The transcriptome of the avian malaria parasite *Plasmodium ashfordi* displays host-specific gene expression

ELIN VIDEVALL,* CHARLIE K. CORNWALLIS,* DAG AHRÉN,*† VAIDAS PALINAUSKAS,‡ GEDIMINAS VALKIUNAS‡ and OLOF HELLGREN*  
*Department of Biology, Lund University, Sölvegatan 37, SE-22362 Lund, Sweden, †National Bioinformatics Infrastructure Sweden (NBIS), Lund University, Sölvegatan 37, SE-22362 Lund, Sweden, ‡Institute of Ecology, Nature Research Centre, Akademijos 2, LT-08412 Vilnius, Lithuania

**Abstract**

Malaria parasites (*Plasmodium* spp.) include some of the world’s most widespread and virulent pathogens. Our knowledge of the molecular mechanisms these parasites use to invade and exploit their hosts other than in mice and primates is, however, extremely limited. It is therefore imperative to characterize transcriptome-wide gene expression from nonmodel malaria parasites and how this varies across individual hosts. Here, we used high-throughput Illumina RNA sequencing on blood from wild-caught Eurasian siskins experimentally infected with a clonal strain of the avian malaria parasite *Plasmodium ashfordi* (lineage GRW2). Using a bioinformatic multistep approach to filter out host transcripts, we successfully assembled the blood-stage transcriptome of *P. ashfordi*. A total of 11 954 expressed transcripts were identified, and 7860 were annotated with protein information. We quantified gene expression levels of all parasite transcripts across three hosts during two infection stages – peak and decreasing parasitemia. Interestingly, parasites from the same host displayed remarkably similar expression profiles during different infection stages, but showed large differences across hosts, indicating that *P. ashfordi* may adjust its gene expression to specific host individuals. We further show that the majority of transcripts are most similar to the human parasite *Plasmodium falciparum*, and a large number of red blood cell invasion genes were discovered, suggesting evolutionary conserved invasion strategies between mammalian and avian *Plasmodium*. The transcriptome of *P. ashfordi* and its host-specific gene expression advances our understanding of *Plasmodium* plasticity and is a valuable resource as it allows for further studies analysing gene evolution and comparisons of parasite gene expression.

**Keywords:** gene expression, host–parasite interaction, malaria, *Plasmodium*, RNA-seq, transcriptome

Received 9 September 2016; revision received 16 January 2017; accepted 17 February 2017

**Introduction**

The apicomplexan parasites of the genus *Plasmodium* (malaria parasites) encompass a worldwide distribution and infect a multitude of vertebrate hosts, including reptiles, birds and mammals (Garnham 1966). Their virulence can be highly variable between different strains and species. Some induce mild pathogenic effects on hosts and some cause severe disease, leading to high mortality rates (Palinauskas et al. 2008). Host individuals and host species also differ in their resistance and tolerance to malaria, and this interaction between host and parasite ultimately determines disease severity. Furthermore, the molecular response of hosts changes during the course of infection and creates a dynamic environment in which the parasites need to
accommodate. Nevertheless, our understanding of how malaria parasites respond molecularly to different host individuals and to changes in the host immune defence over time is very limited.

Parasites of two clades in *Plasmodium* have been extensively studied from a molecular perspective, murine and primate parasites. We have learned a great deal about how malaria parasites of humans evolved and function by studying transcriptomes of their rodent-infecting relatives (Hall et al. 2005; Spence et al. 2013; Otto et al. 2014a). The majority of studies investigating gene expression in human malaria parasites have been conducted using cell lines (in vitro) or tissue cultures (ex vivo), which has provided tremendous insight into the biology of *Plasmodium* life stages (e.g. Bozdech et al. 2003; Otto et al. 2010; Siegel et al. 2014). However, several discrepancies in parasite expression between cultures and live animals (in vivo) have been documented (Lapp et al. 2015) and a wide range of host environmental factors are absent in the in vitro systems. For example, temperature fluctuations, inflammatory and immune effector molecules, hormones, metabolites, microenvironments, and varying levels of oxygen, pH and glucose are difficult to simulate in in vitro settings (LeRoux et al. 2009). Parasites cultured outside hosts reflect this with different expression patterns and markedly downregulate important vaccine candidate genes such as cell surface antigens (Daily et al. 2005; Siau et al. 2008). The natural host environment, which includes genotypic and immunological cues, may therefore strongly affect the transcriptional responses of malaria parasites. To obtain representative transcriptional information from *Plasmodium* parasites, it is therefore valuable to study natural host systems.

The molecular mechanisms that enable successful invasion and establishment of malaria parasites in hosts other than mice and primates are unfortunately poorly known. Which genes are conserved across *Plasmodium* and how do virulence, immune evasion and host specificity vary in species infecting nonmammalian animals? To investigate these questions, it will be necessary to assemble and characterize genomewide expression information from malaria parasites across their host phylogenetic range. With recent developments of high-throughput sequencing techniques, it has now become possible to generate genomic sequences from nonmodel parasites (Martinsen & Perkins 2013). Dual RNA sequencing of hosts and their parasites opens up possibilities of simultaneously studying host–parasite interactions and describing transcriptome expression in both actors. Assembling novel parasite sequences de novo is, however, a very difficult task. Without a reference genome, transcripts from the host and/or other sources of contamination may remain after annotation and can influence downstream analyses. Meticulous filtering of assemblies using bioinformatics is therefore crucial to avoid erroneous conclusions (see e.g., Koutsovoulos et al. 2016). Nevertheless, successfully constructing parasite transcriptome data of high quality from different hosts will provide us with valuable insights into the hidden biology of *Plasmodium*.

Malaria parasites that infect birds provide an excellent opportunity for studying transcriptional parasite responses due to their enormous diversity and large variation in host specificity and virulence (Bensch et al. 2004; Križanauskienė et al. 2006; Lachish et al. 2011). They are closely related to mammalian *Plasmodium*, but it was only recently established that rodent and primate malaria parasites are indeed monophyletic in the *Plasmodium* phylogeny (Bensch et al. 2016). Some avian *Plasmodium* are extreme host generalists, successfully infecting birds over several orders, while other parasites are host specialists infecting a single species (Pérez-Tris et al. 2007; Drovetski et al. 2014). The avian malaria system allows for the possibility of capturing wild birds in a natural setting, and evaluating their status of malaria infection. Additionally, recent infection experiments have illustrated the potential to study *Plasmodium* in passerines under controlled conditions in laboratories (Palinauskas et al. 2008, 2011; Zehntndiev et al. 2008; Cornet et al. 2014; Dimitrov et al. 2015; Ellis et al. 2015).

In this study, we used a bioinformatic multistep filtering approach to assemble the blood transcriptome of the avian malaria parasite *Plasmodium ashfordi*, mitochondrial lineage GRW2. This parasite was first associated with various warblers (families Acrocephalidae and Phylloscopidae) (Valkiūnas et al. 2007), but has been found since 2016 in 15 different host species of two avian orders (Bensch et al. 2009). We used our assembly of *P. ashfordi* to evaluate transcriptome characteristics and genomewide sequence similarity to other apicomplexans and searched for genes known to be involved in the *Plasmodium* red blood cell invasion process. We analysed expression levels of parasite genes in three experimentally infected birds during two infection stages, peak and decreasing parasitemia, where the hosts have previously been shown to exhibit different transcriptome responses (Videvall et al. 2015). This allowed us for the first time to follow and describe the transcriptome of an avian malaria parasite over time in individual hosts.

**Methods**

**Experimental set-up**

We used four wild-caught juvenile Eurasian siskins (*Carduelis spinus*) in an infection experiment. The experimental procedure was carried out 2012 at the Biological
The parasite strain was originally collected in 2011 from a single common cuckoo (Cuculus canorus) that had acquired the infection naturally. It was thereafter multiplied in common crossbills (Loxia curvirostra) in the laboratory and deep frozen in liquid nitrogen for storage. One crossbill was subsequently infected with the parasite, and blood from this single bird was used to infect our experimental birds. Crossbills were used as intermediate hosts because they are both susceptible to this strain and have a large enough body size to provide the amount of blood needed for donations. A subinoculation of a freshly prepared mixture containing infected blood from the donor was made into the pectoral muscle of the recipient birds (details of procedure can be found in Palinauskas et al. 2008). Using a single donor, we ensured that the same clonal parasite strain and parasite quantity was injected into recipient birds.

