Unprecedented genetic diversity suggests importance of understudied PFam54 paralogs to Lyme borreliosis spirochetes

Janna Wülbern[1], Laura Windorfer[2], Kozue Sato[3], Minoru Nakao[4], Sabrina Hepner[5], Gabriele MARGOS[5], Volker Fingerle[5], Hiroki Kawabata[3], Noemie Becker[6], Peter Kraiczy[7], and Robert Rollins[8]

[1]Christian-Albrechts-Universitat zu Kiel
[2]Technical University of Munich School of Life Sciences Weihenstephan
[3]National Institute of Infectious Diseases
[4]Asahikawa Medical University Department of Parasitology
[5]Bavarian Health and Food Safety Authority Munich
[6]Ludwig-Maximilians-Universitat Munchen
[7]Hospital of the Goethe University Frankfurt Institute of Medical Microbiology and Infection Control
[8]Institute of Avian Research

August 28, 2023

Abstract

Lyme borreliosis (LB) is the most common vector-borne disease in the Northern Hemisphere caused by spirochetes belonging to the Borrelia burgdorferi sensu lato (Bbsl) complex. Borrelia spirochetes circulate in obligatory transmission cycles between tick vectors and different vertebrate hosts. To successfully complete this complex transmission cycle, Bbsl encode for an arsenal of proteins including the PFam54 protein family with known, or proposed, influences to reservoir host and/or vector adaptation. Even so, only fragmentary information is available regarding the naturally occurring level of variation in the PFam54 gene array and its impact on Borrelia pathogenesis. Utilizing whole genome data from isolates (n=141) originated from the three major LB-causing Borrelia species across Eurasia (B. afzelii, B. bavariensis, and B. garinii), we aimed to characterize the diversity of the PFam54 gene array in these isolates to facilitate understanding the evolution of PFam54 orthologs on an intra- and interspecies level. We found an extraordinarily high level of variation in the PFam54 gene array with 39 PFam54 paralogs belonging to 23 orthologous groups including five novel paralogs. Even so, the gene array appears to have remained fairly stable over the evolutionary history of these Borrelia species. Interestingly, genes outside Clade IV previously associated with host or, proposed, vector adaptation more frequently displayed signatures of diversifying selection. Taken together, our findings support the idea that non-Clade IV orthologs could play a larger role in host and/or vector adaptation than previously thought.

Introduction

Lyme borreliosis (LB, also termed Lyme disease in North America) is the most common vector-borne disease in the Northern Hemisphere . This disease is caused by spirochetes belonging to the *Borrelia burgdorferi* sensu lato (*Bb* sl) complex and are maintained naturally in complex enzootic transmission cycles between ixodid ticks and various vertebrate hosts . *Borrelia* possess a range of proteins encoded by genes located on a highly fragmented genome allowing spirochetes to interact with different environmental conditions in tick vectors and vertebrate hosts. The borrelial genome consists of a highly conserved, linear chromosome and up to 20 circular and linear plasmids . Although much research has gone into understanding the pathogenesis of spirochetes in humans, many open questions remain regarding the level of naturally-occurring, genetic

variation in genes encoding for proteins associated with *Borrelia* pathogenesis, vector competence, or host adaptation and the impact of this variation on underlying molecular infection mechanisms.

