


Highly Resolved Genomes of Two Closely Related Lineages of the Rodent Louse *Polyplax serrata* with Different Host Specificities

Jana Martinů¹, Hassan Tarabai^{1,2}, Jan Štefka^{1,3}, and Václav Hypša ^{1,3,*}

¹Department of Parasitology, Faculty of Science, University of South Bohemia, České Budějovice, Czech Republic

²Central European Institute of Technology (CEITEC), University of Veterinary Sciences, Brno, Czech Republic

³Institute of Parasitology, Biology Centre, The Czech Academy of Sciences, České Budějovice, Czech Republic

*Corresponding author: E-mail: vacatko@prf.jcu.cz.

Accepted: February 27, 2024

Abstract

Sucking lice of the parvorder Anoplura are permanent ectoparasites with specific lifestyle and highly derived features. Currently, genomic data are only available for a single species, the human louse *Pediculus humanus*. Here, we present genomes of two distinct lineages, with different host spectra, of a rodent louse *Polyplax serrata*. Genomes of these ecologically different lineages are closely similar in gene content and display a conserved order of genes, with the exception of a single translocation. Compared with *P. humanus*, the *P. serrata* genomes are noticeably larger (139 vs. 111 Mbp) and encode a higher number of genes. Similar to *P. humanus*, they are reduced in sensory-related categories such as vision and olfaction. Utilizing genome-wide data, we perform phylogenetic reconstruction and evolutionary dating of the *P. serrata* lineages. Obtained estimates reveal their relatively deep divergence (~6.5 Mya), comparable with the split between the human and chimpanzee lice *P. humanus* and *Pediculus schaeffi*. This supports the view that the *P. serrata* lineages are likely to represent two cryptic species with different host spectra. Historical demographies show glaciation-related population size (N_e) reduction, but recent restoration of N_e was seen only in the less host-specific lineage. Together with the louse genomes, we analyze genomes of their bacterial symbiont *Legionella polyplacis* and evaluate their potential complementarity in synthesis of amino acids and B vitamins. We show that both systems, *Polyplax/Legionella* and *Pediculus/Riesia*, display almost identical patterns, with symbionts involved in synthesis of B vitamins but not amino acids.

Key words: sucking lice, Anoplura, genomics, symbiosis.

Significance

Sucking lice are an extremely specialized group of parasitic insects, with many unique characteristics. Due to the morphological and behavioral changes during evolution of their unique lifestyle, they could provide a valuable model for studying genomic changes and adaptations. Despite this potential, a complete genome is currently available for only a single species, *Pediculus humanus*. Here, we present genomes for two lineages of another, phylogenetically distant species, *Polyplax serrata*, and compare their characteristics to the *P. humanus* genome. Such analyses are required to distinguish between the features common to all sucking lice due to their parasitic lifestyles (e.g. loss of receptor-associated genes) and those unique to a particular species (e.g. reduction of genome size and GC content).

© The Author(s) 2024. Published by Oxford University Press on behalf of Society for Molecular Biology and Evolution.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

Introduction

Among several thousands of insect genome records currently available in the National Center for Biotechnology Information (NCBI), only six represent permanent ectoparasites of the infraorder Phthiraptera, and just two of them the parvorder of sucking lice Anoplura, both representing a single species *Pediculus humanus*. Of these genomes, three have been analyzed and published in the form of a genomic comparative study. One of them, *P. humanus* (Kirkness et al. 2010), represents sucking lice, and the other two, *Columbicola columbae* and *Brueelia nebulosa*, are chewing lice of the family Philopterae (Baldwin-Brown et al. 2021; Sweet et al. 2023). In this study, we sequence genomes of two phylogenetically distinct lineages of the sucking louse *Polyplax serrata*, providing genomic data for a second anopluran species and allowing for the comparison between different sucking louse genera. Such comparison is important to distinguish which features of *P. humanus* stressed in the previous comparative analyses of phthirapteran genomes (Kirkness et al. 2010; Baldwin-Brown et al. 2021; Sweet et al. 2023) represent common anopluran characteristics and which are unique to this species (e.g. highly reduced genome size and loss of sensory genes).

In addition to this general significance, the genomes of the two *P. serrata* lineages provide an important background for evolutionary studies of this louse species. The lice classified as *P. serrata* were shown in series of studies to form a complex assembly of populations with different distributions and ecologies (Stefka and Hypsa 2008; Martinů et al. 2015, 2018, 2020). The most conspicuous difference across this assemblage is the degree of host specificity in two different lineages, namely the “nonspecific lineage” (N lineage) parasitizing two host species (*Apodemus sylvaticus* and *Apodemus flavicollis*) and the “specific lineage” (S lineage) restricted only to *A. flavicollis*. In fact, the phylogenetic and genealogical patterns obtained in the previous analyses indicate that these lineages are likely to represent two different but morphologically indiscernible species.

Finally, like all other sucking lice, *P. serrata* maintains obligate symbiosis with a bacterial symbiont *Legionella polyplacis* (Rihova et al. 2017). Such mutualistic associations with bacteria are known from various ecological types of insects. Genomes of these symbionts typically undergo dramatic changes, mostly losses of genes, and their metabolic capacities evolve to reflect the host’s lifestyle and source of the diet (Kinjo et al. 2022). In blood-feeding insects, complete genomes of the host and the symbiont have so far been available only for tsetse flies of the genus *Glossina* and their symbiont *Wigglesworthia glossinidia* (Akman et al. 2002; Rio et al. 2012) and for *P. humanus* with the symbiont *Candidatus Riesia pediculicola* (Kirkness et al. 2010; *Riesia pediculicola* hereafter).

Results and Discussion

Genome Assemblies and Annotation

Two high-quality genome assemblies were generated combining Oxford Nanopore and Illumina reads, with Nanopore coverage 108x for the “specific lineage” (designated as “S lineage” throughout the text and “*P. serrata S*” in figures and tables) and 42x for the “non-specific lineage” (designated as “N lineage” throughout the text and “*P. serrata N*” in figures and tables). The assemblies yielded similar main characteristics for the two *P. serrata* lineages (Table 1). For S lineage, the total assembly size was 138.66 Mbp with the largest scaffold covering 20.69 Mbp and scaffold N50 of 10.50 Mbp. The genome encoded 14,045 predicted genes for 13,914 mRNAs and 131 tRNAs. For N lineage, the assembly reached 138.54 Mbp, the largest scaffold covered 17.59 Mbp, and scaffold N50 was 13.3 Mbp. Benchmarking Universal Single-Copy Orthologs (BUSCO) genome completeness analysis found 98.7% (1,000/1,013) and 98.9% (1,002/1,013) complete BUSCOs of *P. serrata S* and N lineages, respectively (supplementary table S1, Datasheet S1, Supplementary Material online). Gene prediction of *P. serrata N* revealed 15,132 genes for 14,991 mRNAs and 141 tRNAs. The repeat sequences constituted 4.93 Mbp (3.56%) of the genome size in S lineage and 4.84 Mbp (3.49%) in the N lineage. In both the S and N lineages, simple repeats constituted 49% of all identified repeat elements, while interspersed repeats accounted for 40% in each lineage. A slightly higher difference was observed for the content of long interspersed nuclear elements (LINEs), which constituted 0.52% of total genomic length in S lineage compared with 0.39% in N lineage (see supplementary table S1, Datasheet S2, Supplementary Material online, for detailed breakdown of the identified repeat elements).

In contrast to the high similarity of the two closely related S and N lineages, *P. serrata* differs considerably from *P. humanus*, the only other sucking louse for which genome assembly is available (Table 1). Generally, *P. serrata* genomes are larger, with a higher number of genes and a higher GC content. A striking difference is the presence of 25/26 rRNA genes in the *P. serrata* lineages compared with the 561 rRNA genes in *P. humanus*. The majority of this rRNA contents in the human louse (536 genes) is 5S rRNAs, known to be extremely variable in number of copies (Ding et al. 2022). However, the differences between the two anopluran genera are relatively small when compared with the differences between the two chewing lice included in the study. While the genome of *B. nebulosa* is smaller but comparable in size with the anopluran genomes, the genome of *C. columbae* is almost twofold larger.