All birds were thoroughly screened with both microscopic (Palinauskas et al. 2008) and molecular (Hellgren et al. 2004) methods before the experiment to make sure they had no prior haemosporidian infection. Blood samples for RNA sequencing were taken from birds before infection (day 0), during peak parasitemia (day 21 postinfection) and during the decreasing parasitemia stage (day 31 postinfection). The parasitemia intensity varied substantially in infected birds, with the parasite, and blood from this single bird was used to infect our experimental birds. Crossbills were used as intermediate hosts because they are both susceptible to this strain and have a large enough body size to provide the amount of blood needed for donations. A subinoculation of a freshly prepared mixture containing infected blood from the donor was made into the pectoral muscle of the recipient birds (details of procedure can be found in Palinauskas et al. 2008). Using a single donor, we ensured that the same clonal parasite strain and parasite quantity was injected into recipient birds.

All birds were thoroughly screened with both microscopic (Palinauskas et al. 2008) and molecular (Hellgren et al. 2004) methods before the experiment to make sure they had no prior haemosporidian infection. Blood samples for RNA sequencing were taken from birds before infection (day 0), during peak parasitemia (day 21 postinfection) and during the decreasing parasitemia stage (day 31 postinfection). The parasitemia intensity varied substantially in infected birds, with the parasite, and blood from this single bird was used to infect our experimental birds. Crossbills were used as intermediate hosts because they are both susceptible to this strain and have a large enough body size to provide the amount of blood needed for donations. A subinoculation of a freshly prepared mixture containing infected blood from the donor was made into the pectoral muscle of the recipient birds (details of procedure can be found in Palinauskas et al. 2008). Using a single donor, we ensured that the same clonal parasite strain and parasite quantity was injected into recipient birds.

All birds were thoroughly screened with both microscopic (Palinauskas et al. 2008) and molecular (Hellgren et al. 2004) methods before the experiment to make sure they had no prior haemosporidian infection. Blood samples for RNA sequencing were taken from birds before infection (day 0), during peak parasitemia (day 21 postinfection) and during the decreasing parasitemia stage (day 31 postinfection). The parasitemia intensity varied substantially in infected birds, with the parasite, and blood from this single bird was used to infect our experimental birds. Crossbills were used as intermediate hosts because they are both susceptible to this strain and have a large enough body size to provide the amount of blood needed for donations. A subinoculation of a freshly prepared mixture containing infected blood from the donor was made into the pectoral muscle of the recipient birds (details of procedure can be found in Palinauskas et al. 2008). Using a single donor, we ensured that the same clonal parasite strain and parasite quantity was injected into recipient birds.

All birds were thoroughly screened with both microscopic (Palinauskas et al. 2008) and molecular (Hellgren et al. 2004) methods before the experiment to make sure they had no prior haemosporidian infection. Blood samples for RNA sequencing were taken from birds before infection (day 0), during peak parasitemia (day 21 postinfection) and during the decreasing parasitemia stage (day 31 postinfection). The parasitemia intensity varied substantially in infected birds, with the parasite, and blood from this single bird was used to infect our experimental birds. Crossbills were used as intermediate hosts because they are both susceptible to this strain and have a large enough body size to provide the amount of blood needed for donations. A subinoculation of a freshly prepared mixture containing infected blood from the donor was made into the pectoral muscle of the recipient birds (details of procedure can be found in Palinauskas et al. 2008). Using a single donor, we ensured that the same clonal parasite strain and parasite quantity was injected into recipient birds.

All birds were thoroughly screened with both microscopic (Palinauskas et al. 2008) and molecular (Hellgren et al. 2004) methods before the experiment to make sure they had no prior haemosporidian infection. Blood samples for RNA sequencing were taken from birds before infection (day 0), during peak parasitemia (day 21 postinfection) and during the decreasing parasitemia stage (day 31 postinfection). The parasitemia intensity varied substantially in infected birds, with the parasite, and blood from this single bird was used to infect our experimental birds. Crossbills were used as intermediate hosts because they are both susceptible to this strain and have a large enough body size to provide the amount of blood needed for donations. A subinoculation of a freshly prepared mixture containing infected blood from the donor was made into the pectoral muscle of the recipient birds (details of procedure can be found in Palinauskas et al. 2008). Using a single donor, we ensured that the same clonal parasite strain and parasite quantity was injected into recipient birds.

All birds were thoroughly screened with both microscopic (Palinauskas et al. 2008) and molecular (Hellgren et al. 2004) methods before the experiment to make sure they had no prior haemosporidian infection. Blood samples for RNA sequencing were taken from birds before infection (day 0), during peak parasitemia (day 21 postinfection) and during the decreasing parasitemia stage (day 31 postinfection). The parasitemia intensity varied substantially in infected birds, with the parasite, and blood from this single bird was used to infect our experimental birds. Crossbills were used as intermediate hosts because they are both susceptible to this strain and have a large enough body size to provide the amount of blood needed for donations. A subinoculation of a freshly prepared mixture containing infected blood from the donor was made into the pectoral muscle of the recipient birds (details of procedure can be found in Palinauskas et al. 2008). Using a single donor, we ensured that the same clonal parasite strain and parasite quantity was injected into recipient birds.

All birds were thoroughly screened with both microscopic (Palinauskas et al. 2008) and molecular (Hellgren et al. 2004) methods before the experiment to make sure they had no prior haemosporidian infection. Blood samples for RNA sequencing were taken from birds before infection (day 0), during peak parasitemia (day 21 postinfection) and during the decreasing parasitemia stage (day 31 postinfection). The parasitemia intensity varied substantially in infected birds, with the parasite, and blood from this single bird was used to infect our experimental birds. Crossbills were used as intermediate hosts because they are both susceptible to this strain and have a large enough body size to provide the amount of blood needed for donations. A subinoculation of a freshly prepared mixture containing infected blood from the donor was made into the pectoral muscle of the recipient birds (details of procedure can be found in Palinauskas et al. 2008). Using a single donor, we ensured that the same clonal parasite strain and parasite quantity was injected into recipient birds.

All birds were thoroughly screened with both microscopic (Palinauskas et al. 2008) and molecular (Hellgren et al. 2004) methods before the experiment to make sure they had no prior haemosporidian infection. Blood samples for RNA sequencing were taken from birds before infection (day 0), during peak parasitemia (day 21 postinfection) and during the decreasing parasitemia stage (day 31 postinfection). The parasitemia intensity varied substantially in infected birds, with the parasite, and blood from this single bird was used to infect our experimental birds. Crossbills were used as intermediate hosts because they are both susceptible to this strain and have a large enough body size to provide the amount of blood needed for donations. A subinoculation of a freshly prepared mixture containing infected blood from the donor was made into the pectoral muscle of the recipient birds (details of procedure can be found in Palinauskas et al. 2008). Using a single donor, we ensured that the same clonal parasite strain and parasite quantity was injected into recipient birds.
NR database will not result in any significant blast hits to genes unique in avian malaria parasites. We strictly filtered the initial assembly by only retaining a total of 9015 transcripts (isoforms) that produced significant blast matches against proteins from species in the Apicomplexa phylum. A previous assembly using an earlier version of TRINITY (v. r20140413p1) (Grabherr et al. 2011) performed better when it came to assembling the longest contigs (>6 kbp). Different versions of assembly software may construct de Bruijn graphs somewhat differently, which is why it can be a good idea to make several assemblies and later combine parts of them (Brian Haas, personal communication). The previous assembly had been blasted and screened for Apicomplexa in exactly the same way as described above. In order not to lose these important full-length transcripts, we therefore included the longest contigs from the previous assembly that had (i) not assembled correctly in the current assembly, and (ii) significant blastx hits against Apicomplexa ($n = 10$), resulting in a total of 9025 transcripts. The fact that these contigs contained similar sequences already present in the assembly was dealt with through downstream clustering of the sequences.

Transcriptome cleaning and filtering

Some contigs in the annotated assembly contained poly-A tails which complicated downstream analyses and resulted in biased mapping estimates. We therefore removed all poly-A/T tails using PRINSEQ (v. 0.20.4) (Schmieder & Edwards 2011) with a minimum prerequisite of 12 continuous As or Ts in the 5’ or 3’ end of

Fig. 1 Filtering of host transcripts from the Plasmodium ashfordi transcriptome using gene annotation and GC content. Density curves of contig GC content in (A) the initial, unfiltered assembly, (B) the annotated $P$. ashfordi transcriptome assembly, (C) all contigs giving significant blastx matches to birds and (D) all unknown contigs before GC % filtering. The arrows indicate assembly versions before and after initial filtering and cleaning steps. Both the initial, unfiltered assembly and the assembly with unknown, unfiltered contigs display a bimodal distribution, incorporating both avian and malaria parasite transcripts. The dashed straight line in D indicates the 23% GC cut-off where unknown transcripts with lower GC content were extracted, filtered and later included in the final $P$. ashfordi assembly as unannotated transcripts.
the contigs. A total of 106 202 bases (1.18%) was trimmed and the mean transcript length was reduced from 995.28 to 983.52 bases.

