Genomic insights into the ancient spread of Lyme disease across North America

Katharine S. Walter1, Giovanna Carpi2, Adalgisa Caccone3 and Maria A. Diuk-Wasser4

Lyme disease is the most prevalent vector-borne disease in North America and continues to spread. The disease was first clinically described in the 1970s in Lyme, Connecticut, but the origins and history of spread of the Lyme disease bacteria, Borrelia burgdorferi sensu stricto (s.s.), are unknown. To explore the evolutionary history of B. burgdorferi in North America, we collected ticks from across the USA and southern Canada from 1984 to 2013 and sequenced the, to our knowledge, largest collection of 146 B. burgdorferi s.s. genomes. Here, we show that B. burgdorferi s.s. has a complex evolutionary history with previously undocumented levels of migration. Diversity is ancient and geographically widespread, well pre-dating the Lyme disease epidemic of the past ~40 years, as well as the Last Glacial Maximum ~20,000 years ago. This means the recent emergence of human Lyme disease probably reflects ecological change—climate change and land use changes over the past century—rather than evolutionary change of the bacterium.

The Lyme disease spirochete, Borrelia burgdorferi sensu stricto (s.s.), is maintained in an enzootic transmission cycle, in which the bacteria is transmitted between Ixodes tick vectors and a community of vertebrate reservoir hosts. Humans are incidental hosts; we are susceptible to infection but do not contribute significantly to transmission or spread. Despite its epidemiological importance, little is known about the evolutionary history of B. burgdorferi in North America. This limits our ability to track and predict the direction of ongoing spread and implement public health interventions.

Pathogen genomes can reveal epidemic histories. Phylogeography provides a powerful framework for inferring pathogen evolutionary dynamics including probable epidemic origins, rates and patterns of spread, and the distribution of highly virulent clades9-16. However, the evolutionary history of B. burgdorferi has not been explored using high-resolution molecular markers such as whole genomes sampled from a broad spatial area. Sampling and sequencing B. burgdorferi genomes directly from ticks is challenging; bacterial DNA is swamped by tick and environmental DNA11, and culturing the bacterium introduces biases12. Only 15 genomes of B. burgdorferi s.s. existed previously (Supplementary Table 1) and previous phylogeographic study has relied on single locus5,14 or multi-locus sequence typing (MLST) markers11-17.

Here, we use a population genomics approach to investigate outstanding questions about the processes shaping B. burgdorferi variation, patterns of gene flow across North America and the timescale of bacterial evolution. (1) How is B. burgdorferi variation generated and maintained? Previous study of 13 B. burgdorferi s.s. genomes found frequent small-scale recombination9-11, but the relative importance of mutation to recombination in shaping population-level diversity is unknown. (2) Where did North American B. burgdorferi originate and how has the bacterium spread? A previous MLST-based study suggests that B. burgdorferi was historically introduced to the midwest from the northeast and the two regions are now isolated11. However, rates of gene flow of the bacteria between these two Lyme endemic regions and the relative role of different vertebrate hosts in bacterial dispersal is unknown. (3) Finally, how old is B. burgdorferi diversity? Museum specimens reveal the bacteria was present on Cape Cod, Massachusetts, in the 1890s20 and Long Island, New York, in the 1940s21, long before clinical recognition of the disease. Phylogeography suggests a deeper evolutionary history—ticks and B. burgdorferi may have survived the Last Glacial Maximum in the southern United States and spread across North America after glacial retreat, ~20,000 years ago9,13. Reconstructing the timescale of diversification will indicate if B. burgdorferi variation holds a signature of recent emergence of human Lyme disease in the past 40 years or if diversity reflects more ancient processes.

To address these three epidemiologically relevant aspects of the evolutionary history of B. burgdorferi in North America, we sampled and sequenced 146 B. burgdorferi genomes directly from tick vectors from across the northeastern and midwestern United States and from southern Canada, including samples from 1984. With this genomic collection and all previously published B. burgdorferi genomes, including samples from Europe and the western USA (California), we reconstructed the evolutionary history of the Lyme disease bacteria in North America.

Results

Field sampling and variation. We collected 146 B. burgdorferi-infected nymphal Ixodes scapularis ticks from across Lyme disease endemic areas in the northeast and midwest United States and southern Canada, spanning 30 years (1984–2013) and representing...
the widest spatial and temporal genomic collection of *B. burgdorferi* yet published (Fig. 1; Supplementary Information). We used a hybrid capture method we previously developed to enrich for and efficiently sequence *B. burgdorferi*-derived reads from a mixed DNA template (~73% capture efficiency), allowing us, for the first time, to screen *B. burgdorferi* directly from ticks. Our genomic collection allowed us to simultaneously study DNA from *B. burgdorferi*, the tick vector, and a co-vectored parasite, *Babesia microti*. To extend the geographic range of samples, we added all 15 published *B. burgdorferi* s.s. genomes (Supplementary Table 1), including two samples from Europe and three from the western USA (California), and used *Borrelia finlandensis* as an outgroup.

We identified single nucleotide polymorphisms (SNPs) along the *B. burgdorferi* linear chromosome (910,724 base pairs (bp)) and the two best-characterized and most conserved plasmids (Supplementary Fig. 1), including two recombination events in the chromosome and recombinant regions (Supplementary Fig. 1a). The relative frequency of recombination to point mutation is lower on the plasmids than on the chromosome: $u_{r/m}$ of 0.70 (plasmid cp26) and 0.259 (plasmid lp54) (Supplementary Fig. 1b,c).

