<table>
<thead>
<tr>
<th>Title</th>
<th>Integrated analysis of the Plasmodium species transcriptome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author(s)</td>
<td>Hoo, Regina; Zhu, Lei; Amaladoss, Anburaj; Mok, Sachel; Natalang, Onguma; Lapp, Stacey A.; Hu, Guangan; Liew, Kingsley; Galinski, Mary R.; Bozdech, Zbynek; Preiser, Peter Rainer</td>
</tr>
<tr>
<td>Date</td>
<td>2016</td>
</tr>
<tr>
<td>URL</td>
<td><a href="http://hdl.handle.net/10220/46472">http://hdl.handle.net/10220/46472</a></td>
</tr>
<tr>
<td>Rights</td>
<td>© 2016 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<a href="http://creativecommons.org/licenses/by-nc-nd/4.0/">http://creativecommons.org/licenses/by-nc-nd/4.0/</a>).</td>
</tr>
</tbody>
</table>
Research Paper

Integrated analysis of the Plasmodium species transcriptome

Regina Hoo a, Lei Zhu a, Anburaj Amaladoss a, Sachel Mok a, Anguma NataIang a, Stacey A. Lapp b, Guangan Hu a, Kingsley Liew a, Mary R. Galinski b,c, Zbynek Bozdech a, Peter R. Preiser a,*

a School of Biological Sciences, Nanyang Technological University, Singapore
b Emory Vaccine Center, Yerkes National Primate Research Center, Emory University, Atlanta, GA, USA
c Department of Medicine, Division of Infectious Diseases, Emory University, Atlanta, GA, USA

A R T I C L E   I N F O

Article history:
Received 2 February 2016
Received in revised form 9 April 2016
Accepted 11 April 2016
Available online 22 April 2016

Keywords:
Comparative transcriptomics
Plasmodium species
Evolution
Microarray
Transcriptome
Drug targets

A B S T R A C T

The genome sequence available for different Plasmodium species is a valuable resource for understanding malaria parasite biology. However, comparative genomics on its own cannot fully explain all the species-specific differences which suggests that other genomic aspects such as regulation of gene expression play an important role in defining species-specific characteristics. Here, we developed a comprehensive approach to measure transcriptional changes of the evolutionary conserved syntenic orthologs during the intraerythrocytic developmental cycle across six Plasmodium species. We show significant transcriptional constraint at the mid-developmental stage of Plasmodium species while the earliest stages of parasite development display the greatest transcriptional variation associated with critical functional processes. Modeling of the evolutionary relationship based on changes in transcriptional profile reveal a phylogeny pattern of the Plasmodium species that strictly follows its mammalian hosts. In addition, the work shows that transcriptional conserved orthologs represent potential future targets for anti-malaria intervention as they would be expected to carry key essential functions within the parasites. This work provides an integrated analysis of orthologous transcriptome, which aims to provide insights into the Plasmodium evolution thereby establishing a framework to explore complex pathways and drug discovery in Plasmodium species with broad host range.

© 2016 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Protozoan belonging to the Plasmodium species are obligate intracellular parasites that display substantial developmental complexity during their life cycle in the vertebrate hosts and mosquito vectors. The development of single nucleated parasite cells into multi-nucleated schizonts through several rounds of mitosis closely resembles embryonic development of multicellular organisms with a majority of genes changing their expression during this period (Piras et al., 2014; Quint et al., 2012; Bozdech et al., 2003; Irie and Kuratani, 2011). The intraerythrocytic developmental cycle (IDC) exhibits a tightly regulated transcriptional cascade in which essentially every gene in the genome is targeted to a specific stage of the parasite development. This precise control of gene expression ultimately governs critical functional processes for Plasmodium species to thrive within the host erythrocytes (Bozdech et al., 2003; Le Roch et al., 2003). We have previously reported that orthologous expression between two human Plasmodium species, Plasmodium falciparum and Plasmodium vivax, showed transcriptional divergence in 30% of the genes, some of which could be linked to known functional differences between these two species (Bozdech et al., 2008). This suggested that while there is high conservation of the overall pattern of genome activity during the IDC of the two Plasmodium species, there is transcriptional variation of a subset of genes that enable the parasites to adapt to the individual host niches presumably defined by variations in nutrients, metabolites, as well as other cellular components (Kafsack and Llinas, 2010; Srivastava et al., 2015). However, no studies so far have addressed the extent of transcriptional diversity of Plasmodium species from distinct host erythrocytes. Here, we developed a comprehensive analysis of the IDC transcriptional profiles; from uniform data processing to extensive orthology annotation, which allow us to directly compare variations in mRNA abundance across six different Plasmodium species and significantly extends previous work using pairwise comparison (Bozdech et al., 2008). The Plasmodium conserved syntenic orthologs show significant transcriptional divergence at the earliest stages of parasite development, with transcriptional phylogeny pattern that strictly follows the Plasmodium species.
species mammalian hosts. Furthermore, changes in the expression of key putative transcriptional regulators are implicated in the transcriptional diversity. Critically the work also provides the tools for the identification of new and so far uncharacterized drug targets, as the orthologs that show a conserved transcription pattern across the species are likely to carry out critical conserved functions in all Plasmodium species.