The unknown transcripts that failed to produce significant hits to any organism during the blastx run \((n = 260 162)\) were subsequently cleaned using the following procedure. First, we trimmed them for poly-A tails, resulting in a total of 455 331 bases removed, and a slight decrease of the mean length of the unknown sequences from 555.27 nt before trimming to 553.52 nt after trimming. The majority of these unknown transcripts came from host mRNA, but their GC content displayed a clear bimodal distribution (Fig. 1D), where the contigs with very low GC were strongly suspected to originate from the parasite. To avoid any host contigs, we strictly filtered the unknown transcripts to only include sequences with a mean GC content lower than 23% \((n = 4624)\). This threshold was based on the Api-complexa-matching transcripts (mean GC = 21.22%), the contigs matching birds (class: Aves; \(n = 47 823\); mean GC = 47.65%) and the bird contig with the absolute lowest GC content (GC = 23.48%) (Fig. 1C).

Transcriptome clustering, further filtering and validation

To reduce redundancy of multiple isoforms and transcripts derived from the same gene, we first merged together the annotated and the unknown transcripts with a GC content \(<23%\) \((n = 13 649)\). We then clustered these sequences together to retain most transcripts but group the highly similar ones based on 97% sequence similarity and a k-mer size of 10, using CD-HIT-EST (v. 4.6) (Li & Godzik 2006). The most representative (longest) contig in every cluster was selected to represent those transcripts, resulting in 12 266 contigs/clusters. We further filtered all the short sequences (<200 bases), to obtain a set of 12 182 representative transcripts.

A second blast filtering step with the trimmed representative contigs against the RefSeq genomic database was then employed using BLASTN+ (v. 2.2.29) (Altschul et al. 1990; Camacho et al. 2009) to identify some ambiguous contigs suspected to contain noncoding RNA bird sequences. We removed all contigs that gave significant matches \((e\text{-value} < 1e^{-6})\) against all animals (kingdom: Metazoa), so we could be confident that the assembly only consisted of true parasite transcripts. This last filtering step removed 228 contigs.

The unannotated transcripts \((n = 4094)\) of the final assembly were further validated to originate from the parasite using reads from the six uninfected samples of the same hosts sampled before infection and the control bird. A total of 350 318 482 reads (65 bp and 90 bp) from all uninfected samples were mapped to the unannotated transcripts using Bowtie2 (v. 2.2.5) (Langmead & Salzberg 2012), resulting in the alignment of 90 read pairs (0.000051%). This extremely low mapping percentage from the uninfected samples greatly supported our conclusion that these transcripts had indeed been transcribed by \(P.\) ashfordi. These 4094 representative transcripts with unknown function are referred to throughout the article as unannotated transcripts. The resulting final transcriptome assembly consists of 11 954 representative annotated and unannotated transcripts.

Estimating expression levels

Poly-A tails of 489 million RNA-seq reads from all samples were trimmed as well using PRINSEQ (v. 0.20.4) (Schmieder & Edwards 2011). A minimum prerequisite for trimming was 20 continuous As or Ts in the 5’ or 3’ end of each read. Only trimmed reads longer than 40 bp and still in a pair were retained \((n = 451 684 626)\) to confidently map high-quality reads with good minimum lengths. Bowtie2 (v. 2.2.5) (Langmead & Salzberg 2012) was used to map the trimmed RNA-seq reads of every sample \((n = 8)\) (six biological and two technical replicates) back to the \(P.\) ashfordi transcriptome consisting of the 11 954 representative sequences. We calculated expression levels using RSEM (v. 1.2.21) (Li & Dewey 2011), which produces expected read counts for every contig.

The counts of the 11 954 transcripts were subsequently analysed inside the \(R\) statistical environment (v. 3.2.5) (R Core Team 2015). We tested for expression differences in the malaria parasites between the two time points and between the hosts, using DESeq2 (v. 1.10.1) (Love et al. 2014). The two resequenced samples (technical replicates) of bird 3 and bird 4 during peak parasitemia were handled exactly the same as all other samples, and their respective read count was added to their biological samples, according to the DESeq2 manual. Counts were normalized to account for potential variation in sequencing depth as well as the large differences in number of parasites present in the blood (parasitemia levels). Regularized log transformation of counts was performed to represent the data without any prior knowledge of sampling design in the principal component analysis and sample distance calculations. This way of presenting counts without bias is preferred over variance stabilizing of counts when normalization size factors varies greatly between the samples (Love et al. 2014), as they naturally do in our data.

Variant calling and SNP analyses

Sequence variation in the transcriptome assembly was performed according to the Genome Analysis Toolkit
Best Practices workflow for variant calling in RNA-seq data (v. 2015-12-07). The BAM files for each sample produced by \textit{rsem} were sorted, read groups added and duplicate reads removed using \textit{Picard Tools} (v. 1.76) (https://broadinstitute.github.io/picard/). With the \textit{SplitNCigarReads} tool in \textit{GATK} (v. 3.4-46) (McKenna et al. 2010), the mapped reads were further reassigned mapping qualities, cleaned of Ns and hard-clipped of overhang regions. Variant calling was performed in \textit{GATK} with \textit{HaplotypeCaller} using ploidy = 1 and optimized parameters recommended for RNA-seq data (see \textit{GATK} Best Practices). Indels were excluded using the SelectVariants tool, and variants were filtered on quality using the \textit{VariantFiltration} tool in \textit{GATK} according to the filter recommendation for RNA-seq data. Next, we used \textit{SnpSift} in \textit{SnpEff} (v. 4_3 g) (Cingolani et al. 2012) to filter variants on depths of a minimum of 20 high-quality, nonduplicated reads. Finally, to calculate allele frequencies of all samples in variant positions called in only some hosts, we ran \textit{HaplotypeCaller} again in the --ERC BP\_RESOLUTION mode and extracted the read depths of the nucleotide positions called previously. Nucleotide positions were filtered if they were positioned in the very end of contigs or had a depth of < 20 high-quality, nonduplicated reads according to \textit{GATK}. Allele frequencies were calculated by dividing the number of reads supporting the alternative nucleotide with the total number of reads at each variant site.

\textbf{Transcriptome evaluation}

Transcriptome statistics such as GC content, contig length and assembled bases were calculated using Bash scripts and in the \textit{R} statistical environment (v. 3.2.5) (Pages et al. 2015; R Core Team 2015). \textit{P}-values were corrected for multiple testing with the Benjamini–Hochberg false discovery rate (Benjamini & Hochberg 1995) and corrected values have been labelled as \textit{q}-values throughout. We calculated the GC content of two \textit{Eimeria} transcriptomes downloaded from \textit{ToxODB} (v. 25) (Gajria et al. 2007), initially sequenced by Reid et al. (2014). Transcriptome E90N50 was calculated using \textit{rsem} (v. 1.2.21) and \textit{Trinity} (v. 2.0.6) (Grabherr et al. 2011; Li & Dewey 2011; Haas 2016). Plots were made with the \textit{R} package \textit{ggplot2} (v. 2.2.1) (Wickham 2009). The transcriptome of \textit{Plasmodium falciparum} 3D7 (v. 25) was downloaded from PlasmoDB (Gardner et al. 2002; Aurrecochea et al. 2009) and gene ontology information was derived from UniprotKB (Bateman et al. 2015). The red blood cell invasion genes were searched for in the transcriptome annotation we produced for \textit{P. ashfordi} (Table S1, Supporting information). Only genes with documented involvement in \textit{Plasmodium} erythrocyte invasion were included in the search.

\textbf{Results}

\textit{The Plasmodium ashfordi transcriptome assembly}

The transcriptome of \textit{P. ashfordi} was assembled into two versions to make it transparent and as useful as possible for other researchers. The first assembly version, which we refer to as the annotated assembly, contains the transcripts with annotation information from proteins of Apicomplexa (\textit{n} = 7860) (Fig. 1B; Table S1, Supporting information). The second version that we refer to as the total assembly, also includes the unannotated contigs (\textit{n} = 4094) that were strictly filtered to remove contigs from the host (Fig. 1), resulting in a total of 11 954 representative transcripts (Table 1). The genomes of \textit{Plasmodium} parasites generally contain around 5-6000 protein-coding genes (Kersey et al. 2016), making it reasonable to assume similar gene numbers for \textit{P. ashfordi}.

