# Phylogeographic Reconstruction of *Borrelia Burgdorferi Sensu Stricto* in North America and Europe

Frank Burkhart, Ethan Smith

April 22, 2024

## 1 Introduction

Lyme disease is a vector-borne illness that is transmitted by the Spirochete *Borrelia Burgdorferi*. The disease was first discovered and clinically documented in Lyme, Connecticut in 1976 [6]. Since then, it has become the most prevalent vector-borne disease in North America, with the CDC estimating nearly $500,000$ new cases per year [1]. Since the new implementation of the Lyme disease surveillance case definition in 2022, annual case reports have increased an estimated 69% [3]. Most cases of Lyme disease in North America are caused by the *Borrelia Burgdorferi* species, *Sensu Stricto* (*Borrelia Burgdorferi s.s*) [6]. This species is most commonly transmitted to humans via the *Ixodes scapularis* vector (also known as deer or black-legged ticks). Several incidental host populations of *Borrelia Burgdorferi s.s.* exist, but do not contribute significantly to the transmission of Lyme disease [7]. These hosts include white-legged mice, white-tailed deer, birds, reptiles, and humans [5]. The complex migratory patterns of these hosts make studying the phylogeography of *Borrelia Burgdorferi* of great interest.

Given the complex migratory patterns of this pathogen, working towards understanding its spatial spread can greatly aid in Lyme disease mitigation. Our project seeks to reconstruct the phylogeography of *Borrelia Burgdorferi* across regions in North America and Western Europe. Specifically, we seek to reconstruct the phylogeography of *Borrelia Burgdorferi* using a discrete trait model. Our sequence data includes samples from regions in Western North America, Midwestern North America, Eastern North America, and Western Europe. Our analysis seeks to determine how much migration could be occurring between each of these four regions, if any exists at all. Previous work has shown that there exists bi-directional gene flow between the Midwestern and Eastern United States [7]. However, this analysis did not include regions in Canada, Europe, and the Western United States. Therefore, we seek to verify these prior results about migration between the Midwestern and Eastern United States and also extend the analysis to include regions in Canada and Europe.

# 2 Data Collection and methods

## 2.1 Data Collection and Preprocessing

For our discrete trait phylogeographic reconstruction, we chose to analyze multi-locus sequence typing (MLST) data. Despite there existing a few hundred whole genome sequences for *Borrelia Burgdorferi s.s.*, we found that they were not collected from the regions we sought to analyze in this project. Therefore, we only incorporated MLST data with region attributes in our analysis.

We started by downloading all available MLST sequences of the standard housekeeping genes for *Borrelia Burgdorferi* . These sequences include the clpA, clpX, nifS, pepX, pyrG, recG, rplB, and uvrA genes [4]. Each sequence of concatenated genes is approximately $5,000$ bp long. The metadata and corresponding sequence data were downloaded from the Public database for molecular typing and microbial genome diversity (PubMLST) [2]. We then proceeded to filter our sequences down to only include samples from our regions of interest and to only include the *sensu stricto* species.

Recall that our regions of interest for this analysis are Western, Midwestern, and Eastern North America, as well as Western Europe. We decided to subdivide each sample from these regions into smaller group locations to use as our discrete traits. This resulted in the distribution of samples across the locations specified in Table 1.

| Group Name | Locations |
|---|---|
| **Europe** | |
| UK | UK |
| Mainland Europe | IT, DE, FR, CH |
| **Canada** | |
| Quebec/Newfoundland | QC, NL |
| New Brunswick/PEI/Nova Scotia | NB, PE, NS |
| Ontario | ON |
| Manitoba | MB |
| British Columbia/Alberta | BC, AB |
| **USA** | |
| New England | ME, VT, NH, MA, RI, CT |
| North-East US | NY, PA, NJ |
| DC-Adjacent | MD, VA |
| East Mid-West US | MI, IN, IL, MO |
| West Mid-West US | WI, MN, IA |
| West US | CA |

Table 1: Table of the geographical locations defining each group from the first grouping.

The initial filtering of sequences from PubMLST to our regions of interest resulted in over $1,500$ sequences, which would be computationally infeasible to perform our analysis on. Because of this, we down-sampled these sequences to obtain a reasonable number of samples per location grouping. To do this, we selected up to a maximum of 40 samples per group. If a group did not have more than 40 samples available prior to down-sampling, all of the samples from that group were included in the analysis. To ensure that down-sampling did not significantly influence our findings, we replicated the down-sampling procedure to create two equivalently sized sample pools on which the same analyses would be performed.

| Location Group | Original Count | Down-sampled Count |
|---|---|---|
| UK | 32 | 32 |
| Mainland Europe | 90 | 40 |
| QC/NF | 106 | 40 |
| NB/PEI/NS | 95 | 40 |
| Ontario | 227 | 40 |
| Manitoba | 95 | 40 |
| BC/AB | 33 | 33 |
| New England | 247 | 40 |
| North-East US | 341 | 40 |
| DC-Adjacent | 14 | 14 |
| East Mid-West US | 22 | 22 |
| West Mid-West US | 145 | 40 |
| West US | 35 | 35 |

Table 2: Table of the distribution of original and down-sampled sequence counts for each discrete trait location for the first grouping.

