

Characterization of the family-level *Borreliaceae* pan-genome and development of an episomal typing protocol

Kayla M. Socarras

Center for Genomic Sciences, Institute for Molecular Medicine and Infectious Disease, Drexel University
College of Medicine

Mary C. Marino

Center for Genomic Sciences, Institute for Molecular Medicine and Infectious Disease, Drexel University
College of Medicine

Joshua P. Earl

Center for Genomic Sciences, Institute for Molecular Medicine and Infectious Disease, Drexel University
College of Medicine

Rachel L. Ehrlich

Nicholas A. Cramer

Department of Microbiology and Immunology, Virginia Commonwealth University Medical Center

Joshua C. Mell

Center for Genomic Sciences, Institute for Molecular Medicine and Infectious Disease, Drexel University
College of Medicine

Bhaswati Sen

Center for Genomic Sciences, Institute for Molecular Medicine and Infectious Disease, Drexel University
College of Medicine

Azad Ahmed

Center for Genomic Sciences, Institute for Molecular Medicine and Infectious Disease, Drexel University
College of Medicine

Richard T. Marconi

Department of Microbiology and Immunology, Virginia Commonwealth University Medical Center

Garth D. Ehrlich

ge33@drexel.edu

Center for Genomic Sciences, Institute for Molecular Medicine and Infectious Disease, Drexel University
College of Medicine

Keywords: Spirochete, distributed genome hypothesis, pan-genome, supragenome, comparative genomics, phylogenetics, *Borrelia*, *Borrelia*, *Borrelia*, *Borrelia*, *Borrelia*, taxonomy

Posted Date: June 11th, 2024

DOI: <https://doi.org/10.21203/rs.3.rs-4491589/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

The *Borreliaceae* family includes many obligate parasitic bacterial species which are etiologically associated with a myriad of zoonotic borrelioses including Lyme disease and vector-borne relapsing fevers. Infections by the *Borreliaceae* are difficult to detect by both direct and indirect methods, often leading to delayed and missed diagnoses. Efforts to improve diagnoses center around the development of molecular diagnostics (MDx), but due to deep tissue sequestration of the causative spirochaetes and the lack of persistent bacteremias, even MDx assays suffer from a lack of sensitivity. Additionally, the highly extensive genomic heterogeneity among isolates, even within the same species, contributes to the lack of assay sensitivity as single target assays cannot provide universal coverage. This within-species heterogeneity is partly due to differences in replicon repertoires and genomic structures that have likely arisen to support the complex *Borreliaceae* lifecycle in which these parasites have to survive in multiple hosts each with unique immune responses.

Results

We constructed a *Borreliaceae* family-level pangenome and characterized the phylogenetic relationships among the constituent taxa which supports the recent taxonomy of splitting the family into at least two genera. Gene content profiles were created for the majority of the *Borreliaceae* replicons, providing for the first time their unambiguous molecular typing.

Conclusion

Our characterization of the *Borreliaceae* pan-genome supports the splitting of the former *Borrelia* genus into two genera and provides for the phylogenetic placement of several non-species designated isolates. Mining this family-level pangenome will enable precision diagnostics corresponding to gene content-driven clinical outcomes while also providing targets for interventions.

Background

Our development of the distributed genome hypothesis (1–4) led directly to the concept of the bacterial species-level supragenome/pan-genome (5–16). These types of analyses have been extended to the genus level (17–21), and recently to the family level (22, 23). The identification of novel distributed genes and their association with phenotypic characteristics provides for precision diagnostics and targeted characterization of components of the genomic dark matter (19, 24, 25). Despite the progress made in comparative bacterial genomics, there have been relatively few reports examining the agents of spirochetoses including the *Borreliaceae* (26, 27).

Spirochetoses are characterized as chronic, invasive, debilitating diseases caused by parasitic bacteria within the taxonomic order spirochaetales. Among these diderm helical microbes, the most genomically complex in terms of replicon numbers and heterogeneity are those in the *Borreliaceae* family. Due to climate change and other anthropomorphic environmental disruptions, the *Borreliaceae* have greatly expanded beyond their traditional geographic epidemiological zones and have developed diverse symbiotic interactions with multiple alternative hosts (28). The true taxonomy of the *Borrelia* genus has been a subject of interest due to the genomic and pathogenic disparities among its member species. A recent taxonomic revision led to the proposed division of the *Borrelia* genus into two distinct genera, the *Borrelia* and *Borrelia* (29). In this revised classification, the relapsing fever agents retain the original *Borrelia* taxonomy due to their priority of discovery and the Lyme borreliosis agents became known as *Borrelia* spirochetes. This work did not include a novel subset of isolates known as the reptile-associated *Borrelia*, which to date have not been well characterized.

It is vital to note that this reclassification has not been without controversy (30–32). In the years since the original proposal, some authors have argued that data other than differences in average nucleotide identity (ANI) do not support the split (30). However, these latter authors were themselves criticized for not fully examining the multi-omic data that provided extensive quantitative data in support of the division (33, 34), thus the controversy continues.

Early genomic studies on *Borreliaceae* spirochetes revealed that they possess a complex genome of a single, linear chromosome terminating in covalently bound hairpin ends, along with multiple single-copy episomes that exist as circular or hairpinned-linear molecules (35, 36). While the chromosome contains conserved genes linked to essential functions and survival, the individual episomal elements vary widely in size, structure (linear and circular), and gene content, and facilitate *Borrelia* host adaptation and/or virulence. Historically, the smaller replicons were found to not be as conserved and were not always retained in long-time cultures. The smaller episomes that were retained or found biologically important were principally categorized by molecular size and structure with only minimal use of gene content.

Despite these limitations, previous comparative genomic studies led to the establishment of a *B. burgdorferi sensu lato* pangenome (37). This study utilized 23 *Borrelia* genomes, most of which were *B. burgdorferi*, and were of varying assembly quality and plasmid content. This first attempt to create a genus-level pangenome showed that additional isolates would be necessary to understand the degree of variation among *Borrelia*. However, until recently, there have been very few pan-genomic publications on these clinically important taxa (38, 39). Moreover, there has not been a pangenome constructed for the *Borrelia* genus nor for the entire *Borreliaceae* family. It is essential to better understand these pathogens through the construction of a family-level pangenome using the latest error-correcting long-read DNA sequencing methodologies to ensure capture and single-contig assembly of all genomic elements.

Results

Study design

All available complete or near-complete high-quality *Borreliaceae* genomes, as of the commencement of this study (n = 69) with sequencing coverage greater than 30X and an average N50 of 900 Kb or above were downloaded from the NCBI prokaryotic genome database and subjected to additional QC analyses for completeness. These genomes were then combined with 39 newly sequenced *Borreliaceae* genomes chosen to fill in gaps within the taxonomic coverage of the family. All pan-genome analyses were conducted from this final curated database of 108 strains (Table 1).

Pacific Biosciences Sequel I whole-genome sequencing and validation

A modified gel-plug DNA isolation and pulse field gel electrophoresis (PFGE) analysis was performed on selected *Borreliaceae* isolates (*Borrelia burgdorferi* strain B31 and *Borrelia hermsii* strain HS1) to determine the number of replicons and to serve as a control for the extraction and sequencing of the multiple replicons that make up the *Borreliaceae* genomes (40). Using the same DNAs, we performed whole genome sequencing on a Pacific Biosciences (PacBio) Sequel I using single molecule error-correcting circular consensus sequencing. Quality assurance measures of the WGS's included assembling the linear chromosomes into a single contig ≥ 0.9 Mb and identifying each of the other assembled contigs (9 Kb to 200 Kb) as corresponding to one of the PFGE replicons.

We first performed comparative analyses of our lab-sequenced genomes of the *Borrelia burgdorferi* B31 and *Borrelia hermsii* HS1 strains with their respective NCBI reference genomes to determine if all replicons of both *Borreliaceae* species representing both genera were present and fully sequenced (41, 42). We confirmed that our whole genome sequencing was comparable to previous work by aligning a reference and our sequence of *B. hermsii* strain HS1 (Fig. 1a). Using a progressive Mauve alignment of the reference strain and our sequenced *B. hermsii* strain HS1d, multiple locally colinear blocks (LCB) were noted across the entire length of the genome. The largest LCB showed high homology between our sequence and the reference for the linear chromosome and the large linear plasmid. The linear chromosomes contained 100 single nucleotide polymorphisms (SNPs) and the large linear plasmids had 5 SNPs. The smaller contigs from both genomes were either partially aligned or had no clear matches, likely in part due to progressive Mauve being unable to track duplications and the shared gene content between some plasmids (43). The estimated total number of SNPs between the two HS1 genomes was ~ 1,000. These differences likely stem from small duplications near the telomeres of the linear plasmids since progressive Mauve is ill-suited to handle them. Additionally, novel plasmids within the newly sequenced *B. hermsii* strain HS1d may have arisen from recombination events to which *Borreliaceae* plasmids are prone (38, 44), or alternatively due to plasmid loss during the culturing of the original HS1 isolate. A second alignment using D-Genies corroborated these findings (Fig. 1b).

Finally, whole genome alignments of the reference *B. hermsii* strain HS1 and our sequenced *B. hermsii* strain HS1d confirmed their similarity (Fig. S1) but also demonstrated that they are not identical,

particularly with respect to the episomal elements. Some differences in the episomal elements may be accounted for by genes encoding proteins undergoing antigenic variation or by the variable cassettes which drive genetic conversion.

Coverage of covalently bound hairpin ends of *Borreliaceae* episomes

In our alignments, several reference *B. hermsii* strain HS1 plasmids exhibited homology to the middle of our sequenced HS1d contig pairs (Fig. 1a). This could be an artifact of how these contigs were parsed and assembled. We identified short flanking inverted repeats at the ends of many HS1d contigs that were not included in an LCB with the HS1 reference (Fig. 1 and Fig. 2). These inverted repeats stem from the hairpin present within the ends of all linear *Borreliaceae* replicons or because of artifacts of long-read sequencing (35, 38, 45).

To confirm which regions of the *B. hermsii* strain HS1 and HS1d plasmids were inverted, we paired the contigs and aligned them using the D-Genies program (Fig. 1b). This analysis identified regions with inverted repeats present at the very ends of some HS1d contigs that were termini of linear replicons. Further inspection of the *B. hermsii* strain HS1d contigs using minimap2 and subsequent plotting with pafr confirmed this artifact on both ends of suspected linear plasmids (Fig. 2) (46, 47).

The presence of these inverted repeats results from a specific artifact that arises due to a combination of the actual hairpin topology of linear *Borreliaceae* replicons and the stem-loop adapters used for PacBio long-read sequencing (Fig. 3) (35, 38, 45). The heterogeneous ends of these contigs were closely inspected for the presence of the telomere resolvase (ResT) Box 3 TATA motif found in *Borrelia* species (Fig. 3) (48, 49). These motifs were identified solely on linear *B. hermsii* strain HS1d contigs as well as within our sequenced *Borreliaceae* cohort. To ensure that the inverted repeat sequence in each contig for all the *Borreliaceae* assembly would not skew downstream comparative genomic analyses, each contig that contained a ResT Box 3 motif was trimmed down.

Average Nucleotide Identity Analyses Support a Digeneric *Borreliaceae*

The ANI among all *Borreliaceae* genome assemblies were computed using the pyani average nucleotide identity blast method (BLASTN+). These analyses revealed that the species members of the original single *Borrelia* genus were clearly separated into two groups corresponding to the proposed (50) *Borrelia* and *Borrelia* genera (Fig. 4). Thus, our findings concur with the division of the former *Borrelia* genus into two genera and henceforth we will refer to the Lyme borreliosis-causing spirochetes as *Borrelia* and the relapsing fever-causing spirochetes as *Borrelia*. Additionally, we included in the analysis two non-species designated *Borrelia* spirochetes, B_sp_HM_thM16W and B_sp_FGy1, which are thought to be part of the reptile-associated group (51–55) that were demonstrated to be *sui generis*, but clustered more closely with the *Borrelia* genus than the *Borrelia*.