Building eukaryotic transcriptomes de novo, however, naturally yields many more transcripts than there are genes in the genome (Grabherr et al. 2011), and so the larger number of transcripts in our assembly is a result of isoform varieties, fragmented contigs and potential noncoding RNA.

The size of the total transcriptome assembly is 9.0 Mbp and the annotated version of the assembly is 7.3 Mbp (81.20\%) (Table 1). We calculated assembly statistics using the transcriptome-specific measurement E90N50, which represents the contig N50 value based on the set of transcripts representing 90\% of the expression data, and is preferable over the original N50 when evaluating transcriptome assemblies (Haas 2016). The assembly E90N50 is 1988 bp and the mean transcript length of the annotated transcripts is 930.8 bp (Table 1; Fig. S1, Supporting information). In comparison, the length of coding sequences in the genome of \textit{Haemoproteus tartakovskyi} (bird parasite in a sister genus to

<table>
<thead>
<tr>
<th>Table 1 Plasmodium ashfordi transcriptome assembly statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annotated transcripts</td>
</tr>
<tr>
<td>-----------------------</td>
</tr>
<tr>
<td>Number of contigs</td>
</tr>
<tr>
<td>Number of bases</td>
</tr>
<tr>
<td>Contig length, min–max (bp)</td>
</tr>
<tr>
<td>Median contig length (bp)</td>
</tr>
<tr>
<td>Mean contig length (bp)</td>
</tr>
<tr>
<td>GC content (%)</td>
</tr>
</tbody>
</table>

© 2017 John Wiley & Sons Ltd
**Plasmodium** is mean = 1206 and median = 1809 bp (Bensch et al. 2016). The human parasite *Plasmodium falciparum* has a transcriptome with transcripts of median length 1320 and mean length 2197 bp (Gardner et al. 2002). The longest contig in the *P. ashfordi* assembly, consisting of 26 773 bp, is transcribed from the extremely long ubiquitin transferase gene (AK88_05171), which has a similar transcript length of around 27 400 bp in other *Plasmodium* species. The annotated transcriptome has an exceptionally low mean GC content of 21.22% (Fig. 1B), which is even lower than the highly AT-biased transcriptome of *P. falciparum* (23.80%).

**Functional analysis suggests that many host-interaction genes have evolved beyond recognition**

To evaluate biological and molecular functions of all annotated transcripts in *P. ashfordi*, we first analysed their associated gene ontology. An analysis of the transcriptome of *P. falciparum* was performed simultaneously to get an appreciation of how the *P. ashfordi* assembly compares functionally to a closely related, well-studied species. Overall, the two transcriptomes displayed highly similar gene ontology patterns (Fig. 2A–B; Tables S8–S14, Supporting information). Transcripts primarily belonged to the two major molecular functions: ‘binding’ and ‘catalytic activity’; as well as the biological groups: ‘metabolic process’ and ‘cellular process’. These broad functional classes were also the gene ontology terms where *P. ashfordi* displayed a greater number of transcripts relative to *P. falciparum* (Fig. 2D). They are likely to contain multiple transcripts per gene due to isoforms, fragmented contigs or possible gene duplications. The categories ‘receptor activity’, ‘cell adhesion’ and ‘multi-organism process’ were in contrast almost exclusively occupied by transcripts from *P. falciparum*. Interestingly, these categories predominately relate to the interaction with host cells and host defences (Fig. 2C), which are known to contain genes showing strong evidence for positive selection in *Plasmodium* (Jeffares et al. 2007). It is highly likely that many *P. ashfordi* transcripts that belong to these host-interaction processes have diverged sufficiently from mammalian parasite genes to escape annotation, but are expressed and present in the unannotated portion of the transcriptome assembly.

We explored the metabolic processes of *P. ashfordi* by looking into the child terms of this gene ontology. All metabolic categories of *P. ashfordi* contained similar or slightly more transcripts compared to *P. falciparum*, except the ‘catabolic process’ where *P. ashfordi* had fewer transcripts (Table S12, Supporting information). An investigation of the broad gene ontology category ‘kinase activity’ resulted in a total of 235 transcripts (Table S1, Supporting information). The kinase...
multigene family FIKK has drastically expanded in the genomes of *P. falciparum* and *P. reichenowi* (Otto *et al.* 2014b), but only exists as one copy in the rodent malaria parasites. We used the proteins in the FIKK family from *P. falciparum* (n = 21) and *P. yoelii* (n = 1) to search for matches in our assembly. All 22 protein sequences produced significant blast matches (e-value < 1e-5) against a single *P. ashfordi* transcript (TR71526|c0.g2.i1), further indicating that the FIKK gene expansion in primate malaria parasites likely happened after the split from both avian and rodent *Plasmodium*. Complete results of all gene ontology analyses of *P. ashfordi* and *P. falciparum* can be found in Tables S8–S14 (Supporting information).

**Gene expression is similar across different stages of infection**

Next, we analysed expression levels of the *P. ashfordi* transcripts within individual hosts across the two parasitemia stages. We accounted for differences in parasitemia levels between hosts and time points, and any variation in sequencing depth between samples, by normalizing overall expression values according to the DESeq method (Anders & Huber 2010). We found that the parasites displayed very similar gene expression patterns during peak and decreasing parasitemia stages (Fig. 3A–E). No genes were significantly differentially expressed between the time points (q-value > 0.99), and the correlation in gene expression was extremely high (Pearson’s product–moment correlation = 0.9983, t = 1905.2, df = 11 952, P-value < 2.2e-16) (Fig. 3A; Table S2, Supporting information). Annotated transcripts showing the highest expression fold change (nonsignificantly) between the two parasitemia stages were derived from the following genes (in order of most observed change): rho-GTPase-activating protein 1, 40S ribosomal protein S3a, two uncharacterized proteins, TATA-box-binding protein, heat-shock protein 90 and C50 peptidase (Fig. 3D; Table S2, Supporting information).

**Gene expression is host-specific**

In contrast to the similarities in gene expression between parasitemia stages, the parasite transcriptomes showed much larger differences in expression levels between the different host individuals. A principal component analysis of expression variance clustered parasite samples together within their respective hosts, showing major similarities in expression profiles (Fig. 3B). Samples derived from the same host individual did not separate until the third (15% variance) and fourth (13% variance) principal component dimensions (Fig. 3C). The parasite transcriptome from host four during decreasing parasitemia showed the largest variation in parasite gene expression among all samples, yet it was still most similar to the transcriptome from the same host during peak parasitemia (Fig. 3B, E). In fact, all parasite transcriptomes during the decreasing parasitemia stage demonstrated closest distance to the transcriptome sample within the same host 10 days earlier (Fig. 3E). We further evaluated if specific transcripts contributed to the differences in parasite gene expression levels between individual hosts by performing a likelihood ratio test over all host individuals while controlling for parasitemia stage (time point). This resulted in 28 significant *P. ashfordi* transcripts (q-value < 0.1) displaying very high expression variation between hosts (Fig. 4; Table S3, Supporting information). The most significant transcripts were derived from the genes cytochrome c oxidase subunit 1, 70-kd heat shock-like protein, M1 family aminopeptidase and metabolite/drug transporter (Table S3, Supporting information).

*Plasmodium ashfordi* is genetically identical in all hosts

In an effort to investigate potential mechanisms behind the host-specific gene expression, we first evaluated if different haplotypes (multiclonality) of the parasite were present and had established differentially in the host individuals. As described in the Methods, all hosts were infected with the same clonal parasite isolate determined by the cytochrome b locus and derived from a single donor. This was also independently verified in all samples by examining read mapping of mitochondrial transcripts (Fig. S5, Supporting information). Because sexual recombination of *Plasmodium* takes place in the mosquito, multiple alleles of nuclear genes could have been present in the parasite strain injected into the birds. A sensitive single nucleotide polymorphism (SNP) analysis over the entire *P. ashfordi* transcriptome found, however, extremely little variation in the parasite. During peak infection, we recovered a total of 10 (unfiltered) SNPs in the 11 954 *P. ashfordi* transcripts from the parasites in host 2, 32 SNPs in host 3 and 46 SNPs in host 4. The variation in number of SNPs called between parasites in different host individuals was due to read coverage, which is directly dependent on parasitemia levels, where host 2 had the lowest, host 3 intermediate and host 4 highest parasitemia (see Methods). After filtering on read depth, a total of 19 SNPs were identified and used in analyses (Table S4, Supporting information). We discovered that all SNPs (100%) were present in the parasites of all three host individuals, which carried the exact same allele polymorphism (e.g., C/T). To determine if the two alleles were expressed
differentially in the hosts, allele frequency was calculated for each SNP in the parasite transcriptome from each host during peak infection. An analysis of the SNPs found no differences in allele frequencies between host individuals (Friedman rank sum test, Friedman chi-squared = 1.68, df = 2, P-value = 0.432), indicating similar allele expression levels in all parasite samples (Fig. 3F).

