Integrated transcriptomic, proteomic and epigenomic analysis of *Plasmodium vivax* salivary-gland sporozoites.

Vivax Sporozoite Consortium* (Ivo Muller[1,2,3], Aaron R. Jex[1,3,4], Stefan H. I. Kappe[5], Sebastian A. Mikolajczak[5], Jetsumon Sattabongkot[7], Rapatbhorn Patrapuvich[6], Scott Lindner[8], Erika L. Flannery[5], Cristian Koepfli[1], Brendan Ansell[4], Anita Lerch[1], Kristian Swearingen[5], Robert Moritz[9], Michaela Petter[10], Michael Duffy[10], Vorada Chuenchob[5]).

*Group authorship – all others are equal contributors (order per author contributions section below).

1. Population Health and Immunity Division, The Walter and Eliza Hall Institute for Medical Research, 1G Royal Parade, Parkville, Victoria, 3052, Australia.
2. Malaria: Parasites & Hosts Unit, Institut Pasteur, 28 Rue de Dr. Roux, 75015, Paris, France.
3. Department of Medical Biology, The University of Melbourne, Victoria, 3010, Australia.
4. Faculty of Veterinary and Agricultural Sciences, The University of Melbourne, Corner of Park and Flemington Road, Parkville, Victoria, 3010, Australia.
5. Center for Infectious Disease Research, 307 Westlake Avenue North, Suite 500, Seattle, WA 98109, USA;
6. Department of Global Health, University of Washington, Seattle, WA 98195, USA.
7. Mahidol Vivax Research Center, Faculty of Tropical Medicine, Mahidol University, Bangkok 10400, Thailand.
8. Department of Biochemistry and Molecular Biology, Center for Malaria Research, Pennsylvania State University, University Park, PA 16802, USA.
9. Institute for Systems Biology, Seattle, WA, 98109, USA.
10. Department of Medicine Royal Melbourne Hospital, The Peter Doherty Institute, The University of Melbourne, 792 Elizabeth Street, Melbourne, Victoria 3000, Australia.

Abstract

Background: *Plasmodium vivax* is the key obstacle to malaria elimination in Asia and Latin America, largely attributed to its ability to form resilient 'hypnozoites' (sleeper-cells) in the host liver that escape treatment and cause relapsing infections. The decision to form hypnozoite is made early in the liver infection and may already be set in sporozoites prior to invasion. To better understand these early stages of infection, and the potential mechanisms through which the development may be pre-programmed, we undertook a comprehensive transcriptomic, proteomic and histone epigenetic characterization of *P. vivax* sporozoites.

Results: Our study highlights the loading of the salivary-gland sporozoite with proteins required for cell traversal and invasion and transcripts for infection of and development within hepatocytes. We characterise histone epigenetic modifications in the *P. vivax* sporozoite and explore their role in regulating transcription. This work shows a close correlation between H3K9ac marks and transcriptional activity, with H3K4me3 and H3K9me3 appearing to act as general markers of euchromatin and heterochromatin respectively. We also identify the remarkable transcriptional silence in the (sub)telomeres and discuss potential roles of AP2 transcription factors, specifically ApiAP2-SP and L in regulating this stage.

Conclusions: Collectively, these data indicate the sporozoite as a tightly programmed stage primed to infect the human host and identifies key targets to be further explored in liver stage models.

Background

Malaria is among the most significant infectious diseases impacting humans globally, with 3.3 billion people at risk of infection, 381 million suspected clinical cases and up to ~660,000 deaths attributed to malaria globally in 2014 [1]. Two major parasite species contribute to the vast majority of human malaria, *Plasmodium falciparum* and *P. vivax*. Historically, *P. falciparum* has attracted the majority of global attention, due to its higher contribution to morbidity and mortality. However, *P. vivax* is broadly distributed, more pathogenic than previously thought, and is recognised as the key obstacle to malaria elimination in the Asia-Pacific and Americas [2]. Unlike *P. falciparum*, *P. vivax* can establish long-lasting 'sleeper-cells' (= hypnozoites) in the host liver that emerge weeks, months or years after the primary infection (= relapsing malaria) [3]. Primaquine is the only approved drug that prevents

relapse. However, the short half-life, long dosage regimens and incompatibility of primaquine with glucose-6-phosphate-dehydrogenase deficiency (which requires pre-screening of recipients [4]) makes it unsuitable for widespread use. As a consequence, *P. vivax* is overtaking *P. falciparum* as the primary cause of malaria in a number of co-endemic regions [5]. Developing new tools to diagnose, treat and/or prevent hypnozoite infections is considered one of the highest priorities in the malaria elimination research agenda [6].

When *Plasmodium* sporozoites are deposited by an infected mosquito, they likely traverse the skin cells, enter the blood-stream and are trafficked to the host liver, as has been shown in rodent malaria parasites [7]. Upon reaching the liver, sporozoites traverse Kupffer and endothelial cells to reach the parenchyma, moving through several hepatocytes before invading a final hepatocyte suitable for liver stage development [7, 8]. Within hepatocytes, these parasites replicate, and undergo further development and differentiation to produce tens of thousands of merozoites that emerge from the liver and infect red blood cells. However, *P. vivax* sporozoites are able to commit to two distinct developmental fates within the hepatocyte: they either immediately continue development as replicating schizonts and establish a blood infection, or delay replication and persist as hypnozoites. Regulation of this major development fate decision is not understood and this represents a key gap in current knowledge of *P. vivax* biology and control.

The sporozoites' journey from skin deposition to hepatocytes takes less than a few minutes [9]. It has been hypothesized that *P. vivax* sporozoites exist within an inoculum as replicating 'tachysporozoites' and relapsing 'bradysporozoites [10] and that these subpopulations may have distinct a developmental fate as schizont or hypnozoites, thus contributing to their relapse phenotype [10-12]. This observation is supported by the stability of different hypnozoite phenotypes in *P. vivax* infections of liver-chimeric mouse models [13]. Sporozoites prepare for mammalian host infection while still residing in the mosquito salivary glands. Studies using rodent malaria parasites have identified genes [14], that are transcribed in sporozoites but translationally repressed (i.e., present as transcript but un- or under-represented as protein), via RNA-binding proteins [15], and ready for just-in-time translation after the parasites infection of the mammalian host [13, 16]. Translational repression (i.e., the blocking of translation of present and retained transcripts) and other mechanisms of epigenetic control may contribute to the *P. vivax* sporozoite fate decision and hypnozoite formation, persistence and activation. Supporting this hypothesis, histone methyltransferase inhibitors stimulate increased activation of *Plasmodium cynomolgi* hypnozoites in macaque hepatocytes [17, 18]. Epigenetic control of stage development is further evidenced in *Plasmodium* through chromatin structure controlling expression of PfAP2-G, a specific transcription factor that, in turn, regulates gametocyte (dimorphic sexual stages) development in blood-stages [19]. It is well documented that *P. vivax* hypnozoite activation patterns stratify with climate and geography [11] and recent modelling suggests transmission potential selects for hypnozoite phenotype [20]. Clearly the ability for *P. vivax* to dynamically regulate hypnozoite formation and relapse phenotypes in response to high or low transmission periods in different climate conditions would confer a significant evolutionary advantage.

Unfortunately, despite recent advances [21] current approaches for *in vitro P. vivax* culture do not support routine maintenance in the laboratory and tools to directly perturb gene function are not established. This renders studies on *P. vivax*, particularly its sporozoites and liver stages, exceedingly difficult. Although *in-vitro* liver stage assays and humanised mouse models are being developed [13], they cannot yet support 'omics analysis of *P. vivax* liver stage dormancy. Recent characterization [22] of liver-stage (hypnozoites and schizonts) of *P. cynomolgi* (a related and relapsing parasite in macaques) provides valuable insight, but investigations in *P. vivax* directly are clearly needed. The systems analysis of *P. vivax* sporozoites that reside in the mosquito salivary glands and are poised for transmission and liver infection offer a key opportunity to gain insight into *P. vivax* infection. To date, such characterization of *Plasmodium vivax* sporozoites is limited [23], and only one recent study, of *P. falciparum* [24], has undertaken exploration of epigenetic regulation in sporozoites of any *Plasmodium* species. Here, we present a detailed characterization of the *P. vivax*

sporozoite transcriptome, proteome and epigenome and use these data to better understand this key infective stage and the role of sporozoite programming in invasion and infection of the human host, and development within the host liver.

Results and Discussion

We quantified transcript abundance for 5,714 *P. vivax* genes (4,991 with a mean transcript per million (TPM) count ≥ 1.0) at a mean estimated abundance of 175.1 TPM (Additional File 1: Figure S1 and Additional File 2: Table S1) for *P. vivax* sporozoites isolated from *Anopheles dirus* salivary glands using the recently completed *P. vivax* P01 assembly and gene models (see methods). For ease of reference, where one-to-one orthologs are established between the P01 and previous *P. vivax* (Sal1) reference, we use the Sal1 gene names in text (both the P01 and Sal1 gene names are provided for all genes in the supplementary information). Mosquito infections were generated by membrane feeding of blood samples taken from *P. vivax* infected patients in western Thailand (n = 9). Among the most highly transcribed genes in the infectious sporozoite stage are *csp* (circumsporozoite protein), five *etramps* (early transcribed membrane proteins), including *uis3* (up-regulated in infective sporozoites), *uis4* and *lsap-1* (liver stage associated protein 1), a variety of genes involved in cell transversal and initiation of invasion, including *celtos* (cell traversal protein for ookinetes and sporozoites), *gest* (gamete egress and sporozoite traversal protein), *spect1* (sporozoite protein essential for cell traversal) and *siap-1* (sporozoite invasion associated protein), and genes associated with translational repression (*alba1*, *alba4* and *Puf2*). Collectively, these genes account for >1/3rd of all transcription in the sporozoite. We found moderate agreement (R^2 = 0.35; Additional File 1: Figure S2) between our RNA-seq data and previous microarray data for *P. vivax* sporozoites [23]. Improved transcript detection and quantitation is expected with the improved technical resolution of RNA-seq over microarray. Supporting this, we find higher correlation between RNA-seq data from *P. vivax* and *P. falciparum* (single replicate sequenced herein for comparative purposes) sporozoite datasets (R^2 = 0.42), compared to either species relative to published microarray data (Additional File 1: Figure S2). Although microarray supports the high transcription in sporozoites of genes such as *uis4*, *csp*, *celtos* and several other *etramps*, 27% and 16% of the most abundant 1% of transcribed genes in our sporozoite RNA-seq data are absent from the top decile or quartile respectively in the existing *P. vivax* sporozoite microarray data [23]. Among these are genes involved in early invasion/hepatocyte development, such as *lsap-1*, *celtos*, *gest* and *siap-1*, or translational repression (e.g., *alba-1* and *alba-4*); orthologs of these genes are also in the top percentile of transcripts in RNA-seq (see [24] and Additional File 2: Table S2) and (see [25] and Additional File 2: Table S3) and previous microarray data [26, 27] for *P. falciparum* and *P. yoelii* sporozoites respectively, suggesting many are indeed more abundant than previously characterized.

