3 TRANSCRIPTOME PROFILING BEFORE INVASION

3.1 Introduction

Previous work aimed at understanding *P. vivax* merozoite invasion of reticulocytes has focused on a narrow set of genes, almost all of which are homologs of *P. falciparum* genes known to be involved in invasion, primarily because *P. vivax* is very difficult to study directly. There is no established *in vitro* culture system for *P. vivax*, despite repeated efforts by multiple groups, meaning that almost all *P. vivax* research relies on samples taken directly from clinical infections. However, *P. vivax* infections are often of low parasitemia due to the parasite's preference for invading reticulocytes, which limits the amount of material obtainable from clinical samples. A search of PubMed finds nearly 50% (311/637) of publications from 1970-2016 related to *P. vivax* merozoites and/or invasion focus on just 3 proteins: DBP, MSP1, and AMA1. These 3 proteins, and several others with shorter publication lists such as the reticulocyte binding proteins (RBPs), are certainly important for invasion, but there are almost certainly other genes encoding for proteins involved in invasion that are not currently being studied at all. Many of these genes may well be *P. vivax*-specific, and so will not be picked up by the approach of focusing only on those genes with clear homologues in *P. falciparum*. To provide a more unbiased assessment of genes that may be important for *P. vivax* merozoite invasion of reticulocytes, I used RNA-Seq to identify genes upregulated during

the schizont stage of the intraerythrocytic development cycle (IDC), when invasive merozoites are developing.

There are several published transcriptome studies for *P. falciparum* (Bozdech et al., 2003, Otto et al., 2010), which used microarray technology and RNA-Seq respectively (described in section 1.2.2 in detail). Current analyses of *P. vivax* transcription rely primarily on 2 microarray experiments from Bozdech *et al*. (Bozdech et al., 2008) and Westenberger *et al*. (Westenberger et al., 2010). The data in these studies provide a strong foundation for understanding transcription during invasion, but each lacks some information. The Bozdech *et al*. microarray experiment lacks complete genome coverage, and so any genes not present in the initial *P. vivax* Sal 1 reference genome annotation published in 2008 will not have a transcriptional profile. The Westenberger *et al*. microarray experiment covers additional unannotated *P. vivax* genome sections; however, it lacks abundancy comparisons during the asexual life cycle, which is critical for understanding invasion-related transcription. We chose to use RNA-Seq, rather than microarray technology, because it will enable us to produce unbiased transcript abundancy data during the schizont stage that is not constrained by the specific probes used on a microarray chip. RNA-Seq will also allow definition of the boundaries of genes, such that it could uncover novel gene transcripts, alternative splicing events, validate and/or correct current gene models, and predict 5' and 3' untranslated regions. RNA-Seq therefore provides an opportunity to generate a more definitive list of potential invasion associated genes in *P. vivax*.

No RNA-Seq data had yet been published for *P. vivax* at the time of this work, from either clinical isolates or primate models. As *P. vivax* has no reliable long-term *in vitro* culture system, samples could only be obtained from laboratory-controlled primate infections or clinical samples from patients from *P. vivax* endemic areas. Given that we wanted to understand erythrocyte invasion in naturally-occurring human infections, which may differ from erythrocyte invasion in artificially infected primates, samples from human patients were preferable. However, such a study is technically and logistically challenging. Firstly, processing field isolates can be difficult because -80^{o}C storage and transport are not always available. We were aided in this aspect through access to the NIH laboratory field site in Pursat Province, Cambodia (with Rick Fairhurst at LMVR/NIH and Socheat Duong at Cambodian National Center for Parasitology, Entomology, and Malaria Control), which gave us access to *P. vivax* patients and

laboratory facilities to process and ship samples. Secondly, the preference of *P. vivax* for reticulocytes means that parasitemias are regularly under 1%. This leaves little parasite RNA and an overwhelming majority of human RNA (from leukocytes and erythrocytes). We therefore included sample processing steps to remove host leukocytes, and further enrich for the parasite and hence parasite RNA. Finally, *P. vivax* clinical isolates are often asynchronous, with all stages present in a single blood draw. Given that we were most interested in the schizont stage transcriptome for our study, we included an *ex vivo* culturing step to mature the parasites to the schizont stage, followed by Percoll enrichment of schizonts (Figure 3.1 below). A protocol published near the time of our planning stages provided a starting point to use for our study (Russell et al., 2011).