Table 1

Overview of the genome characteristics of the analyzed Phthiraptera species

Genome characteristics	<i>P. serrata</i> S	<i>P. serrata</i> N	<i>P. humanus</i>	<i>B. nebulosa</i>	<i>C. columbae</i>
Assembly length (bp)	138,661,288	138,547,654	110,770,411	113,965,818	207,887,661
Num scaffolds	89	70	1,873	1,684	384
GC content (%)	36.90	36.92	26.87	37.84	36.11
N50	10,507,470	13,306,208	497,057	636,874	17,673,050
Num genes	14,045	15,132	12,130	15,901	25,246
Num protein-coding genes	13,914	14,991	12,004	15,814	25,113
Num tRNA-coding genes	131	141	126	87	133
Num rRNA-coding genes	26	25	561	34	36
Average gene length (bp)	2,667	2,335	2,574	2,248	2,098
Transcript CDS	13,914	14,991	12,004	15,814	25,113
CDS complete	13,848	14,520	11,493	15,216	24,485
Total exon no.	76,836	78,485	71,562	83,006	109,157
Total exon CDS	76,836	78,485	71,562	83,006	109,157

Comparative Genomic Analysis

The genomes of the two phylogenetically distant species, *P. serrata* and *P. humanus*, share a high proportion of their protein contents. Over 93% of the annotated proteins from the Pfam and InterPro databases were conserved across the three analyzed anopluran genomes (Fig. 1, [supplementary table S1](#), [Datasheets S3 and S4](#), [Supplementary Material](#) online). The analyses using several different databases provided similar results, indicating a high number of shared features and only few unique features (Fig. 1; [supplementary table S1](#), [Datasheets S5 to S7](#), [Supplementary Material](#) online), with the exception of predicted signal peptides, showing lower number in *P. humanus* ($n = 702$) compared with that of *P. serrata* S ($n = 1,049$) and N ($n = 1,063$) (Fig. 1, [supplementary table S1](#), [Datasheet S7](#), [Supplementary Material](#) online).

The comparison of the three anopluran genomes with the two available genomes of chewing lice and other blood-feeders is more complex ([supplementary fig. S1](#), [Supplementary Material](#) online; see [supplementary table S1](#), [Datasheet S8](#), [Supplementary Material](#) online, for list and references of all compared genomes). In their analysis, Baldwin-Brown et al. (2021) pointed out that chewing louse *C. columbae* and sucking louse *P. humanus* possess reduced numbers of opsins (two and three, respectively). This is a much lower number than we detected in some blood-feeding dipterans (e.g. 20 in *Aedes aegypti*, 11 in *Glossina morsitans*) but comparable with blood-feeding heteropterans *Cimex lectularius* (2) and *Rhodnius prolixus* (5). In the *P. serrata* genomes presented here, the opsin genes (InterPro IDs: IPR027430, IPR001760, and IPR001391) were entirely missing, reflecting the fact that in this species eyes are lost.

An even stronger effect of the permanent parasitism on gene loss was observed in the repertoire of olfactory receptors (IPR004117). We obtained this annotation for only 18

genes in *B. nebulosa*, 21 in *C. columbicola*, 13 in *P. humanus*, and 15 in both *P. serrata* lineages, compared with much higher numbers in temporary insect parasites (which regularly have to search for the host, in contrast to the permanent ectoparasites), specifically 38 in *C. lectularius*, 49 in *G. morsitans*, 66 in *A. aegypti*, and 153 in *R. prolixus*. We also detected a single gene associated with taste receptor activity (GO: 0008527) in each of the *P. serrata* lineages. In agreement with the Baldwin-Brown et al. (2021) report, we found two such genes in the *C. columbae* genome, but we did not detect this GO in *P. humanus*. However, when considering all annotations defined as “taste connected,” we found 16 genes in *B. nebulosa* and *C. columbae*, 5 in *P. humanus*, 13 in *P. serrata* N, and 10 in *P. serrata* S. Apart from these differences in the repertoire of insect genes, we also observed various numbers of Rhabdovirus-related genes. Specifically, we identified 33 hits in *P. serrata* N lineage, compared with 10 hits in S lineage, 3 hits in *P. humanus*, and single hit in *C. columbae* ([supplementary table S1](#), [Datasheet S4](#), [Supplementary Material](#) online). This group of viruses is known to be broadly distributed across a wide range of organisms, including plants, insects, and vertebrates (Ammar et al. 2009). It was also demonstrated that in insects the rhabdoviruses can be vertically transmitted to progeny (Longdon et al. 2017). It is therefore difficult to hypothesize on a possible evolutionary or ecological significance of this finding in our data.

When visualizing genome dissimilarities of the five phthirapteran species by principal coordinate analysis (PCoA), the plots well reflected their phylogenetic relationships and evolutionary distances. For the results obtained from the family-centered Pfam database, the pattern was straightforward with the two *P. serrata* positioned as two closest points and *B. nebulosa* as the genome most distant from all others (Fig. 2A). In the InterPro-based PcoA, the distant position of *B. nebulosa* lowers resolution between the

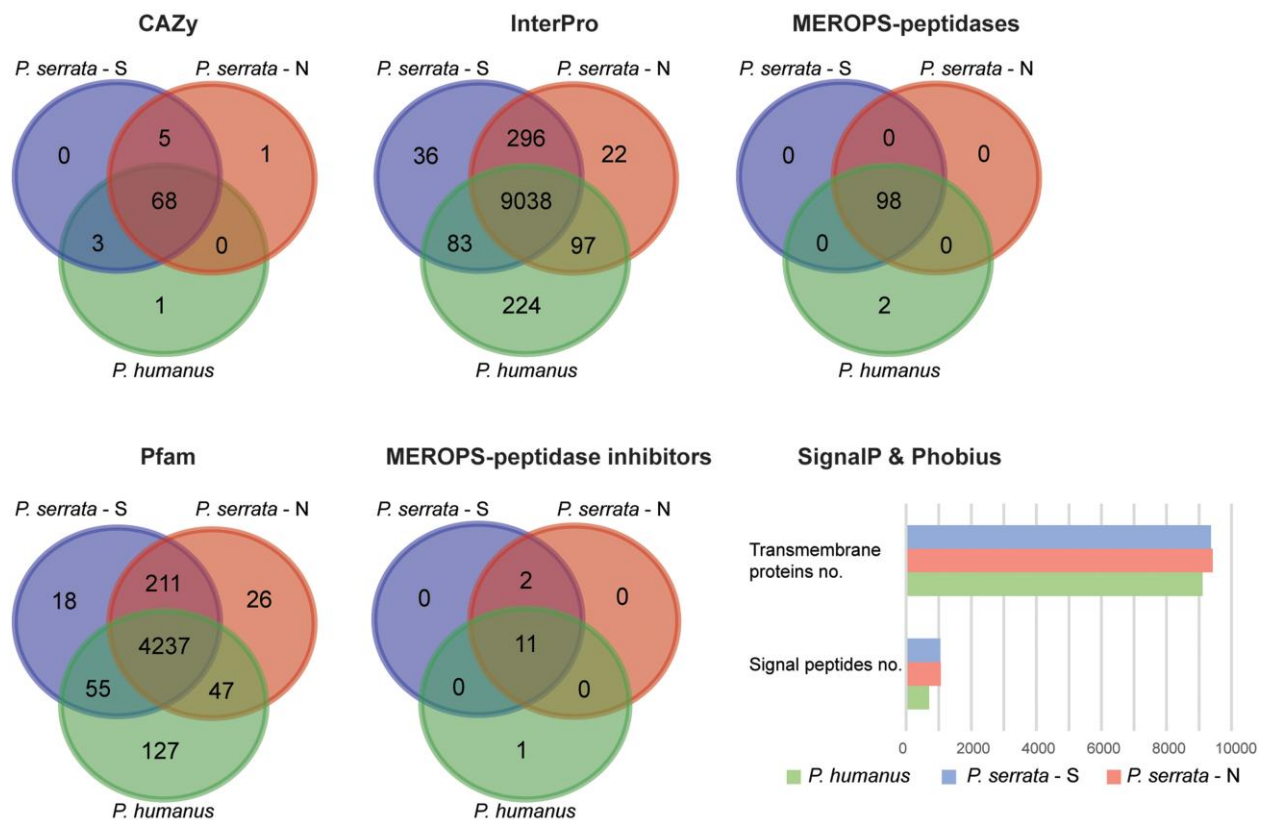


Fig. 1.—Comparison of genome contents of the three sucking lice (Anoplura). For the databases CAZy, InterPro, MEROPS, and Pfam, the numbers represent unique IDs identified in the genomes. For SignalP and Phobius, the plot shows total numbers of the transmembrane proteins and signal peptides identified by the databases.

anopluran genera (Fig. 2B). However, after the removal of the *B. nebulosa* outlier, the analysis reveals close proximity of the two *P. serrata* lineages in comparison with significantly more distant *P. humanus* (Fig. 2C).

Comparative analysis of clusters of orthologous groups (COGs) revealed that 8,138 of the total 13,129 COGs were shared by all genomes, while the 1,181 COGs shared by the 2 *P. serrata* lineages formed the second largest subset (supplementary fig. S2A, Supplementary Material online). Surprisingly, only 116 COGs were unique among Anoplura compared with chewing lice species. Contraction/expansion analysis of gene families indicated considerable changes along the evolution of the analyzed genomes (supplementary fig. S2B and table S2, Datasheets S1 to S4, Supplementary Material online). Compared with other nodes presented in the supplementary fig. S2B, Supplementary Material online, Anoplura underwent the lowest number of changes, ten of them statistically significant (supplementary table S2, Datasheet S5, Supplementary Material online). However, additional genomic data encompassing a broader phylogenetic sample of Anoplura are necessary

to make any biological interference and to identify changes potentially driving Anoplura's adaptations.

