Title: Complex exchanges among plasmids and clonal expansion of lineages shape the population structure and virulence of *Borrelia burgdorferi*

Rachel A. Laing*[1,2], Michael J. Foster*[1,2], M. Amine Hassani*[1,2], Benjamin Kotzen*[1,2], Weihua Huang[3,4], Terrence Shea[2], Stephen F. Schaffner[2], Tjasa Cerar[5], Lisa Freimark[2], Eva Ruzic-Sabljic[5], Dionysios Liveris[3], Kurt D. Reed[6], John A. Branda[1], Allen C. Steere[1], Gary P. Wormser[3], Franc Strle[7], Pardis C. Sabeti[2,8,9], Ashlee Earl[2], Ira Schwartz[3], Klemen Strle[10], Jacob E. Lemieux[1,2]

[1]Massachusetts General Hospital, Harvard Medical School, [2]Broad Institute of MIT and Harvard, [3]New York Medical College, [4]East Carolina University, [5]University of Ljubljana, [6]University of Wisconsin, [7]University Medical Center Ljubljana, [8]Harvard University, [9]Harvard T.H.Chan School of Public Health, [10]Tufts University.

*Contributed equally to this work

Correspondence to: Jacob E. Lemieux, jelemieux@mgb.org

Key words: Lyme disease, *Borrelia burgdorferi*, Whole Genome Sequencing; Virulence; Pathogenicity, recombination, genomics, lipoproteins

ABSTRACT

Background: In the United States, *Borrelia burgdorferi* (*Bb*) is the principal etiologic agent of Lyme disease. The complex structure of *Bb* genomes has posed challenges for genomic studies because homology among the bacterium's many plasmids, which account for ~40% of the genome by length, has made them difficult to sequence and assemble.

Results: We used long-read sequencing to generate near-complete assemblies of 62 isolates of human-derived *Bb* and collected public genomes with plasmid sequences. We characterized genetic diversity and population structure in the resulting set of 82 plasmid-complete *Borrelia burgdorferi* sensu stricto genomes. The *Bb* core genome is encoded by a chromosome and the conserved plasmids cp26, lp54, and lp17; the accessory genome is encoded by all other plasmids and the distal arm of the chromosome. Near-complete genomes reveal that the most granular *Bb* genotypes are clonal expansions of complex rearrangements among accessory genome elements. Ribosomal spacer types (RST) represent multiple collections of such genotypes, whereas OspC types are usually clonal. Structural rearrangements are non-randomly distributed throughout the genome, with cp32 plasmids undergoing dense exchanges and most linear plasmids, except lp54, sharing blocks among themselves and with the distal arm of the chromosome. OspC type A strains, known to possess greater virulence in humans, are distinguished by the presence of lp28-1 and lp56. Rearrangements among plasmids tended to preserve gene content, suggesting functional constraints among gene networks. Using k-partite graph decompositions, we identified gene sets with correlation patterns suggestive of conserved functional modules.

Conclusions: Long-read assemblies reveal that *Bb* population genetic structure results from clonal expansion of lineages that have undergone complex rearrangements among plasmid-encoded accessory genome elements. Genetic structure is preserved among genes even when plasmid rearrangements occur, suggesting that selection among epistatic loci maintains functional genetic networks. The analysis of near-complete genomes assembled using long-read sequencing methods advances our understanding of *Bb* biology and Lyme disease pathogenesis by providing the first detailed view of population variation in previously inaccessible areas of the *Bb* genome.

BACKGROUND

Lyme disease, the most common tick-borne disease in North America, is principally caused by the spirochete *Borrelia burgdorferi (Bb)*. The United States Centers for Disease Control and Prevention estimates that approximately 476,000 new cases of Lyme disease are diagnosed annually in the United States[1]. In humans, Lyme disease is variable and progresses in stages. Disease begins in the skin with the bite of an infected tick and, if untreated, *Bb* spirochetes may disseminate to other sites resulting in multifocal cutaneous disease, aseptic meningitis, carditis, arthritis, and other complications [2–4]. Most patients respond well to standard courses of oral antibiotic therapy, as detailed in the guidelines of the Infectious Diseases Society of America. However, some patients with Lyme arthritis, a late disease manifestation, require intravenous antibiotic therapy for successful treatment of the infection. Others develop post-infectious, antibiotic-refractory Lyme arthritis in which autoimmunity appears to play a role in disease pathogenesis. Moreover, in an estimated 5-20% of patients, Lyme disease is complicated by persistent pain, neurocognitive, or fatigue symptoms after treatment, termed post-treatment Lyme disease syndrome (PTLDS), which is difficult to treat effectively [5].

The complexity of Lyme disease is due in part to genetic diversity among the spirochetes that cause disease. The major divisions among *Borreliaceae* spirochetes that cause Lyme disease are referred to as genospecies, genomic species, or subspecies (reviewed in [6]), and they are collectively referred to as *Borrelia burgdorferi* sensu lato (“in the general sense”) to distinguish them from the original *Borrelia burgdorferi* sensu stricto (“in the strict sense”). Some investigators have proposed replacing *Borrelia burgdorferi* sensu lato with a new genus, *Borreliella*, because of the extent of genetic variation among the included spirochetes [7], but this proposal remains controversial [8,9]. Here, we use *Bb* to refer to *B. burgdorferi* sensu

stricto, the agent responsible for nearly all Lyme disease cases in the United States and the focus of this manuscript.

Bb has been divided into genotypes using multiple methods including serotyping [10], genotyping the 16S-23S intergenic ribosomal spacer type (RST) [11], sequencing of the outer surface protein C [12], sequencing of multiple genes to perform multi-locus sequence typing (MLST) [13], and sequencing complete genomes [14]. Analysis of *Bb* genotypes along with data on clinical outcomes has revealed a correlation between genotypes and clinical manifestations of Lyme disease [15–20]. For example, certain genotypes of *Bb* are more likely to disseminate [19–22], more likely to cause Lyme arthritis [23], and cause greater inflammation [24].

Defining the complete set of genomic differences between *Bb* strains that cause Lyme disease has been challenging for two reasons. First, human isolates of *Bb* are difficult to obtain because they typically require an invasive sample (such as a skin biopsy) and specialized culture methods. Second, the *Bb* genome has been particularly challenging to sequence, assemble, and analyze because *Bb* strains possess a unique genomic structure consisting of a single chromosome of approximately 910 kb and numerous plasmids (21 in the B31 reference strain) [25] that encode approximately one-third of gene content. These include both linear plasmids (named with the prefix lp) and circular plasmids (named with the prefix cp). Next-generation sequencing (NGS) methods have been used to generate partial genome sequences for a large number of isolates [14,26,27], but short reads cannot assemble most plasmids due to their similarity and complexity [28]. Yet characterizing plasmids is critical to understanding the phenotypic properties of *Bb* because plasmids encode the majority of surface antigens and determine virulence [3].

Recent advances in next-generation sequencing producing long reads can resolve the homologous regions in *Bb* plasmids [28,29]. In this study, we used long-read sequencing to characterize genetic diversity among *Bb* isolates, including plasmid sequences.

RESULTS

Long-read sequencing improves genome assembly and reveals *Bb* population structure.

We used long-read methods to sequence and assemble 62 *Bb* strains isolated from humans in the US and Slovenia and also downloaded all available assemblies with complete plasmid sequences available on Genbank (n = 20) (Table S1) [30]. We analyzed patterns of plasmid and genome diversity among these 82 genomes with complete plasmid sequences. Compared to a matched subset of 49 short-read assemblies (SRAs) generated in earlier work [14], long-read assemblies (LRAs) yielded entire replicons (Figure S1A), resulted in longer contigs (Figure S1B), and led to fewer incomplete genes (Figure S1C).

We annotated LRAs to determine the plasmids present in each isolate (Figure 1). Following earlier work [31], we named plasmids using their PFam32 plasmid compatibility sequences (PF-32 genes) when present (PF-32 genes are found in all replicons except cp9 and lp5) [28] or BLAST homology when absent. We estimated the frequency of plasmids among major genotypes and in the dataset overall (Figure S2). We classified plasmids as either core (present in >95% of isolates) or accessory (<95% of isolates). Based on this criterion, the only core plasmids are the 26 kilobase (Kb) circular plasmid (cp26), a 54 Kb linear plasmid (lp54) and a 17 Kb linear plasmid (lp17). Cp26 and lp54 are known to be essential [32–34]. Lp17, lp36, lp28-3, lp25, cp32-7, lp28-4, and lp38 were present in >75% of isolates (Figure 1, Figure S2). In contrast, many accessory plasmids were present in the sample at low frequencies and in certain genotypes only. For example, lp28-2, and lp56 were more common among RST1 and OspC type A isolates; lp28-1 was more common among RST1 and RST2 samples than among RST3.

whereas lp28-5 and lp28-6 were more common among RST2 and OspC type K isolates (Figure S2). Plasmid count within isolates ranged from 10 (ASM1913465v1, ASM215148v1, ESI361H) to 23 (URI40H).

To quantify the sharing of genetic content between replicons, we aligned each assembly to the B31 reference and also against all other assemblies in an “all vs one” and “all vs all” comparison scheme. We assessed pairwise homology between each assembly and the B31 reference (Figure S3A). We also computed normalized mutual information to compare the distribution of genes across replicons in the reference versus assembled genomes (Figure S3B-C). Normalized mutual information (NMI) is an information-theoretic concept reflecting the reduction in uncertainty between two dependent random variables when the outcome of one of them is known; an NMI of 0 indicates no reduction in uncertainty while an NMI of 1 indicates a complete reduction in uncertainty. The mean NMI across all isolates was 0.84 nats (a unit of information), and increased to 0.87 nats among OspC Type A isolates, the genotype of the reference isolate, B31. This indicates a greater similarity among genetic architecture for isolates of the same genotype. We also computed pairwise homology among all assemblies (Figure 4A). In contrast to other replicons, cp26 and lp54 do not share DNA content with each other or any other replicons (Figure S4A-B). Other linear and circular plasmids frequently exchange genetic content (Figure S3C). The degree of exchange between cp32 plasmids was high, making it difficult to distinguish between cp32 plasmids based on their associated gene content. The distal 3’ arm of the chromosome is also found to be homologous with various linear plasmids, as reported previously [35,36].

The *Bb* pangenome structures population genomic variation

We analyzed pangenome content and patterns of gene sharing across strains and plasmids. We constructed a pangenome by identifying ortholog groups and dividing them into

core (genes shared by >95% isolates) and accessory genes (ortholog groups present in the sample but not in the core) using Roary [37]. The *Bb* pangenome consists of approximately 900 core ortholog groups and 1100 accessory groups when paralogs are grouped together (Table 1). Separating paralogs into different groups, the core genome remains at approximately 900 gene groups, whereas the accessory genome expands to approximately 4000 groups (Table 1), pointing to the importance of paralogous gene families in accessory genome diversification in *Bb*.