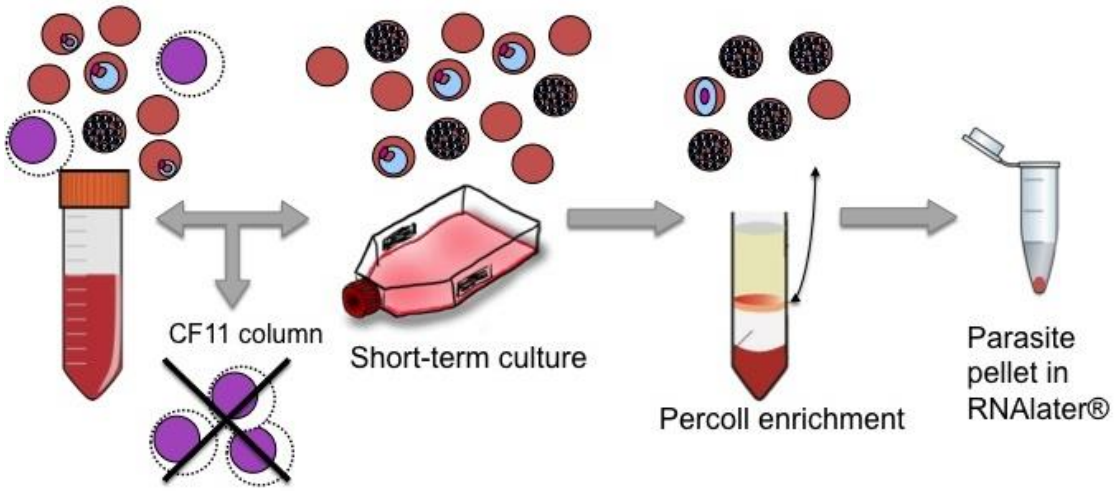

Figure 3.1: Short term *ex vivo* culture of *P. vivax* clinical isolates
The starting material, 16-32 ml of patient blood, contained leukocytes and mixed *P. vivax* life stages. CF11 columns were used to filter leukocytes, which remain in the cellulose matrix as the erythrocytes pass through. Isolates were cultured short-term (11-39 hours) until the majority of parasites were matured to the schizont stage. The culture was enriched for schizonts (with some gametocyte contamination) with a Percoll® gradient. Parasite pellets (up to 200 µl packed cells) were placed in RNAlater® and stored at -20^{o}C. (See section 2.1.3 for greater detail.)

Plasmodium genomes are relatively compact, meaning that the transcribed 5' and 3' UTRs of adjacent genes can overlap. Deconvoluting which gene an RNA-Seq read comes from requires strand-specific Illumina RNA sequencing, but stand-specific RNA-Seq for *Plasmodium* samples was not a standard process during the time of our study. Lia Chappell, a PhD student in the Berriman Laboratory at the WTSI, was actively developing several *Plasmodium* RNA-Seq Illumina library construction protocols at the

start of our study. She provided protocols, training, and guidance for completing the *P. vivax* RNA-Seq libraries before they were sequenced in the WTSI sequencing pipelines (See section 2.1.6 and Figure 3.2 below).