Further inspection of the 65 *Borrelia* strains included in the analyses, covering nine named species, demonstrated that the vast majority have been placed within the correct taxa with essentially all pairwise

strain analyses within a species showing > 96% ANI (Fig. 5). All the *B. burgdorferi* and *B. afzelii* were correctly assigned. One of the two strains typed as *B. bavariensis* clustered within the *B. garinii*. The other typed with one of the two strains typed as *B. valaisiana*, forming a small *sui generis* group which may correspond to a hybrid species between *B. garinii* and *B. afzelii* as the hierarchical clustering placed them between the proposed parent species. These findings are in line with previous reports that had suggested that *B. bavariensis* was a subspecies of *B. garinii*. *B. maritima* was equidistant between *B. afzelii* and *B. burgdorferi*. In addition, we characterized the phylogenetic relationship of several novel isolates included in this study. These included the lagomorph isolate, *Borrelia andersonii*, which at the time of its discovery was misclassified as a *B. burgdorferi* strain (56) but was later reclassified as *B. andersonii* (57). In our ANI analyses, the *B. andersonii* strain MOK_3a clustered closely to *B. burgdorferi* and *B. maritima*. Of the remaining three *Borrelia* species (*B. andersonii*, *B. bissetii*, and *B. mayonii*), all were most closely related to *B. burgdorferi* but each were sufficiently distant to warrant their species designations.

Compared to the *Borrelia*, the *Borrelia* spirochetes displayed greater genetic heterogeneity. The *B. miyamotoi* strains clustered strongly together, displaying an average pairwise ANI of 97% (Fig. 6) but were distinct from the cluster of species containing *B. hermsii*, *B. parkeri*, and *B. turicatae*. The reptile-associated *Borrelia* species were both remarkably distant from these other major relapsing fever spirochetes and highly divergent among themselves with only B_sp_FGy1 and *B. turcica* strains sharing any substantial similarity at 94% (Fig. 7). Through ANIb (blast), we determined that these two genomes bracket the *Borrelia* genera (Fig. 4). The isolate B_sp_HM_thM16w was also divergent from the group and was closest to *B. recurrentis* (Fig. 4, 6).

Following our initial ANI analysis, the evolutionary relatedness of these spirochetes was checked further by constructing a family phylogeny (Fig. 7). This revealed both a division amongst the *Borreliaceae* spirochetes and the same clustering for *Borrelia* and *Borrelia* spirochetes. Interestingly, we observed that some more recently identified *Borrelia* and *Borrelia* isolates cluster within their respective genera. This is most clearly illustrated with *B. andersonii* which, based on the single-copy core phylogeny, was most closely related to *B. burgdorferi* and *B. bissetii*. Other novel genomes, like those seen within the reptile-associated *Borrelia* spirochetes, stood apart from the majority of the *Borrelia* genus as had been previously seen in the ANI. Only the novel *B. sp strain thHMw* intercalated within *B. miyamotoi* and *B. hermsii* strains. The rest of the reptile-associated spirochetes formed a separate clade within *Borrelia* spirochetes.

Borreliaceae family-level pangenome

Following our initial ANI analyses, we wanted to understand how similar the genera were based on their overall gene similarities. The percentage of shared genes within the family was approximately 50% (Fig. 9). To begin to understand which genes are shared between the different genera, the *Borreliaceae* family-level pangenome was multiply calculated at 5% intervals of increasing similarity (Fig. 9). At 75% BlastP, clear separation between the core and cloud components of the *Borreliaceae* pangenome is seen

(Fig. 9). At the 75% BlastP threshold, there are 322 core genes, 8 soft-core genes, 1,352 genes in the shell, and 3,424 genes in the cloud. There are 1,253 unique genes for all 109 isolates included in this study. Interestingly, the *Borreliaceae* pangenome had a different composition when analyzed with the EggNOG algorithm. Through this second approach, EggNOG estimates that *Borreliaceae* contains 997 unique genes for 109 isolates. Furthermore, its pangenome distribution consists of 577 core, 30 soft-core, 204 shell, and 213 cloud genes. Although each of these pangenome analysis tools were given the same dataset, they yielded different results. These discrepancies are due to underlying differences in the alignment methods used by the algorithms. EggNOG relies upon alignment to a database of proteins which have been pre-clustered, rather than relying on a specific blast threshold. Additionally, EggNOG largely ignores genes of unknown function rather than labeling it as “unknown,” which artificially decreases the number of unique genes.

Previously, a *Borreliella* genospecies complex pangenome was constructed through the use of 22 genomes comprised of *B. burgdorferi*, *B. afzelii*, and *B. garinii* (37). In this work, they found that the *Borreliella* pangenome was open. To ascertain if this was true within our own curated collection of *Borreliaceae* genomes, we next tested for changes in the pangenome using both data acquired from Roary at 75% BlastP and EggNOG (Fig. 10). It is evident that the number of gene clusters present within each pangenome compartment is subject to change at the incorporation of a novel pangenome. This suggests that the family-level pangenome is open with a growing distributed component reminiscent of what was found within the early *Borreliella* comparative genomic study.

Next, we visualized the *Borreliaceae* pangenome using a heat map for all included isolates (Fig. 11). The family-level core genome consists of 488 genes possessed by all isolates in both genera. Interestingly, the family-level distributed pangenome is fragmented into multiple components. This division reveals two additional groups of genes that on further inspection were identified as the core genes of either the *Borreliella* or *Borrelia* genera.

The *Borreliaceae* episomes can be grouped through gene content

Borreliaceae spirochetes thrive throughout their complex lifecycles during which they undergo multiple transmissions between phylogenetically distinct hosts including arthropod vectors and multiple classes of vertebrates. Their ability to survive in various hosts is dependent upon a large set of distributed genes that are mostly encoded on episomal replicons. These episomes have traditionally been typed primarily based on molecular size, structure (linear or circular), and alleles of proteins in the PFam32 family (58). To ensure proper episomal identification, this system needs to be expanded to include information on the entire gene repertoire of all available episomes. To this end, we ran a pyani ANIb on approximately 1,800 *Borreliaceae* replicons from our curated genome collection. As pyani ANIb may not factor all gene content within the replicons, we also employed a Hadamard matrix (which interprets identity and coverage simultaneously) for this purpose. In the Hadamard matrix heatmap, multiple subgroups for nearly all of the *Borreliaceae* family replicons were identified (Fig. 12).

The most similar episomes clustered closely together. The three most conserved groups are the *Borrelia* chromosome, cp26, and lp54. Other *Borrelia* episomal groups with relatively higher percentages of similarity were lp25 and lp36. Interestingly, there was very broad and diffuse clustering for the clinically significant family of cp32 plasmids which are known to contain genes encoding host-specific proteins, indicating both their relatedness and the very high degrees of heterogeneity necessary to ensure survival during a lifecycle that includes multiple hosts. Unlike cp32, the lp28 episomes do not likely represent a single family of replicons as they did not all cluster together and instead separated out into smaller clusters.

For the *Borrelia* genus, a high degree of similarity was seen amongst the chromosomes as well as for the large linear plasmids. Additionally, species-specific groupings amongst the episomes were also observed. This is best illustrated in *B. miyamotoi* strains lp6 and lp12. It is important to note that while the reptile-associated *Borrelia* spirochetes were included in these analyses, they were the only isolates to have extraordinarily dissimilar replicons. This is likely attributable to the incompleteness of the B_sp_A_FGy1 genome as indicated by its large number of constituent contigs.

Discussion

The debilitating diseases caused by *Borreliaceae* spirochetes have remained on the periphery of societal memory for centuries. Despite our current knowledge of the etiological agents of these spirochetoses, there is still a paucity of information on how each disease is caused and why their presentations are so variable among individuals. In the work presented in this study, multiple pressing points within the field of *Borreliaceae* comparative genomics were addressed to help provide a framework going forward to be able to associate microbial genotypes with disease phenotypes in the host.

The first topic we addressed was the composition of the *Borreliaceae* family. It was recently proposed that the previous single genus, *Borrelia*, be divided into two genera based on differences in overall genomic content (29, 34). The well-studied and more widely known Lyme borreliosis spirochetes were given the genus name of *Borrelia*, while the older and less clinically studied species complex comprising the relapsing fever spirochetes retained the *Borrelia* designation. Since the establishment of this proposed dichotomy, there has been animated discussion among scientists both for and against the split (30–34).

Pairwise ANI analyses of all 108 sequenced isolates (N = 11556 genome comparisons) in this study revealed a sharp division between the two proposed genera indicating that the recent taxonomic reclassification is correct. Even isolates from within each of the genera often did not share sequence similarities above 85%. While it may be true that the ANI cut-off can vary based on species as seen with *Stenotrophomonas maltophilia*, *Escherichia* spp, or *Lactobacillus* spp., our results do not support including the Lyme borreliosis spirochetes and relapsing fever spirochetes within a single genus (59, 60).

Additional analyses revealed that the recently sequenced novel spirochaetal species B_sp_FGy1 and B_sp_HM_thM16w clustered most closely with *B. anserina* and *B. recurrentis*, respectively. These two

genomes clustered within the known reptile- associated *Borreliaceae*. As these isolates are novel, it is unknown if they are pathogenic to humans.

We constructed a family-level pan-genome to characterize the similarities and differences between the two genera concerning gene content. In doing so, we found that the family-level *Borreliaceae* core genome was very small, consisting of only 488 genes, but that each of the genus-level core genomes were each more than double that size and that the family-level pangenome is still open. This indicates that there are large numbers of distributed genes still to be discovered. The number of core genes possessed by both genera were similar.

It was previously known that there is some similarity between *Borrelia* and *Borrelia* replicons with the *Borrelia* replicons containing regions of similarity to the *Borrelia* cp26 and lp54 (27, 44, 61, 62). To better place the various replicons within a gene-functional framework, we established a system for typing the *Borreliaceae* plasmids. Previously, these episomes have been typed based on size, structure, and the presence of specific alleles of particular protein families found on many of the plasmids (26, 58). While this technique helped somewhat in the identification of the many *Borreliaceae* episomes, it did little to identify them by the biological functions they encode. In our work, all of the available episomes for each of the isolates in our data set were used in an ANI analysis via pyani. Doing so for over 1,800 replicons resulted in a complex matrix that did little to clarify how these episomes related to each other (Fig. 12). However, in building a Hadamard matrix from these analyses it was possible to visualize each of the known plasmids with respect to their overall hierarchical clustering, even for the most heterogenous replicon, cp32. Thus, we have established a universal *Borreliaceae* plasmid typing system based on gene content that captures all of the replicons produced via our PacBio-based long-read sequencing protocols.

Finally, we identified an artifact of sequencing linear bacterial replicons with hairpin termini with the PacBio SMRTbell technology. In our sequencing, it was noted that some of the many linear episomes had regions of heterogeneity. After closer inspection, it became apparent that these regions were inverted repeats that could emerge as artifacts in long-read sequencing platforms (38, 63). These regions were identified as the telomeric ends of all linear *Borreliaceae* replicons, a feature not unlike those seen in higher eukaryotes (41, 45, 64, 65). To prevent false inflation in the genes within our data set, we searched for and confirmed the location of the ResT Box 3 site (49, 65). From this, we have begun to build the first system to handle these *Borreliaceae* long-read sequencing artifacts that could be applied to any replicon with similar features. These data also suggest that raw PacBio circular consensus sequences may contain hidden information about hairpin and cruciform DNA structures in other genomes.

This work on the *Borreliaceae* family-wide pangenome is pivotal and the first of its kind. More sequencing is required to ensure that the pangenome is truly reflective of all *Borreliaceae* spirochetes. Furthermore, more work is required for the typing of all *Borreliaceae* episomes. This family-level pangenome analysis provides new tools and information in a field that requires more insight on how to

handle such evasive and persistent pathogens. More importantly, this work can later culminate into effective therapeutics.

Conclusions

Through the application of multiple comparative genomic methods including phylogenetics, ANI, gene content, and core genome analyses, we demonstrate that the *Borreliaceae* are composed of at least two genera. We also developed a gene-based plasmid typing protocol to replace the current confusing method that relies solely on size and topography. This approach to episomal typing revealed an extremely high degree of allelic heterogeneity in the plasmid(s) that contain the genes that encode the major host-interacting proteins which are reflective of the extremely broad host-range of these parasitic bacteria. Additional analyses with a larger number of genomes will be required to determine if the reptile-associated strains and other underrepresented groups in the current analysis define additional genera.