We mapped the location of each ortholog group and plotted the presence of ortholog groups by isolate. We also annotated the encoding plasmid by color (Figure 2). As previously observed [14], ortholog groups were present in blocks that correlated with clades in the phylogenetic tree. LRAs revealed ortholog occupancy relationships across plasmids. In most instances, this block structure correlated with the encoding plasmid, but in several notable cases the patterns of ortholog group presence were maintained, whereas the encoding plasmid differed, indicating genetic structure among gene content that was preserved despite plasmid changes (Figure 2) and gene flow between replicons. Thus, although plasmid presence is correlated with gene content, plasmid occupancy among phylogenetic clades does not fully explain gene occupancy among phylogenetic clades. This block structure was also seen specifically among surface lipoproteins (Figure S5A-B). Surface lipoproteins were partitioned by plasmid along the phylogenetic tree, indicating that plasmids are a key mechanism for the heritability of surface antigens and other phenotypes linked to host interaction.

Certain genotypes, particularly RST1 strains and the subset characterized by OspC type A, are known to be more invasive. Invasive genotypes possess larger genomes with more lipoproteins (Figure S6A-B). For example, RST1 strains, particularly OspC type A genotype, encoded a greater number of lipoproteins than other genotypes (Figure S6B). OspC type K genomes, also known to be invasive [20], also had larger numbers of surface lipoproteins and

longer genomes than other genotypes. RST3 isolates had fewer plasmids than RST1 and RST2 isolates (Figure S6C).

The behavior of individual genes typified the pattern of general phylogenetic consistency disrupted by episodes of gene flow. Lp28-4 genes, such as group 2443, group 2444, and group 2448, were commonly found on other linear plasmids including lp25, lp36, lp28-3, or lp28-2 (Figure S5B). Other plasmids encoded different sets of genes along different branches of the phylogenetic tree. For example, multiple groups of genes (e.g., group 2189) occurred on lp38 primarily in RST2 and 3 isolates, while other groups (e.g., group 2190) occurred on lp38 in most RST1 isolates. Within this bifurcation of lp38 content, there was evidence of exchange to other plasmids. For example, group 1543 was encoded by lp38 in most RST2 and RST3 isolates but was encoded on lp28-8 in a subset of RST3 isolates from Europe. Numerous other examples of gene flow are apparent in Figures 2 and S5. These patterns were observed among isolates sequenced using different long-read sequencing technologies and assemblies obtained from Genbank, making sequencing assembly artifacts less likely.

We characterized genetic structure using several approaches. We typed isolates using single genetic markers (RST and OspC), applied multi-locus sequence typing (MLST) classification, and applied two clustering methods that incorporate whole genome sequence: a Bayesian clustering analysis (BAPS[38]) and kmer distance (PopPUNK[39]). The existence of multiple genetic subtypes of *Bb* was observed with all typing systems. Consistent with short-read analyses [14], single-locus typing methods were predictive of genetic structure but had limitations. RST1 is monophyletic, whereas RST2 and RST3 are polyphyletic markers that overlap among closely related populations. OspC types were generally monophyletic (with notable exceptions, e.g., type L); however, the large number of types without a measure of distance between them limits the utility of this typing system when the numbers of strains and types are large. Model-based clustering procedures generally identified genetic structure that

was intermediate between the coarseness of RST and the granularity of MLST. We also decomposed the structure of both the plasmid presence/absence matrix and the gene presence/absence matrix using principal component analysis (Figure S7). Variation in the gene group presence/absence matrix was better approximated by low-dimension projections (Figure S7) than variation in the plasmid presence/absence matrix, one indication that the genetic structure among genes is simpler than that among plasmids.

Genomic similarity analysis reveals clonal expansion

Previous reports using single genetic markers [40] and whole genome sequencing with short reads [14] have identified a high linkage disequilibrium between markers of different replicons and interpreted this as an indication of a near-clonal structure among *Bb* genotypes. To assess this possibility using LRAs and quantify similarity between the genomes, we computed pairwise average nucleotide identity (ANI) among the sequenced isolates and plotted these measurements alongside the core genome phylogeny (Figure 3A-B). ANI divergence was, expectedly, a strong correlate of phylogenetic distance, but patterns of ANI contained blocks indicative of structure within the phylogeny (Figure 3A). Blocks tended to correspond to OspC types. To further support these observations, we plotted ANI scores for a given OspC type (Figure 3B and Figure S8). Isolates from the same OspC types had a high degree of genomic similarity (Figure 3B), compared to isolates from different OspC types (Figure S8). These results indicate that they are nearly clonal. In contrast, isolates from the same RST contain multiple blocks of nucleotide similarity, with RST1 being the least diverse and RST3 the most diverse.

We projected the isolates into two dimensions using principal component analysis of the pangenome (Figure 4A). This two-dimensional embedding revealed that RST1 isolates formed a distinct subpopulation compared to RST2 and RST3 (Figure 4A). Consistent with the phylogenetic tree, RST2 and RST3 isolates predominantly separate into distinct clusters, but

there were some RST3 isolates in the RST2 cluster and vice versa, consistent with mixing of these types when genome-wide divergence is considered [14]. We also compared core genome divergence to accessory divergence (Figure 4B). There was an approximately linear relationship between core and accessory genome divergence, but accessory genome divergence was, at a given core genome distance, frequently multimodal (Figure 4B), indicating episodes of rapid and marked changes in the accessory genome. RST3 in particular appeared to be genetically heterogenous. Divergence of core and accessory distances within RST3 were equivalent to inter-RST distances (Figure 4C). To further characterize divergence in the accessory genome, we computed gene cluster occupancy (i.e., occurrence frequencies) within each RST type (Figure 4D) and indicated those shared across and between RST types (Figure 4E-F). Although no particular cluster of genes (COG) function [41] was enriched in an RST type, the ternary plot revealed gene clusters within the accessory pangenome that are shared across RST types, whereas others were RST type specific (Figure 4D). We observed 313 gene clusters shared by at least one strain from each RST type, as well as a large number of gene clusters unique to an RST type (Figure 4E). Notably, 149 gene clusters were unique to the RST3 strains (Figure 4E), indicative of the diversity of the accessory genome of RST3. Further inspection of set relationships among all gene clusters revealed that 32 out 69 and 45 out 72 were core to RST1 and RST2, respectively (Figure 4F). Strikingly, none of the 149 gene clusters unique to RST3 (Figure 4E) were shared across all RST3 strains (Figure 4D). These results further corroborate the contribution of the accessory genome in the diversification of the Bb genome and the heterogeneity of RST3. We also identified the full set of *Bb* gene groups associated with specific RSTs (Table S4).

Genetic changes within and between RSTs corresponded to blocks of homoplasy and complex rearrangements among plasmids, with columns (ortholog groups) with multiple colors (Figure 2), indicative of gene flow across plasmids and suggestive of recombination. To further

characterize recombination events, we used Gubbins [42], a phylogenetic approach to identify bases with high substitution rates (Figure S9). This analysis revealed recombination events that are widespread across the isolates (Figure S9A). We noted an average of 1.8 r/m and 36.5 kbp total recombination length (Figure S9A). Although we could detect recombination patterns across the phylogeny, there were no significant differences in total recombination length (Figure S9B) and r/m (Figure S9C) between the different RST and OspC types. Further analysis revealed that 58% (773/1340) of the genes had recombined at least one time (Figure S9A). Among these genes, we noted 88 predicted lipoprotein genes, including 35 surface lipoproteins, such as ospC, cspZ and cspA (Figure S9A). Recombination events were detected across the whole genome, except in lp5, cp32-1, cp32-4, cp32-6, cp32-9 (Figure S9D). Percent of genes that recombined per replicon varied, but we noted recombination events across genes with different occurrence frequency (Figure S9D).

Network analysis reveals genetic modules of *Bb* genotypes.

The block pattern of ortholog presence/absence (Figure 2) is maintained despite changes in the encoding plasmids. This phenomenon indicates that complex rearrangements such as plasmid fusion and/or genetic exchange preserve gene content. Such patterns are often indicative of functional relationships among gene networks [43,44]. To identify such networks, we used graph theory methods to create correlation networks and visualized their structure using force-directed algorithms (Supplemental Note). Correlation networks contained dense clusters corresponding to modules of correlated genes (Figure 5A-B). We colored genes by their occupancy relationships across RST (Supplemental Note). Under the Fructerman-Reingold layout with this colorspace representation, the network contains a well-defined structure. There is a core network of orthologs present in nearly all strains along with subclusters that correspond to RST, indicating that *Bb* gene relationships exist in modules that are expanded in genetic subtypes.

RST1 and RST2 isolates encode large gene networks, but there are only a small number of genes specific to RST3, as seen by the limited nodes in dark blue. The yellow region between RST1- (red) and RST2- (green) dominant genes represents a transition region in which genes are shared between RST1 and RST2, but not RST3, indicating genetic sharing between these types. PCA recapitulates the separation between gene groups prevalent in different RST types (Figure S10-S12), indicating that RST effectively captures much of the statistical variation in gene content in US *Bb* populations even though RST2 and RST3 are polyphyletic. Together, these analyses demonstrate a modular pattern of functional networks in *Bb*, with genotypes encoding distinct modules.

To characterize this network structure further, we decomposed the *Bb* pangenome correlation graph into k-partite subgraphs, which revealed 82 modules (Figure 6). We colored modules using the same colorspace mapping of RST as used for the network. Color coding of the modules by genotype revealed that modules are generally encoded by a single RST or may be shared by two RSTs. This effect was also apparent when we plotted the genes that comprise modules alongside the phylogeny (Figure S13) and colored modules by the plasmid families that encode them (Figure S14). The modular structure of *Bb* gene networks effectively recovers the genotypic patterns of *Bb* populations, consistent with previous analysis and emphasizing the way in which genetic variation in *Bb* results from clonal amplification of genotypes. Modules recovered from correlation networks are also more granular than RST, corresponding more closely to OspC types (Figure S13), indicating that more granular typing schemes capture intrinsic genetic variation as defined by the correlation of gene content.