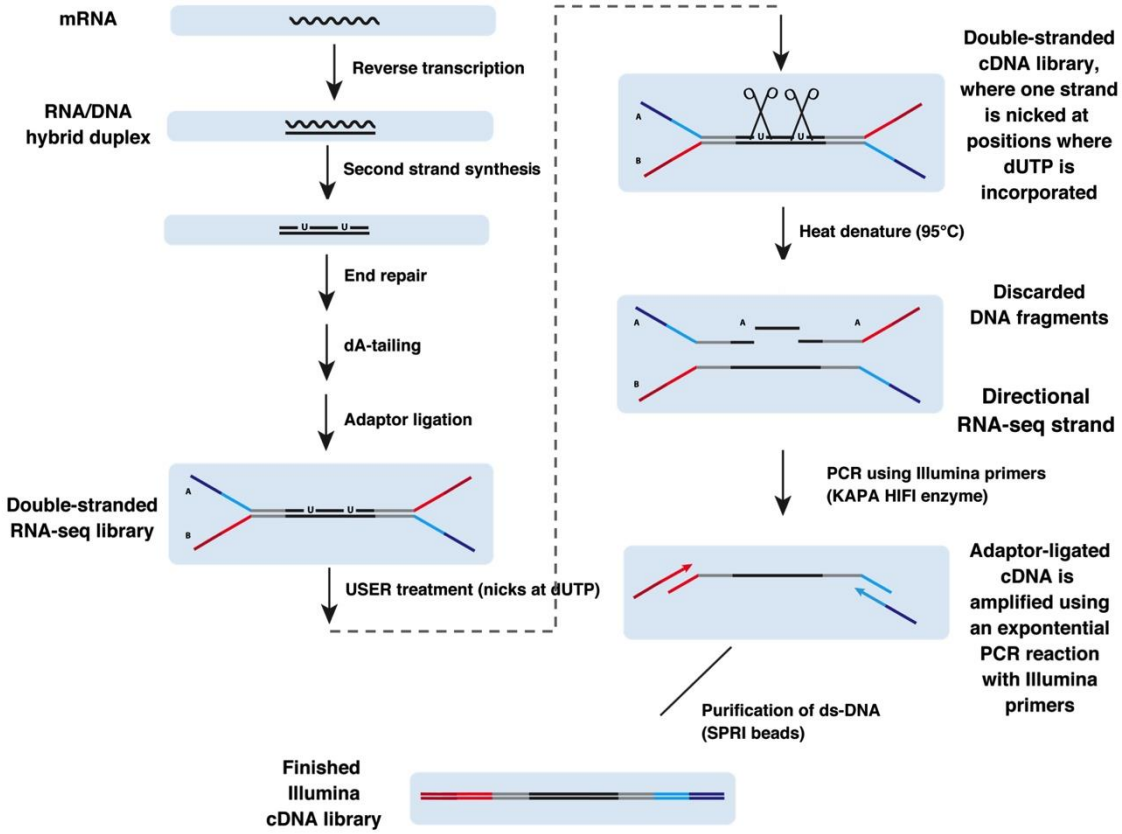

Figure 3.2: RNA-Seq Illumina library construction

Illumina strand-specific library construction protocol developed by Lia Chappell for *Plasmodium* RNA (manuscript in preparation). Messenger RNA (mRNA) transcripts are processed while preserving strand information through the use of second strand synthesis incorporating dUTPs instead of dTTPs. The strand containing dUTPs is ultimately nicked and discarded. An optional final PCR step allows for recovery of samples with concentrations too low for downstream Illumina sequencing (See section 2.1.6 for greater detail). Figure modified from images provided from Lia Chappell.

While obtaining and processing samples was potentially problematic, it was also clear that the downstream data processing and analysis would also be challenging. The *P. vivax* Sal 1 reference genome, published in 2008, represents the *P. vivax* isolate Salvador 1 from an El Salvadoran patient in 1972 that was passed through *Saimiri boliviensis boliviensis* monkeys to generate enough material for whole-genome sequencing (Carlton et al., 2008, Collins et al., 1972). Given its geographical and temporal distance from the

isolates we obtained in Cambodia, there may be significant differences at the genome sequence level. This could substantially hinder read mapping and downstream analysis, particularly in areas of the genome under diversifying selective pressure. Because they are exposed to the adaptive immune system, the merozoite proteins in which we were most interested are particularly likely to suffer from this problem. The MSP3 family serves as a well-known example, where numbers of paralogs differ in different isolates, and large regions are frequently unalignable between isolates (Neafsey et al., 2012, Rice et al., 2014). The *P. vivax* Sal 1 reference genome is also incomplete, particularly at the telomeric ends of chromosomes, and is known to be missing at least 1 gene, EBP or DBL2, which has an erythrocyte-binding domain and may relate to *P. vivax* invasion (Hester et al., 2013). Gene families with high sequence similarities between members, and therefore potentially difficult to assemble properly, are particularly hard to compare. Recently, a new reference genome from a 2014 Indonesian isolate has become available (Auburn et. al, unpublished), containing much more complete telomere assemblies and the most updated gene-model annotation to date. This resource provided a potentially improved assembly to which our sequencing data could be mapped.

3.1.1 Benefits of this study

Transcriptomes of *P. vivax* clinical isolates are not routinely explored, and at the time of this writing, no published RNA-Seq datasets yet exist. Our study aims to benefit the scientific community in several ways. Firstly, the study provides a guide for processing *P. vivax* clinical isolates from collection in the field through Illumina library construction. Secondly, it provides publicly available transcriptome sequencing datasets for *P. vivax* clinical isolates, which will be used for the improvement of reference gene models, as well as evaluating 5' and 3' untranslated regions, novel gene transcripts, and alternative splicing events. Finally, the transcriptomes will inform our understanding of *P. vivax* invasion by identifying genes that are up-regulated in schizonts and enabling inter-isolate comparisons of expression patterns, including the evaluation of multi-gene families known to be important for erythrocyte invasion.

3.1.2 Objectives

i. To identify genes expressed during the *P. vivax* schizont stage, just prior to invasion of reticulocytes.

ii. To provide a data resource for building a library of merozoite surface and/or invasion genes, potential asexual stage vaccine candidates, for further study.

3.2 Results

3.2.1 High-quality RNA extracted from *P. vivax* clinical isolates

Chanaki Amaratunga, a staff scientist in the Fairhurst laboratory, selected 4 Cambodian *P. vivax* clinical isolates with high concentrations of ring and/or trophozoite parasite stages, and then cultured the samples *ex vivo* for 11-38.5 hours in the field (Section 2.1.3, Figure 3.1 above), sending parasite pellets in RNAlater® to me at the WTSI (Table 3.1 below). All patients were males aged 14-32 years.

Table 3.1: Patient and sample profiles for *P. vivax* clinical isolates

ID	Sex	Age	Temp (ºC)	Parasite density/µl				Volume of blood (mL)	Hours in culture	Parasit-emia %
				Ring	Troph	Schiz.	G'cyte			
PV0563	M	14	38	5544	6099	1504	316	16	38.5	0.3
PV0565	M	32	39	4750	12785	1821	678	32	12	0.4
PV0568	M	28	39	6202	6766	669	387	32	11	0.3
PV0417-3	M	21	38.5	5615	8730	461	384	32	18	0.3

Parasite density counts from starting samples counted from thick smears for rings, trophozoite (Troph.), schizonts (Schiz.), and gametocytes (G'cyte).

