                    | NA         |
| T14   | Human  | CDC    | EM                           | −                                | +                    | NA         |
| T15   | Human  | CDC    | Lyme arthritis               | +                                | +                    | NA         |
| T16   | Human  | CDC    | EM                           | +                                | +                    | NA         |
| T17   | Human  | CDC    | EM                           | +                                | +                    | NA         |
| T18   | Human  | CDC    | Neurological Lyme            | +                                | +                    | NA         |
| T19   | Human  | CDC    | Cardiac Lyme                 | +                                | +                    | NA         |
| T20   | Human  | CDC    | EM                           | +                                | +                    | NA         |
| T21   | Human  | CDC    | Neurological Lyme            | −                                | +                    | NA         |
| T22   | Human  | CDC    | EM                           | +                                | +                    | NA         |
| T23   | Human  | CDC    | EM                           | −                                | −                    | NA         |
| T24   | Human  | CDC    | EM                           | +                                | +                    | NA         |
| T25   | Human  | CDC    | EM                           | +                                | +                    | NA         |
| T26   | Human  | CDC    | EM                           | −                                | +                    | NA         |
| T27   | Human  | CDC    | EM                           | +                                | −                    | NA         |
| T29   | Human  | CDC    | EM                           | +                                | +                    | NA         |
| T30   | Human  | CDC    | Neurological Lyme            | +                                | +                    | NA         |
| T31   | Human  | CDC    | EM                           | −                                | −                    | NA         |
| T32   | Human  | CDC    | Cardiac Lyme                 | +                                | +                    | NA         |
| H01   | Human  | Stony Brook | Late Lyme  | NA | NA | 0.924 |
| H04   | Human  | Stony Brook | Late Lyme  | NA | NA | 1.947 |
| H07   | Human  | Stony Brook | Late Lyme  | NA | NA | 0.555 |
| H08   | Human  | Stony Brook | Late Lyme  | NA | NA | 0.260 |
| H09   | Human  | Stony Brook | Late Lyme  | NA | NA | 0.013 |
| H10   | Human  | Stony Brook | Late Lyme  | NA | NA | 0.491 |
| P01   | P. leucopus | Millbrook, NY | Field reservoir | NA | NA | 0.392 |
generated with random permutations of the scaled OD_{450} readings among the same set of serum samples.

A binary model and computer simulations of maximum antigen separation

To explore immunological and molecular mechanisms underlying the broad antigenicity of evolutionary centroids, we construct a mathematical model and computationally simulated maximum antigen diversification (MAD) and evolutionary centroids. Following the multiple-epitope extension of the Gupta et al. model [3], we represented antigen variants in a pathogen population as binary strings.

In generating binary strings, let \( p \) be the probability of the binary state 1 (and \( 1 - p \) the probability of the binary state 0). Without loss of generality, we may assume that \( p \leq 0.5 \) since, otherwise, switching the roles of 0s and 1s preserves all distances (we use the Hamming distance) and changes \( p \) to \( 1 - p \). Following the convention of using zeroes to represent ancestral states in evolutionary analysis, we designate the zero string (the string with all bits set to 0) as a "centroid". Mathematically, this choice of the centroid is general because flipping all bits in a given position preserves all distances. Two random binary strings of length \( n \) are expected to differ in \( 2p(1-p) \) bits. Similarly, a random binary string of length \( n \) is expected to differ in \( p \) bits from the centroid. While this behavior is only in expectation, when \( n \) is large we can achieve it with high probability. Define distance as the Hamming distance normalized by the length \( n \). Given a set of randomly generated binary strings, let \( D \) be the minimum distance among all strings, and \( d \) the maximum distance of any string to the centroid. If the set of strings is reasonably small and the length of strings is long, standard probability theory tells us that, with high probability, \( D \) will be close to \( 2p(1-p) \) and \( d \) will be close to \( p \), resulting in \( D = 2D - 2d \). We conjecture that this curve imposes a theoretical bound when given a large enough set of strings. To create a population of maximally separated (a large \( D \)) strings with minimal distances (a small \( d \)) to the centroid, one would seek to maximize the ratio \( D/d \) for a given value of \( p \). Here, \( d \), which approaches the frequency of derived states \( p \), is a measure of sequence divergence from the centroid. Thus, the maximized \( D/d \) represents the maximal possible sequence divergence among the evolved strings at a given level of evolutionary divergence. Without constraining on \( d \) (as in a separate algorithm shown below), the strings would diverge without regard to any biologically realistic constraints including the time since the evolutionary origin and functional and structural conservation. We used a genetic algorithm to validate the theoretical boundary with \( n = 100 \) bits and \( N = 10 \) strings. Distances optimized with the genetic algorithm (GA package in R [67]) were compared with empirical results based on the natural OspC variants as well as with results of randomly generated binary strings without \( D/d \) maximization.

In a separate experiment, we used a genetic algorithm to search for a sample of maximally separated antigen variants to represent a MAD population, without consideration of a centroid. The searching was performed using the GA package in R [67] to maximize \( D \) (as defined above) \( \text{fit} = \min(d_{ij}) \). We then searched for centroid variants, using a separate genetic algorithm to minimize \( d = \max(d_{ij} - 1.10) \). Evolutionary analogs were validated with a neighboring tree based on pairwise Hamming distances. An R markdown of the simulation protocol is included as Supplementary Information Text S2.

We note the similarity between the simulated maximally diversified antigens and well-separated binary codewords under the Hamming distance. We further note that the problem of finding centroids given a set of strings is known as the Closest String or the Hamming Centroid problem in computer science. While coding theory provides various techniques for generating well-separated codewords, and many algorithms for finding centroids exist [68, 69], we approached both problems in one stochastic framework based on genetic algorithms as described above.

RESULTS

Lack of immune cross-protection among *B. burgdorferi* strains in nature

In a previous study, *B. burgdorferi* strain diversity was quantified with high-throughput sequencing of the ospC locus at the level of single *I. scapularis* ticks [54]. Consistent with earlier results based on DNA cloning and DNA-DNA hybridization, the results reaffirmed a largely uniform distribution of a diverse set of *B. burgdorferi* strains identifiable by ~16 major-group ospC alleles in the highly endemic regions of Lyme disease in the Northeast US [22, 42]. Using the same data set (Supplementary Information Data S1), we tested if frequencies of pairs of *B. burgdorferi* strains co-infecting a single tick were higher, lower, or equivalent relative to the expectation of random allelic association. With a sample of \( n = 119 \) infected *I. scapularis* ticks, we found that the majority of strain pairs were overrepresented relative to random expectations and no pair was significantly underrepresented in infected ticks (Fig. 2). In particular, strain pairs containing the OspC variants F and J were the most overrepresented. To account for the possible contribution of the multiple blood meals of the adult ticks to the observed overrepresentation of mixed infections in single ticks, we repeated the permutation test using only the 25 nymphal ticks. All strain pairs were
overrepresented, eight of which significantly so (results not shown). Since the *B. burgdorferi* strain diversity in individual ticks are driven mainly by the pathogen diversity in the host, the predominance of overrepresentation of mixed strain pairs supported immunological distinctness of OspC variants and a lack of immunological cross-protection against superinfection of the reservoir hosts [22, 23, 53].