Transcription in P. vivax *relative to other plasmodia*

To gain insight into species-specific aspects of the *P. vivax* transcriptome, we qualitatively compared these data with available data from *P. falciparum* and *P. yoelii* sporozoites (single replicate only) for 4,220 and 4,067 single-copy orthologs (SCO) (transcribed at ≥ 1 TPM in *P. vivax* infectious sporozoites) shared with *P. falciparum* (Additional File 2: Table S3) and with both *P. falciparum* and *P. yoelli* (Additional File 2: Table S4) respectively. Genes highly transcribed in salivary-gland sporozoites of all three species include *celtos*, *gest*, *trap*, *siap1*, *spect1* and *puf2*. There are 696 *P. vivax* genes shared as orthologs between *P. vivax* P01 and *P. vivax* Sal1 lacking a defined SCO in *P. falciparum* or *P. yoelli* transcribed at a mean of ≥ 1 TPM in *P. vivax* salivary-gland sporozoites (Additional File 2: Table S5). Prominent among these are *vir* (n=25) and *Pv-fam* (41 fam-e, 16 fam-b, 14 fam-a, 8 fam-d and 3 fam-h) genes, as well as, hypothetical proteins or proteins of unknown function (n=212) and, interestingly, a number of 'merozoite surface protein' 3 and 7 homologs (n=5 of each). Both *msp3* and *msp7* have undergone significant expansion in *P. vivax* relative to *P. falciparum* and *P. yoelii* [28] and may have repurposed functions in sporozoites. In addition, there are 69 *P. vivax* P01

genes lacking a defined ortholog in *P. vivax* Sal1, *P. falciparum* or *P. yoelli* transcribed at ≥ 1 TPM in infectious *P. vivax* sporozoites; most of which are *Plasmodium* interspersed repeat (PIR) genes [28] found in telomeric regions of the P01 assembly and likely absent from the Sal1 assembly but present in the Sal1 genome.

P. vivax sporozoite transcriptional enrichment

To comprehensively identify sporozoite enriched transcripts, we compared the *P. vivax* sporozoite transcriptome (Additional File 2: Table S6) to RNA-seq data for *P. vivax* blood-stages [29] (the only other RNA-seq data presently available for *P. vivax*; Fig. 1 and Additional File 1: Figures S3-5). We identified 1,672 up (Additional File 2: Table S7) and 1,958 down-regulated (Additional File 2: Table S8) transcripts (FDR ≤ 0.05; minimum 2-fold change in Counts per Million (CPM)) and next explored patterns among these differentially transcribed genes (DTGs) by protein family (Fig. 1C and Additional File 2: Table S9) and Gene Ontology (GO) classifications (Additional File 2: Table S10). RNA recognition motifs (RRM-1 and RRM-6) and helicase domains (Helicase-C and DEAD box helicases) are over-represented (p-value <0.05) among sporozoite-enriched transcripts, consistent with translational repression through ribonucleoprotein (RNP) granules [30]. Transcripts encoding nucleic acid binding domains, such as bromodomains (PF00439; which can also bind lysine-acetylated proteins), zinc fingers (PF13923) and EF hand domains (PF13499) are also enriched in sporozoites. Included among these proteins are a putative ApiAP2 transcription factor (PVX_083040) and a homologue of the *Drosophila* zinc-binding protein 'Yippee' (PVX_099695). Thrombospondin-1 like repeats (TSR: PF00090) and von Willebrand factor type A domains (PF00092) are enriched in sporozoites as well. In sporozoites, *P. falciparum* genes enriched in TSR domains are important in invasion of the mosquito salivary gland (e.g., *trap*) and secretory vesicles released by sporozoites upon entering the vertebrate host (e.g., *csp*) [31]. By comparison, genes up-regulated in blood-stages are enriched for *vir* gene domains (PF09687 and PF05796), Tryptophan-Threonine-rich *Plasmodium* antigens (PF12319; which are associated with merozoites [32]), markers of cell-division (PF02493),[33] protein production/degradation (PF00112, PF10584, PF00152, PF09688 and PF00227) and ATP metabolism (PF08238 and PF12774). 47 of the 343 transcripts unique to *P. vivax* sporozoites relative to *P. falciparum* or *P. yoelii* are enriched in sporozoites compared to *P. vivax* blood stages. Nine of these are in the top decile of transcription, and include a Pv-fam-e (PVX_089880), a Pf-fam-b homolog (PVX_001710) and 7 proteins of unknown function. A further nine have an ortholog in *P. cynomolgi* (which also forms hypnozoites) but not the closely related *P. knowlesi* (which does not form hypnozoites) and include '*msp7*'-like (PVX_082685, PVX_082650 and PVX_082670) and '*msp3*'-like (PVX_097705) and Pv-fam-e genes (PVX_001100, PVX_089860 and PVX_089810), a serine-threonine protein kinase (PVX_081395) and a RecQ1 helicase homolog (PVX_099345). Notably, the *P. cynomolgi* ortholog of PVX_081395, PCYB_021650, is transcriptionally enriched in hypnozoites relative to replicating schizonts [22], indicating a target of significant interest when considering hypnozoite formation and/or biology.

Translational repression machinery

In *Plasmodium*, translational repression regulates key life-cycle transitions coinciding with switching between the mosquito and the mammalian host (either as sporozoites or gametocytes) [30]. For example, although *uis4* is the most abundant transcript in the infectious sporozoite ([23, 27]; Additional File 2: Table S1), UIS4 is translationally repressed in this stage [15] and only expressed after hepatocyte invasion [34]. In sporozoites, it is thought that PUF2 binds to mRNA transcripts and prevents their translation [25], and SAP1 stabilises the repressed transcripts and prevents their degradation [34]. Consistent with this, *Puf2* and *SAP1* are among the more abundant *P. vivax* transcripts enriched in the sporozoite relative to blood-stages. Indeed, *Puf2* is among the top percentile of transcripts in infectious sporozoites. However, our data implicate other genes, many already known to be involved in translational repression in other *Plasmodium* stages and other protists [30], that may act in *P. vivax* sporozoites. Notable among these are *alba-2* and *alba-4*, both of which are among the

top 2% of genes transcribed in sporozoites and ~14 to 20-fold more highly transcribed in sporozoites relative to blood-stages. In addition, *P. vivax* sporozoites are enriched for genes encoding RRM-6 RNA helicase domains. Intriguing among these genes are HoMu (homolog of Musashi) and ptbp (polypyrimidine tract binding protein). Musashi is a master regulator of eukaryotic stem cell differentiation through translational repression [35] and HoMu localizes with DOZI and CITH in *Plasmodium* gametocytes [36]. PTBP is linked to mRNA stability, splice regulation and translational initiation [37] and may perform a complementary role to SAP1.

Translational repression in *P. vivax* sporozoites

More than 700 genes have been identified as being translationally repressed in *Plasmodium berghei* ('rodent malaria') gametocytes based on DOZI pulldowns [38]. In contrast, translationally repressed genes have not been characterized in sporozoites in a comprehensive way. As a step in addressing this, we analysed the *P. vivax* sporozoite proteome (Additional File 1: Figure S6 and Additional File 2: Table S11) by mass spectrometry and identified peptide signals for 2,640 proteins. Among the most highly expressed proteins in sporozoites were those associated with the apical complex (AMA1, GAMA, RON12, RON3, RON5), motility / cell traversal (MYOSIN A, PLP1, TRAP, SIAP1, GEST, SPECT1, CELTOS) and the inner membrane complex (ISP1/3, IMC1a, e, g, h, m and k), which has a key role in motility and invasion [39]. We identified 2,402 *P. vivax* genes transcribed in the sporozoite (TPKM > 1) for which no protein expression was detected. In considering genes that may be translational repressed (i.e., transcribed but not translated) in the *P. vivax* sporozoite, we confine our observations to those transcripts representing the top decile of transcript abundance to ensure their lack of detection as proteins was not due to limitations in the detection sensitivity of the proteomic dataset. Notably, ~1/3rd of transcripts in the top decile of transcriptional abundance (n = 173 of 558) in *P. vivax* sporozoites were not detectable as peptides in multiple replicates (Additional File 2: Table S12). Of these 173 putatively repressed transcripts, 156 and 154 have orthologs in *P. falciparum* and *P. yoelii* respectively, with 89 and 118 of these also not detected as proteins in *P. falciparum* and *P. yoelii* salivary-gland sporozoites [40] despite being identified as transcribed in these stages (see [24, 25]; Additional File 2: Tables S2-4). In addition, a number of genes (e.g., *uis4*) are expressed in infectious *P. vivax* sporozoites at levels many fold lower than their transcription might indicate (bottom quartile of protein expression, compared with top decile of transcript abundance). While each putatively repressed transcript will require validation, this system level approach is supported by immunofluorescent imaging (Additional File 1: Figure S7) of UIS4 and LISP1 (one known and one proposed here as translationally repressed in *P. vivax* sporozoites) relative to TRAP and BiP (which are both transcribed and expressed as protein in the *P. vivax* sporozoite; Additional File 2: Table S12).

Development within the host hepatocyte

Following cell traversal and hepatocyte invasion, *P. vivax* sporozoites establish their intracellular niche, which includes modification of the parasitophorous vacuole membrane (PVM) and the parasite then proceeds to replicate as a liver stage. UIS3 and UIS4 are resident PVM-proteins and are the best characterized proteins under translational repression by Puf2/SAP1 in infectious sporozoites [41], both of which are essential for liver stage development [14]. In the present study, *uis4* represents 18.8% of transcripts but just 0.06% of proteins in the sporozoites. Similarly, *uis3* is the 7th most abundant transcript in sporozoites, but represented only by a single peptide count in one proteomic replicate. In addition to *uis3* and *uis4*, genes involved in liver stage development and under apparent translational repression in the *P. vivax* sporozoites include *lsap1* (liver stage associated protein 1), *zipco* (ZIP domain-containing protein), several other *etramps* (PVX_118680, PVX_003565, PVX_088870 and PVX_086915), *pv1* (parasitophorous vacuole protein 1) and *lisp1* and *lisp2* (PVX_085550 and PVX_000975). The *lisp1* gene is an intriguing find, and may have an altered role in *P. vivax* liver stages (Additional File 1: Figure S7). In *P. berghei*, *lisp1* is essential for rupture of the PVM during liver stage development allowing release of the

merozoite into the host blood stream. *Pv-lisp1* is ~350-fold and ~1,350-fold more highly transcribed in *P. vivax* sporozoites compared to sporozoites of either *P. falciparum* or *P. yoelli* (see Additional File 2: Table S4). Also notable among translationally repressed genes in sporozoites is a putative 'Yippee' homolog (PVX_099695). Yippee is a DNA-binding protein that, in humans (YPEL3), suppresses cell growth [42]. Its specific function in *Plasmodium*, either in parasite development or on the host interactions, is not yet known. However, that Yippee-like proteins suppress cell growth/division and appear to be regulated through histone acetylation [43] is intriguing in the context of a potential role in *P. vivax* hypnozoite developmental arrest.