Recombinant hotspots on the plasmids contain several known antigens. A recombination peak exists at the *ospC* (outer surface protein C) gene on plasmid cp26 (Supplementary Fig. 1b). *ospC* is a major *B. burgdorferi* antigen required for transmission from tick to vertebrate host and dissemination within vertebrates. *ospC* defines *B. burgdorferi* serotype: many *ospC* serotypes (alleles) circulate and vertebrates develop type-specific immunity protecting them from re-infection with the same *ospC* serotype. *ospC* serotypes vary in virulence in humans: ‘disseminating’ serotypes are associated with spread from the skin (the initial site of infection) through the bloodstream, causing more severe disease, while non-disseminating serotypes may remain as localized skin infections. Recombinant tracts vary in length: some are tightly centred on *ospC* (632 bp), while others are up to 5,000 bp long, including the neighbouring *guaA* (GMP synthase) and *guaB* (dehydrogenase) genes essential for infectivity in vertebrate hosts. A recombination hotspot on plasmid cp26 also includes *resT*, a gene involved in recombination.

**Recombination detection.** Both de novo mutation and recombination contribute to *B. burgdorferi* variation. Early studies described *B. burgdorferi* as clonal, while recent studies identified small-scale recombination events. The contribution of recombination to *B. burgdorferi* genomic variation has not been examined in a large population sample. We evaluated the contribution of both processes to simultaneously infer recombinant tracts and the underlying maximum likelihood phylogeny of the *B. burgdorferi* chromosome and plasmids cp26 and lp54 (Supplementary Fig. 1).

Recombination is frequent along the chromosome (763 recombination events) and plasmids cp26 (68 recombination events) and lp54 (53 recombination events) (Supplementary Fig. 1), and it has shaped the *B. burgdorferi* genome over its evolutionary history. For each branch of the chromosomal and plasmid trees, we determined the ratio of SNPs introduced via recombination to the number introduced via point mutations, $u_{r/m}$, as well as the total number of recombination events per branch. The chromosome had a mean $u_{r/m}$ of 1.55 (standard deviation = 3.83) and had a mean of 2.63 recombination events per branch. Despite the recombination detected, 93.18% (846,801 of 910,724 bp) of the chromosome lies within a clonal frame, outside predicted recombinant regions (Supplementary Fig. 1a). The relative frequency of recombination to point mutation is lower on the plasmids than on the chromosome: $u_{r/m}$ of 0.70 (plasmid cp26) and 0.259 (plasmid lp54) (Supplementary Fig. 1b,c).

Two sharp recombination peaks on plasmid lp54 correspond to two known *B. burgdorferi* antigens (Supplementary Fig. 1c): *dbpA*, an adhesin that enables *B. burgdorferi* dissemination in vertebrate hosts; and BB_A05, an antigen involved in transmission from ticks to mammals. Recombinant tracts in *dbpA* also vary in size from 70 to ~8,000 bp and frequently include the neighbouring *dbpB*, another adhesin. A second recombination hotspot includes two other antigens that elicit human antibody responses: S1 (*BB_A05*), highly expressed in feeding nymphal ticks during transmission to mammals; and S2 (*BbuZS7_A03*).

**Phylogenetic structure.** To investigate *B. burgdorferi* population structure, we built a maximum likelihood phylogeny of the 16,370 SNP recombination-free alignment. The chromosomal phylogeny reveals a high degree of *B. burgdorferi* population structure divided into four major clades, which do not always reflect geography (Fig. 2a). Samples from the northeast and midwest are not monophyletic, in contrast to previous MLST-based analysis suggesting the two regions had a shared past and are now isolated.

**Fig. 1 | Map of *B. burgdorferi* samples collected in North America.** a, Distribution of 146 *I. scapularis* field-collected ticks. Samples are coloured by major sampling region. b, Distribution of samples collected each year.
Similarly, southern samples are not monophyletic and are distributed across the tree.

To explore the population structure of B. burgdorferi on a broader spatial scale, we built a maximum likelihood phylogeny including all 15 previously published B. burgdorferi s.s. genomes in addition to the 146 samples collected for this study (Supplementary Fig. 2). The three samples from the western USA (California) fall in two distinct clades, indicating that western B. burgdorferi are not genetically isolated from midwestern and northeastern samples (Supplementary Fig. 2). Although B. burgdorferi is maintained in a different ecological cycle in the western USA, where it primarily cycles between Ixodes pacificus ticks and wood rats32, the bacteria do not seem to be genetically differentiated from other North American samples. The two European B. burgdorferi s.s. genomes are closely related to each other and monophyletic (Supplementary Fig. 2, pink tips), nesting in a clade that contains midwestern samples. Previous MLST-based analysis suggested a European origin of all B. burgdorferi33. Although we cannot formally test this hypothesis because of the limited European sampling for this study, the placement of the two European samples within a clade including only midwestern samples suggests a more complex pattern of genetic differentiation than the one hypothesized with MLST-based analysis.

B. burgdorferi chromosomal and plasmid phylogenies are broadly similar, though there is evidence of several historic plasmid exchanges between lineages (Supplementary Fig. 3).