Methods

Borreliaceae pangenome project design

This study included all available reference *Borreliaceae* genomes obtained from NCBI RefSeq on October 17th, 2021. The downloaded genomes were checked for completeness, sequencing platform, sequencing coverage, and if the average number of contigs greatly superseded the number of included replicons. For example, selected genomes were disqualified if more than 2 or 3 contigs were needed for a single replicon. Additionally, they were run through a QC pipeline to ensure usability in later studies. The remainder of the genomes were used henceforth. Novel isolates collected by collaborators at Virginia Commonwealth University were then sequenced at the Center for Genomic Sciences core facility at Drexel University College of Medicine. These newly sequenced members of the *Borreliaceae* are included in BioProject PRJNA861274.

Borreliaceae culturing

All *Borrelia* and *Borrelia* isolates were cultivated in 50 mL of BSK-II media supplemented with 6 or 12% rabbit serum (Sigma, MA), respectively. The culture was then monitored using wet-mounts and dark-field microscopy. Once spirochetes grew to the concentration of 5×10^7 cells/mL at late log phase, they were harvested by centrifugation, resuspended in 1X sterile PBS, and then recovered by centrifugation.

Pulse-Field Gel plasmid verification

DNA isolation of selected *Borreliaceae* isolates was done through a modified BioNano plug prep (Bionano Genomics, Bionano Prep cell Culture DNA Isolation Protocol) (Bionano Genomics, CA). Cultured *Borreliaceae* were pelleted at 5,000 xg for 10 mins and sent overnight to the Center for Genomic Sciences at Drexel University College of Medicine. Received pellets were immediately thawed, resuspended with a total volume of 66 μ L of cell buffer, and then 40 μ L of 2% agarose was added. The

contents were mixed 10 times via pipette and then 100 μ L of each sample was aliquoted into a plug mold. The filled plug mold was incubated at 4°C for 15 mins before starting protein digestion. This process was done by placing the plug into a solution of 167 μ L proteinase K (Qiagen, DE) and 2.5 mL Lysis buffer, then incubated for 2 hours at 50°C in a thermomixer at intermittent mixing. After incubation, the digestion solution was replaced and incubated overnight on the thermomixer as described above.

Following the overnight incubation, plugs were allowed to cool at room temp for 5 mins and then we added 50 μ L of RNase A before placing the plugs in the thermomixer at 37°C with intermittent mixing. Plugs were then washed 3 times with 10 mL 1X Wash buffer before being washed once more with 10 mL and set on an orbital shaker for 15 mins at 180 RPM. This latter step was repeated 3 more times. DNA recovery was initiated by adding 10 mL of 1X TE buffer to wash plugs and shaking at 180 RPM for 15 mins a total of 5 times. Cleaned plugs were then inserted into a 1% low melt agarose gel and run on a pulse-field gel at 5V for 21 hrs. Pulse field gel was stained with ethidium bromide and visualized under UV.

In silico Borreliaceae replicon verification

The *Borreliaceae* genome assembly pairs were aligned through progressiveMauve (v.2.4.0), NUCmer (MUMer 3.0) and D-Genies (v1.4.0) (43, 66, 67). Additionally, each spirochete genome FASTA file that had been used for the above initial validation work was split by contig, imported into R, and then aligned to itself using pafr (v 0.0.2) (46).

Borreliaceae DNA extraction

Once received, frozen *Borreliaceae* pellets were thawed on ice and centrifuged at 13,000 rpm for 1 min. Following pelleting, the remaining solvent was discarded before the pellets were resuspended in 200 μ L of 1X PBS, pipette mixed, and transferred to a 2 mL bead beating tube (Matrix E) (MP Biomedicals, CA). Following the transfer, 20 μ L of Proteinase K (Qiagen, DE) was added, and cells were homogenized using a SPEX 1600 MiniG (Fisher Scientific, MA) for 1 min at 1500 Hz. Subsequent DNA extraction was performed using the Qiagen™ DNeasy Blood & Tissue Kit according to the manufacturer's instructions (Qiagen, DE).

DNA preparation and sequencing

Extracted *Borreliaceae* DNA was quantified following ThermoFisher Scientific™ 1X dsDNA HS kit per the manufacturer's instructions (Biotium Inc, CA) on Qubit. DNA from each Borrelial/Borrelia specimen was prepped with the SMRTbell Template Prep Kit 2.0 (Pacific Biosciences, CA) to make PacBio SMRTbell libraries with barcodes sourced from the Barcoded Overhang Adaptor Kit 8A and 8B (Pacific Biosciences, CA). The sequencing primers were then annealed and bound to Polymerase 3.0 using the Sequel Binding Kit 3.0 (Pacific Biosciences, CA). The final bound complex was then purified and later sequenced on PacBio Sequel I using the SMRT Cell M1 v3 tray (Pacific Biosciences, CA). The spike-in controls for each PacBio Sequel I run were from the Internal Control Kit 3.0 (Pacific Biosciences, CA).

Genome assembly, annotation, and pangenome construction

Borreliaceae isolates that were sequenced on the PacBio Sequel I platform were processed using pbcromwell (v 1.0.4) (Pacific Biosciences, CA). All the barcoded data were demultiplexed using pb_demux_subreads and assembled using pb_assembly_microbial. The base modification motifs for each assembly were computed using pb_basemods. The species of each genome was verified using taxator and GTDB-Tk (v1.7.0) with repack (v r202). All reference and sequenced *Borreliaceae* assemblies were annotated with Prokka (v 1.11) and then homologous genes were clustered with Roary (v 3.5.1) (68, 69). The BlastP threshold was calculated consecutively at 5% intervals to determine the threshold. The final BlastP threshold, 75%, was selected based on the separation of the core pangenome from the cloud and accessory pangenomes. All CDS identified by Prokka were also annotated using EggNOG (70).

Phylogenetic tree construction

Single-copy core genes of each *Borreliaceae* genome were aligned using MAFFTT (v7.490-1) and then the concatenated alignment was run through Randomized Accelerated Maximum Likelihood (RaxML) to generate a Newick string (71, 72). The resulting file was then ported into RStudio (v2022.07.2 + 576), a tree was generated using ape and then visualized using ggtree (v3.15) R packages (73, 74).

Average Nucleotide Identity and percent similarity or difference analysis

All pairs of *Borreliaceae* genomes, both sequenced and downloaded were checked for overall similarity by ANI using the pyani (v2.12) with the ANIb method (75). The resulting whole-genome ANI matrix (or Hadamard matrix) was read into RStudio (v.4.1.1), clustered with adapted RaxML phylogeny, and visualized with pheatmap (v1.0.12).

The *Borreliaceae* percent gene similarity was calculated using RStudio in a pairwise fashion for each isolate in this study. The final list of shared percent identity was then pivoted wider to generate a matrix. Any pairs that did not have a gene percent similarity value were assigned a value of 0. The final matrix was then clustered with hclust and visualized with pheatmap (v1.0.12).

Borreliaceae pangenome mapping

To find pangenome distribution, all Roary at BlastP 75% or EggNOG gene clusters were imported into R (70). Then each gene was associated with a corresponding strain and/or replicon before being used to construct a presence-absence matrix. This matrix was then visualized in pheatmap to see the gene presence and absence across all species used in this study. To determine if the *Borreliaceae* pangenome was either open or closed, the composition was checked by shuffling strains, incorporating them one at a time into the analysis, and then calculating the core/distributed component. The combinations of strains incorporated were shifted prior to being plotted in each iteration.

Declarations

Funding

This work was supported by grants from the Bill and Marian Cook Foundation; the Oskar Fischer Project, a James Truchard Philanthropy; Cure; the Coulter-Drexel Translational Research Partnership; and NIH grant Nos: DC 02148 and DK082316 to GDE. This work was also supported by NIH grants 1R01AI165876-01 and 1R01AI141801-01A1 to RTM.

Availability of data and materials

The sequencing data used in this study is available under Bioproject PRJNA1026537.

Competing interests

(NONE)

Author details

(NA)

Author Contribution

The main manuscript text was written by KMS, GDE, and MCM. Study strategy was conceived by GDE. Study design was conceived by KMS, JPE, JCM, RLE, RTM, and GDE. Experimental work was performed by KMS, SB, AA, NAC, RLE, JPE, and JCM. Data analysis was performed by KMS, JPE, RLE, and JCM. Figure construction was performed by KMS, JPE, JCM, and MCM. Funding was provided through grants awarded to GDE and RTM.

References

1. Shen K, Wang X, Post JC, Ehrlich GD. Molecular and Translational Research Approaches for the study of Bacterial Pathogenesis in Otitis Media. In: Rosenfeld R, Bluestone CD, editors. Evidence-based Otitis Media. 2nd ed. Hamilton, London: B.C. Decker Inc.; 2003. pp. 91–119.
2. Ehrlich GD, Hu ZF, Post JC. Role for Biofilms in Infectious Disease. In: Ghannoum M, O'Toole GA, editors. Microbial Biofilms. Washington, D.C.: ASM; 2004. pp. 332–58.
3. Ehrlich GD, Hu FZ, Shen K, Stoodley P, Post JC. Bacterial plurality as a general mechanism driving persistence in chronic infections. Clin Orthop Relat Res. 2005(437):20–4.
4. Ehrlich GD, Ahmed A, Earl J, Hiller NL, Costerton JW, Stoodley P, et al. The distributed genome hypothesis as a rubric for understanding evolution in situ during chronic bacterial biofilm infectious processes. FEMS Immunol Med Microbiol. 2010;59(3):269–79.

5. Shen K, Antalis P, Gladitz J, Sayeed S, Ahmed A, Yu S, et al. Identification, distribution, and expression of novel genes in 10 clinical isolates of nontypeable *Haemophilus influenzae*. *Infect Immun*. 2005;73(6):3479–91.
6. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, et al. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial pan-genome. *Proc Natl Acad Sci U S A*. 2005;102(39):13950–5.
7. Shen K, Gladitz J, Antalis P, Dice B, Janto B, Keefe R, et al. Characterization, distribution, and expression of novel genes among eight clinical isolates of *Streptococcus pneumoniae*. *Infect Immun*. 2006;74(1):321–30.
8. Shen K, Sayeed S, Antalis P, Gladitz J, Ahmed A, Dice B, et al. Extensive genomic plasticity in *Pseudomonas aeruginosa* revealed by identification and distribution studies of novel genes among clinical isolates. *Infect Immun*. 2006;74(9):5272–83.
9. Hogg JS, Hu FZ, Janto B, Boissy R, Hayes J, Keefe R, et al. Characterization and modeling of the *Haemophilus influenzae* core and supragenomes based on the complete genomic sequences of Rd and 12 clinical nontypeable strains. *Genome Biol*. 2007;8(6):R103.
10. Hiller NL, Janto B, Hogg JS, Boissy R, Yu S, Powell E, et al. Comparative genomic analyses of seventeen *Streptococcus pneumoniae* strains: insights into the pneumococcal supragenome. *J Bacteriol*. 2007;189(22):8186–95.
11. Donati C, Hiller NL, Tettelin H, Muzzi A, Croucher NJ, Angiuoli SV, et al. Structure and dynamics of the pan-genome of *Streptococcus pneumoniae* and closely related species. *Genome Biol*. 2010;11(10):R107.
12. Hall BG, Ehrlich GD, Hu FZ. Pan-genome analysis provides much higher strain typing resolution than multi-locus sequence typing. *Microbiol (Reading)*. 2010;156(Pt 4):1060–8.
13. Boissy R, Ahmed A, Janto B, Earl J, Hall BG, Hogg JS, et al. Comparative supragenomic analyses among the pathogens *Staphylococcus aureus*, *Streptococcus pneumoniae*, and *Haemophilus influenzae* using a modification of the finite supragenome model. *BMC Genomics*. 2011;12:187.
14. Davie JJ, Earl J, de Vries SP, Ahmed A, Hu FZ, Bootsma HJ, et al. Comparative analysis and supragenome modeling of twelve *Moraxella catarrhalis* clinical isolates. *BMC Genomics*. 2011;12:70.
15. Ahmed A, Earl J, Retchless A, Hillier SL, Rabe LK, Cherpes TL, et al. Comparative genomic analyses of 17 clinical isolates of *Gardnerella vaginalis* provide evidence of multiple genetically isolated clades consistent with subspeciation into genovars. *J Bacteriol*. 2012;194(15):3922–37.
16. Earl JP, de Vries SP, Ahmed A, Powell E, Schultz MP, Hermans PW, et al. Comparative Genomic Analyses of the *Moraxella catarrhalis* Serosensitive and Seroresistant Lineages Demonstrate Their Independent Evolution. *Genome Biol Evol*. 2016;8(4):955–74.
17. Nistico LEJ, Hiller L, Ahmed A, Retchless A, Janto B, Costerton JC, Hu FZ, Ehrlich GD. Using the core and supra genomes to determine diversity and natural proclivities among bacterial strains. In:

- Torben L, Skovhus SMCaCRJH, editors. Applications of Molecular Microbiological Methods. U.K.: Caister Academic; 2014.
18. Innamorati KA, Earl JP, Aggarwal SD, Ehrlich GD, Hiller NL. The Bacterial Guide to Designing a Diversified Gene Portfolio. In: Tettelin H, Medini D, editors. The Pangenome: Diversity, Dynamics and Evolution of Genomes. Cham (CH)2020. pp. 51–87.
 19. Hammond JA, Gordon EA, Socarras KM, Mell JC, Ehrlich GD. Beyond the pan-genome: current perspectives on the functional and practical outcomes of the distributed genome hypothesis. *Biochem Soc Trans* 1–192020.
 20. Otani H, Uduary DW, Mouncey NJ. Comparative and pangenomic analysis of the genus *Streptomyces*. *Sci Rep*. 2022;12(1):18909.
 21. Kim M, Cha IT, Lee KE, Li M, Park SJ. Pangenome analysis provides insights into the genetic diversity, metabolic versatility, and evolution of the genus *Flavobacterium*. *Microbiol Spectr*. 2023;11(5):e0100323.
 22. Socarras KM, Haslund-Gourley BS, Cramer NA, Comunale MA, Marconi RT, Ehrlich GD. Large-Scale Sequencing of *Borreliaceae* for the Construction of Pan-Genomic-Based Diagnostics. *Genes (Basel)*. 2022;13(9).
 23. Rajput A, Chauhan SM, Mohite OS, Hyun JC, Ardalani O, Jahn LJ, et al. Pangenome analysis reveals the genetic basis for taxonomic classification of the *Lactobacillaceae* family. *Food Microbiol*. 2023;115:104334.
 24. Kress-Bennett JM, Hiller NL, Eutsey RA, Powell E, Longwell MJ, Hillman T, et al. Identification and Characterization of *msf*, a Novel Virulence Factor in *Haemophilus influenzae*. *PLoS ONE*. 2016;11(3):e0149891.
 25. Hu FZ, Krol JE, Tsai CHS, Eutsey RA, Hiller LN, Sen B, et al. Deletion of genes involved in the ketogluconate metabolism, Entner-Doudoroff pathway, and glucose dehydrogenase increase local and invasive virulence phenotypes in *Streptococcus pneumoniae*. *PLoS ONE*. 2019;14(1):e0209688.
 26. Casjens SR, Mongodin EF, Qiu WG, Luft BJ, Schutzer SE, Gilcrease EB, et al. Genome stability of Lyme disease spirochetes: comparative genomics of *Borrelia burgdorferi* plasmids. *PLoS ONE*. 2012;7(3):e33280.
 27. Kneubehl AR, Krishnavajhala A, Leal SM, Replogle AJ, Kingry LC, Bermudez SE, et al. Comparative genomics of the Western Hemisphere soft tick-borne relapsing fever borreliae highlights extensive plasmid diversity. *BMC Genomics*. 2022;23(1):410.
 28. Pfeiffer MB, Lyme. The First Epidemic of Climate Change. Washington, D.C.: Island; 2018.
 29. 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 *Borrelia* *gen. nov.* containing the members of the Lyme disease *Borrelia* (*Borrelia burgdorferi sensu lato* complex). *Antonie Van Leeuwenhoek*. 2014;105(6):1049–72.

30. Margos G, Gofton A, Wibberg D, Dangel A, Marosevic D, Loh SM, et al. The genus *Borrelia* reloaded. *PLoS ONE*. 2018;13(12):e0208432.
31. Winslow C, Coburn J. Recent discoveries and advancements in research on the Lyme disease spirochete *Borrelia burgdorferi*. *F1000Res*. 2019;8.
32. Estrada-Pena A, Cabezas-Cruz A. Phyloproteomic and functional analyses do not support a split in the genus *Borrelia* (phylum Spirochaetes). *BMC Evol Biol*. 2019;19(1):54.
33. Barbour AG, Adeolu M, Gupta RS et al. Division of the genus *Borrelia* into two genera (corresponding to Lyme disease and relapsing fever groups) reflects their genetic and phenotypic distinctiveness and will lead to a better understanding of these two groups of microbes (Margos. (2016) There is inadequate evidence to support the division of the genus *Borrelia*. *Int. J. Syst. Evol. Microbiol*. 10.1099/ijsem.0.001717). *Int J Syst Evol Microbiol*. 2017;67(6):2058-67.
34. Gupta RS et al. Distinction between *Borrelia* and *Borrelia* is more robustly supported by molecular and phenotypic characteristics than all other neighbouring prokaryotic genera: Response to Margos'. The genus *Borrelia* reloaded (*PLoS ONE* 13(12): e0208432). *PLoS One*. 2019;14(8):e0221397.
35. Barbour AG, Garon CF. Linear plasmids of the bacterium *Borrelia burgdorferi* have covalently closed ends. *Science*. 1987;237(4813):409–11.
36. Baril C, Richaud C, Baranton G, Saint Girons IS. Linear chromosome of *Borrelia burgdorferi*. *Res Microbiol*. 1989;140(8):507–16.
37. Mongodin EF, Casjens SR, Bruno JF, Xu Y, Drabek EF, Riley DR, et al. Inter- and intra-specific pan-genomes of *Borrelia burgdorferi* sensu lato: genome stability and adaptive radiation. *BMC Genomics*. 2013;14:693.
38. Schwartz I, Margos G, Casjens SR, Qiu WG, Eggers CH. Multipartite Genome of Lyme Disease *Borrelia*: Structure, Variation and Prophages. *Curr Issues Mol Biol*. 2021;42:409–54.
39. 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(8):e1011243.
40. Schwartz DC, Cantor CR. Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis. *Cell*. 1984;37(1):67–75.
41. 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(6660):580–6.
42. Barbour AG. Chromosome and Plasmids of the Tick-Borne Relapsing Fever Agent *Borrelia hermsii*. *Genome Announc*. 2016;4(3).
43. Darling AE, Mau B, Perna NT. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE*. 2010;5(6):e11147.
44. Kingry LC. Genomics of Relapsing Fever Spirochetes. In: Samuels JDRDS, editor. *Lyme Disease and Relapsing Fever Spirochetes: Genomics, Molecular Biology, Host Interactions and Disease. Pathogenesis*: Caister Academic; 2021. pp. 63–72.

45. Chaconas G. Hairpin telomeres and genome plasticity in *Borrelia*: all mixed up in the end. *Mol Microbiol.* 2005;58(3):625–35.
46. D W. pafr: reading, manipulating and plotting genome alignments in the PAF format 2020 [https://cran.r-project.org/web/packages/pafr/vignettes/Introduction_to_pafr.html].
47. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics.* 2018;34(18):3094–100.
48. Tourand Y, Lee L, Chaconas G. Telomere resolution by *Borrelia burgdorferi* ResT through the collaborative efforts of tethered DNA binding domains. *Mol Microbiol.* 2007;64(3):580–90.
49. Kobryn K, Chaconas G. ResT, a telomere resolvase encoded by the Lyme disease spirochete. *Mol Cell.* 2002;9(1):195–201.
50. Gupta RS, Mahmood S, Adeolu M. A phylogenomic and molecular signature based approach for characterization of the phylum Spirochaetes and its major clades: proposal for a taxonomic revision of the phylum. *Front Microbiol.* 2013;4:217.
51. Guner ES, Watanabe M, Hashimoto N, Kadosaka T, Kawamura Y, Ezaki T, et al. *Borrelia turcica* sp. nov., isolated from the hard tick *Hyalomma aegyptium* in Turkey. *Int J Syst Evol Microbiol.* 2004;54(Pt 5):1649–52.
52. Loh SM, Gofton AW, Lo N, Gillett A, Ryan UM, Irwin PJ, et al. Novel *Borrelia* species detected in echidna ticks, *Bothriocroton concolor*, in Australia. *Parasit Vectors.* 2016;9(1):339.
53. Loh SM, Gillett A, Ryan U, Irwin P, Oskam C. Molecular characterization of 'Candidatus *Borrelia tachyglossi*' (family Spirochaetaceae) in echidna ticks, *Bothriocroton concolor*. *Int J Syst Evol Microbiol.* 2017;67(4):1075–80.
54. Panetta JL, Šíma R, Calvani NED, Hajdušek O, Chandra S, Panuccio J, et al. Reptile-associated *Borrelia* species in the goanna tick (*Bothriocroton undatum*) from Sydney, Australia. *Parasites Vectors.* 2017;10(1):616.
55. Takano A, Goka K, Une Y, Shimada Y, Fujita H, Shiino T, et al. Isolation and characterization of a novel *Borrelia* group of tick-borne borreliae from imported reptiles and their associated ticks. *Environ Microbiol.* 2010;12(1):134–46.
56. Anderson JF, Magnarelli LA, LeFebvre RB, Andreadis TG, McAninch JB, Perng GC, et al. Antigenically variable *Borrelia burgdorferi* isolated from cottontail rabbits and *Ixodes dentatus* in rural and urban areas. *J Clin Microbiol.* 1989;27(1):13–20.
57. Marconi RT, Liveris D, Schwartz I. Identification of novel insertion elements, restriction fragment length polymorphism patterns, and discontinuous 23S rRNA in Lyme disease spirochetes: phylogenetic analyses of rRNA genes and their intergenic spacers in *Borrelia japonica* sp. nov. and genomic group 21038 (*Borrelia andersonii* sp. nov.) isolates. *J Clin Microbiol.* 1995;33(9):2427–34.
58. 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(1):165.

59. Ciufo S, Kannan S, Sharma S, Badretin A, Clark K, Turner S, et al. Using average nucleotide identity to improve taxonomic assignments in prokaryotic genomes at the NCBI. *Int J Syst Evol Microbiol*. 2018;68(7):2386–92.
60. Konstantinidis KT, Tiedje JM. Genomic insights that advance the species definition for prokaryotes. *Proc Natl Acad Sci U S A*. 2005;102(7):2567–72.
61. Lescot M, Audic S, Robert C, Nguyen TT, Blanc G, Cutler SJ, et al. The genome of *Borrelia recurrentis*, the agent of deadly louse-borne relapsing fever, is a degraded subset of tick-borne *Borrelia duttonii*. *PLoS Genet*. 2008;4(9):e1000185.
62. Miller SC, Porcella SF, Raffel SJ, Schwan TG, Barbour AG. Large linear plasmids of *Borrelia* species that cause relapsing fever. *J Bacteriol*. 2013;195(16):3629–39.
63. Casjens S, Palmer N, van Vugt R, Huang WM, Stevenson B, Rosa P, et al. A bacterial genome in flux: the twelve linear and nine circular extrachromosomal DNAs in an infectious isolate of the Lyme disease spirochete *Borrelia burgdorferi*. *Mol Microbiol*. 2000;35(3):490–516.
64. Hinnebusch J, Barbour AG. Linear plasmids of *Borrelia burgdorferi* have a telomeric structure and sequence similar to those of a eukaryotic virus. *J Bacteriol*. 1991;173(22):7233–9.
65. Chaconas G, Kobryn K. Structure, function, and evolution of linear replicons in *Borrelia*. *Annu Rev Microbiol*. 2010;64:185–202.
66. Cabanettes F, Klopp C. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *PeerJ*. 2018;6:e4958.
67. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. *Genome Biol*. 2004;5(2):R12.
68. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*. 2014;30(14):2068–9.
69. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, et al. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*. 2015;31(22):3691–3.
70. Jensen LJ, Julien P, Kuhn M, von Mering C, Muller J, Doerks T, et al. eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res*. 2008;36(Database issue):D250–4.
71. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. 2014;30(9):1312–3.
72. Nakamura T, Yamada KD, Tomii K, Katoh K. Parallelization of MAFFT for large-scale multiple sequence alignments. *Bioinformatics*. 2018;34(14):2490–2.
73. Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*. 2004;20(2):289–90.
74. Yu G. Using ggtree to Visualize Data on Tree-Like Structures. *Curr Protoc Bioinf*. 2020;69(1):e96.
75. Pritchard LGR, Humphris S, Elphinstone JG, Toth IK. Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. *Anal Methods*. 2016(1).

