

SCIENTIFIC REPORTS



OPEN

Whole genome sequencing and phylogenetic analysis of strains of the agent of Lyme disease *Borrelia burgdorferi* from Canadian emergence zones

Shaun Tyler¹, Shari Tyson¹, Antonia Dibernardo², Michael Drebot², Edward J. Feil³, Morag Graham¹, Natalie C. Knox¹, L. Robbin Lindsay², Gabriele Margos^{4,5}, Samir Mechai⁶, Gary Van Domselaar¹, Harry A. Thorpe³ & Nick H. Ogden⁶

Lyme disease is emerging in southern Canada due to range expansion of the tick vector, followed by invasion of the agent of Lyme disease *Borrelia burgdorferi* sensu stricto. Strain diversity, as determined by Multi Locus Sequence Typing, occurs in this zone of emergence, and this may have its origins in adaptation to ecological niches, and have phenotypic consequences for pathogenicity and serological test performance. Sixty-four unique strains were cultured from ticks collected in southern Canada and the genomes sequenced using the Illumina MiSeq platform. A maximum likelihood phylogenetic tree of the chromosome revealed two large clades with multiple subclades. Consistent with previous studies on this species, the clades were not geographically defined, and some Canadian strains were highly divergent from previously sequenced US strains. There was evidence for recombination in the chromosome but this did not affect the phylogeny. Analysis of chromosomal genes indicated that these are under intense purifying selection. Phylogenies of the accessory genome and chromosome were congruent. Therefore strain differences identified in the phylogeny of chromosomal genes likely act as a proxy for genetic determinants of phenotypic differences amongst strains that are harboured in the accessory genome. Further studies on health implications of strain diversity are needed.

Lyme disease is a tick-borne zoonosis that in North America is caused by the bacterial pathogen *Borrelia* (or *Borrelia*) *burgdorferi* sensu stricto (henceforth shortened to *B. burgdorferi*). This bacterium is maintained in nature by wild animal reservoirs that live in woodlands¹, and is emerging in eastern and central Canada due to northwards spread of the geographic range of the tick vector *Ixodes scapularis*². The tick vector *Ixodes scapularis* and *B. burgdorferi* are species invading natural sylvatic communities in southern Canada, and the evolutionary history of *B. burgdorferi* associated with past expansions/invasions appears linked to phenotypes of public health importance³; both pathogenicity⁴ and detectability of infections by sero-diagnostic tests^{5,6} may vary with infecting strain. Whole genome sequencing (WGS) of bacterial pathogens has been successfully deployed in a public health context for high resolution genotyping, phylogeography, detecting emergent high-risk clones and source attribution^{7,8}. The data can also be used to address fundamental evolutionary and ecological questions^{9–12}. WGS of *B. burgdorferi* is set to also shed light on the public health risk in zones of emergence such as southern Canada

¹Genomics Core Facility, National Microbiology Laboratory, Public Health Agency of Canada, Canadian Science Centre for Human and Animal Health, 1015, Arlington St., Winnipeg, Manitoba, Canada. ²Zoonotic Diseases and Special Pathogens Division, National Microbiology Laboratory, Public Health Agency of Canada, Winnipeg, Manitoba, Canada. ³Department of Biology and Biochemistry, University of Bath, Claverton Down, Bath, United Kingdom. ⁴Ludwig Maximilians Universität München, Department for Infectious Diseases and Zoonoses, Munich, Germany. ⁵National Reference Centre for Borrelia, Oberschleissheim and Bavarian Health and Food Safety Authority, Oberschleissheim, Germany. ⁶Public Health Risk Sciences Division, National Microbiology Laboratory, Public Health Agency of Canada, Saint-Hyacinthe, Québec, J2S 2M2, Canada. Correspondence and requests for materials should be addressed to N.H.O. (email: nicholas.ogden@canada.ca)

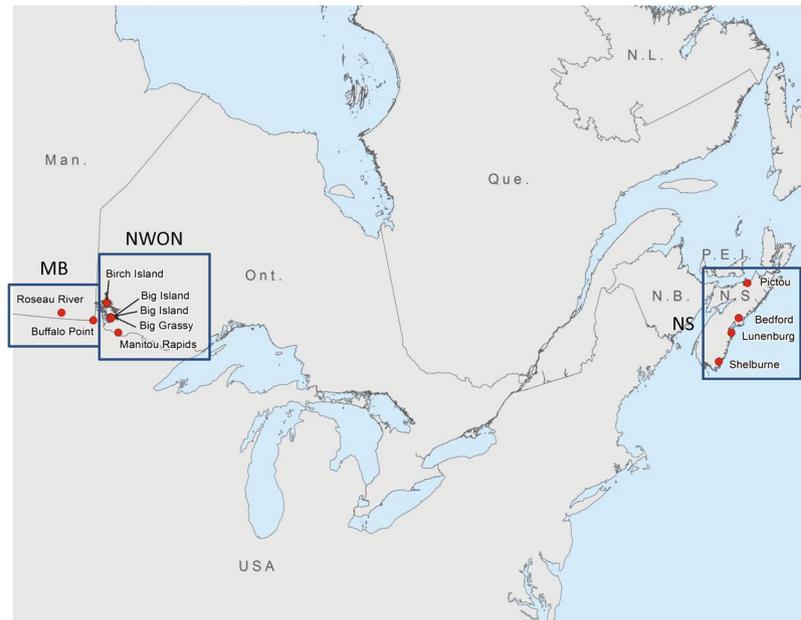


Figure 1. Location of sites from which host-seeking *I. scapularis* ticks were collected to obtain *B. burgdorferi* cultures. The circles indicate the groups of sample sites in southern Manitoba (MB), northwest Ontario (NWO), and Nova Scotia (NS). The Canadian provinces identified on the map are Man. = Manitoba, Ont. = Ontario, Que. = Quebec, N.L. = Newfoundland and Labrador, P.E.I. = Prince Edward Island, N.B. = New Brunswick, N.S. = Nova Scotia. The map was created in ArcGIS Version 10.5 (ESRI, Redlands, CA: <https://support.esri.com/en/products/desktop/arcgis-desktop/arcmap/10-5>).

by allowing us to identify sources and predict invasion patterns, and better understand the evolutionary ecology of *B. burgdorferi* and its associations with phenotype variations of ecological and public health significance¹³.

In recent years it has become recognised that *B. burgdorferi* is genetically diverse and a range of typing methods have been employed. Currently three main methods of typing are used because of their utility in identifying phenotypic traits such as pathogenicity and/or ecological, geographical and evolutionary patterns¹⁴. These are sequences of the 16S-23S *rrs-rrlA* intergenic spacer (IGS), *ospC* sequences, and multilocus sequence typing/multilocus sequence analysis (MLST/MLSA)¹⁵. IGS and *ospC* sequences are routinely clustered for strain typing using (i) Ribosomal Sequence Types (RSTs) or ribosomal spacer identification numbers (RSPs) for IGS^{16,17}, and (ii) *ospC* major groups for *ospC*¹⁸. The diversity at *ospC* is believed to be maintained by balancing selection¹⁸ or niche specialisation¹⁹. Due to the focus on several housekeeping genes, phylogenetic trees constructed using the MLST sequences are likely to reflect the evolutionary history of *B. burgdorferi*, although there is considerable linkage disequilibrium in the *B. burgdorferi* genome and phylogenies constructed using IGS trees can show broadly similar patterns to those obtained by MLST trees. For *ospC*, however, the phylogenetic branching pattern was different from that of MLST genes suggesting differences in the evolutionary histories of these loci¹⁵.

MLST has provided data consistent with a south-to-north invasion of *B. burgdorferi* in Canada, in that many of genotypes detected in southern Canada are identical to those observed immediately to the south in northern USA²⁰. However, in some parts of Canada (particularly southern Manitoba, north-west Ontario and the Maritime provinces) there are many diverse strains of *B. burgdorferi* in ticks and animal reservoir hosts, and some strains have to date only been found in Canada²⁰⁻²². The possible ecological origins of this variation in strain diversity, and its consequences with respect to public health risk, specifically pathogenicity and serological test performance, are only just starting to be explored^{3,23}.

