Delineation of the Genera Haemoproteus and Plasmodium Using RNA-Seq and Multi-gene Phylogenetics

Jasper Toscani Field, Josh Weinberg, Staffan Bensch, Nubia E. Matta, Gediminas Valkiūnas & Ravinder N. M. Sehgal
Your article is protected by copyright and all rights are held exclusively by Springer Science+Business Media, LLC, part of Springer Nature. This e-offprint is for personal use only and shall not be self-archived in electronic repositories. If you wish to self-archive your article, please use the accepted manuscript version for posting on your own website. You may further deposit the accepted manuscript version in any repository, provided it is only made publicly available 12 months after official publication or later and provided acknowledgement is given to the original source of publication and a link is inserted to the published article on Springer’s website. The link must be accompanied by the following text: “The final publication is available at link.springer.com”.
Delineation of the Genera *Haemoproteus* and *Plasmodium* Using RNA-Seq and Multi-gene Phylogenetics

Jasper Toscani Field¹ · Josh Weinberg¹ · Staffan Bensch² · Nubia E. Matta³ · Gediminas Valkiūnas⁴ · Ravinder N. M. Sehgal¹

Received: 5 April 2018 / Accepted: 1 November 2018
© Springer Science+Business Media, LLC, part of Springer Nature 2018

Abstract

Members of the order Haemosporida are protist parasites that infect mammals, reptiles and birds. This group includes the causal agents of malaria, *Plasmodium* parasites, the genera *Leucocytozoon* and *Fallisia*, as well as the species rich genus *Haemoproteus* with its two subgenera *Haemoproteus* and *Parahaemoproteus*. Some species of *Haemoproteus* cause severe disease in avian hosts, and these parasites display high levels of diversity worldwide. This diversity emphasizes the need for accurate evolutionary information. Most molecular studies of wildlife haemosporidians use a bar coding approach by sequencing a fragment of the mitochondrial cytochrome *b* gene. This method is efficient at differentiating parasite lineages but insufficient for accurate phylogenetic inferences in highly diverse taxa such as haemosporidians. Recent studies have utilized multiple mitochondrial genes (*cyt b*, *cox1* and *cox3*), sometimes combined with a few apicoplast and nuclear genes. These studies have been highly successful with one notable exception: the evolutionary relationships of the genus *Haemoproteus* remain unresolved. Here we describe the transcriptome of *Haemoproteus columbae* and investigate its phylogenetic position recovered from a multi-gene dataset (600 genes). This genomic approach restricts the taxon sampling to 18 species of apicomplexan parasites. We employed Bayesian inference and maximum likelihood methods of phylogenetic analyses and found *H. columbae* and a representative from the subgenus *Parahaemoproteus* to be sister taxa. This result strengthens the hypothesis of genus *Haemoproteus* being monophyletic; however, resolving this question will require sequences of orthologs from, in particular, representatives of *Leucocytozoon* species.

Keywords *Haemoproteus* · Phylogenetics · Transcriptome · Avian parasitology


Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00239-018-9875-3) contains supplementary material, which is available to authorized users.