To establish an infection, *Bb* sl must evade host immune responses including complement, an important pillar of innate immunity. This may be either indirectly through the acquisition of complement regulators or directly through interactions with different complement components . The complement system consists of three distinct pathways (classical, lectin, and alternative) which all converge to the cleavage of the central component C3 to form activated C3b . Host cells control excessive complement damage by utilizing membrane-bound or fluid-phase regulator such as CR1 (CD35), MCP (CD46), DAF (CD55), protectin (CD59), C1q, C1-esterase inhibitor, Factor H, C4b-binding protein, clusterin, vitronectin, and FHR-1 which all can terminate the complement cascade at specific activation levels to protect self-cells . Lyme borreliosis spirochetes produce diverse outer surface proteins that bind distinct host complement components resulting in complement inactivation . One well-studied protein, CspA (*bba68* , Clade IV), capable of binding Factor H and Factor H-like protein 1 (FHL-1), belongs to the paralogous protein family PFam54. Members of the PFam54 are encoded by genes predominantly arranged in a multi-gene array located at the 5´-terminal end of the linear plasmid 54 (lp54) in the majority of *Bb* sl isolates studied . The PFam54 gene array can be separated into five major Clades where Clades I, II, III, and V share one-to-one orthology among the *Bb* sl species studied to date . Clade IV, however, contains a variable number of paralogs and many species display unique PFam54 paralogs not found in other *Bb* sl species . The PFam54 gene family is supposed to have evolved through duplication and corresponding diversification . Furthermore, gene array members have been linked to specific host adaptations as well as being differentially expressed in ticks and during infection of the vertebrate host placing them as suitable candidate genes to study their role in relation to host and/or vector adaptation.

Currently, three major *Borrelia* species act as LB-causing agents across Eurasia (*B. afzelii* , *B. bavariensis,* and *B. garinii*) offering a unique opportunity to study how spirochetes have adapted to numerous environmental conditions including diverse vertebrate hosts and tick vectors . All three of these species share an Asian origin and have expanded into at least two tick transmission cycles (*I. persulcatus* , *I. ricinus*) but only in the case of *B. bavariensis* is this expansion thought to be adaptive towards colonizing a new tick vector (*I. ricinus*) . In both transmission cycles, these three *Borrelia* species utilize either rodent (*B. afzelii* , *B. bavariensis*) or avian (*B. garinii*) reservoir hosts further suggesting variable adaptation to infect different host types . Based on these characteristics, these species offer a unique opportunity to study host and vector adaptation through comparative genomics especially in relation to candidate gene families (e.g., PFam54). Through analysis of whole genome sequencing data recently published from 136 Eurasian *Borrelia* isolates along with published reference genomes (n =5), we aimed to quantify standing, genetic variation in the PFam54 gene family across and within *B. afzelii* , *B. bavariensis* , and *B. garinii* . Our comparative analyses revealed a high level of diversity in the architecture of the PFam54 gene array including absence of genes, presence of novel paralogs, and evidence of diversifying selection along multiple PFam54 genes. These findings have the potential to open the door for novel directions in future research to determine the role of natural, genetic variation in *Bb* sl host and vector adaptation.

Methods

Isolates and reconstruction of lp54 sequences

In this study, we aimed to determine the amount of genetic variation along the PFam54 gene array in *B. afzelii, B. bavariensis* , and*B. garinii* . For reconstruction of the lp54 sequences, on which the PFam54 gene array is located, we utilized the isolate library described in containing MiSeq sequencing data for 136 Eurasian*Borrelia* isolates: *B. afzelii* (total n=33, Asian n=20, European n=13), *B. bavariensis* (total n=46, Asian n=27, European n=19), and *B. garinii* (total n=57, Asian n=25; European n=32). Asian or European here refers to an isolate arising either from the*I. persulcatus* or *I. ricinus* transmission cycle, respectively.

Sequences of the linear plasmid lp54 were assembled based on the mapping protocol recently outlined . In brief, Illumina reads (paired-end reads of 250bp) were first trimmed for Illumina MiSeq adapter sequences

using Trimmomatic v. 0.38 before being assembled using SPAdes v. 3.13.0 (Bankevich et al., 2012). SPAdes contigs were then mapped to reference lp54 sequences (PacBio Sequences: PBi, A104S, NT24, PHei, PBr, and NT31 (; GenBank: PKo (CP002950.1), K78 (CP009059.1), and ACA-1 (CP001247)) using NUCmer v. 3.23 from the package MUMmer . Lp54 identity was confirmed through searching for plasmid partitioning genes belonging to PFam32, 49, 50, and 57.62 families using BLAST v.2.8.1 (algorithm:*blastn*) as outlined in . Final lp54 sequences were determined in one of two ways: 1) a complete contig covering the full reference sequence and containing one or more of the partitioning genes mentioned above was taken without modification as the complete lp54 sequence, or 2) the lp54 was fragmented over two or more contigs for which the best reference (highest percent identity, highest overall coverage, and fewest structural variations) was used to guide reconstruction as outlined in . In this way, 136 lp54 sequences were reconstructed from NGS data to which the GenBank references used for mapping (n=3) and additional GenBank lp54 sequences (n=2; *B. bavariensis,* BgVir, CP003202.1; *B. garinii* , Far04, CP001318.1) were included in all future analyses bringing the total sample set analyzed to 141 sequences.