Here we describe development and analysis of WGS data for 64 unique strains of *B. burgdorferi* collected in Canada in 2016. Phylogenetic analysis of strains from Canada, the US and Europe using WGS was compared against analyses obtained from sequences that are frequently currently used for strain typing and phylogenetic analysis of this species. Evidence for recombination, and its potential effect on strain phylogeny was explored. The strength and direction of selection across the chromosome was investigated to see if chromosomal genes could drive differences amongst strains in phenotypes of importance for public health (e.g., pathogenicity), that are reflected in the phylogeny. The sequence data we obtained will also provide a baseline resource for future studies of the evolutionary history of *B. burgdorferi*, and simultaneously allow exploration of what the observed diversity may mean for inter-strain phenotypic variations that are of public health significance.

Results

Samples collected and success of culture. A total of 483 host-seeking blacklegged ticks were collected from 10 sites in Canada during October and November 2016 (Fig. 1), comprising two sites in Manitoba (Buffalo Point [n = 78] and Roseau River [70], MB), 4 sites in northwest Ontario (Big Grassy [16], Big Island [11], Manitou Rapids [3], and Birch Island [12], ON), and 4 sites in Nova Scotia (in the counties of Bedford [60],

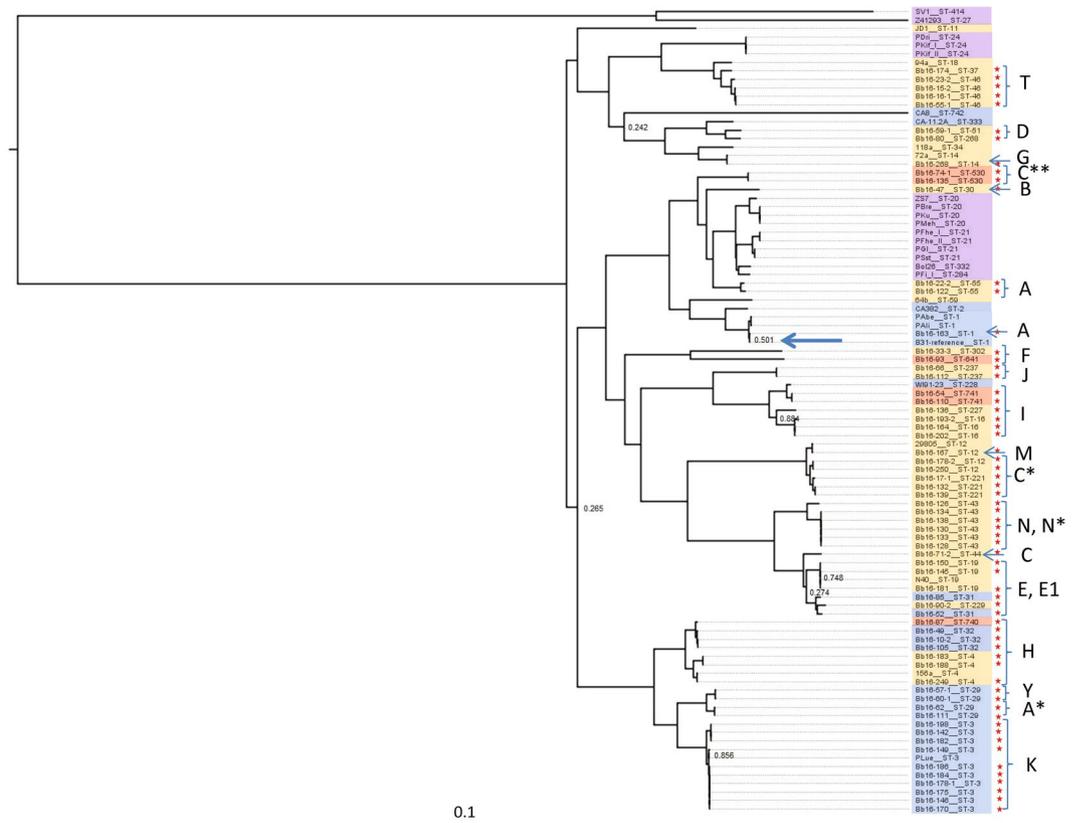


Figure 2. A midpoint rooted phylogenetic tree of the *B. burgdorferi* chromosome generated via the SNVPhyl pipeline. All sequences obtained in this study and currently available published full sequences from samples collected elsewhere in North America and Europe were used. Colour coding indicates the geographic occurrence of the MLST types of the sequences: blue = USA only; red = Canada only, yellow = USA and Canada; violet = Europe. Red stars to the right of the tree indicate sequences obtain from Canadian sites in this study. Letters to the right of the tree indicate the *ospC* major groups of the strains. Letters with asterisks indicate novel *ospC* major group sequences as described in the text. The position of the reference strain, B31 is shown by a blue arrow. Likelihood values for branches with a value less than 0.9 are shown.

Lunenburg [113], Pictou [34] and Shelburne [86], NS). From these samples, 64 unique single-strain cultures were obtained in the laboratory: 26 cultures from MB, 13 from ON, and 25 from NS (sample details are presented in Table S1). No cultures from solid media (the first step in the isolation procedure) contained more than one strain.

Whole genomes sequenced. The chromosomes for all strains assembled as single contigs and were found to be collinear and between approximately 899 kb and 912 kb with a plasmid content ranging between 335 kb and 475 kb (Table S1). As previously reported²⁴ the primary difference in chromosomal arrangement occurred at the 3' (right) end. It is likely that the well documented repeat region in the *lmp1* gene on the chromosome²⁵ contains inaccuracies. PCR performed using primers flanking this repeat region produced amplicons of variable size which were consistently larger than that predicted from the assemblies indicating that the repeat was likely collapsed during the assembly process. Unfortunately we were unable to resolve the exact number of repeats after conventional ABI sequencing. Estimates of PCR amplicon size ranged from approximately 950 bp–1600 bp corresponding to 4–8 copies. Because sequence data for all isolates assembled into a single, large, chromosomal contig; the remaining assembled contigs were considered to be plasmid content.

Using the paired end data, the strains were typed using the standard *Borrelia* MLST scheme. This revealed 27 different MLST Sequence Types (STs). There were 15 STs from MB, 7 STs from ON, 8 STs from NS. Of these, 7/20 (35%) STs from MB and NW ON are Canada-specific STs new to this study or found during recent studies^{21,22} (Table S1).

Phylogenetic analysis. The maximum likelihood phylogenetic tree (Fig. 2), constructed with the Canadian strains, and published sequences from Europe and the US (Table S2) comprises two well-supported major clades, within which are well-supported subclades. There is relatively high diversity of strains occurring in Canada, with sequences in almost all major and minor clades, as well as strains that cluster with the reference strain (B31). The clades and subclades in the phylogenetic tree are not defined geographically, with strains from Canada and the US in most. Two clades of European strains nested within clades containing strains from North America. Some Canada-specific strains are very similar to other known strains but some are quite divergent (Fig. 2). A