Several modules were enriched in surface lipoproteins (e.g. Modules 72, 10, 4, 60, 23, 12, and 13; p values < 0.05 by hypergeometric test), consistent with putative functional properties and indicating their potential role in the display of surface antigens and/or host interactions. In some cases, modules that differ by genotype appeared to have related functions:

For example, module 12 (an RST1-specific module) encodes multiple Erp and Mlp proteins, similar to module 22, an RST2-specific module. Thus, these modules appear to capture the functions of host tissue adhesion and complement inhibition in separate genotypes. A complete list of modules and their gene groups is provided in Table S6. These networks reveal the modular nature of *Bb* genetic programs which are genotype-specific and encoded on groups of plasmids. A breakdown of gene group connections by plasmid shows differing structure and typing prevalence within (Figure S15). For example, cp32-3 demonstrates clear subregions of RST1-, RST2-, and RST3-prevalent genes, in contrast with cp32-12 which appears bimodal between RST2 and RST3, and cp32-2 which is dominated by RST2. Further some plasmids such as cp32-7 and lp28-5 demonstrate a more orderly structure with the spatial distribution of their gene groups, versus the chaotic mixture of various typing schemes in cp32-1. We also note that lp54 and cp26 both contain several near-core genes (large, off-white nodes), but those in lp54 tend to be lipoproteins while those in cp26 are not. All of these observations highlight the value of a plasmid-agnostic perspective at a high level; while plasmid presence / absence can provide some insight, it is really the underlying gene groups and the interactions between them that tell the full story.

The core genome phylogenetic tree, decomposition of *Bb* into k-partite modules, analysis of single genetic markers, and average nucleotide identity statistics all reveal that *Bb* is grouped into discrete clusters whose within-cluster variation is minimal whereas between-cluster variation is quite substantial. Although there is core genome divergence, nearly all of the divergence that contributes to population structure occurs within the accessory genome. Taken together, these analyses reveal a picture of *Bb* in diversity and population structure resulting from clonal amplification of new genotypes that contain complex reorganizations of plasmid-encoded accessory genome elements.

DISCUSSION

Using long-read sequencing, we present here a comprehensive analysis of the population genomic variation of *Bb* isolates. These detailed maps overcome the major challenges associated with the plasmid-rich, multipartite structure of *Bb* genomes, confirm existing observations about *Bb* population structure with new detail, and reveal novel insights into *Bb* evolution and genetic diversity.

First, we find that the inclusion of complete accessory genome content into analyses reaffirms that *Bb* contains a strong population structure. This is consistent with multiple decades of previous studies including those that have used serotyping [10,45], a variety of genetic markers (RST [11], OspC [12], MLST [13]), and short read sequencing [14]. This analysis confirms that RST1, particularly the subset marked by OspC type A, is a homogenous group of isolates that is genetically separated from other *Bb*. This especially virulent subset of strains is set apart from other *Bb* genotypes by a larger genome, a larger surface lipoproteome, and greater numbers of plasmids. OspC type A strains possess unique genetic content encoded by lp28-1 and lp56 plasmids, which are more commonly present in RST1 and Lp28-1 is also encoded by OspC type K strain. These findings underscore a paradigm in which a greater number of lipoproteins, which are associated with immune evasive and host invasive functions, is associated with a greater degree of virulence. The differences in gene content between virulent and less-virulent genotypes also provides a way to identify genes potentially involved in virulence (Table S5). Lp28-1 is known to encode the antigenic variation locus (vls), and its loss significantly reduces *B. burgdorferi* infectivity [46,47]. On the other hand, the loss of lp56, which is known to encode for DNA restriction and/or modification systems [48], correlated with a lower number of Borrelia cells inside ticks after feeding on live mice [49]. Gene content on these plasmids may confer virulence and survival and are consistent with the observations in humans

and in mice which demonstrate that OspC type A (RST1) and OspC type-K (RST2) strains have a greater propensity to disseminate [20,21,24].

It has been previously observed that *Bb* genotypes are nearly clonal [40]. Genome-wide assemblies confirm the relatively high genetic similarity of most OspC types (Figure 3B), but also demonstrate clonal expansion within these types. Observations of gene flow among plasmids (Figure 2, S3A) and genome-wide analysis of recombination (Figure S9) indicates that recombination is widespread among *Bb*, particularly among plasmid-encoded accessory genome elements, and contributes to the genome diversification of this microorganism through facilitating complex exchanges, particularly among plasmid-encoded accessory elements, which then undergo clonal expansion. The discrete jumps in accessory divergence at a given core genome divergence indicate that the evolutionary processes shaping the Lyme disease agent is saltatory, a phenomenon that has been observed in other bacteria [50,51].

How many genetic subtypes of *Bb* are there? While most of the variation in the pangenome can be described using dimensionality reduction methods with 3 dimensions (Figure S7,10-12), the large number and near-clonality of OspC types demonstrate the potential for a large number of rare genotypes (Figure S7). Single-locus typing systems, which capture much of the genetic variation in *Bb [14]* recover this variation reasonably well. However, while low-dimensional typing schemes capture the dominant structure among *Bb* genotypes, the modular nature of the accessory genome means there is a nearly limitless capacity to generate new types; as the principal component analysis (Figure S7) shows, higher-dimensional typing schemes continue to be needed to capture the full extent of genetic variation among *Bb* isolates. The large number of OspC types with substantial divergence between them and meaningful phenotypic differences in, for example, dissemination [14,20,21], imply that this variation is phenotypically important.

Consistent with prior observations [14,26,52], essentially all strain-variable genetic content is encoded on plasmids. Only the chromosome and a small number of plasmids such as cp26 and lp54 encode core genome content, observations which were known from short read sequencing and earlier techniques such as plasmid typing and genomic DNA hybridization. What is revealed by assembling complete plasmids is the extent to which conserved patterns among the accessory genome also emerge and the complex role that plasmids play in organizing these patterns.

As expected, plasmids are strongly correlated with accessory genome structure, an effect that is best seen by the consistent coloring of blocks in the map of orthologs by phylogeny (Figure 2). Yet, perhaps surprisingly, patterns of ortholog presence/absence persist even when the encoding plasmid changes (Figure 2). This indicates that evolutionary forces beyond physical linkage maintain this structure. Although the genetic information is inherited through plasmid transmission [53], these plasmids can be lost [54]. Gene flow could be seen as a preemptive mechanism of genetic loss and maintenance of important features, such as lipoproteins. Phages are one potential mechanism for gene flow in *Borrelia* [55,56], but given that gene flow is not limited to phage-packaged cp32 plasmids [55], other mechanisms likely contribute and need to be identified. The most likely candidate for this is selection acting on groups of alleles, i.e., epistasis, indicative of shared functional structure. We used graph theory to identify the genes that likely comprise these functional networks. These gene sets, which are modular and structured by genotype, may be valuable to further research into genetic and phenotypic analysis of *Bb*. For example, investigation into the greater bloodstream invasiveness and severity of OspC type A / RST1 strains [19–22,24] can begin with networks that are expanded in these subgroups of strains. Methods to identify substructure such as network centrality can be used to prioritize genes for functional studies.

Patterns of genetic diversity among core and accessory genomes are driven by accessory-weighted recombination and complex exchanges among plasmids. LRAs reveal extensive diversity in plasmid structure, with frequent episodes of plasmid fusion and partial sequence exchange. An example is lp38, which encodes different content in RST1 vs most RST2 strains (Figure 2). Such reconfigurations highlight the need for a subtyping system to add granularity not well-captured by PF-32 type. A corollary is that knowledge of the presence or absence of a plasmid is not a reliable predictor of gene content in *Bb* without additional information such as the strain genotype obtained via a high-resolution typing system such as whole genome sequencing or MLST.

These observations suggest that the plasticity of the genetic structure of the accessory genome itself is a mechanism to introduce diversity into the *Bb* genome. There is an approximately linear relationship between average genetic variation in the core and accessory genomes. The core genome, by separating core metabolic and structural genes on a conserved chromosome and a small number of conserved plasmids, provides a stable framework for core *Bb* functions. The accessory genome may serve as a genetic ratchet in which discrete jumps in genetic variation are encoded through complex rearrangements. Such complex changes may dramatically change the antigenicity, host tropism, and tissue distribution of the spirochete in ways that may be advantageous during adaptation to different ecological niches and environments, while insulating the core metabolic and reproductive functions of the spirochete, increasing the likelihood that complex mutants are viable.

This study has several limitations. The sampling is non-random and weighted toward North American clinical isolates. Rare OspC types are represented by one or only a few sequences. Isolates were cultured and despite attempts to minimize spontaneous plasmid loss by limiting passaging, they may have lost plasmids during culture adaptation [54,57]. In addition, although we generated assemblies by multiple different methods, we were not able to complete

the genomes fully for certain highly homologous plasmids. Finally, given the complexity of *Bb* genomes, assembly errors are possible. Long-read assemblers struggle with short plasmids, often requiring special considerations to be taken into account [58]. However, the consistency of phylogenetic patterns across multiple sources of isolates, including those sequenced for this study and published assemblies, suggests that misassemblies are not a major source of error.

A major strength of this work is that it provides the first detailed view of genome diversity among plasmid sequence diversity and genetic structure among *Bb*. The genomes and analysis presented here complements a recent report [59] on complete *B. burgdorferi* sensu lato genomes as systematic surveys of *Bb* evolution using LRAs confirm and expand observations from previous short read analyses [14,26,27,30,52] and earlier analyses of plasmid diversity [31,60,61]. In particular, LRAs reveal the ways in which plasmids provide a mechanism for strain-specific inheritance of accessory genome elements and reveal the complex interchanges that occur among plasmids, while genetic networks remain relatively stable. In addition to this panoramic representation of genome diversity, other strengths of this analysis include open-source pipelines for assembly and genotyping of *Bb* genomes and a large set of isolates from humans.

CONCLUSION

In summary, this survey of 82 near-complete assemblies provides the largest and most detailed survey to date of *Bb* genomic diversity among plasmids, which encode ~40% of *Bb* genetic content, and reveals that genome diversification occurs through complex rearrangement of plasmid sequences and clonal expansion. Accessory genome content is encoded by strain-variable plasmids. It is notable that genetic structure is preserved among genes even when plasmid rearrangements occur, suggesting that selection among epistatic loci maintains functional genetic networks. The network analysis indicated a modular functionalization of gene

content that is preserved when physical linkage on plasmid sequence changes. OspC type A strains, a virulent genotype, contain the largest genome size, greatest number of lipoproteins, and greatest number of plasmids and exemplify a model in which the number surface lipoproteins correlates with spirochetal dissemination and greater inflammation [14,24]. It will also be important to learn whether specific plasmids convey specific phenotypes in human disease. Thus, the map of genetic variation in *Bb* provided here will facilitate future efforts to investigate the biology and pathogenicity of the Lyme disease agent. This study advances our understanding of *Bb* biology and Lyme disease pathogenesis by providing a detailed view of population variation in previously inaccessible areas of the *Bb* genome and uncover potentially novel mechanisms of genome diversification.

METHODS

PacBio assemblies: Genomic DNA was extracted using the QIAamp DNA kit (Qiagen). Sequencing library was prepared using the SMRTbell template preparation kit (Pacific Biosciences). Sequencing was performed in the RSII SMRT sequencing system. After sequencing, reads were trimmed, filtered, and assembled using the Hierarchical Genome Assembly Pipeline (HGAP). Resulting contigs were trimmed for linear plasmids and the chromosome, and circularized for circular plasmids. All the contigs were manually polished and corrected by BWA alignment and mapping [62] and Integrative Genomics Viewer (IGV, [63]) visualization using the previously generated Illumina short-reads [14].