Figures

Figure 1.

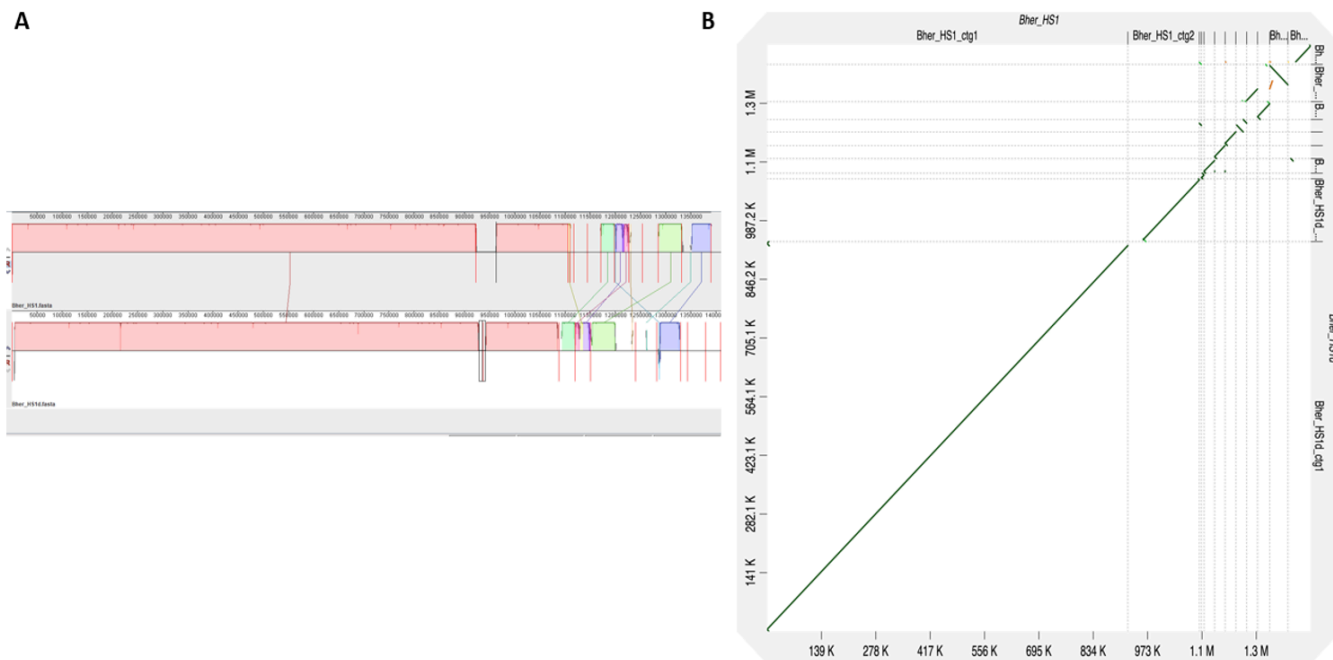


Figure 1

In silico sequencing validation of *B. hermsii* strain HS1 pair shows homology throughout several replicons and indicates regions of heterogeneity. Alignments of *B. hermsii* strain HS1 show nearly identical genomes. Similar pairwise locally colinear blocks are indicated in the same colors (pink to pink, blue to blue etc.). Contig breaks are indicated by red lines. b D-Genies dot plot of whole-genome complete reference *B. hermsii* strain HS1 (query, x-axis) to our sequenced *B. hermsii* strain HS1d (target, y-axis). Overall percent similarity was indicated from red (low = 0) to green (high = 100%). All contigs of the two whole genome sequences are portrayed in the figure. Each contig pair is separated by a dotted line. The position of the matches between strains is indicated by colored dots and lines (yellow, orange, green).

Figure 2

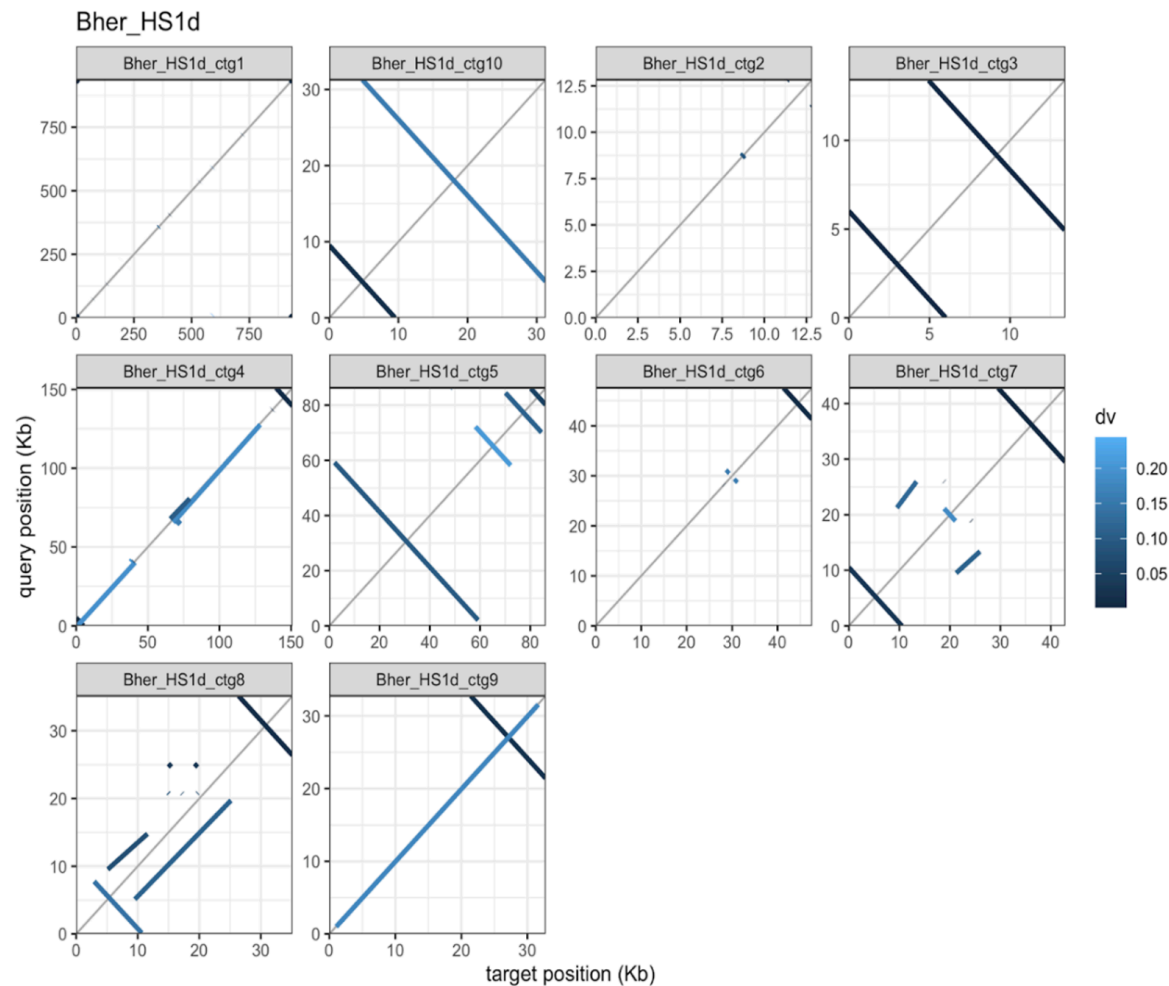


Figure 2

Self-alignment of *B. hermsii* strain HS1d contigs reveals inverted repeats at linear contig ends. A faceted pair dot plot of each contig within the *B. hermsii* strain HS1d whole-genome sequence. The overall percent similarity was indicated from light blue (low = 0) to dark blue (high = 100%). All contigs were represented within each individual facet of the plot. Within each plot X and Y-axes are the overall lengths of the compared contigs.

Figure 3

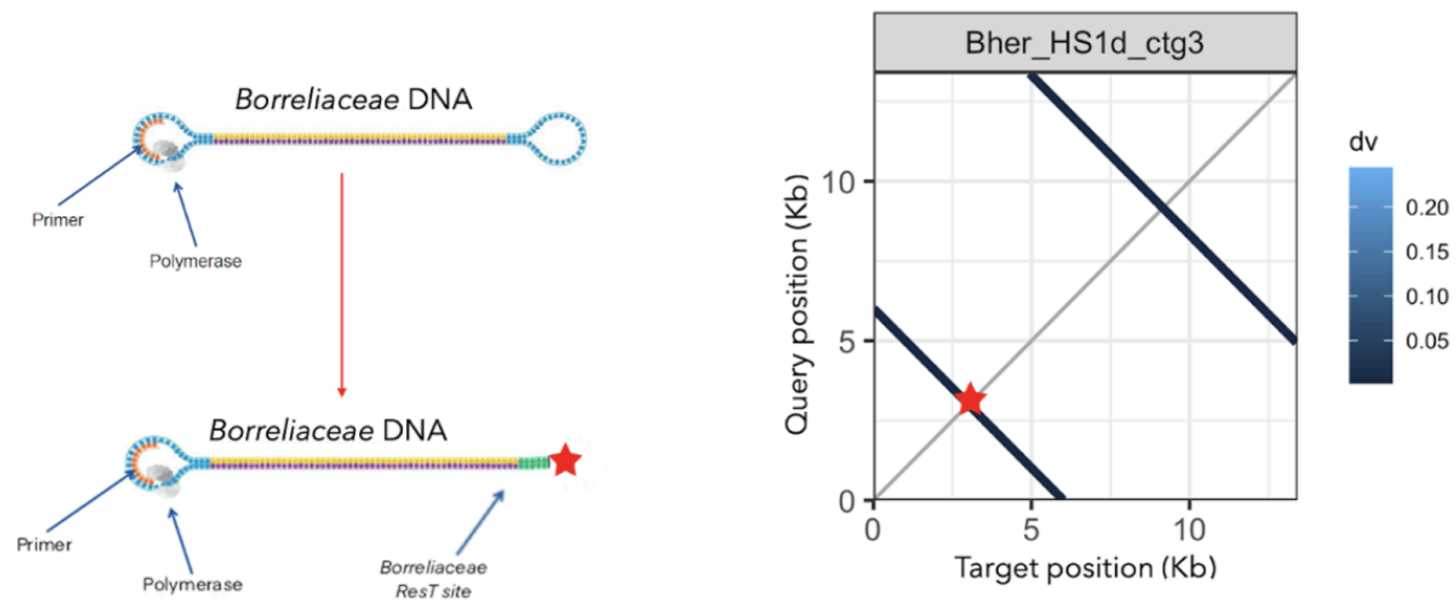


Figure 3

Blast search of *B. hermsii* strain HS1 linear contigs reveals ResT site located on ends of sequenced replicons. A graphic depiction of Pacific Biosciences long-read sequencing platform and modified prepared Borreliaceae template. Alongside sequencing schematic is one of the faceted plots of *B. hermsii* strain HS1d (contig 3). This plot shows the percent similarity of replicon along its entire length with regions of similarity removed. Perpendicular lines are inverted repeats. The overall percent similarity was indicated from light blue (low = 0) to dark blue (high = 100%). All contigs were represented within each individual facet of the plot.