To assess whether clades vary in virulence, we investigated their ospC serotype composition by serotyping samples in silico. ospC serotypes26 cluster imperfectly within the B. burgdorferi tree, reflecting recombination discussed above (Fig. 2a; Supplementary Fig. 2). Virulent, disseminating ospC serotypes A, B, H, I and K occur within each phylogenetic cluster. However, the clades have different virulence profiles. Disseminating strains make up the majority of samples in clade I (72.1%, 95% confidence interval (CI): 56.1–84.2%) and clade II (72.5%, 95% CI: 55.9–84.9%) but only a minority of clade III (27.8%, 95% CI: 14.8–45.4%) and clade IV (7.7%, 95% CI: 1.34–26.6%) (Fig. 2a). Virulent serotypes are more common in the northeast, where 35.8% (95% CI: 28.6–43.6%) serotyped samples are disseminating compared with 15.8% (7.91–28.4%) disseminating in the midwest.

**Discordant tick and pathogen phylogenies.** Although the efficiency of our hybrid capture protocol was high (~70% of reads mapped to B. burgdorferi)35, the majority of remaining sequence data corresponded to the tick, I. scapularis, genome33 (Supplementary Fig. 4).

Coverage of three tick mitochondrial genes (16S, cytochrome oxidase II and the control region; Supplementary Table 2) enabled us to examine potential co-evolution of the bacteria and its vector46. Although the efficiency of our hybrid capture protocol was high (~70% of reads mapped to B. burgdorferi)35, the majority of remaining sequence data corresponded to the tick, I. scapularis, genome33 (Supplementary Fig. 4).

Coverage of three tick mitochondrial genes (16S, cytochrome oxidase II and the control region; Supplementary Table 2) enabled us to examine potential co-evolution of the bacteria and its vector46. Although the efficiency of our hybrid capture protocol was high (~70% of reads mapped to B. burgdorferi)35, the majority of remaining sequence data corresponded to the tick, I. scapularis, genome33 (Supplementary Fig. 4).
To test for evidence of shared evolutionary history of the two pathogens, we constructed a maximum likelihood tree based on the complete genome of the B. microti apicoplast (a small organelle found in most Apicomplexa parasites), 28.7 kbp, and compared it with the B. burgdorferi phylogeny (Supplementary Fig. 7). Despite our limited sample size of nine co-infected ticks, due to the relatively low prevalence of B. microti, we found incongruence between the two parasite phylogenies.

Migration patterns. We used ancestral state reconstruction to explore the geographic origins and gene flow patterns of North American B. burgdorferi (Fig. 2b).

The most recent common ancestor of all sampled B. burgdorferi most likely existed in the northeast (posterior probability, 77.2%), as did the most recent common ancestor of each B. burgdorferi clade (clade I: 78.4%; clade II: 89.3%; clade III: 76.7%; and clade IV: 77.6%; Fig. 2b). Our phylogeny reveals a pattern of gene flow more complex than the previously described unidirectional pattern from northeast to midwest. We uncover a signal of bidirectional gene flow between the three regions surveyed (Supplementary Fig. 9). Although migration rates between regions are not statistically different, there is a clear hierarchy in estimated rates of gene flow between regions. Migration between the northeast and midwest is most frequent, followed by migration between the northeast and south, and finally migration between the south and midwest. The most likely ancestral location for interior nodes within clades I, III and IV (Figs. 2b and 3, nodes labelled 'MW') existed in the midwest, revealing historic gene flow from the northeast to the midwest. This means that midwestern B. burgdorferi populations are not recently introduced from the northeast and that genetically distinct B. burgdorferi populations from clades I, II and IV were historically endemic both in the northeast and the midwest (Fig. 2b). Although the northeastern USA is currently the region with the highest density of tick vectors, highest prevalence of infected ticks and the greatest epidemiological burden of Lyme disease, complex patterns of historic gene flow across North America have shaped current patterns of diversity (Fig. 2b).

Emergence timing. Estimating the timing of the North American B. burgdorferi diversification requires knowledge of how fast the bacteria evolves. However, the substitution rate of this bacteria has not been measured directly, and available estimates vary widely from 10^{-6} to 10^{-4} substitutions per site per year^{13,15}. In bacteria, where fossils are not available to calibrate node ages, serially sampled tips (that is, samples collected at several time points) may enable estimation of substitution and diversification rates along the rest of the tree. This is possible if sufficient substitutions accumulate over the sampling period, within a measurably evolving population^{9,40}. To maximize our sampling period, we included four ticks from 1984, the oldest available B. burgdorferi DNA in our dataset (Fig. 1b) and the oldest existing B. burgdorferi whole genomes besides the B31 reference strain cultured from a tick in 1982^{24}. We simultaneously estimated phylogenies, substitution rates and demography with a Bayesian approach implemented in BEAST^{41}.

The observed substitution rate of 7.49×10^{-6} substitutions per site per year (95% highest posterior density (HPD) 3.32×10^{-6}–1.34×10^{-5}; Supplementary Table 3) is faster than previously estimated for B. burgdorferi^{13}. This probably reflects the bacterium’s ecology. When B. burgdorferi is transmitted from tick to vertebrate host to tick, its population experiences severe bottlenecks, so that only a fraction of the variants transmit to the next host. Genetic drift in this severely bottlenecked population may increase the observed substitution rate^{41}, although the within-host B. burgdorferi substitution rate is likely to be much slower.