**ELISA testing of antigenic specificity of OspC variants with variant-specific sera**

In a previous study performed in one of our labs, C3H mice were immunized with the 16 recombinant OspC natural variants [47]. These variant-specific sera were used to test the cross-reactivity of the OspC variants by ELISA. The authors identified five OspC variants (B, E, F, I, and K) that were most broadly reactive with the variant-specific sera, consistent with results using naturally infected sera from human patients, dogs, and *P. leucopus* mice [47]. Here we re-analyzed the ELISA data set (Supplementary Information Data S2) by correcting for serum-to-serum variation (Supplementary Information Data S2) by correcting for serum-to-serum variation [47]. Here we re-analyzed the ELISA data set (Supplementary Information Data S2) by correcting for serum-to-serum variation. The OD_{450} readings showed stronger reactivity of homologous than heterologous bindings but large serum-specific variability (Fig. 3A). Some variant-specific sera (e.g., sK, for K-specific serum) showed consistently higher readings than others, reflecting experiment-specific factors such as a strong immune response of an animal. Without normalization to remove such experiment-specific background noise, the antigenicity of purified recombinant OspC displayed large variability and thus lacked statistical power for comparison (Fig. 3B). Indeed, the raw readings of homologous bindings were generally lower than the readings of heterologous bindings for the OspC variants. Normalization with respect to sera removed serum-specific background without altering the reactivity rankings or variance for each variant-specific serum (Fig. 3C). Critically, normalization restored the expected stronger reactivity of homologous than heterologous bindings while greatly reducing the reactivity variance for individual OspC variants (Fig. 3D). Thus, antigenicity of OspC variants could be compared with greater statistical confidence, such as the top cross-reactivity of rE (for recombinant E variant), rK, rI, rF, and rB variants, in increasing order of cross-reactivity with a median value of z > 0. Normalization did not nullify but increased statistical confidence of the conclusion of the original study, which identified the same set of variants among the top cross-reactive variants [48].

The serum-normalized ELISA readings showed that, with two exceptions, rOspCs reacted significantly (i.e., with z ~ 2.0, two standard deviations above the mean) with homologous sera, indicating high antigenic specificity of rOspCs (Fig. 4, bar plot). The two exceptions included the variant F, which reacted significantly with both the F- and the B-specific sera, and the variant J, which reacted more strongly with the M-specific serum than with the J-specific serum. The high antigenic specificity of rOspCs is alternatively visualized with a heat map, which shows a strong diagonal line indicating the highest reactivity of rOspCs with homologous sera (Fig. 4, heat map). Note the absence of reactivity with the L-specific serum in both the bar graph and the heat map. This is because the strain expressing the L variant was not available for generating the L-specific serum at the time [48]. Note also that although heterologous bindings were generally weaker than homologous bindings, rOspCs nonetheless reacted with heterologous sera. Notice that a binding value of z = 0 represented the average reactivity, not an absence of antigen- serum binding.

An antigen reaction characteristic (ARC) curve was a way to summarize the overall reactivity of a rOspC with all serum samples (Fig. 4, ARC curves). In addition, the ARC curves provided a quantitative measure of antigen specificity and cross-reactivity, showing the most broadly reactive antigens at the top and the most specific antigens at the bottom. For example, the ARC curves show B, F, K, E, and I being the most broadly reactive variants, in agreement with the conclusion of the original study [47].

**Immunoblot assays with variant-specific sera**

We further tested the antigenic specificity of rOspCs with the use of immunoblot assays and a full set of 16 variant-specific sera (the L-specific sera included) from immunized C3H and *P. leucopus* mice (Fig. 5, top). The raw immunoblot images showed strong specific reactions of rOspC with homologous sera (diagonal) and weak reactions with heterologous sera (off-diagonal). As in the ELISA analysis, we corrected for serum-to-serum variation by normalizing binding intensities with respect to sera (Supplementary Information Data S3). Consistent with ELISA results, the re-scaled intensities
variants with a median normalization, the OspC variants showed significant among the sera (F = 51.46, p < 2.2e−16, by an ANOVA). B OD₄₅₀ readings with respect to the 15 recombinant OspC variants (x-axis), ordered by the medians. Without correcting for the serum-by-serum variability, the OspC variants did not vary significantly in reactivity with the variant-specific sera (F = 1.04, p = 0.42). C Normalized reactivity (z-score, y-axis) with respect to the variant-specific sera (x-axis). Serum-specific variability was removed (F = 0, p = 1). D Normalized reactivity (z-score, y-axis) with respect to the OspC variants (x-axis). After normalization, the OspC variants showed significant variability in reactivity with the variant-specific sera (F = 6.17, p = 8.3e−11). Five OspC variants with a median z > 0, indicating above-average reactivity, were highlighted with shaded boxes. The same five OspC variants ranked as the most reactive without normalization. Thus, normalization did not change the ranking but greatly improved statistical confidence and precision for comparing the antigenicity among the antigen variants.

To conclude, testing on the basis of ELISA and immunoblots showed the strongest bindings between rOspCs with homologous sera (Fig. 5, heat map). However, the ARC curves showed a lack of consistency in the topmost cross-reactive rOspC variants between the immunoblot using the C3H mice (rH and rI at the top) and the immunoblot using the P. leucopus mice (rT and rJ at the top) (Fig. 5, ARC curves). Further, the rOspC rankings of the immunological breadth as quantified by the ARC curves in both immunoblots were different from the ranking from the ELISA experiment using the C3H mice (rB and rF at the top, Fig. 4, ARC curves).

To conclude, testing on the basis of ELISA and immunoblots and with the use of OspC variant-specific sera from two mouse species all showed the strongest reactions of OspC variants with homologous sera. Reactions of OspC variants with heterologous sera, however, were weaker and inconsistent between experiments and between the two mouse species.

Centroids reacted broadly with naturally infected human and mouse sera

We designed six evolutionary analogs (Supplementary Information Text S1) expected to show broad antigenic cross-reactivity with the 16 natural OspC variants with the use of three evolution-based algorithms (Fig. 1). The root analog (“Root”) is defined as the maximum-likelihood sequence of the hypervariable phylogenetic root of the 16 natural OspC variants.

The consensus sequence (“Consense”) consisted of majority amino-acid residues at individual alignment positions. The centroid analogs (“Centroid”) were computationally derived sequences that were minimized for sequence differences with the 16 natural variants. Whereas the root and consensus algorithms each generated a single OspC analog, the centroid algorithm generated ten optimized sequences from each run. To increase the diversity of candidate sequences, the algorithm was run with repetition and the most optimized sequence was chosen form each run. We chose four centroids among a dozen candidates for further experimentation on the basis of their distinct phylogenetic positions.