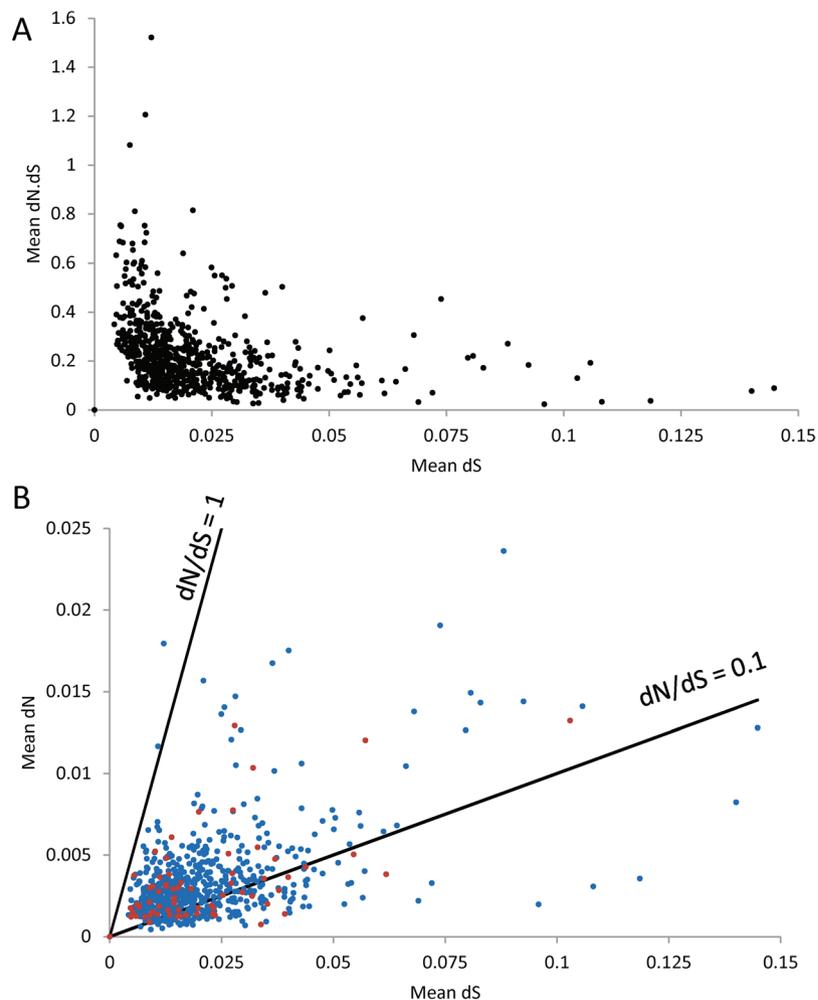


Figure 3. Output of assessment of strength and direction of selection on chromosomal genes. **(A)** A plot of pairwise dN/dS versus divergence time (dS) over all chromosomal genes. **(B)** A plot of dN versus dS for all chromosomal genes. Those genes encoding surface exposed proteins, or other genes which are putatively immunogenic (see Table S2) are shown in blue, all other genes are shown in red. The lines represent two interpretations of the neutral expectation ($dN/dS = 1$ and 0.1). The single gene with $dN/dS > 1$ is *rpsH* which encodes a ribosomal protein hence would be expected to be under strong stabilising selection.

similar phylogenetic tree was obtained using some of the sequences from Walter *et al.*²⁶, although some of the new sequences from this study did cluster with the divergent Canadian strains (Fig. S1).

Strains of the same MLST ST always clustered together in the whole chromosome tree (Fig. S2) and the tree constructed only using the MLST sequences formed the same major clades as the whole chromosome tree with some minor differences (Fig. S3). Analysis of outer surface protein C (*ospC*) sequences revealed that the Canadian strains collectively carried 14 previously described *ospC* major groups (A, B, C, D, E, E1, F, G, H, I, J, K, M, N). In addition, 11 strains carried *ospC* sequences that had similarities to existing major group sequences that were too low (<90%) to be classified as belonging to an existing *ospC* major group. These sequences have been temporarily named according to the closest major group as A*, C*, C** and N* (Figs 2 and S4). The strains carried IGS sequences determined as belonging to 12 IGS genotypes: 1A, 2A, 2D, 4, 4A, 5, 6A, 6B, 7A, 7B, 8C and 9 (Fig. S5). The clade structures in the trees of the *ospC* and IGS sequences had a greater number of differences than the MLST tree when compared to the whole chromosome tree (Figs S4 and S5).

Evidence of strong purifying selection on the chromosomal genes. In order to examine the strength and direction of selection on the chromosomal genes we calculated dN and dS for all genes in turn using the method of Nei & Gojobori²⁷. Non-synonymous changes are on average more likely to be deleterious than synonymous changes, hence the dN/dS ratio provides evidence as to the strength of purifying selection (where $dN/dS < 1$) or positive selection ($dN/dS > 1$). We first examined the trajectory of dN/dS over divergence time (dS). Purifying selection operating over time will lower dN/dS at a rate determined by the selection co-efficients of the non-synonymous changes (*s*) and the effective population size (N_e)²⁸. Figure 3a shows pairwise dN/dS against dS over all chromosomal genes. Although there is weak evidence of a slight increase in dN/dS over very short

divergence times (low dS), in general dN/dS is very low (~0.1) regardless of dS. This points to very strong purifying (stabilising) selection acting on the chromosomal genes as a whole.

We then considered whether we could detect any single genes for which there was evidence of positive rather than purifying selection by plotting, for each gene, dN/dS (Fig. 3b), and comparing these ratios with the neutral expectation (dN/dS = 1). We separately considered all those genes listed in Table S3 which are surface exposed or else putatively immunogenic. These genes might be expected to be under positive selection, in which case we might expect dN/dS > 1. Figure 3b reveals that for all genes, including those encoding surface exposed proteins (shown in blue) fall well below the neutral expectation (dN/dS = 1), which is consistent with strong purifying selection across the chromosome. The only possible exception is the *rpsH* gene (locus tag BB0493), which encodes the 30S ribosomal protein S8, which has an overall dN/dS > 1. As there is no obvious reason why this gene should be subject to positive selection, this analysis again points to very strong stabilising selection acting on the chromosomal genes. To support this further, we note that there are only 11 nonsense mutations in the data, out of a total of 13,507 SNPs (synonymous = 8463, non-synonymous = 4193, intergenic = 903). Thus only 0.081% of the total SNPs potentially resulted in a loss of gene function, which is approximately half the proportion we observed in comparable datasets for *S. aureus* (0.19%) and *E. coli* (0.19%). This suggests that a high proportion of the genes retained on the chromosome are essential, which again is consistent with strong purifying selection. It has been suggested that, for within-species inter-strain comparisons, a threshold dN/dS value of 0.1 rather than 1 may be more appropriate for identifying genes under positive selection²⁹. There were 574 genes with dN/dS values above this threshold, but the proportion of genes likely encoding surface-exposed proteins above the threshold (53/74, 71.6%) was not significantly different from that for genes that likely encode non-surface-exposed proteins (521/674, 77.3%, $\chi^2 = 1.2$, degrees of freedom = 1, $P > 0.1$).

Evidence for recombination in the sequences. Recombination analysis using Gubbins³⁰ identified potential regions of recombination occurring throughout chromosome, with 61% of chromosomal genes being implicated in recombination in at least one strain (Fig. 4). The PhiTest implemented in SplitsTree4 (version 4.1.3.1)³¹ detected statistically significant evidence for recombination for the chromosome, whether or not *B. bissetiae* strain DN127 was used as outgroup (Table 1). There was a mean 3.58 recombination events per branch and the mean ratio of SNPs introduced by recombination to those introduced by point mutations was 1.12. The regions of recombination along the chromosome are similar amongst strain of the same clade but more different in strains in different clades (Figs 4 and S6). Excluding regions of recombination identified by Gubbins from the phylogenies did not significantly alter the phylogenetic relationship among the strains as determined by the core genome SNP analysis (Fig. 5). Similarly, strains that clustered in the phylogenetic tree also clustered in the NeighbourNet network analysis (Fig. 6).

A highly variable accessory genome. The accessory genome was found to be highly variable among the isolates. Of the 1042 gene clusters identified by Roary³² only 127 were found to be present in greater than 95% of the isolates, 446 of the clusters fell between 15% and 95% of the isolates and 469 were present in 15% or less (Fig. 7). However, the clustering of the strains based on the presence or absence of these genes was consistent with that found via SNP analysis of the chromosome, with the exception of a difference in the position in the tree of one clade (Fig. 8).

Discussion

At the time of writing, in this study we have more than doubled the currently available high quality whole genome sequences of *B. burgdorferi*. The data provided by these sequences will provide a resource for study of phylogenetics and functional genomics of this bacterium (e.g. mechanisms of host associations, strain-specific pathogenicity, strain-specific detectability by diagnostic tests) in the future. The high yield of single-strain cultures using the semi-solid phase plating technique demonstrated its utility for obtaining cultured strains from field-collected samples for WGS.

The sequences obtained were concordant in structure to those previously sequenced (Table S2) with sequences being assembled into a single chromosomal contig, a large number of sequences forming the accessory genome, greater inter-strain variation at the 3' end of the chromosome³³ and variation in length associated with a variable-length repeat region²⁵.