Identification and characterization of PFam54 gene array

Sequences for lp54-located PFam54 paralogs described in for *B. afzelii* PKo (*pko2060 -pko2071*), *B. bavariensis* PBi (*bga63-bga73*), *B. garinii* ZQ1 (*zqa66-zqa73*) and*B. burgdorferi* B31 (*bba64-bba66, bba68-70, bba73*) were downloaded from GenBank (Accession Numbers: PKo, CP002950.1; PBi, CP000015.1; ZQ1, AJ786369.1; B31, AE000790.2) and used as queries (Note: GenBank and designations are not identical but numbering is). We used BLAST v.2.8.1 (algorithm:*blastn*) to search for paralogs described above. Blast hits shorter than 500bp and with a percentage identity lower than 80% compared to the reference were not considered as paralogous. Further BLAST hits were removed if they were overlapping with regions already designated as a different paralog. BLAST hit lists were manually checked for intergenic regions >1000 bp, which were extracted and scanned for open reading frames in Aliview v1.28 . Final PFam54 gene sequences were compared against their own reference and used to produce final gene lists. Gene assignments were checked through phylogenetic reconstruction (see below). Final gene lists were used to define PFam54 gene array architecture types (i.e., structure of gene array taking gene order, content, and gene/intergenic space length into account). An architecture type was defined based on the following rules: 1) same paralogs present in the same order, 2) gene length does not differ by more than ±50bp, 3) intergenic spaces do not differ by more than ±100bp.

Phylogenetic reconstruction and selection testing

To understand the evolution of the PFam54 gene family, a robust reconstruction of the evolutionary history of the orthologous genes in all borrelial isolates analyzed was required. We opted for phylogenetic reconstruction taking codon variation into account, which should capture the evolution of protein coding sequences better than simple nuclear substitution models . Only unique PFam54 sequences (n=524 of 1308 sequences) were used for phylogenetic reconstruction. Final stop codons were removed before aligning as amino acids using MUSCLE v3.8.425 implemented in Aliview v1.28 . Phylogenetic reconstruction was run in MrBayes v. 3.2.6 with the following parameters: ploidy = haploid, codon substitution model with inverse gamma-distributed rate variation, genetic-code = universal, and assuming equal site selection (ω) . Three independent runs were launched for 50 million generations each. Convergence was checked with Tracer v. 1.7.1 . Consensus trees were built using the *sumt* command from MrBayes using a respective burn-in of 25%. Convergence to a single topology in all three independent runs was checked manually in FigTree v. 1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/). Gene orthology of individual paralogs (i.e., orthology groups) was based on the monophyletic clustering of individual gene copies either for single or multiple *Borrelia* species.

In addition to reconstructing the gene phylogeny, we further aimed to test for instances of diversifying selection. As signals of positive selection can often be hidden by negative selected sites and selection coefficients (ω_n) could vary along different branches within the reconstructed gene phylogeny, we chose a model for selection inference that allows for variation in selection pressure between branches and sites. We tested for instances of diversifying selection along the tree reconstructed with the codon model in MrBayes using

aBSREL v2.2 from the HyPhy package (https://www.hyphy.org/) using the universal genetic code and not allowing for multiple hits. Positive selection on a branch is indicated by a significant result of the LRT performed by aBSREL after multiple testing correction by the Holm–Bonferroni sequential rejection procedure. Four different kinds of branches were tested: branches separating species which utilize different vertebrate hosts (bird vs. rodent; i.e., host adaptation), branches separating isolates arising from transmission cycles involving different tick vectors (*I. persulcatus* , *I. ricinus* ; i.e., vector adaptation), genes which appear to have arisen through recombination based on our analysis (e.g.,*bga67b*), or genes known to encode proteins that have been shown to provide protection from host-immune mediated killing (*zqa68, bga66 bga71, pko2068*) . This resulted in a total of 44 branches being tested.