Effectively, the algorithms drastically cut the sequence differences of the evolutionary analogs with the natural OspC variants (e.g., d = 0.182 ± 0.024 for the consense) to approximately half of the sequence differences among the natural variants themselves (d = 0.260 ± 0.033). Based on immunological models suggesting a tight correlation between sequence and antigenic distances [70], we expected a similar level of reduction in the antigenic distances of each evolutionary analog to the natural OspC variants. Phylogenetic analysis of these evolutionary analogs validated their expected central positions among the OspC sequence diversity (Fig. 6C). Sequence differences of the evolutionary centroids with the 16 natural variants were more

---

**Fig. 3** Removal of serum-specific background noise with normalization. A Raw OD₄₅₀ readings (y-axis) from a previously published ELISA of the binding between 16 purified recombinant OspC antigens with 15 OspC variant-specific mouse sera (x-axis) [48]. Homologous bindings (solid dots) were between an antigen variant and a serum from a C3H mouse immunized with the same recombinant variant. Heterologous bindings (open dots) were between an antigen variant and a serum from a mouse immunized with a different variant. ELISA readings varied significantly among the sera (F = 51.46, p < 2.2e−16, by an ANOVA). B OD₄₅₀ readings with respect to the 15 recombinant OspC variants (x-axis), ordered by the medians. Without correcting for the serum-by-serum variability, the OspC variants did not vary significantly in reactivity with the variant-specific sera (F = 1.04, p = 0.42). C Normalized reactivity (z-score, y-axis) with respect to the variant-specific sera (x-axis). Serum-specific variability was removed (F = 0, p = 1). D Normalized reactivity (z-score, y-axis) with respect to the OspC variants (x-axis). After normalization, the OspC variants showed significant variability in reactivity with the variant-specific sera (F = 6.17, p = 8.3e−11). Five OspC variants with a median z > 0, indicating above-average reactivity, were highlighted with shaded boxes. The same five OspC variants ranked as the most reactive without normalization. Thus, normalization did not change the ranking but greatly improved statistical confidence and precision for comparing the antigenicity among the antigen variants.
uniform while the consensus analog showed a lower average difference (Fig. 6A, B; Supplementary Information Data S4). The root analog showed the highest average sequence difference as well as the highest variability in sequence differences with the natural variants (Fig. 6B).

We cloned, overexpressed, and purified the six evolutionary analogs and the 16 natural variants as recombinant proteins (Supplementary Information Fig. S1). Antigenicity of each rOspC was quantified by its reactions with OspC-positive sera (Table 1) from naturally infected human patients (n = 41) and P. leucopus

Fig. 4 ELISA of OspC variants with variant-specific sera. Fifteen sera ("sA" through "sU") from mice, each immunized with a specific recombinant OspC variant, were previously assayed for reactivity with the 16 OspC variants ("rA" through "rU") using ELISA [48]. (Top) Each panel shows binding intensities (normalized z-scores, y-axis) of an OspC variant with a panel of OspC-specific sera (x-axis). Error bars show one standard deviation above and below the mean from two replicated assays. A value above the z = 2 line (dashes) indicates a highly significant reaction. (Bottom left) A heat map representation of the mean z-scores. (Bottom right) Antigen reaction characteristics (ARC) curves, similar to the receiver-operation characteristics (ROC) curves, is a measure of antigen specificity. Each curve traced the cumulative z-scores (y-axis) of an OspC variant's binding intensities with the sera samples, ordered by the lowest to the highest reactivity. The ARC curve rises with an above-average binding value (z > 0) and drops with a below-average binding value (z < 0). Thus, a high-rising curve (e.g., for rB) indicated consistently above-average reactivity with sera samples, suggesting a broadly cross-reactive antigen. Conversely, a low-lying curve (e.g., for rG) indicated consistently below-average reactivity, suggesting a relatively specific antigen. Curves close to the zero line (the majority of variants) indicated antigens with an average level of cross-reactivity.
Fig. 5  Immunoblot testing of OspC variants with variant-specific sera. (Top) Immunoblot images of OspC-specific sera (x-axis) from the C3H mice (left) and the P. leucopus mice (right) reacting with recombinant OspC variants (y-axis). The last column (labeled with “−”) is the negative control, showing reactions of sera from un-immunized mice. (Middle) Corresponding heatmaps. The binding intensity values on the immunoblot images were captured by ImageJ [64]. Values were then normalized by subtracting intensities from the negative controls and by scaling to z-scores. (Bottom) ARC curves. Some of the most (topmost) and the least (bottom-most) reactive recombinant OspC variants were labeled.
mice \((n = 10)\) using ELISA (Supplementary Information Data S5). Natural OspC variants (gray bars) reacted with the human serum samples with visible variability, so did the root (orange bars) and the consensus (blue bars) analogs (Supplementary Information Fig. S2). One centroid ("CT1") reacted poorly with the majority of mice sera (Supplementary Information Fig. S3). In contrast, the other three centroids ("CT2", "CT3", and "CT4") reacted consistently at high levels with all sera. The mouse sera reacted with rOspCs in a more variant-specific manner than the human sera. For example, the mouse sera P03, P04, P06, P08, and P09 reacted strongly with one to four natural rOspC variants while weakly with other natural variants (Supplementary Information Fig. S3). Although the natural rOspC variants reacted strongly with some of the murine sera, the three centroids reacted consistently high with all murine sera.

The antigenic breaths of the OspC variants were further quantified with the use of heat map (Supplementary Information Fig. S4) and the normalized z-scores (Fig. 7). In the heat map, the OD450 readings were scaled with respect to individual sera and, subsequently, both the sera (in columns) and the rOspCs (in rows) were grouped according to pairwise similarities in reactivity (Supplementary Information Fig. S4). The three centroids (CT2, CT3, and CT4) showed as a distinct cluster that reacted with the human and mouse sera at levels that were consistently above the average.

The boxplots of the serum-normalized z-scores confirmed significantly broader reactivity of all evolutionary analogs relative to the natural variants with the naturally infected \(P.\ leucopus\) sera, with \(p = 0.031\) for the root analog, \(p = 4.9e^{-03}\) for the consensus analog, \(p = 0.041\) for CT1, \(p = 5.2e^{-05}\) for CT2, \(p = 1.4e^{-06}\) for CT3, and \(p = 5.6e^{-05}\) for CT4 (Fig. 7 top left, boxplot). With the use of naturally infected human sera, reactivity of CT2 \((p = 9.1e^{-10})\), CT3 \((p = 2.4e^{-06})\), and CT4 \((p = 1.1e^{-09})\) was significantly higher than the natural variants. The reactivity of CT1 \((p = 2.3e^{-06})\) was significantly lower than the antigenicity of the natural variants while the reactivity of the root \((p = 0.31)\) and consensus \((p = 0.065)\) analogs was not significant (Fig. 7 bottom left, boxplot). The ARC curves summarized the strong reactions of the three evolutionary centroids with the sera (the top three curves) and the weak reaction of the root and consensus analogs (Fig. 7 top and bottom right, ARC curves). The ARC curves of the cross-reactivity of natural OspC variants with the human sera (gray lines) showed R8 as a top cross-reactive and R9 as the least cross-reactive variant, consistent with the rankings of these two
variants in the ARC curve based on ELISA with variant-specific sera (Fig. 4 bottom right, ARC curves). We conclude from the ELISA testing that the evolutionary centroids with computationally optimized sequence differences tended to react more broadly with naturally infected human and mouse sera than natural OspC variants.