While our initial fine-grain location grouping gave us a uniform distribution of the number of samples across locations, the large number of groups have the potential to produce a messy, unreadable tree. Because of this, we created a second set of location groupings to use in a second analysis. These groups were made by summing the previous locations into the larger, pooled groups of Western North America, Midwestern North America, Eastern North America, and Europe. The resulting distribution of sequences for each group is illustrated in Table 3.

| Pooled Group | Original Groups | Sequence Count |
|---|---|---|
| EU | UK, Mainland Europe | 72 |
| NA East | New England, North-East US, QC/NF, NB/PEI/NS, Ontario | 214 |
| NA Midwest | East Mid-West US, West Mid-West US, Manitoba | 102 |
| NA West | West US, BC/AB | 68 |

Table 3: Table of the distribution of sequence counts for each discrete trait location of the second grouping.

Here we are sacrificing the previously achieved uniform distribution of each group's sequence count for the benefit of readability. Either way, each grouping should result in a tree that indicates whether or not there exists gene transfer between regions of North America and Western Europe. With these sequence groupings in place, we proceeded with our discrete trait analyses.

### 2.2 Methodologies

To investigate cross continental and intercontinental migration of *Borrelia Burgdorferi s.s.*, we performed a discrete trait analysis using BEAST2. We performed reconstruction twice, once for the groupings described in Table 2, and again for the groupings from Table 3.

For each set of groupings, we performed the following steps. To begin, we aligned our sequences using MAFFT, specifically the FFT-NS-i iterative refinement method with a maximum of 2 iterations. From here, we set up a discrete trait model in BEAUTi. We used a Gamma site model with 4 categories and a HKY substitution model. A strict clock rate for time calibration was chosen with a default rate of 1.0. For our tree prior, we used a constant coalescent population model with all other priors left at their default state. Our discrete traits are the location groupings previously mentioned. We finished our analysis setup by using a symmetric substitution model, a Poisson prior for nonZero rates, and a log-normal distribution for geographical rate prior. We chose an MCMC chain length of $10,000,000$. The result of this section is two XML files for each of the location groupings, which we imported into BEAST2 to perform the MCMC.

## 3 Results and Discussion

### 3.1 Results for the first grouping

Our resulting tree for the analysis run on the 13 groups is provided in Figure 1. The issue of readability becomes immediately obvious when viewing the tree, however, it is still clear that not only does there exist migration between the major regions across North America, but also migration between North America and Europe. The tree has an initial split between the New England and Western Canada groupings.

Additionally, while internal nodes closer to the root appear very noisy in

Figure 1, some terminal clades provide useful insights into the evolutionary relationship between our North American and European samples. Specifically, we can observe that some of the samples from the Mainland Europe group appear very far diverged from the rest of the samples. This would suggest an early introduction of *Borrelia Burgdorferi s.s.* into Europe from North America. Additionally, we can see that the majority of the UK samples, along with a couple of Mainland Europe samples have formed their own clade towards the bottom middle of the tree. This would suggest that the introduction of *Borrelia Burgdorferi s.s.* into the UK was also independent of the introduction events responsible for lineages present in the Mainland Europe samples at the bottom and top of the tree. Overall, while the tree produced from this first grouping did not yield the clearest insights into the phylogeography of *Borrelia Burgdorferi s.s.* on each continent, it did provide us with evidence of gene transfer between regions of North America and Western Europe, as well as the multiple distinct introduction events of *Borrelia Burgdorferi s.s.* into Europe.