Synteny and Chromosomes

Considering the genome size around 139 Mbp in the *P. serrata* S and N lineages (Table 1) and the number of chromosomes in *Polyplax* lice (Golub and Nokkala 2004), we presume that the longest scaffolds in both *P. serrata* lineages (17 to 20 Mbp) likely represent almost complete chromosomes. High contiguity of the two assemblies allowed for comparing their synteny on a macroscale level. Contigs longer than 0.7 Mbp were chosen for collinearity analyses. Pair-wise comparison of the 18 longest contigs of S lineage (99.3% of the genome) and 21 of N lineage (98.7%) revealed a high degree of synteny with 82.78% of collinear genes (Fig. 3A). The only observed structural rearrangement was a translocation of a short fragment (44.5 kb) between scaffolds PS6 in *Polyplax* S lineage and the region covering scaffolds PN12 and PN18 of *Polyplax* N lineage (Fig. 3B). As high level of synteny is crucial for sexually reproducing species during recombination, its

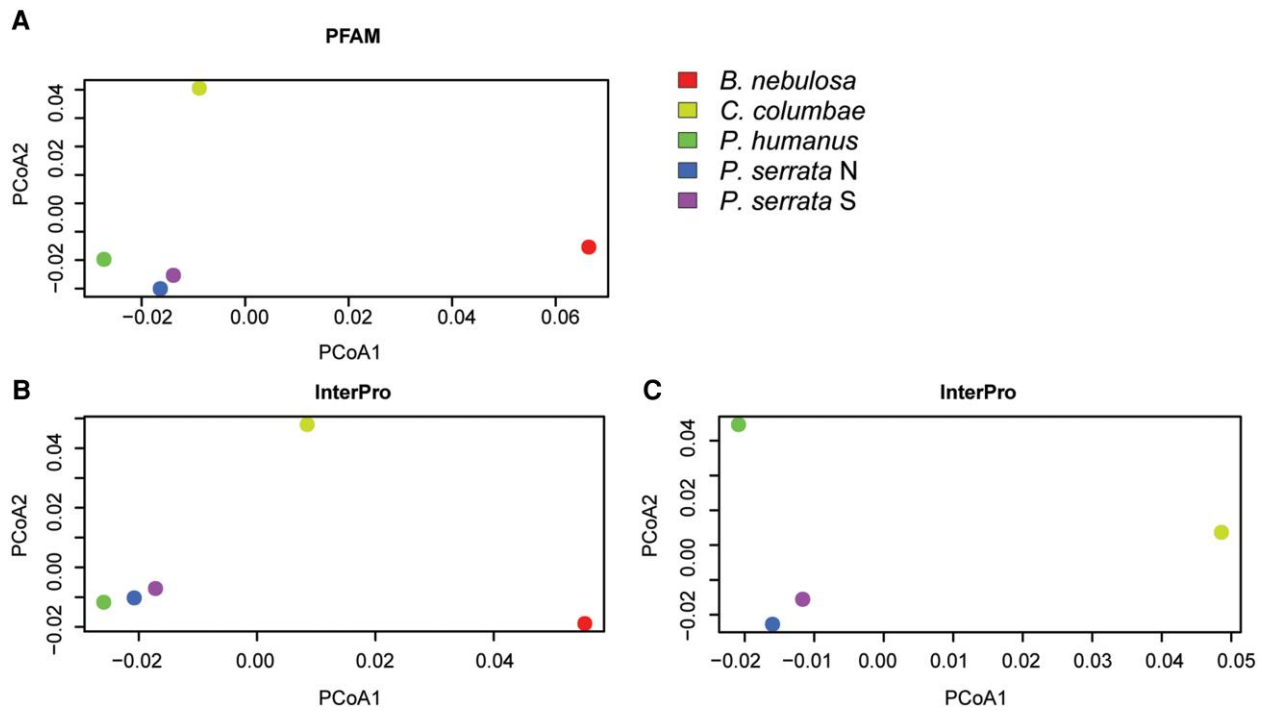


FIG. 2.—PcoA analysis of the five phthirapteran genomes. A) PCoA based on the results of the family-centered Pfam database. B) PCoA based on the InterPro database. C) PCoA based on the InterPro database.

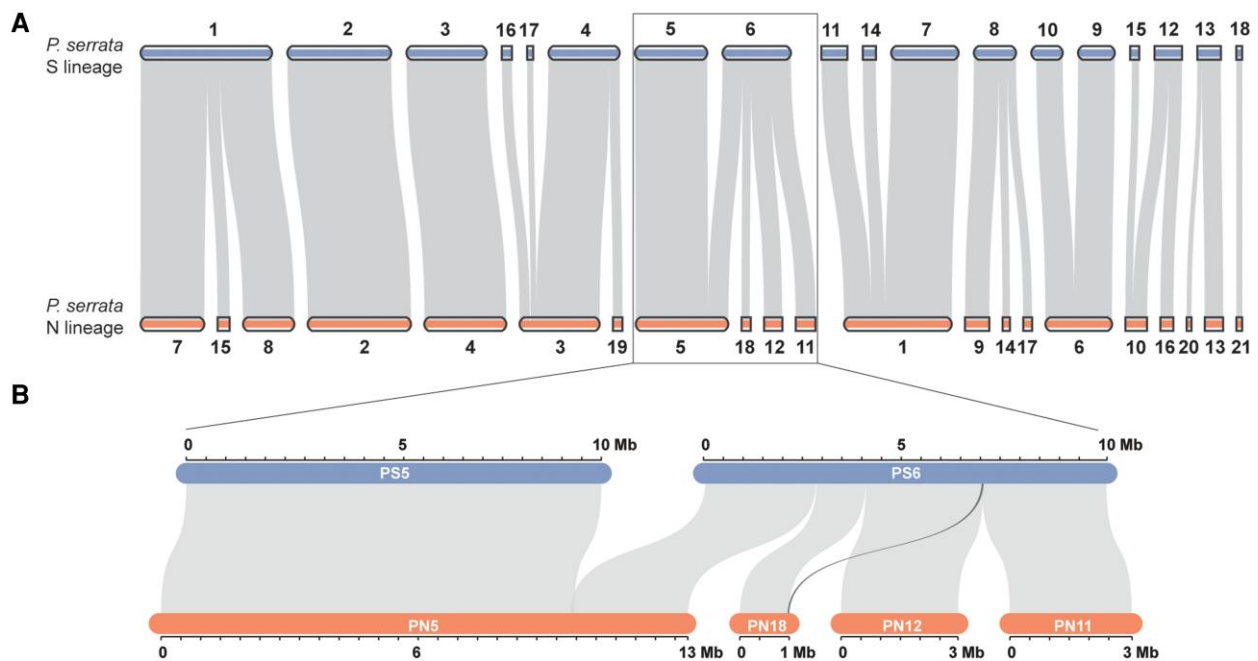


FIG. 3.—Dual synteny plot of longest scaffolds of *P. serrata* S lineage (upper set of contigs) and N lineage (lower set of contigs). A) Syntenic regions identified by McScanX are highlighted by light gray. B) Detailed synteny and dot plots of structural variations between the scaffolds PS5-PS6 from S lineage and PN5-PN11-PN12-PN18 from N lineage visualized by SynVisio; the translocation is highlighted by dark gray.

loss can have large impact on the success of mating and viability of the progeny (Feulner and De-Kayne 2017; Simakov et al. 2020).

Genome of *L. polyplacis* and Host–Symbiont Complementarity

Both sucking lice, *P. serrata* and *P. humanus*, have previously been found to live in obligate symbiosis with bacteria *L. polyplacis* and *R. pediculicola*, respectively (Allen et al. 2007; Rihova et al. 2017). While our *P. serrata* hybrid Nanopore–Illumina assemblies contained only fragmented and incomplete genomes of *L. polyplacis*, SPAdes assemblies of Illumina reads produced complete circular genomes (529,751 and 530,980 bp). The genomes were highly similar to those of the *L. polyplacis* samples reported previously (Rihova et al. 2017; Martinů et al. 2020). To compare the host–symbiont associations in *Polyplax* and *Pediculus*, we analyzed metabolic capacities in two categories usually considered in relation to the insect–bacteria symbiosis, B vitamins, and amino acids. For both categories, the two anopluran genera show high similarity (Table 2). The capacity for amino acids synthesis is in both lice determined almost strictly by the host genomes, while in the symbionts, the pathways are largely deteriorated. The only difference between the two systems consists in the enzyme allowing conversion between serine and glycine, present in the *Riesia* genome of *Pediculus*. This amino acid pattern corresponds to the general view that, in contrast to the insects feeding on plant saps, the blood-feeding groups do not depend on their symbionts for amino acids. In contrast, at least three B vitamins (riboflavin, B2; biotin, B7; and folate, B9) are provided by the symbionts. In both louse systems, the bacterial folate pathway lacks only one of the ten required enzymes (according to the Kyoto Encyclopedia of Genes and Genomes [KEGG] module M00126). This high degree of conservation indicates that the pathway is functional and the missing reaction is fulfilled by an unknown gene/enzyme (Říhová et al. 2023). Comparison with the assembled *P. serrata* genomes shows that one possible candidate is the louse alkaline phosphatase. This makes the folate biosynthesis the only possible candidate for complementarity between the louse and the symbiont (i.e. the gene missing in the pathway encoded by the symbiont is supplemented by the host). The simple lipoic acid (LA) pathway is coded independently by both lice and the symbionts. A clear difference between the *Polyplax* and *Pediculus* systems represents the pantothenate pathway. In *Polyplax*, the *Legionella* symbiont lacks the key genes for the pantothenate synthesis and is therefore not able to produce this vitamin and provide it to the host. In *Pediculus*, *Riesia* has been reported to carry a plasmid with the pantothenate genes, suggesting that this symbiont may serve as a pantothenate source for the host (Boyd et al. 2014).