2. Materials and methods

2.1. Sample collection for microarray analysis

2.1.1. Rodent malaria parasites

All studies involving mice were approved by the institutional animal care and use committee (IACUC) of the Nanyang Technological University, Singapore. Male BALB/c mice 6–7 weeks old, bred specific pathogen free (SPF) at the Nanyang Technological University Animal Resource Facility, were infected with either cryopreserved stocks of parasites or by syringe passage from a pre-existing infected mouse. Mice were infected by intraperitoneal injection of Plasmodium berghei ANKA, Plasmodium chabaudi AS or Plasmodium yoelii 17× parasitized erythrocytes and parasitaemia and parasite stages were monitored by thin blood smears stained with Giemsa. For P. berghei and P. yoelii infection, mice were terminal bled and the stage-specific parasitized erythrocytes were separated via Nycodenz density gradient. The ring stage interface was isolated, washed and subjected to ex vivo culture, which was then collected every 2 hr over the course of 24 hr for a complete IDC cycle. Mice infected with P. chabaudi were terminal bled every 2 hr under anesthesia over the course of 24 hr. Blood was collected and filtered through Plasmipur filters (Eurodiagnostica, Netherlands) to remove white blood cells and then washed with PBS. The washed blood was flash-frozen in liquid nitrogen and stored at −80 °C until further use.

2.2. RNA extraction, cDNA preparation and DNA microarray hybridization

Total RNA was isolated using a standard protocol using trizol/chloroform extraction as described by (Bozdech et al., 2003). For preparation of the target DNA for microarray hybridization, Switch Mechanism was synthesized at the 5’ end of Reverse Transcription (SMART) PCR approach was employed (Petalidis et al., 2003). Thereafter, cDNA was synthesized by using the reverse transcriptase (PowerScript, 2.2. RNA extraction, cDNA preparation and DNA microarray hybridization

2.3. Microarray data processing and analysis

2.3.2. Normalize and data filtering

All microarray hybridization spots obtained from all six Plasmodium species were subjected to “normexp” background correction followed by LOWESS (locally weighted scatterplot smoothing) normalization within each array and quantile normalization between arrays using Limma package of R. Log2 ratios of Cy5 over Cy3 intensities were calculated for each spot to represent expression value of a particular probe except those with signal intensity <1.5 times the background intensity for both Cy5 and Cy3 fluorescence. For each gene, the expression value was estimated as the average of all probes representing it. Overall, 4750 (90% of genes designed on the microarray) P. falciparum genes, 4670 (99%) P. knowlesi genes, 4884 (97%) P. vivax genes, 5486 (88%) P. yoelii genes, 3990 (89%) P. chabaudi genes and 3787 (86%) P. berghei genes display expression profiles with one or zero missing value across each IDC life cycle. These “processed” microarray expression dataset was used for subsequent analysis. The raw and processed microarray data for P. yoelii, P. berghei and P. chabaudi have been deposited into Gene Expression Omnibus (GEO accession number: GSE80015).

2.3.3. Phaseogram

The expression profile of each gene is modeled using sine function which has been described in detail (Lapp et al., 2015). Briefly, the formula is E(t) = A × sin(ω×t−α) + C. Where E(t) is the log2 ratio sample/reference control at the t time point of sample collection, A is the amplitude of expression profile across life cycle, C is the vertical offset of profile from zero, ω is the angular frequency given by 2π/h and h is the length of a complete IDC duration expressed in hours (details see below), and α is the horizontal offset of profile from zero. α was projected into an interval ranging from 0 to 2π with, π/2 of α representing the peak expression at early ring stage matching to 0. The converted α was subsequently used as phaseogram (Ph) to indicate the IDC timing where gene expression peaks. Ph for each gene was sorted from early to late IDC for complete visualization of genes expressed in the complete IDC.

2.3.4. Estimation of IDC duration and adjustment of phaseogram

The timing for complete IDC varies between the six Plasmodium species and the time points collected for each sample were different. Therefore, h, which is the length of complete IDC duration was optimized and projected within 44 hr to 54 hr for P. falciparum, P. vivax and 20 hr to 36 hr for P. knowlesi and the other three rodent strains to determine the best fit sine function model (Lapp et al., 2015). As a result, the optimized h is 48 hr for P. falciparum, 49 hr for P. vivax, 29 hr for P. knowlesi, 27 hr for P. yoelii, 28 hr for P. chabaudi genes and 29 hr P. berghei. To minimize the effect of asynchronous parasites between species samples, the phaseogram was adjusted for P. vivax, P. knowlesi, P. yoelii, P. chabaudi and P. berghei to P. falciparum reference. For example, for P. vivax phaseogram adjustment, the best matching time points with P. falciparum for each P. vivax time point was estimated by the highest Spearman Rank Correlation Coefficient (SRCC) values between global transcription profiles of syntenic orthologs. Next, the average shift of time using those best matching time point pairs with SRCC > 0.2 was calculated, which resulted in 25 time points for each gene per species on a sine wave model.
2.3.5. Delta phase, $\Delta$Ph

The dissimilarity of transcription profiles between two genes is given by:

$$\Delta \text{Ph}(a, b) = \min(\text{Ph}_a - \text{Ph}_b, 2\pi - |\text{Ph}_a - \text{Ph}_b|), a \neq b$$

Where $\Delta \text{Ph}$ is the distance of Ph expression timing between gene $a$ and gene $b$, $\text{Ph}_a$ and $\text{Ph}_b$ is the expression timing of gene $a$ and $b$ respectively.