We estimate the most recent common ancestor of the B. burgdorferi analysed in this study existed ~60,000 years ago (95% HPD 20,000–98,000) (Fig. 3; Supplementary Table 3). B. burgdorferi diversity is ancient and long pre-dates not only the Lyme epidemic of the past ~40 years but also the Last Glacial Maximum ~20,000 years ago^{13}. Current Lyme disease foci are in regions once covered by the Pleistocene ice sheets^{13}, Earlier studies suggest that the south was a refugia for many North American animals during the Last Glacial Maximum^{13,61} and, after glacial retreat, southern migrants colonized the northeast and midwest^{13,51}. The ancient timescale of B. burgdorferi diversification in North America suggests that this species, in addition to ticks, was endemic in North America before the Pleistocene glaciation and its evolutionary history was shaped by ancient geological events.

Our estimates of substitution rate and tree age have substantial uncertainty (Fig. 3; Supplementary Table 3), as we do not have independent data to calibrate divergence time estimates, and the 30-year sampling interval that separate our samples is orders of magnitude shorter than the estimated tree age. However, even if B. burgdorferi substitution rates are slower than estimated in this study (that is, ~10^{-7}–10^{-8} substitutions per site per year), the age of the most recent common ancestor would still be quite old, dating to several hundred thousand or a million years old. Thus, our finding that B. burgdorferi diversity is ancient and pre-dates the ongoing Lyme epidemic seems to be robust to uncertainty in the substitution rate estimate.

Demographic change. B. burgdorferi population size expanded dramatically ~20,000 years ago and remained relatively stable.
thereafter (Supplementary Fig. 10a) according to the skyline plot, a piecewise population model that allows effective population size to vary over time\textsuperscript{44}, estimated in BEAST\textsuperscript{45}. This signal may reflect population expansion from refugial populations in the south, after Pleistocene ice sheets retreated, allowing ticks and vertebrate hosts of \textit{B. burgdorferi} to repopulate wide regions of northern North America\textsuperscript{13}. This is consistent with previous studies reporting historic population size expansions in \textit{B. burgdorferi}\textsuperscript{13–15}. However, this result must be interpreted with caution because the bacterial population is geographically structured both across space and between hosts, violating the coalescent model assumption of a single, panmictic population, which can potentially bias effective population size estimates.

To complement the skyline plot analyses, we used the mismatch distribution—the distribution of pairwise SNP differences between samples—to test for historic changes in \textit{B. burgdorferi} population size (Methods). The mismatch distribution was unimodal (Supplementary Fig. 10b) and consistent with past demographic or spatial expansion. The peak of the mismatch distribution (\(t = 389\)) corresponds to the timing of historic population expansion, in units of mutational distance. Separate analyses of the mismatch distribution for samples from the northeast, midwest and south reveal a similar pattern of historic population expansion in each region.

Discussion

Here, we reconstruct the evolutionary history of the Lyme disease bacteria in North America from the widest yet collection of \textit{B. burgdorferi} genomes sequenced directly from field-collected ticks. By mining sequence by-catch and co-captured DNA in our genomic collection, we explore patterns of potential co-evolution with the tick vector and another co-infecting human parasite, \textit{B. microti}.

We find that although \textit{B. burgdorferi} is largely clonal, both recombination of short genomic tracts along the chromosome and plasmids (Supplementary Fig. 1) and shuffling of entire plasmids (Supplementary Fig. 3) shape genomic diversity. We also found that recombination hotspots include the major \textit{B. burgdorferi} antigens \textit{ospC}, \textit{dbpA} and \textit{dbpB} (Supplementary Fig. 1b,c). These surface-expressed antigens are probably experiencing balancing selection imposed by vertebrate host immune responses.

The complex phylogenetic structure of \textit{B. burgdorferi} (Fig. 2a) together with the migration rate analyses (Fig. 2b) suggest previously undocumented levels of gene flow across North America. As \textit{Ixodes} ticks move little, \textit{B. burgdorferi} spreads through the movement of its vertebrate hosts, small mammals and birds. In contrast to previous findings of strong barriers to gene flow between the northeast and midwest\textsuperscript{13,15}, we find evidence of long-distance migration events between the three major geographic regions sampled (Fig. 2b), probably due to long-distance, bird-mediated dispersal. Local dispersal by small mammals probably also contributes to gene flow on much smaller spatial scales. One limitation to this study is that phylogenies are inferred from consensus bacterial sequences from each individual tick sample. The consensus \textit{B. burgdorferi} sequence from a tick with a mixed infection\textsuperscript{46} often represents the majority strain infecting a tick (that is, the strain comprising the largest proportion of infection). However, the consensus sequence may alternatively represent a chimeric sequence, a combination of segments of the multiple co-infecting haplotypes. With short-read sequence data used here, we are unable to reconstruct multiple bacterial haplotypes within a single sample. Longer-read sequencing will enable sequencing multiple bacterial genomes from individual tick vectors in the future.