I first tested RNA extraction methods to determine which method provided both the highest yield and highest quality. I extracted RNA from a laboratory-adapted *P. falciparum* clone using 3 methods: RNeasy Plus Mini kit (Qiagen), RiboPure Blood kit (Ambion), and a TRIzol (Invitrogen) extraction method adapted from Kyes *et al*. (Kyes et al., 2000) (Table 3.2**,** below and Sections 2.1.1 and 2.1.2). Overall, the quality of RNA from each method appeared sufficiently high prior to DNA digestion, but the TRIzol method and RiboPure Blood kit performed best in overall yield and quality after 2 rounds of DNA digestion. The RiboPure Blood kit protocol was faster and did not require the use of a fume hood compared to the TRIzol method tested. The TRIzol method, alternatively, was much more economical for large numbers of samples. Given that only 4 samples

were processed, the RiboPure Blood kit was selected for RNA extraction of the *P. vivax* isolates.

Table 3.2: RNA extraction results

Sample	Method	Before DNA digestion		After DNA digestion	
		RNA concentration (ng/µl)*	RNA integrity number (RIN)*	RNA concentration (ng/µl)	RNA integrity number (RIN)
1	RNeasy Plus Mini kit (Qiagen)	134	9.5	73	NA
2	RiboPure Blood kit (Ambion)	302	9.6	210	9.4
3	TRIzol (Invitrogen)	451	9.4	203	9.2

*RNA concentration and RNA integrity number (RIN) computed by Bioanalyzer® (Agilent Technologies, Inc.) using an RNA Nano Chip. RIN is scored 1-10. NA=Not available.

I extracted RNA from the 4 schizont-enriched *P. vivax* isolates, which had been stored in RNAlater® (Ambion) after collection from patients with *P. vivax* malaria and short-term culture carried out in Pursat, Cambodia (ClinicalTrials.gov Identifier: NCT00663546): PV0417-3, PV0563, PV0565, and PV0568. RNA extraction using the RiboPure Blood kit and 2 rounds of DNA digestion (Section 2.1.4) yielded high-quality RNA with concentrations abundant for downstream RNA library construction **(**Table 3.3 below).

Table 3.3: RNA extraction results after DNA digestion

Patient ID	Blood volume (ml)	Hours in culture	Parasite count (x107/ µl)	RNA conc. (ng/µl)	Total volume (µl)	Total RNA (µg)	RIN*
PV0563	16	38.5	2.815	369	80	29.52	9.8
PV0565	32	11	0.387	133	80	10.64	9.7
PV0568	32	18	7.3	408	80	32.64	10
PV0417-3	32	12	2.975	231	80	18.48	9.8

*RNA concentration and RNA Integrity Number (RIN) computed by Bioanalyzer® (Agilent Technologies, Inc.) using an RNA Nano Chip.

To test for contamination with genomic DNA, I used an aliquot of extracted RNA to make cDNA using the High Capacity cDNA Reverse Transcription Kit (Applied Biosystems). I then used PCR to amplify a region of the *PvDBP* gene (PVX_110810) containing an intron (Section 2.1.5, Figure 3.3 below). The RNA extracts yielded no

discernible bands to indicate genomic DNA contamination; however, all 4 cDNA samples showed both a bright band corresponding to the product with the intron excised (211 bp) and a faint band corresponding to the expected product if the intron was present (346 bp). This indicated either some genomic DNA might still be present and/or a proportion of RNA was incompletely spliced. While this could raise some problems in interpreting sequencing reads, it was judged that the risk of additional loss in quality from a third round of defrosting and digestion was too great, and no further DNA digestion was performed. Potential DNA contamination in the sequencing output was analysed further below.

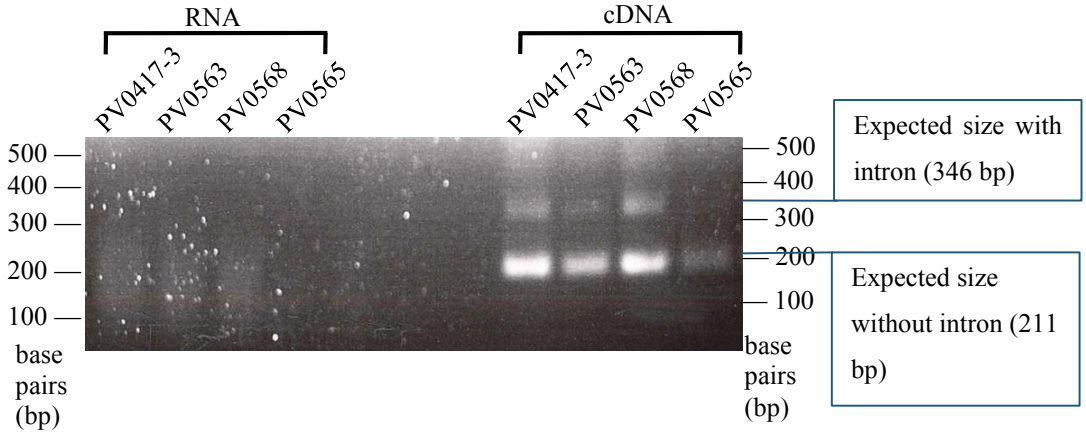

Figure 3.3: Minor DNA contamination and/or incompletely spliced transcripts in *P. vivax* RNA extractions

Polyacrylamide gel showed no discernible bands when testing RNA directly from the 4 Cambodian clinical isolates (PV0417-3, PV0563, PV0568, PV0565) but faint bands were visible after converting RNA to cDNA and amplifying a region of *PvDBP* containing an intron.