2.3.6. Medoid gene and Ds value

The metric of transcription divergence among multiple genes, given by Ds, is the average $\Delta \text{Ph}$ of each gene to the medoid gene. K-medoids clustering methods was applied, where K equal to 1, and took the genes in question as a cluster. The medoid gene of this cluster was determined based on the dissimilarity matrix, which contained $\Delta \text{Ph}$ of all gene pairs. Ds value is the average dissimilarity of this cluster based on its medoid. Calculation was conducted using the package 'cluster' of R.

2.3.7. Detection of outlier gene and species

For a syntenic orthologs, outlier of one species is defined as the ortholog gene with the most divergent expression timing or Ph compare to the other species member within an ortholog group. To detect the outlier for each ortholog group, we construct the dissimilarity matrix of species based on $\Delta \text{Ph}$. Outlier is define as the ortholog gene of a

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Summary table of number of genes that are syntenic orthologs and with specific transcriptional profile across all six Plasmodium species.</th>
</tr>
</thead>
</table>
| Microarray data analysis | \begin{tabular}{|l|l|l|l|l|l|} \hline
| Pf | Pk | Pv | Py | Pb | Pc | \hline
| After data normalization and filtering (related to Fig. 1) | 4750 & 4670 & 4884 & 5486 & 3787 & 3990 | \hline
| Syntenic orthologs and species-species specific genes | \hline
| Syntenic orthologs present in at least one or more species (OrthoMCL DB) | 4148 & 4286 & 4275 & 3577 & 3708 & 3783 | \hline
| Syntenic orthologs present in all six species (OrthoMCL DB) | 3374 & 3374 & 3374 & 3374 & 3374 & 3374 | \hline
| Syntenic orthologs present in at least one or more species with specific transcriptional profile | 3175 & 3182 & 3190 & 3161 & 2951 & 3025 | \hline
| Syntenic orthologs present in all six species with specific transcriptional profile | 2312 & 2312 & 2312 & 2312 & 2312 & 2312 | \hline
| Species-specific genes with specific transcriptional profile (non-orthologs, non-syntenic) | 84 & 245 & 457 & 1436 & 29 & 136 | \hline
\end{tabular} |
particular species which maximally contributes to the sum of $\Delta \Phi$ of that dissimilarity matrix.

2.3.8. Dendrogram of transcriptome relationship

Dendrogram was constructed by applying Ward hierarchical clustering method based on the dissimilarity matrix containing the distance of each pair of species, $\Delta \Phi$. The distance of two species is defined as;

$$D(A, B) = \sqrt{ \sum_{i=0}^{n} \Delta \Phi_i(A, B)^2 }, A \neq B$$

where $D(A,B)$ is the distance of species $A$ and $B$, $n$ is the total number of syntenic orthologous gene which is equal to 2312 and $\Delta \Phi_i$ is the distance of expression timing between the $i$th gene of species $A$ and its orthologous gene in species $B$. The dendrogram cluster was subjected to 100 times bootstrapping to estimate the percentage of approximately unbiased (AU), $p$-value (in red) and bootstrap probability (BP) (in green).

2.3.9. $K_a/K_s$ ratio

$K_a/K_s$ ratio is the ratio of the number of nonsynonymous substitutions per non-synonymous site ($K_a$) to the number of synonymous substitutions per synonymous site ($K_s$). Selective pressure of protein-coding genes between six Plasmodium species was estimated by calculating $K_a/K_s$ ratio for each pair of syntenic orthologs for a total of 2312 orthologs. The syntenic orthologs were aligned using ClustalW (Larkin et al., 2007) and $K_a/K_s$ were calculated using package ‘seqinr’ of R (Charif and Lobry, 2007). Dendrogram indicating evolution relationship between species was constructed based on the dissimilarity matrix in which the distance of two species was defined by the mode of $K_a$ values.

2.3.10. Functional enrichment analysis

Functional enrichment analysis was carried out by identifying syntenic orthologs sets that are significantly over-represented in studied genes list by comparing to annotated gene sets from P. falciparum MPM database. Hypergeometric test was applied to calculate the level of significance of indicated orthologous gene sets from the MPM database.
3. Results

3.1. Lack of conservation of specific ortholog expression in Plasmodium species’ life cycle

Despite the long evolutionary time scale of about 100 million years (Escalante and Ayala, 1995; Escalante and Ayala, 1994), the overall genome organization and content across the sequenced *Plasmodium* species is highly conserved, with about 4000 conserved syntenic genes located within the central core regions of the 14 chromosomes (Hall et al., 2002; Carlton et al., 2008a; Pain et al., 2008; Gardner et al., 2002; Hall et al., 2005). On the other hand, species-specific differences are attributed to the variant multigene families that reside predominantly at the subtelomeric regions. In this study, we generated the IDC transcriptomes for the rodent malaria parasites *P. berghei*, *P. yoelii* and *P. chabaudi* and compared these with the previously generated data for the human malaria species *P. falciparum* (Foth et al., 2011), *P. vivax* (Bozdech et al., 2008), and the simian malaria species *P. knowlesi* (Lapp et al., 2015) (Supplementary Table 1, Supplementary Fig. 1a).