Oxford Nanopore Technologies (ONT) assemblies: ONT assemblies were generated as previously described [64]. Briefly, samples were barcoded using the Oxford Nanopore Rapid Barcoding kit SQK-RBK004 and run in batches of six samples per flow cell on a GridIon sequencer (Oxford Nanopore Technologies Ltd, Science Park, UK). Oxford Nanopore (ONT) reads were demultiplexed using Deepbinner (v0.2.0)[65], trimmed any remaining adapter using Porechop (v0.2.3)[66], and subsampled to ~50× depth of genome coverage. Illumina reads were trimmed of adapters using Trim Galore (v0.5.0) and subsampled to ~100× depth of genome coverage. Two Unicycler (v0.4.4, with default settings)[67] hybrid assemblies were generated for each sample, one assembly combining the Illumina 100× dataset with the 50× subsampled ONT dataset and another assembly combining the Illumina 100× dataset with the full set of ONT reads (if >50×). ONT reads were aligned to Unicycler contigs using minimap2 (v2.15) [68]. Illumina reads were aligned to Unicycler contigs using BWA MEM (v0.7.17) [62], and the resulting alignments were input to Pilon (v1.23) [69] for assembly polishing.

Assembly Quality Control: All assembled contigs were validated to be free of contamination using Kraken2 using the “standard” database configuration [70]. Resulting assemblies were

then analyzed using QUAST v5.2 to generate assembly metrics with Biomark for gene prediction counts [71]. *Bb* B31 (Genbank assembly ASM868v2, NCBI RefSeq assembly GCF_000008685.2) was used as reference for preliminary determination of misassembly and relative genome content.

Genotyping: For determination of OspC typing, a BLAST database was constructed using sequences obtained from NCBI [72]. Each assembly was annotated via Bakta as previously described and the corresponding sequences annotated as OspC were aligned against this database. Alignments with highest identity were used as the OspC allele type and assemblies with multiple OspC alleles were determined to be contaminated and were dropped from further analysis. For determination of MLST type, a docker image, sanger-pathogens/mlst_check, was utilized [73]. Each assembly was then typed and the resulting output table merged with existing metadata. RST, MLST, and typing information was obtained from reference [14] and studies reported therein. BAPS was implemented in R using the package rhierbaps.[74]

Assembly annotation and pangenome construction: Generated assemblies were annotated using Bakta v1.9.1.[75]. The resulting gff3 annotations were then fed into Roary v3.11.2 [37] within a docker container for pan genome construction. We ran Roary with and without splitting paralogs (the -e flag). The pangenomic reference generated by Roary was annotated using Bakta. The presence of lipoproteins in the resulting pangenome references was initially evaluated via signal peptide prediction by DeepSig[76] as part of Bakta annotation. We also predicted lipoproteins from the pangenome using an internally developed Python implementation of the SpLip algorithm [77] that classifies putative signal peptides. Sorenson and Jaccard distances were computed using gene cluster presence/absence provided by Anvio pangenome analysis [78] implemented in “vegan” R package. Cophonetic distances were computed using the “stats” R package. Gene clusters derived from Anvio pangenome analysis were annotated using COG2020 database [41].

Tree construction: The core genome alignment of Roary used to generate phylogenetic trees using FastTree2[79] and RAxML[80]. Additionally, Geneious was used to generate various trees to assist manual review.

Plasmid Annotation: When possible, contigs were named using their PF-32 plasmid compatibility gene by aligning the contig to a database of PF-32 genes from Genbank and previous work [14]. If no PF-32 sequence was present, contigs were named according to BLAST homology with known plasmids from Genbank. Contigs shorter than 1kb were considered too fragmented to reasonably annotate and were left unannotated. See Data Availability to access tools for annotating plasmids.

Lipoprotein Identification: We identify lipoproteins by taking the union of those from the Dowdell paper [81] and our internal lipoprotein annotation program. Surface lipoproteins were further designated using the annotations from [81] (S, P-IM, P-OM), and additionally classifying all gene groups relating to Osps, Mlps, and Erps as surface lipoproteins [82].

Population structure analysis. To identify population structure in the 82 B*. burgdorferi sensu stricto* genomes, we used hierarchical Bayesian Analysis of Population Structure (BAPS) implemented in R using 2 levels of clustering and 9 initial clusters. The probability of cluster assignment was calculated for each genome [74]. We used Population Partitioning Using Nucleotide K-mers (PopPUNK) which is a workflow that uses variable length k-mers to define core and accessory genome clusters. To calculate distances between the genomes, we used variable k-mers lengths between 13 and 29 and used HDBSCAN for model fitting and refinement [39]. To compute genomic similarity between the genomes, we used PyANI program [83] implemented in Anvio and displayed full average nucleotide identity scores with a cutoff of 0.95 [78].

Prediction of recombination events. To identify recombination loci, we used ***G****enealogies* ***U****nbiased* ***B****y recom****B****inations In* ***N****ucleotide* ***S****equences* (Gubbins v8) on the whole-genome alignment of 82 strains [42]. Genomes were mapped against the reference *B. burgdorferi* B31 (accession AE000783) using SKA2 (https://github.com/bacpop/ska.rust) to generate the whole-genome alignment. Fasttree was used as the first model with 100 bootstraps, then RAxMLng with the model GTRGAMMA for later iterations.

Phylogenetic Visualisation: Tree figures are created using R version 4.3.0 (https://www.R-project.org/), with the packages ggplot2, ggtree, ggtreeExtra, ggnewscale, and tidytree. All trees are midpoint rooted and formatted with the standard rectangular layout. Metadata columns are converted into factor variables and affixed to the tree object, then plotted as separate layers.

Homology Network Construction: All genomes were concatenated into a single file that was then aligned against itself using MUMmer (v4, nucmer --maxmatch) [84]. These alignments were then parsed and used to construct a network where each node is an individual contig and edges represent weighted shared homology. This network was then visualized in 3D using Plotly and in 2D using Cytoscape [85]. Nodes were filtered with a length cutoff of 1500bp, alignment length of 500bp, and a weight of 0.01.

Graph Theory: We explore the relationships between genes using graph theory concepts and the NetworkX package (v2.7.1) in Python 3.11.3. Gene analysis is performed at the gene group level using presence/absence data from pangenome analysis. For all pairwise combinations of non-invariant gene groups, Pearson's correlation is calculated. These correlations can then be displayed via network format by representing the sign and strength of the correlation through the color and thickness of the edge between the two nodes. Further details on node color-mapping and module analysis can be found in the supplement.

Acknowledgments:

This work was supported by the National Institute of Allergy and Infectious Diseases (K99/R00148604 to J.E.L, U19AI110818 and U01AI151812 to P.C.S.; R01AI045801 to I.S., and R21AI144916 to K.S.), the National Institute of Arthritis and Musculoskeletal and Skin Diseases (R01AR41511 to I.S. and K01AR062098 to K.S.), the Bay Area Lyme Foundation (to P.C.S. and J.E.L.), the Howard Hughes Medical Institute (P.C.S.), the Arthritis Foundation Fellowship (to K.S.) and the Wadsworth Center startup funds (to K.S.). We gratefully acknowledge the investigators who generated and submitted *Bb* genome data to genbank.

Author Contributions:

Analyzed data: R.A.L., M.J.F., M.A.H., B.K., W.H., T.S., S.F.S., A.E., I.S, K.S., J.E.L. M.J.F. and B.K. curated the data through bioinformatics software development, and performed homology analyses. R.A.L. performed phylogenetic and statistical analyses, and developed network analysis methodology. M.A.H. performed population structure and recombination analyses. M.J.F., T.S., and W.H. performed assemblies. T.C., E.R.S., D.L., K.D.R., J.A.B., A.C.S., G.P.W., F.S., K.S., I.S. contributed materials. L.F. generated sequencing libraries. R.A.L., M.J.F., M.A.H., and B.K. generated all figures. R.A.L., M.J.F., M.A.H., B.K., I.S., K.S., and J.E.L. wrote the first draft of the manuscript and incorporated feedback from all authors. J.E.L. supervised the project. All authors read and approved the final manuscript.

Declaration of interests:

P.C.S. is a co-founder of, shareholder in, and consultant to Sherlock Biosciences and Delve Bio, as well as a board member of and shareholder in Danaher Corporation. K.S. served as a consultant for T2 Biosystems, Roche, BioMerieux, and NYS Biodefense Fund, for the development of a diagnostic assay in Lyme borreliosis. F.S. served on the scientific advisory board for Roche on Lyme disease serological diagnostics and on the scientific advisory board

for Pfizer on Lyme disease vaccine, and is an unpaid member of the steering committee of the ESCMID Study Group on Lyme Borreliosis/ESGBOR. J.A.B. has received research funding to his institution from Analog Devices Inc., Zeus Scientific, Immunetics, Pfizer, DiaSorin, bioMerieux and the Steven & Alexandra Cohen Foundation, and has been a paid consultant to T2 Biosystems, DiaSorin, Flightpath Biosciences and Roche Diagnostics. G.P.W. reports receiving research grants from Biopeptides, Corp. He has been an expert witness in malpractice cases involving Lyme disease and babesiosis; and is an unpaid board member of the non-profit American Lyme Disease Foundation. J.A.B. and J.E.L. are co-authors on a provisional patent application for the diagnosis of Lyme Disease unrelated to this work.