3.2.2 Library and mapping statistics

Since the extracted RNA was very high quality (Bioanalyzer RIN > 9.7), I proceeded with the strand-specific Illumina library protocol developed and shared by Lia Chappell with a starting input of 5 µg per sample. The vast majority of the RNA was ribosomal (as with any initial extraction), and therefore, we needed a strategy for its removal. Based on Lia's prior experiments, I removed rRNA through the use of oligo(dT) conjugated to magnetic beads. This step selected for transcripts with poly-A tails, but resulted in the loss of any non-coding RNAs. Future experiments using other methods, such as exonuclease treatment, are planned to study the non-mRNA transcriptome for these

isolates. The yield after the final adapter ligation step necessitated the use of 8 cycles of PCR to reach the DNA quantification required for sequencing on the Illumina HiSeq. The final barcoded strand-specific Illumina RNA sequencing libraries (section 2.1.6) for the 4 *P. vivax* clinical isolates were sequenced together in a single lane on an Illumina HiSeq.

The number of paired-end reads per sample ranged from 55 million to 63 million (Table 3.4). The sequencing and mapping statistics were remarkably similar between the samples, with an average genome coverage ranging from 151x to 186x. The overall amount of ribosomal RNA contamination was very low at 1% per sample, but these rRNA regions corresponded to the regions with the highest coverage overall (maximum coverage, genome-wide in Table 3.4). The samples were also highly pure, with less than 1% human or *P. falciparum* contamination (Figure 3.4).

Table 3.4: RNA-Seq mapping statistics to the *P. vivax* P01 genome

	PV0563	PV0565	PV0568	PV0417-3
Total reads	55720594	52651506	63124660	64316506
Mapped to unique locations[a]	47405586	45030514	53779293	55393526
% Mapped to unique locations[a]	85	86	85	86
Reads mapped to rRNA[b]	833178	656333	606261	818503
% Reads mapped to rRNA[b]	1	1	1	1
Fold coverage[b]	159	151	180	186
% Genome not covered[b]	23	23	23	23
% Genome covered, 1-fold[b]	77	77	77	77
% Genome covered, 20-fold[b]	59	57	62	61
% Genome covered, 100-fold[b]	33	29	34	36
Max. coverage across exon sequences[b]	53426	55738	63269	65164
Max. coverage, genome-wide[b]	99569	81806	177536	69293

[a]Reads mapped using TopHat including only reads mapping to unique locations

[b]Coverage determined using Bedtools

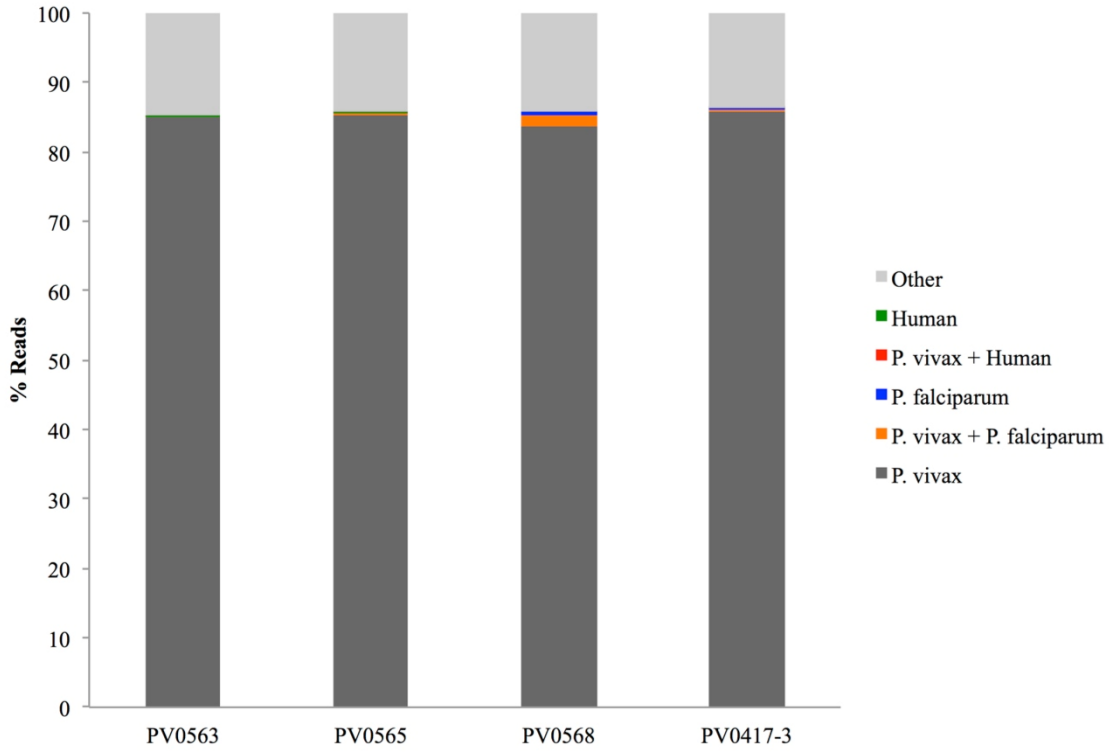

Figure 3.4: Illumina read alignments against the *P. vivax*, *P. falciparum*, and human reference genomes

P. vivax clinical isolate RNA-Seq data contained less than 1% contamination with either human or *P. falciparum* RNA/DNA. A portion of reads, 14-15% per isolate, did not map to *P. vivax* P01, *P. falciparum* 3D7, or human reference genomes.