Determining gene gain and loss events along the PFam54 gene array

Gene gain and loss events were determined by reconstructing the phylogeny of the entire lp54 after removing any recombining region following the four-gamete condition test as outlined in . Full, corrected lp54 sequences were aligned using MAFFT v. 7.407 . Phylogenetic reconstruction was performed in MrBayes v. 3.2.6 with ploidy set to haploid and a GTR substitution model with inverse gamma-distributed rate variation. Three independent runs were launched and ran for 10 million generations. Convergence was checked with Tracer v. 1.7.1 . Consensus trees were built using the *sumt* command from MrBayes using a respective burn-in of 25%. Convergence to a single topology in all three independent runs was checked manually in FigTree v. 1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/). Gene gain or losses were then mapped onto the final lp54 tree using the maximum parsimony principle.

Statistical analysis

Clustering based on a binary matrix of orthology group (OG) presence or absence was run using classical multidimensional scaling (MDS) run with the *cmdscale* function using the base R package on a distance matrix calculated from the binary presence/absence plasmid data per isolate. Saturation curves were produced by randomly sampling architecture types 50 times for each number of isolates less than or equal to the total number of isolates present from a single species in each transmission cycle from all possible combinations of architecture types. The number of unique architecture types was determined per each random sample. The mean and standard deviation were calculated for each sampling event in R v.3.5.2 .

Results

In total, 23 unique OGs could be described in our analysis with 19 OGs present in the 141 Eurasian *B. afzelii* , *B. bavariensis* , and *B. garinii* isolates studied (Table S1). In comparison to the three *Borrelia* species analyzed, *B. burgdorferi* sensu stricto (s.s.) strain B31, used as a reference, showed four unique OGs including OG14, OG15, OG19, and OG23 (Table S1). Five novel PFam54 orthologs were identified in all species but only in isolates arising from the *I. persulcatus* transmission cycle (*pko2065b, bga67b* , *bga68b* , *bga71b* , and *zqa66b*) most of which belong to Clade IV (n=4). Certain reference genes (*pko2069* ,*pko2070* , *zqa73*) based on could not be detected in any of the 141 *Bb* sl isolates (Table S2). Both, *B. afzelii* and*B. bavariensis* displayed one major architecture type (Ba_A5,*B. afzelii* ; Bba_A11, *B. bavariensis*) (Figure 1A) while*B. garinii* displayed multiple architecture types at similar frequencies (Figure 1A). In general, we observed a high number of variable PFam54 gene array architecture types with 8, 18, and 27 different architectures found in *B. afzelii* , *B. bavariensis* , and *B. garinii,* respectively (Figure 1A, S1-S3; Table S3). Within these, Clade I, II, III, and V genes were generally present in all *Borrelia* isolates with the majority of variability arising due to differences in gene length (e.g., *zqa66* in*B. garinii* , Figure S3) or absence/presence of Clade IV genes (Figure S1-S3, Table S2). Only *pko2060* and *zqa65* were found in all *B. afzelii* and *B. garinii* isolates, respectively, with all other identified orthologs being absent from at least one *Borrelia* isolate (Table S2). MDS analysis based on the binary string of OG presence/absence showed that *Borrelia* species generally form independent groups with only a few samples not displaying clear clustering based on species identity (Figure 1B). In an effort to determine if our samples represented all possible PFam54 architecture types, we performed a saturation curve analysis. Only Asian *B. bavariensis* and *B. garinii* isolates appeared to not reach an asymptote in this analysis suggesting there could be further architecture types present in

both populations (Figure 1C, see Table S3 for additional information of individual architecture types per isolate).