The *P. vivax* ortholog (PVP01_1016100; no corresponding ortholog is identified in the *P. vivax* Sal1 assembly) of the *P. cynomolgi* AP2 transcription factor, PCYB_102390, which was recently designated AP2-Q (i.e., 'quiescent') due to its enriched transcription in *P. cynomolgi* hypnozoites [22], is also detectable as transcripts but not proteins in *P. vivax* sporozoites. This may support a specific role for this transcription factor in hypnozoites. However, as Pv-AP2-Q is transcribed at an abundance (~50 TPM) at or below which ~≥50% of *P. vivax* genes are detectable as transcripts but not as proteins, the lack of detected AP2-Q protein could as likely result from the detection sensitivity of the proteomics data-set as from translation repression. Furthermore, while AP2-Q is proposed in *P. cynomolgi* as a possible hypnozoite marker in part due to its presence in *P. cynomolgi*, *P. vivax* and *P. ovale* (all of which generate hypnozoites) and reported absence from other *Plasmodium* species [22]. However, orthologs of this gene are also identified in PlasmoDB for several non-hypnozoite producing *Plasmodium* species, such as *P. knowlesi*, *P. gallinaceum* and *P. inui*, raising questions in regard to its function in these parasite species.

Lastly, while *Plasmodium* species lack a classical Golgi body, some genes (e.g., *golgi reassembly stacking protein*) functioning in protein transport between the Golgi body and the endoplasmic reticulum have been repurposed for vesicular transport and protein secretion during invasion [44]. Noting this, several homologs of genes associated with cycling of proteins between the Golgi body and the ER in other eukaryotes, including COPI-associated protein (PVX_100850), a putative STF2 (PVX_116780) and Got1 (PVX_090050) appear under translational repression in *P. vivax* sporozoites. Interestingly, in liver cells, the membrane of the parasitophorous vacuole, in which *Plasmodium* resides, often associates with the host cell ER and Golgi apparatus and may exploit this association to hijack host secretory pathways [45]. This may represent a key mechanism underpinning development in hepatocytes meriting further study.

Apoptosis-inhibition

Also notable among genes apparently translationally repressed in sporozoites are two putative Bax1 (Bcl-2 associated X protein) inhibitors (PVX_117470 and PVX_101315). Bax1 dimerizes with Bcl-2 to promote intrinsic apoptosis, leading to destruction of the mitochondrial membrane, caspase release and cell death. Bax1 inhibitor is a component of the cell stress response to prevent Bax1 from prematurely triggering cell death. When Bax1 is blocked, Bcl-2 switches from a cell-death to a pro-survival/anti-apoptotic role [46]. Intriguingly, specific suppression of mitochondrial-induced apoptosis has been demonstrated in liver-cells infected with *P. yoelii* [47] and this anti-apoptotic signal is blocked by Bcl-2 family inhibitors [48]. Orthologs of both *P. vivax* encoded Bax1 inhibitors are found in all *Plasmodium* species, suggesting a conserved function across the genus. Nonetheless, it is attractive to contemplate a potential role for these genes in promoting survival of host hepatocytes following the initial parasite invasion. Notably, the *P. cynomolgy* orthology of PVX_101315, PCYB_147290, is ~2-fold enriched in transcript abundance in schizonts compared to hypnozoites, which may indicate a role in repressing hepatocyte cell death during parasite replication rather than extending its life-span during parasite dormancy. This is to be explored.

Potential binding motif for Pv-Puf2

Research in *Toxoplasma gondii*, has identified a repetitive UGU motif in coding regions of translationally repressed genes bound by *Tg*-Puf2 [49] and, presumably, mediating repression. A similar UGU motif has been identified in the 3'UTR of *P. falciparum* transcripts (e.g., pfs25 and pfs28) and shown to bind PfPUF2 leading to their translational repression [50]. The binding motif for *Pv*-PUF2 has not been described. We found one motif (AGAT[TAC]G; Additional File 1: Figure S8) over-represented in coding regions of putatively repressed sporozoite transcripts relative to similarly highly transcribed but also translated genes e-value: $1.9e^{-9}$). We note the complementarity between AGAT and UGUA, however no over-represented motifs were detected in the 3'UTRs of these genes. Intriguingly, translational repression of *uis4* in *P. berghei* does not require the UTR [15]. It may be that the location of the *Puf2*-binding motif is somewhat flexible in *Plasmodium* and other apicomplexan species. We also identified a similarly over-represented motif ([GT]CGTC[CT]) within 500bp upstream of putatively repressed genes (p-value: 2.2e-9). It is possible this motif is a binding site for an as yet unattributed transcription factor co-ordinating genes destined for translational repression in the sporozoite. This motif is comparable to the [AG]C[AG]TGC motif identified for Pf-AP2-Sp [24], a transcription factor that is required for sporozoite development in *P. berghei* [51], and transcriptionally enriched in *P. falciparum* [24] and *P. vivax* (Additional File 2: Table S7) sporozoites relative to oocysts or blood stages respectively.

Histone modifications in *P. vivax* sporozoites

No epigenetic data are currently available for any *P. vivax* life-cycle stage. Studies of *P. falciparum* blood-stages have identified the importance of histone modifications as a primary epigenetic regulator [52, 53] and characterized key markers of heterochromatin ($H3K9me^3$) and euchromatin/transcriptional activation ($H3K4me^3$ and H3K9ac). Recently, these marks have been explored with the maturation of *P. falciparum* sporozoites in the mosquito [24]. Here, we characterize these marks in *P. vivax* sporozoites and assess their relationship to transcript abundance. Clearly this is of particular interest as a potential mechanism for dynamic regulation of sporozoite development in human hepatocytes. We identified 1,506, 1,999 and 5,262 ChIP-seq peaks stably represented in multiple *P. vivax* sporozoite replicates and associated with $H3K9me^3$, H3K9ac and $H3K4me^3$ histone marks respectively (Fig. 2). Peak width, spacing and stability differed with histone mark type (Additional File 1: Figures S9 and S10). $H3K4me^3$ peaks were significantly broader (mean width: 1,985 bp) than H3K9 peaks, and covered the greatest breadth of the genome; 36.0% of all bases were stably associated with $H3K4me^3$ marks. This mark was also most stable among replicates, with just ~16% of bases associated with an $H3K4me^3$ not supported by more than one biological replicate. By comparison $H3K9me^3$ marks were narrowest (mean width: 796 bp) and least stable, with 46% of bases associated with this mark supported by just one replicate. Consistent with observations in *P. falciparum* $H3K9me^3$ 'heterochromatin' marks primarily clustered in telomeric and subtelometric regionsv (Additional File 1: Figure S11). In contrast, the 'euchromatin' / transcriptionally open histone marks, $H3K4me^3$ and H3K9ac clustered around genic regions and did not overlap with regions under $H3K9me^3$ suppression. Both $H3K9me^3$ and $H3K4me^3$ marks were reasonably uniformly distributed (mean peak spacing ~500bp for each) within their respective regions of the genome. In contrast, H3K9ac peaks were spaced farther apart (mean: ~2kb), but also with a greater variability in spacing (likely reflecting their association with promoter regions [54]). The instability of $H3K9me^3$ may reflect its use in *Plasmodium* for regulating variegated expression of contingency genes from multigene families whose members have overlapping and redundant functions [55] and confer phenotypic plasticity [56].

Genes under histone regulation

We explored an association between these histone marks and the transcriptional behaviour of protein coding genes (Fig. 2 and Additional File 2: Tables S13-17). 485 coding genes stably intersected with an $H3K9me^3$ mark; all are located near the ends of the chromosomal scaffolds (i.e., are (sub)telomeric). On average, these genes are transcribed at ~30 fold lower

levels (mean <3 TPKMs) than genes not stably intersected by H3K9me^3 marks. These data clearly support the function of this mark in transcriptional silencing. This is largely consistent with observations in *P. falciparum* sporozoites [24], however, we observe no instances of genes that are stably marked by H3K9me^3 and moderately or highly transcribed regardless. Whether this relates to differences in epigenetic control between the species is not clear. We note that (sub)telomeric genes are overall transcriptionally silent in *P. vivax* sporozoites relative to blood-stages (Fig. 2a and 2b and Additional File 2: Tables S18-20). Consistent with observations in *P. falciparum* [52], the bulk of these genes include complex protein families, such as *vir* and *Pv-fam* genes, which function primarily in blood-stages. Also notable among the genes are several reticulocyte-binding proteins, including RBP2, 2a, 2b and 2c. Strikingly, we find no exceptions to this trend in our data, indicating the (sub)telomeres are remarkably transcriptionally silent in the sporozoite stage. By comparison, H3K4me^3 marks are stably associated with the Transcription Start Site (TSS) and/or 5' UTRs of 3,677 genes. We also identified 1,284 coding genes stably associated with an H3K9ac mark within 1kb or the TSS, with 179 of these genes stably marked also by H3K4me^3. The average transcription of these genes is 116, 180 and 199 TPKMs respectively (39, 60 and 66-fold higher than H3K9me^3 marked genes). These data support the role of these marks in transcriptional activation, the lower abundance of H3K4me^3 marker, compared with H3K9ac or H3K9ac and H3K4me^3 marked genes suggest these marks work synergistically and that H3K9ac is possibly the better single mark indicator of transcriptional activity in *P. vivax*. This is consistent with recent observations in *P. falciparum* sporozoites [24].

Interestingly, H3K9ac-marked genes ranged in transcriptional activity from the most abundantly transcribed genes to many in the lower 50% and even lowest decile of transcription. This suggests more contributes to transcriptional activation in *P. vivax* than, simply, gene accessibility through chromatin regulation. Specific activation by a transcription factor (e.g., ApiAP2s [57]) is the most obvious candidate. To explore this, we compared upstream regions (within 1kb of the TSS or up to the 3' end of the next gene upstream, whichever was less) of highly (top 10%) and lowly (bottom 10%) transcribed H3K9ac marked genes for over-represented sequence motifs that might coincide with known ApiAP2 transcription factor binding sites [58]. We identified these based on the location of the nearest stable H3K9ac peak relative to the transcription start site for each gene (Additional File 1: Figure S12). In most instances, these peaks were within 100bp of the TSS and, consistent with data from *P. falciparum* [54], *P. vivax* promoters appear to be no more than a few hundred to a maximum of 1000 bp upstream of the TSS. Exploring these regions, we identified two over-represented motifs: TGTACMA (e-value 2.7e^{-2}) and ATATTTH (e-value 3.3e^{-3}) (Fig. 2D). TGTAC is consistent with the known binding site for *Pf*-AP2-G, which regulates sexual differentiation in gametocytes [59], but its *P. vivax* ortholog (PVX_123760) is neither highly transcribed nor expressed in sporozoites. It may be that some genes encoding this domain are active in both sporozoites and gametocytes, but regulated by different mechanisms in each stage. Alternatively, this motif may represent a binding site for another, as yet uncharacterized transcription factor (e.g., PVX_083040). ATATTTH is similar to the binding motif for *Pf*-AP2-L (AATTTCC), a transcription factor that is important for liver stage development in *P. berghei* [60]. In contrast to AP2-G, *Pv*-AP2-L (PVX_081180) is in the top 10% of transcription and expression in *P. vivax* sporozoites and enriched relative to blood-stages. In *P. vivax* sporozoites, the ATATTTH motif is associated with a number of highly transcribed genes, including *lisp1* and *uis2-4*, known to be regulated by AP2-L in *P. berghei* [60] as well as many of the most highly transcribed, H3K9ac marked genes, including two *etramps* (PVX_086815 and PVX_088870), several RNA-binding proteins, including *Puf2*, *ddx5* and a dead-box helicase (PVX_123240), as well as one of the putative *bax1* inhibitors (PVX_101315). Interestingly, a number of highly transcribed and translationally repressed genes associated with the ATATTTH motif, including *uis4*, *siap2* and *pv1*, are not stably marked by H3K9ac in all replicates (i.e., there is significant variation in the placement of the H3K9ac peak or their presence/absence among replicates for these genes). It may be that additional histone modifications, for example H3K27me or H2 or H4 modifications, are involved in regulating transcription of these genes. Certainly H2A.Z, which is present in *P.*

falciparum, and controls temperature responses in plants [61] is intriguing as a potential mark regulating sporozoite fate in *P. vivax* considering the association between hypnozoite activation rate and climate [11].