Data and code availability:

Code is available at https://github.com/broadinstitute/Bb_pangenome

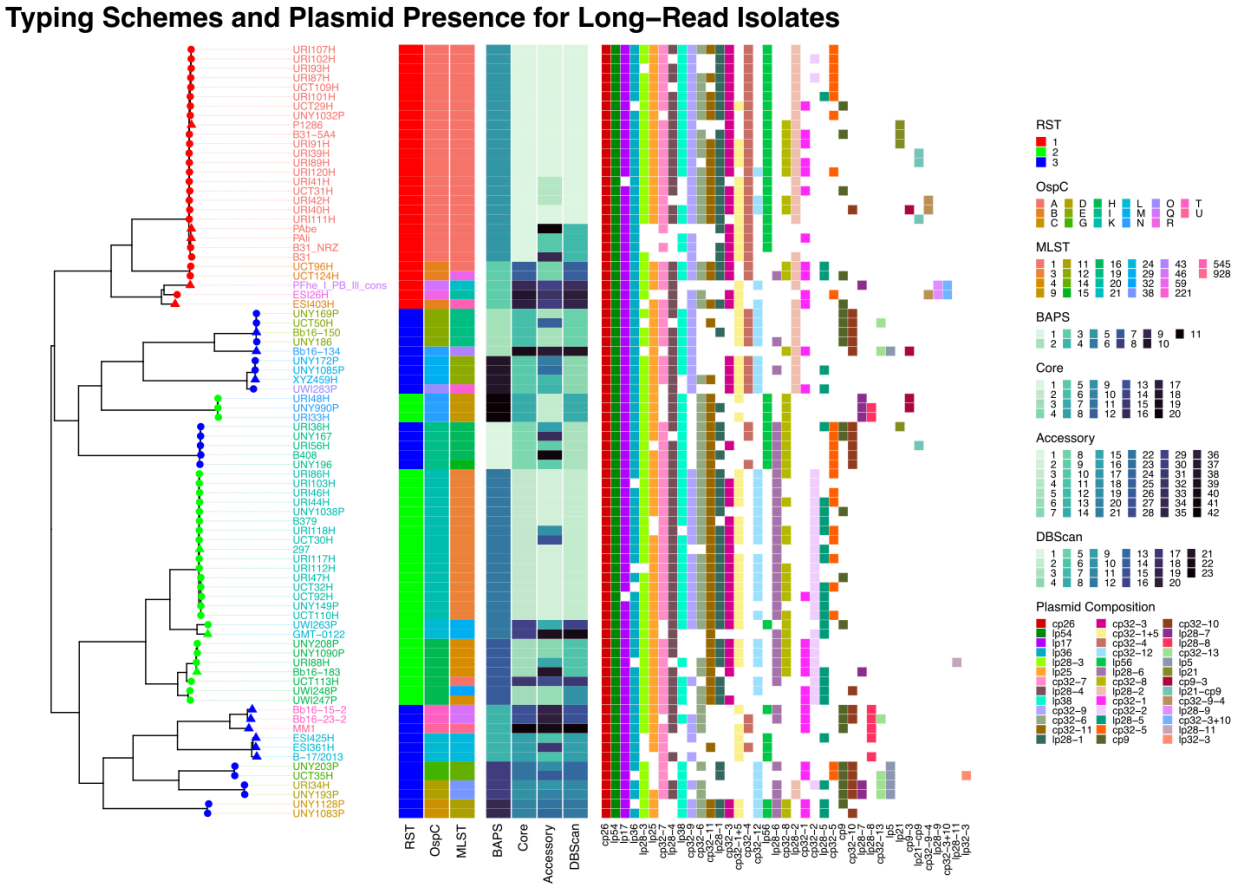


Figure 1: The phylogenetic structure of genetic markers and plasmid presence. A phylogenetic tree assembled from core genome sequences is shown. Tips are colored by RST and tip labels are colored by OspC type. To the right of the tree, the results of genetic marker and cluster annotation from several typing schemes (RST, OspC, MLST, BAPS, PopPUNK core clusters, PopPUNK accessory clusters, DBScan). Plasmids are annotated by color and ordered by prevalence.

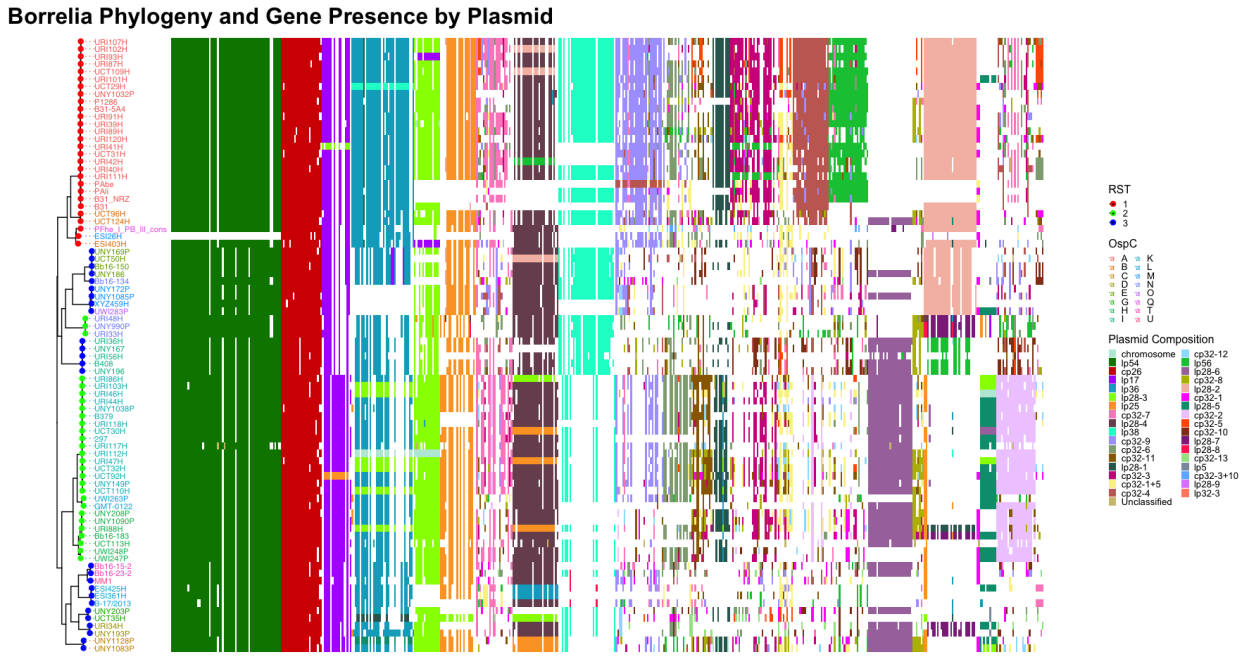
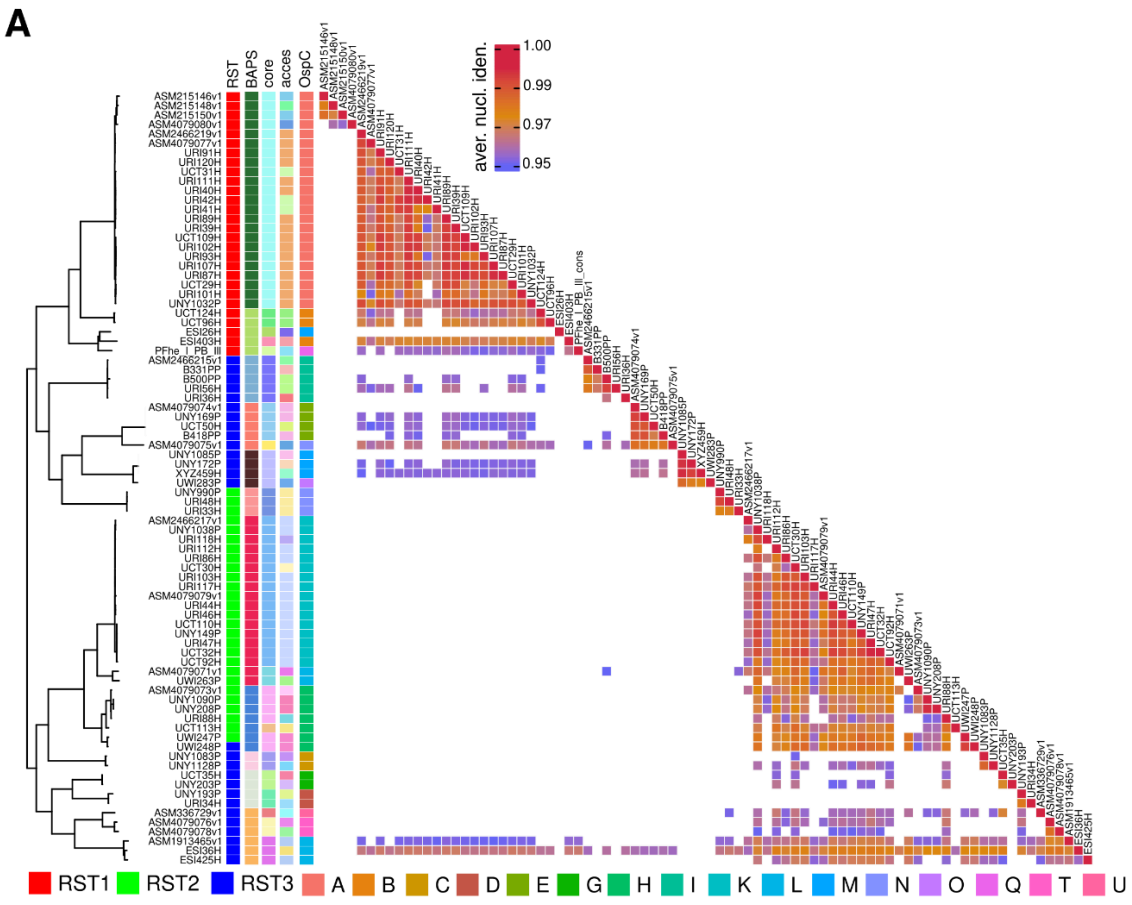


Figure 2: Phylogenetic structure and plasmid occupancy patterns of *Bb* orthologs. A phylogenetic tree assembled from core genome sequences is shown at left. Tips are colored by RST and tip labels are colored by OspC type. To the right of the phylogenetic tree, orthologs present in each isolate are shown in a matrix. For visual clarity, only orthologs present in 26 or more isolates are shown, with this threshold being selected due to the partitioning of isolates into 28 RST1, 28 RST2, and 26 RST3. Columns of the matrix are grouped by best hit replicon, with replicons ordered by presence rates across our sample. The plasmid encoding the ortholog is annotated by color.



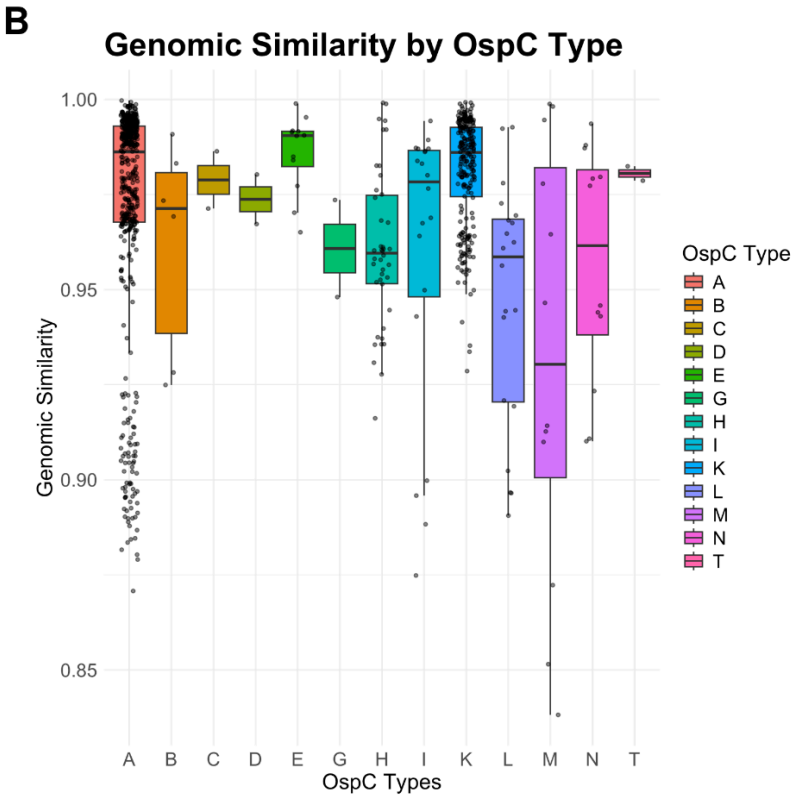


Figure 3: Matrix of pairwise average nucleotide identity, clustered by phylogeny. (A) A phylogenetic tree constructed from core genome sequences is shown at left. Genetic markers and cluster membership are shown next to the phylogenetic tree. At the right, the lower triangle portion of a matrix whose elements correspond to the pairwise calculation of average nucleotide identity (ANI) is shown. Each element in the matrix is colored by the ANI value. (B) The boxplots depict genomic similarity (full average nucleotide identity) within each OspC type.