Figure 4

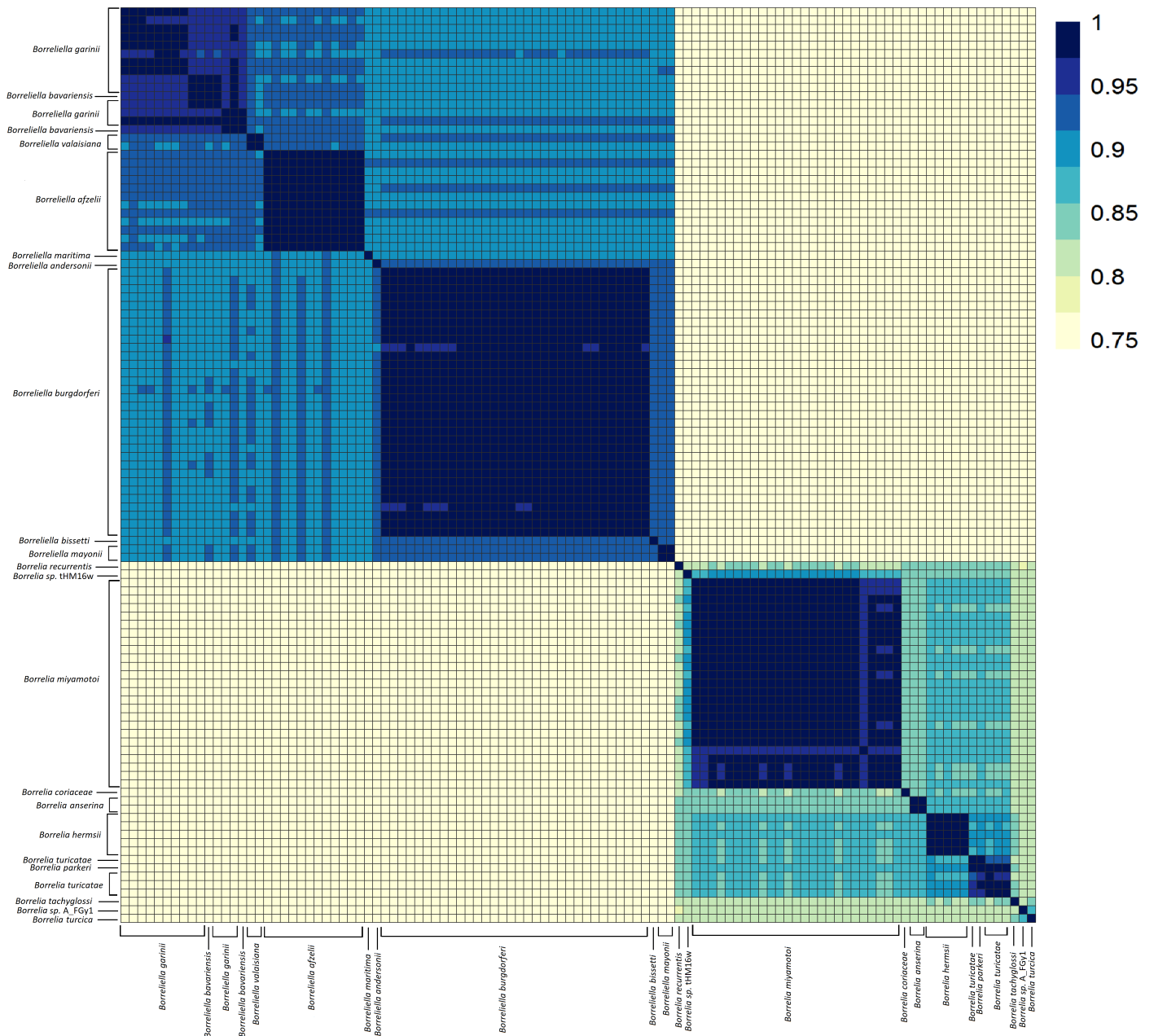


Figure 4

Average Nucleotide Identity of all Borreliaceae species and strains examined demonstrated genera-specific clustering. The range of ANI is from least (yellow) to greatest (blue). Each species is represented on the x- and y-axes. The diagonal line represents identity. The dendrograms were made with complete-linkage hierarchical clustering. The key color is noted on the right-hand side. The annotations indicate the generic groupings.

Figure 5

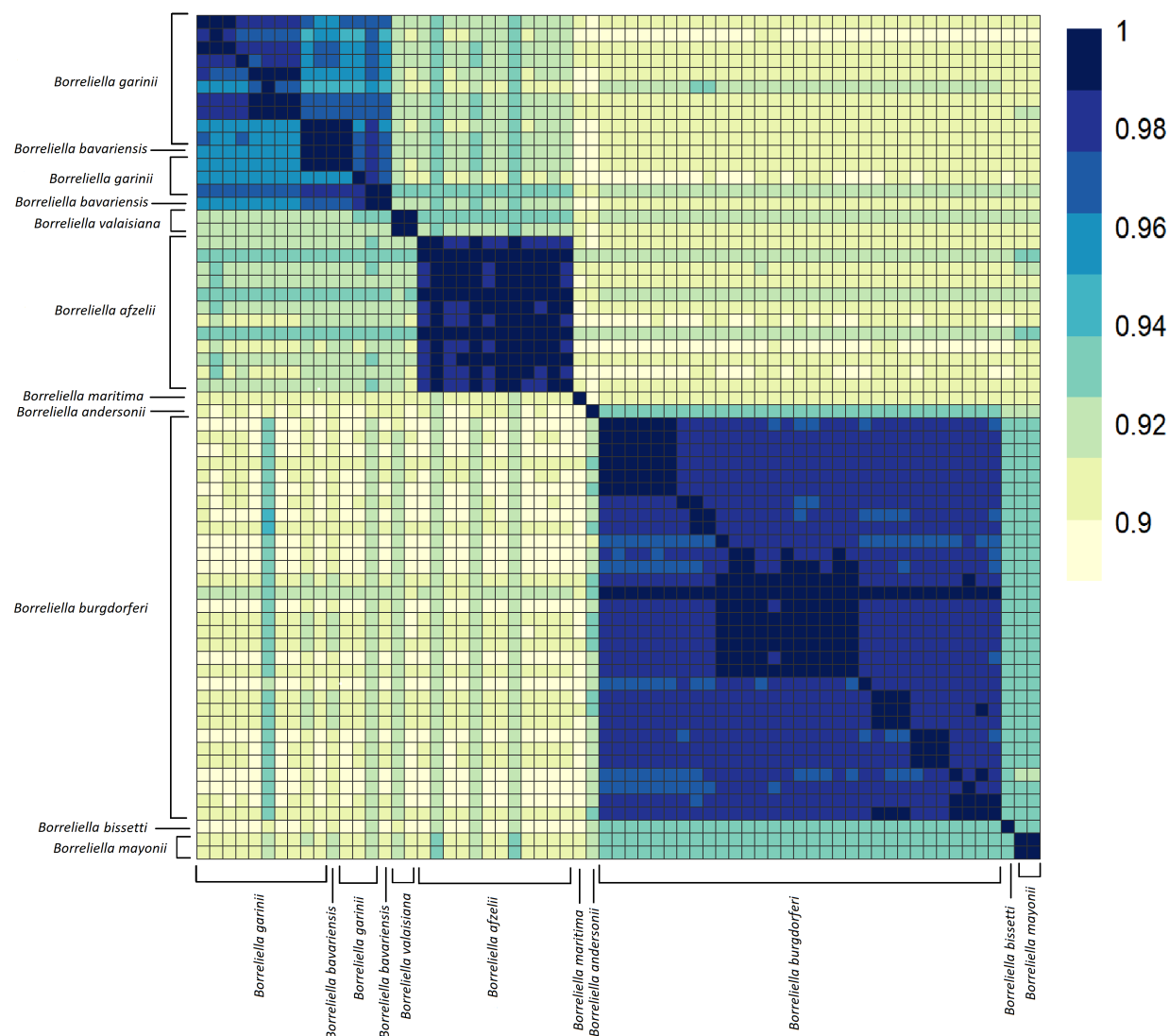


Figure 5

Average Nucleotide Identity analyses of *Borreliella* genomes shows a high degree of similarity amongst the major species of the genus. The range of ANI is from least (yellow) to greatest (blue). Each of the *Borreliella* spirochetes are represented on the x- and y-axes. The diagonal line represents identities. The dendrograms were made with complete-linkage hierarchical clustering. The color bar key is noted on the right-hand side. The annotations indicate the species group.

Figure 6

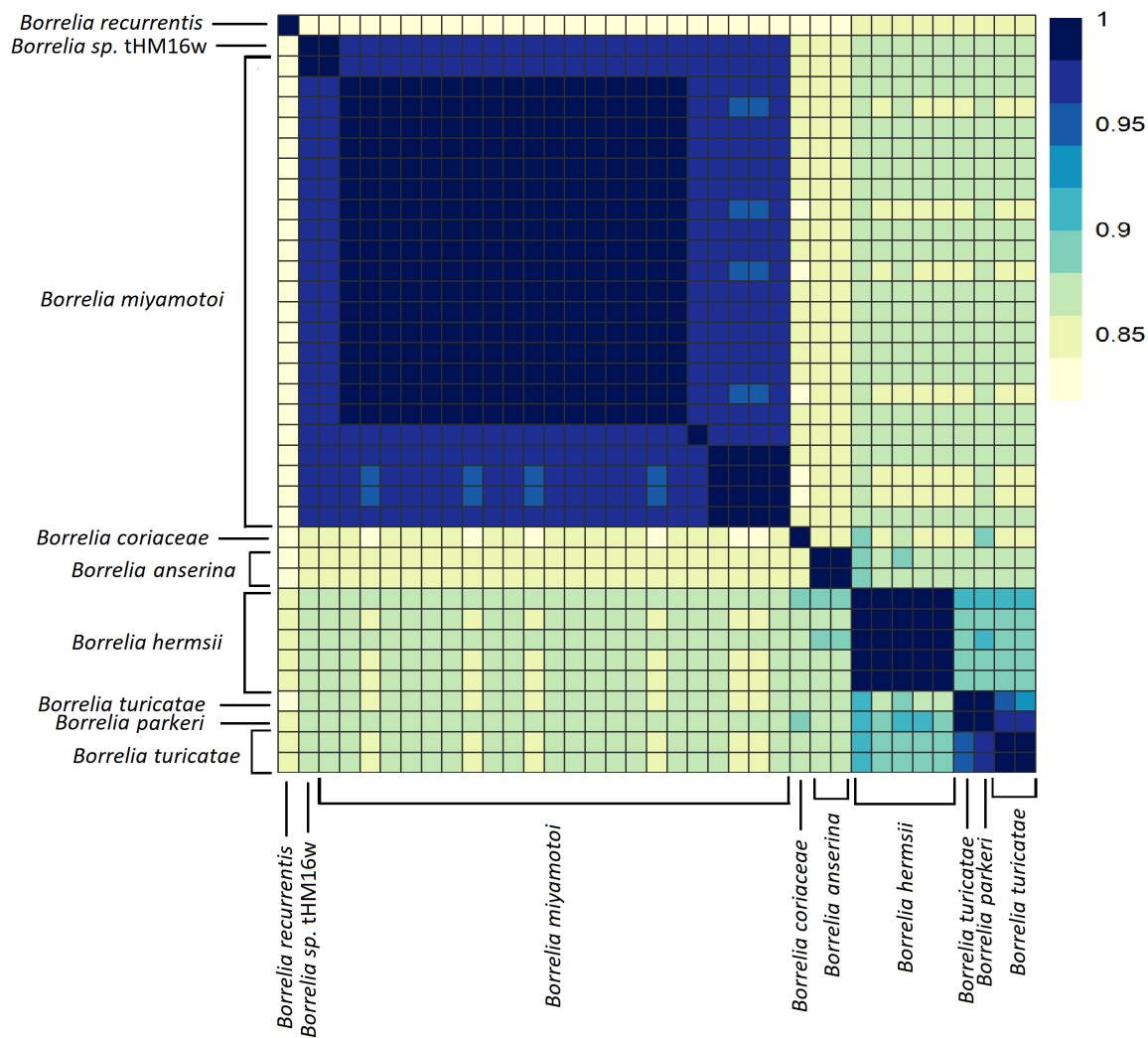


Figure 6

Average Nucleotide Identity of *Borrelia* genome spirochete shows separation of major species. The range of ANI is from least (yellow) to greatest (blue). Each *Borrelia* spirochete was represented on both axes. The diagonal line is where sample similarity values are to itself and beyond said line is for other genus members. Complete-linkage hierarchical clustering. The color bar key is noted on the right-hand side. The annotations indicate the species group.