3.2.3 Assessing DNA contamination

In order to understand the quality of the sequences and assess the amount of genomic DNA contamination, I compared the depths of coverage across 3 types of genomic regions: exons, introns, and intergenic ("other"). The intergenic region includes both true intergenic sequences, but also includes 5' and 3' UTR regions, non-coding sequence regions, and rRNA regions. Some of the intergenic region would therefore still be expected to have a significant amount of sequence coverage in RNA-Seq, whereas coverage of introns should be almost completely absent in sequence reads from RNA, so can be used to measure the level of genomic DNA contamination in the samples. Note that this approach to analysing potential DNA contamination is complicated by the presence of incompletely spliced transcripts, which would produce reads from some introns, but is still a useful proxy.

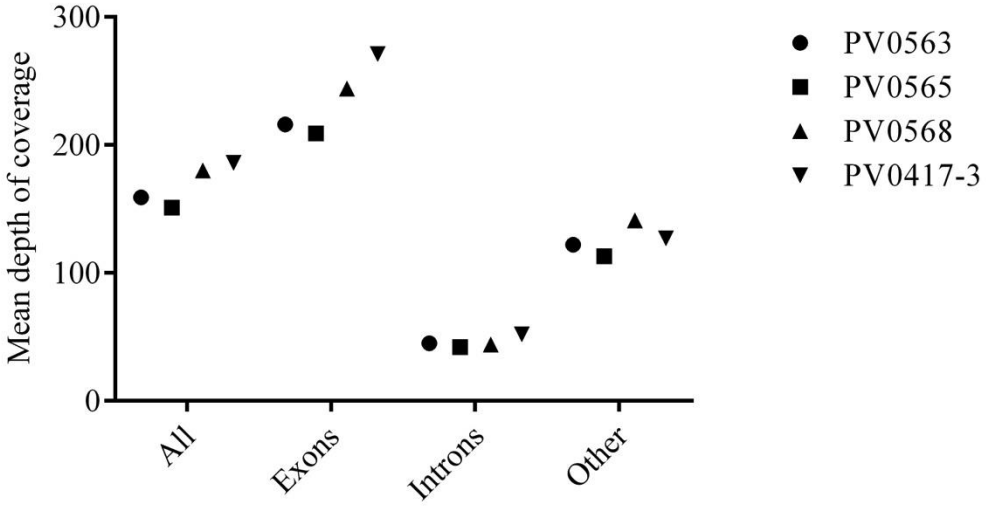

Figure 3.5: Sequence breadth distributions across the *P. vivax* P01 reference genome in exons, introns, and other regions

Mean depth of RNA-Seq coverage for each *P. vivax* clinical isolate in exons, introns, and all remaining regions ("other"). Between isolates, average exon coverage varied from 209- to 271-fold, and average intron coverage varied 42- to 52-fold.

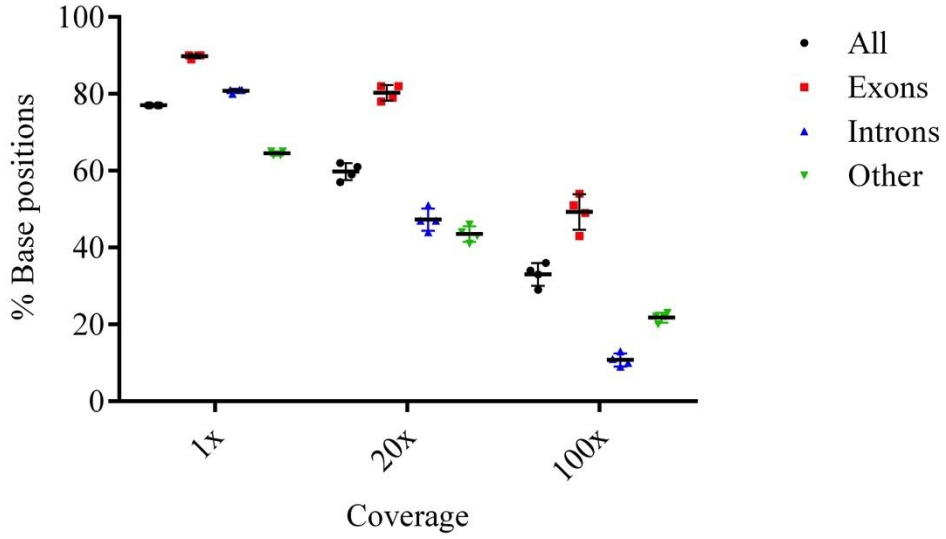

Figure 3.6: Sequence depth distributions across the *P. vivax* P01 reference genome in exons, introns, and other regions

Each isolate is represented by a point, with mean and SD shown. The percentage of bases covered at 1-fold, 20-fold, and 100-fold coverage for exons, introns, and all remaining regions ("other"). Exons consistently had the highest percentage of bases covered at every threshold.