The phylogenetic tree obtained from the whole chromosome is consistent with previous trees made using MLST^{15,22} in that there is a strong clade structure but this is not based on the geographic origin of strains (i.e. isolation by distance). However, the WGS provided trees of finer precision in identifying strain differences, by having better support of internal branches, than any method used to date. It should be noted that one clade of strains (comprising strains PALi, PAbe und PLue) that was found in European patients resemble North American strains. These strains have never been found in ticks in Europe and it has been suggested that the affected people acquired their infection in North America³⁴. There was evidence for recombination in the chromosome (which occurred at rates similar to those in other studies)²⁶ but this had only very slight impact on the structure of the phylogenetic tree, which may be due to recombination events occurring in similar regions within-clades, but in more different regions between clades. This suggests that both point mutations and recombination events have occurred contemporaneously in broad terms during the population processes that drove development of the clades.

Some of these Canada-specific strains are very similar to strains occurring in the US. However some Canada-specific strains (e.g. 96, 74-1 and 135: Fig. 2) are quite divergent from the strains occurring in the US for which full sequences are available. Strains 74-1 and 135 also carried a novel *ospC* allele (Fig. 2). We have in the past suggested that strains novel to Canada are not from refugees in Canada, but represent strains endemic to the northern US that have yet to be discovered there³. Consistent with this, some strains from ticks in the northern US identified by Walter *et al.*²⁶ clustered with the divergent Canadian strains (Fig. S1). The WGS supports previous studies using

	With <i>B. bissettiae</i> as outgroup	Without outgroup
Informative sites	23953	22542
Window	100 with k as 38	100 with k as 69
Mean	0.243	0.170
Variance	1.338×10^{-7}	5.741×10^{-8}
Observed	0.203	0.144
<i>P</i>	<0.001	<0.001

Table 1. Results of the PhiTest for evidence for recombination for the chromosome using SplitsTree.

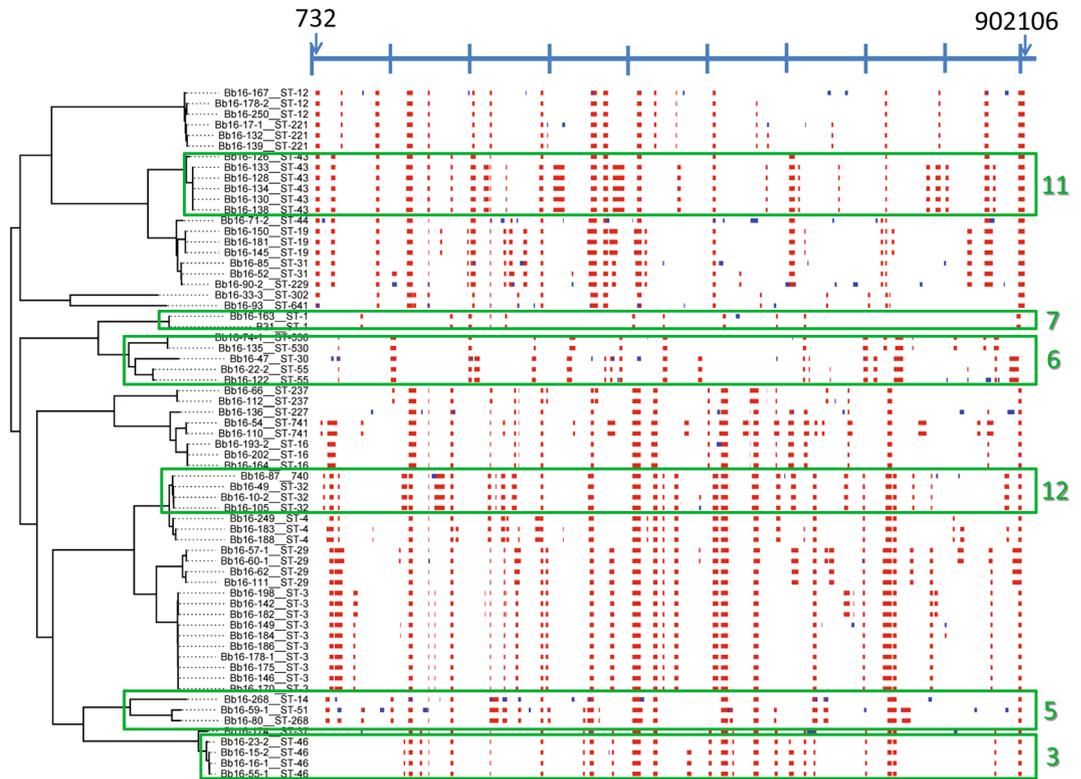


Figure 4. Output from Gubbins analysis showing a map of possible regions of recombination. The scale bar at the top of the map shows the length of the chromosome (with tick marks every 100 Kb) and arrows indicate the first and last position where recombination events were detected. The analysis included all sequences obtained in this study. Red regions are those that are common to multiple strains, the blue regions are those found only in one strain, and white areas indicate where recombination was not detected. Green boxes illustrate some of the clades that had patterns of recombination loci that were similar within-clade but different to adjacent clades. The numbers refer to the numbering of clades in the phylogenetic tree in Fig. 6.

MLST in terms of (i) the topology of the phylogenetic tree, which was similar to the tree generated using MLST sequences alone, and in which strains of the same MLST type always clustered together, and (ii) in the identification of strains (and MLST types) that are (to date) unique to Canada. The study also found some of the Canada-specific MLST STs that we identified in recent studies^{20–22} confirming the reproducibility of this work. The ecological and evolutionary origins of the tree structure continue to be intriguing and will continue to be a focus for research.

Of the 64 unique strains, 22 carried *ospC* alleles of major groups considered as being particularly associated with disseminated Lyme disease (A, B, I, K and N: reviewed in³⁵). To what extent the other *ospC* sequences, and the new *ospC* sequences are associated or not with disseminated disease, and where they occur, remains to be studied in Canada.

There was evidence for very strong stabilising selection acting on chromosomal genes, which suggests that the vast majority of adaptive changes occur within the much more dynamic accessory genome carried on the plasmids. The distinction between the core and accessory genomes in *B. burgdorferi* is therefore particularly striking as these two components are physically unlinked. In this way the bacteria can maintain essential genes on the chromosome, which are subject to strong purifying selection and only low rates of recombination, whilst ecological adaptations are made possible largely through the much more “sexual” accessory genome borne on the plasmids. This suggests that associations between strain type and phenotypic characteristics (interactions with different host species²², pathogenicity⁴, and possibly the detectability of antibody responses to some surface