Introduction

Haemosporidian parasites are notable for their diversity and cosmopolitan distribution. These globally distributed parasites are members of the phylum *Apicomplexa* (Garnham 1966; Valkiūnas 2005). Host species include mammals, lizards and birds (Ricklefs and Fallon 2002; Valkiūnas 2005), and several haemosporidian species have been reported in amphibians and fish (Valkiūnas 2005). Genera infecting avian hosts include *Plasmodium*, *Fallisia*, *Leucocytozoon* and *Haemoproteus*. The genus *Haemoproteus* currently contains two subgenera: *Haemoproteus* and *Parahaemoproteus*. Members of the subgenus *Haemoproteus* are vectored by hippoboscid flies (Hippoboscidae) and primarily infect Columbiformes such as pigeons and doves but have also been found in marine birds belonging to Pelecaniformes and Charadriiformes (Levin et al. 2012). *Parahaemoproteus* parasites are spread by biting midges (Ceratopogonidae) and
infect many bird species, in particular among passerines, raptors and waterfowl. Members of both subgenera produce hemozoin and undergo merogony in cells of fixed endothelial tissues with erythrocyte invasion producing only gametocytes (Valkiūnas 2005; Levin et al. 2012; Palinauskas et al. 2013). The avian haemosporidian sequence database MalAvi (Version 3.2.2, April 4th, 2018) shows 1148 cytochrome b lineages of *Haemoproteus*. Of these lineages, 97.38% belong to the subgenus *Parahaemoproteus* and 30 belong to subgenus *Haemoproteus*. Of these 30 lineages, five have been morphologically described as belonging to *Haemoproteus columbae*. *H. columbae* can cause obstructions in pulmonary, myocardial and hepatic tissue of its hosts as seen in an examination of a deceased Bleeding Heart Dove (*Galliformes* (*columbae*) (Earlé et al. 1993). Marked enlargement of spleen and liver has been reported during heavy *H. columbae* parasitemia, which might exceed 50% of parasitized red blood cells and cause anemia in rock pigeons (Garnham 1966; Valkiūnas 2005). This parasite is globally distributed and has become one of the most well studied *Haemoproteus* species (Valkiūnas 2005; Earlé et al. 1993; Adrianov and Cordeiro 2001; Waldenström et al. 2002; Santiago-Alarcón et al. 2010; Waite et al. 2012, 2014).

Focusing on parasite diversity, there are hundreds of morphologically defined species within the genera *Haemoproteus*, *Leucocytozoon* and *Plasmodium* and they are further diversified with multiple genetically unique lineages (Ricklefs and Fallon 2002; Valkiūnas 2005; Martinsen et al. 2008; Santiago-Alarcón et al. 2010; Borner et al. 2016). The effects of each species on its host can be extremely variable, with some species causing severe anemia and death while others persist in barely detectable chronic infections (Palinauskas et al. 2008; Dimitrov et al. 2015; Palinauskas et al. 2015). While valuable, this kind of analysis lacks the depth necessary to answer important questions regarding parasite relationships (Hellgren et al. 2013; Outlaw and Ricklefs 2014; Borner et al. 2016; Bensch et al. 2016). Unresolved phylogenetic relationships can be addressed with multi-gene datasets (Borner et al. 2016). The primary limitation of such methods is that genomic and transcriptomic datasets, required for phylogenomic analyses, are particularly difficult to obtain from haemosporidian parasites (Lauron et al. 2014; Bensch et al. 2016; Videvall et al. 2017; Böehme et al. 2018; Videvall 2018). Recent multi-gene phylogenetic analyses of the avian apicomplexans performed with multiple statistical methods produced well-resolved phylogenies with one exception in the case of haemosporidian parasites, i.e., the placement of subgenus *Haemoproteus* in relation to subgenus *Parahaemoproteus* and genus *Plasmodium* (Martinsen et al. 2008; Bensch et al. 2016; Borner et al. 2016; Pacheco et al. 2018). An early multi-gene study by Martinsen et al. (2008) showed the genus *Haemoproteus* as paraphyletic with *Plasmodium* forming a sister relationship with subgenus *Parahaemoproteus*. In work by Borner et al. (2016), depending on the evolutionary model, taxa and genes used for analyses, the subgenus *Haemoproteus*, represented by *H. columbae*, formed a sister relationship with either subgenus *Parahaemoproteus* or with the genus *Plasmodium*. Furthermore, the relationships of the *Haemoproteus* subgenera have been open to debate, with one subgenus forming a sister relationship with *Plasmodium* as in Fig. 1a (Martinsen et al. 2008; Santiago-Alarcón et al. 2010; Martinez-de la Puente et al. 2011; Carlson et al. 2013; Jasper et al. 2014; Outlaw and Ricklefs 2014; Palinauskas et al. 2015). While valuable, this kind of analysis lacks the depth necessary to answer important questions regarding parasite relationships (Hellgren et al. 2013; Outlaw and Ricklefs 2014; Borner et al. 2016; Bensch et al. 2016). Unresolved phylogenetic relationships can be addressed with multi-gene datasets (Borner et al. 2016). The primary limitation of such methods is that genomic and transcriptomic datasets, required for phylogenomic analyses, are particularly difficult to obtain from haemosporidian parasites (Lauron et al. 2014; Bensch et al. 2016; Videvall et al. 2017; Böehme et al. 2018; Videvall 2018). Recent multi-gene phylogenetic analyses of the avian apicomplexans performed with multiple statistical methods produced well-resolved phylogenies with one exception in the case of haemosporidian parasites, i.e., the placement of subgenus *Haemoproteus* in relation to subgenus *Parahaemoproteus* and genus *Plasmodium* (Martinsen et al. 2008; Bensch et al. 2016; Borner et al. 2016; Pacheco et al. 2018). An early multi-gene study by Martinsen et al. (2008) showed the genus *Haemoproteus* as paraphyletic with *Plasmodium* forming a sister relationship with subgenus *Parahaemoproteus*. In work by Borner et al. (2016), depending on the evolutionary model, taxa and genes used for analyses, the subgenus *Haemoproteus*, represented by *H. columbae*, formed a sister relationship with either subgenus *Parahaemoproteus* or with the genus *Plasmodium*. Furthermore, the relationships of the *Haemoproteus* subgenera have been open to debate, with one subgenus forming a sister relationship with *Plasmodium* as in Fig. 1a (Martinsen et al. 2008; Santiago-Alarcón et al. 2010; Martinez-de la Puente et al. 2011; Pacheco et al. 2018) or with both subgenera forming a monophyletic clade as in Fig. 1b (Valkiūnas et al. 2010, 2016; Levin et al. 2012; Carlson et al. 2013; Palinauskas et al. 2015; Lutz et al. 2016). While the dataset used in Borner et al. (2016) provided strong resolution for all other clades, it is interesting that *H. columbae*’s placement could not be resolved. Currently available sequencing technology provides an opportunity to address this issue but certain challenges must be acknowledged.