Table 2 Comparison of metabolic capacities for amino acids, B vitamins, and LA in the two louse/symbiont systems

	Amino acids										B vitamins																			
	Ala	Arg	Asn	Asp	Cys	Glu	Gln	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyrosine	Val	B1	B2	B3	B5	B6	B7	B9	LA		
<i>Polyplax/Legionella</i>	H	H	H	H	H	B	H	H	H	H	H	H	H	*	H	H	H	H	*	S	S	S	S	S	S	S	S	S	C	B
<i>Pediculus/Riesia</i>	H	H	H	H	H	H	H	H	H	H	H	H	H	*	H	H	H	H	*	S	S	S	SP	S	S	S	S	C	C	B

The potentially functional pathways are designated by the gray background, while contribution of the louse and symbiont genomes is indicated by the letters: H, full pathway is encoded by the host genome; S, the pathway is coded by the symbiont; C, possible complementarity (see text); B, full pathway is present in both counterparts; SP, pathway encoded on symbiont's plasmid; *, enzyme for Phe/Tyr conversion present in *Polyplax*.

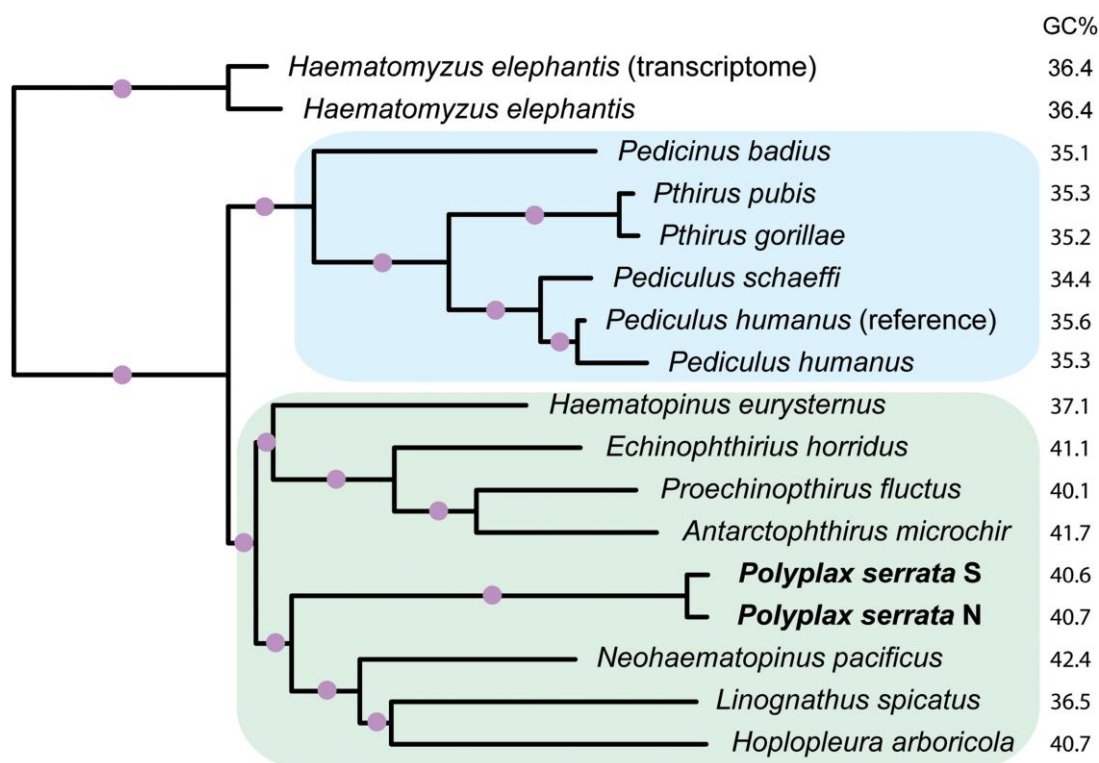


FIG. 4.—Phylogenetic tree derived by ML (IQ-TREE 2) from a matrix of 1049 orthologs (only second codon positions; 536,051 sites). The new *Polyplax* genomes printed in bold. GC% = GC content in full concatenated matrix (i.e. all codon positions). All bootstrap values reached 100 (indicated by the purple dots). The colored background designates the two branches with different ranges of GC content: blue, primate-associated lice; green, lice associated with carnivores, ungulates, and rodents.

Phylogeny and Evolutionary Dating

Using maximum likelihood (ML)-based analysis, we placed *P. serrata* lineages into the phylogeny of Anopluran taxa for which genomic data are available. The analysis produced a tree with strong bootstrap supports (Fig. 4). Its topology agrees with the arrangement of anopluran species in the tree published by de Moya et al. (2021). The first dichotomy lies between the primate-associated lice (*Pedicinus*, *Pediculus*, *Pthirus*) and the remaining taxa. In the latter clade, the two *Polyplax* lineages branched as a sister taxon to the *Hoplopleura* + *Linognathus* + *Neohaematopinus* cluster. This position of *P. serrata* is in conflict with the phylogeny previously published (Light et al. 2010) which placed *P. serrata* as a sister group of the primate-associated genera. Since the tree we present here is based on a large amount of data and supported by high bootstrap values, we consider this topology a reliable representation of the *P. serrata* position within Anoplura. Moreover, this topology better reflects differences in the GC content of the analyzed genomes. Since complete genomes are not available for most of the included Anoplura taxa, we used GC content of the selected set of genes as a proxy. The comparison shows that the striking difference between the *P. humanus* and *P. serrata* genomes in GC content

(Table 1) fits into the general pattern of the GC along the phylogenetic tree (Fig. 4).

Dating analysis based on the calibrations known for the primate-associated lice produced an estimate of ~6.5 Mya for the split between the N and S lineages of *P. serrata* (supplementary fig. S3, Supplementary Material online). This dating places the origin of the N and S lineages considerably deeper than the estimate 1.5 Mya reported previously (Stefka and Hypsa 2008). Similar to the phylogenetic reconstruction described above, our time estimate obtained here is based on considerably larger data than in the previous study (1,049 genes compared with 3). Moreover, the branch lengths of the *P. serrata* lineages shown in Fig. 4 are comparable with those of different *Pediculus* species (this is also reflected in divergences of *cox1* genes, 17% between the two *Pediculus* species, and 18.5% between the two *Polyplax* species). This estimate and the argument should however be considered with caution. On one hand, the comparison between the two MCMCtree runs (regression of the date series; supplementary fig. S4, Supplementary Material online) shows that the analysis reached good convergence and the estimates are not affected by sampling error. On the other hand, however, the estimate falls into a broad confidence interval (95% highest probability density

[HPD] = 1.8 to 11 Mya). Moreover, the calculation is based on calibration points available only for the primate-associated cluster, presuming similar evolutionary tempo in both louse clusters shown in Fig. 4.

Regardless of the exact divergence time, the *P. serrata* lineages N and S seem to represent two distinct species, which differ in an important parameter of their lifestyles, namely host specificity/spectrum (with the S lineage strictly specific to a single host *A. flavicollis* while N lineage capable to live also on *A. sylvaticus*). Retention of a close morphological similarity over millions of years is not exceptional (Shin and Allmon 2023). Also, the two *Pediculus* species, human louse *P. humanus* and chimpanzee louse *P. schaeffi*, are difficult to distinguish despite ~6 My of independent evolution (Reed et al. 2007), although a few morphological features differentiate them (e.g. the width of the thorax). In connection to the *P. serrata* lice, it is pertinent to note that it is extremely difficult (in some cases impossible) to distinguish morphologically also their two host mouse species, estimated to have diverged ~4 Mya (Michaux and Pasquier 1974 as cited in Michaux et al. 2005). Thus, it is possible that the morphological similarity of *P. serrata* hosts and the fact that they share one of the hosts (*A. flavicollis*) conserved morphology of the N and S lineages via stabilizing selection.