Evolutionary analogs are structurally similar to native OspC variants
The CT1 centroid was more cross-reactive than the natural variants with the mouse sera but, unlike the CT2, CT3, and CT4, less cross-reactive than many natural variants with the human sera (Fig. 7, top and bottom right, ARC curves). We had expected CT1 to be the most cross-reactive among the four centroids because it had the lowest variance in sequence differences to the natural variants (Fig. 6A). It appeared that low sequence differences with the natural variants were an essential but not sufficient predictor of broad OspC cross-reactivity. To investigate structural factors contributing to the antigenic breadth of OspC variants, we obtained a structural alignment of the evolutionary analogs with a solved OspC structure (PDB ID: 1F1M) [71] (Supplementary Information Fig. S5). Measurements of structural variability showed the high structural similarity of the evolutionary analogs with the natural OspC variants as well as among the evolutionary analogs themselves (Supplementary Information Data S6). The structural alignment provided a basis for further comparative analysis to identify the amino-acid residues associated with the variability in antigenicity among the natural and synthetic OspC variants.

Fig. 7 Reactivity of synthetic analogs with naturally infected human and mouse sera. (Top left) Reactivity (z-score, y-axis) of natural OspC variants (n = 16) and evolutionary analogs (n = 6) (x-axis) with sera from naturally infected P. leucopus mice (n = 10). All six evolutionary analogs showed significantly higher (with t tests) reactivity with the sera of the reservoir host species than the reactivity of the natural variants as a group. (Top right) Antigen reaction characteristics (ARC) curves showed consistently high reactivity of the six evolutionary analogs, indicating their broader antigenicity relative to the natural OspC variants in reacting with the sera of the reservoir hosts of B. burgdorferi. (Bottom left and right) Corresponding graphs with sera from naturally infected human patients (n = 41). Three centroids (CT2, CT3, and CT4) showed significantly higher (with t tests) reactivity than the reactivity of the natural variants as a group. Reactivity of the other three evolutionary analogs (CT1, Consense, and Root) was significantly lower than or not significantly different from the reactivity of the natural antigen variants.
DISCUSSION

High antigenic specificities of natural OspC variants

Previous field-based studies have established an overabundance of ticks infected by a mixture of Lyme pathogen strains identified by their ospC alleles [23, 45]. In the present study, we further tested immunological distinctness of diverse B. burgdorferi strains co-existing in the Northeast US using field-collected I. scapularis ticks. Composition of B. burgdorferi strains in individual infected ticks especially in nymphs—having fed on a single blood meal from a single host—faithfully reflects the spirochete composition in reservoir hosts [23, 53]. As such, we expected that the frequency from a single host overabundance of pairs of ticks especially in nymphs we conclude the immunological distinctness of pathogens tend to be infected by multiple strains, indicating a lack of cross-protective immunity in reservoir hosts. By extension, we conclude the immunological distinctness of B. burgdorferi strains carrying different ospC alleles in nature.

Experimental infection in laboratories using B. afzelii, a Lyme pathogen common in Europe and Asia, showed that mice immunized with one recombinant OspC variant protected the host from infection by a strain carrying the homologous OspC variant but not by the strain carrying a heterologous OspC variant [37]. These strains, however, differed in genomic background besides the ospC sequences. Immunological mechanisms by which the host serum neutralizes spirochetes carrying a homologous but not a heterologous ospC allele was elucidated using genetic manipulations and immunodeficient mice, firmly establishing the causal role of the OspC molecule in eliciting strain-specific protective humoral immunity in B. burgdorferi hosts [39]. Sequences that are conserved among OspC variants, e.g., the C7 and C10 domains, are unlikely to be the targets of NFDS and indeed do not elicit protective immunity [46, 49]. Instead, immunodominant epitopes have been mapped to the highly variable regions including the C-terminus domains [47, 72]. Furthermore, conformational epitopes and structural integrity of the OspC molecules are required to trigger protective immunity [60, 73, 74].

By immunizing the C3H mice and the reservoir species P. leucopus with recombinant OspC proteins and quantifying antigenic reactions using ELISA and immunoblots, a previous study [48] and the present work demonstrated the high antigenic specificities of natural OspC variants with the homologous sera, and the much diminished reactivity of OspC variants with the heterologous sera (Figs. 4, 5).

To summarize these field-based and lab-based studies, we use the term MAD to describe the immunological distinction of natural OspC variants and their dominant role in maintaining B. burgdorferi diversity in nature. Evidence of antigenic separation among natural OspC variants emerged first from population genetic surveys of ospC sequence variability and allele frequencies in natural B. burgdorferi populations, which showed strong balancing selection driving genetic diversity at the ospC locus mediated by ecological mechanisms including immune escape, host species specialization, or both [22, 25, 42]. Subsequent whole-genome sequencing revealed frequent recombination among co-existing strains and ospC being a recombination hotspot as well as the most polymorphic single-copy gene in the B. burgdorferi genome [40, 75]. We showed by forward-evolution simulation that the combined forces of homologous recombination and negative-frequency-dependent selection were sufficient to explain the seemingly paradoxical pattern of the high recombination rate at ospC and the sequence hyper-variability at the same locus [40].

An epidemiological model offers a more intuitive understanding of the paradox of sustained linkage disequilibrium in the presence of genetic recombination at an antigen locus [3]. Using a token antigen consisting of two bi-allelic epitope sites (e.g., A1 and A2 at site A, B1 and B2 at site B), Gupta et al. predicted complete linkage disequilibrium resulting in a population consisting of only A1B1 and A2B2 haplotypes without the crossover A1B2 and A2B1 haplotypes, if it could be assumed that the host antibodies neutralize A1 and A2 (as well as B1 and B2) specifically without cross-reactivity (i.e., anti-A1 not binding A2 and vice versa). This is because, in such a system the A1B1-genotyped microbes would survive the host producing antibodies against A2 and B2, the A2B2-genotyped microbes would survive the host producing antibodies against A1 and B1, but the A1B2- or A2B1-genotyped microbes would not survive either host. This simple epidemiological model thus predicts maximum antigenic divergence (two bits of difference between A1B1 and A2B2) when the host immunity is highly allele-specific. This model, known as the strain theory, has been further refined and used to understand the stable coexistence and the temporal persistence of diverse strains in natural pathogen populations including the influenza A (H3N2) virus and malaria [6, 76].

Mechanism of maximum antigen diversification: a mathematical model

To explore immunological and molecular mechanisms underlying the broad antigenicity of evolutionary centroids, we proposed a mathematical model in which antigen sequences in a pathogen population were represented by binary strings of 1 s and 0 s (see Methods). Although there are potentially 20 amino-acid states at each alignment site, the binary representation of antigen sequences is justified on the basis of sequence variability within pathogen populations which consists predominantly of single-nucleotide changes. We subsequently simulated MAD and evolutionary centroids using genetic algorithms (simulation code available as Supplementary Information Text S2). Simulating OspC sequence evolution with genetic algorithm, which generates diverse binary strings through genetic mechanisms including random mutation and recombination, is justified on the basis that the ospC locus is a recombination hotspot on the circular plasmid cp26 in B. burgdorferi [14, 40, 77]. Note that in general genetic algorithms attain local maximal divergence but not a global maximum.