Most paralogs described (74%) are hypothesized to be present at the base of the lp54 phylogeny (i.e., present for the entire evolutionary history of each *Borrelia* species studied) including two novel PFam54 orthologs, *bga68b* in *B. bavariensis* and*pko2065b* in *B. afzelii* (Figure 2). Variation in which orthologs were found per isolate appear to have arisen mostly through individual gene gain (n=12) or loss (n=56) events with losses being more common (Figure 2). *Borrelia afzelii* displayed the lowest number of gene gains or losses (n=16) followed by *B. garinii* (n=23) with*B. bavariensis* displaying the highest (n=29), but with the majority of gains/losses in *B. bavariensis* isolates coming from the *I. persulcatus* transmission cycle.

All five clades (I-V) originally described in were found in our dataset although structuring of the clades deviated with Clade III being the most basal group (Figure 3) instead of Clade I as previously described . The genes *pko2063* and *pko2064* , which were previously not assigned to a Clade, form sister clades to Clade IV and I respectively but with lower node probability (*pko2063* , p =0.57;*pko2064* , p =0.77) and remain unique genes to *B. afzelii* (Figure 3). In the phylogenetic reconstruction of the PFam54 gene family, most genes form monophyletic clades by *Borrelia*species within each OG (Figure 3), although some orthologs do not follow this pattern. For example, *zqa66* (*B. garinii*) and*bga65* (*B. bavariensis*) form a monophyletic clade (Figure 3) with all *zqa66* sequences forming a monophyletic clade within the larger *zqa66/bga65* clade (Figure 3). A similar pattern can be seen for other *B. bavariensis* and *B. garinii* orthologs (*bga68* and *zqa69* ; *bga67* and *zqa67* ;*bga64* and *zqa65*) where the genes do not form monophyletic clades purely based on species (Figure 3). Additionally, *pko2071*of *B. afzelii* and *bga73* of *B. bavariensis* form a monophyletic clade with the *B. garinii* ortholog (*zqa73*) as a sister clade to the *pko2071* /*bga73* monophyletic clade (Figure 3). Novel PFam54 orthologs clustered throughout the phylogeny with *bga67b* and *bga68b* of *B. bavariensis* being orthologs of *B. garinii* genes *zqa68* and *zqa70* , respectively (Figure 3). *Borrelia bavariensis* gene *bga71b*clusters within *bga71* but *bga71b* is paraphyletic (Figure 3). *Borrelia afzelii* gene *pko2065b* forms a sister clade to*pko2065* with *pko2066* being basal to both genes (Figure 3). Finally, *zqa66b* forms a single monophyletic clade which is a sister clade to the OG24 group (Clade III) containing *pko2062* ,*bga65* , and *zqa66* (Figure 3).

Some proteins, predominantly encoded by Clade IV PFam54 genes, are known to play a role in immune evasion with influences to host adaptation and, potentially, vector adaptation . To determine whether PFam54 orthologs of *B. afzelii* ,*B. bavariensis* , and *B. garinii* may also possess similar function, we tested 44 branches for signatures of diversifying selection. Of the branches tested, 11 branches showed significant evidence for diversifying selection (Table S4). These were found in five OGs including OG24 (Clade III), OG3 (Clade 1), OG4 (Clade V), OG10 (Clade IV), and OG21 (Figure 3). Of these, four are hypothesized to be in relation to vector adaptation as these branches separate isolates vectored either by *I. ricinus* (Europe) or *I. persulcatus*(Asia) (N588, *B. afzelii* on *pko2062* ; N526, *B. bavariensis* on *bga65* ; terminal branch leading to *B. bavariensis* isolate PBi on *bga63* and *B. garinii* isolate 20047 on *zqa68*) (Figure 3; indicated by red stars). The other seven branches are hypothesized to be in relation to vertebrate host adaptation as the branches separate *Borrelia* species which differ in their major reservoir hosts (rodent or bird) (Figure 3). Specifically, OG24 (Clade V) and OG3 (Clade I) where both branches leading either to the rodent associated clade (*B. afzelii* and/or*B. bavariensis* ; N850 & N905) and the bird-adapted clade (*B. garinii* ; *zqa63* terminal branch for ZQ1 & N884) displayed significant instances of diversifying selection. Genes with known functionality to human or host adaptation predominantly did not show signatures of diversifying selection (*pko2068* , *B. afzelii* ; *bga66* & *bga71* , *B. bavariensis*). Only the monophyletic clade containing *zqa68* and the novel *B. bavariensis* gene *bga67b* showed evidence for diversifying selection but *bga67b* alone did not (Figure 3, Table S4).