The IDC transcriptional profiles of each gene were first best fitted onto a sine-wave function as described in the experimental procedure section (Fig. 1a, Supplementary Fig. 1b). This allowed us to define several parameters that reflect the temporal aspects of gene expression and thus make direct species-to-species comparisons. The quality of the newly generated data was further verified by comparisons with the recently published *P. berghei* RNA-seq data showing high levels of correlation (Supplementary Fig. 1c) (Otto et al., 2014). Overall, 3787 to 5486 genes (including species-specific genes) exhibited specific IDC transcriptional profiles in the six *Plasmodium* species (Table 1). In contrast, between 29 and 1436 genes are unique to individual species with specific transcriptional profiles (Table 1, Supplementary Fig. 1d), where the majority of them are genes restricted to subtelomeric ends and are involved in antigenic variation, host-parasite interactions, cytoadherence and erythrocyte aggregation for all six species, similar to previous genomic studies (Kooij et al., 2005; Kyes et al., 2001; Dzikowski et al., 2006; Galinski and Corredor, 2004). Interestingly, while the distribution of peak time expression remains relatively constant for the syntenic orthologs, the species-specific genes show remarkable variability (Supplementary Fig. 1d).

Previous studies using Pearson correlation coefficient (PCC) to compare *P. falciparum* to *P. vivax* had indicated that PCC between 1 and 0.5 represent strongly correlated pattern, 0.5 to −0.2 represent a moderate timing shift in gene expression, while PCC < −0.2 indicate dramatic changes of gene expression along the IDC (Bozdech et al., 2008) (Supplementary Fig. 2). Using the same approach, we compared all orthologous gene pairs with *P. falciparum* as the reference (PF-Pk, PF-Pc, PF-Pb and PF-Py) and showed high correlations between 50 and 65%, intermediate to low correlations for 18% to 23% and anti-correlation for 18% to 26% (Fig. 1b). *P. yoelii* appears most divergent from *P. falciparum* with 725 (23.1%) and 815 (25.9%) gene pairs with...
low and anti-correlations, respectively (Fig. 1b). The overall distribution of highly correlated orthologous genes (PCC 1–0.50) showed only 474 genes conserved among the six species (Fig. 1c). Even when using any group of five out of the six species the total number of conserved genes would increase by only 104 to 341 genes (Fig. 1c). This suggests that only a relatively small subset of genes remained transcriptionally

![Diagram](https://via.placeholder.com/150)

**Summary of the Malaria Parasite Metabolic Pathways (MPM)**

<table>
<thead>
<tr>
<th>Pathway Categories</th>
<th>PF</th>
<th>PK</th>
<th>PV</th>
<th>PY</th>
<th>PB</th>
<th>PC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intracellular traffic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transcription: biogenesis of RNA</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mitochondrial functions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Translation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Replication</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Histones and their modifications</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transcription: RNA modification and degradation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Post-translational: single amino acid metabolisms</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chromosome structure: mitosis and chromosome separation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Post-translational: ubiquitin-dependent processes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amino acid metabolism</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cofactors &amp; other substances</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lipid metabolism</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Post-translational: chaperones &amp; protein structure modifications</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nucleotide metabolism</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Motility</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other functional processes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Post-translational: other protein modifications</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protein export</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transport through nuclear pore</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Redox metabolism</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Post-translational: single amino acid modification</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Carbohydrates: glycolysis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other organelles (acidocalcisome)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The parasite cell membrane: transporters of the plasma membrane</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
conserved throughout Plasmodium evolution while the majority underwent (some level of) transcriptional diversification. The results also indicate a relatively limited constraint on transcriptional timing across the syntenic orthologs providing a considerable flexibility for species-specific adaptation among Plasmodium species.

3.2. Pattern of evolution in Plasmodium transcriptional regulation

To compare the temporal character of expression between orthologous genes, we define the time-shift (delta phase, ΔPh) relationship between the ortholog pairs derived from their IDC gene expression profiles. ΔPh between the species’ orthologs for each gene expression profile was first projected onto a polar coordinate system with timing of gene expression being depicted by values ranging from 0 to 2π (Supplementary Fig. 3). Phase adjustment, Ph was performed by offsetting the peak of gene expression timing to the peak gene expression time point at the early ring-stage of development (Bozdech et al., 2008) (Supplementary Table 2). This method takes advantage of a uniform periodic character of the gene expression profiles that allows much more efficient and representative normalization of gene expression profiles compared to PCC-based similarity or Euclidean distance tests generally used to study other organisms (Pereira et al., 2009; Glazko and Mushegian, 2010). In total, we established ΔPh for the 2312 syntenic ortholog groups with IDC expression datasets for all six Plasmodium species. K-means clustering for the ΔPh values revealed eight distinct clusters that represent the overall evolutionary pattern of transcriptional regulation in Plasmodium (Fig. 2a). The eight clusters consist of gene groups that are (i) fully conserved in all six species; (ii) show no similarity in gene expression in any species pair; (iii) genes with diverse expression in one of the species while similar in the other five. In particular, (i) there are 818 genes whose transcriptional timing have been maintained throughout evolution (Fig. 2a), representing 35.4% of the 2312 syntenic orthologs. This transcriptionally conservation suggests a crucial role of these genes in Plasmodium biology. On the other hand, (ii) there are 309 genes that are completely devoid of any transcriptional conservation among the Plasmodium species. Although the role(s) of these genes is generally unknown, this large diversity suggests their involvement in highly dynamic evolutionary processes in Plasmodium. Finally, (iii) there are six clusters ranging from 136 to 357 genes, in which one Plasmodium species shows a diversion from the other five (Fig. 2a). To further clarify these findings we calculated the overall transcriptional divergence (Ds score) among orthologs as the average ΔPh between each species gene to the medoid species gene (Fig. 2b). When including all 6 species, about 50% of the orthogroups have a Ds6 value of <0.75 (conserved transcription). This can be significantly increased to approximately 80% if one (Ds5) or two (Ds4) outlier gene(s) with the most divergent Ph are removed (Fig. 2b, Supplementary Table 2). This supports the model that the majority of transcriptional divergence occurred in individual single species diverging gene expression from a putative ancestral transcriptional profile. Finally, the ΔPh distance measure separates the Plasmodium species into two distinct clades that delineates precisely along the mammalian host species (Fig. 2c). This suggests that transcriptome-based phylogeny is shaped by the adaptation of the different Plasmodium species in their respective mammalian host erythrocytes; the human P. falciparum and P. vivax, the rhesus macaque P. knowlesi and the rodent Plasmodium species.