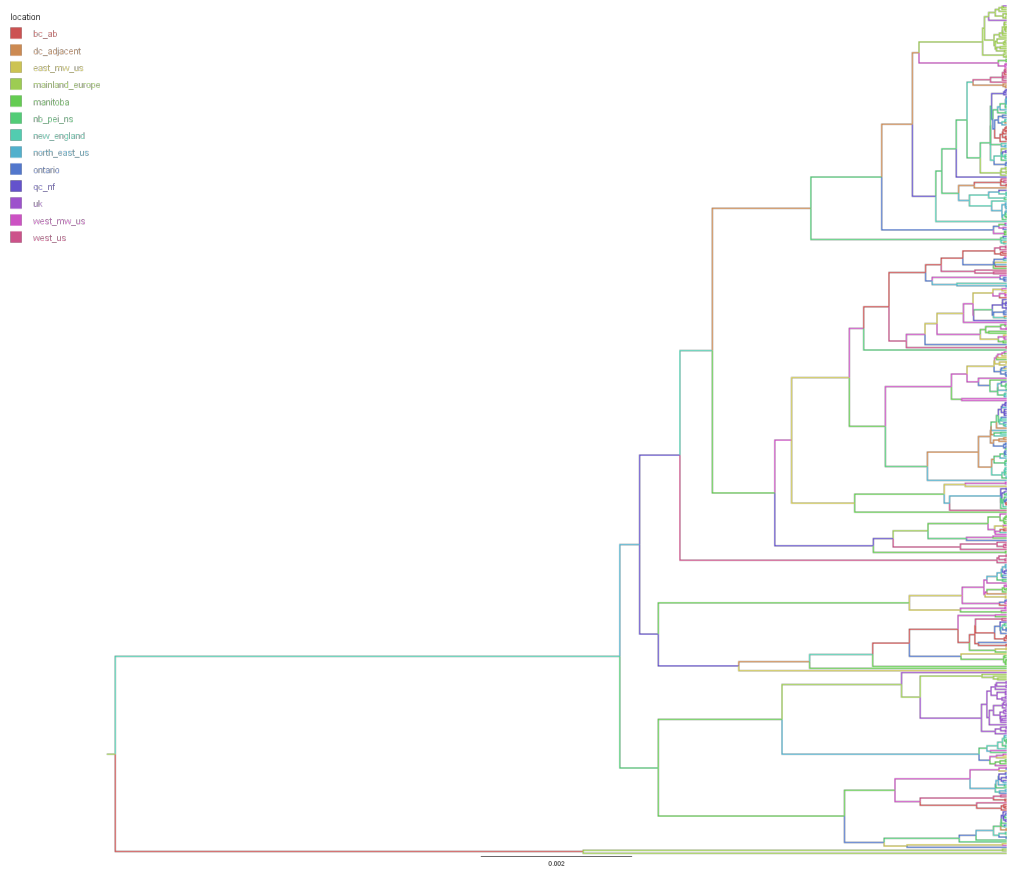


Figure 1: Phylogeographic reconstruction of 13 groupings from regions in North America and Western Europe.

## 3.2 Results for the second grouping

Despite our primary biological question being answered from our first tree, the number of groups made readability very difficult. Thus, we re-ran the analysis using the groupings outlined in Table 3 in order to achieve the following tree.

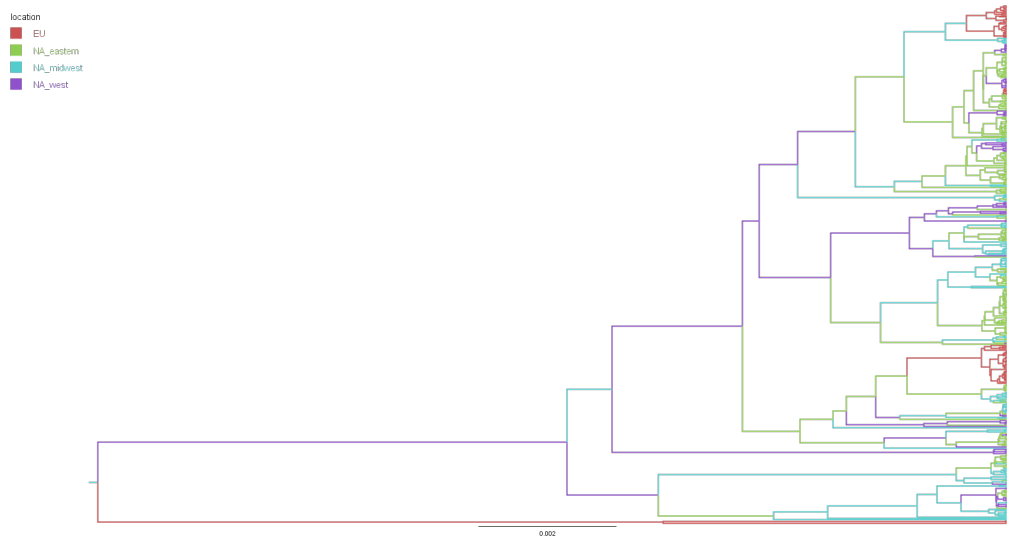


Figure 2: Phylogeographic reconstruction of 4 pooled groupings from regions in North America and Western Europe.

To validate the results we obtained in Figure 2, we used the second batch of randomly down-sampled samples and obtained the tree seen in Figure 3.