The binary model revealed a theoretical bound $D = 2d (1 - d)$ at a particular evolutionary distance $p$, where $D$ is the minimum sequence distance among the binary strings, $d$ is the maximum distance of these strings to a centroid (represented by a string of 0 s, i.e., all ancestral states, at the root), and $p$ is the probability of 1 (i.e., a derived state) (Fig. 6D). In this formulation, $p = d$, leading to $D = 2p(1 - p)$. As the population evolves, $p$ increases and results in increasing sequence diversity among the strings ($D$) as well as increasing distances to the centroid ($d$), as shown by the random points in Fig. 6D. When the sequences are under selection for diversification from one another, the $D$ value deviates far above the random points and is maximized to $D = 0.5$ at $p = 0.5$. Indeed, the natural OspC variants are separated from one another at an average sequence difference of $D = 0.462$ in fraction of variable sites (not counting the constant sites), in agreement with the maximum sequence separation driven by diversifying selection (Fig. 6B). Presence of recombination, more than mutation, is a key force driving the maximal and approximately uniform sequence divergence among the OspC variants. Simulated sequence divergence without recombination resulted in lower and more dispersed pairwise sequence differences (Supplementary Information Text S2).

Furthermore, the binary model indicated that the genetic algorithms we designed was effective in generating centroids close to the theoretical bounds (Fig. 6D). With a maximum distance of $d = 0.415$, the binary model suggests that OspC centroids with smaller distances to the natural variants are unlikely...
to be discovered given the length of variable site (~100 amino acids) and this set of 16 natural variants. In summary, the binary model elucidates a theoretical boundary of OspC sequence variability within the B. burgdorferi populations as well as a theoretical limit in sequence distances of possible OspC centroids.

In a token model with the use of ten 20-bit strings, we simulated a population of maximally diverged antigens (Table 2) and validated the central positions of the consensus and centroid variants with a neighbor-joining tree based on pairwise Hamming distances (Fig. 8 top right, tree). Simulation results were further validated by tabulating pairwise Hamming distances into a distance matrix, which showed a narrow range of distances (6 to 10) between the simulate results and experimental natural OspC variants (Supplementary Information Fig. S5). We obtained z-scores by normalizing the Hamming distance scores to the native variants (Supplementary Information Fig. S5).

Validation of structural similarity of evolutionary analogs to the natural OspC variants requires experimental interrogation with e.g., circular dichroism (CD) and nuclear magnetic resonance spectroscopy [46]. Computational and experimental structural analyses are needed to identify the structural determinants of variability in antigenicity among the OspC variants.

Implications to diagnostic and vaccine development

A new class of broad-spectrum diagnostics and vaccines could be designed by counteracting the evolutionary trend of maximum antigenic divergence in local B. burgdorferi populations. In diagnosis, the standard two-tiered testing (STTT) is based on EIA and immunoblots and lacks sensitivity for patients who develop acute erythema migrans, an early-stage Lyme disease [62, 78]. The newly recommended modified two-tiered testing (MTTT) protocol consisted of two EIAs without immunoblot and improved the sensitivity of detecting early Lyme disease cases [79, 80]. The use of multiple OspC variants may further improve diagnostic sensitivity with their broad reactivity with diverse B. burgdorferi strains [48]. With a similar ability to react with diverse B. burgdorferi strains and with a single antigen, the centroid antigens are novel diagnostic candidates if they pass specificity tests [61].

Currently there is no human-use vaccine against Lyme pathogens on the US market [81–83]. The design of currently available OspC/OspA-based vaccines for canine use was based on identification of immunodominant epitopes in individual OspC variants and concatenating them into linear multivalent super-antigens [46, 72, 84]. A multivalent vaccine consisting of as many as eight OspC-type specific epitopes has been shown to be immunogenic [49]. Because of the large number of OspC variants co-circulating in a local endemic area (e.g., ~20 in the Northeast US), it is unclear the efficacy of chimeric vaccines to elicit broadly protective immunity in humans [50, 85]. Critically, studies have shown that immune protection of OspC-based vaccines required experimental interrogation with e.g., circular dichroism (CD) and nuclear magnetic resonance spectroscopy [46]. Computational and experimental structural analyses are needed to identify the structural determinants of variability in antigenicity among the OspC variants.

Table 2. Simulated maximally diverged variants and evolutionary analogs.

| Variants | Epitope sequencesb (20-bit strings) | Hamming distancesc |
|----------|-----------------------------------|-------------------|
| A01      | 010001011101001010               | 0.70 0.40 0.70 0.55 0.55 0.55 0.55 0.60 0.65 0.50 0.45 |
| A02      | 110100001101100100               | 14 0.40 0.30 0.55 0.55 0.55 0.45 0.50 0.45 0.30 0.45 |
| A03      | 110001001110111001               | 8 8 0.40 0.75 0.55 0.55 0.65 0.60 0.35 0.40 0.45 |
| A04      | 110100101101100110               | 14 6 8 0.55 0.45 0.35 0.35 0.50 0.45 0.20 0.35 |
| A05      | 000010111001110110               | 11 11 15 11 11 11 0.50 0.40 0.40 0.35 0.60 0.45 0.40 |
| A06      | 100111101011100110111           | 11 11 11 9 10 10 0.60 0.50 0.75 0.50 0.55 0.40 |
| A07      | 110000101010000100              | 11 7 11 7 8 12 0.50 0.45 0.50 0.25 0.30 |
| A08      | 001100000111101111111           | 11 9 13 7 8 10 10 0.45 0.50 0.25 0.40 |
| A09      | 1010001101010110101             | 12 10 12 10 10 7 15 9 9 0.55 0.40 0.45 |
| A10      | 1110100011111101110100         | 13 9 7 9 12 10 10 10 10 11 0.35 0.30 |
| Consensec | 1100000011100111100            | 10 6 8 4 9 11 5 5 8 7 0.15 |
| Centroidc | 110000011011001110001          | 9 9 9 7 8 8 6 8 9 6 3 |

aBoth the maximally divergent variants (A01 through A10) and the evolutionary analogs (“Consense”/“CS” and “Centroid”/“CT”) were generated using genetic algorithms (analysis shown in Fig. 8; code shown in Supplementary Information S4 R Markdown).

bEach string represents an antigen consisting of 20 epitopes with two possible states (0 and 1). Substrings identical to those in the Centroid are underlined to highlight the interleaved nature of antigen similarities.

cHamming distances pairwise string differences (lower triangle) and length-adjusted relative distances (upper triangle).
In the current study, we described three evolutionary algorithms, proposed a theoretical model, and presented the initial proof-of-concept results demonstrating the broad antigenicity of the evolutionary centroids. The broader antigenicity of the centroids relative to the natural OspC variants makes the centroids promising candidates for improved diagnostics for Lyme disease. For vaccine development, it is further necessary to quantify the immunogenicity and test the protective efficacy of the centroids after immunization of mice with the synthetic OspC variants. Reactivity of the centroid-specific antibodies would then be tested against diverse B. burgdorferi strains through, e.g., immunofluorescence assays of cells present in field-collected ticks. Indeed, experiments are under way in our labs to generate centroid-specific antibodies, measure their immunogenicity and bacterial neutralization capability, and evaluate vaccine efficacy by challenging centroid-immunized mice with infected ticks carrying diverse B. burgdorferi strains. If validated, the OspC centroids would constitute a novel class of Lyme disease vaccines for humans and animals. If used as reservoir-targeted vaccines [86], the centroid antigens have the potential to reduce spirochete...
loads in natural reservoir hosts by eliciting immunity against all Lyme pathogenic strains.