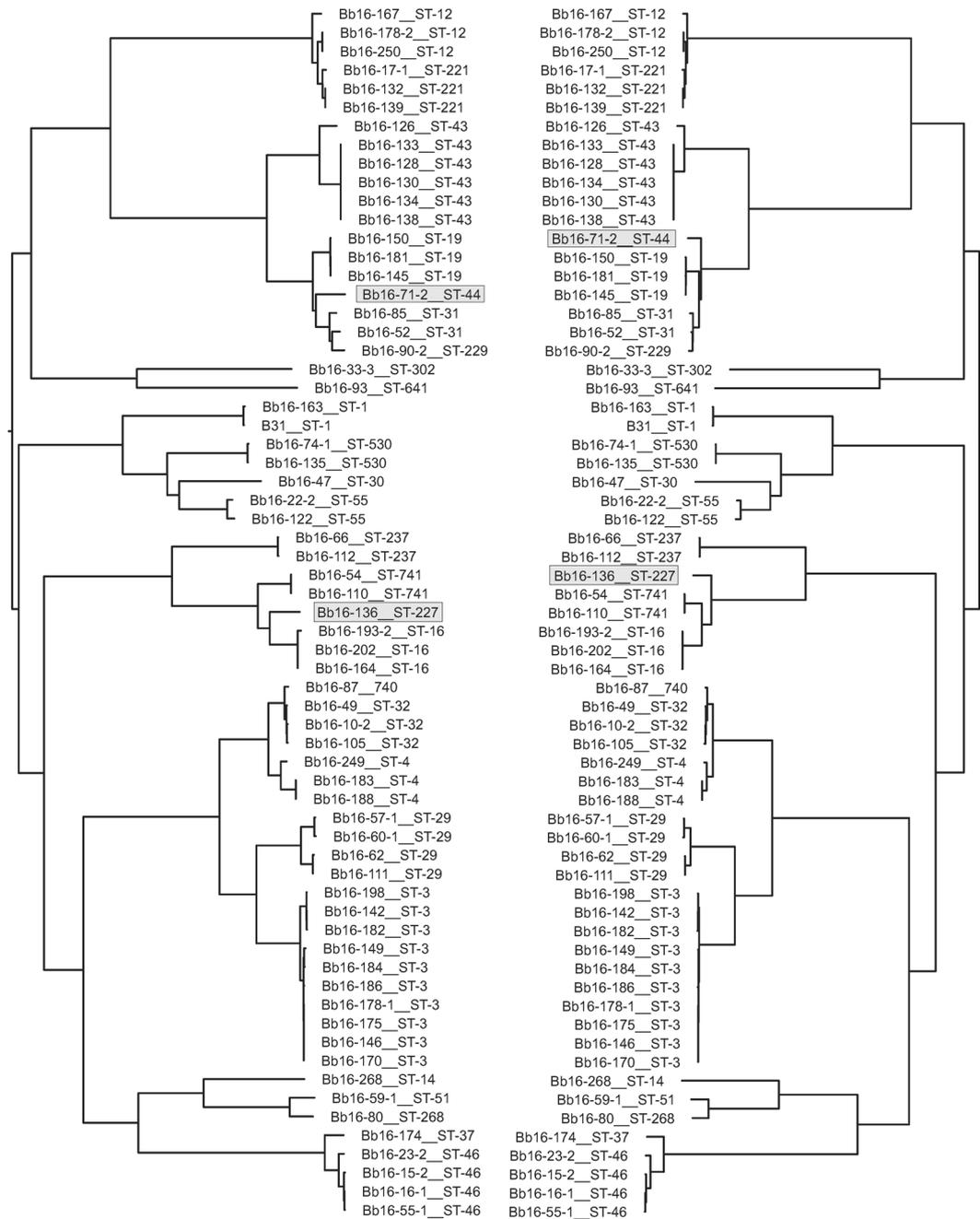


Figure 5. A comparison of phylogenies with and without potential recombinant regions. Phylogenetic tree of the *B. burgdorferi* ss chromosome generated via the SNVPhyl pipeline is shown on the left and a tree constructed using Gubbins after exclusion of potential recombination regions is shown for comparison on the right.

proteins by serological tests)⁵ are mostly driven by genetic variations in the accessory genome rather than by those in the chromosome. The phylogenies of the accessory/plasmid content and the chromosome were consistent (as in other recent studies)³⁶ so strain differences identified in the phylogeny of chromosomal genes likely act as a proxy for the genetic determinants of phenotypic differences that are mostly in the accessory genome. This is consistent with our previous conclusions regarding associations of MLST types with phenotypic differences^{4,22}. However, the plastic nature of plasmid content amongst the strains means that even strains that are phylogenetically closely related could harbor differences in gene content that have important impacts on the phenotypes.

The analysis of the accessory genome is consistent with previous studies showing that plasmids within the genospecies *B. burgdorferi* show a high degree of diversity^{36,37}. Although some linear plasmids can be assembled from short reads only, long-read methods such as Pacific Bioscience SMRT technology or Nanopore Sequencing technology are likely to be required to explore the plasmid content of *Borrelia* strains in detail³⁸.

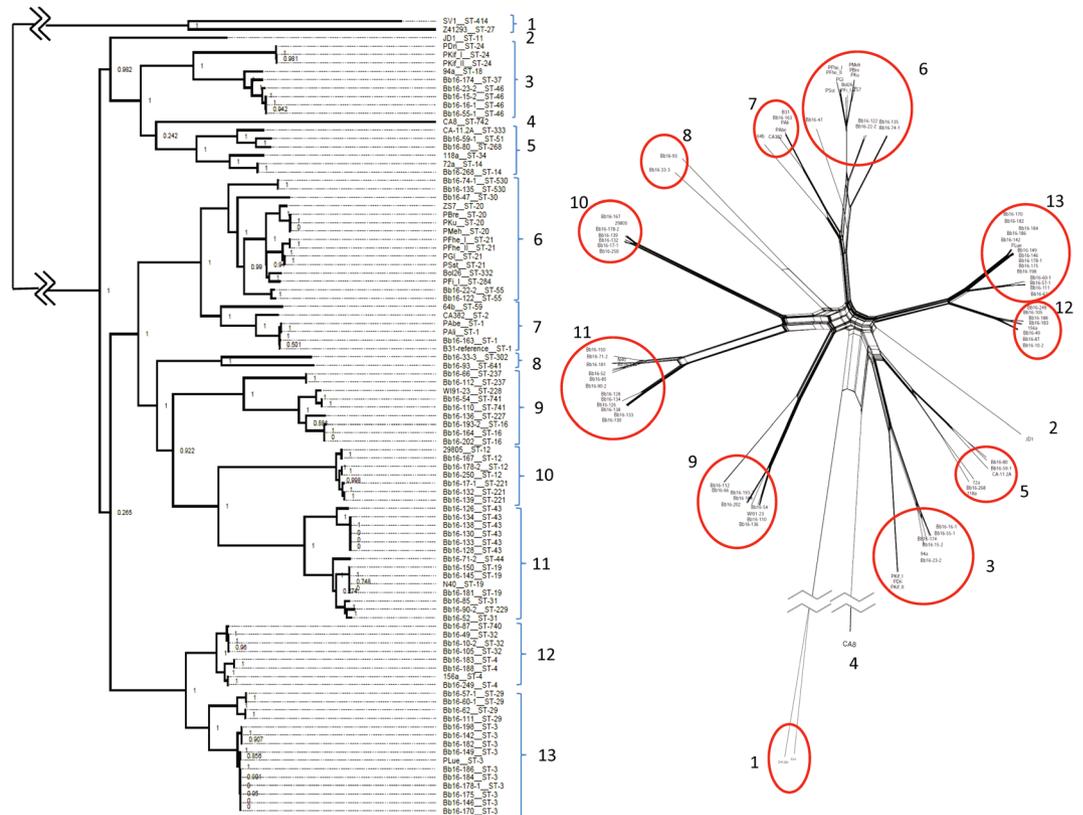


Figure 6. A comparison of clades identified by maximum likelihood via the SNVPhyl pipeline and by network analysis. The orientation of clades and divergent single strains are identified by brackets in the maximum likelihood phylogenetic tree (left) and clusters are identified by red circles in the network analysis (right). The numbers identify the clades and corresponding clusters.

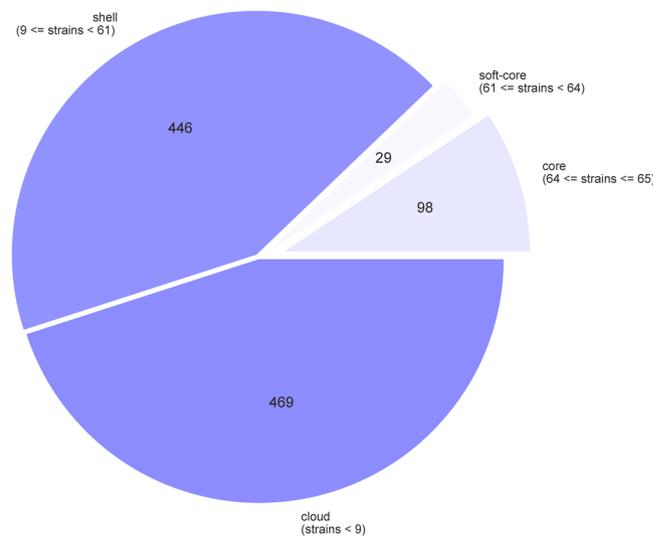


Figure 7. Pie chart of the breakdown of plasmid genes and the number of isolates in which they are present.

In this study we have sequenced the genome of 64 strains of *B. burgdorferi* s.s. adding considerably to our data on, and knowledge of, this bacterium. These data will help us explore questions regarding pathogenicity, ecology and epidemiology that are relevant for public health responses to the increasing threat from Lyme disease.