The coverage across exons was about 5 times that of introns on average for each isolate, suggesting that while the majority of reads originated from RNA, some incompletely spliced transcripts and/or genomic DNA may have persisted despite 2 rounds of DNA digestion (Figure 3.5). However, these samples largely represent a single life stage, and will have a highly skewed expression profile, with some genes highly expressed and others not at all. A comparison of the average coverage across introns and exons across the whole genome is therefore not necessarily the most useful approach.

Considering the coverage for each group at low, medium, and high coverage depth cut-offs (1x, 20x, 100x), more exon bases are covered at all the coverage depths (Figure 3.6). At the lowest, 1x threshold, most introns and exon bases are covered (80-90%). At the higher thresholds, however, about 35% more exon bases are consistently covered than intron bases (i.e., at 20x, about 80% of exon bases are covered compared to about 45% of intron bases for each sample; at 100x, about 45% exon bases are covered compared to about 10% of intron bases). This may indicate that genomic DNA contamination is present as 80% of intron bases have some coverage, but the fact that 10% of intron bases have coverage of over 100x (far above the 42-52x average), may indicate that incompletely spliced transcripts play a role in covering introns.

To investigate this more carefully, I compared the average depth of exons and introns for 3 groups: 50 multi-exon genes (~1% of total genome) with the highest, middle, and lowest coverage (setting a minimum coverage depth of 20x, to eliminate any genes with mapping issues rather than true low expression). If genomic DNA represented the primary cause of the coverage of introns, one would expect lower abundance genes to contain a higher ratio of exon-to-intron coverage than higher abundance genes. If intron coverage related primarily to incompletely-spliced transcripts, the exon-to-intron ratio would be similar between the groups. The results suggested that both genomic DNA and incompletely-spliced transcripts contribute to intron coverage (Table 3.5). The exon-to-intron ratio increased from the lowest to highest covered genes, indicative of genomic DNA contamination. However, after the coverage values are normalized by subtracting the average depth of the introns from the lowest covered genes group (a surrogate for the general level of genomic DNA contamination of each sample), the exon-to-intron ratios are much more similar.

Table 3.5: Exon-to-intron coverage comparison for 50 (~1% of the genome) lowest, middle and highest coverage genes

	Average coverage						Exon-to-intron ratio		
	L		M		H		L	M	H
	exon	intron	exon	intron	exon	intron			
PV0563	24	14	174	29	3520	288	1.7	6	12.2
PV0565	22	14	141	28	3518	277	1.6	5	12.7
PV0568	37	20	188	34	3262	267	1.9	5.5	12.2
PV0417.3	37	19	209	35	3746	307	1.9	6	12.2
Normalized									
PV0563	10	0	160	15	3506	274	NA	10.7	12.8
PV0565	8	0	127	14	3504	263	NA	9.1	13.3
PV0568	17	0	168	14	3242	247	NA	12	13.1
PV0417.3	18	0	190	16	3727	288	NA	11.9	12.9

L=lowest, M=Middle, H=Highest; NA=Not applicable

Thus, there appears to be a lower level of genomic DNA contamination contributing 15-20x coverage throughout the genome, or roughly 10% of the average coverage. Above this threshold, coverage in the introns is likely caused by incompletely spliced transcripts. Overall, the isolates appear to have similar levels of both DNA contamination and incompletely spliced transcripts as evidenced by the similar exon-to-intron coverage ratios (1.6-1.9 for genes with lowest coverage; 5-6 for genes with middle coverage; 12.2-12.7 for genes with highest coverage), ensuring that the downstream expression analysis will not be impacted. However, since genomic DNA contamination would mostly impact genes with lower coverage/expression it would be prudent to set a conservative minimum coverage threshold (at least 20x) for including genes in any expression analysis. With the coverage generated in these isolates, this corresponds to a minimum FPKM of 5, and therefore a conservative FPKM lower limit of 10 was set for downstream analysis.

3.2.4 Assessing asexual stage time point

While thin blood smears made after short-term culture and prior to the addition of RNAlater® to the parasite pellets suggested that samples were highly enriched for schizonts, it was also important to test this computationally. In order to assess this, I compared the expression values (FPKMs) for all genes to their one-to-one orthologs in both the published *P. falciparum* microarray dataset (53 time points across the IDC, using enrichment values) and *P. falciparum* RNA-Seq dataset (7 time points across the IDC, using RPKMs) (Figure 3.7). I also compared my data to the published *P. vivax* microarray dataset (9 time points across the IDC, using enrichment values) (Figure 3.8). This work was significantly aided by Lia Chappell who provided an initial file listing one-to-one orthologs between *P. falciparum* 3D7 and *P. vivax* Sal 1 as well as some R scripts.