Population Demography

To reveal congruence or divergence in coalescence rates for genomic data between the S and N lineages, we used Multiple Sequentially Markovian Coalescent (MSMC2) software (Schiffels and Wang 2020), which estimates effective population size (N_e) changes during time using Markovian approach while taking into account the connections between linked single-nucleotide polymorphisms (SNPs). Thus, it reconstructs demographic information not only from coalescence but considering recombination events as well. Demographic history analyses of both S and N lineages revealed considerable differences in population sizes during the same period of time (Fig. 5). *Polyplax* S showed gradual decline of N_e in the past, whereas *Polyplax* N suffered a more dramatic population collapse with subsequent increase. The lack of possibility to robustly phase *Polyplax* genome did not allow us to relate changes in N_e to exact historical events (Schiffels and Durbin 2014). Moreover, due to the absence of mutation rate estimate in closely related taxa, coalescence times were rescaled using the mutation rate of *Drosophila melanogaster*, which together with the unphased genome reduced the ability to provide absolute timing of demographic events. However, possible errors in timing do not affect interpretation of relative differences in N_e over time between the lineages. Life history traits (such as lifespan and fecundity) were proposed to represent the most important drivers of

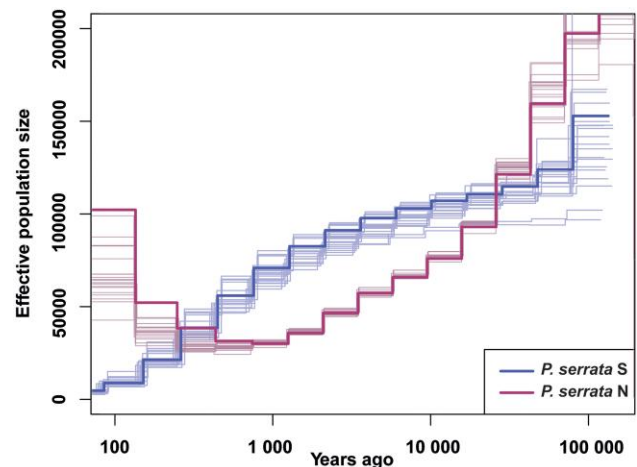


FIG. 5.—Changes of effective population sizes through time evaluated by MSMC2 for *P. serrata* S (bold blue) and *P. serrata* N_e (bold red) lineages. Bootstrap replicates (20 for each lineage) are plotted in lighter lines. Coalescence rates were rescaled using the mutation rate of *Drosophila* according to Wang et al. (2023) (3.3×10^{-9}), and generation time was estimated as 12 generations per year.

effective population size changes in animals (Romiguier et al. 2014). Given that we compared two closely related lineages with similar life history strategies, except for the difference in host specificities, we expect the shape of the N_e curves to be mostly influenced by varying demographic histories of their hosts. It was shown that the two host species reacted in a different way to the last glaciation period, retreating to different refugia (possibly more fragmented for *A. sylvaticus* than for *A. flavicollis*), and had different recolonization histories (Michaux et al. 2005). In accordance with that, after the glaciation-related decline, the *Polyplax* N lineage could have been able to restore its population size more quickly (when both its hosts recolonized central and northern parts of Europe) compared with the *Polyplax* S, which retained strict specificity to a single host. Correspondingly, our earlier microsatellite-based study (Martinů, et al. 2018) showed consistent population genetic diversity differences across several sympatric pairs of *Polyplax* N and S lineage populations, with the S lineage always possessing lower local diversity than the N lineage.

Methods

Tissue Samples and Isolation of Genetic Material

P. serrata lice were collected in the northwest of the Czech Republic (CZ) and the Šumava mountains region (CZ) from *A. flavicollis* hosts caught in wooden snap traps. Permission for field studies was provided by the Committee on the Ethics of Animal Experiments of the University of South Bohemia, by the Ministry of the Environment of the Czech Republic, and by the Ministry of the Agriculture of

the Czech Republic (Nos. MZP/2017/630/854, 43873/2019-MZE-18134, MZP/2021/630/2459). Lice were gathered from the mouse fur by brushing and stored in 100% ethanol at -20°C . Genomic DNA from individual louse specimens was extracted using the Qiagen QIAamp DNA Micro Kit (Qiagen). Before the next steps, lice were assigned to S and N lineages by sequencing a fragment of the mitochondrial cytochrome oxidase subunit I gene (COI, 379 bp) as in Martinů et al. (2018).

Oxford Nanopore and Illumina Genomic DNA and RNA Sequencing and Assembly

One specimen from each lineage (98c_Pro_SE for S lineage and HR10 for N lineage) with sufficient concentration measured on Qubit 2.0 Fluorometer (Invitrogen) were sequenced using Oxford Nanopore Technology (ONT) on one MinION II flowcell. Preparation of Oxford Nanopore 1D low input libraries and sequencing were performed at the Roy J. Carver Biotechnology Center (University of Illinois at Urbana-Champaign, USA). Altogether, 16.7 Gbp of data were produced for the S lineage and 10.3 Gbp for the N lineage (3.3 and 2.6 million reads of average size of 5 kbp). Data sets were basecalled with Fast-Bonito basecaller (Xu et al. 2021). Raw reads were trimmed with filtlong (version 0.2.0; <https://github.com/rrwick/Filtlong>) to obtain reads longer than 4,000 bp and with a phred score of 20 and higher. Genome assemblies were reconstructed with flye assembler (version 2.5; Kolmogorov et al. 2019) using ONT reads with genome size estimation parameter of 110 Mbp. Assemblies were subsequently corrected once with racon (version 1.3.3; <https://github.com/isovic/racon>) and twice with medaka (version 0.6.5; <https://github.com/nanoporetech/medaka>) using the ONT reads. The last polishing step was performed again with racon, this time using Illumina reads obtained from the same specimen as ONT reads. Illumina reads from S and N samples were sequenced on a NovaSeq lane as part of a different study, and details are described in Martinů et al. (2020). Completeness of the assemblies was checked with BUSCO using Arthropoda gene set (version 3; Waterhouse et al. 2018).

For RNA sequencing of the S lineage, 34 lice of different life stages were gathered from 1 specimen of the field mouse *A. flavicollis* caught in the proximity of a game preserve Flaje (CZ). Lice were preserved in RNAlater (Thermo Fischer Scientific) and isolated with phenol–chloroform extraction in the laboratory. Preparation of the cDNA library and sequencing of the 150-bp-long PE reads on Illumina NovaSeq 6000 sequencer were performed by Novogene (United Kingdom). Adapters and low-quality reads were removed using Trimmomatic v0.39 (Bolger et al. 2014). Obtained reads were then checked for quality with FastQC (Andrews 2010) and assembled with Trinity

v2.15.1 with default settings (Grabherr et al. 2011). RNA data for N lineage were not obtained due to the lack of sufficient amount of input material.

Since the assemblies combining Illumina and Oxford nanopore reads contained incomplete genomes of the symbiont *L. polyplacis*, fragmented into several contigs, we used (based on our previous experience) SPAdes 3.10 (Bankevich et al. 2012) to assemble complete symbionts' genomes from the Illumina short reads. The reads were trimmed by Trimmomatic (Bolger et al. 2014) with default parameters. SPAdes was run with the option `--meta`. In the resulting assemblies, the contigs representing complete *L. polyplacis* genomes were identified based on their lengths and open reading frames (ORFs) distribution (i.e. densely arranged ORFs as typical for prokaryotes) and were annotated in Prokka (Seemann 2014).

Gene Prediction, Annotation, and Functional Comparative Analysis

To compare *P. serrata* genomes with the other three available phthirapterans (*P. humanus*, *B. nebulosa*, and *C. columbae*), we reannotated and reanalyzed all genomes, rather than using the published data for the comparison. This ensures a more consistent approach. The results we obtained from these reannotations differ in some aspects from those published previously, particularly in the numbers of genes identified in various categories and families. Considering the complexity of eukaryotic genomes, the level of uncertainty during the annotation process, and the differences in annotation approaches/programs, such inter-study differences are to be expected. However, they also provide warning that the results of the annotation step and identification of gene functions must be taken with caution (Salzberg 2019; Scalzitti et al. 2020).