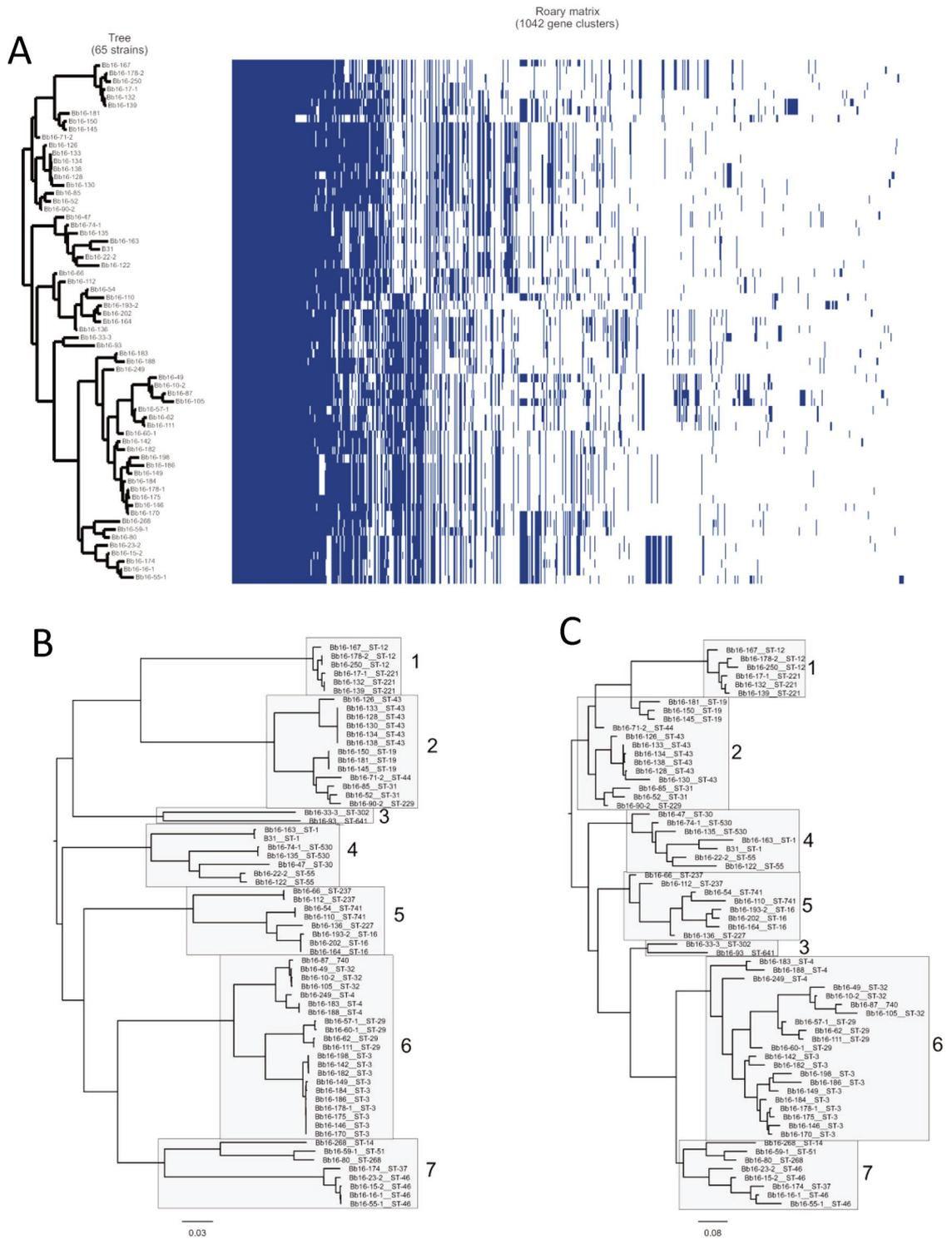


Figure 8. Similarity in phylogeny generated from the chromosome and the accessory genome. (A) Phylogenetic relationship of plasmid genome as determined by Roary compared to a matrix with the presence (blue blocks) and absence (grey areas) of core and accessory genes found in the pangenome. Correlation between clades as determined by (B) SNP analysis of the chromosome and (C) pangenome analysis of the plasmid genome is illustrated. The relationship among the strains is consistent between the two methods although the absolute placement of the clade denoted as #3 differs.

Materials and Methods

Collection and selection of samples for culture. Samples for the study comprised host-seeking ticks collected by drag sampling³⁹ from areas of *I. scapularis* and *B. burgdorferi* invasion in south central Canada

(southeastern Manitoba [MB] and northwestern Ontario [ON]) and the Maritimes (Nova Scotia [NS]) (Fig. 1). Ticks collected from the environment were kept alive at room temperature by placing 15 ml vials containing ticks in plastic storage bags provisioned with moistened paper towel to maintain humidity over 90%. Ticks were transported by surface transport (from Manitoba and northwestern Ontario) or courier (from Nova Scotia) to the National Microbiology Laboratory in Winnipeg. At the laboratory, ticks were surface decontaminated by immersion in the following sequence of solutions prepared with sterile nuclease-free water; 0.5% bleach containing one drop of Tween 80 per 10 ml, 0.5% benzalkonium chloride, 3% hydrogen peroxide, and 70% ethanol, and then rinsed in nuclease free water. Ticks were disinfected individually or in pools of up to ten specimens if collected from the same site. Individual ticks were transferred to sterile 1.5 ml tubes containing Barbour-Stoenner-Kelly (BSK) medium prepared in-house and supplemented with rifampicin, kanamycin, phosphomycin, amphotericin B, and heat-inactivated rabbit serum (Millipore Sigma, Oakville, Canada). Ticks were then cut into small pieces using a sterile scalpel and a 100 µl aliquot of the tick macerate was removed for DNA extraction using a Qiagen DNeasy 96 kit. Extracted DNA was tested by real-time PCR targeting the 23S rRNA gene of *B. burgdorferi* s.l.⁴⁰. The remaining material was held at 4 °C until PCR was complete and samples that were positive for *B. burgdorferi* by PCR were cultured in semi-solid agar.

Semi-solid phase plating for recovery of single-strain colonies. In our experience, approximately 12% of ticks from collection sites in Canada show evidence of mixed-strain infections²². In order to obtain clonal populations of single strains of *B. burgdorferi*, tick homogenates were cultured in semi-solid agar. Undiluted, 10^{-1} and 10^{-2} dilutions of tick homogenates were added to a solution of BSK and 0.7% agarose held at 42 °C. The mixtures were then poured into plastic petri plates and allowed to solidify. Plates were incubated at 37 °C with 5% CO₂ for up to 28 days. Two detectable colonies per tick were excised from the agar and seeded into separate vials containing liquid BSK-H (Millipore Sigma, Oakville, Canada) for expansion of the culture at 35 °C for 7 days. Once cultures were in exponential phase of growth, an aliquot was removed for DNA extraction for genome sequencing. The remaining sample was centrifuged and resuspended in cBSK supplemented with 10% glycerol for cryopreservation in liquid nitrogen.

Genome sequencing and analysis. Multiplexed libraries were created with TruSeq sample preparation kits (Illumina, San Diego, CA) and paired-end, 300 bp, indexed reads were generated on the Illumina MiSeq platform (Illumina, San Diego, CA). Genome assembly was performed using SPAdes v3.9^{40,41} with a minimum contig size cut off of 1000 bp.

For MLST determination, paired end reads for each isolate were directly compared to the eight housekeeping loci (*clpA*, *clpX*, *nifS*, *pepX*, *pyrG*, *recG*, *rplB*, and *uvrA*) for other *B. burgdorferi* sensu stricto isolates in the MLST database at the University of Oxford (<http://www.pubmlst.org>) using SRST2 v.0.3.6. Novel MLST types found in this study were submitted to the database curators for verification, classification, and inclusion in the database.