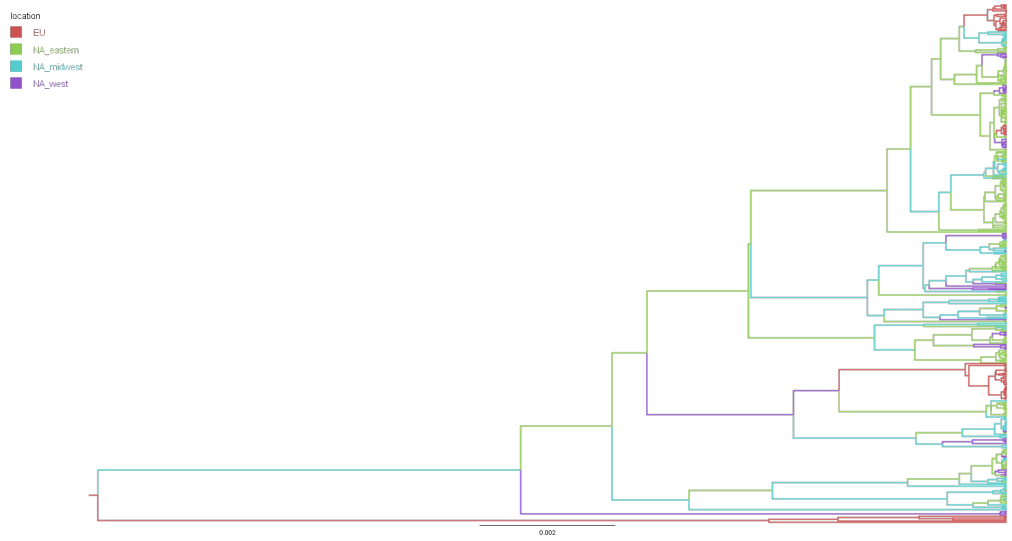


Figure 3: Phylogeographic reconstruction of 4 pooled groupings from regions in North America and Western Europe using alternate samples.

The similarity between the tips of the trees observed in Figure 2 and Figure 3 suggests that the down-sampling we performed from our original dataset did not meaningfully impact our ability to infer migration between regions. Additionally, it confirms our observation suggesting that there were at least four separate introduction events of *Borrelia Burgdorferi s.s.* into Western Europe from North America, as the same four European clades are observed in both trees. Further down the tree, we observe evidence which confirms the prior work that there is bi-directional gene transfer occurring between regions in Eastern and Midwestern North America. To add to this previous knowledge, our tree suggests that this transfer is occurring between all three highlighted regions of North America.

While the tree topology towards the tips of the tree remains relatively constant for each group of samples, we observe some differences towards the root ends of each tree, indicating that the origins of *Borrelia Burgdorferi s.s.* are unclear. Figure 1 suggests a Midwestern origin that splits off into Europe and Western North America. On the contrary, Figure 3 indicates a European origin with an early split into North Eastern North America. While conflicting roots for each tree make the origin of *Borrelia Burgdorferi s.s.*unclear, we can still conclude from any of the three trees that migration between each region is occurring.

# References


[1] Centers for Disease Control and Prevention. How many people get lyme disease? Accessed on April 21, 2024.

[2] Keith A Jolley, James E Bray, and Martin CJ Maiden. Open-access bacterial population genomics: Bigsdb software, the pubmlst. org website and their applications. *Wellcome open research*, 3, 2018.

[3] Kiersten J Kugeler. Surveillance for lyme disease after implementation of a revised case definition—united states, 2022. *MMWR. Morbidity and Mortality Weekly Report*, 73, 2024.

[4] Gabriele Margos, Anne G Gatewood, David M Aanensen, Klára Hanincová, Darya Terekhova, Stephanie A Vollmer, Muriel Cornet, Joseph Piesman, Michael Donaghy, Antra Bormane, et al. Mlst of housekeeping genes captures geographic population structure and suggests a european origin of borrelia burgdorferi. *Proceedings of the National Academy of Sciences*, 105(25):8730–8735, 2008.

[5] Gerold Stanek, Gary P Wormser, Jeremy Gray, and Franc Strle. Lyme borreliosis. *The Lancet*, 379(9814):461–473, 2012.

[6] Anna Szymanska, Anna E Platek, Miroslaw Dluzniewski, and Filip M Szymanski. History of lyme disease as a predictor of atrial fibrillation. *The American Journal of Cardiology*, 125(11):1651–1654, 2020.

[7] Katharine S Walter, Giovanna Carpi, Adalgisa Caccone, and Maria A Diuk-Wasser. Genomic insights into the ancient spread of lyme disease across north america. *Nature ecology & evolution*, 1(10):1569–1576, 2017.