Producing large haemosporidian sequence databases (genomes or transcriptomes) requires knowledge of both parasite genome structure and life cycle (Lauron et al. 2014; Bensch et al. 2016; Böehme et al. 2018). Researchers pursuing genome sequencing must consider that current technology will provide sequences for both host and
parasite. Ratios of host-to-parasite DNA generally favor the host by wide margins due to avian host erythrocyte nucleation (Auburn et al. 2011; Oyola et al. 2012; Bensch et al. 2016; Böehme et al. 2018). The presence of nuclei in avian erythrocytes hinders the parasite’s whole-genome sequencing, although recently next generation sequencing and sophisticated parasite DNA isolation protocols have resulted in the sequencing of some avian haemosporidian genomes (Bensch et al. 2016; Böehme et al. 2018). In addition to parasite-host genome ratios, transcriptome assembly requires selecting either mRNA or total organismal RNA for preferential sequencing (Lauron et al. 2014; Videvall et al. 2017). Here we describe the processing of RNA-seq data to produce the first transcriptome of *H. columbae*. We also implement a large gene data set gathered from the transcriptome to attempt to resolve the *Haemoproteus* phylogenetic relationships described in Fig. 1. Included in the data set are sequences from 9 other haemosporidian parasites and 8 more distantly related apicomplexan parasites as detailed in Bensch et al. (2016).

**Methods**

**RNA Collection and Extraction**

A Rock Pigeon, *Columba livia domestica*, was captured at the campus of Universidad Nacional de Colombia located at 2560 m above sea level in the city of Bogotá, Colombia. Infection was verified by examining Giemsa stained slides. Whole blood was collected and stored in Trizol® LS reagent (Invitrogen, Grand Island, NY, USA). Samples were imported to San Francisco State University (USDA veterinary permit 114165). Giemsa slides were re-examined to verify infection by a single parasite lineage. RNA was extracted using a Trizol® LS (Invitrogen, Grand Island, NY, USA) extraction protocol in which phase-lock gel tubs were used to separate RNA from DNA and proteins in aqueous phases. Isopropyl ethanol and a high-salt solution were used to precipitate suspended RNA. After re-suspension, the RNA was treated with Ambion® TurboDNase™ before a size-separation step using Agencourt® RNA Clean® XP (Agencourt Bioscience Corporation, Beverly, Massachusetts, USA). Beads were applied to remove degraded RNA before re-suspension in DEPC treated water. Additional verification of parasite species was obtained by polymerase chain reaction of the cytochrome *b* gene (Hellgren et al. 2004).