Phylogenetic analyses. A maximum likelihood phylogenetic tree of the bacterial chromosome was created with SNVPhyl, a whole genome SNV phylogenomics pipeline⁴² using B31 (Table S2) as the reference sequence, which was originally isolated from a tick in northeastern US^{43,44}. The analysis was performed with all strains previously described³³ from Europe and the USA excluding two European strains (PHas and PMi), which were omitted due to the low coverage present in the SRA data. In cases where the raw reads were unavailable these were simulated based on the WGS scaffolds using ART⁴⁵.

Recently a large number of sequences of the *B. burgdorferi* genome obtained from ticks collected in North America have been published²⁶. However these sequences were obtained by a hybrid capture method that has some limitations in terms of sequence quality that make these sequences incompatible for analysis with those generated here. These include low coverage in approximately 10% of strains, absence of sequences that are not present in the reference strain B31 (which is used to assemble the reads), an unknown number of strains that have chimeric sequences that are obtained because the method obtains reads from all strains in mixed infections (which were found in about 50% of the ticks), as well as the possibility of chimeras that include sequences of genes of other bacteria species present in the ticks, where these are conserved across bacterial species^{26,46}. Nevertheless, we selected sequences from this study that we consider the least ambiguous (as described in supplementary information), to include in a maximum likelihood phylogenetic tree (also created with SNVPhyl) with our own sequences and previously published sequences to see if their inclusion markedly alters the phylogeny we obtained without them.

A Bayesian phylogenetic tree using the MLST sequences¹⁵ of the strains was constructed using MrBayes v3.2.1⁴⁷ to compare the tree topology obtained against the tree generated from the whole chromosome. Markov chain Monte Carlo (MCMC) samplings were run for 500,000 generations with trees being sampled every 1,000th generation, and convergence diagnostics conducted as previously described^{47–49}.

Because *ospC* – coding sequences (which occur on the circular cp26 plasmid), and the chromosomal 16S–23S *rrs-rrlA* IGS sequences are frequently used in strain typing of *B. burgdorferi*, these sequences were obtained and analysed as follows. To obtain the *ospC* sequences for each strain, paired end reads were mapped to the B31 reference strain including plasmid sequences. The consensus for the *ospC* gene from plasmid cp26 (nucleotides 16903 through 17535 as per AE000792.1 for a total of 632 bp), obtained as described below, was pulled out for each isolate. The sequences were aligned using ClustalW2 (v2.1, EMBL-EBI, Hinxton, Cambridgeshire, UK) and typed to *ospC* major groups using the similarity criteria of Qiu *et al.*¹⁸. The IGS region for each isolate was extracted in an identical manner as that for *ospC* based on annotation from the B31 reference chromosomal sequence. Sequences were aligned, trimmed, and genotyped according to the criteria of Bunikis *et al.*⁵⁰. Bayesian phylogenetic trees of the *ospC* and IGS sequences were also constructed using MrBayes v3.2.1 for comparison against the tree generated from the whole chromosome of the strains.

The strength and direction of selection on the chromosomal genes. We examined the degree to which the purifying and positive selection is operating on chromosomal encoded genes by plotting dN (the number of non-synonymous SNPs per non-synonymous site) against dS (the number of synonymous SNPs per synonymous site). dN and dS were calculated using the method of Nei and Gojobori²⁷. In order to detect whether potentially immunogenic genes were under unusually strong positive selection, we separately considered surface expressed proteins, and/or proteins included in the two-tier serological test for Lyme disease³¹, and compared these against all other chromosomal genes. The proteins are listed in supplementary information Table S3.

As dN/dS has also been shown to decrease with increasing divergence (dS) in bacteria, owing to ongoing purifying selection of slightly deleterious non-synonymous changes²⁸, we also plotted pairwise dN/dS against dS using averages of all chromosomal genes.

Assessment of the degree of recombination in the sequences. An alignment was generated by incorporating the variable sites identified in each strain via the SNVPhyl analysis using the B31 sequence as the reference sequence using in-house scripts. This was then used to assess the effect of recombination on the phylogenies using Gubbins. The PhiTest implemented in Split Tree 4 software was also used to evaluate phylogenetic networks based on the alignment. Multiple alignments of strains (from this and previous studies, Table S1) in Phylip were uploaded and a network analysis (NeighborNet) was conducted in Splitstree⁴³ using standard settings. The Phi Test was conducted using *B. bissettiae* strain DN127 as outgroup or with only *B. burgdorferi* strains.

Analysis of the accessory genome. The plasmid content among the strains was compared by first assembling the genomes using SPAdes v3.9⁴¹ with a minimum contig size cut off of 1000 bp. The chromosomal contigs were removed from the collection for each strain and those remaining were assessed via a pangenome approach using Roary³² allowing for the collapse of paralogs. All 21 complete plasmid sequences associated with the B31 reference sequence were included (Table S1). Phylo.io⁵² was used to compare the resulting phylogenetic tree, based on gene possession in the accessory genome, to the one based on the chromosome generated using SNVPhyl.

Data availability. All data have been submitted to Genbank under BioProject PRJNA416494, while the MLST and isolate data are available at the MLST website <http://pubmlst.org/borrelia>.

References

- Kilpatrick, A. M. *et al.* Lyme disease ecology in a changing world: consensus, uncertainty and critical gaps for improving control. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **372**, 1722 (2017).
- Ogden, N. H., Koffi, J. K., Pelcat, Y. & Lindsay, L. R. Environmental risk from Lyme disease in central and eastern Canada: a summary of recent surveillance information. *Can. Commun. Dis. Rep.* **40**, 74–82 (2014).
- Ogden, N. H. *et al.* Evolutionary aspects of emerging Lyme disease in Canada. *Appl. Environ. Microbiol.* **81**, 7350–7359 (2015).
- Hanincová, K. *et al.* Multilocus sequence typing of *Borrelia burgdorferi* suggests existence of lineages with differential pathogenic properties in humans. *PLoS One* **8**, e73066 (2017).
- Wang, G. *et al.* Impact of genotypic variation of *Borrelia burgdorferi* sensu stricto on kinetics of dissemination and severity of disease in C3H/HeJ mice. *Infect. Immun.* **69**, 4303–4312 (2001).
- Wormser, G. P. *et al.* Effect of *Borrelia burgdorferi* genotype on the sensitivity of C6 and 2-tier testing in North American patients with culture-confirmed Lyme disease. *Clin. Infect. Dis.* **47**, 910–914 (2008).
- Gilmour, M. W., Graham, M., Reimer, A. & Van Domselaar, G. Public health genomics and the new molecular epidemiology of bacterial pathogens. *Public Health Genomics* **16**, 25–30 (2013).
- Aanensen, D. M. *et al.* Whole-genome sequencing for routine pathogen surveillance in public health: a population snapshot of invasive *Staphylococcus aureus* in Europe. *MBio* **7**, e00444–16 (2016).
- Hossain, M. J. *et al.* An Asian origin of virulent *Aeromonas hydrophila* responsible for disease epidemics in United States-farmed catfish. *MBio* **5**, e00848–14 (2014).
- Jarvis, E. D. *et al.* Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* **346**, 1320–1331 (2014).
- Katz, L. S. *et al.* Evolutionary dynamics of *Vibrio cholerae* O1 following a single-source introduction to Haiti. *MBio* **4**, e00398–13 (2013).
- Thorpe, H. A., Bayliss, S. C., Hurst, L. D. & Feil, E. J. Comparative analyses of selection operating on nontranslated intergenic regions of diverse bacterial species. *Genetics* **206**, 363–376 (2017).
- Kurtenbach, K. *et al.* Key processes in the evolutionary ecology of Lyme borreliosis. *Nat. Rev. Microbiol.* **4**, 660–669 (2006).
- Margos, G., Vollmer, S. A., Ogden, N. H. & Fish, D. Population genetics, taxonomy, phylogeny and evolution of *Borrelia burgdorferi* sensu lato. *Infect. Genet. Evol.* **11**, 1545–1563 (2011).
- Margos, G. *et al.* MLST of housekeeping genes captures geographic population structure and suggests a European origin of *Borrelia burgdorferi*. *Proc. Natl. Acad. Sci. USA* **105**, 8730–8735 (2008).
- Liveris, D., Gazumyan, A. & Schwartz, I. Molecular typing of *Borrelia burgdorferi* sensu lato by PCR-restriction fragment length polymorphism analysis. *J. Clin. Microbiol.* **33**, 589–595 (1995).
- Hanincova, K., Liveris, D., Sandigursky, S., Wormser, G. P. & Schwartz, I. *Borrelia burgdorferi* sensu stricto is clonal in patients with early Lyme borreliosis. *Appl. Environ. Microbiol.* **74**, 5008–5014 (2008).
- Qiu, W. G., Dykhuizen, D. F., Acosta, M. S. & Luft, B. J. Geographic uniformity of the Lyme disease spirochete (*Borrelia burgdorferi*) and its shared history with tick vector (*Ixodes scapularis*) in the Northeastern United States. *Genetics* **160**, 833–849 (2002).
- Brisson, D., Baxamusa, N., Schwartz, I. & Wormser, G. P. Biodiversity of *Borrelia burgdorferi* strains in tissues of Lyme disease patients. *PLoS One* **6**, e22926 (2011).
- Ogden, N. H. *et al.* Investigation of genotypes of *Borrelia burgdorferi* in *Ixodes scapularis* ticks collected in surveillance in Canada. *Appl. Environ. Microbiol.* **77**, 3244–3254 (2011).
- Mechai, S., Margos, G., Feil, E. J., Lindsay, L. R. & Ogden, N. H. Complex population structure of *Borrelia burgdorferi* in southeastern and south central Canada as revealed by phylogeographic analysis. *Appl. Environ. Microbiol.* **81**, 1309–1318 (2015).
- Mechai, S. *et al.* Evidence for host-genotype associations of *Borrelia burgdorferi* sensu stricto. *PLoS One* **11**, e0149345 (2016).
- Ogden, N. H., Arsenaault, J., Hatchette, T. F., Mechai, S. & Lindsay, L. R. Antibody responses to *Borrelia burgdorferi* detected by Western blot vary geographically in Canada. *PLoS One* **12**, e01717312017 (2017).
- Schutzer, S. E. *et al.* Whole-genome sequences of thirteen isolates of *Borrelia burgdorferi*. *J. Bacteriol.* **193**, 1018–1020 (2011).
- Yang, X. *et al.* Characterization of unique regions of *Borrelia burgdorferi* surface-located membrane protein 1. *Infect. Immun.* **78**, 4477–4487 (2010).