3.3. Sequence and transcription-based evolution in Plasmodium species

In Plasmodium, diversity of genomic sequence is considered to be a major factor of speciation, in particular involving species-specific gene families encoding of factors of host-parasite interaction (Freh and Chen, 2011). In the next step we wished to evaluate the evolutionary role of transcription in the context of genome sequence divergence. For that, we first estimated the number of non-synonymous substitutions per non-synonymous site (Ks) and the number of synonymous substitutions per synonymous site (Ks) for each species pair (Fig. 3a). The distribution of Ks/Ks ratio reflects the level of selection pressure on protein-coding genes throughout the evolution of Plasmodium parasites species (Fig. 3a), with the rodent malaria parasites showing a greater difference in the change of protein-coding genes within the species group. One of the speculated factors of the rodent malaria parasites evolution is the geographical aspect of adaptation to their hosts that subsequently led to significant changes in the coding regions within the rodent species compared to their primate counterparts. Accordingly, the overall transcriptome in the rodent Plasmodium species exhibits a greater divergence compared to the primate-infecting species (Fig. 3b). This suggests broader diversification of the rodent-infecting species driven by both sequence diversity and transcriptional divergence. We next assembled a dendrogram based on non-synonymous Ks ratio, which reflects the phylogenetic relationships among the Plasmodium species based on the sequence homology. Here, P. vivax and P. knowlesi are more closely related than P. falciparum and P. berghei, which forms a separate cluster together with P. yoelii and slightly more distant P. chabaudi (Fig. 3c). This is consistent with the previously constructed phylogenetic topology based on partial mitochondrial genomes (Escalante et al., 1998; Carlton et al., 2008b) as well as genes encoding surface antigens (Weedall et al., 2008). The sequence-based phylogenetic tree(s) are different from the transcriptional differences that is reflective of its mammalian host lineage (as mentioned above) (Fig. 2c). The sequence-based relationships seem to be driven by other (than host) factors such as AT content of the genome (with P. falciparum having a much higher AT content than P. knowlesi and P. vivax) and possibly reticulocyte preferences (Craig et al., 2012).

3.4. Evolutionary divergence and conservation of the Plasmodium transcriptome occurs at specific IDC timing and within selected functional pathways

We examined IDC progression to evaluate whether particular stages are more prone to transcriptional conservation/diversity. We analyzed the Ds6 values as a function of gene expression timing, where Ds6 were equally divided into low [0–π/6, medium [π/6–π/3] and high values [π/3–2π/6]) and segregated equally into six IDC unit time points from 0 to 2π (in a range of π/3) according to the peak of gene expression corresponding to the individual IDC stages (Fig. 4a). Genes whose expression peaks at the early ring stage for any of the Plasmodium species showed significantly higher Ds (P < 0.001). This implies that corresponding orthologs are likely to be expressed at different IDC stages in other species (Fig. 4a). In contrast, the late ring and early trophozoite stages express genes with overall low Ds (P > 0.001), which suggest that in other species these genes are expressed at a similar IDC stage. Genes whose expression peaks at the late trophozoite, early schizont and late schizont stages showed variable enrichment in conservation and
divergence across the species (Fig. 4a). For instance, genes expressed at the late schizont stage are only significantly enriched with low Ds in *P. falciparum* (*P* < 0.01), *P. knowlesi* (*P* < 0.01), *P. vivax* (*P* < 0.05), and *P. berghei* (*P* < 0.001). The early developmental stage (rings) exhibit the highest diversity, the mid stage (late ring trophozoite) appear to be the most constrained and the late stage (schizonts) are characterized by intermediate levels of diversion, demonstrating that transcriptional conservation of *Plasmodium* genes is not uniform throughout the IDC. To investigate the biological significance of transcriptional conservation in *Plasmodium*, we carried out pathway enrichment analyses of the orthologs with an “outlier” in one *Plasmodium* species; where the gene is expressed at a different set of genes in the otherwise overlapping pathways. It is feasible to speculate that these genes represent rate limiting steps of these biological processes that facilitate temporal shifts of their overall function. On the other hand there are several pathways with significantly diversified orthologs that are exclusive to one species including protein export (*P. chabaudi*), redox metabolism (*P. yoelii*), glycocysis and protein transport (*P. knowlesi*) and post-translational modifications (*P. vivax*) (Fig. 4b). This exclusivity may indicate the importance of these pathways for the evolutionary adaptation.