Vaccines based on centroid antigens would be similar to the COBRA (Computationally Optimized Broadly Reactive Antigen) vaccines against influenza viruses and the vaccine candidates against HIV-1 viruses based on the "center-of-tree" ancestral sequences [8, 52]. All three approaches are based on principles of antigen evolution and use automated computational design. While the COBRA design is based on consensus sequences and the center-of-tree algorithm infers ancestral sequences, the centroid design uses genetic algorithms to minimize sequence differences to the natural antigen variants. In the present study, the centroid algorithm did not enforce any structural constraints on the OspC molecule beyond the primary sequences. Additional functional and structural constraints to OspC diversification certainly exist. Indeed, the synthetic OspC centroids were less soluble than native variants under laboratory conditions, suggesting reduced structural stability of the synthetic OspC analogs. One approach of identifying additional constraints to OspC evolution is to develop a computational classifier by fine-tuning the pretrained universal protein models with a large number of sequences of natural OspC variants [87, 88]. Such an OspC-specific classifier should help identify centroids with improved functional and structural integrity while maintaining broad antigenicity.

Stable coexistence of antigen variants like OspC variants in B. burgdorferi is widespread in natural pathogen populations. The Dengue viral populations consist of four antigenically distinct serotypes associated with sequence variations of the envelope protein [89]. The influenza B viral populations contain two evolutionary lineages associated with sequence variations of hemagglutinin [90]. The malaria parasite populations are structured into antigenic groups associated with genetic variations of the var genes encoding an erythrocyte membrane protein [6]. If these pathogen strains indeed represent ecological niches shaped by host immunity [3, 91], evolutionary centroids would be a novel and effective strategy against a broad range of microbial pathogens.

**DATA AVAILABILITY**

All datasets are included in the Supplementary Information.

**CODE AVAILABILITY**

Source codes of the Perl and Python implementations of the centroid algorithm are available in a Github repository (https://github.com/veigangag/qg-div). Also available in the same Github repository are R scripts for generating the figures.

**REFERENCES**

1. Allen JA, Clarke BC. Frequency dependent selection: homage to E. B. Poulton. Biol J Linn Soc. 1984;23:15–8.

2. Papkou A, Guzella T, Yang W, Koepper S, Pees B, Schalkowski R, et al. The genomic basis of Red Queen dynamics during rapid reciprocal host-pathogen coevolution. Proc Natl Acad Sci USA. 2019;116:923–8.

3. Gupta S, Maiden MC, Feavers IM, Nee S, May RM, Anderson RM. The maintenance of genetic isolation among species and underscores benefits and constraints to studying intra-specific epidemiological processes. PLoS ONE. 2014;9:e94384.

4. Crevar CJ, Carter DM, Lee KYJ, Ross TM. Cocktail of H5N1 COBRA HA vaccines elicit protective antibodies against H5N1 viruses from multiple clades. Hum Vaccines Immunother. 2015;11:572–83.

5. Houser K, Subbarao K. Influenza vaccines: challenges and solutions. Cell Host Microbe. 2015;17:295–300.

6. Schwartz AM, Hinckley AF, Mead PS, Hook SA, Kugeler KJ. Surveillance for Lyme disease—United States, 2008–2015. Morb Mortal Wkly Rev Surveill Summ Wash DC. 2002;51:661–2.

7. Adeoula M, Gupta RS. A phylogenomic and molecular marker based proposal for the division of the genus Borrelia into two genera: the emended genus Borrelia containing only the members of the relapsing fever Borrelia, and the genus Borrelia gen. nov. containing the members of the Lyme disease Borrelia (Borrelia burgdorferi sensu lato complex). Antimicrob Agents Chemother. 2014;58:1049–72.

8. Margos G, Marosevic D, Cutler S, Derdakova M, Diuk-Wasser M, Emler S, et al. There is inadequate evidence to support the division of the genus Borrelia. Int J Ecol. 2017;2017:1081–4.

9. Devlin K, Bestor A, Rosa PA. Lipoprotein succession in Borrelia burgdorferi: similar but distinct roles for OspC and VlsE at different stages of mammalian infection. Mol Microbiol. 2013;89:216–27.

10. Aslam B, Nisar MA, Khurshid M, Farooq, Salamat MK. Immune escape strategies of Borrelia burgdorferi. Future Microbiol. 2017;12:1219–37.

11. Coutte L, Botkin DJ, Gao L, Norris SJ. Detailed analysis of sequence changes occurring during vlsE antigenic variation in the mouse model of Borrelia burgdorferi infection. PLoS Pathog. 2009;5:e1000293.

12. Jacquot M, Gonnert M, Ferquel E, Abrial D, Claude A, Gasqui P, et al. Comparative population genomics of the Borrelia burgdorferi species complex reveals high degree of genetic isolation among species and underscores benefits and constraints to studying intra-specific epidemiological processes. PLoS ONE. 2014;9:e94384.

13. Selfert SN, Khatchikian CE, Zhou W, Brisson D. Evolution and population genomics of the Lyme borreliosis pathogen, Borrelia burgdorferi. Trends Genet TIG. 2015;31:201–7.

14. Schwartz I, Margos G, Casjens SR, Qiu W-G, Eggers CH. Multiparigene genome of Lyme disease Borrelia: Structure, Variation and Prophages. Curr Issues Mol Biol. 2012;40:49–54.

15. Pritt BS, Respicio-Kingry LB, Sloan LM, Schrieve ME, Replogle AJ, Bjork J, et al. Borrelia mayonii sp. nov., a member of the Borrelia burgdorferi sensu lato complex, detected in patients and ticks in the upper midwestern United States. Int J Ecol. 2016;66:4878–80.

16. Wang I-N, Dykhuizen DE, Qiu W, Dunn JJ, Bosler EM, Luft BJ. Genetic diversity of OspC in a local population of Borrelia burgdorferi sensu stricto. Genetics. 1999;151:15–30.

17. Tilly K, Bestor A, Rosa PA. Lipoprotein succession in Borrelia burgdorferi: similar but distinct roles for OspC and VlsE at different stages of mammalian infection. Mol Microbiol. 2013;89:216–27.