To perform genomic comparison of *P. serrata* with other Phthiraptera, we included into our analyses the three previously published genomes, i.e. *P. humanus*, *B. nebulosa*, and *C. columbae*. While there are three other phthirapteran genomes deposited in the NCBI GenBank, we did not include them due to their lower quality. To compare several specific genes/functions of the studied lice with other blood-feeding insects, we also included genomes of *A. aegypti*, *C. lectularius*, *G. morsitans*, and *R. prolixus* (see [supplementary table S1](#), [Datasheet S8](#), [Supplementary Material](#) online, for the NCBI GenBank accession numbers and transcriptome references of the genomes). As a preparatory step for the gene prediction, we identified repeat contents in our de novo assemblies of *P. serrata* utilizing RepeatModeler v2.0.3 (Flynn et al. 2020), which was followed by soft masking complex repeats using RepeatMasker v4.1.2-p1 (Tarailo-Graovac and Chen 2009). To maintain a methodologically consistent approach to the downstream analysis, we applied the same prediction and annotation

process to all of the included genomes. This procedure ensured methodological uniformity and minimized variability that might arise from using different tools, databases, or varying versions of the databases in the previous studies. Funannotate v1.18.14 (<https://github.com/nextgenusfs/funannotate>) was employed to perform gene prediction and functional annotation on the analyzed genomes. Briefly, ab initio gene prediction was performed by “funannotate predict” command that employed Augustus v3.5.0 (Hoff and Stanke 2019), glimmerHMM v3.0.4 (Majoros et al. 2004), snap v2013_11_29 (Kolesov et al. 2001), and GeneMark v4.71 (Lukashin and Borodovsky 1998) gene predictors. Transcript assemblies were used as transcript evidence to enhance gene prediction (with exception of *B. nebulosa* and *G. morsitans* for which no transcriptomic evidence is available in public databases). The derived gene models underwent annotation via the “funannotate annotate” command, which invokes InterProScan v5.60-92.0 (Zdobnov and Apweiler 2001), eggNOG-mapper v2.1.10 (Cantalapiedra et al. 2021), SignalP v5.0b (Almagro Armenteros et al. 2019), and Phobius v1.0.1 (Kall et al. 2004) tools for gene annotation. Ribosomal RNAs were predicted using Barrnap v0.9 (<https://github.com/tseemann/barrnap>). Comparative analysis was performed using the “funannotate compare” function of Funannotate v1.8.14 (<https://github.com/nextgenusfs/funannotate>).

Comparative Functional Genomic Analysis of Anoplura and Chewing Lice

To elucidate shared and unique protein families and domains across the compared Anoplura (*P. serrata* S and N lineage and *P. humanus*) and chewing lice (*C. columbae* and *B. nebulosa*) genomes, Venn diagrams were generated based on functional comparison outputs (supplementary table S1, Datasheets S3 to S5 and S7, Supplementary Material online) from different protein annotation databases. This included Pfam and InterPro (Paysan-Lafosse et al. 2023) databases for annotated protein families, domains, and conserved sites, the carbohydrate-active enzyme (CAZy) database for classifications of carbohydrate-active enzymes (Drula et al. 2022), and MEROPS database for peptidases and peptidase inhibitors. In addition, a PCoA was conducted on comparison output from the Pfam and InterPro databases (supplementary table S1, Datasheets S3 to S4, Supplementary Material online), which utilized Bray–Curtis dissimilarity metric in vegan package v2.6-4 (Oksanen 2022) within the R environment (R Core Team 2013). Genome-wide analysis of COGs in compared Anoplura and chewing lice species was performed and visualized using OrthoVenn3 (Sun et al. 2023). CAFE v5 (Mendes et al. 2020) tool was utilized to infer patterns of gene family expansion and contraction among

compared species and their parent nodes. Changes in gene family size and their associated statistical significance were visualized using CafePlotter v0.2.0 (<https://github.com/moshi4/CafePlotter>).

Metabolic Reconstructions of Host–Symbiont Complementarity

To evaluate and compare possible complementarity of the louse hosts and their obligate symbionts in production of amino acids and B vitamins (the compounds typically considered in the insect–bacteria symbiosis), we analyzed metabolic capacities using the KEGG database (Kanehisa, Sato, Kawashima, et al. 2016). For the genomes of both *P. serrata* lineages and their *L. polyplacis* symbionts, we assigned K numbers (KEGG orthology identifiers) to all annotated proteins in their genomes by the web-based program BlastKoala (Kanehisa, Sato and Morishima 2016) and mapped these numbers on the biosynthetic pathways using the KEGG mapper tool. For *P. humanus* and its symbiont *R. pediculicola*, we used the pathway maps already available in the KEGG database.

Phylogeny and Dating of the *P. serrata* Lineages

To reconstruct phylogenetic position of the 2 *P. serrata* lineages within Anoplura, we build a matrix of 15 Anoplura species and 2 sequences of *Haematomyzus elephantis* as outgroups. Using a locally generated pipeline (refer to the Data availability section), we extracted orthologs from the alignment published by de Moya et al. (2021) in their phylogenomic analysis of Psocodea. (downloaded from <https://datadryad.org/stash/dataset/doi:10.5061/dryad.c59zw3r50>). Since we focused on Anoplura, we selected for the following analysis only the 13 anopluran species and the outgroup. These data were extended with the genes extracted from the *P. serrata* genomes. To obtain sets of single-copy orthologs, we translated all sequences into amino acids using the EMBOSS v6.6.0.0 (Li et al. 2015) transeq function and searched the orthologs by OrthoFinder v2.5.5 (Emms and Kelly 2015). For the total of the 1,049 identified single-copy orthologs, we made alignment of their nucleotide forms using MAFFT v7.520 (Katoh et al. 2002) implemented in Geneious (Kearse et al. 2012). The alignments were concatenated, and a matrix was built from all second codon positions, resulting in a matrix of 536,051 positions. The tree was inferred by ML using IQ-TREE 2 v. 2.2.0 (Minh et al. 2020), with 1,000 ultrafast bootstrap replicates. Optimal combination of the 1,049 partitions, and selection of the model for each partition set (based on BIC), was performed by the program (supplementary table S3, Supplementary Material online). The resulting topology was further used as a constraint for the dating analysis in the MCMCtree program (Puttick 2019). Calibration for two nodes within Anoplura was adopted from the de

Moya et al. (2021) analysis, namely *Pedicinus* + (*Phthirus* + *Pediculus*) (20 to 25 Mya) and *P. schaeffi* + *P. humanus* (5 to 7 Mya). To derive the estimates for the other nodes, we ran the MCMCtree analysis with approx. likelihood calculation and the mcmc process set to 500,000 samples, burnin 50,000, and sample frequency 10. Convergency of the mcmc chains was checked by comparing convergence between the dating sets from both chains, as recommended in the MCMCtree manual. To measure the divergence in the DNA barcoding gene *cox1*, we retrieved the *cox1* sequences from our assemblies for both *P. serrata* samples and obtained the divergencies from the alignment tool MAFFT implemented in Geneious. For comparison, we also obtained the *cox1* divergence between *P. humanus* (accession KC685844) and *P. schaeffi* (AY695999).

Synteny Analysis

For comparative analysis of *Polyplax* S and N lineages, contigs longer than 0.7 Mbp were chosen. The level of synteny of the 18 longest scaffolds of *Polyplax* S and 21 of *Polyplax* N was analyzed using McScanX (Wang et al. 2012) method considering orthologous genes as anchors. Collinearity for the ortholog synteny blocks on contigs was evaluated using default settings. SynVisio program (Bandi and Gutwin 2020) was used for detailed visualization of the McScanX outputs in regions where structural rearrangements occurred.

Population Demography

To compare population demography of the two *P. serrata* lineages, we prepared genomic DNA libraries for individual louse samples with insert size of 450 bp. The libraries were sequenced by paired-end Illumina process on NovaSeq 6000, yielding ~59.5 million paired-end reads per sample. Preparation of libraries and sequencing were provided by the W. M. Keck Center (University of Illinois, Urbana, IL, USA). For the following analysis, we selected eight samples for each lineage (the S lineage data were also part of the previously published study [Martinů et al. 2020]). The demographic history of both lineages was evaluated from the whole genome sequences by MSMC2 software (Schiffels et al. 2020). The data were adaptor and size filtered with bbtools (<https://jgi.doe.gov/data-andtools/bbtools/>), and reads were mapped against *Polyplax* S genome using bowtie2 (Langmead and Salzberg 2012). Duplicated reads were removed with PICARD (<http://broadinstitute.github.io/picard/>), and SNP calling was performed using GATK Genome Analysis Toolkit following the “Best Practices” guide from the Broad Institute (Van der Auwera et al. 2013). Data sets of *Polyplax* S and N were separately filtered for quality with GATK, and then, minor allele frequencies (MAF) equaled to 0.05 were

removed in PLINK 1.9 (<https://www.cog-genomics.org/plink/1.9/>). Data were then converted according to MSMC2 instructions using SAMtools v.1.8 (Li et al. 2009), BCFtools v.1.8 (Danecek et al. 2021), and scripts available at (msmc-tools/msmc-tutorial/guide.md at master · stschiif/msmc-tools · GitHub). MSMC2 analyses assessed coalescence rates between haplotypes within the S and N lineages as well as 20 bootstrap replicates based on default values except for time segment patterning parameter (-p). To avoid overfitting, the default 32 time segments were lowered to 18 (-p 1*2+15*1+1*2), due to the small size of the *Polyplax* genome (139 Mbp). Results were plotted in R Studio, where they were scaled based on the mutation rate of *Drosophila* (3.3×10^{-9}) (Wang et al. 2023) and generation time of 12 generations per year.

Supplementary Material

Supplementary material is available at *Genome Biology and Evolution* online.

Acknowledgments

We thank our colleagues Masoud Nazarzadeh, Jakub Vlček, and Milena Nováková for their advice on bioinformatic procedures. Access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum provided under the program “Projects of Large Research, Development, and Innovations Infrastructures” (CESNET LM2015042) is greatly appreciated.