**Library Preparation and Sequencing**

Library preparation was performed at the University of California, Berkeley Functional Genomics Laboratory. PolyA selection was used for mRNA enrichment via Invitrogen Dynabeads® mRNA Direct™ kit (Life Technologies, Carlsbad, CA, USA). Next, the Ovation® RNA-seq system (NuGEN Technologies, Inc, San Carlos, CA, USA) was used for cDNA synthesis and SPIA amplification. A S220 Focused-Ultrasonicator (Covaris, inc., Woburn, Massachusetts, USA) was used to fragment the cDNA which was then cleaned and concentrated using the MinElute® PCR Purification kit (Qiagen, Valencia, CA, USA). The sequencing library was prepared on an Apollo 324TM™ (Wafergen Biosystems, inc, Fremont, CA, USA) with PrepX ILM 32i DNA Library Kit (Wafergen Biosystems, inc, Fremont, CA, USA) and nine (9) cycles of polymerase chain reaction for library enrichment. Libraries were sequenced on an Illumina Hiseq-4000™ (Illumina, inc, San Diego, CA, USA) with read size selection of 100 base-pairs, paired-end. Raw reads were deposited to the NCBI biosample sequence archive (GenBank Accession No. SAMN06899305).
Raw Sequence Data Processing and Host Separation

The resulting reads from the sequencing described above were collected, and quality was assessed with the FastQC program (Babraham and Bioinformatics 2017). Read quality scores were high enough that very little quality trimming was necessary. Trimming was performed with BBDUK (Joint Genome Institute 2017), a quality filter and adapter removal program, removing sequencing adapters only. The genome of sample host C. livia was downloaded from NCBI (Accession: GCA_001887795.1) (National Center for Biotechnology Information 2017). The filtered reads were mapped to the C. livia genome using HISAT2 with the “very sensitive” option (Kim et al. 2015). Reads mapping to the C. livia genome were not exported, and only unmapped reads were used in the rest of the pipeline. Read output fastq files were relayed to the Trinity (2.1.1) program for de novo contig assembly (Haas et al. 2013). The resulting contigs were clustered using CD-HIT-EST (Li and Godzik 2005) to merge clusters with 97% similarity. The genome of Haemoproteus tartakovskyi was downloaded from the MalAvi database (Bensch et al. 2009). The clustered contigs were aligned to the H. tartakovskyi genome using BlastX (Altschul et al. 1990) and custom python scripts pulled the aligning sequences with significant matches (e-value: 1e-6). A second round of BlastN (Altschul et al. 1990) and separation was preformed to remove sequences matching to C. livia with an identity >= 90%. These successive removal steps ensured that the remaining contigs could be confidently associated with the H. columbae parasite. Transcriptome statistics (GC content, contig length, assembled bases and other standard assembly metrics) were calculated using Trinity’s built in statistical assessment program and bash scripts (Table 1). This transcriptome shotgun assembly project has been deposited at DDBJ/ENA/GenBank under the accession GGWD00000000. The version described in this paper is the first version, GGWD01000000.

Dataset Alignment and Phylogenetic Analysis

The dataset included 600 protein coding sequences for 17 taxa. The species names and original sources of each taxa dataset are provided in supplementary file (1) taxa sequence datasets were originally collected from the following databases: PlasmoDB (Aurrecoechea et al. 2008), ToxoDB (Gajria et al. 2007), PiroplasmaDB (PiroplasmaDB 2014), CryptoDB (Puig et al. 2004) and MalAvi (Bensch et al. 2009). The corresponding proteins were found in the transcriptome of H. columbae using TblastX (Altschul et al. 1990), and the resulting protein sequences were collected with custom python scripts and added to the Bensch dataset. Sequence alignment was performed with T-Coffee (Notredame et al. 2000) using default parameters. Gene IDs associated with each protein block are listed in supplementary file (2) Aligned protein files were filtered using Gblock (Castresana 2000) to remove gaps and regions of poor alignment. Gblock was run twice, once with default settings allowing no gaps in sequence products. The second Gblock run allowed gaps in alignments using the 50% allowance setting. Gblock alignments were imported to Geneious (version 7.1.0) for final sequence concatenation.