Discussion

Utilizing recently published whole genome sequences (n=136) and GenBank reference genomes (n=5) obtained from 141 Eurasian *B. afzelii* ,*B. bavariensis* , and *B. garinii* isolates , we aimed to quantify the variation along the PFam54 gene array and to analyze how PFam54 genes have evolved in these three *Bor-*

relia species. Our analyses highlighted the PFam54 gene array is more variable than previously thought, with novel paralogs found in all three*Borrelia* species studied. We also provide some of the first evidence that genes displaying signatures of diversifying selection in association with vector and/or host adaption lie outside of Clade IV which has received the majority of interest in relation to*Borrelia* infection mechanisms.

The isolates studied displayed a high number of different PFam54 architecture types which has also been partially observed for *B. burgdorferi* s. s., although based on analysis of five isolates . Architecture types do appear to be geographically restricted suggesting that they potentially have evolved over time due to the selection pressures associated with various biotic and abiotic conditions arising from different tick-host transmission cycles. Of particular interest is that *B. garinii* appears to maintain many equally frequent PFam54 architecture types in contrast to *B. afzelii* and *B. bavariensis* (Figure 1A & Figure S3). This species-specific pattern supports the notion for an underlying mechanism that triggers the adaptation of certain *Borrelia* species to specific vertebrate hosts (i.e., individual bird species) . This pattern is mirrored in the full genome where the plasmid content of *B. garinii* isolates tends to be smaller than in other *Borrelia* species but without a reduction in overall plasmid diversity when all isolates are considered, suggesting a trend towards smaller but, potentially, specialized genomes . This type of genome reduction is commonly found in pathogenic organisms and is thought to be due to adaptation to a pathogenic lifestyle . If true, one could hypothesize that *B. garinii*isolates would be variable in their ability to infect specific bird species. As it is known that *B. garinii* appears to selectively bind the complement regulator, Factor H, of avian origin to protect itself from complement mediated killing and faciliate infection of the avian host, differences in susceptibility to complement *in vitro*could be a viable proxy to determine infection capacity of *B. garinii* isolates. Recent data has demonstrated variation in the susceptibility of *B. garinii* to avian complement from different terrestrial European bird species suggesting that host association towards specific bird species could exist.

Our data revealed, regardless of variability, that the PFam54 gene array has remained relatively stable over time as most paralogs (74% of all) are predicted to be present at the base of the lp54 phylogeny. Even so, only two PFam54 paralogs, *pko2060* and *zqa65* , have been detected in all *B. afzelii* or *B. garinii* isolates, respectively, which could show that the other genes are dispensable for completing the spirochetes tick-host transmission cycle. This includes the loss of genes encoding for proteins known to interact with the complement system of vertebrates, namely BGA66 and BGA71 of *B. bavariensis* , PKO2068 of *B. afzelii* , and ZQA68 of *B. garinii* . This is supported by recent work which showed that the loss of the whole PFam54 gene array in two *B. bavariensis* isolates (PBN, PNi) impacted susceptibility to human complement but did not influence infection of mice further highlighting the probable, functional redundancy within the *Borrelia* genome . Additionally, we identified a number of, so far, undescribed PFam54 paralogs including*pko2065b* (*B. afzelii*), *bga67b* (*B. bavariensis*), *bga68b* (*B. bavariensis*), *bga71b*(*B. bavariensis*), and *zqa66b* (*B. garinii*). As*pko2065b* and *bga68b* are suggested to be present at the base of the species-specific lp54 phylogeny, it could be hypothesized that both genes could play a more integral role in the evolutionary history of *B. afzelii* and *B. bavariensis* . For this,*bga68b* , and additionally *bga67b* , are orthologs of*zqa68* and *zqa70* from *B. garinii* . Concerning the history of *bga67b* , this gene could have arisen through a recombination event between *B. bavariensis* and *B. garinii*,as suggested previously . Of note, the ortholog of BGA67b in *B. garinii* , ZQA68, displays selective binding properties to the complement regulator Factor H of avian origin, and thus protects spirochetes from complement-mediated killing by avian innate immunity . Polymorphisms in complement-interacting proteins of *B. burgdorferi* s.s. are known to influence potential host associations through complement binding Even though multiple single nucleotide polymorphisms are present when comparing the *zqa68* and *bga67b* sequences, it is tempting to speculate that the this paralog could enable Asian *B. bavariensis* isolates to circulate in birds in contrast to European*B. bavariensis* which is most likely a mammalian-associated*Borrelia* species . However, further *in vitro* studies are warranted to understand the role(s) of these novel orthologs in host adaptation and immune evasion.