Conclusions

We provide the first comprehensive study of the transcriptome, proteome and epigenome of infectious *Plasmodium vivax* sporozoites and the only study to integrate 'omics investigation of the sporozoite of any *Plasmodium* species. These data support the proposal that the sporozoite is a highly-programmed stage that is primed for invasion of and development in the host hepatocyte. Translational repression clearly plays a major role in shaping this stage, with many of the genes proposed here as being under translational repression are involved in hepatocyte infection and early liver-stage development. We highlight a major role for RNA-binding proteins, including PUF2, ALBA2/4 and, intriguingly, 'Homologue of Musashi' (HoMu). Noting that HoMu uses translational repression to regulate, in *Drosophila*, stem cell, and, in *Plasmodium*, gametocyte differentiation, it is intriguing to contemplate its potential role in setting liver-stage developmental fate. Identifying the sporozoite transcripts regulated by HoMu and other RNA binding proteins should be a key priority. As should in-depth comparative analysis using similar approaches of differences between/among relapsing and non-relapsing *Plasmodium* species, as well as, *P. vivax* field isolates with distinct, hypnozoite phenotypes. Our study provides a key foundation for understanding the early stages of hepatocyte infection and the developmental switch between liver trophozoite and hypnozoite formation. Importantly, it is a major first step in rationally prioritizing targets underpinning liver-stage differentiation for functional evaluation in humanized mouse and simian models for relapsing *Plasmodium* species and identifying novel avenues to understand and eradicate liver-stage infections.

Methods

Material collection, isolation and preparation

Nine field isolates (PvSpz-Thai 1 to 9), representing symptomatic blood-stage malaria infections were collected as venous blood (20 mL) from patients presenting at malaria clinics in Tak and Ubon Ratchatani provinces in Thailand. Each isolate was used to establish, infections in *Anopheles dirus* colonized at Mahidol University (Bangkok) by membrane feeding [13], after14-16 days post blood feeding, ~3-15 million sporozoites were harvested per field isolate from the salivary glands of up to 1,000 of these mosquitoes as per [62] and shipped in preservative (trizol (RNA/DNA) or 1% paraformaldehyde (DNA for ChiP-seq)) to the Walter and Eliza Hall Institute (WEHI).

Transcriptomics sequencing and differential analysis

Upon arrival at WEHI, messenger RNAs were purified from an aliquot (~0.5-1 million sporozoites) of each *P. vivax* field isolate as per [29] and subjected to RNA-seq on Illumina NextSeq using TruSeq library construction chemistry as per the manufacturer's instructions. Raw reads for each RNA-seq replicate are available through the Sequence Read Archive (XXX-XXX). Sequencing adaptors were removed and low quality reads trimmed and filtered using Trimmomatic v. 0.36 [63]. To remove host contaminants, processed reads were aligned, as single-end reads, to the *Anopholes dirus* wrari2 genome (VectorBase version W1) using Bowtie2[64] (--very-sensitive preset). All non-host reads were then aligned to the manually curated transcripts of the *P. vivax* P01 genome (http://www.genedb.org/Homepage/PvivaxP01) using RSEM [65] (pertinent settings: --bowtie2 --bowtie2-sensitivity-level very_sensitive --calc-ci --ci-memory 10240 --estimate-rspd --paired-end). Transcript abundance for each gene in each replicate was calculated by RSEM as raw count, posterior mean estimate expected counts (pme-EC) and transcripts per million (TPM).

Transcriptional abundance in *P. vivax* sporozoites was compared qualitatively (by ranked abundance) with previously published microarray data for *P. vivax* salivary-gland sporozoites [23]. As a further quality control, these RNA-seq data were compared also with

previously published microarray data for *P. falciparum* salivary-gland sporozoites [26], as well as RNA-seq data from salivary-gland sporozoites generated here for *P. falciparum* (single replicate generated from *P. falciparum* 3D7 lab cultures isolated from *Anopholes stephensi* and processed as above) and previously published for *P. yoelii* [25]. RNA-seq data from these additional *Plasmodium* species were (re)analysed from raw reads and transcriptional abundance for each species was determined (raw counts and pme-EC and TPM data) as described above using gene models current as of 04-10-2016 (PlasmoDB release v29). Interspecific transcriptional behaviour was qualitatively compared by relative ranked abundance in each species using TPM data for single copy orthologs (SCOs; defined in PlasmoDB) only, shared between *P. vivax* and *P. faliciparum* or shared among *P. vivax*, *P. falciparum* and *P. yoelii*.

To define sporozoite-enriched transcripts, we remapped raw reads representing early (18-24 hours post-infection (HPI)), mid (30-40 HPI) and late (42-46 HPI) *P. vivax* blood-stage infections recently published by Zhu *et al* [29] to the *P. vivax* P01 transcripts using RSEM as above. All replicate data was assessed for mapping metrics, transcript saturation and other standard QC metrics using QualiMap v 2.1.3 [66]. Differential transcription between *P. vivax* salivary-gland sporozoites and mixed blood-stages [29] was assessed using pme-EC data in EdgeR [67] (differential transcription cut-off: $\geq$ 2-fold change in counts per million (CPM) and a False Discovery Rate (FDR) $\leq$ 0.05). Pearson Chi squared tests were used to detect over-represented Pfam domains and Gene Ontology (GO) terms among differentially transcribed genes in sporozoites (Bonferroni-corrected $p < 0.05$), based on gene annotations in PlasmoDB (release v29).

Proteomic sequencing and quantitative analysis

Aliquots of ~10^7 salivary-gland sporozoites were generated from PvSpz-Thai1 and PvSpz-Thai6 isolates, purified on an Accudenz gradient per [62] and shipped on dry ice (protein) to the Center for Infectious Disease Research (CIDR). These cells were lysed in 2x Sample Buffer and their proteins separated by SDS-PAGE per [40]. For the whole proteome analysis, each gel was run out 52 mm and cut into 27-29 fractions using a grid cutter (Gel Company, San Francisco, CA). Pooled peptides in each gel fraction were reduced in dithiothreitol / ammonium bicarbonate, and digested for 4.5 hours at 36 °C in 6.25 ng/mL trypsin under vortex at 700 RPM. The supernatant was recovered and peptides were extracted by incubating the gel in 2% (v/v) acetonitrile/1% (v/v) formic acid. Supernatant after three extractions was combined with the digest supernatant, evaporated to dryness in a rotary vacuum, and reconstituted in HPLC loading buffer consisting of 2% (v/v) acetonitrile/0.2% (v/v) trifluoroacetic acid. Nanoflow liquid chromatography (nanoLC) was performed using an Agilent 1100 nano pump with electronically controlled split flow or a Proxeon Easy nLC. Peptides were separated on a column with an integrated fritted tip (360 µm outer diameter (O.D.), 75 µm inner diameter (I.D.), 15 µm I.D. tip; New Objective) packed in-house with a 20 cm bed of C18 (Dr. Maisch ReproSil-Pur C18-AQ, 120 Å, 3 µm; Ammerbuch-Entringen, Germany). Tandem mass spectrometry (MS/MS) was performed with an LTQ Velos Pro-Orbitrap Elite (Thermo Fisher Scientific). Two nanoLC-MS technical replicates were performed for each fraction, with roughly half the available sample injected for each replicate. The mass spectrometry data generated for this manuscript, along with the search parameters, analysis parameters and protein databases can be downloaded from PeptideAtlas (www.peptideatlas.org) using the identifier #####.

Mass spectrometer output files were converted to .mZML format using MSConvert version 2.2.0 (whole proteome data) or 3.0.5533 (surface-labeled data) [68] and searched with X!Tandem [69] version 2013.06.15.1 JACKHAMMER and Comet version 2015.02 rev.0.[70] MS/MS data were analyzed using the Trans-Proteomic Pipeline[71] version 4.8.0 PHILAE. Peptide spectrum matches (PSM) generated by each search engine were analyzed separately with PeptideProphet [72] and combined in iProphet.[73] Protein identifications were inferred with ProteinProphet [74]. In the case that multiple proteins were inferred at equal confidence by a set of peptides, the inference was counted as a single identification and all relevant protein IDs were listed. Only proteins with ProteinProphet probabilities corresponding to a

model-estimated false discovery rate (FDR) less than 1.0 % were reported. Spectra were searched against a protein sequence database comprised of *P. vivax* P01 (version 29, www.plasmodb.org), *An. stephensi* SDA 500 (version 1.3, www.vectorbase.org), and a modified version of the common Repository of Adventitious Proteins (version 2012.01.01, www.thegpm.org/cRAP) with the Sigma Universal Standard Proteins removed and the LC calibration standard peptide [Glu-1] fibrinopeptide B appended. Label-free proteomics methods based on spectral counts (SpC) were used to identify proteins that were significantly more abundant in labeled samples compared to unlabeled controls. The SpC for a given protein in a given biological replicate was taken as the number of PSM used by ProteinProphet to make the protein inference. All SpC values were increased by one in order to give all proteins non-zero SpC values for log-transformation [75]. The spectral abundance factor (SAF) for a given protein was calculated as the quotient of the SpC and the protein's length and natural log-transformed to ln(SAF) [76]. For a more detailed description of the proteome data collection process and analysis please refer to manuscript by Swearingen *et al (submitted)*.

To identify genes likely under translational repression in the *P. vivax* sporozoite, we examined these data for genes that were highly transcribed (top 10 percentile) but for which we could find no evidence of protein expression in any sporozoite replicate. In addition, we conducted abundance ranked comparisons between the mean transcriptional abundance of each *P. vivax* gene in sporozoites (see above) and the mean quantitative abundance of its protein in our expressional data. Genes were sorted on the differential between their relative transcription and relative expression ranking to identify highly transcribed genes with substantially lower expression relative to their transcriptional abundance.

Salivary-gland sporozoite and liver-stage immunofluorescence assays (IFAs)

IFAs were performed as per [13]. Liver stages were obtained from 10µm formalin fixed paraffin embedded day 7 liver stages generated previously [13] from FRG knockout huHep mice;[13] these were deparaffinized prior to staining. Fresh salivary-gland sporozoites were fixed in acetone per [13]. All cells were incubated twice for 3 minutes in Xylene, then 100% Ethanol, and finally once for 3 minutes each in 95%, 70%, and 50% Ethanol. The cells were rinsed in DI water and permeabilized immediately in 1XTBS, containing Triton X-100 and 30% hydrogen peroxide. The cells were blocked in 5% milk in 1XTBS. The hepatocytes were stained overnight with a rabbit polyclonal LISP1 antibody (A), a rabbit polyclonal UIS4 antibody (B), and a rabbit polyclonal BIP antibody (C) in blocking buffer. The cells were washed with 1XTBS and the primary antibodies were detected with goat anti-rabbit Alexa Fluor 488 antibody (Life Technologies). The cells were washed in 1XTBS. The hepatocytes were rinsed in KMNO4 and washed in 1XTBS. The cells were incubated in DAPI for 5 minutes.