We find no support for co-evolution of \textit{B. burgdorferi} and its tick vector (Supplementary Fig. 6). While captured sequence data enabled us to examine phylogenetic structure at only a few tick genes, we do not expect further analysis to reveal a strong pattern of co-emergence of the bacteria and its vector. In contrast to obligate parasites that co-diverge with their hosts, \textit{B. burgdorferi} cycles between tick vectors and vertebrate hosts, decoupling tick and bacterial evolutionary histories. Our finding of a lack of association between tick and bacterial evolutionary histories suggests that invasion of ticks and bacteria is not coupled and may reflect distinct ecological processes. This is consistent with observations that \textit{B. burgdorferi} invasion often lags behind tick invasion. This means that epidemiological surveillance should focus on areas with established \textit{I. scapularis} tick populations that are potential sites for \textit{B. burgdorferi} introduction.

Finally, we find \textit{B. burgdorferi} diversity is ancient (~60,000 years old) (Fig. 3). Although there is substantial uncertainty in estimated divergence times, our phylogeny shows a clear signature of ancient diversification of \textit{B. burgdorferi} that long pre-dates the Lyme epidemic of the past ~40 years. Our reconstruction of the probable geographic origins of this diversity shows that, historically, it was geographically widespread across the northeast and midwest (Fig. 2b). The recent geographic and population size expansion (Supplementary Fig. 10) of \textit{B. burgdorferi} reflects the spread of diverse \textit{B. burgdorferi} already present in each region (Fig. 2a,b). This means that the recent emergence of Lyme disease does not reflect the spread of a single epidemic \textit{B. burgdorferi} lineage across North America or recent diversification of the bacteria, but rather the spread of pre-existing, geographically widespread bacterial diversity.

Our finding of ancient \textit{B. burgdorferi} diversification suggests that the recent Lyme disease epidemic does not reflect evolutionary processes but rather was driven by the ecological change in North America beginning in the colonial period ~700 years ago. Deforestation and intensive hunting\textsuperscript{47} during the colonial period followed by population explosion of white-tailed deer\textsuperscript{48,49} and climate change\textsuperscript{50–52} in the past century probably enabled dramatic range expansion of \textit{Ixodes} spp. ticks\textsuperscript{53}. The spread of ticks into environments with high densities of competent vertebrate hosts substantially widened the potential geographic range of \textit{B. burgdorferi} and of Lyme disease. Ticks and \textit{B. burgdorferi} continue to spread into southern Canada and across the United States, putting a greater population at risk of Lyme disease.

Although the evolutionary history of \textit{B. burgdorferi} occurs at a deeper timescale than the recent Lyme disease epidemic, our phylogeographic analysis provides insights into the epidemiology of Lyme disease. The high levels of \textit{B. burgdorferi} genomic diversity found across North America and continued long-distance gene flow suggests that humans may be exposed to diverse bacterial lineages regardless of where they are infected. \textit{B. burgdorferi} strains vary substantially in virulence and we find that disseminating strains associated with more severe disease are found in all regions surveyed and can occur on diverse genomic backgrounds (Fig. 2a). The patterns of migration uncovered here suggest continued spread of diverse \textit{B. burgdorferi} not only from the northeast across the rest of North America but also in several other directions (Fig. 2b), with gene flow occurring not only locally but also on a continental scale (Supplementary Fig. 2). This finding has important epidemiological consequences as it suggests that wide regions with established tick and vertebrate host populations are potential sites of \textit{B. burgdorferi} invasion and future Lyme disease.

Methods

\textbf{Bacterial sampling and sequencing.} We collected 146 \textit{B. burgdorferi}-infected nymphal \textit{I. scapularis} ticks from the widest available spatial and temporal range (including ticks sampled from 1984 to 2013) (Fig. 1; Supplementary Information). Tick sampling, DNA extractions and quantitative PCR testing for \textit{B. burgdorferi} infection followed described protocols\textsuperscript{46}. Genomic libraries were prepared from infected tick samples. \textit{B. burgdorferi} DNA was captured using a custom hybridization capture array method\textsuperscript{48}. Sequencing of 75-bp paired end reads was conducted on an Illumina HiSeq 2500 at the Yale Center for Genomic Analysis. Short-read sequence data were submitted to the NCBI Short Read Archive.
Articles

Read alignment and SNP detection. We focused analysis on SNPs on the B. burgdorferi linear chromosome (910,724 bp) and the two best-characterized and most conserved plasmids, lp54 (53,657 bp) and cp26 (26,498 bp). This represents 65% of the total B31 reference genome.

Raw sequence reads for each sample were aligned to the B. burgdorferi reference genome strain B31, using WMA mem (v. 2.7.7). Duplicate sequence reads were realigned with the Picard Suite (v. 1.117) MarkDuplicates (http://picard.sourceforge.net). Only samples with mean coverage > 10× were retained for analysis. Variants with respect to strain B31 were identified with GATK HaplotypeCaller (ploddy set to 1), generating sample-specific gVCF files. We conducted joint genotyping of samples with GATK X Genotyper2. We excluded indels (insertions and deletions) and SNPs with signals of low mapping or genotyping quality with GATK VariantFiltering, using the following filters recommended by GATK: quality by depth (QD < 2.0), Fisher strand bias (FS > 60.0), mapping quality (MQ < 40.0), mapping quality rank sum test (MQRankSum < -12.5), the Mann–Whitney rank sum test (ReadPosRankSum < -8.0), low genotype call (GQ < 20) and strand odds ratio test (SOR > 4.0), and set filtered genotypes to no call (GATK option --setFilteredGtToNcall).