The 19 SNPs were located inside nine of 11 954 contigs in the P. ashfordi transcriptome, resulting in an allelic occurrence of 0.075%. For a visual example of a contig with three SNPs present, see Fig. S6 (Supporting information). Four of the nine transcripts containing SNPs were unannotated, and the others were derived from the genes merozoite surface protein 9 (MSP9), surface protein P113, ubiquitin-related protein, multidrug resistance protein and a conserved Plasmodium protein with unknown function (Table S4, Supporting information).

The presence of extremely few sequence polymorphisms in the transcriptome of P. ashfordi demonstrates that the parasite strain used was not only clonal with respect to the mitochondrial lineage, but homogeneous over the vast majority of the transcriptome.
Furthermore, the detection of 19 SNPs that were identical in the parasites from all host individuals verifies that the parasite is genetically identical in all hosts. Intriguingly, it also means that the *P. ashfordi* isolate originally consisted of a minimum of two different haplotypes which differed at nine genes before the experiment and that this genetic variation managed to survive in the parasite population during the three-week-long infection, remaining in similar frequency in all hosts.

**Parasite sexual development does not differ between hosts**

We further evaluated the possibility that transcriptome host clustering were due to consistent differences relating to the sexual development of *P. ashfordi* during both time points in the host individuals. Lemieux *et al.* (2009) found that differences in *P. falciparum* gene expression between blood samples from children could be partially explained because the parasite expressed different genes during its sexual development (gametocyte) stage. We counted gametocytes in blood slides from the different samples using microscopy (Table S5, Supporting information) and found no differences in gametocyte proportions between host individuals and time points (Pearson’s chi-squared = 1.59, df = 2, P-value = 0.452, mean peak parasitemia = 6.74%, mean decreasing parasitemia = 10.24%). Furthermore, we directly compared the *P. ashfordi* transcripts showing significant expression differences between hosts (n = 28) to a list of sexual development genes exhibiting gametocyte-specific expression patterns in *P. falciparum* (n = 246) (Young *et al.* 2005). Only one of the 28 genes had a match in the sexual development gene set, namely the metabolite/drug transporter MFS1 (Table S3, Supporting information). However, this single match is not greater than expected by chance (hypergeometric test, P-value = 0.372), given probability of a match across all protein-coding genes in the *P. falciparum* genome (n = 5344). Together, these results show that the host-specific expression pattern in *P. ashfordi* cannot be explained due to differences in sexual development.

**Plasmodium ashfordi shows sequence similarities to primate malaria parasites**

Almost all annotated contigs (99.59%; n = 7828) resulted in a best blast hit against a species within the genus *Plasmodium* (Fig. 5A). The remaining contigs had matches against species within the genera of *Eimeria* (n = 12), *Cryptosporidium* (n = 6), *Neospora* (n = 5), *Babesia* (n = 4), *Hammondia* (n = 2), *Ascogregarina* (n = 1), *Theileria* (n = 1) and *Toxoplasma* (n = 1) (Table S6, Supporting information). The great majority (73.59%) of the contig blast matches were proteins originating from primate parasites, while 25.34% matched rodent parasites, and only 0.92% parasites of birds (Fig. 5B).

At the species level, most contigs (29.91%) resulted in best blast hit against *P. falciparum* (n = 12), *Cryptosporidium* (n = 6), *Neospora* (n = 5), *Babesia* (n = 4), *Hammondia* (n = 2), *Ascogregarina* (n = 1), *Theileria* (n = 1) and *Toxoplasma* (n = 1) (Table S6, Supporting information). The great majority (73.59%) of the contig blast matches were proteins originating from primate parasites, while 25.34% matched rodent parasites, and only 0.92% parasites of birds (Fig. 5B).

To compare across genes, expression levels have been normalized with respect to library size, regularized log-transformed, and scaled and centred around zero to give Z-scores.

© 2017 John Wiley & Sons Ltd
The contigs giving matches to avian Plasmodium were primarily derived from commonly sequenced apicomplexan genes and therefore available in public databases, for example cytochrome c oxidase subunit 1 (COX1; *P. lutzii*), merozoite surface protein 1 (MSP1; *P. relictum*), thrombospondin-related anonymous protein (TRAP; *P. relictum*) and cytochrome b (CYTB; *P. gallinaceum*) (Table S1, Supporting information).

The five contigs with highest GC content in the *P. ashfordi* transcriptome (47.7%–56.4%) all had matches against the avian parasites *Eimeria*, despite them only comprising 0.15% (*n* = 12) of the total annotation. *Eimeria* species have a very high transcriptome GC content (*E. acervulina*: 55.98%; *E. mitis*: 57.30%), and the *P. ashfordi* transcripts matching this genus consist predominantly of ribosomal and transporter genes (Table S1, Supporting information). The *P. ashfordi* contigs with highest expression levels were primarily annotated by uncharacterized protein matches to the rodent parasite *P. yoelii* (Table S7, Supporting information). In fact, the six most highly expressed transcripts that were annotated all gave significant blast matches to *P. yoelii*. Further investigation revealed that these transcripts are most likely derived from ribosomal RNA.

**Identification of conserved Plasmodium invasion genes**

Finally, to assess molecularly conserved strategies of *P. ashfordi* compared to mammalian malaria parasites,
we searched for annotated genes known to be involved in the red blood cell invasion by *Plasmodium* (Bozdech et al. 2003; Beeson et al. 2016). We discovered successfully assembled *P. ashfordi* transcripts from a whole suite of host cell invasion genes (Table 2). This includes, for example, the genes merozoite surface protein 1 (MSP1), apical membrane antigen 1 (AMA1), merozoite adhesive erythrocytic binding protein (MAEBL), GPI-anchored micronemal antigen (GAMA) and the rhoptry neck binding proteins 2, 4 and 5 (RON2, RON4 and RON5). Interestingly, the *P. ashfordi* RON genes in particular seemed to slightly decrease expression levels in all hosts over the two time points (Fig. 6). In general, however, the invasion genes showed a range of expression patterns over time, going in various directions (Fig. 6).

All genes known to be involved in the *Plasmodium* motor complex (Opitz & Soldati 2002; Baum et al. 2006), driving parasite gliding motion and enabling host cell invasion, were discovered in *P. ashfordi*. These include actin (ACT1), actin-like protein (ALP1), aldolase (FBPA), myosin A (MyoA), myosin A tail interacting protein (MTIP), gloidesome-associated proteins 45 and 50 (GAP45 and GAP50) and thrombospondin-related anonymous protein (TRAP) (Table 2). We also found the bromodomain protein 1 (BDP1), which has been directly linked to erythrocyte invasion by binding to chromatin at transcriptional start sites of invasion-related genes and controlling their expression (Josling et al. 2015).

We found two transcripts matching the low molecular weight rhoptry-associated proteins 1 and 3 (RAP1 and RAP3) that are secreted from the rhoptry organelles during cell invasion. The genomes of human malaria parasites contain a paralog gene called RAP2 as well, whereas rodent malaria parasites have a single gene copy that is a chimera of RAP2 and RAP3 (Counihan et al. 2013). The *P. ashfordi* transcript in question (TRI3305_c0_g1_i1) matches *P. falciparum* RAP3 better than the rodent parasite version of RAP2/3. The three high molecular weight rhoptry proteins (RhopH1, RhopH2, RhopH3) which bind to the erythrocyte plasma membrane and transfer to the parasitophorous vacuole membrane upon invasion (Vincensini et al. 2008; Counihan et al. 2013) were all identified in *P. ashfordi*. RhopH1 encompasses the multigene family of cytoadherence-linked asexual proteins (CLAGs), present in varying copy number across *Plasmodium*.