Histone ChIP sequencing and analysis

Aliquots of 2 – 6 million freshly isolated sporozoites were fixed with 1% paraformaldehyde for 10 min at 37°C and the reaction subsequently quenched by adding glycine to a final concentration of 125 mM. After three washes with PBS, sporozoite pellets were stored at -80°C and shipped to Australia. Nuclei were released from the sporozoites by dounce homogenization in lysis buffer (10 mM Hepes pH 7.9, 10 mM KCl, 0.1 mM EDTA, 0.1 mM EDTA, 1 mM DTT, 1x EDTA-free protease inhibitor cocktail (Roche), 0.25% NP40). Nuclei were pelleted by centrifugation at 21,000 g for 10 min at 4°C and resuspended in SDS lysis buffer (1% SDS, 10 mM EDTA, 50 mM Tris pH 8.1, 1x EDTA-free protease inhibitor cocktail). Chromatin was sheared into 200–1000 bp fragments by sonication for 16 cycles in 30 sec intervals (on/off, high setting) using a Bioruptor (Diagenode) and diluted 1:10 in ChIP dilution buffer (0.01% SDS, 1.1% Triton X-100, 1.2 mM EDTA, 16.7 mM Tris pH 8.1, 150 mM NaCl). Chromatin was precleared for 1 hour with protein A/G sepharose (4FastFlow, GE Healthcare) equilibrated in 0.1% BSA in ChIP dilution buffer. Chromatin from 3 x 10^5 nuclei was taken aside as input material. Chromatin from approximately 3 x 10^6 sporozoite nuclei was used for each ChIP. ChIP was carried out over night at 4°C with 5 µg of antibody

(H3K9me3 (Active Motif), H3K4me3 (Abcam), H3K9ac (Upstate), H4K16ac (Abcam)) and 10 µl each of equilibrated protein A and G sepharose beads (4FastFlow, GE Healthcare). After washes in low-salt, high-salt, LiCl, and TE buffers (EZ-ChIP Kit, Millipore), precipitated complexes were eluted in 1% SDS, 0.1 M $NaHCO_3$. Cross-linking of the immune complexes and input material was reversed for 6 hours at 45°C after addition of 500 mM NaCl and 20 µg/ml of proteinase K (NEB). DNA was purified using the MinElute® PCR purification kit (Qiagen) and paired-end sequenced on Illumina NextSeq using TruSeq library construction chemistry as per the manufacturer's instructions. Raw reads for each ChIP-seq replicate are available through the Sequence Read Archive (XXX-XXX).

Fastq files were checked for quality using fastqc (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and adapter sequences were trimmed using cutadapt [77]. Paired end reads were mapped to the *P. vivax* P01 strain genome annotation using Bowtie2 [64]. The alignment files were converted to Bam format, sorted and indexed using Samtools [78]. ChIP peaks were called relative to input using MACS2[79] in paired end mode with a q value less than or equal to 0.01. Peaks and peak summits were converted to sorted BED files. Bedtools intersect[80] was used to identify genes that intersected H3K9me3 peaks and Bedtools closest was used to identify genes that were closest to and downstream of H3K9ac and H3K4me3 peak summits.

Sequence motif analysis

Conserved sequence motifs were identified using the program DREME [81]. Only genes in the top decile of transcription showing no evidence of protein expression in multiple salivary-gland sporozoite replicates were considered as putatively translationally repressed (n = 170). We queried coding regions and regions upstream of the transcriptional start site (TSS) for each gene, defined by Zhu *et al* [29] and/or predicted here from all RNA-seq data using the Tuxedo suite [82], for enriched sequence motifs in comparison to 170 genes found to be in the top decile of both transcriptional and expressional abundance in the same sporozoite replicates. In searching for motifs associated with highly transcribed genes with stable H3K9ac marks within 1kb of the TSS (or up to the 3' end of the next gene upstream), we compared H3K9ac marked genes in the top decile of transcription to the same number of H3K9ac marked genes in the bottom decile of transcription. In both instances, an e-value threshold of 0.05 was considered the minimum threshold for statistical significance.

Author contributions: Study design and development: Ivo Muller[1,2,3] (IM), Aaron R. Jex[1,3,4] (AJ), Stefan H. I. Kappe[5] (SK) and Sebastian A. Mikolajczak[5] (SM); Parasite collection and sporozoite production and purification: Jetsumon Sattabongkot[7] (JSP), Rapatbhorn Patrapuvich[6] (RP), SK, SM, Scott Lindner[8] (SL) and Erika L. Flannery[5] (EF); DNA/RNA isolation and sequence library preparation: Cristian Koepfli[1] (CK) and EF; Transcriptomics analysis: AJ, Brendan Ansell[4] (BA) and Anita Lerch[1] (AL); Proteomics analysis: Kristian Swearingen[5] (KS), Robert Moritz (RM)[9] SL, SM and EF; ChIP-seq preparation and analyses: Michaela Petter[10] (MP) and Michael Duffy[10] (MD); Immunofluorescence assays: Vorada Chuenchob[5]; Data integration and interpretation: AJ, IM, EF, SM and SK; Manuscript preparation: AJ, IM, SM, SK, SL, and EF.

Author affiliations: 1. Population Health and Immunity Division, The Walter and Eliza Hall Institute for Medical Research, 1G Royal Parade, Parkville, Victoria, 3052, Australia; 2. Malaria: Parasites & Hosts Unit, Institut Pasteur, 28 Rue de Dr. Roux, 75015, Paris, France; 3. Department of Medical Biology, The University of Melbourne, Victoria, 3010, Australia; 4. Faculty of Veterinary and Agricultural Sciences, The University of Melbourne, Corner of Park and Flemington Road, Parkville, Victoria, 3010, Australia; 5. Center for Infectious Disease Research, 307 Westlake Avenue North, Suite 500, Seattle, WA 98109, USA; 6. Department of Global Health, University of Washington, Seattle, WA 98195, USA; 7. Mahidol Vivax Research Center, Faculty of Tropical Medicine, Mahidol University, Bangkok 10400, Thailand; 8. Department of Biochemistry and Molecular Biology, Center for Malaria Research, Pennsylvania State University, University Park, PA 16802, USA. 9. Institute for

Systems Biology, Seattle, WA, 98109, USA. 10. Department of Medicine Royal Melbourne Hospital, The Peter Doherty Institute, The University of Melbourne, 792 Elizabeth Street, Melbourne, Victoria 3000, Australia

Acknowledgements: The authors acknowledge funding from the National Health and Medical Research Council (NHMRC, APP1021544, 1043345, & 1092789), the Australian Research Council (ARC), the Victorian State Government Operational Infrastructure Support and Australian Government National Health and Medical Research Council Independent Research Institute Infrastructure Support Scheme, the Ian Potter Foundation, the National Institute of Health, the Bill and Melinda Gates Foundation, the US Department of Defense (W81XWH-15-1-0249) and the Office of the Assistant Secretary of Defence for Health Affairs through the Peer Reviewed Medical Research Program (PRMRP).

Competing Interests: The authors declare that no author of this manuscript has a competing financial or non-financial interest related to this work.

References

1. Organization WH: World Malaria Report 2015. WHO, Geneva. 2015.
2. Feachem RG, Phillips AA, Hwang J, Cotter C, Wielgosz B, Greenwood BM, et al: Shrinking the malaria map: progress and prospects. Lancet. 2010;376(9752):1566-78.
3. Price RN, Douglas NM, Anstey NM: New developments in *Plasmodium vivax* malaria: severe disease and the rise of chloroquine resistance. Curr Opin Infect Dis. 2009;22(5):430-5.
4. Baird KJ: Malaria caused by *Plasmodium vivax*: recurrent, difficult to treat, disabling, and threatening to life - averting the infectious bite preempts these hazards. Pathogens Global Health. 2013;107475-9.
5. Sattabongkot J, Tsuboi T, Zollner GE, Sirichaisinthop J, Cui L: *Plasmodium vivax* transmission: chances for control? Trends Parasitol. 2004;20(4):192-8.
6. Mueller I, Galinski MR, Baird JK, Carlton JM, Kochar DK, Alonso PL, et al: Key gaps in the knowledge of *Plasmodium vivax*, a neglected human malaria parasite. Lancet Infect Dis. 2009;9(9):555-66.
7. Lindner SE, Miller JL, Kappe SH: Malaria parasite pre-erythrocytic infection: preparation meets opportunity. Cell Microbiol. 2012;14(3):316-24.
8. Mota MM, Pradel G, Vanderberg JP, Hafalla JC, Frevert U, Nussenzweig RS, et al: Migration of *Plasmodium* sporozoites through cells before infection. Science. 2001;291(5501):141-4.
9. Shin SC, Vanderberg JP, Terzakis JA: Direct infection of hepatocytes by sporozoites of *Plasmodium berghei*. J Protozool. 1982;29(3):448-54.
10. Lysenko AJ, Beljaev A, Rybalka V: Population studies of *Plasmodium vivax*: 1. The theory of polymorphism of sporozoites and epidemiological phenomena of tertian malaria. Bulletin WHO. 1977;55(5):541.
11. White NJ: Determinants of relapse periodicity in *Plasmodium vivax* malaria. Malar J. 2011;10297.
12. Price RN, Tjitra E, Guerra CA, Yeung S, White NJ, Anstey NM: Vivax malaria: neglected and not benign. Amer J Trop Med Hyg. 2007;77(6 Suppl):79-87.
13. Mikolajczak SA, Vaughan AM, Kangwanrangsan N, Roobsoong W, Fishbaugher M, Yimamnuaychok N, et al: *Plasmodium vivax* liver stage development and hypnozoite persistence in human liver-chimeric mice. Cell Host Microbe. 2015;17(4):526-35.
14. Mueller A-K, Camargo N, Kaiser K, Andorfer C, Frevert U, Matuschewski K, et al: *Plasmodium* liver stage developmental arrest by depletion of a protein at the parasite–host interface. Proc Natl Acad Sci U S A. 2005;102(8):3022-7.