Model selection was preformed using Modelgenerator (version 85, Keane et al. 2006). Maximum likelihood analysis was performed on the concatenated gapped and ungapped datasets in RAxML with 100 bootstrapped phylogenetic trees (version 8.2.10, Stamatakis 2014) using an LG + G + F model. Bayesian analysis was performed with Mrbayes (version 3.2.6, Ronquist and Huelsenbeck 2003) using the LG + G + F evolutionary model. Mrbayes was run for 2 million generations, sampling every 100 generations before a burn in of 25%. Both analysis methods used Cryptosporidium parvum and Cryptosporidium muris as outgroups as detailed in Bensch et al. (2016). Individual gene trees were produced with RAxML for all 600 genes in the dataset, and a comparison of tree topologies matching the concatenated gapped dataset analysis was performed with the Sumbtree program (Sukumaran and Holder 2010).

Results

Sample Parasitemia

The parasitemia was 1.68%, calculated as the number of different parasite stages of H. columbae in 10,000 red blood cells.
Read Processing and Transcriptome Assembly

Sequences produced by the Hiseq-4000 totaled 110,331,097 100 nucleotide (nt) paired-end reads (220,662,194 total) equaling 22 Gbp. The BBDUK program removed 84,000 reads as Illumina adapter sequences (0.00038% of total reads). The mapping of reads to the genome of *C. livia* using HISAT2 resulted in 102 million reads remaining unmapped (46% of total reads) that were used for assembly. The assembly resulted in 267,604 contigs. After CD-HIT-EST, 220,867 (46% of total reads) that were used for assembly. The assembly resulted in 267,604 contigs. After CD-HIT-EST, 220,867 contigs remained after isoform clustering. The resulting sequences were mapped to the *H. tartakovskyi* genome using BlastX and contigs with significant e-value (< 1e−6) were separated with custom scripts. Only 26,781 contigs passed this filter. At this stage, the GC content was ~27% with an average contig length of 716 bp. The separate sequences were then mapped to *C. livia* genome using BlastN and any sequences with 90% identity to *C. livia* were removed. This led to 17,238 contigs passing filter with a GC of 17.78% and an average contig length of 769 bp. For all further purposes, this dataset is referred to as the *H. columbae* transcriptome.

Ortholog Clustering and Phylogenetic Analysis

Ortholog trimming with Gblock produced two distinct datasets. The first dataset consisted of 458 ortholog clusters. No missing data (sequence gaps) were permitted in this dataset. The second dataset contained 600 ortholog clusters but contained gaps in alignments. The discrepancy of 153 sets. The second dataset contained 600 ortholog clusters but contained gaps in alignments. The resulting sequences were mapped to the *H. tartakovskyi* genome using BlastX and contigs with significant e-value (< 1e−6) were separated with custom scripts. Only 26,781 contigs passed this filter. At this stage, the GC content was ~27% with an average contig length of 716 bp. The separate sequences were then mapped to *C. livia* genome using BlastN and any sequences with 90% identity to *C. livia* were removed. This led to 17,238 contigs passing filter with a GC of 17.78% and an average contig length of 769 bp. For all further purposes, this dataset is referred to as the *H. columbae* transcriptome.

Discussion

This study provides the first transcriptome data from a *Haemoproteus* parasite. This type of large genomic datasets improves the accuracy of evolutionary reconstructions. Haemosporidian genomic-scale data have also been instrumental in the discovery of many invasion genes in avian parasites as well as resolving the positions of mammalian and avian parasites (Martinez et al. 2013; Lauron et al. 2014; Videvall et al. 2017; Borner et al. 2016; Bensch et al. 2016). Transcriptomic approaches have been effective at characterizing *P. ashfordi* (Videvall et al. 2017), *P. gallinaceum* (Lauron et al. 2014), investigating the effects of parasitism on host gene expression (Videvall et al. 2015) and guiding genome assemblies (Boehme et al. 2018). The transcriptome of *H. columbae* opens new research opportunities for examining invasion gene variation between this and previously published subgenera (Lauron et al. 2014; Videvall et al. 2017). Additionally, sequence-based approaches to protein–protein interactions are feasible, such as ortholog based comparisons and phylogenetic mirror tree methods (Rao et al. 2014). It should be possible to utilize genomic datasets for either *P. ashfordi* or another avian parasite with well-characterized invasion genes, such as *P. gallinaceum* or *P. relictum*, to establish orthologous parasite proteins in *H. columbae* (Boehme et al. 2018). In a preliminary exploration of the transcriptome using Blast (Altschul et al. 1990) sequence matching, we discovered contigs matching portions of the *ama1* and *ron2* genes. Using this knowledge of ortholog data, numerous avian genomes may be used to infer host receptor proteins (Lee et al. 2008; Rao et al. 2014; Videvall et al. 2017).