Other assembled *P. ashfordi* orthologs of genes involved in host cell invasion were the rhoptry-associated leucine zipper-like protein 1 (RALP1), rhoptry-associated membrane antigen (RAMA), armadillo-domain-containing rhoptry protein (ARO), RH5 interacting protein (RIPR), TRAP-like protein (TLP), merozoite TRAP-like protein (MTRAP), thrombospondin-related apical membrane protein (TRAMP), subtilisin proteases 1 and 2 (SUB1 and SUB2) and merozoite surface proteins 8 and 9 (MSP8 and MSP9). MTRAP and TRAMP are proteins that belong to the TRAP family and are released from the microneme organelles during invasion (Green et al. 2006; Cowman et al. 2012), and the subtilisin proteases SUB1 and SUB2 are heavily involved in the processing and cleavage of immature merozoite antigens, for example MSP1 and AMA1 (Beeson et al. 2016). ARO plays a crucial role in positioning the rhoptry organelles within the apical end of the parasite to enable the release of rhoptry-associated molecules (Mueller et al. 2013) such as RAMA and RALP1, which then bind to the erythrocyte surface.

We furthermore discovered transcripts of several reticulocyte-binding proteins (RBP/RH) thought to be absent in the genomes of avian malaria parasites (Lauren et al. 2015). These particular transcripts, together with RAMA, showed much higher e-values than other invasion genes (Table 2), indicating high differentiation between avian and mammalian *Plasmodium* RH genes. Finally, two rhomboid proteases (ROM1 and ROM4) have been linked to host cell invasion in *P. falciparum* via cleavage of transmembrane adhesins (Baker et al. 2006; Santos et al. 2012). We found both of these genes, together with other rhomboid proteases (ROM2, ROM3, ROM6, ROM8 and ROM10) expressed in *P. ashfordi*. More information about the assembled genes can be found in Table S1 (Supporting information).

**Discussion**

In this study, we assembled and characterized a blood-stage transcriptome with quantified gene expression of an avian malaria parasite, *P. ashfordi*. By developing a bioinformatic filtering method capable of dealing with dual RNA-seq data, we effectively removed contigs originating from the host and other sources of contamination in a multistep approach (Fig. 1). This resulted in a transcriptome with 7860 annotated transcripts and an additional 4094 unannotated transcripts. We discovered that 19 SNPs were not only present, but identical, in the transcriptomes of all six parasite samples from the three hosts, which corroborates that these are indeed true sequence polymorphisms and is an excellent independent verification that all host transcripts have been successfully removed from the assembly (Table S4, Supporting information). The gene expression of *P. ashfordi* displayed strikingly similar patterns during peak and decreasing infection stages and within individual hosts (Fig. 3). Furthermore, *P. ashfordi* shows most
sequence similarities to the human malaria parasite *P. falciparum* (Fig. SC), but specific genes involved in host interaction and defence seem to be highly differentiated between the two parasites (Fig. 2C). Nonetheless, the assembly supports several important erythrocyte invasion genes (Table 2), indicating evolutionary conserved cell invasion strategies across the phylogenetic host range of *Plasmodium* parasites.

© 2017 John Wiley & Sons Ltd

---

### Table 2 Assembled transcripts of genes involved in *Plasmodium* invasion of red blood cells

<table>
<thead>
<tr>
<th>Gene name</th>
<th>Gene product</th>
<th>Represented by transcripts</th>
<th>Species match (blastx)</th>
<th>Bit score</th>
<th>e-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACT1</td>
<td>Actin 1</td>
<td>TR215622</td>
<td>0</td>
<td>_g1</td>
<td>_i1</td>
</tr>
<tr>
<td>ALP1</td>
<td>Actin-like protein 1</td>
<td>TR55613</td>
<td>0</td>
<td>_g1</td>
<td>_i1</td>
</tr>
<tr>
<td>AMA-1</td>
<td>Apical membrane antigen 1</td>
<td>TR118224</td>
<td>0</td>
<td>_g1</td>
<td>_i1</td>
</tr>
<tr>
<td>ARO</td>
<td>Armadillo-domain-containing rhoptry protein</td>
<td>TR18124</td>
<td>0</td>
<td>_g1</td>
<td>_i1</td>
</tr>
<tr>
<td>BDP1</td>
<td>Bromodomain protein 1</td>
<td>TR66841</td>
<td>0</td>
<td>_g2</td>
<td>_i1</td>
</tr>
<tr>
<td>FBPA</td>
<td>Fructose-bisphosphate aldolase</td>
<td>TR12911</td>
<td>0</td>
<td>_g1</td>
<td>_i1</td>
</tr>
<tr>
<td>GAMA</td>
<td>GPI-anchored micronemal antigen</td>
<td>TR16322</td>
<td>0</td>
<td>_g2</td>
<td>_i1</td>
</tr>
<tr>
<td>GAP45</td>
<td>Glideosome-associated protein 45</td>
<td>TR34884</td>
<td>0</td>
<td>_g1</td>
<td>_i1</td>
</tr>
<tr>
<td>GAP50</td>
<td>Glideosome-associated protein 50</td>
<td>TR144721</td>
<td>0</td>
<td>_g1</td>
<td>_i1</td>
</tr>
<tr>
<td>MAEBL</td>
<td>Merozoite adhesive erythrocytic binding protein</td>
<td>TR234951</td>
<td>0</td>
<td>_g1</td>
<td>_i1</td>
</tr>
<tr>
<td>MCP1</td>
<td>Merozoite capping protein 1</td>
<td>TR198550</td>
<td>0</td>
<td>_g1</td>
<td>_i1</td>
</tr>
<tr>
<td>MSP1</td>
<td>Merozoite surface protein 1</td>
<td>TR112241</td>
<td>0</td>
<td>_g2</td>
<td>_i1</td>
</tr>
<tr>
<td>MTIP (MLC1)</td>
<td>Myosin A tail domain interacting protein</td>
<td>TR99718</td>
<td>0</td>
<td>_g1</td>
<td>_i1</td>
</tr>
<tr>
<td>MTRAP</td>
<td>Merozoite TRAP-like protein</td>
<td>TR215622</td>
<td>0</td>
<td>_g1</td>
<td>_i1</td>
</tr>
<tr>
<td>MYOA</td>
<td>Myosin A</td>
<td>TR188691</td>
<td>0</td>
<td>_g2</td>
<td>_i1</td>
</tr>
<tr>
<td>RAMA</td>
<td>Rhotry-associated membrane antigen</td>
<td>TR144799</td>
<td>0</td>
<td>_g1</td>
<td>_i1</td>
</tr>
<tr>
<td>RALP1</td>
<td>Rhotry-associated leucine zipper-like protein 1</td>
<td>TR7446</td>
<td>0</td>
<td>_g1</td>
<td>_i1</td>
</tr>
<tr>
<td>RAP1</td>
<td>Rhotry-associated protein 1</td>
<td>TR115690</td>
<td>2</td>
<td>_g1</td>
<td>_i1</td>
</tr>
<tr>
<td>RAP3</td>
<td>Rhotry-associated protein 3</td>
<td>TR13035</td>
<td>0</td>
<td>_g1</td>
<td>_i1</td>
</tr>
<tr>
<td>RBP1 (RH1)</td>
<td>Reticulocyte-binding protein 1</td>
<td>TR53756</td>
<td>1</td>
<td>_g1</td>
<td>_i1</td>
</tr>
<tr>
<td>RBP2 (RH2)</td>
<td>Reticulocyte-binding protein 2</td>
<td>TR208311</td>
<td>0</td>
<td>_g1</td>
<td>_i1</td>
</tr>
<tr>
<td>RBP2b (RH2b)</td>
<td>Reticulocyte-binding protein 2 homolog B</td>
<td>TR66282</td>
<td>1</td>
<td>_g2</td>
<td>_i1</td>
</tr>
<tr>
<td>RHOPH1</td>
<td>High molecular weight rhotry protein 1 (CLAG9)</td>
<td>TR7367</td>
<td>0</td>
<td>_g1</td>
<td>_i1</td>
</tr>
<tr>
<td>RHOPH2</td>
<td>High molecular weight rhotry protein 2</td>
<td>TR200056</td>
<td>0</td>
<td>_g1</td>
<td>_i1</td>
</tr>
<tr>
<td>RHOPH3</td>
<td>High molecular weight rhotry protein 3</td>
<td>TR233854</td>
<td>0</td>
<td>_g1</td>
<td>_i1</td>
</tr>
<tr>
<td>RIPR</td>
<td>RH5 interacting protein</td>
<td>TR258538</td>
<td>0</td>
<td>_g3</td>
<td>_i1</td>
</tr>
<tr>
<td>ROM1</td>
<td>Rhomboid protease 1</td>
<td>TR178976</td>
<td>0</td>
<td>_g1</td>
<td>_i1</td>
</tr>
<tr>
<td>ROM4</td>
<td>Rhomboid protease 4</td>
<td>TR92176</td>
<td>0</td>
<td>_g1</td>
<td>_i1</td>
</tr>
<tr>
<td>RON2</td>
<td>Rhotry neck protein 2</td>
<td>TR62544</td>
<td>0</td>
<td>_g1</td>
<td>_i1</td>
</tr>
<tr>
<td>RON4</td>
<td>Rhotry neck protein 4</td>
<td>TR145809</td>
<td>0</td>
<td>_g1</td>
<td>_i1</td>
</tr>
<tr>
<td>RON5</td>
<td>Rhotry neck protein 5</td>
<td>TR67313</td>
<td>2</td>
<td>_g1</td>
<td>_i1</td>
</tr>
<tr>
<td>SUB1</td>
<td>Subtilisin 1</td>
<td>TR16466</td>
<td>4</td>
<td>0</td>
<td>_g1</td>
</tr>
<tr>
<td>SUB2</td>
<td>Subtilisin 2</td>
<td>TR16922</td>
<td>0</td>
<td>_g1</td>
<td>_i1</td>
</tr>
<tr>
<td>TLP</td>
<td>TRAP-like protein</td>
<td>TR153597</td>
<td>0</td>
<td>_g1</td>
<td>_i1</td>
</tr>
<tr>
<td>TRAMP</td>
<td>Thrombospondin-related apical membrane protein</td>
<td>TR179194</td>
<td>0</td>
<td>_g1</td>
<td>_i1</td>
</tr>
<tr>
<td>TRAP</td>
<td>Thrombospondin-related anonymous protein</td>
<td>TR16495</td>
<td>0</td>
<td>_g1</td>
<td>_i1</td>
</tr>
</tbody>
</table>
P. ashfordi displays host-specific gene expression