15. Silvie O, Briquet S, Muller K, Manzoni G, Matuschewski K: Post-transcriptional silencing of UIS4 in *Plasmodium berghei* sporozoites is important for host switch. Mol Microbiol. 2014;91(6):1200-13.
16. Mackellar DC, O'Neill MT, Aly AS, Sacci JB, Jr., Cowman AF, Kappe SH: *Plasmodium falciparum* PF10_0164 (ETRAMP10.3) is an essential parasitophorous vacuole and exported protein in blood stages. Eukaryot Cell. 2010;9(5):784-94.
17. Dembele L, Franetich JF, Lorthiois A, Gego A, Zeeman AM, Kocken CH, et al: Persistence and activation of malaria hypnozoites in long-term primary hepatocyte cultures. Nat Med. 2014;20(3):307-12.
18. Malmquist NA, Moss TA, Mecheri S, Scherf A, Fuchter MJ: Small-molecule histone methyltransferase inhibitors display rapid antimalarial activity against all blood stage forms in *Plasmodium falciparum*. Proc Natl Acad Sci U S A. 2012;109(41):16708-13.
19. Josling GA, Llinas M: Sexual development in *Plasmodium* parasites: knowing when it's time to commit. Nat Rev Microbiol. 2015;13(9):573-87.
20. White MT, Karl S, Battle KE, Hay SI, Mueller I, Ghani AC: Modelling the contribution of the hypnozoite reservoir to *Plasmodium vivax* transmission. Elife. 2014;3.
21. Roobsoong W, Tharinjaroen CS, Rachaphaew N, Chobson P, Schofield L, Cui L, et al: Improvement of culture conditions for long-term in vitro culture of *Plasmodium vivax*. Malaria J. 2015;14(1):1.
22. Cubi R, Vembar SS, Biton A, Franetich JF, Bordessoulles M, Sossau D, et al: Laser capture microdissection enables transcriptomic analysis of dividing and quiescent liver stages of *Plasmodium* relapsing species. Cell Microbiol. 2017.
23. Westenberger SJ, McClean CM, Chattopadhyay R, Dharia NV, Carlton JM, Barnwell JW, et al: A systems-based analysis of *Plasmodium vivax* lifecycle transcription from human to mosquito. PLoS Negl Trop Dis. 2010;4(4):e653.
24. Gomez-Diaz E, Yerbanga RS, Lefevre T, Cohuet A, Rowley MJ, Ouedraogo JB, et al: Epigenetic regulation of *Plasmodium falciparum* clonally variant gene expression during development in *Anopheles gambiae*. Sci Rep. 2017;740655.
25. Lindner SE, Mikolajczak SA, Vaughan AM, Moon W, Joyce BR, Sullivan WJ, Jr., et al: Perturbations of *Plasmodium* Puf2 expression and RNA-seq of Puf2-deficient sporozoites reveal a critical role in maintaining RNA homeostasis and parasite transmissibility. Cell Microbiol. 2013;15(7):1266-83.
26. Le Roch KG, Johnson JR, Florens L, Zhou Y, Santrosyan A, Grainger M, et al: Global analysis of transcript and protein levels across the *Plasmodium falciparum* life cycle. Genome Res. 2004;14(11):2308-18.
27. Mikolajczak SA, Silva-Rivera H, Peng X, Tarun AS, Camargo N, Jacobs-Lorena V, et al: Distinct malaria parasite sporozoites reveal transcriptional changes that cause differential tissue infection competence in the mosquito vector and mammalian host. Mol Cell Biol. 2008;28(20):6196-207.
28. Carlton JM, Adams JH, Silva JC, Bidwell SL, Lorenzi H, Caler E, et al: Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*. Nature. 2008;455(7214):757-63.
29. Zhu L, Mok S, Imwong M, Jaidee A, Russell B, Nosten F, et al: New insights into the *Plasmodium vivax* transcriptome using RNA-Seq. Sci Rep. 2016;620498.
30. Kramer S: RNA in development: how ribonucleoprotein granules regulate the life cycles of pathogenic protozoa. WIR: RNA. 2014;5(2):263-84.
31. Tucker RP: The thrombospondin type 1 repeat superfamily. Int J Biochem Cell Biol. 2004;36(6):969-74.
32. Ntumngia FB, Bouyou-Akotet MK, Uhlemann AC, Mordmuller B, Kremsner PG, Kun JF: Characterisation of a tryptophan-rich *Plasmodium falciparum* antigen associated with merozoites. Mol Biochem Parasitol. 2004;137(2):349-53.

33. Gubbels MJ, Vaishnava S, Boot N, Dubremetz JF, Striepen B: A MORN-repeat protein is a dynamic component of the *Toxoplasma gondii* cell division apparatus. J Cell Sci. 2006;119(Pt 11):2236-45.
34. Aly AS, Lindner SE, MacKellar DC, Peng X, Kappe SH: SAP1 is a critical post-transcriptional regulator of infectivity in malaria parasite sporozoite stages. Mol Microbiol. 2011;79(4):929-39.
35. Okano H, Imai T, Okabe M: Musashi: a translational regulator of cell fate. J Cell Sci. 2002;115(7):1355-9.
36. Cui L, Lindner S, Miao J: Translational regulation during stage transitions in malaria parasites. Annals N Y Acad Sci. 2015;1342(1):1-9.
37. Lasko P: Gene regulation at the RNA layer: RNA binding proteins in intercellular signaling networks. Sci STKE. 2003;179RE6.
38. Guerreiro A, Deligianni E, Santos JM, Silva PA, Louis C, Pain A, et al: Genome-wide RIP-Chip analysis of translational repressor-bound mRNAs in the *Plasmodium* gametocyte. Genome Biol. 2014;15(11):493.
39. Kappe SH, Buscaglia CA, Bergman LW, Coppens I, Nussenzweig V: Apicomplexan gliding motility and host cell invasion: overhauling the motor model. Trends in parasitology. 2004;20(1):13-6.
40. Lindner SE, Swearingen KE, Harupa A, Vaughan AM, Sinnis P, Moritz RL, et al: Total and putative surface proteomics of malaria parasite salivary gland sporozoites. Mol Cell Proteomics. 2013;12(5):1127-43.
41. Silvie O, Briquet S, Müller K, Manzoni G, Matuschewski K: Post-transcriptional silencing of UIS4 in *Plasmodium berghei* sporozoites is important for host switch. Molecular microbiology. 2014;91(6):1200-13.
42. Kelley KD, Miller KR, Todd A, Kelley AR, Tuttle R, Berberich SJ: YPEL3, a p53-regulated gene that induces cellular senescence. Cancer Res. 2010;70(9):3566-75.
43. Tuttle R, Simon M, Hitch DC, Maiorano JN, Hellan M, Ouellette J, et al: Senescence-associated gene YPEL3 is downregulated in human colon tumors. Ann Surg Oncol. 2011;18(6):1791-6.
44. Struck NS, de Souza Dias S, Langer C, Marti M, Pearce JA, Cowman AF, et al: Re-defining the Golgi complex in *Plasmodium falciparum* using the novel Golgi marker PfGRASP. J Cell Sci. 2005;118(Pt 23):5603-13.
45. Graewe S, Stanway RR, Rennenberg A, Heussler VT: Chronicle of a death foretold: *Plasmodium* liver stage parasites decide on the fate of the host cell. FEMS Microbiol Rev. 2012;36(1):111-30.
46. Bruchhaus I, Roeder T, Rennenberg A, Heussler VT: Protozoan parasites: programmed cell death as a mechanism of parasitism. Trends Parasitol. 2007;23(8):376-83.
47. Albuquerque SS, Carret C, Grosso AR, Tarun AS, Peng X, Kappe SH, et al: Host cell transcriptional profiling during malaria liver stage infection reveals a coordinated and sequential set of biological events. BMC genomics. 2009;10(1):1.
48. Kaushansky A, Metzger PG, Douglass AN, Mikolajczak SA, Lakshmanan V, Kain HS, et al: Malaria parasite liver stages render host hepatocytes susceptible to mitochondria-initiated apoptosis. Cell Death Dis. 2013;4e762.
49. Liu M, Miao J, Liu T, Sullivan WJ, Cui L, Chen X: Characterization of TgPuf1, a member of the Puf family RNA-binding proteins from *Toxoplasma gondii*. Parasites & vectors. 2014;7(1):1.
50. Miao J, Fan Q, Parker D, Li X, Li J, Cui L: Puf mediates translation repression of transmission-blocking vaccine candidates in malaria parasites. PLoS Pathog. 2013;9(4):e1003268.
51. Yuda M, Iwanaga S, Shigenobu S, Kato T, Kaneko I: Transcription factor AP2-Sp and its target genes in malarial sporozoites. Mol Microbiol. 2010;75(4):854-63.

824 52. Lopez-Rubio J-J, Mancio-Silva L, Scherf A: Genome-wide analysis of
825 heterochromatin associates clonally variant gene regulation with perinuclear
826 repressive centers in malaria parasites. Cell Host Microbe. 2009;5(2):179-90.
827 53. Duffy MF, Selvarajah SA, Josling GA, Petter M: Epigenetic regulation of the
828 *Plasmodium falciparum* genome. Brief Funct Genomics. 2014;13(3):203-16.
829 54. Cui L, Miao J, Furuya T, Li X, Su XZ, Cui L: PfGCN5-mediated histone H3
830 acetylation plays a key role in gene expression in *Plasmodium falciparum*.
831 Eukaryot Cell. 2007;6(7):1219-27.
832 55. Guizetti J, Scherf A: Silence, activate, poise and switch! Mechanisms of antigenic
833 variation in *Plasmodium falciparum*. Cell Microbiol. 2013;15(5):718-26.
834 56. Rovira-Graells N, Gupta AP, Planet E, Crowley VM, Mok S, de Pouplana LR, et al:
835 Transcriptional variation in the malaria parasite *Plasmodium falciparum*.
836 Genome Res. 2012;22(5):925-38.
837 57. De Silva EK, Gehrke AR, Olszewski K, León I, Chahal JS, Bulyk ML, et al: Specific
838 DNA-binding by apicomplexan AP2 transcription factors. Proc Natl Acad Sci U S
839 A. 2008;105(24):8393-8.
840 58. Painter HJ, Campbell TL, Llinás M: The Apicomplexan AP2 family: integral factors
841 regulating *Plasmodium* development. Mol Biochem Parasitol. 2011;176(1):1-7.
842 59. Kafsack BF, Rovira-Graells N, Clark TG, Bancells C, Crowley VM, Campino SG, et
843 al: A transcriptional switch underlies commitment to sexual development in
844 human malaria parasites. Nature. 2014;507(7491):248.
845 60. Iwanaga S, Kaneko I, Kato T, Yuda M: Identification of an AP2-family protein that
846 is critical for malaria liver stage development. PLoS One. 2012;7(11):e47557.
847 61. Boden SA, Kavanova M, Finnegan EJ, Wigge PA: Thermal stress effects on grain
848 yield in Brachypodium distachyon occur via H2A.Z-nucleosomes. Genome Biol.
849 2013;14(6):R65.
850 62. Kennedy M, Fishbaugher ME, Vaughan AM, Patrapuvich R, Boonhok R,
851 Yimamnuaychok N, et al: A rapid and scalable density gradient purification
852 method for *Plasmodium* sporozoites. Malar J. 2012;11421.
853 63. Bolger AM, Lohse M, Usadel B: Trimmomatic: a flexible trimmer for Illumina
854 sequence data. Bioinformatics. 2014;30(15):2114-20.
855 64. Langmead B, Salzberg SL: Fast gapped-read alignment with Bowtie 2. Nat
856 Methods. 2012;9(4):357-9.
857 65. Li B, Dewey CN: RSEM: accurate transcript quantification from RNA-Seq data
858 with or without a reference genome. BMC Bioinformatics. 2011;12(1):323.
859 66. Okonechnikov K, Conesa A, Garcia-Alcalde F: Qualimap 2: advanced multi-
860 sample quality control for high-throughput sequencing data. Bioinformatics.
861 2016;32(2):292-4.
862 67. Nikolayeva O, Robinson MD: edgeR for differential RNA-seq and ChIP-seq
863 analysis: an application to stem cell biology. Methods Mol Biol. 2014;115045-79.
864 68. Kessner D, Chambers M, Burke R, Agus D, Mallick P: ProteoWizard: open source
865 software for rapid proteomics tools development. Bioinformatics.
866 2008;24(21):2534-6.
867 69. Craig R, Beavis RC: TANDEM: matching proteins with tandem mass spectra.
868 Bioinformatics. 2004;20(9):1466-7.
869 70. Eng JK, Jahan TA, Hoopmann MR: Comet: an open-source MS/MS sequence
870 database search tool. Proteomics. 2013;13(1):22-4.
871 71. Deutsch EW, Mendoza L, Shteynberg D, Slagel J, Sun Z, Moritz RL: Trans-
872 Proteomic Pipeline, a standardized data processing pipeline for large-scale
873 reproducible proteomics informatics. Proteomics Clin Appl. 2015;9(7-8):745-54.
874 72. Keller A, Nesvizhskii AI, Kolker E, Aebersold R: Empirical statistical model to
875 estimate the accuracy of peptide identifications made by MS/MS and database
876 search. Anal Chem. 2002;74(20):5383-92.