18. Norris SJ. vls antigenic variation systems of Lyme disease Borrelia: eluding host immunity through both random, segmental gene conversion and framework heterogenity. Microbiol Spectr. 2014;2:MMDNA3-0038-2014.

19. Lin T, Gao L, Edmondson DG, Jacobs MB, Philipp MT, Norris SJ. Central role of the Holliday junction helicase RuvA/B in vlsE recombination and infectivity of Borrelia burgdorferi. PLoS Pathog. 2009;5:e1000679.

20. Zhang JR, Hardham AM, Barbour AG, Norris SJ. VlsE antigenic variation in Lyme disease Borrelia by promiscuous recombination of VMP-like sequence cassettes. Cell. 1997;89:275–85.

21. Glöckner G, Schulte-Spechtel U, Schlüchter M, Felder M, Sühnel J, Wilke B, et al. Comparative genomic analysis: selection pressure on the Borrelia vls cassettes is essential for infectivity. BMC Genom. 2006;7:211.

22. Gravès CI, Carter DM, Lee KYJ, Ross TM. Cocktail of H5N1 COBRA HA vaccines elicit protective antibodies against H5N1 viruses from multiple clades. Hum Vaccines Immunother. 2015;11:572–83.

23. Houser K, Subbarao K. Influenza vaccines: challenges and solutions. Cell Host Microbe. 2015;17:295–300.
33. Carrasco SE, Troxell B, Yang Y, Brandt SL, Li H, Sandusky GE, et al. Outer surface protein OspC is an antiphagocytic factor that protects Borrelia burgdorferi from phagocytosis by macrophages. Infect Immun. 2015;83:4848–60.

34. Ønder Ö, Humphrey PT, McOmber B, Korobova F, Francella N, Greenbaum DC, et al. OspC is potent plasminogen receptor on surface of Borrelia burgdorferi. J Biol Chem. 2012;287:16860–8.

35. Rolland M, Jensen MA, Nickle DC, Yan J, Learn GH, Heath L, et al. Borrelia burgdorferi strain-specific OspC-mediated immunity in mice. Infect Immun. 1997;65:4661–7.

36. OspC is a major target of humoral immunity in Borrelia burgdorferi-infected mice. Infect Immun. 2007;75:4766–72.

37. Stajich JE. An introduction to BioPerl. Methods Mol Biol Clifton NJ. 2007;406:535

38. Hoen AG, Margos G, Bent SJ, Diuk-Wasser MA, Barbour A, Kurtenbach K, et al. Lyme disease risk not amplified in a species-poor vertebrate community: similar spillover to Ixodes scapularis. Proc Natl Acad Sci 2009; 106:15013–8.

39. Krupka M, Masek J, Barkocziova L, Knotigova PT, Kulich P, Plockova J, et al. The risk of misdiagnosis of indirect ELISA by normalizing serum-specific background noise: The example of detecting anti-FGFR3 autoantibodies. J Immunol Methods. 2019;446:52–60.

40. CDC. Recommendations for test performance and interpretation from the Second National Conference on Serologic Diagnosis of Lyme Disease. MMWR Mort Wkly Rep. 1995;44:590–1.

41. Ivanova LB, Tomova A, Gonzalez-Acuña D, Murúa R, Moreno CX, Hernández C, et al. 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. 2013;15:1609–18.

42. Schindelin J, Rueden CT, Hiner MC, Elcisy KE. The ImageJ ecosystem: An open platform for biomedical image analysis. Mol Reprod Dev. 2015;82:518–29.

43. Önder Ö, Humphrey PT, McOmber B, Korobova F, Francella N, Greenbaum DC, et al. OspC is potent plasminogen receptor on surface of Borrelia burgdorferi. J Biol Chem. 2012;287:16860–8.

44. Hoen AG, Margos G, Bent SJ, Diuk-Wasser MA, Barbour A, Kurtenbach K, et al. Lyme disease risk not amplified in a species-poor vertebrate community: similar spillover to Ixodes scapularis. Proc Natl Acad Sci 2009; 106:15013–8.

45. Durand J, Herrmann G, Genné O, Voordouw MJ. Cross-reactive acquired immunity influences transmission success of the Lyme disease pathogen, Borrelia afzelii. Infect Genet Evol J Mol Epidemiol Evol Genet Infect Dis. 2013;13:613–40.

46. Jacob J, Durand J, Rais O, Voordouw MJ. Cross-reactive acquired immunity influences transmission success of the Lyme disease pathogen, Borrelia afzelii. Infect Genet Evol J Mol Epidemiol Evol Genet Infect Dis. 2013;13:613–40.

47. Jackel S, Durand J, Rais O, Voordouw MJ. Cross-reactive acquired immunity influences transmission success of the Lyme disease pathogen, Borrelia afzelii. Infect Genet Evol J Mol Epidemiol Evol Genet Infect Dis. 2013;13:613–40.

48. Igarashi M, Ander KO, Shank J, Wannemuehler KA, Jolly MT, et al. Lyme disease vaccine: approaches for improving the immune response to chimeric vaccinogens. Vaccine. 2007;25:3419–26.

49. Izac JR, Marconi RT. Diversity of the Lyme disease spirochetes and its influence on immune responses to infection and vaccination. Vet Clin N Am Small Anim Pract. 2019;49:671–86.

50. Baum E, Randall AZ, Zeller M, Barbour AG. Inferring epitopes of a polymorphic antigen amid broadly cross-reactive antibodies using protein microarrays: A study of OspC proteins of Borrelia burgdorferi. PLoS ONE. 2013;8:e67445.

51. Johnson DL, Gomes-Solecki M. Oral immunization with OspC does not prevent tick-borne Borrelia burgdorferi infection. PLoS ONE. 2016;11:e0151850.

52. Rolland M, Jensen MA, Nickle DC, Yan J, Learn GH, Heath L, et al. Multistain Infections with Lyme Borreliosis Pathogens in the Tick Vector. Appl Environ Microbiol. 2014;81:7740–5.

53. Heaton K, Vargas LC, Gionfriddo ME, Rapoport JL, Newport JG. Evidence for lateral transfer and recombination within the feeding tick vector. PLOS Pathog. 2018;14:e1006959.

54. Önder Ö, Humphrey PT, McOmber B, Korobova F, Francella N, Greenbaum DC, et al. OspC is potent plasminogen receptor on surface of Borrelia burgdorferi. J Biol Chem. 2012;287:16860–8.

55. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32:1792–7.

56. Hoen AG, Margos G, Bent SJ, Diuk-Wasser MA, Barbour A, Kurtenbach K, et al. Lyme disease risk not amplified in a species-poor vertebrate community: similar spillover to Ixodes scapularis. Proc Natl Acad Sci 2009; 106:15013–8.

57. Stajich JE. An introduction to BioPerl. Methods Mol Biol Clifton NJ. 2007;406:535–48.

58. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30:1312–3.

59. Önder Ö, Humphrey PT, McOmber B, Korobova F, Francella N, Greenbaum DC, et al. OspC is potent plasminogen receptor on surface of Borrelia burgdorferi. J Biol Chem. 2012;287:16860–8.