Interestingly, and contrary to our expectations, P. ashfordi showed highly similar expression profiles inside the same host, despite being sampled 10 days apart during two different disease stages. All birds were inoculated with the same malaria strain derived from a single donor bird, and our SNP analysis verified that the genetic composition of P. ashfordi is identical in all hosts. The mechanism behind this host-specific expression pattern is unknown, but there is no reason why six independently sampled parasite populations should cluster transcriptionally based on host individual unless expression levels are somehow regulated in response to hosts. The expression pattern can potentially be caused by genotype-by-genotype interactions between the host and the parasite, modulation of parasite expression by the host, or plasticity of the parasite to different host environments. This result has potentially important implications for our understanding of the evolution of host–parasite interactions, and as a result warrants further research extending the limited sample size to more hosts and more time points throughout the infection.

Host genotype-by-parasite genotype interactions are complicated and not that well documented in malaria parasite systems. Studies with different genotypes of both host and parasite have found effects of host genotype, but not parasite genotype, on factors such as host resistance and parasite virulence (Mackinnon et al. 2002; de Roode et al. 2004; Grech et al. 2006; see also Idaghdour et al. 2012). Less is known about the transcriptome responses of malaria parasites to different host individuals. Some studies have found differential gene expression responses of Plasmodium to resistant vs. susceptible mice strains (see e.g., Lovegrove et al. 2006), and Daily et al. (2007) discovered host-specific distinct transcriptional states of P. falciparum in the blood of Senegalese children. However, a reanalysis of the data by Daily et al. (2007) suggested that differences in sexual development of the parasite may have contributed to the transcriptional states (Lemieux et al. 2009). With respect to our results, we found no evidence of differences in sexual development of the parasite and our SNP analysis showed that parasite haplotypes did not establish differentially in the host individuals. This suggests that sexual development and haplotype differences had little influence on our results and that the most likely explanation for the host-specific transcriptome patterns is plasticity in parasite gene expression.

The expression profiles of P. ashfordi did not exhibit any significant differences between peak and decreasing parasitemia stages (days 21 and 31 postinfection). The hosts in our study experienced relatively high parasitemia levels during the decreasing parasitemia stage as well (see Methods), so it is possible that these specific time points do not provide very different environmental host cues to the parasites. However, the concurrent transcriptomes of the avian hosts (analysed

![Individual gene expression plots for some of the Plasmodium ashfordi transcripts involved in red blood cell invasion. Line plots displaying normalized parasite gene expression in each individual host over the two sampled parasitemia stages. Host 2 is depicted in green, host 3 in orange and host 4 in purple. Thick horizontal lines indicate mean expression levels in each stage.](https://example.com/figure6.png)
in Videvall et al. 2015) displayed large differences in gene expression between these two parasitemia stages, notably with a reduced immune response during decreasing parasitemia. This is important because it appears that P. ashfordi does not adjust its gene expression in response to the decreasing immune levels of the hosts, but instead conforms to the specific environment of individual hosts.

Plasmodium falciparum evades the human immune defence via intracellularity, clonal antigenic variation (Guizetti & Scherf 2013), transcriptional antigenic switches (Recker et al. 2011), splenic avoidance by erythrocytic adherence to the endothelium (Craig & Scherf 2001) and sequestration in organ microvasculature (Sila-mut et al. 1999), erythrocytic rosetting (Niang et al. 2014), host immunosuppression (Hisaeda et al. 2004) and manipulation of host gene expression. It is possible that P. ashfordi shares several of these evasion strategies with P. falciparum, although this remains unknown. One example of immune evasion by manipulation of host gene expression is the parasite gene macrophage migration inhibitory factor homologue (MIF), which contributes to Plasmodium parasites’ ability to modulate the host immune response by molecular mimicry (Cordery et al. 2007). This gene was discovered transcribed in P. ashfordi as well (TR2046|c0_g1_i1), suggesting that a similar immune manipulation strategy is possible (Table S1, Supporting information).

Similarities to P. falciparum and other malaria parasites

The majority of all annotated contigs (73.59%) resulted in a best blast hit against primate parasites (Fig. 5B). Curiously, the human malaria parasite P. falciparum comprised the majority of all matches with almost a third of the transcripts (29.91%) (Fig. 5C). This is likely because P. falciparum currently constitutes the organism with most sequence similarities to P. ashfordi based on publically available sequences (Bensch et al. 2016). The chimpanzee parasite P. reichenowi had the second most blast matches to P. ashfordi (Fig. 5C), and it is the closest living relative to P. falciparum based on current genomic data (Otto et al. 2014b). Furthermore, both P. falciparum and P. ashfordi share the genome characteristics of being extremely AT-biased, with P. ashfordi reaching a remarkably low transcriptomic GC content of 21.22% (Table 1) compared to the already GC-poor P. falciparum (23.80%). Lastly, because of its role in human disease, P. falciparum is the most sequenced Plasmodium species (172 official strains in the NCBI taxonomy database as of May 2016) (Gardner et al. 2002), resulting in the greatest opportunity for transcript sequences to find significant blast matches.

Less than 1 per cent of all contigs resulted in a best hit against avian parasites (0.92%). This is due to the fact that almost no genomic resources are available for avian Plasmodium. Despite their enormous diversity, worldwide distribution and harmful effects on susceptible bird populations, genomic studies of avian malaria parasites have been largely nonexistent until now. The genome of P. gallinaceum, the malaria parasite of chickens, has been sequenced but not published and we were therefore not able to use it in our analyses. A transcriptome assembly of P. gallinaceum is available for download (Lauron et al. 2014, 2015), although still contains a large proportion of contigs matching birds, making comparisons with P. ashfordi difficult (see Fig. S4, Supporting information). Dual RNA sequencing of a more distantly related apicomplexan parasite, Leucocytozoon buteonis, and its buzzard host was recently described (Pauli et al. 2015), although no publically available transcriptome exists. Finally, both 454 RNA sequencing and Illumina genome sequencing of the generalist avian malaria parasite P. relictum (lineage SG5) have been performed (Hellgren et al. 2013; Bensch et al. 2014; Lutz et al. 2016), but the extremely low sequence coverage in both cases does unfortunately not allow for assembly nor any genomewide analyses. We hope that future sequencing of these avian parasites will enable genomeric comparisons.