To create an alignment of consensus sequences from the variant file, we used the GATK tool FastAlternateReferenceMaker. This resulted in multiple alignments of consensus sequences of the chromosome (910,724 bp), plasmid cp26 (26,498 bp) and plasmid lp54 (53,657 bp).

Phylogenetic analysis. It is critical to account for recombination when reconstructing bacterial evolutionary histories because horizontal gene transfer can overwhelm the evolutionary signal from vertically inherited mutations and skew inferred phylogenies. Therefore, we used GUBINS to infer recombinant tracts along the chromosome in addition to plasmids cp26 and lp54. We used RaxML to construct maximum likelihood phylogenetic trees from the 16,370 bp recombination-free SNP alignment of all 146 isolates. We used the generalized time-reversible model and a gamma distribution to model site-specific rate variation (the GTR + F substitution model) and corrected for using only variant sites in the tree, using the Lewis ascertainment bias correction. Support for each branch was assessed using 100 bootstrapped samples of the multiple alignment. Recombination hotspots and RAXML phylogenies are visualized in Supplementary Fig. 1 with Phandango (http://jameshadfield.github.io/phandango/).

To infer rates of gene flow between geographic regions, we performed discrete ancestral state reconstruction, coding the region of origin of tick samples as a discrete trait. We fitted a Markov model of character state evolution to the underlying BEAST phylogeny with the divForest package in R and found that posterior probability distributions of migration rates between regions overlapped significantly (Supplementary Fig. 8). Therefore, we fitted an equal rates model that considered a single long-distance migration rate between regions in phylomaps and simulated 1,000 stochastic character maps, or reconstructions of ancestral geographic location, on our phylogeny with the phylotools function make.simmap (Fig. 2b).

We identified ospC serotypes in silico with srst2. Trees were visualized in R with the packages phytools and ape and in R.

Sequence by-catch and co-capture. Although our hybrid capture approach was efficient and the majority of sequence reads corresponded to B. burgdorferi, ~30% of reads did not map to B. burgdorferi and constitute metagenomic data from our tick samples. We tested whether by-catch—unintentionally captured sequence data—corresponded to the tick vector by mapping sequence reads to the tick, I. scapularis, genome1 with WMA mem (v. 2.7.7), as described above. To investigate whether by-catch sequence contained sufficient coverage of I. scapularis genes for phylogenetic inference, we mapped reads to ten tick genes previously used for population genetic study (Supplementary Table 2). Genbank accession numbers of the I. scapularis haplotypes used as references for mapping are in Supplementary Table 2.

Three tick mitochondrial genes (16S, cytochrome oxidase II and the control region) had a mean coverage > 1.5 × (Supplementary Fig. 5, Supplementary Table 2), enabling us to use them in phylogenetic analyses to examine potential co-evolution of the bacteria and its vector1. We called variants on the three mitochondrial genes and used RaxML to infer a maximum likelihood tree, using the methods described above and including all tick samples with <40% missing data.

As we simultaneously co-captured both B. burgdorferi and the co- vectored pathogen B. microti1, we were able to explore patterns of potential co-evolution of the two co vectored pathogens. Reading variant calling and tree building of the B. microti apicoplast, a small organelle found in most Apicomplexa parasites (28.7 kbp), was conducted as previously described.

We tested for evidence of co-evolution of B. burgdorferi and both I. scapularis and B. microti with a global test of co-evolution implemented in ParaFit2. The two null hypotheses that B. burgdorferi and its tick vector and that B. burgdorferi and B. microti have independent evolutionary histories were tested by permuting the observed associations of parasite-vector or parasite-parasite- parasite 999 times.

Phylogeographic and demographic analysis. We coalesced approach implemented in BEAST 2 to simultaneously explore the phylogenetic and demographic history of sampled B. burgdorferi. Recombination-free SNP alignments were used to fit four alternative models including either a strict molecular clock or a relaxed (uncorrelated log-normal) molecular clock by allowing substitution rates to vary across tree branches and a constant population size or a Bayesian skyline model of changing population size. We used dates of tick collection to calibrate tips. We used an ascertainment bias correction to account for the Invariant sites not included in the SNP alignment. All phylogenographic models used the HKY substitution model, allowing for gamma-distributed rate variation across sites.

For each model, we ran chains of 100 million iterations or until convergence. Chains were thinned by sampling every 1,000 iterations and 10% of each chain was discarded as burn-in. Maximum clade credibility (MCC) trees were generated using TreeAnnotator, Tracer v. 1.6 was used to assess convergence visually and confirm effective sample sizes were greater than 200 for each parameter. To test the influence of prior distributions specified for evolutionary and demographic model parameters, we ran BEAST analyses for each model with no input alignment, sampling only from prior distributions and compared results with the posterior parameter estimates of fitted models (Supplementary Fig. 11).

To identify the best-fitting model, we used path sampling to estimate the marginal likelihood of each model. We compared alternative models with Bayes factors and identified the relaxed (uncorrelated log-normal) clock, skyline population size as the best-fitting model for the recombination-free chromosomal SNP alignment (Bayes factors > 80; Supplementary Table 3), used in all above analyses. All four coalescent models estimated highly congruent tree topologies and differed only in placement of deep branches, expected because of the high degree of uncertainty at deeper timescales within the B. burgdorferi tree.