Our study has added to the work of Bensch et al. (2016) by expanding it as a potential standard dataset for haemosporidian evolutionary analyses. While we must acknowledge that the taxon sampling in this dataset is low, we believe that it represents a firm foundation for future studies when genomes and transcriptomes will be available from other haemosporidian taxa. Here we also address the polytomy found in work by Borner et al. (2016) regarding the evolutionary relationships of *H. columbae* (Borner et al. 2016). The issue of the relationships of genus *Haemoproteus* requires significant literature review. Phylogenetic reconstructions in Martinsen et al. (2008), using a multi-gene analysis, showed *Parahaemoproteus* forming a sister relationship with *Plasmodium*, and the subgenus *Haemoproteus*, represented by *H. columbae*, forming a separate clade. A study by Santiago-Alarcón et al. (2010) describing the relationships of New World columbiform parasites focused on the phylogenetic relationships of subgenus *Haemoproteus* members. The author’s results
supported the sister relationship of Parahaemoproteus and Plasmodium while the subgenus Haemoproteus, including H. columbae, formed a separate clade. The authors performed analyses on two genes, the mitochondrial cyt b and the apicoplast caseinolytic protease C (ClpC) gene. Despite examining the product of the two-gene analyses, the recovered phylogenetic tree contained multiple polytomies (Santiago-Alarcón et al. 2010). Work by Valkiūnas et al. (2010) again supported a sister relationship of the subgenera Haemoproteus and Parahaemoproteus: this study focused on parasites primarily found in Columbiformes of the Galapagos Islands with most of the subgenus Haemoproteus being represented by H. multipigmentatus (Valkiūnas et al. 2010). Investigations by Martinez-de la Puente et al. (2011) found the subgenus Parahaemoproteus sister to Plasmodium and the subgenus Haemoproteus as a more distantly related clade. This study was also based only on cyt b analyses and is interesting due to the diversity of Parahaemoproteus members included in the study, while Haemoproteus was represented by only three lineages of H. columbae (Martinez-de la Puente et al. 2011). A recent study by Palinauskas et al. (2015) examined the differentiation of the cryptic parasite species Plasmodium homocircumflexum and Plasmodium circumflexum and included a Haemoproteus lineage for the phylogenetic analysis. This work showed the subgenus Haemoproteus and Parahaemoproteus as sister clades and Plasmodium as more distantly related (Palinauskas et al. 2015).
discussed earlier, work by Borner et al. (2016) displayed the instability of inferred *Haemoproteus* relationships. Depending on the analysis methods used, the topology vacillated to place the subgenus *Haemoproteus* as a sister clade to either *Parahaemoproteus* or *Plasmodium*. The methods included both maximum likelihood analysis and Bayesian inference of nucleotides and proteins (Borner et al. 2016). Work by Lutz et al. (2016) showed a monophyletic relationship for the genus *Haemoproteus*. The authors used mitochondrial cyt *b* sequences along with sequences for a single nuclear and an apicoplast gene. Interestingly, the authors found *Plasmodium* as a paraphyletic clade (Lutz et al. 2016). Recent work by Valkiūnas describing a new malaria parasite *Plasmodium delichoni* included phylogenetic analyses describing a sister relationship of the subgenus *Haemoproteus* and *Parahaemoproteus*, and a more distant relationship with *Plasmodium* (Valkiūnas et al. 2016). Finally, work by Pacheco et al. (2018) used 114 complete mitochondrial genomes of many taxa to infer the relationships of haemosporidians for the genera *Leucocytozoon*, *Plasmodium*, *Haemoproteus* and *Hepatocystis*. These phylogenetic reconstructions based on both Bayesian and likelihood methods found a sister relationship between *Plasmodium* and *Parahaemoproteus*. To summarize previous research, four articles supported a *Parahaemoproteus*/subgenus *Haemoproteus* sister relationship (Valkiūnas et al. 2010, 2016; Palinauskas et al. 2015; Lutz et al. 2016), five articles supported a *Plasmodium*/*Parahaemoproteus* sister relationship (Martinsen et al. 2008; Santiago-Alarcón et al. 2010; Martínez-de la Puente et al. 2011; Pacheco et al. 2018) and one article obtained mixed results depending on the taxa, evolutionary model and genes analyzed (Borner et al. 2016). Pacheco et al. (2018) bears special mention as the most recent study with the largest dataset in terms of complete mitochondrial genes (Pacheco et al. 2018). With relatively few deep-sequenced haemosporidian taxa available, we can only speculate as to the difference such a selection can have on phylogenetic inference. It should be noted that while much research has been performed supporting both topologies, most articles only used one method of phylogenetic analysis. Additionally, many studies also focused only on cyt *b* analyses, specifically a 478 bp sequence of cyt *b*. Pacheco et al. (2018) stress the necessity of expanding data collection standards beyond sequencing only the cyt *b* gene (Pacheco et al. 2018). Our attempt to clarify the relationships between the subgenus *Haemoproteus* and *Parahaemoproteus* by using a large number of nuclear genes strengthens the claim of a monophyletic *Haemoproteus* clade.