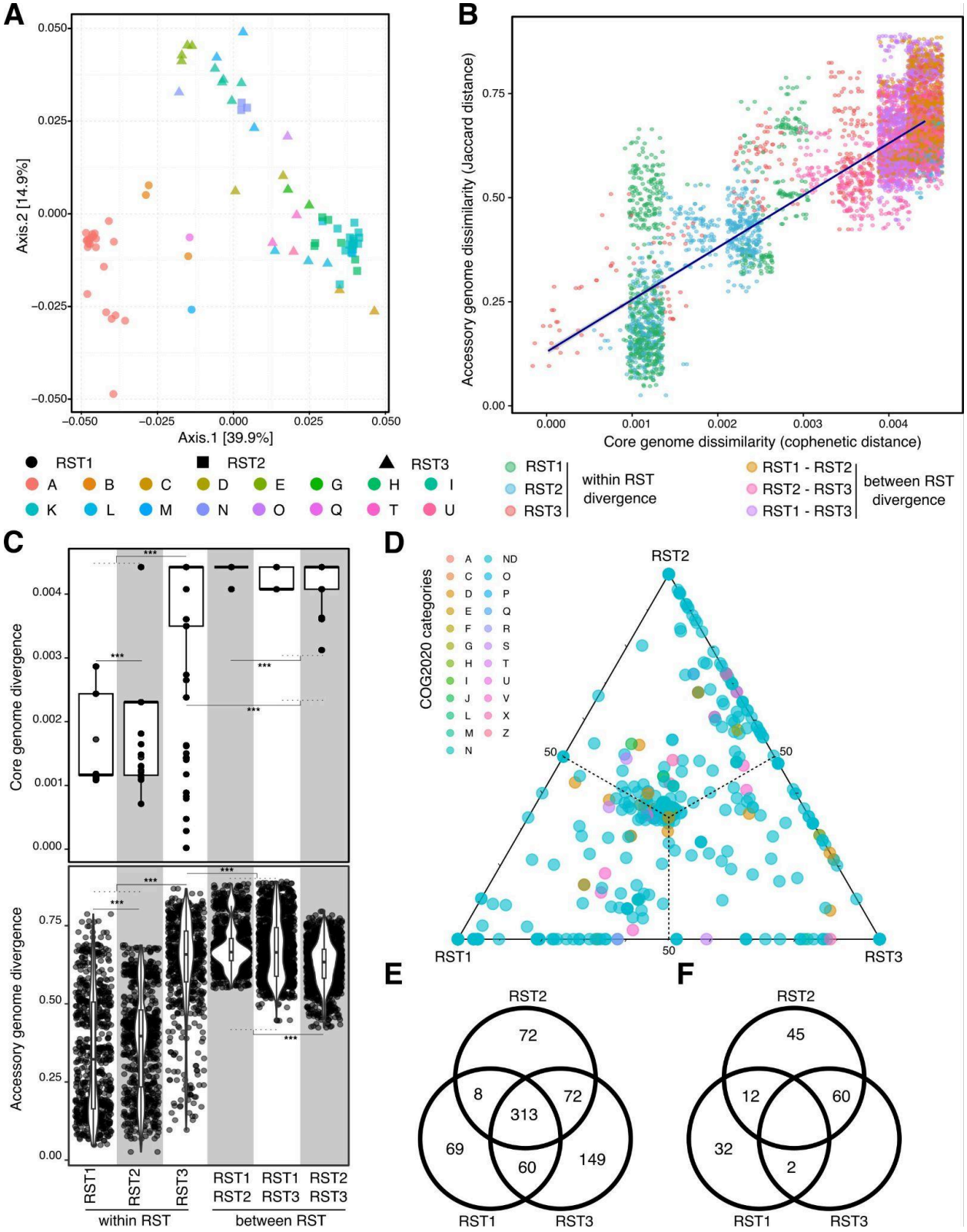


Figure 4: Genetic structure and patterns of divergence in the core and accessory *Bb* pangenome. (A) Principal component analysis shows the projection of Sorensen distances for the collection of isolates. (B) Core vs. accessory divergences. (C) Core (top panel) and accessory (lower panel) divergence grouped by RST. (D) COG2020 categories and their occupancy according to RST. (E and F) Venn diagrams of pangenome orthologs by RST (left panel) and pangenome orthologs with 100% prevalence in at least one RST (right panel). *** $p < 0.001$

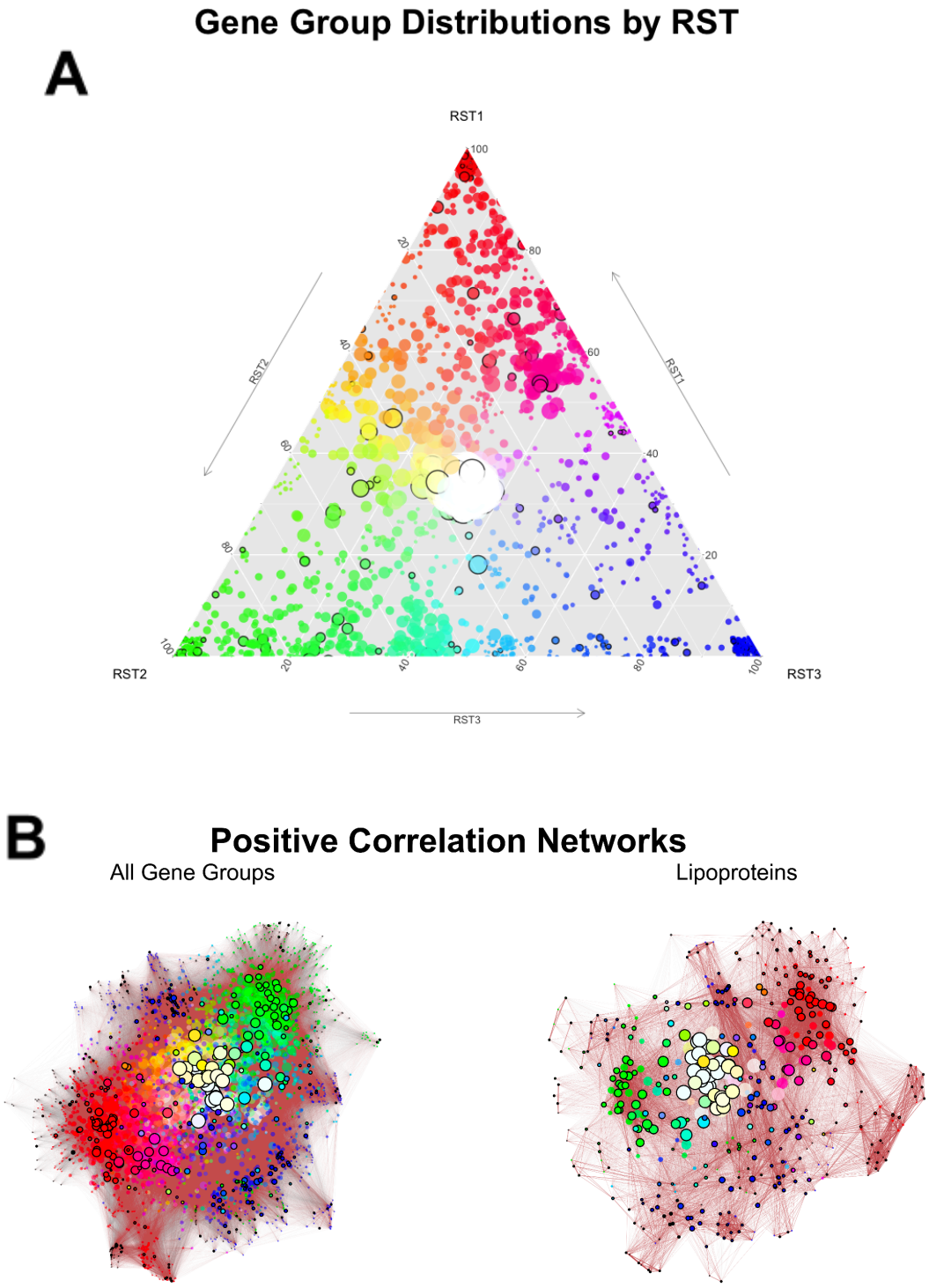


Figure 5: Gene group networks provide a structural view of variation by sequence type.

(A) Ternary plot of gene groups by RST types. Points are sized by gene frequency across the

sample, and coloring is determined with a 3-dimensional RGB color cube mapping based on gene distribution amongst RST types. Surface lipoproteins are circled in black. For visibility, a small jitter has been applied to the points. (B) Positive correlation networks for (i) all gene groups and (ii) gene groups associated with lipoproteins. Nodes are sized by gene frequency across the sample, and coloring is determined with a 3-dimensional RGB colorcube mapping based on gene distribution amongst RST types. Edge weights indicate correlation strength. Gene groups encoding for surface lipoproteins are circled in black.