73. Shteynberg D, Deutsch EW, Lam H, Eng JK, Sun Z, Tasman N, et al: iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. Mol Cell Proteomics. 2011;10(12):M111 007690.
74. Nesvizhskii AI, Keller A, Kolker E, Aebersold R: A statistical model for identifying proteins by tandem mass spectrometry. Anal Chem. 2003;75(17):4646-58.
75. Hendrickson EL, Xia Q, Wang T, Leigh JA, Hackett M: Comparison of spectral counting and metabolic stable isotope labeling for use with quantitative microbial proteomics. Analyst. 2006;131(12):1335-41.
76. Zybailov B, Mosley AL, Sardiu ME, Coleman MK, Florens L, Washburn MP: Statistical analysis of membrane proteome expression changes in Saccharomyces cerevisiae. J Proteome Res. 2006;5(9):2339-47.
77. Martin M: Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet journal. 2011;17(1):pp. 10-2.
78. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al: The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.
79. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al: Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9(9):R137.
80. Quinlan AR, Hall IM: BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841-2.
81. Bailey TL: DREME: motif discovery in transcription factor ChIP-seq data. Bioinformatics. 2011;27(12):1653-9.
82. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al: Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2012;7(3):562-78.

Figures

Fig. 1 Differential transcription between *Plasmodium vivax* salivary-gland sporozoites and blood-stages. **a** BCV plot showing separation between blood-stage (black) and salivary-gland sporozoite (red) biological replicates. **b** Volcano plot of distribution of fold-changes (FC) in transcription between blood-stages and salivary-gland sporozoites relative to statistical significance threshold (False Discovery Rate (FDR) ≤ 0.05). Positive FC represents enriched transcription in the sporozoite stage. **c** Mirror plot showing pFam domains statistically significantly (FDR ≤ 0.05) over-represented in salivary-gland sporozoite enriched (red) or blood-stage enriched (black) transcripts. Scale bar truncated for presentation. * - 55 PRESAN domains are in this dataset. ** - 99 Vir domains are in this dataset.