The research provided here adds to the greater knowledge of haemosporidian genetics but many questions remain. The lack of a *Leucocytozoon* genome or transcriptome prevents us from addressing the parasite relationships of the two subgenera of *Haemoproteus*. Additionally, the low taxon sampling implies that the relationships found here may change as more genomes and transcriptomes from haemosporidians become available. Additionally, coevolutionary studies between parasites and their vectors could add valuable pieces to parasite evolutionary reconstructions, as it seems likely parasites would adapt to closely related vectors (Lauron et al. 2015). Vector involvement in the evolution of haemosporidians remains an understudied area of research. We speculate that a method of sequencing both erythrocyte parasite life cycle stages and the oocyte or sporozoite stages present in the vector would prove advantageous in obtaining complete genetic data.

**Conclusions**

In closing, our findings are twofold. This study provides the first data about the transcriptome of blood parasites belonging to the family. The transcriptome of *H. columbae* represents a valuable resource for the advancement of haemosporidian genomic studies. Research on parasite antigen polymorphisms from a globally distributed parasite such as *H. columbae* could provide a significant system to study conserved invasion mechanisms. Additionally, our work on evolutionary relationships using large-scale transcriptomic datasets adds to our knowledge of potential haemosporidian evolutionary histories. Our dataset, combined with the original dataset from Bensch et al. (2016), will provide a stable framework to clarify the relationships of apicomplexan parasites in years to come.

**Acknowledgements** This work was supported by Research Council of Lithuania (Grant MIP-045/2015). This study was also supported by the grant to RNMS, NIH 1SC3GM118210-01A1. The authors would like to thank Elvin J. Lauron, Bradley Bowser and Dr. Frank Cipriano for helping in planning rna-seq and project logistics. We are grateful to Dr. Greg Spicer, Andrew Ontano, Dr. Emily Jane McTavish and Trent Liu for assistance on phylogenetic and scripting advisement.

**References**


Babraham Bioinformatics (2017) http://www.bioinformatics.babraham.ac.uk/


P projectId -change Malaria parasites. Nucleic Acids Res 37(Suppl_1):D539–D543

Babraham Bioinformatics (2017) http://www.bioinformatics.babraham.ac.uk/


National Center for Biotechnology Information (NCBI)[Internet] Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; Accessed Apr 12 2017


PiroplasmaDB (2014), http://piroplasmadb.org


Valkiūnas G (2005) Avian malaria parasites and other haemosporidia. CRC Press, Boca Raton


Videvall E (2018) Plasmodium parasites of birds have the most AT-rich genes of eukaryotes. Microb Genomics 4:1–9