The lack of genomewide sequence data from malaria parasites of hosts other than mice and primates means that little is known about which genes across Plasmodium are conserved and which that are unique. Our gene ontology results confirm that the transcriptomes of P. ashfordi and P. falciparum overall are functionally similar (Fig. 2A–B), although specific genes involved in host interaction and receptor binding could not be directly located in the P. ashfordi assembly (Fig. 2C).

The transcriptome of P. falciparum is certainly more complete because it is based on the genome sequence (Gardner et al. 2002) and therefore includes genes expressed from the entire life cycle. As a result, any P. falciparum genes not found in P. ashfordi, are either specifically transcribed during certain life stages (e.g. in the mosquito), not present in the genome, or too diverged to be detected with sequence similarity searches. Seeing how this particular group of genes involved in host interaction are under strong positive and diversifying selection pressure in Plasmodium (Hall et al. 2005; Jeffares et al. 2007; Otto et al. 2014b), it is likely that many are indeed present in the P. ashfordi assembly, but unannotated due to evolutionary divergence.

As a step to investigate genes involved in host interaction in more detail, we searched in the P. ashfordi assembly for genes specifically known to be involved in
the merozoite invasion of host red blood cells. Previously, only a handful of studies have sequenced candidate invasion genes in avian malaria parasites; these include MAEBL (Martinez et al. 2013), AMA1 and RON2 (Lauron et al. 2014), MSP1 (Hellgren et al. 2013, 2015), RIPR (Lauron et al. 2015) and TRAP (Templeton & Kaslow 1997; Farias et al. 2012). Due to the evolutionary distance between mammals and saurians, and their inherent blood cell differences (birds/reptiles have erythrocytes with a nucleus and mitochondria while mammalian cells are anucleated), we might expect to find few and highly differentiated gene orthologs. Instead, we discovered a large number of red blood cell invasion genes expressed in P. ashfordi (Table 2), indicating that most of these specific invasion genes are conserved across both mammalian and avian Plasmodium.

The invasion genes that were most differentiated between birds and mammals were the rhoptry-associated membrane antigen (RAMA) and the reticulocyte-binding proteins (RBP/RH), which had diverged almost beyond recognition. These RH genes, together with other erythrocyte binding antigens (EBA), have been assumed to be absent in the genomes of avian malaria parasites (Lauron et al. 2015). However, our result suggests that several are not only present, but also transcribed, although with high sequence divergence. It is possible that additional erythrocyte binding proteins are present in the P. ashfordi assembly, although ortholog searches for these genes will become complicated if they have evolved under especially strong selection pressure in avian Plasmodium.

Conclusion

In this study, we have de novo assembled, characterized and evaluated the blood transcriptome of the avian parasite Plasmodium ashfordi. By developing a rigorous bioinformatic multistep approach, the assembly was successfully cleaned of host sequences and contains high numbers of important genes in, for example, red blood cell invasion. We have shown that P. ashfordi displays similar expression profiles within individual hosts during two different stages of the infection but different expression patterns between individual hosts indicating possible host-specific parasite gene regulation. The expression information of all transcripts will assist researchers studying genes involved in, for example, immune evasion, host specificity and parasite plasticity. In addition, our results show that our isolate of P. ashfordi originally contained a minimum of two alleles at nine loci and has managed to maintain this low-level genomic heterozygosity throughout the infection in all host individuals. The results presented here and the associated assembly will help improve our understanding of host–parasite interactions, evolutionary conserved Plasmodium strategies and the phylogenetic relationships between apicomplexans.

Acknowledgements

We would like to thank Staffan Bensch for stimulating discussions and comments on this manuscript. We are also grateful to Brian Haas for advice on transcriptome assemblies, and the director of the Biological Station ‘Rybachy’, Casimir V. Bolkhakov, for generously providing facilities for the experimental research. Two anonymous reviewers provided valuable feedback which significantly improved the manuscript. The assembly and blastn computations were performed on resources provided by SNIC through Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) (Lampa et al. 2013) under project b2014134; all other computational analyses were performed by EV on a local machine.

Funding

This work was supported by the Swedish Research Council (grant 621-2011-3548 to O.H. and 2010-5641 to C.K.C.), the Crafoord Foundation (grant 20120630 to O.H.), European Social Fund under the global grant measure (grant VPI-3.1.-ŠMM-07-K-01-047 to G.V.), Research Council of Lithuania (grant MIP-045/2015 to G.V.) and a Wallenberg Academy Fellowship to C.K.C.

References

Bensch S, Pérez-Tris J, Waldenström J, Hellgren O (2004) Linkage between nuclear and mitochondrial DNA sequences in...


Koutsovoulos G, Kumar S, Laetsch DR *et al.* (2016) No evidence for extensive horizontal gene transfer in the genome


**Data accessibility**

The supplementary tables and figures supporting this article have been uploaded as part of the online supporting information. The sequence reads of both host and parasite have been deposited at the NCBI Sequence Read Archive (SRA) under the accession no. PRJNA311546. The assembled *P. ashfordi* transcriptome is available for download as Appendix S1 and Appendix S2 (Supporting information), or alternatively at http://mbio-serv2.mbioekol.lu.se/Malavi/Downloads.

The study design was initially conceived by O.H., V.P. and G.V. and further developed together with E.V. and C.K.C. The infection experiment was planned by G.V. and V.P. and performed by V.P. O.H. performed the RNA extractions. E.V. performed the assembly and all bioinformatic and statistical analyses. D.A. advised in the trimming of the contigs and in the mapping of sequence reads. O.H., C.K.C. and E.V. planned the study. E.V. wrote the manuscript with extensive input from all authors.

**Supporting information**

Additional supporting information may be found in the online version of this article.

**Fig. S1** Length distribution of contigs in the *Plasmodium ashfordi* transcriptome assembly.

**Fig. S2** Multidensity plot showing density of transcripts over log-transformed normalized expression values.

**Fig. S3** Dot plot showing allele frequencies of 19 SNPs in the parasites of three hosts during peak parasitemia.

**Fig. S4** GC content distribution of a transcriptome assembly of the chicken parasite *Plasmodium gallinaceum* (Lauron et al. 2015) highlights the difficulties in assembling clean parasite transcriptomes from dual RNA-seq data.

**Fig. S5** Example of transcript sequence homogeneity from the *P. ashfordi* transcript TR213315(c0_g1_i1) derived from the mitochondrial gene cytochrome c oxidase subunit 1 (COX1).

**Fig. S6** Example of transcript sequence heterogeneity showing the presence of three SNPs in the *P. ashfordi* transcript TR73260(c0_g1_i1).

**Table S1** Information of all annotated *P. ashfordi* transcripts (n = 7860).

**Table S2** Normalized expression levels of all *P. ashfordi* transcripts (n = 11 954) in individual hosts during peak and decreasing parasitemia stages.

**Table S3** *P. ashfordi* transcripts that were significantly differentially expressed between host individuals (n = 28).

**Table S4** Results of SNPs discovered in the *P. ashfordi* transcriptome (n = 19).
**Table S5** Gametocyte and meront proportions in the *P. ashfordi* blood smears.

**Table S6** Species distribution matches from the annotated transcripts of *P. ashfordi*.

**Table S7** Most highly expressed transcripts in the *P. ashfordi* transcriptome.

**Table S8** Gene ontology terms with associated numbers of *P. ashfordi* and *P. falciparum* transcripts in the category ‘Biological process’.

**Table S9** Gene ontology terms with associated numbers of *P. ashfordi* and *P. falciparum* transcripts in the category ‘Molecular function’.

**Table S10** Gene ontology terms with associated numbers of *P. ashfordi* and *P. falciparum* transcripts in the category ‘Multorganism process’.

**Table S11** Gene ontology terms with associated numbers of *P. ashfordi* and *P. falciparum* transcripts in the category ‘Catalytic activity’.

**Table S12** Gene ontology terms with associated numbers of *P. ashfordi* and *P. falciparum* transcripts in the category ‘Metabolic process’.

**Table S13** Gene ontology terms with associated numbers of *P. ashfordi* and *P. falciparum* transcripts in the category ‘Cellular process’.

**Table S14** Gene ontology terms with associated numbers of *P. ashfordi* and *P. falciparum* transcripts in the category ‘Binding’.

**Appendix S1** Sequences in the annotated *P. ashfordi* transcriptome assembly (n = 7860) as a gzipped multi-fasta file.

**Appendix S2** Sequences in the total *P. ashfordi* transcriptome assembly (n = 11 954) as a gzipped multi-fasta file.