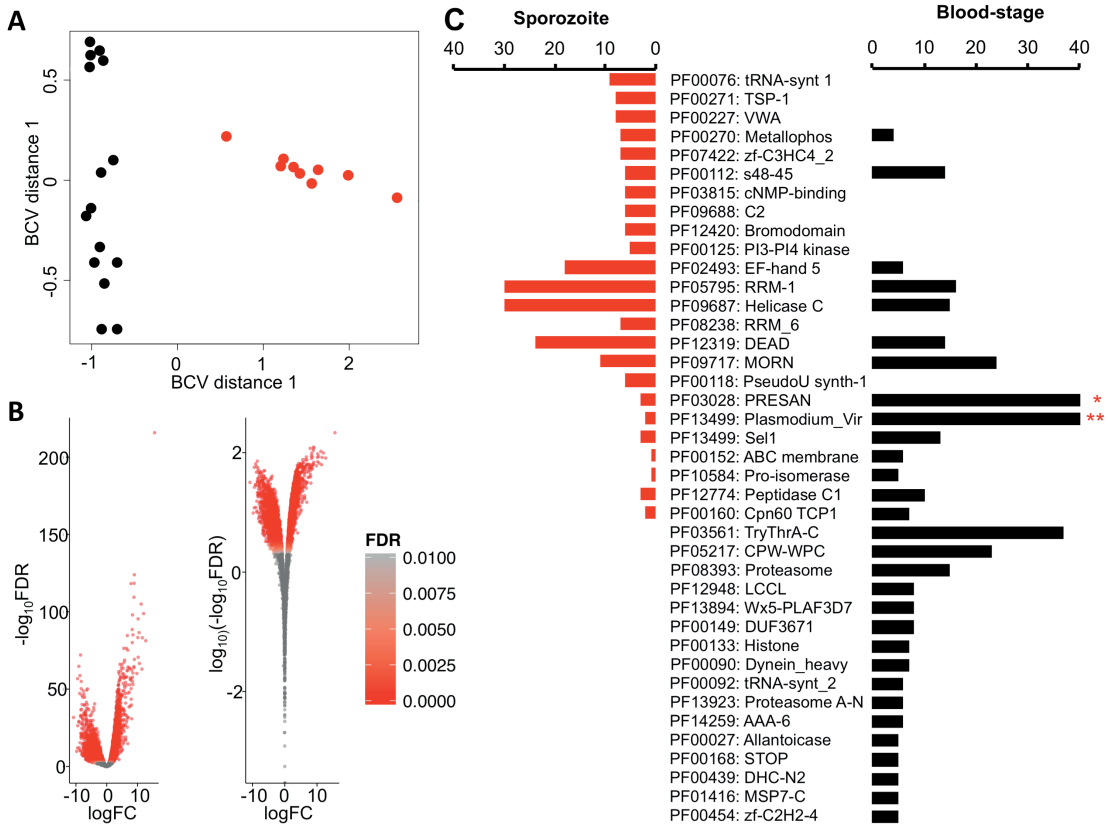
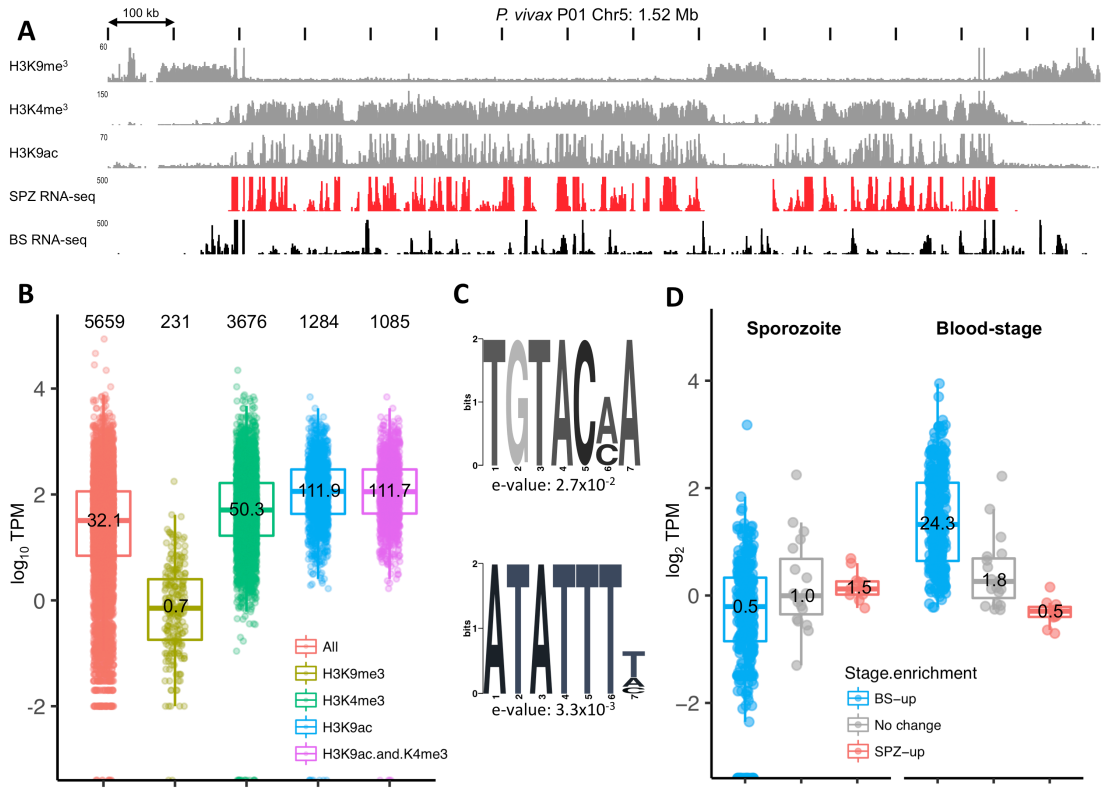
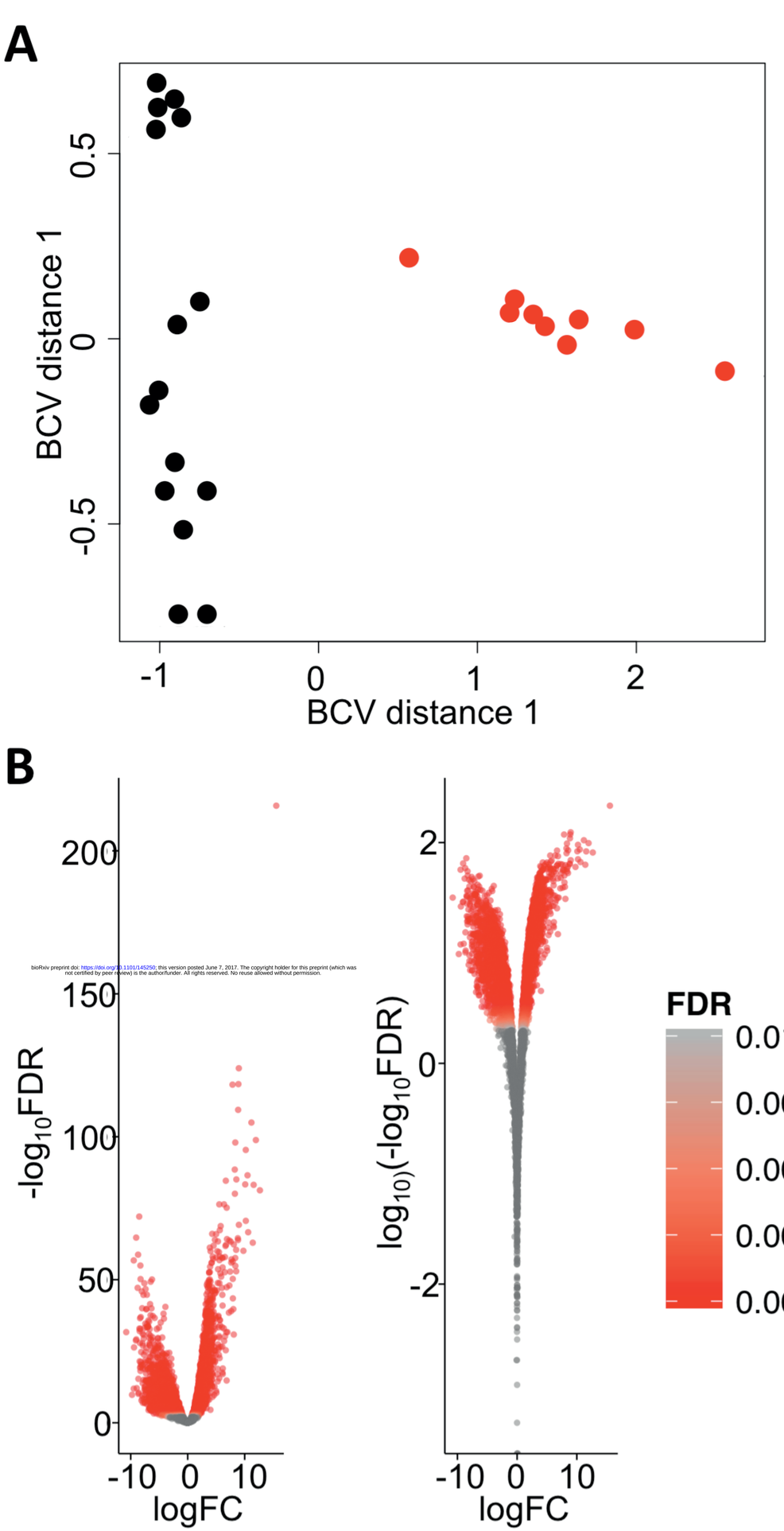
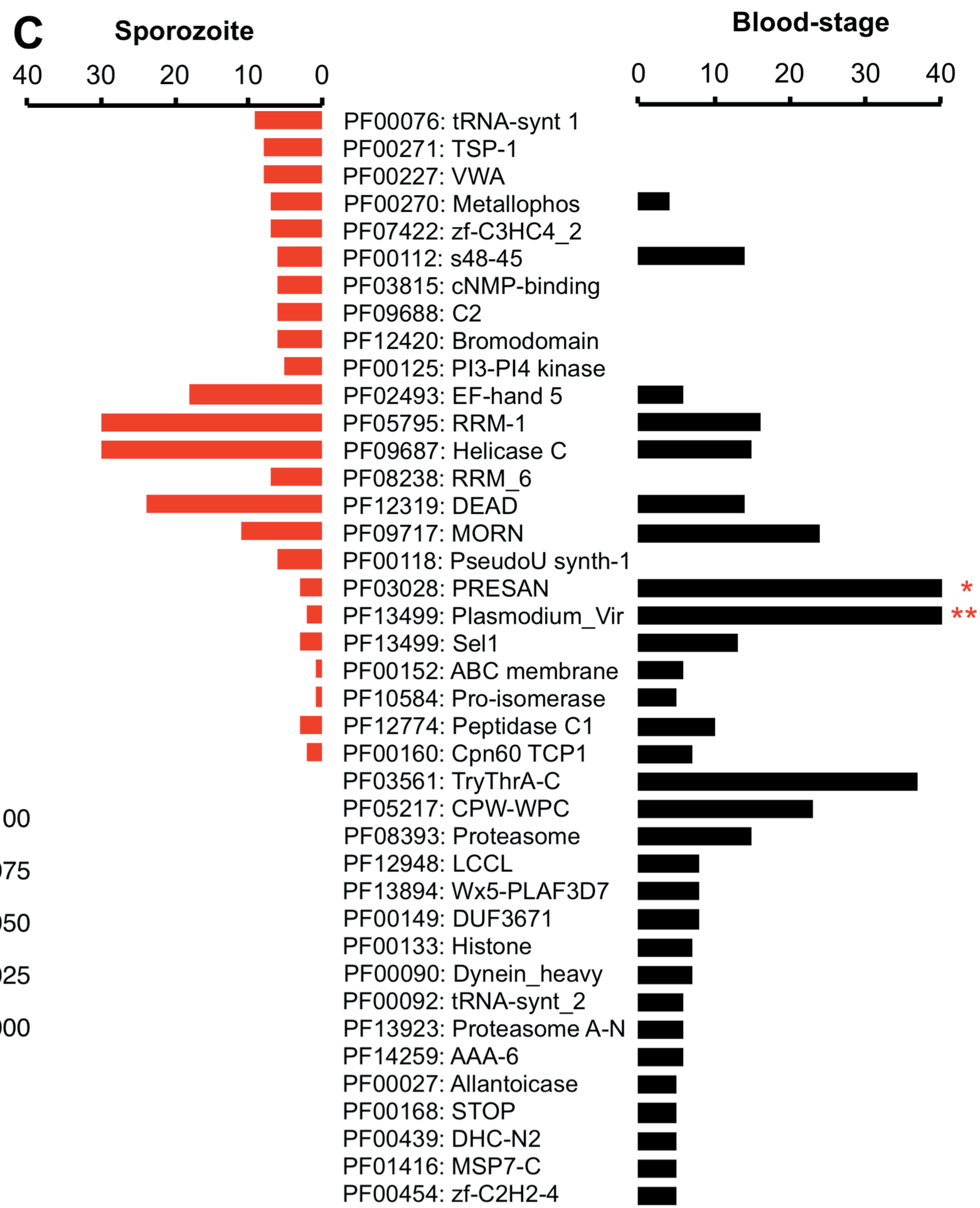
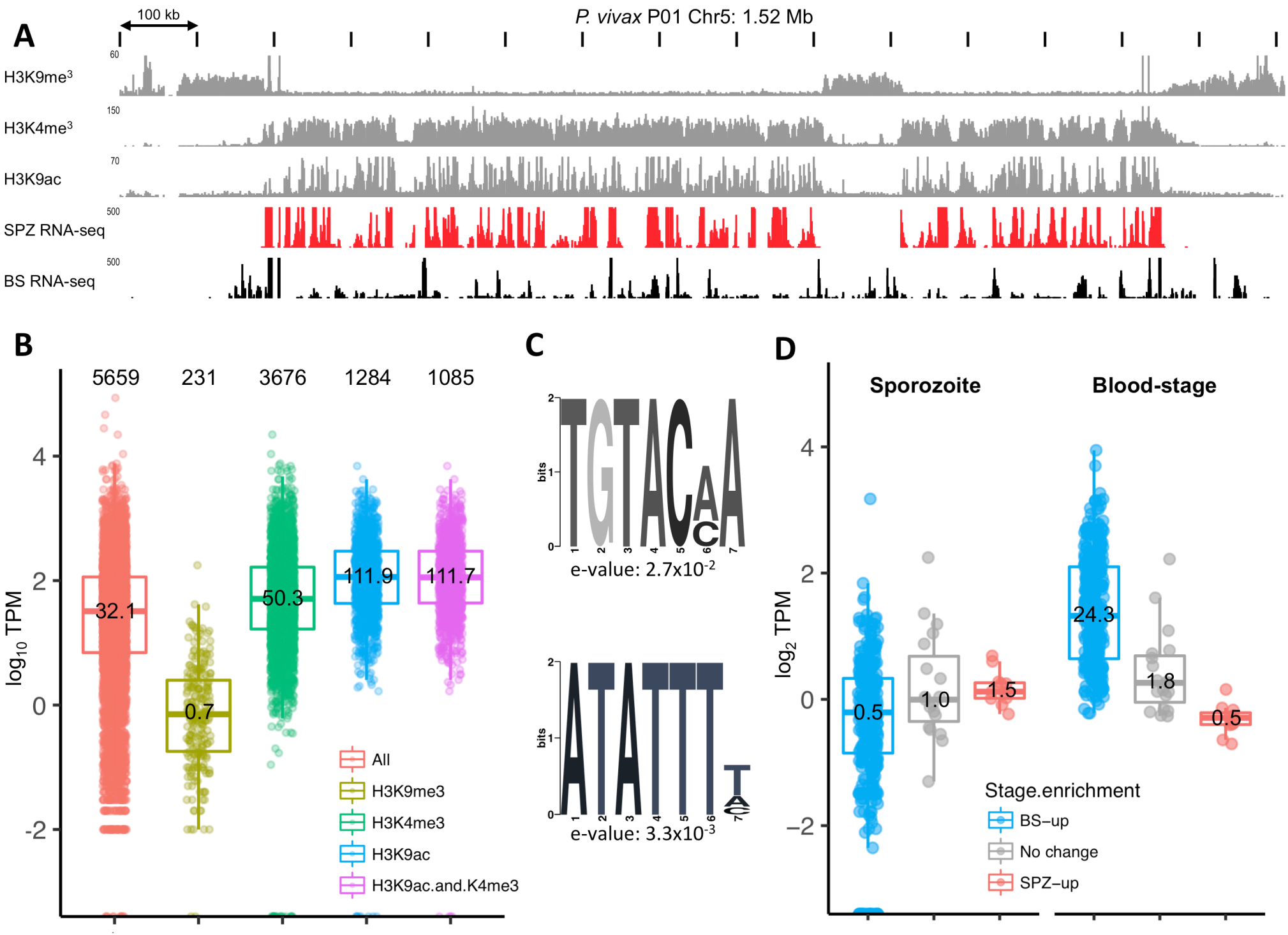


Fig. 2 Histone epigenetics relative to transcriptional behaviour in salivary-gland sporozoites. **a** Representative H3K9me^3, H3K4me^3 and H3K9ac ChIP-seq data (grey) from a representative chromosome (*P. vivax* P01 Chr5) relative to mRNA transcription in salivary-gland sporozoites (black) and blood-stages (black). Small numbers to top left of each row show data range. **b** Salivary-gland sporozoite transcription relative to nearest stable histone epigenetic marks. Numbers at the top of the figure represent total genes included in each category. Numbers within in box plot represent mean transcription in transcripts per million (TPM). **c** Sequence motifs enriched within 1kb upstream of the Transcription Start Site of highly transcribed (top 10%) relative to lowly transcribed genes associated with H3K9ac marks in salivary-gland sporozoites. **d** Relative transcription of (sub)telomeric genes in *P. vivax* salivary-gland sporozoites and blood-stages categorized by gene sets enriched in blood-stages (blue), salivary sporzoites (red) or not stage enriched (grey). Numbers in each box show mean transcription in TPM.








A
100 kb
P. vivax P01 Chr5: 1.52 Mb
H3K9me3
H3K4me3
H3K9ac
SPZ RNA-seq
BS RNA-seq
B
5659
231
3676
1284
1085
log10 TPM
32.1
0.7
50.3
111.9
111.7
All
H3K9me3
H3K4me3
H3K9ac
H3K9ac.and.K4me3
C
bits
e-value: 2.7x10-2
e-value: 3.3x10-3
D
Sporozoite
Blood-stage
log2 TPM
0.5
1.0
1.5
24.3
1.8
0.5
Stage.enrichment
BS−up
No change
SPZ−up