### 3.5. Transcription conservation and divergence of the AP2 transcriptional regulators and its cis-regulatory region

To better understand the possible mechanism(s) that drive transcriptional divergence, we analyzed the transcriptional profile of the 26 putative and known AP2 transcriptional regulators in all six species (Painter et al., 2011). Of these 26 AP2 genes, only 19 are present as syntenic orthologs in all six species. Four AP2 orthologs were subjected to preliminary analyses of the orthologs present in all 6 *Plasmodium* species. (a) Heatmap shows ΔPh profile for 13 AP2 syntenic orthologs present in all 6 *Plasmodium* species. Statistical significance of differences was measured between the ΔPh of the highly divergent and conserve AP2 orthologs expression, highlighted with red box, using one-way ANOVA analysis (*P* < 0.05). (b) Barplot and scatterplot (median value indicated by black line) show Ds6 and Ph respectively for the 13 AP2 syntenic orthologs. Statistical significance of differences was measured between Ds6 of the highly divergent and conserve AP2 orthologs expression using nonparametric t-test (**P < 0.01**). (c) Alignment plot of coding region and -1000 bp upstream region from the ATG start site of four AP2 with the most conserved and diverged gene expression using mLAGAN and wAlignAce. Cutoff of >50% similarities are highlighted in pink. AP2 orthologs are represented by *P. falciparum* gene ID on the top-left of each alignment plot.
bioinformatics analysis on the conservation of cis-regulatory regions within the promoters of AP2 orthologs in different Plasmodium species. Phylogenetic footprint analysis on the AP2 promoters was performed to estimate the percentage of best-matching regions within the 1000 bp nucleotide frame (Janky and van Helden, 2008) (Fig. 5c). Using P. falciparum as the reference, the alignment analysis revealed variable patterns and percentages of sequence conservation upstream of the ATG start sites (Fig. 5c). Genes with greater transcriptional conservation across species (PFF0550w and PF11_0404) tend to have greater conservation at the upstream non-coding sequence, while the transcriptionally diverged genes (PF10_0075 and PF14_0271) showed a low level of conservation (Fig. 5c). In particular, PFF0550w orthologs displayed a clear conserved, interspecies “footprint” at the promoter intergenic region, about −500 and −100 bp upstream of the ATG start site. This region may contain putative conserved regulatory elements for transcriptional regulation (Chen and Jiang, 2006). In contrast limited conservation of regulatory sequences indicates transcriptional divergence across species.

3.6. Druggability prediction of the conserved ortholog transcriptome

Our analysis thus far implies that the different Plasmodium species have evolved to conserve a subset of transcriptional responses with analogous cellular functions and pathways, suggesting a possible selective advantage required for the parasites’ development and survival within the host cell. Considering the relatively low number of genes that are transcriptionally conserved across Plasmodium species where about 818 orthologs are transcriptionally conserved (Fig. 2), those that

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Gene product</th>
<th>Gene tmd count</th>
<th>Gene signal peptide</th>
<th>Druggability evidence index</th>
</tr>
</thead>
<tbody>
<tr>
<td>PF10_0084</td>
<td>tubulin beta chain, putative</td>
<td>0</td>
<td>N</td>
<td>1</td>
</tr>
<tr>
<td>PFC0525c</td>
<td>glycogen synthase kinase 5</td>
<td>0</td>
<td>N</td>
<td>0.9</td>
</tr>
<tr>
<td>MAL13P1.279</td>
<td>protein kinase 5</td>
<td>0</td>
<td>N</td>
<td>0.9</td>
</tr>
<tr>
<td>PFE1050w</td>
<td>adenosylhomocysteinase(5-adenosyl-L-homocysteine hydrolase)</td>
<td>0</td>
<td>N</td>
<td>0.8</td>
</tr>
<tr>
<td>PFO380c</td>
<td>formylmethionine deformylase, putative</td>
<td>1</td>
<td>Y</td>
<td>0.8</td>
</tr>
<tr>
<td>PFL1370w</td>
<td>NIMA-related protein kinase, Pfenk-1</td>
<td>0</td>
<td>N</td>
<td>0.7</td>
</tr>
<tr>
<td>PFL2275c</td>
<td>FK506-binding protein (FKBP)-type peptidyl-propyl isomerase</td>
<td>0</td>
<td>N</td>
<td>0.6</td>
</tr>
<tr>
<td>PF14_0641</td>
<td>1-deoxy-D-xylulose 5-phosphate reductoisomerase</td>
<td>0</td>
<td>N</td>
<td>0.6</td>
</tr>
<tr>
<td>PF10_0086</td>
<td>adenylate kinase</td>
<td>0</td>
<td>N</td>
<td>0.6</td>
</tr>
<tr>
<td>PF11_0282</td>
<td>deoxyuridine 5'-triphosphate nucleotidohydrolase, putative</td>
<td>0</td>
<td>N</td>
<td>0.6</td>
</tr>
<tr>
<td>PF0925w</td>
<td>gamma-glutamylcysteine synthetase</td>
<td>0</td>
<td>N</td>
<td>0.6</td>
</tr>
<tr>
<td>PF10_0123</td>
<td>GMP synthetase</td>
<td>0</td>
<td>N</td>
<td>0.6</td>
</tr>
<tr>
<td>PF11_0301</td>
<td>spermidine synthase</td>
<td>0</td>
<td>N</td>
<td>0.6</td>
</tr>
</tbody>
</table>