Figure 7

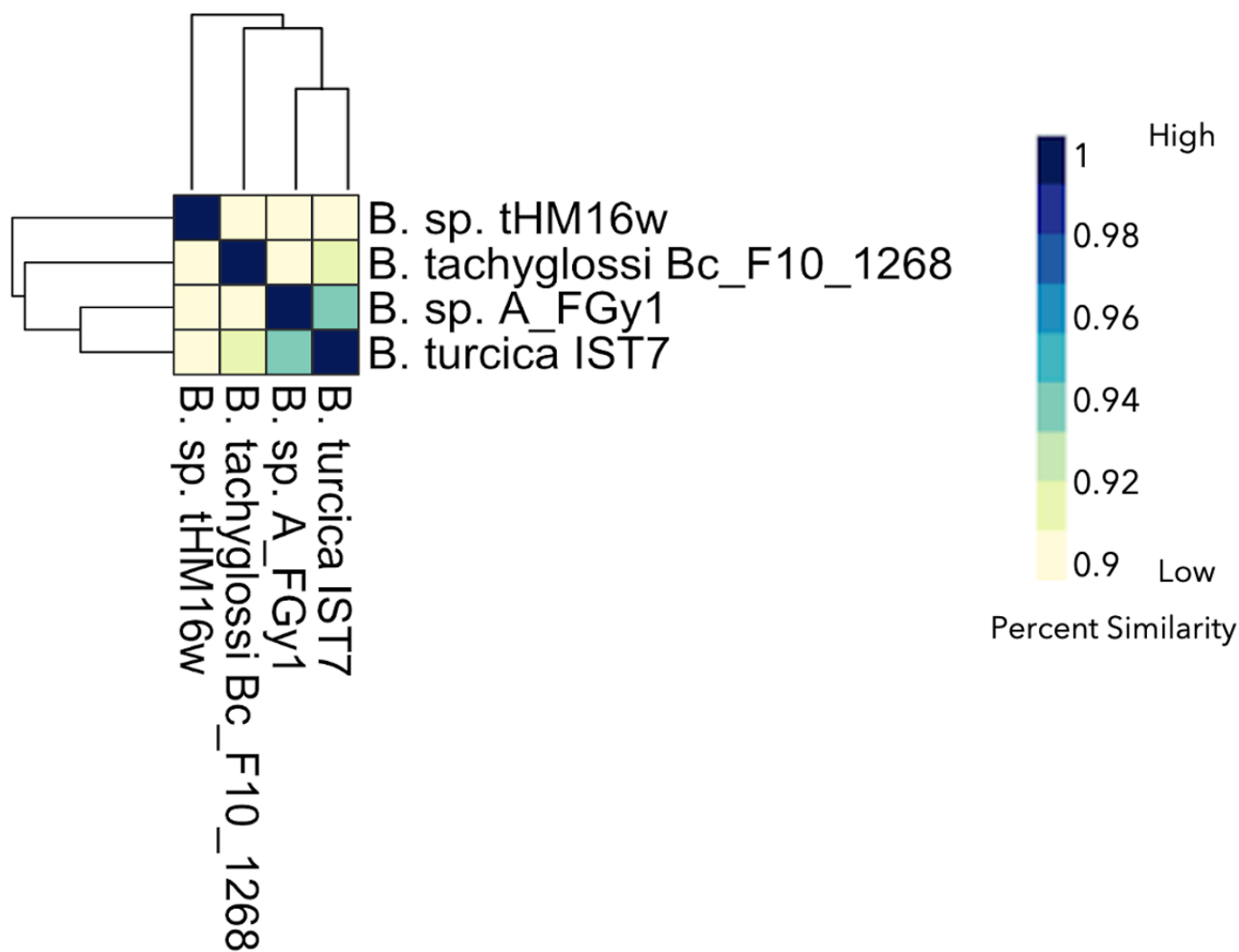


Figure 7

Average Nucleotide Identity of reptile-associated *Borrelia* spirochetes suggests placement of non-species designated isolates. The range of ANI is from least (yellow) to greatest (blue). Each isolate is represented on the x- and y-axis. The diagonal line is where sample similarity values are to itself and beyond said line is for other genus members. Complete-linkage hierarchical clustering. The annotation color bars for each spirochete are noted on the right-hand side.

Figure 8

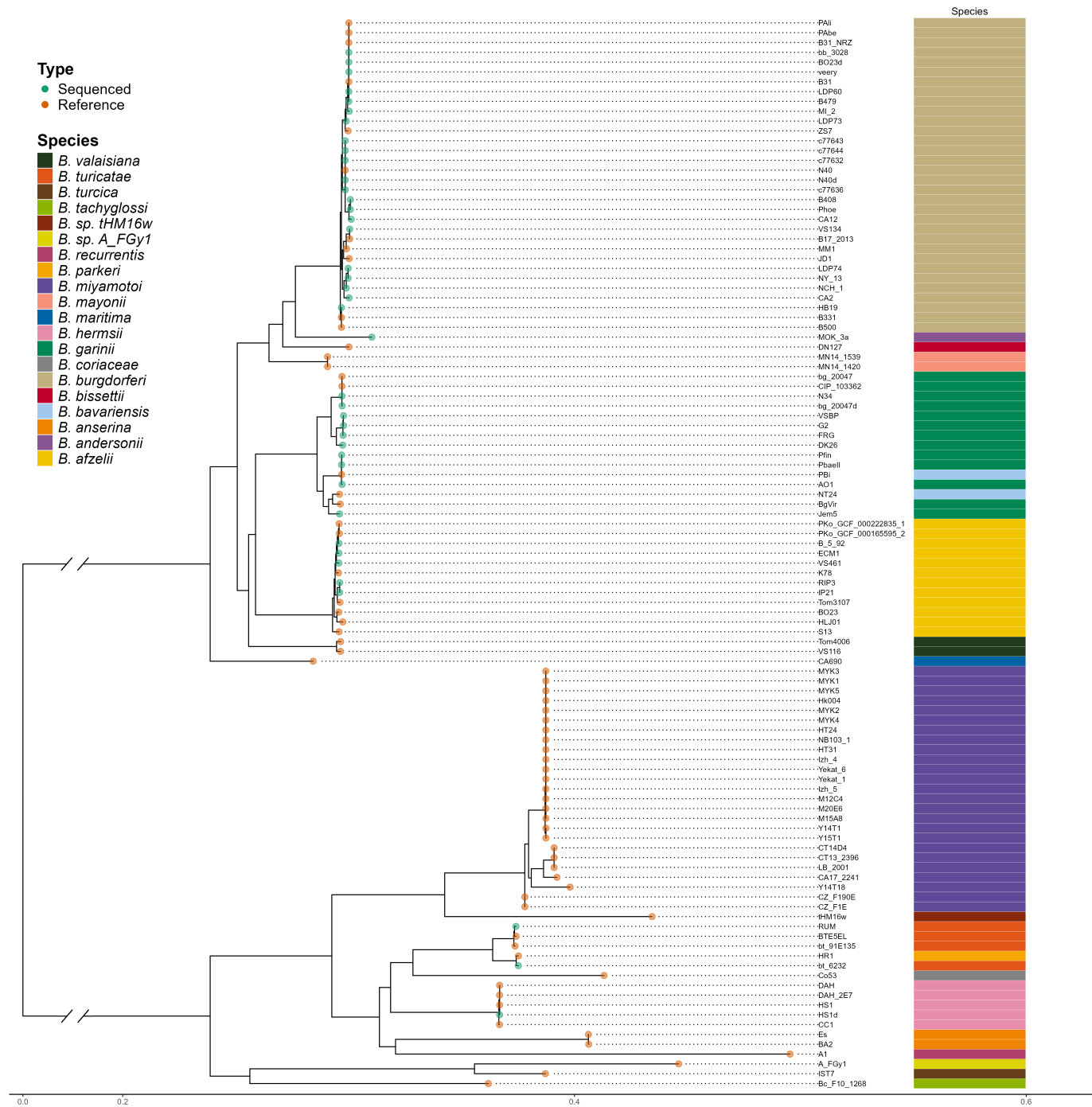


Figure 8

Midpoint rooted single-copy core gene tree of Borreliaceae shows the division between the Borreliella and Borrelia genera. The origin of the spirochete is annotated in blue (VCU) or red (NCBI) spheres at the nodes. The annotation bar indicates the species.

Figure 9

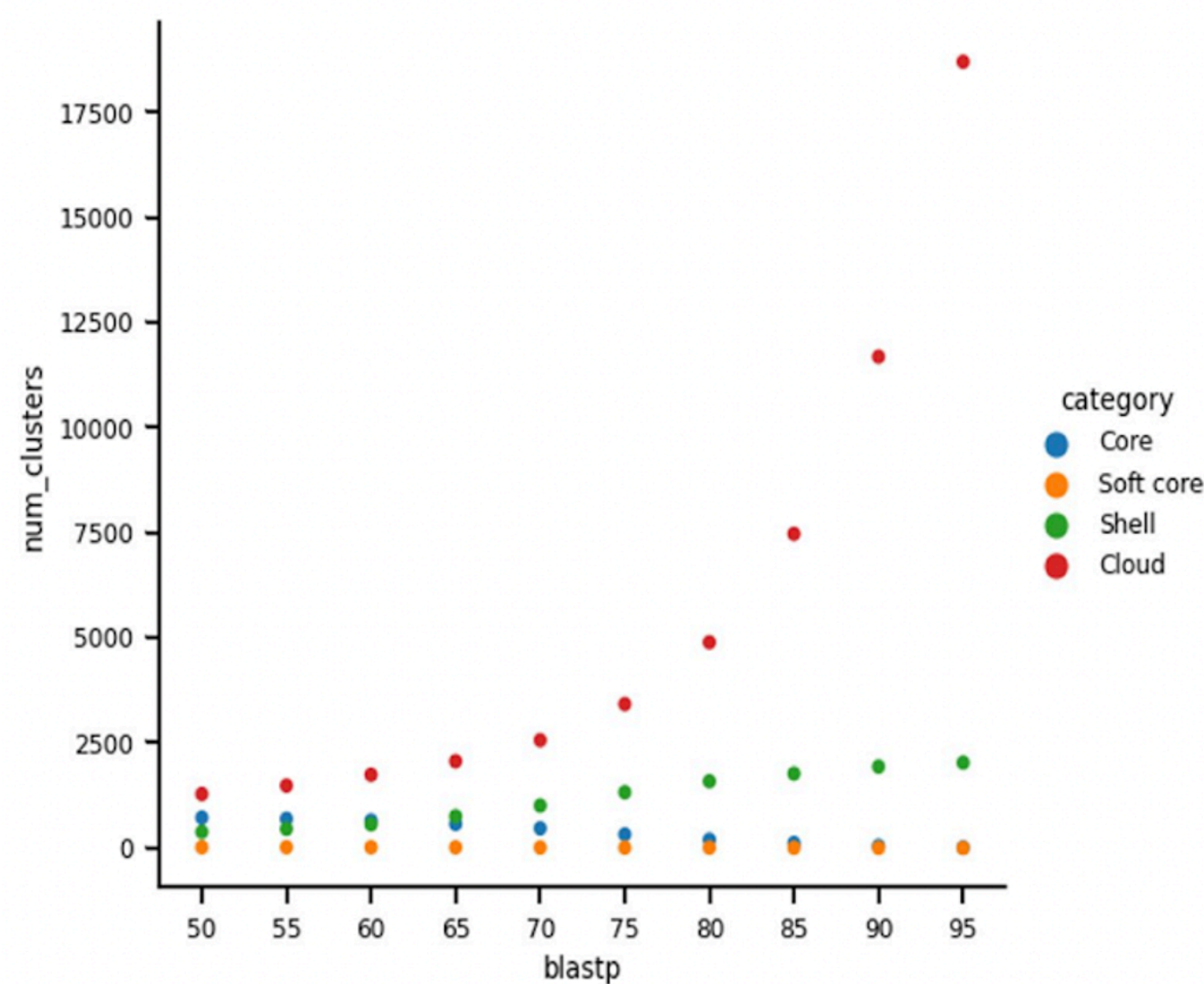


Figure 9

The gene cluster frequency of the Borreliaceae family pangenome reveals a large pangenome. Each colored circle signifies different components of the pangenome [core (blue), soft-core (orange), shell (green), and cloud (red)]. X-axis signifies the BlastP threshold. Y-axis signifies the number of genes within the pangenome.