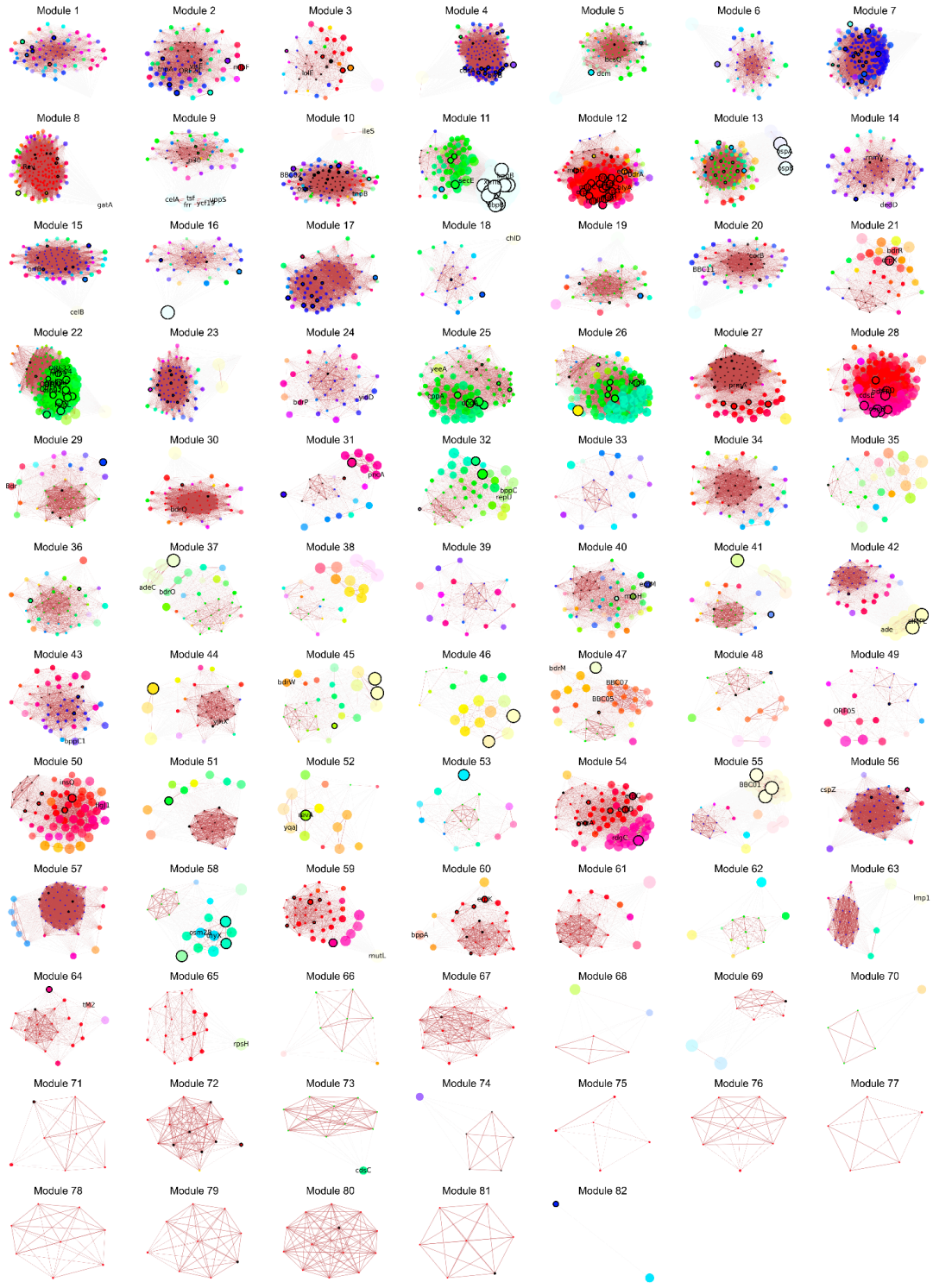


Figure 6: K-partite analysis reveals 82 distinct modules within the *Bb* genome. K-partite sets were determined using the negative correlation network, then induced subgraphs were created for each group of nodes using positive correlations.

Table 1

# Genes (% of isolates)	**Split paralog pangenome**	**No split paralog pangenome**	**Sig. level**	**P Value**	**Test**
Core (99%+)	801	823	***	5.16e-80	proportions z-test
Soft core (95-99%)	97	115	***	7.26e-13	proportions z-test
Shell (15%-95%)	961	491	**	0.049	proportions z-test
Cloud (<15%)	2747	661	***	4.12e-109	proportions z-test
Total	**4505**	**2090**	***	**3.53e-122**	**χ^2 test**

Table 1: Counts of ortholog groups in the pangenome, with and without splitting of paralogous gene families. The p value reports the test of hypothesis that counts of split and unsplit paralogs are the same, using a z test of proportions.

References

1. Kugeler KJ, Schwartz AM, Delorey MJ, Mead PS, Hinckley AF. Estimating the Frequency of Lyme Disease Diagnoses, United States, 2010-2018. Emerg Infect Dis. 2021;27:616–9.

2. Steere AC, Strle F, Wormser GP, Hu LT, Branda JA, Hovius JWR, et al. Lyme borreliosis. Nat Rev Dis Primers. 2016;2:16090.

3. Radolf JD, Strle K, Lemieux JE, Strle F. Lyme Disease in Humans. Curr Issues Mol Biol. 2021;42:333–84.

4. Arvikar SL, Steere AC. Lyme arthritis. Infect Dis Clin North Am. 2022;36:563–77.

5. Marques A. Persistent Symptoms After Treatment of Lyme Disease. Infect Dis Clin North Am. 2022;36:621–38.

6. Stanek G, Reiter M. The expanding Lyme Borrelia complex—clinical significance of genomic species? Clin Microbiol Infect. 2011;17:487–93.

7. Adeolu M, Gupta RS. A phylogenomic and molecular marker based proposal for the division of the genus Borrelia into two genera: the emended genus Borrelia containing only the members of the relapsing fever Borrelia, and the genus Borreliella gen. nov. containing the members of the Lyme disease Borrelia (*Borrelia burgdorferi* sensu lato complex). Antonie Van Leeuwenhoek. 2014;105:1049–72.

8. Margos G, Castillo-Ramirez S, Cutler S, Dessau RB, Eikeland R, Estrada-Peña A, et al. Rejection of the name Borreliella and all proposed species comb. nov. placed therein. Int J Syst Evol Microbiol. 2020;70:3577–81.

9. Stevenson B, Fingerle V, Wormser GP, Margos G. Public health and patient safety concerns merit retention of Lyme borreliosis-associated spirochetes within the genus Borrelia, and rejection of the genus novum Borreliella. Ticks Tick Borne Dis. 2019;10:1–4.

10. Wilske B, Preac-Mursic V, Göbel UB, Graf B, Jauris S, Soutschek E, et al. An OspA serotyping system for *Borrelia burgdorferi* based on reactivity with monoclonal antibodies and OspA sequence analysis. J Clin Microbiol. 1993;31:340–50.

11. Liveris D, Gazumyan A, Schwartz I. Molecular typing of *Borrelia burgdorferi* sensu lato by PCR-restriction fragment length polymorphism analysis. J Clin Microbiol. 1995;33:589–95.

12. Wang IN, Dykhuizen DE, Qiu W, Dunn JJ, Bosler EM, Luft BJ. Genetic diversity of ospC in a local population of *Borrelia burgdorferi* sensu stricto. Genetics. 1999;151:15–30.

13. Margos G, Gatewood AG, Aanensen DM, Hanincová K, Terekhova D, Vollmer SA, et al. MLST of housekeeping genes captures geographic population structure and suggests a European origin of *Borrelia burgdorferi*. Proc Natl Acad Sci U S A. 2008;105:8730–5.

14. Lemieux JE, Huang W, Hill N, Cerar T, Freimark L, Hernandez S, et al. Whole genome sequencing of human *Borrelia burgdorferi* isolates reveals linked blocks of accessory genome elements located on plasmids and associated with human dissemination. PLoS Pathog.

2023;19:e1011243.

15. Coipan EC, Jahfari S, Fonville M, Oei GA, Spanjaard L, Takumi K, et al. Imbalanced presence of *Borrelia burgdorferi* s.l. multilocus sequence types in clinical manifestations of Lyme borreliosis. Infect Genet Evol. 2016;42:66–76.

16. van Dam AP, Kuiper H, Vos K, Widjojokusumo A, de Jongh BM, Spanjaard L, et al. Different genospecies of *Borrelia burgdorferi* are associated with distinct clinical manifestations of Lyme borreliosis. Clin Infect Dis. 1993;17:708–17.

17. Balmelli T, Piffaretti JC. Association between different clinical manifestations of Lyme disease and different species of *Borrelia burgdorferi* sensu lato. Res Microbiol. 1995;146:329–40.

18. Grillon A, Scherlinger M, Boyer P-H, De Martino S, Perdriger A, Blasquez A, et al. Characteristics and clinical outcomes after treatment of a national cohort of PCR-positive Lyme arthritis. Semin Arthritis Rheum [Internet]. 2018; Available from: http://dx.doi.org/10.1016/j.semarthrit.2018.09.007

19. Wormser GP, Liveris D, Nowakowski J, Nadelman RB, Cavaliere LF, McKenna D, et al. Association of specific subtypes of *Borrelia burgdorferi* with hematogenous dissemination in early Lyme disease. J Infect Dis. 1999;180:720–5.

20. Wormser GP, Brisson D, Liveris D, Hanincová K, Sandigursky S, Nowakowski J, et al. *Borrelia burgdorferi* genotype predicts the capacity for hematogenous dissemination during early Lyme disease. J Infect Dis. 2008;198:1358–64.

21. Dykhuizen DE, Brisson D, Sandigursky S, Wormser GP, Nowakowski J, Nadelman RB, et al. The Propensity of Different *Borrelia burgdorferi* sensu stricto Genotypes to Cause Disseminated Infections in Humans. Am J Trop Med Hyg. 2008;78:806–10.

22. Jones KL, Glickstein LJ, Damle N, Sikand VK, McHugh G, Steere AC. *Borrelia burgdorferi* genetic markers and disseminated disease in patients with early Lyme disease. J Clin Microbiol. 2006;44:4407–13.

23. Jones KL, McHugh GA, Glickstein LJ, Steere AC. Analysis of *Borrelia burgdorferi* genotypes in patients with Lyme arthritis: High frequency of ribosomal RNA intergenic spacer type 1 strains in antibiotic-refractory arthritis. Arthritis Rheum. 2009;60:2174–82.

24. Strle K, Jones KL, Drouin EE, Li X, Steere AC. *Borrelia burgdorferi* RST1 (OspC type A) genotype is associated with greater inflammation and more severe Lyme disease. Am J Pathol. 2011;178:2726–39.

25. Fraser CM, Casjens S, Huang WM, Sutton GG, Clayton R, Lathigra R, et al. Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. Nature. 1997;390:580–6.

26. Tyler S, Tyson S, Dibernardo A, Drebot M, Feil EJ, Graham M, et al. Whole genome sequencing and phylogenetic analysis of strains of the agent of Lyme disease *Borrelia burgdorferi* from Canadian emergence zones. Sci Rep. 2018;8:10552.

27. Becker NS, Margos G, Blum H, Krebs S, Graf A, Lane RS, et al. Recurrent evolution of host and vector association in bacteria of the *Borrelia burgdorferi* sensu lato species complex. BMC

Genomics. 2016;17:734.

28. Margos G, Hepner S, Mang C, Marosevic D, Reynolds SE, Krebs S, et al. Lost in plasmids: next generation sequencing and the complex genome of the tick-borne pathogen *Borrelia burgdorferi*. BMC Genomics. 2017;18:422.