60. Krupka M, Masek J, Barkocziova L, Knotigova PT, Kulich P, Plockova J, et al. The risk of misdiagnosis of indirect ELISA by normalizing serum-specific background noise: The example of detecting anti-FGFR3 autoantibodies. J Immunol Methods. 2019;446:52–60.

61. CDC. Recommendations for test performance and interpretation from the Second National Conference on Serologic Diagnosis of Lyme Disease. MMWR Mort Wkly Rep. 1995;44:590–1.

62. CDC. Recommendations for test performance and interpretation from the Second National Conference on Serologic Diagnosis of Lyme Disease. MMWR Mort Wkly Rep. 1995;44:590–1.

63. CDC. Recommendations for test performance and interpretation from the Second National Conference on Serologic Diagnosis of Lyme Disease. MMWR Mort Wkly Rep. 1995;44:590–1.

64. CDC. Recommendations for test performance and interpretation from the Second National Conference on Serologic Diagnosis of Lyme Disease. MMWR Mort Wkly Rep. 1995;44:590–1.

65. CDC. Recommendations for test performance and interpretation from the Second National Conference on Serologic Diagnosis of Lyme Disease. MMWR Mort Wkly Rep. 1995;44:590–1.

66. CDC. Recommendations for test performance and interpretation from the Second National Conference on Serologic Diagnosis of Lyme Disease. MMWR Mort Wkly Rep. 1995;44:590–1.

67. CDC. Recommendations for test performance and interpretation from the Second National Conference on Serologic Diagnosis of Lyme Disease. MMWR Mort Wkly Rep. 1995;44:590–1.

68. CDC. Recommendations for test performance and interpretation from the Second National Conference on Serologic Diagnosis of Lyme Disease. MMWR Mort Wkly Rep. 1995;44:590–1.

69. CDC. Recommendations for test performance and interpretation from the Second National Conference on Serologic Diagnosis of Lyme Disease. MMWR Mort Wkly Rep. 1995;44:590–1.

70. CDC. Recommendations for test performance and interpretation from the Second National Conference on Serologic Diagnosis of Lyme Disease. MMWR Mort Wkly Rep. 1995;44:590–1.

71. CDC. Recommendations for test performance and interpretation from the Second National Conference on Serologic Diagnosis of Lyme Disease. MMWR Mort Wkly Rep. 1995;44:590–1.

72. CDC. Recommendations for test performance and interpretation from the Second National Conference on Serologic Diagnosis of Lyme Disease. MMWR Mort Wkly Rep. 1995;44:590–1.

73. CDC. Recommendations for test performance and interpretation from the Second National Conference on Serologic Diagnosis of Lyme Disease. MMWR Mort Wkly Rep. 1995;44:590–1.

74. CDC. Recommendations for test performance and interpretation from the Second National Conference on Serologic Diagnosis of Lyme Disease. MMWR Mort Wkly Rep. 1995;44:590–1.

75. CDC. Recommendations for test performance and interpretation from the Second National Conference on Serologic Diagnosis of Lyme Disease. MMWR Mort Wkly Rep. 1995;44:590–1.

76. CDC. Recommendations for test performance and interpretation from the Second National Conference on Serologic Diagnosis of Lyme Disease. MMWR Mort Wkly Rep. 1995;44:590–1.

77. CDC. Recommendations for test performance and interpretation from the Second National Conference on Serologic Diagnosis of Lyme Disease. MMWR Mort Wkly Rep. 1995;44:590–1.

78. CDC. Recommendations for test performance and interpretation from the Second National Conference on Serologic Diagnosis of Lyme Disease. MMWR Mort Wkly Rep. 1995;44:590–1.

79. CDC. Recommendations for test performance and interpretation from the Second National Conference on Serologic Diagnosis of Lyme Disease. MMWR Mort Wkly Rep. 1995;44:590–1.

80. CDC. Recommendations for test performance and interpretation from the Second National Conference on Serologic Diagnosis of Lyme Disease. MMWR Mort Wkly Rep. 1995;44:590–1.
85. Earnhart CG, Marconi RT. An octavalent Lyme disease vaccine induces antibodies that recognize all incorporated OspC type-specific sequences. Hum Vaccin. 2007;3:281–9.
86. Schuijt TJ, Hovius JW, van der Poll T, van Dam AP, Fikrig E. Lyme borreliosis vaccination: the facts, the challenge, the future. Trends Parasitol. 2011;27:40–7.
87. Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci. 2021;118:e2016239118.
88. Strodthoff N, Wagner P, Wenzel M, Samek W. UDSMProt: universal deep sequence models for protein classification. Bioinformatics. 2020;36:2401–9.
89. Chen R, Vasilakis N. Dengue—Quo tu et quo vadis? Viruses. 2011;3:1562–608.
90. van de Sandt CE, Bodewes R, Rimmelzwaan GF, de Vries RD. Influenza B viruses: not to be discounted. Future Microbiol. 2015;10:1447–65.
91. Buckee CO, Recker M, Watkins ER, Gupta S. Role of stochastic processes in maintaining discrete strain structure in antigenically diverse pathogen populations. Proc Natl Acad Sci. 2011;108:15504–9.

ACKNOWLEDGEMENTS
This work was supported by the Public Health Service awards AI139782 (to WQ) and AI072810, AI074092 and AI155211 (to MGS) from the National Institute of Allergy and Infectious Diseases (NIAID) of the National Institutes of Health (NIH) of the United States of America. Additional funding includes the grant CK000107 (to MGS) from the US Centers for Disease Control and Prevention (CDC) and the award 7F2400001 (to SM) from the PhRMA Foundation. LD and SK are supported in part by the award EB030275 from the National Institute of Biomedical Imaging and Bioengineering (NIBIB, to Brian Zeglis and WQ) of the National Institutes of Health (NIH). The content of this paper is solely the responsibility of the authors and does not necessarily represent the official views of NIAID, NIH, or CDC. SA is supported in part by the Doctoral Program in Biology of the City University of New York. We thank Dr Mirella Salvatore (Weil Cornell Medical College) for introducing us to the study of influenza B viruses. Dr Christopher Sexton and Dr Jeannine Petersen (Division of Vector-Borne Diseases, CDC) prepared a custom panel of human sera for this study. Roman Shimonov and Justin Hiraldo (both of Hunter College) digitalized immunoblot images. We thank three anonymous referees for careful reading and constructive critiques of the original draft.

AUTHOR CONTRIBUTIONS
Conceptualization, funding acquisition, and supervision: MGS, WQ. Model development: SA, BS, SM, and WQ. Experimental Investigations: LD, SA, LI, and BW. Software implementation: LD, EB. Data analysis: LD, SA, EB, BS, LI, and WQ. Writing—original draft: WQ. Writing—review and editing: LD, SA, BS, EB, SM, and MGS.

COMPETING INTERESTS
The following authors declare potential competing interests: MGS (patents) and WGQ (patents). All other authors declare no competing interests in relation to this study.

ADDITIONAL INFORMATION
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41396-021-01089-4.
Correspondence and requests for materials should be addressed to W.-G.Q.
Reprints and permission information is available at http://www.nature.com/reprints
Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.