Figure 10

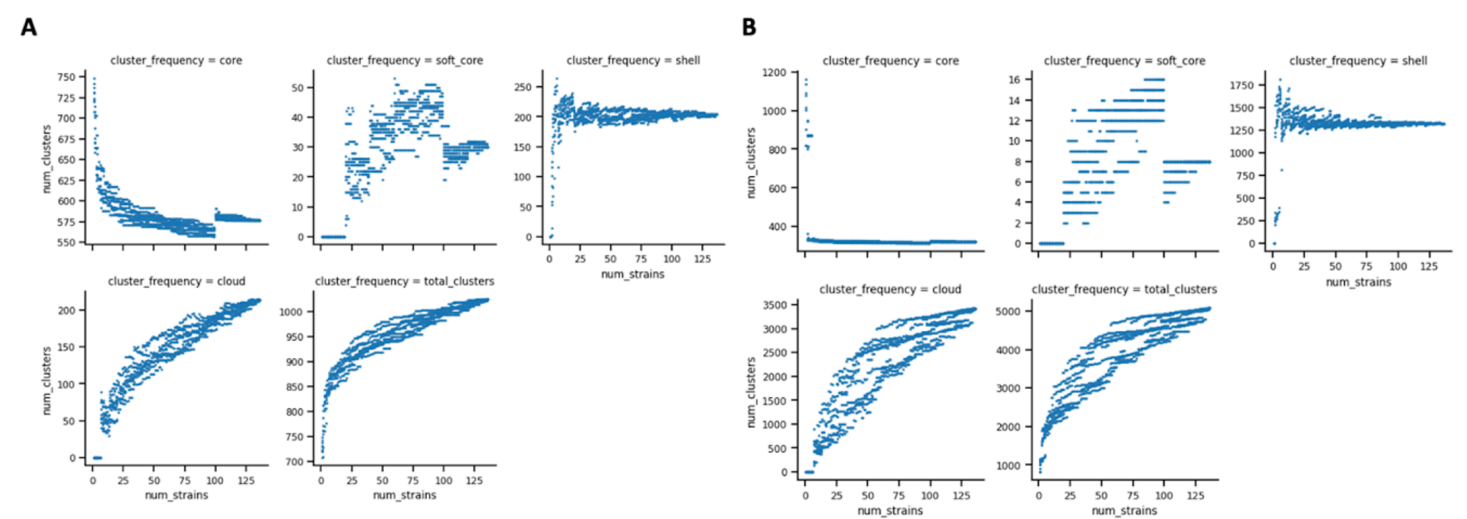


Figure 10

Roary and EggNOG depict an expanding Borreliaceae pangenome over the number of included isolate genomes. X- axis represents the number of isolates and Y-axis represents the number of gene clusters within pangenome. Each panel-faceted plot is dedicated to different components of the pangenome. **a**Roary BlastP 75% was chosen as the threshold. Each light blue dot indicates an incorporated isolate. **b** EggNOG depiction of pangenome fluctuation with each incorporated genome.

Figure 11

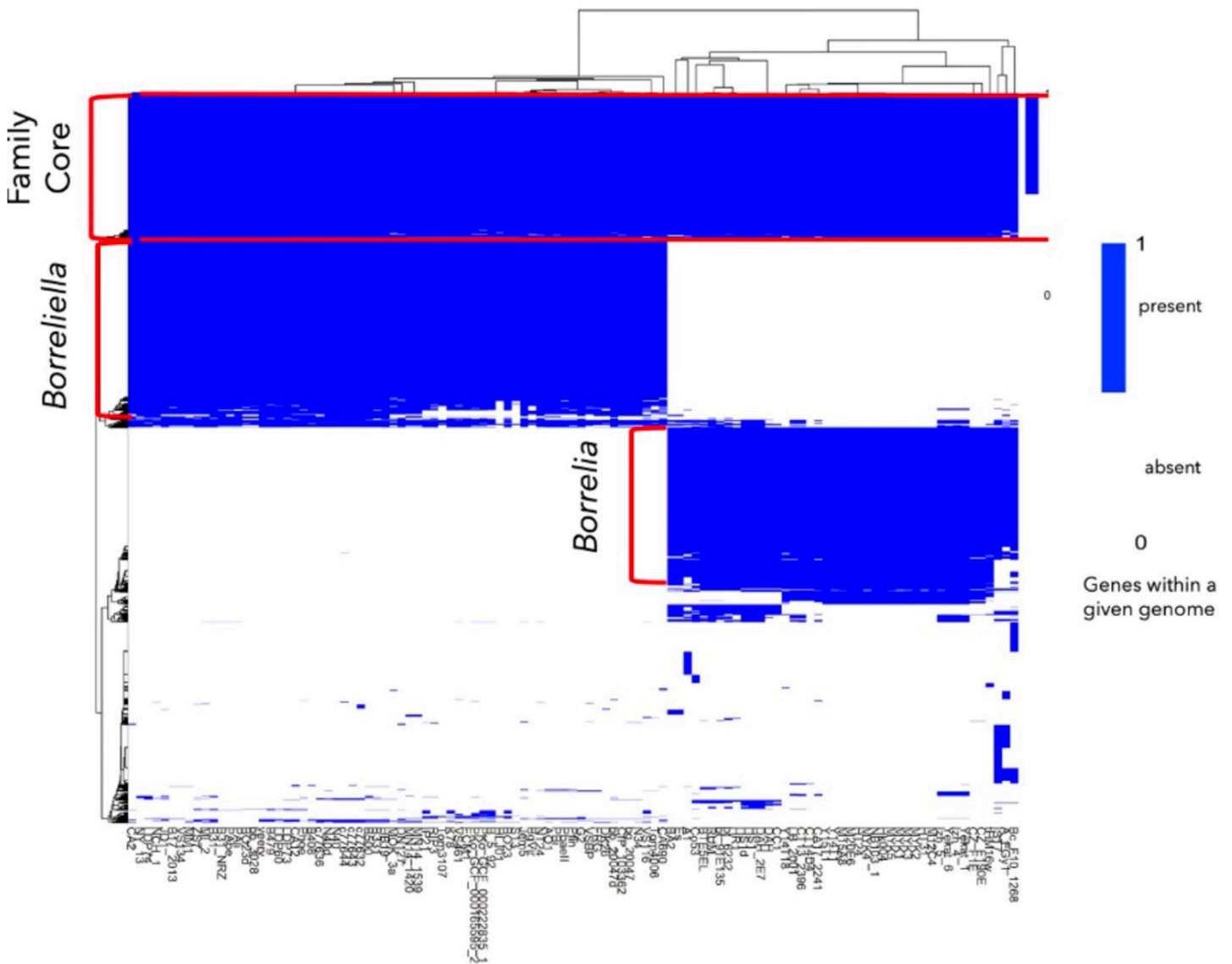


Figure 11

Borreliaceae gene presence/absence heatmap. Genes are colored blue if present and white if absent. Homologs were defined as having Roary BlastP of $\geq 75\%$. Y-axis represents gene clusters within the Borreliaceae pangenome. X-axis depicts all Borreliaceae isolates. The red annotations indicate the gene groups. The Borreliaceae pangenome has a relatively small core with each of the two genera-defined core genomes being more than twice as large. Only the most prevalent of the distributed genes (not present in either of the three core genomes) are included in the heatmap.

Figure 12

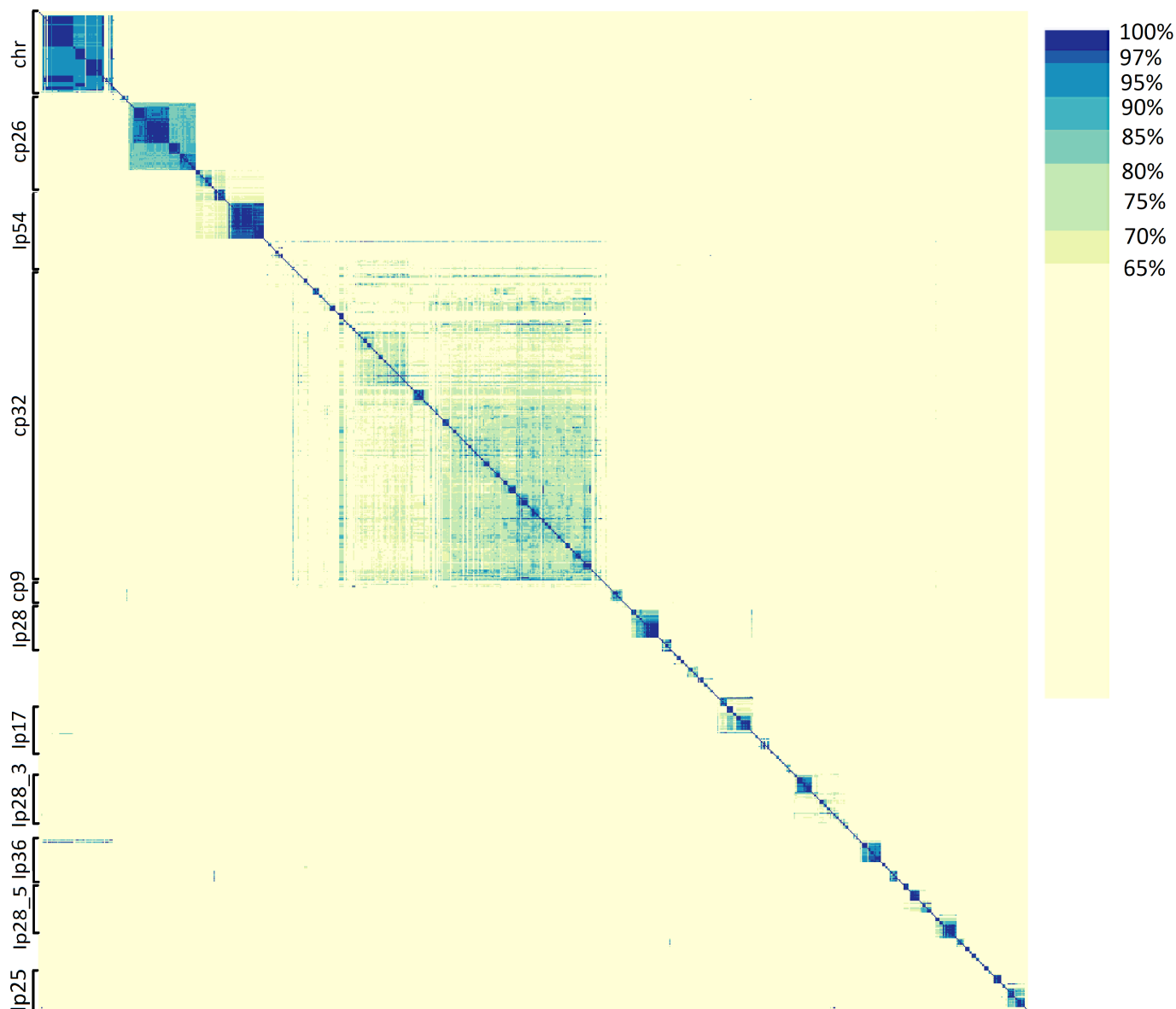


Figure 12

Large-scale average nucleotide identity of Borreliaceae replicons reveals definitive groupings across the family. The range of ANI is from least (yellow) to greatest (blue). Each of the Borreliaceae replicons is represented on the x- and y-axes based on complete-linkage hierarchical clustering. The annotations indicate the replicon group.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [FigureS1.png](#)
- [TableS1.docx](#)

- [TableS2.docx](#)