To speed computational time, we performed BEAST analyses on randomly sampled 5,000-bp subsets of the full 16,370-bp SNP alignment. We resampled and reran analyses ten times for the best-fitting model and compared MCC trees to confirm that subsampling the alignment did not distort phylogenetic inference. The MCC tree depicted in Figs. 2b and 3 was inferred from a single 5,000-bp subset of the SNP alignment.

To investigate B. burgdorferi demographic history, we complemented the coalescent-based approach with mismatch distribution analysis, as implemented in DnaSP v58, and tested if our observed data were consistent with a mismatch distribution expected from a sudden population expansion or a spatial expansion. The mismatch distribution and expected mismatch distribution for a stable population was visualized in the pegas package in R.

Data availability. Short-read sequence data were submitted to the NCBI Short Read Archive (SRA; https://www.ncbi.nlm.nih.gov/sra/), SRA accession: SRP058536 (see also Supplementary Information).

Received: 17 February 2017; Accepted: 18 July 2017; Published online: 28 August 2017

References

1. Steere, A., Coburn, J. & Glickstein, L. The emergence of Lyme disease. J. Clin. Invest. 113, 1093–1101 (2004).
2. Steere, A. C. C., Malawista, S. E., Snydman, D. R. & Andiman, W. A. A cluster of arthritis in children and adults in Lyme, Connecticut. Arthritis Rheumat. 20, 7–17 (1977).
3. Steere, A. C. & Malawista, S. E. Cases of Lyme disease in the United States: locations correlated with distribution of Ixodes dammini. Ann. Intern. Med. 91, 730–733 (1979).
4. Walter, K. S. et al. Invasion of two tick-borne diseases across New England: harnessing human surveillance data to capture underlying ecological invasion processes. Proc. R. Soc. B 283, S301–S327 (2016).
5. How Many People Get Lyme Disease? https://www.cdc.gov/lyme/stats/humancases.html (Centers for Disease Control and Prevention, 2015).
6. Spielman, A. & Wilson, M. Ecology of Ixodes dammini-borne human babesiosis and Lyme disease. Ann. Rev. Enolomol. 30, 439–460 (1985).
7. Wobeser, M. et al. Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. Nature 455, 661–664 (2008).
8. Harris, S. R. et al. Evolution of MRSA during hospital transmission and intercontinental spread. Science 327, 469–474 (2010).