Funding

This work was supported by the Grant Agency of the Czech Republic (grant number 21-02532S to V.H.).

Data Availability

The raw genomic data, the annotated genomes of *P. serrata* S and N lineages, the resequenced samples for the demography reconstruction, and the genomes of *L. polyplax* were deposited on GenBank database under BioProject PRJNA1018720, with GenBank accession numbers SRR27586360 to SRR27586363, JAWJWF000000000, JAWJWE000000000, and CP135136 and CP135137, respectively. In addition, rRNA-seq raw reads of *P. serrata* were deposited on Sequence Read Archive (SRA) database with accession number SRR27590290 under BioProject PRJNA1018720. The *cox1* sequences for the *P. serrata* lineages were deposited in the same bioproject under the accession numbers PP112155 and PP112156. Data sets of *P. serrata* S and N lineages and other compared genomes are available at Zenodo (<https://zenodo.org/records/10523744>) under doi:10.5281/zenodo.10523744. This

includes fasta format files of genomes, transcriptome, and proteome and annotation tables obtained from gene prediction and annotation workflow. In addition, Zenodo-deposited data sets include gbk format files for all compared genomes and identified repeat families in fasta format files for *P. serrata* S and N lineages. Helper python, bash, and R scripts employed in this study are available at <https://github.com/hassantarabai/MS-Pserrata-S-N-2023>.

Literature Cited

- Akman L, Yamashita A, Watanabe H, Oshima K, Shiba T, Hattori M, Aksoy S. Genome sequence of the endocellular obligate symbiont of tsetse flies, *Wigglesworthia glossinidia*. *Nat Genet.* 2002;32(3):402–407. <https://doi.org/10.1038/ng986>.
- Allen JM, Reed DL, Perotti MA, Braig HR. Evolutionary relationships of “*Candidatus* Riesia spp.,” endosymbiotic enterobacteriaceae living within hematophagous primate lice. *Appl Environ Microbiol.* 2007;73(5):1659–1664. <https://doi.org/10.1128/AEM.01877-06>.
- Almagro Armenteros JJ, Tsirigos KD, Sonderby CK, Petersen TN, Winther O, Brunak S, von Heijne G, Nielsen H. SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat Biotechnol.* 2019;37(4):420–423. <https://doi.org/10.1038/s41587-019-0036-z>.
- Ammar E, Tsai C, Whitfield A, Redinbaugh M, Hogenhout S. Cellular and molecular aspects of rhabdovirus interactions with insect and plant hosts. *Annu Rev Entomol.* 2009;54(1):447–468. <https://doi.org/10.1146/annurev.ento.54.110807.090454>.
- Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010. <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>
- Baldwin-Brown J, Villa S, Vickrey A, Johnson K, Bush S, Clayton D, Shapiro M. The assembled and annotated genome of the pigeon louse *Columbicola columbae*, a model ectoparasite. *G3.* 2021;11(2):jkab009. <https://doi.org/10.1093/g3journal/jkab009>.
- Bandi V, Gutwin C. Proceedings of the 46th Graphics Interface Conference on Proceedings of Graphics Interface. Waterloo (ON): Canadian Human-Computer Communications Society; 2020. p. 74–83.
- Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Pribelski AD, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol.* 2012;19(5):455–477. <https://doi.org/10.1089/cmb.2012.0021>.
- Bolger A, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics.* 2014;30(15):2114–2120. <https://doi.org/10.1093/bioinformatics/btu170>.
- Boyd BM, Allen JM, de Crécy-Lagard V, Reed DL. Genome sequence of *Candidatus* Riesia pediculischaeffi, endosymbiont of chimpanzee lice, and genomic comparison of recently acquired endosymbionts from human and chimpanzee lice. *G3.* 2014;4(11):2189–2195. <https://doi.org/10.1534/g3.114.012567>.
- Cantalapiedra C, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol Biol Evol.* 2021;38(12):5825–5829. <https://doi.org/10.1093/molbev/msab293>.
- Danecek P, Bonfield J, Liddle J, Marshall J, Ohan V, Pollard M, Whitwham A, Keane T, McCarthy S, Davies R, et al. Twelve years of SAMtools and BCFtools. *GigaScience* 2021;10(2):giab008. <https://doi.org/10.1093/gigascience/giab008>.
- de Moya R, Yoshizawa K, Walden K, Sweet A, Dietrich C, Johnson KP. Phylogenomics of parasitic and nonparasitic lice (Insecta: Psocodea): combining sequence data and exploring compositional bias solutions in next generation data sets. *Syst Biol.* 2021;70(4):719–738. <https://doi.org/10.1093/sysbio/syaa075>.
- Ding Q, Li R, Ren X, Chan L, Ho V, Xie D, Ye P, Zhao Z. Genomic architecture of 5S rDNA cluster and its variations within and between species. *BMC Genomics.* 2022;23(1):238. <https://doi.org/10.1186/s12864-022-08476-x>.
- Drula E, Garron ML, Dogan S, Lombard V, Henrissat B, Terrapon N. The carbohydrate-active enzyme database: functions and literature. *Nucleic Acids Res.* 2022;50(D1):D571–D577. <https://doi.org/10.1093/nar/gkab1045>.
- Emms D, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* 2015;16(1):157. <https://doi.org/10.1186/s13059-015-0721-2>.
- Feulner P, De-Kayne R. Genome evolution, structural rearrangements and speciation. *J Evol Biol.* 2017;30(8):1488–1490. <https://doi.org/10.1111/jeb.13101>.
- Flynn J, Hubley R, Goubert C, Rosen J, Clark A, Feschotte C, Smit A. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci USA.* 2020;117(17):9451–9457. <https://doi.org/10.1073/pnas.1921046117>.
- Golub N, Nekkala S. Brief report—Chromosome numbers of two sucking louse species (Insecta, Phthiraptera, Anoplura). *Hereditas* 2004;141(1):94–96. <https://doi.org/10.1111/j.1601-5223.2004.01859.x>.
- Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol.* 2011;29(7):644–652. <https://doi.org/10.1038/nbt.1883>.
- Hoff KJ, Stanke M. Predicting genes in single genomes with AUGUSTUS. *Curr Protoc Bioinformatics.* 2019;65(1):e57. <https://doi.org/10.1002/cpbi.57>.
- Kall L, Krogh A, Sonnhammer EL. A combined transmembrane topology and signal peptide prediction method. *J Mol Biol.* 2004;338(5):1027–1036. <https://doi.org/10.1016/j.jmb.2004.03.016>.
- Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* 2016;44(D1):D457–D462. <https://doi.org/10.1093/nar/gkv1070>.
- Kanehisa M, Sato Y, Morishima K. BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *J Mol Biol.* 2016;428(4):726–731. <https://doi.org/10.1016/j.jmb.2015.11.006>.
- Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 2002;30(14):3059–3066. <https://doi.org/10.1093/nar/gkf436>.
- Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, et al. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics.* 2012;28(12):1647–1649. <https://doi.org/10.1093/bioinformatics/bts199>.
- Kinjo Y, Bourguignon T, Hongoh Y, Lo N, Tokuda G, Ohkuma M. Coevolution of metabolic pathways in Blattodea and their Blattabacterium endosymbionts, and comparisons with other insect-bacteria symbioses. *Microbiology Spectrum.* 2022;10:e02779-22.
- Kirkness EF, Haas BJ, Sun WL, Braig HR, Perotti MA, Clark JM, Lee SH, Robertson HM, Kennedy RC, Elhaik E, et al. Genome sequences of