Fig. 6. Druggability prediction from the conserved ortholog transcriptome. (a) Proportion of transcriptionally conserved orthologs with positive druggability index score and essentiality data from RMgmDB database. (b) Overall frequency distribution of peak in IDC expression timing (phase) and (c) metabolic pathway enrichment analysis (PlasmoDB) of 240 genes that are both transcriptionally conserved and with positive druggability evidence index score from TDR database.
show a conserved transcription pattern would be expected to carry out critical time-dependent functions within the parasite and host cells. Drug target prioritization databases from online resources such as TDRv5 (Magarinos et al., 2012; Aguero et al., 2008) short-listed ~30% (240 genes) of the transcriptionally conserved orthologs are potentially druggable with druggability index score (DIS) > 0, where almost one-third (91 genes) was not assigned to any putative function (Fig. 6a, Supplementary Table 4). 17 of the transcriptionally conserved, potentially druggable orthologs including a group of kinases and proteases were found to be essential in P. bergheri from the rodent orthologs database (RMgmDB) (Janse et al., 2011) (Fig. 6a, Supplementary Table 4). The majority of these potentially druggable orthologs are expressed from trophozoite to schizont stages in all six species (Fig. 6b). Metabolic pathway enrichment assessments suggest that these orthologs are involved in nucleotide metabolism, energy dependent processes, tRNA biosynthesis and carbohydrate metabolism (Fig. 6c) supporting a conserved expression for selected functional pathways, which may be relevant for core essential processes across species. Furthermore, 13 of the transcriptionally conserved, druggable orthologs have been characterized and pharmacologically validated as antimalarial drug targets in blood stage P. falciparum and P. vivax orthologs according to TDR databases (Table 2). This includes the PfKFBP3 petidyl-propyl isomerase encoded by PFL2275c (Harikishore et al., 2013), PFK5 or cyclin dependent protein kinase 5 encoded by MAL13P1.279 (Harmsse et al., 2001), tubulin beta chain encoded by PF10_0084 (Fennell et al., 2006), and glycogen synthase kinase 3 encoded by PFC0525c (Drouiche et al., 2004), all which inhibition of function affected the cellular division of the parasites. Moreover, the conserved blood stage-specific expression of this particular group of orthologs suggests there is potential cross-species anti-malarial targets, in particular against P. vivax and P. falciparum both, the prominent etiological agent of human malaria.

4. Discussion

Comparative analysis of large-scale genome wide expression datasets across multiple species, especially time-series expression datasets, was often complicated by several parameters such as oligonucleotide probe affinity, background noises and heterogeneity of samples from different organisms (Kuo et al., 2006). In this study, microarray analysis was performed on Plasmodium oligonucleotides and pan-rodent oligonucleotides designed by OligoRankPick as previously described (Hu et al., 2007; Liew et al., 2010). Taking into account that majority of the genes in P. falciparum, P. vivax and P. knowlesi are expressed in a periodic manner during the IDC stage (Bodzech et al., 2003; Bodzech et al., 2008; Lapp et al., 2015), including the rodent malaria parasite genes, the raw microarray data for each gene from all six species were best-fitted into a sine-wave phaseogram. Phaseogram correction was performed to adjust for the variation in IDC timing among the different Plasmodium species as described in detail elsewhere (Lapp et al., 2015). Nevertheless, the phaseogram correction has its own caveats; not all of the Plasmodium genes will display a distinct periodic expression manner throughout the IDC, particular for very lowly expressed genes, which may exclude their consideration in the analysis of overall transcriptional divergence. Future work using RNA-seq analysis could more effectively assess the contribution that lowly expressed genes make to the overall transcriptional variation. It is interesting to note that for two of the species the IDC microarray data (P. bergheri and P. vivax) is highly correlated with the periodic life-cycle transcriptional profiles from RNA-seq studies of the corresponding species (Otto et al., 2014; Zhu et al., 2016), suggesting that a large proportion of the overall gene expression detected by microarray are credible estimates and that only a small fraction of the transcripts are of such low abundance as not to be accurately detected in the microarray. Second, variation in synchrony of the parasites cultures, specifically for the rodent malaria parasites, could potentially contribute to the overall transcriptional variation observed in this study. To minimize such variation, we sampled the Nycozenid ring stage parasitized rodent erythrocytes from different mice infected with either P. bergheri or P. yoelii from ex vivo culture, which were then collected for every 2 hr until late schizont stage. A similar ex vivo study was performed for the transcriptome study of P. knowlesi ex vivo culture (Lapp et al., 2015). Furthermore, the complete IDC timing for each species was also adjusted to minimize synchrony issues and sampling time variation, which resulted in 25 time points for each gene. The rigorous effort to minimize these potential limitations, makes us confident that our analysis provides a framework for further hypothesis testing across different malaria parasites.