The majority of branches which have experienced diversifying selection potentially in relation to host and/or vector adaptation belonged to non-Clade IV PFam54 which have not been linked to host and vector adaptation so far . Indeed, Clade IV paralogs encoding for proteins capable of binding vertebrate complement

proteins/regulators only contained two branches displaying instances of diversifying selection both in the *zqa68 /bga67b* clade (Figure 3). The majority of observations of diversifying selection in relation to vector adaptation occurred in paralogs *bga63* , *bga65, pko2062* belonging to PFam54 Clades I and III . Previous work revealed that PFam54 Clade I and III paralogs *bba64* and *bba66* from *B. burgdorferi* s. s. are variably regulated during tick infection and additional studies have shown that some PFam54 members grouped in Clade I and III are important for tick transmission of *B. burgdorferi* s. s. to the mammalian host . Clade I, III, and V contained branches also displaying instances of diversifying selection potentially linked to host adaptation. Clade I, III, and V paralogs of *B. burgdorferi* s.s. have been shown to be upregulated during infection in tick-infected mice , but loss of these genes did not significantly impede host infection although some variation in tissue tropism was observed . These results highlight that PFam54 genes outside of Clade IV are most likely important for transmission from the vector to the host and more research is needed to assess their role in host and/or vector adaptation. Further studies are ongoing to assess the function of these particular paralogs regarding their complement-inhibitory capacity.

Taken together, our results show that the PFam54 gene array of the three main causative agents of LB in Europe (*B. afzelii* , *B. bavariensis* , *B. garinii*) is highly variable in genetic architecture including gene composition and content. Moreover, the signatures of diversifying selection identified emphasize a role of paralogs belonging to Clades I, III, and V in host and vector adaptation, and thus, potentially, in functionality in the natural transmission cycles of *Borrelia* . Utilizing genomic data, we were able to elucidate the evolution of an important gene family and were able to generate testable hypotheses regarding which genes should be studied in relation to host and vector adaptation. Taken together, our comparative analyses highlight the importance of investigating individual diversity in *Borrelia* species from a population genetics perspective. So as to increase our understanding and guide future studies into *Borrelia* pathogenesis, with the goal to determine how vector-borne pathogens evolve leading to the emergence of wildlife and human disease.

Acknowledgments

We would like to thank all past members of the Evolutionary Biology group at the LMU; Cecilia Hizo-Teufel, Wiltrud Strehle, Christine Hartberger, and Sylvia Stockmeier and Nikolas Alig from the National Reference Centre for *Borrelia* , the National Institute of Infection Diseases in Tokyo for allowing us to work in their facilities, all members and especially Dr. Niels Dingemanse of the Behavioural Ecology group at LMU for allowing us to collect ticks in their study plots. We would further like to thank Georg Manthey and Dr. Kristian Ullrich for his support in bioinformatical analyses.

Author contributions

RER, JW, NSB, and PK designed the study concept. RER, NSB, HK, SK, SH, GM, and VF produced and/or provided sequencing data. RER assembled all sequence data with the guidance of NSB. PFam54 gene extraction and analysis was done be JW with guidance from RER & NSB. RER & JW ran all phylogenetic analyses with help from LW who ran all selection analyses. RER and PK wrote the manuscript which was approved and read and approved by all co-authors.