26. Walter, K. S., Carpi, G., Caccone, A. & Diuk-Wasser, M. A. Genomic insights into the ancient spread of Lyme disease across North America. *Nat. Ecol. Evol.* **1**, 1569–1576 (2017).
27. Nei, M. & Gojobori, T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3**, 418–426 (1986).
28. Rocha, E. P. *et al.* Comparisons of dN/dS are time dependent for closely related bacterial genomes. *J. Theor. Biol.* **239**, 226–235 (2006).
29. Kryazhimskiy, S. & Plotkin, J. B. The population genetics of dN/dS. *PLoS Genet.* **4**, e1000304 (2008).
30. Croucher, N. J. *et al.* Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res.* **43**, e15 (2015).
31. Huson, D. H. & Bryant, D. Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* **23**, 254–267 (2006).
32. Page, A. J. *et al.* Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691–3693 (2015).
33. Schutzer, S. E. *et al.* Whole-genome sequences of *Borrelia bissettii*, *Borrelia valaisiana*, and *Borrelia spielmanii*. *J. Bacteriol.* **194**, 545–546 (2012).
34. Castillo-Ramírez, S. *et al.* Trans-Atlantic exchanges have shaped the population structure of the Lyme disease agent *Borrelia burgdorferi* sensu stricto. *Sci. Rep.* **6**, 22794 (2016).
35. Vuong, H. B. *et al.* Influences of host community characteristics on *Borrelia burgdorferi* infection prevalence in blacklegged ticks. *PLoS One* **12**, e0167810 (2017).
36. Casjens, S. R. *et al.* Plasmid diversity and phylogenetic consistency in the Lyme disease agent *Borrelia burgdorferi*. *BMC Genomics* **18**, 165 (2017).
37. Casjens, S. R. *et al.* Genome stability of Lyme disease spirochetes: comparative genomics of *Borrelia burgdorferi* plasmids. *PLoS One* **7**, e33280 (2012).
38. Margos, G. *et al.* Completed genome sequences of *Borrelia burgdorferi* sensu stricto B31 (NRZ) and closely related patient isolates from Europe. *Genome Announc.* **5**, e00637–17 (2017).
39. Ogden, N. H., Koffi, J. K. & Lindsay, L. R. Assessment of a screening test to identify Lyme disease risk. *Can. Commun. Dis. Rep.* **40**, 83–87 (2014).
40. Courtney, J. W., Kostelnik, L. M., Zeidner, N. S. & Massung, R. F. Multiplex real-time PCR for detection of *Anaplasma phagocytophilum* and *Borrelia burgdorferi*. *J. Clin. Microbiol.* **42**, 3164–3168 (2004).
41. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
42. Petkau, A. *et al.* SNVPhyl: a single nucleotide variant phylogenomics pipeline for microbial genomic epidemiology. *Microb. Genom.* **3**, e000116 (2017).
43. Burgdorfer, W. *et al.* Lyme disease—a tick-borne spirochetosis? *Science* **216**, 1317–1319 (1982).
44. Johnson, R. C., Schmidt, G. P., Hyde, F. W., Steigerwalt, A. G. & Brenner, D. J. *Borrelia burgdorferi* sp. nov.: etiological agent of Lyme disease. *Int. J. Syst. Bacteriol.* **34**, 496–497 (1984).
45. Huang, W., Li, L., Myers, J. R. & Marth, G. T. ART: a next-generation sequencing read simulator. *Bioinformatics* **28**, 593–594 (2012).
46. Walter, K. S., Carpi, G., Evans, B. R., Caccone, A. & Diuk-Wasser, M. A. Vectors as epidemiological sentinels: patterns of within-tick *Borrelia burgdorferi* diversity. *PLoS Pathog.* **12**, e1005759 (2016).
47. Ronquist, F. *et al.* MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **61**, 539–542 (2012).
48. Hall, B. G. *Phylogenetic trees made easy: a how-to manual*, 4th ed. (Sinauer Associates, Sunderland, MA, 2011).
49. Gelman, A. & Rubin, D. B. Inference from iterative simulation using multiple sequences. *Stat. Sci.* **7**, 457–511 (1992).
50. Buninkis, J. *et al.* Sequence typing reveals extensive strain diversity of the Lyme borreliosis agents *Borrelia burgdorferi* in North America and *Borrelia afzelii* in Europe. *Microbiology* **150**, 1741–1755 (2004).
51. Centers for Disease Control and Prevention (CDC). Recommendations for test performance and interpretation from the Second National Conference on Serologic Diagnosis of Lyme Disease. *Morb. Mortal. Wkly. Rep.* **44**, 590–591 (1995).
52. Robinson, O., Dylus, D. & Dessimoz, C. Phylo.io: interactive viewing and comparison of large phylogenetic trees on the web. *Mol. Biol. Evol.* **33**, 2163–2166 (2016).

Acknowledgements

This study was by the Public Health Agency of Canada. We thank Yann Pelcat for his help preparing Figure 1. We thank the communities and private land owners who kindly provided access so that ticks could be collected for this study.

Author Contributions

N.H.O., S.Tyl., G.vD., M.G., G.M. and E.J.F. conceived the study. L.R.L., A.D. and M.D. ensured collection of ticks, PCR and culture of strains. S.Tyl., S.Tys., G.vD., M.G. enabled and undertook sequencing and bioinformatics analysis. S.Tyl., S.Tys., N.C.K., S.M., G.M., N.H.O., E.J.F. undertook sequence and phylogenetic analyses and interpretation. H.A.T. and E.J.F. undertook analysis of selection on the chromosomal genes. N.H.O. led writing the manuscript to which all contributed. All authors reviewed and accepted the final version of this manuscript.

Additional Information

Supplementary information accompanies this paper at <https://doi.org/10.1038/s41598-018-28908-7>.

Competing Interests: The authors declare no competing interests.

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



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2018