Multi-gene comparisons provided a robust evolutionary topology of the malaria parasite species by dividing them into several distinct monophyletic clades; the human parasite P. falciparum, a separate clade containing both P. vivax and P. knowlesi, and the rodent parasites clade (Martinsen et al., 2008). In this study, we show that Plasmodium species’ phylogeny based on transcriptional divergence contradicts the established sequence based evolutionary model (Perkins and Schall, 2002; Martinsen et al., 2008), suggesting that transcriptional and sequence evolution in Plasmodium species are under different selection pressures. Phylogenetic evolution across species can occur at many different levels; from the changes in coding sequences to protein structure and to alteration in the levels and timing of gene and protein expression (Harrison et al., 2012; Carroll, 2008). Moreover, orthologous expression studies from other higher eukaryotes such as humans and mice have highlighted minor selective constraints in the evolution of both species (Su et al., 2002; Yanai et al., 2004). One crucial finding of this study was the observation that particularly the early ring stage, appears to be under the greatest transcriptional divergence. This suggests that host specific variations that the parasite encounters within a newly infected erythrocyte rather than variation of host cell receptors is a key driver of evolution. The high level of transcriptional divergence in the early ring stage, as opposed to other stages of the parasite’s development, is intriguing and similar to several other ontology studies based on other model organisms such as Drosophila, zebrafish and Arabidopsis, where transcriptional conservation is the highest during the mid-embryonic stage of development (Kalinka et al., 2010; Domazet-Loso and Tautz, 2010; Quint et al., 2012; Irie and Kuratani, 2011). The Plasmodium parasites, like higher multicellular organisms, may undergo initial adaptation responses to the cellular environment through genetic control, in this case within the host erythrocytes at the very early stage of their development.

Our data suggests that rather than coding sequence differences other fundamental features such as transcription factors, upstream promoter regions, or post-transcriptional control such as epigenetic changes, chromatin remodeling events or noncoding RNAs may play a larger part in the modulation of transcription across the Plasmodium species (Coulsion et al., 2004; Milletello et al., 2004; van Noort and Huyten, 2006; Templeton et al., 2004; Gupta et al., 2013; Vembar et al., 2014; Ay et al., 2015). The ApiAP2 family of proteins is the largest putative transcriptional regulators identified in Plasmodium species (Balaji et al., 2005) and has been described as the master regulator in both the sexual and asexual stages (Sinha et al., 2014; Kafoe et al., 2014). Our findings here are consistent with changes in AP2 expression driving broader transcriptional changes in the different species. Although the evolutionary role(s) of cis-regulatory elements among the Plasmodium species are relatively unknown as they are not homologous to any known eukaryotic regulatory elements, the significance of variation in cis-regulatory sequences on transcription factor binding and divergent gene expression towards species phenotypic evolution has been largely described in higher eukaryotes (Gompel et al., 2005; Shirangi et al., 2009; Deplancke et al., 2006). The variation in putative cis-regulatory elements in the AP2 promoter regions of different Plasmodium species with transcriptional divergence is consistent with recent findings in P. falciparum supporting the role of cis-trans regulatory control events (Russell et al., 2015).
With mounting reports of anti-malarial drug resistance, there is an urgent need to identify new drug targets in the malaria parasite that will lead to the development of new therapeutic compounds to complement the existing repertoire of available drugs. The utility of cross-species transcription factors for assessment of potential drug targets against common biological processes is recently gaining attention (Okyere et al., 2014; Foth et al., 2014). The conserved orthologous transcriptome of Plasmodium species provides abundant molecular targets, including both annotated and hypothetical gene targets, for further functional studies and target-based approaches to drug discovery. We propose that the transcriptionally conserved orthologs carry out crucial time-dependent functions within parasite development that have yet to be studied and at the same time also represent possible anti-malarial targets for parasite intervention. The common biological processes with conserved IDT transcriptomes are likely to be most critically functional for the parasites, as the different species have strictly maintained these gene expression patterns over the course of evolution. By combining the comparative transcriptome data with web-based resources, we have short-listed a consensus group of potential anti-malarial targets that increases the likelihood of identifying critical functions in parasite biology, and eventually potential anti-malarial drug targets. Ultimately, the comparative transcriptome analysis across multiple species provides important tools and resources for other post-genomic studies and continued assessments of the evolution of Plasmodium parasites.

Author contributions

R.H. and L.Z. analyzed and interpreted the data. L.Z. developed computational algorithms. A.A., O.N., S.M., S.A.L., G.H., and K.L. performed the microarray experiments. M.R.G. and Z.B. reviewed the manuscript critically and provided intellectual content. R.H. and P.R.P. wrote the manuscript with contributions from all authors.

Disclosure

The author(s) declare(s) that there is no conflict of interests.

Acknowledgements and funding sources

This work was supported by the National Medical Research Council (Singapore) grant number NMRC/1292/2011, the US National Institute of Health/National Institute of Allergy and Infectious Diseases, Grant numbers 1R01AI24710 and R01-A065961, and the Yerkes National Primate Research Center Base Grant (ORIP/OD P51OD11132) awarded by the National Center for Research Resources of the National Institutes of Health.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.jebiomed.2016.04.011.

References