- the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle. *Proc Natl Acad Sci USA*. 2010;107(27):12168–12173. <https://doi.org/10.1073/pnas.1003379107>.
- Kolesov G, Mewes HW, Frishman D. SNAPping up functionally related genes based on context information: a colinearity-free approach. *J Mol Biol*. 2001;311(4):639–656. <https://doi.org/10.1006/jmbi.2001.4701>.
- Kolmogorov M, Yuan J, Lin Y, Pevzner P. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol*. 2019;37(5):540–546. <https://doi.org/10.1038/s41587-019-0072-8>.
- Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012;9(4):357–359. <https://doi.org/10.1038/nmeth.1923>.
- Li W, Cowley A, Uludag M, Gur T, McWilliam H, Squizzato S, Park Y, Buso N, Lopez R. The EMBL-EBI bioinformatics web and programmatic tools framework. *Nucleic Acids Res*. 2015;43(W1):W580–W584. <https://doi.org/10.1093/nar/gkv279>.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>.
- Light J, Smith V, Allen J, Durden L, Reed D. Evolutionary history of mammalian sucking lice (Phthiraptera: Anoplura). *BMC Evol Biol*. 2010;10(1):292. <https://doi.org/10.1186/1471-2148-10-292>.
- Longdon B, Day J, Schulz N, Leftwich P, de Jong M, Breuker C, Gibbs M, Obbard D, Wilfert L, Smith S, et al. Vertically transmitted rhabdoviruses are found across three insect families and have dynamic interactions with their hosts. *Proc R Soc B Biol Sci*. 2017;284(1847):20162381. <https://doi.org/10.1098/rspb.2016.2381>.
- Lukashin AV, Borodovsky M. GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res*. 1998;26(4):1107–1115. <https://doi.org/10.1093/nar/26.4.1107>.
- Majoros WH, Pertea M, Salzberg SL. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics*. 2004;20(16):2878–2879. <https://doi.org/10.1093/bioinformatics/bth315>.
- Martinů J, Hypša V, Štefka J. Host specificity driving genetic structure and diversity in ectoparasite populations: coevolutionary patterns in *Apodemus* mice and their lice. *Ecol Evol*. 2018;8(20):10008–10022. <https://doi.org/10.1002/ece3.4424>.
- Martinů J, Roubová V, Nováková M, Smith VS, Hypša V, Štefka J. Characterisation of microsatellite loci in two species of lice, *Polyplax serrata* (Phthiraptera: Anoplura: Polyplacidae) and *Myrsidea nesomimi* (Phthiraptera: Amblycera: Menoponidae). *Folia Parasitol*. 2015;62:16. <https://doi.org/10.14411/fp.2015.016>.
- Martinů J, Štefka J, Poosakkannu A, Hypša V. “Parasite turnover zone” at secondary contact: a new pattern in host-parasite population genetics. *Mol Ecol*. 2020;29(23):4653–4664. <https://doi.org/10.1111/mec.15653>.
- Mendes F, Vanderpool D, Fulton B, Hahn M. CAFE 5 models variation in evolutionary rates among gene families. *Bioinformatics*. 2020;36(22-23):5516–5518. <https://doi.org/10.1093/bioinformatics/btaa1022>.
- Michaux J, Libois R, Filippucci M. So close and so different: comparative phylogeography of two small mammal species, the yellow-necked fieldmouse (*Apodemus flavicollis*) and the woodmouse (*Apodemus sylvaticus*) in the Western Palearctic region. *Heredity* (Edinb). 2005;94(1):52–63. <https://doi.org/10.1038/sj.hdy.6800561>.
- Michaux J, Pasquier L. Dynamique des populations de mulots (Rodentia, *Apodemus*) en Europe durant le Quaternaire; premières données. *Bull Soc Géol France*. 1974;S7-XVI(4):431–439. <https://doi.org/10.2113/gssgfbull.S7-XVI.4.431>.
- Minh B, Schmidt H, Chernomor O, Schrempf D, Woodhams M, von Haeseler A, Lanfear R. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol Biol Evol*. 2020;37(8):2461–2461. <https://doi.org/10.1093/molbev/msaa131>.
- Oksanen J. vegan: an R package for community ecologists. 2022. <https://github.com/vegandevs/vegan>
- Paysan-Lafosse T, Blum M, Chuguransky S, Grego T, Pinto B, Salazar G, Bileschi M, Bork P, Bridge A, Colwell L, et al. InterPro in 2022. *Nucleic Acids Res*. 2023;51(D1):D418–D427. <https://doi.org/10.1093/nar/gkac993>.
- Puttick M. MCMCtreeR: functions to prepare MCMCtree analyses and visualize posterior ages on trees. *Bioinformatics*. 2019;35(24):5321–5322. <https://doi.org/10.1093/bioinformatics/btz554>.
- R Core Team R. R: a language and environment for statistical computing. 2013. <https://www.r-project.org/>
- Reed D, Light J, Allen J, Kirchman J. Pair of lice lost or parasites regained: the evolutionary history of anthropoid primate lice. *BMC Biol*. 2007;5(1):7. <https://doi.org/10.1186/1741-7007-5-7>.
- Řihová JM, Gupta S, Darby AC, Nováková E, Hypša V. *Arsenophonus* symbiosis with louse flies: multiple origins, coevolutionary dynamics, and metabolic significance. *mSystems*. 8(5):e0070623. <https://doi.org/10.1128/msystems.00706-23>.
- Řihová J, Nováková E, Husník F, Hypša V. *Legionella* becoming a mutualist: adaptive processes shaping the genome of symbiont in the louse *Polyplax serrata*. *Genome Biol Evol*. 2017;9(11):2946–2957. <https://doi.org/10.1093/gbe/evx217>.
- Rio RVM, Symula RE, Wang JW, Lohs C, Wu YN, Snyder AK, Bjornson RD, Oshima K, Biehl BS, Perna NT, et al. Insight into the transmission biology and species-specific functional capabilities of tsetse (Diptera: Glossinidae) obligate symbiont *Wigglesworthia*. *mBio*. 2012;3(1):e00240-11. <https://doi.org/10.1128/mBio.00240-11>.
- Romiguier J, Lourenco J, Gayral P, Faivre N, Weinert L, Ravel S, Ballenghien M, Cahais V, Bernard A, Loire E, et al. Population genomics of eusocial insects: the costs of a vertebrate-like effective population size. *J Evol Biol*. 2014;27(3):593–603. <https://doi.org/10.1111/jeb.12331>.
- Salzberg S. Next-generation genome annotation: we still struggle to get it right. *Genome Biol*. 2019;20(1):92. <https://doi.org/10.1186/s13059-019-1715-2>.
- Scalzitti N, Jeannin-Girardon A, Collet P, Poch O, Thompson J. A benchmark study of ab initio gene prediction methods in diverse eukaryotic organisms. *BMC Genomics*. 2020;21(1):293. <https://doi.org/10.1186/s12864-020-6707-9>.
- Schiffels S, Durbin R. Inferring human population size and separation history from multiple genome sequences. *Nat Genet*. 2014;46:919–925. <https://doi.org/10.1038/ng.3015>.
- Schiffels S, Wang K. MSMC and MSMC2: the Multiple Sequentially Markovian Coalescent. In: Duthel JY, editor. *Statistical population genomics. Methods in Molecular Biology*. New York (NY): Humana; 2020. p. 147–1662090.
- Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*. 2014;30(14):2068–2069. <https://doi.org/10.1093/bioinformatics/btu153>.
- Shin C, Allmon W. How we study cryptic species and their biological implications: a case study from marine shelled gastropods. *Ecol Evol*. 2023;13(9):e10360. <https://doi.org/10.1002/ece3.10360>.
- Simakov O, Marlétaz F, Yue J, O’Connell B, Jenkins J, Brandt A, Calef R, Tung C, Huang T, Schmutz J, et al. Deeply conserved synteny resolves early events in vertebrate evolution. *Nat Ecol Evol*. 2020;4(6):820–830. <https://doi.org/10.1038/s41559-020-1156-z>.
- Štefka J, Hypša V. Host specificity and genealogy of the louse *Polyplax serrata* on field mice, *Apodemus* species: a case of parasite duplication or colonisation? *Int J Parasitol*. 2008;38(6):731–741. <https://doi.org/10.1016/j.ijpara.2007.09.011>.
- Sun J, Lu F, Luo Y, Bie L, Xu L, Wang Y. OrthoVenn3: an integrated platform for exploring and visualizing orthologous data across

- genomes. *Nucleic Acids Res.* 2023;51(W1):W397–W403. <https://doi.org/10.1093/nar/gkad313>.
- Sweet A, Browne D, Hernandez A, Johnson K, Cameron S. Draft genome assemblies of the avian louse *Brueelia nebulosa* and its associates using long-read sequencing from an individual specimen. *G3.* 2023;13(4):jkad030. <https://doi.org/10.1093/g3journal/jkad030>.
- Tarailo-Graovac M, Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinform* 2009: Chapter 4:4.10.1–4.10.14. <https://doi.org/10.1002/0471250953.bi0410s25>.
- Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Current Protoc Bioinform.* 2013;43(1):11.10.1–11.10.33. <https://doi.org/10.1002/0471250953.bi1110s43>.
- Wang Y, McNeil P, Abdulazeez R, Pascual M, Johnston S, Keightley P, Obbard D. Variation in mutation, recombination, and transposition rates in *Drosophila melanogaster* and *Drosophila simulans*. *Genome Res.* 2023;33(4):587–598. <https://doi.org/10.1101/gr.277383.122>.
- Wang Y, Tang H, DeBarry J, Tan X, Li J, Wang X, Lee T, Jin H, Marler B, Guo H, et al. *MCScanX*: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* 2012;40(7):e49. <https://doi.org/10.1093/nar/gkr1293>.
- Waterhouse R, Seppey M, Simao F, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva E, Zdobnov E. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol Biol Evol.* 2018;35(3):543–548. <https://doi.org/10.1093/molbev/msx319>.
- Xu Z, Mai Y, Liu D, He W, Lin X, Xu C, Zhang L, Meng X, Mafofo J, Zaher WA, et al. Fast-bonito: a faster deep learning based basecaller for nanopore sequencing. *Artif Intell Life Sci.* 2021;1:100011. <https://doi.org/10.1016/j.ails.2021.100011>.
- Zdobnov EM, Apweiler R. InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics.* 2001;17(9):847–848. <https://doi.org/10.1093/bioinformatics/17.9.847>.
- Associate editor:** Toni Gossmann