Research funding

The project was funded through the German Research Foundation (DFG Grant No. BE 5791/2-1) (NSB, RER). The National Reference Centre for*Borrelia* was funded by the Robert-Koch-Institut, Berlin (VF, GM). Part of this work was supported by the LOEWE Center DRUID (Novel Drug Targets against Poverty-Related and Neglected Tropical Infectious Diseases), projects C3 (PK).

Data availability

All lp54 sequences and SRA files are in the process of being uploaded to GenBank under the BioProject numbers PRJNA327303, PRJNA449844, and PRJNA722378. All alignments, finalized PFam54 gene lists, finalized trees, MrBayes trace files, and HyPhY output files will be uploaded to a Dryad repository upon acceptance of the manuscript.

Conflicts of interest

The authors have no conflicts of interests to state.

References

Figures and Tables:

Figure 1. Diversity and prevalence of PFam54 architecture types identified in the 141 Eurasian isolates of *B. afzelii* , *B. bavariensis* , and *B. garinii* analyzed in this study. A) Frequency of different architecture types present in the studied isolates (n=141). Architecture types are separated by species (*B. afzelii* , *B. bavarinesis* , *B. garinii*). Colors correspond to tick transmission cycle (dark grey, *I. persulcatus* ; light grey, *I. ricinus*). For information on the architecture of individual isolates see Table S3. B) MDS clustering based on presence/absence matrix of orthology groups (OG) in each isolate. Here each point corresponds to a single*Borrelia* isolate. C) Saturation curve analysis produced per species per transmission. In panels B and C, colors correspond to*Borrelia* species (purple, *B. afzelii* ; orange, *B. bavariensis* ; blue, *B. garinii*) and shapes correspond to transmission cycle (*I. persulcatus* ,*I. ricinus*∗).

Figure 2. Gene gain and loss events of PFam54 paralogs based on the principle of maximum parsimony mapped onto the phylogenetic tree reconstructed based on full lp54 sequences corrected for recombination based on the four-gamete condition test described in and with the PFam54 gene array removed. Phylogenetic reconstruction was performed in MrBayes v. 3.2.6 with ploidy set to haploid and a GTR substitution model with inverse gamma-distributed rate variation. Three independent runs were launched and ran for 10 million generations at which point convergence of parameters was checked with Tracer v. 1.7.1 . Consensus trees were built using the *sumt* command from MrBayes using a respective burn-in of 25%. Convergence to a single topology in all three independent runs was checked manually in FigTree v. 1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/). Colors correspond to which tick transmission cycle an isolate comes from (*I. persulcatus* or *I. ricinus*).

Figure 3. Phylogenetic tree of all, unique PFam54 paralogs identified in our analysis. In total, 1302 paralogs were identified which represented 524 unique sequences. Phylogenetic reconstruction was run in MrBayes v. 3.2.6 with ploidy set to haploid and a codon substitution model with inverse gamma distributed rate variation, the universal genetic code, and assuming equal selection (ω) . Three independent runs were launched and ran for 50 million generations at which point convergence of parameters was checked with Tracer v. 1.7.1 . Consensus trees were built using the *sumt* command from MrBayes using a respective burn-in of 25%. Convergence to a single topology in all three independent runs was checked manually in FigTree v. 1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/). All internal nodes which had probabilities lower than 0.95 are shown in light grey. Orthology groups (OGs), based on the monophyletic clustering of an individual gene copy within the tree either for a single or multiple species, are shown in the outer ring. Tick transmission cycle is shown either was dark grey (*I. persulcatus*) or light grey (*I. ricinus*) in the inner ring. Individual paralog placements in the phylogeny are marked outside of these rings. *Borrelia s* pecies are denoted by tip end color: purple (*B. afzelii*), orange (*B. bavariensis*), blue (*B. garinii*). Roman numerals (I-V) denote the PFam54 Clades as described in . Red stars denote branches which were found to show significant instances of diversifying selection as determined by aBSREL v2.2 from the HyPhy package (https://www.hyphy.org/) using the universal genetic code and not allowing for multiple hits.