29. Hepner S, Kuleshov K, Tooming-Kunderud A, Alig N, Gofton A, Casjens S, et al. A high fidelity approach to assembling the complex Borrelia genome. BMC Genomics. 2023;24:401.

30. Combs M, Marcinkiewicz AL, Dupuis AP 2nd, Davis AD, Lederman P, Nowak TA, et al. Phylogenomic Diversity Elucidates Mechanistic Insights into Lyme Borreliae-Host Association. mSystems. 2022;7:e0048822.

31. Casjens SR, Di L, Akther S, Mongodin EF, Luft BJ, Schutzer SE, et al. Primordial origin and diversification of plasmids in Lyme disease agent bacteria. BMC Genomics. 2018;19:218.

32. Byram R, Stewart PE, Rosa P. The essential nature of the ubiquitous 26-kilobase circular replicon of *Borrelia burgdorferi*. J Bacteriol. 2004;186:3561–9.

33. Jewett MW, Byram R, Bestor A, Tilly K, Lawrence K, Burtnick MN, et al. Genetic basis for retention of a critical virulence plasmid of *Borrelia burgdorferi*. Mol Microbiol. 2007;66:975–90.

34. Bestor A, Stewart PE, Jewett MW, Sarkar A, Tilly K, Rosa PA. Use of the Cre-lox recombination system to investigate the lp54 gene requirement in the infectious cycle of *Borrelia burgdorferi*. Infect Immun. 2010;78:2397–407.

35. Casjens S, Murphy M, DeLange M, Sampson L, van Vugt R, Huang WM. Telomeres of the linear chromosomes of Lyme disease spirochaetes: nucleotide sequence and possible exchange with linear plasmid telomeres. Mol Microbiol. 1997;26:581–96.

36. Banik S, Terekhova D, Iyer R, Pappas CJ, Caimano MJ, Radolf JD, et al. BB0844, an RpoS-regulated protein, is dispensable for *Borrelia burgdorferi* infectivity and maintenance in the mouse-tick infectious cycle. Infect Immun. 2011;79:1208–17.

37. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015;31:3691–3.

38. Corander J, Waldmann P, Sillanpää MJ. Bayesian analysis of genetic differentiation between populations. Genetics. 2003;163:367–74.

39. Lees JA, Harris SR, Tonkin-Hill G, Gladstone RA, Lo SW, Weiser JN, et al. Fast and flexible bacterial genomic epidemiology with PopPUNK. Genome Res. 2019;29:304–16.

40. Hanincová K, Liveris D, Sandigursky S, Wormser GP, Schwartz I. *Borrelia burgdorferi* sensu stricto is clonal in patients with early Lyme borreliosis. Appl Environ Microbiol. 2008;74:5008–14.

41. Galperin MY, Wolf YI, Makarova KS, Vera Alvarez R, Landsman D, Koonin EV. COG database update: focus on microbial diversity, model organisms, and widespread pathogens. Nucleic Acids Res. 2021;49:D274–81.

42. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res. 2015;43:e15.

43. Erciyes K. Graph-Theoretical Analysis of Biological Networks: A Survey. Computation. 2023;11:188.

44. Voy BH, Scharff JA, Perkins AD, Saxton AM, Borate B, Chesler EJ, et al. Extracting gene networks for low-dose radiation using graph theoretical algorithms. PLoS Comput Biol. 2006;2:e89.

45. Wilske B, Jauris-Heipke S, Lobentanzer R, Pradel I, Preac-Mursic V, Rössler D, et al. Phenotypic analysis of outer surface protein C (OspC) of *Borrelia burgdorferi* sensu lato by monoclonal antibodies: relationship to genospecies and OspA serotype. J Clin Microbiol. 1995;33:103–9.

46. Lawrenz MB, Wooten RM, Norris SJ. Effects of vlsE complementation on the infectivity of *Borrelia burgdorferi* lacking the linear plasmid lp28-1. Infect Immun. 2004;72:6577–85.

47. Purser JE, Norris SJ. Correlation between plasmid content and infectivity in *Borrelia burgdorferi*. Proc Natl Acad Sci U S A. 2000;97:13865–70.

48. Lawrenz MB, Kawabata H, Purser JE, Norris SJ. Decreased electroporation efficiency in *Borrelia burgdorferi* containing linear plasmids lp25 and lp56: impact on transformation of infectious *B. burgdorferi*. Infect Immun. 2002;70:4798–804.

49. Jacobs MB, Norris SJ, Phillippi-Falkenstein KM, Philipp MT. Infectivity of the highly transformable BBE02^{-}lp56^{-}mutant of *Borrelia burgdorferi*, the Lyme disease spirochete, via ticks. Infect Immun. 2006;74:3678–81.

50. Frazão N, Konrad A, Amicone M, Seixas E, Güleresi D, Lässig M, et al. Two modes of evolution shape bacterial strain diversity in the mammalian gut for thousands of generations. Nat Commun. 2022;13:5604.

51. Weimann A, Dinan AM, Ruis C, Bernut A, Pont S, Brown K, et al. Evolution and host-specific adaptation of Pseudomonas aeruginosa. Science. 2024;385:eadi0908.

52. Castillo-Ramírez S, Fingerle V, Jungnick S, Straubinger RK, Krebs S, Blum H, et al. Trans-Atlantic exchanges have shaped the population structure of the Lyme disease agent *Borrelia burgdorferi* sensu stricto. Sci Rep. 2016;6:22794.

53. Takacs CN, Wachter J, Xiang Y, Ren Z, Karaboja X, Scott M, et al. Polyploidy, regular patterning of genome copies, and unusual control of DNA partitioning in the Lyme disease spirochete. Nat Commun. 2022;13:7173.

54. Biškup UG, Strle F, Ružić-Sabljić E. Loss of plasmids of *Borrelia burgdorferi* sensu lato during prolonged in vitro cultivation. Plasmid. 2011;66:1–6.

55. Faith DR, Kinnersley M, Brooks DM, Drecktrah D, Hall LS, Luo E, et al. Characterization and genomic analysis of the Lyme disease spirochete bacteriophage ϕBB-1. PLoS Pathog. 2024;20:e1012122.

56. Eggers Christian H., Kimmel Betsy J., Bono James L., Elias Abdallah F., Rosa Patricia, Samuels D. Scott. Transduction by φBB-1, a Bacteriophage of *Borrelia burgdorferi*. J Bacteriol. 2001;183:4771–8.

57. Schwan TG, Burgdorfer W, Garon CF. Changes in infectivity and plasmid profile of the Lyme

disease spirochete, *Borrelia burgdorferi*, as a result of in vitro cultivation. Infect Immun. 1988;56:1831–6.

58. Johnson J, Soehnlen M, Blankenship HM. Long read genome assemblers struggle with small plasmids. Microb Genom [Internet]. 2023;9. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10272865/

59. Akther S, Mongodin EF, Morgan RD, Di L, Yang X, Golovchenko M, et al. Natural selection and recombination at host-interacting lipoprotein loci drive genome diversification of Lyme disease and related bacteria. MBio. 2024;e0174924.

60. Casjens SR, Gilcrease EB, Vujadinovic M, Mongodin EF, Luft BJ, Schutzer SE, et al. Plasmid diversity and phylogenetic consistency in the Lyme disease agent *Borrelia burgdorferi*. BMC Genomics. 2017;18:165.

61. Schutzer SE, Fraser-Liggett CM, Casjens SR, Qiu W-G, Dunn JJ, Mongodin EF, et al. Whole-genome sequences of thirteen isolates of *Borrelia burgdorferi*. J Bacteriol. 2011;193:1018–20.

62. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.

63. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011;29:24–6.

64. Desjardins CA, Cohen KA, Munsamy V, Abeel T, Maharaj K, Walker BJ, et al. Genomic and functional analyses of Mycobacterium tuberculosis strains implicate ald in D-cycloserine resistance. Nat Genet. 2016;48:544–51.

65. Wick RR, Judd LM, Holt KE. Deepbinner: Demultiplexing barcoded Oxford Nanopore reads with deep convolutional neural networks. PLoS Comput Biol. 2018;14:e1006583.

66. Wick R. Porechop: adapter trimmer for Oxford Nanopore reads [Internet]. Github; [cited 2024 Dec 26]. Available from: https://github.com/rrwick/Porechop

67. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol. 2017;13:e1005595.

68. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100.

69. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 2014;9:e112963.

70. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol. 2019;20:257.

71. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29:1072–5.

72. Hanincova K, Mukherjee P, Ogden NH, Margos G, Wormser GP, Reed KD, et al. Multilocus sequence typing of *Borrelia burgdorferi* suggests existence of lineages with differential

pathogenic properties in humans. PLoS One. 2013;8:e73066.

73. Page AJ, Alikhan N-F, Carleton HA, Seemann T, Keane JA, Katz LS. Comparison of classical multi-locus sequence typing software for next-generation sequencing data. Microbial Genomics [Internet]. 2017;3. Available from: https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000124

74. Tonkin-Hill G, Lees JA, Bentley SD, Frost SDW, Corander J. RhierBAPS: An R implementation of the population clustering algorithm hierBAPS. Wellcome Open Res. 2018;3:93.

75. Schwengers O, Jelonek L, Dieckmann MA, Beyvers S, Blom J, Goesmann A. Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification: Find out more about Bakta, the motivation, challenges and applications, here. Microbial Genomics [Internet]. 2021;7. Available from: https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000685

76. Savojardo C, Martelli PL, Fariselli P, Casadio R. DeepSig: deep learning improves signal peptide detection in proteins. Bioinformatics. 2018;34:1690–6.

77. Setubal JC, Reis M, Matsunaga J, Haake DA. Lipoprotein computational prediction in spirochaetal genomes. Microbiology. 2006;152:113–21.

78. Eren AM, Kiefl E, Shaiber A, Veseli I, Miller SE, Schechter MS, et al. Community-led, integrated, reproducible multi-omics with anvi'o. Nat Microbiol. 2021;6:3–6.

79. Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One. 2010;5:e9490.

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

81. Dowdell AS, Murphy MD, Azodi C, Swanson SK, Florens L, Chen S, et al. Comprehensive Spatial Analysis of the *Borrelia burgdorferi* Lipoproteome Reveals a Compartmentalization Bias toward the Bacterial Surface. J Bacteriol [Internet]. 2017;199. Available from: http://dx.doi.org/10.1128/JB.00658-16

82. Toledo A, Crowley JT, Coleman JL, LaRocca TJ, Chiantia S, London E, et al. Selective association of outer surface lipoproteins with the lipid rafts of *Borrelia burgdorferi*. MBio. 2014;5:e00899–14.

83. Pritchard L, Glover RH, Humphris S, Elphinstone JG, Toth IK. Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. Anal Methods. 2016;8:12–24.

84. Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. MUMmer4: A fast and versatile genome alignment system. PLoS Comput Biol. 2018;14:e1005944.

85. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13:2498–504.