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.
9. Gray, R. R. et al. Testing spatiotemporal hypothesis of bacterial evolution using methicillin-resistant Staphylococcus aureus ST239 genome-wide data within a bayesian framework. Mol. Biol. Evol. 28, 1593–1603 (2011).
10. Biek, R. et al. Whole genome sequencing reveals local transmission patterns of Mycobacterium bovis in sympatric cattle and badger populations. PLoS Pathog. 8, e1003008 (2012).
11. Carpi, G. et al. Whole genome capture of vector-borne pathogens from mixed DNA samples: a case study of Borrelia burgdorferi. BMC Genomics 16, 434 (2015).
12. Xu, G. et al. Detection of heterogeneity of Borrelia burgdorferi in Ixodes ticks by culture-dependent and culture-independent methods. J. Clin. Microbiol. 51, 615–617 (2013).
13. Humphrey, P. T., Caporale, D. A. & Brisson, D. Uncoordinated phylogeography of Borrelia burgdorferi and its tick vector, Ixodes scapularis. Evolution 64, 2653–2667 (2010).
14. Qiu, W.-G., Dykhuijzen, D. E., Acosta, M. S. & Luft, B. J. Geographic uniformity of the Lyme disease spirochete (Borrelia burgdorferi) and its shared history with tick vector (Ixodes scapularis). Proc. Natl Acad. Sci. USA 108, 1309–1318 (2011).
15. Haven, J. et al. Pervasive recombination and sympatric genome diversification driven by frequency-dependent selection in Borrelia burgdorferi, the Lyme disease bacterium. Genetics 189, 951–966 (2011).
16. Mongedon, E. E. F. et al. Inter- and intra-specific pan-genomes of Borrelia burgdorferi sensu lato: genome stability and adaptive radiation. BMC Genomics 14, 693 (2013).
17. Marshall, W. F. et al. Detection of Borrelia burgdorferi DNA in museum specimens of Peromyscus leucopus. J. Infect. Dis. 170, 1027–1032 (1994).
18. Persing, D. H. et al. Detection of Borrelia burgdorferi DNA in museum specimens of Ixodes dammini ticks. Science 249, 1420–1423 (1990).
19. Casjens, S. R. et al. Whole genome sequence of an unusual Borrelia burgdorferi sensu lato isolate. J. Bacteriol. 193, 1489–1490 (2011).
20. Casjens, S. R. et al. Plasmid diversity and phylogenetic consistency in the Lyme disease agent Borrelia burgdorferi. BMC Genomics 18, 165 (2017).
21. Fraser, C. M. et al. Genomic sequence of a Lyme disease spirochaete, Borrelia burgdorferi. Nature 390, 580–586 (1997).
22. Qiu, W.-G. et al. Genetic exchange and plasmid transfers in Borrelia burgdorferi sensu stricto revealed by three-way genome comparisons and multilocus sequence typing. Proc. Natl Acad. Sci. USA 101, 14150–14155 (2004).
23. Schwan, T. G. et al. Detection of Borrelia burgdorferi sensu lato in the tick feeding. Proc. Natl Acad. Sci. USA 92, 2999–2993 (1995).
24. Barbour, A. G. & Travinsky, B. Evolution and distribution of the ospC gene: a transferable genotype determinant of Borrelia burgdorferi. mBio 1, e00153–10 (2010).
25. Wormser, G. P. et al. Borrelia burgdorferi genotypes predict the capacity for hematogenous dissemination during early Lyme disease. J. Infect. Dis. 198, 1358–1364 (2008).
26. Jewett, M. W. et al. GuaA and GuaB are essential for Borrelia burgdorferi survival in the tick-mouse infection cycle. J. Bacteriol. 191, 6231–6241 (2009).
27. Shi, Y., Xu, Q., McShan, K. & Fang, T. L. Both decorin-binding proteins A and B are critical for the overall virulence of Borrelia burgdorferi. Infect. Immun. 76, 1239–1246 (2008).
28. Xu, H. et al. Characterization of the highly regulated antigen BBA05 in the enzootic cycle of Borrelia burgdorferi. Infect. Immun. 78, 1067–1073 (2010).
29. Brodzik, R. N. & D pus, R. S. Lyme disease in California: a novel enzootic transmission cycle of Borrelia burgdorferi. Science 256, 1439–1442 (1992).
30. Gullia-Nuss, M. et al. Genomic insights into the Ixodes scapularis tick vector of Lyme disease. Nat. Commun. 7, 10507 (2016).
31. Khatchikian, C. E. et al. Recent and rapid population growth and range expansion of the Lyme disease tick vector, Ixodes scapularis, in North America. Evolution 69, 1678–1689 (2015).
32. Carpi, G. et al. Babesia microti from humans and ticks hold a genomic signature of strong population structure in the United States. BMC Genomics 17, 888 (2016).
33. Duick-Wasser, M. A. et al. Field and climate-based model for predicting the density of host-seeking nymphal Ixodes scapularis, an important vector of tick-borne disease agents in the eastern United States. Glob. Ecol. Biogeogr. 19, 504–514 (2010).
34. Kugeler, K. J., Farley, G. M., Forrester, J. D. & Mead, P. S. Geographic distribution and extension of human Lyme disease, United States. Emerg. Infect. Dis. 21, 1455–1457 (2015).
35. Qiu, W.-G. et al. Wide distribution of a high-virulence Borrelia burgdorferi clone in Europe and North America. Emerg. Infect. Dis. 14, 1097–1104 (2008).
36. Biek, R., McShan, K., Xu, Q., Gullia, M. & Fang, T. L. Measurably evolving pathogens in the genomic era. Trends Ecol. Evol. 30, 306–313 (2015).
37. Driscoll, A. et al. Invasive vector evolution of Borrelia garinii. Trends Ecol. Evol. 26, 47–55 (2011).
38. Driscoll, A. et al. Invasive vector evolution of Borrelia burgdorferi. Trends Ecol. Evol. 26, 47–55 (2011).
39. LeDuc, M. A. et al. Field and climate-based model for predicting the density of host-seeking nymphal Ixodes scapularis, an important vector of tick-borne disease agents in the eastern United States. Glob. Ecol. Biogeogr. 19, 504–514 (2010).
40. LeDuc, M. A. et al. Field and climate-based model for predicting the density of host-seeking nymphal Ixodes scapularis, an important vector of tick-borne disease agents in the eastern United States. Glob. Ecol. Biogeogr. 19, 504–514 (2010).
68. Baele, G. et al. Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Mol. Biol. Evol.* **29**, 2157–2167 (2012).
69. Librado, P. & Rozas, J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**, 1451–1452 (2009).
70. Rogers, A. R. & Harpending, H. Population growth makes waves in the distribution of pairwise genetic differences. *Mol. Biol. Evol.* **9**, 552–569 (1992).
71. Slatkin, M. & Hudson, R. R. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* **129**, 555–562 (1991).
72. Paradis, E. pegas: an R package for population genetics with an integrated-modular approach. *Bioinformatics* **26**, 419–420 (2010).

**Acknowledgements**

This study was supported by the National Institutes of Health (NIH) Ecology and Evolution of Infectious Disease Program grant R01 GM105246 and by NIH grant R21AI12938. K.S.W. was supported by the NIH Ruth L. Kirschstein National Research Service Award (F31 AI118233-01A1) and the NSF Doctoral Dissertation Improvement Grant (DEB-1401143). G.C. was supported by the Gaylord Donnelley Postdoctoral Environmental Fellowship (the Yale Institute for Biospheric Studies). We thank N. Ogden, J. Brinkerhoff, S. Paskewitz, D. Fish and S. Bent for providing tick samples; P. Flynn and J. Underwood for laboratory work; C. Ben Mamoun and P. Krause for discussions; and R. Bjornson and the Yale High Performance Computing Center for computational support.

**Author contributions**

K.S.W., M.A.D.-W., A.C. and G.C. conceived of and designed the experiments. K.S.W. performed the experiments and analysed the data. M.A.D.-W., A.C., G.C. and K.S.W. contributed reagents/materials/analysis tools. K.S.W., M.A.D.-W. and A.C. wrote the manuscript.

**Competing interests**

The authors declare no competing financial interests.

**Additional information**

Supplementary information is available for this paper at doi:10.1038/s41559-017-0282-8.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to K.S.W.

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.