Pearson correlation coefficients using microarray enrichments values and the clinical isolates' FPKMs showed the highest similarity to the latest time points of the IDC, the schizont stage of the life cycle, for both *P. falciparum* 3D7 and *P. vivax* (Figure 3.7A and Figure 3.8A,C). The correlation to the second *P. vivax* microarray sample SD2 (Figure 3.8B) also showed high similarity to time-point 6, corresponding to the late trophozoite stage. However, this sample correlates less well to both the other 2 *P. vivax* microarray samples and the *P. falciparum* microarray datasets, and may have represented a less synchronous sample (Bozdech et al., 2008). Therefore, this sample is likely a less reliable sample to use for comparisons. The correlations with the *P. falciparum* RNA-Seq time course showed no obvious best match, though with the weakest match to the 16-hour time point corresponding to the early trophozoite (Figure 3.7B). Overall, these comparisons suggested that schizonts are the dominant stage in all 4 clinical samples.

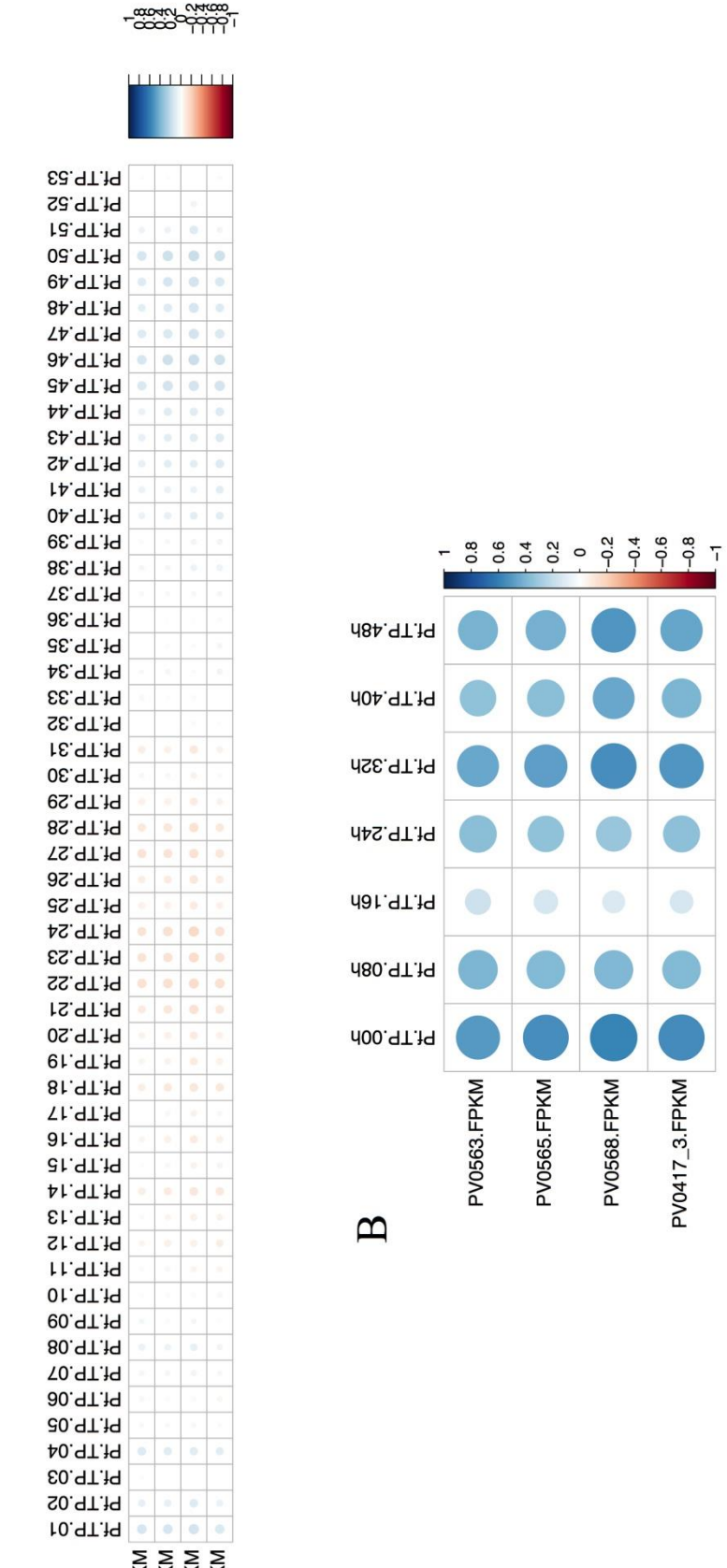

Figure 3.7: Gene expression in *P. vivax* clinical isolates is most similar to the schizont stage in *P. falciparum*

(A) Pearson correlations between the one-to-one orthologs from the *P. falciparum* 3D7 microarray (enrichment values) and the *P. vivax* clinical isolates (FPKMs) showed weak, but best correlations with the late life-cycle time points corresponding to the schizont stage. (B) Pearson correlations coefficients between the normalized *P. falciparum* 3D7 RNA-Seq dataset (RPKMs) and the *P. vivax* clinical isolate RNA-Seq data (FPKMs) showed no best-matching time point, though with the weakest correlations to the early trophozoite stage, 16 hours. *P. vivax* RNA-Seq data mapped to *P. vivax* P01 with TopHat, and FPKMs were calculated using Cufflinks.

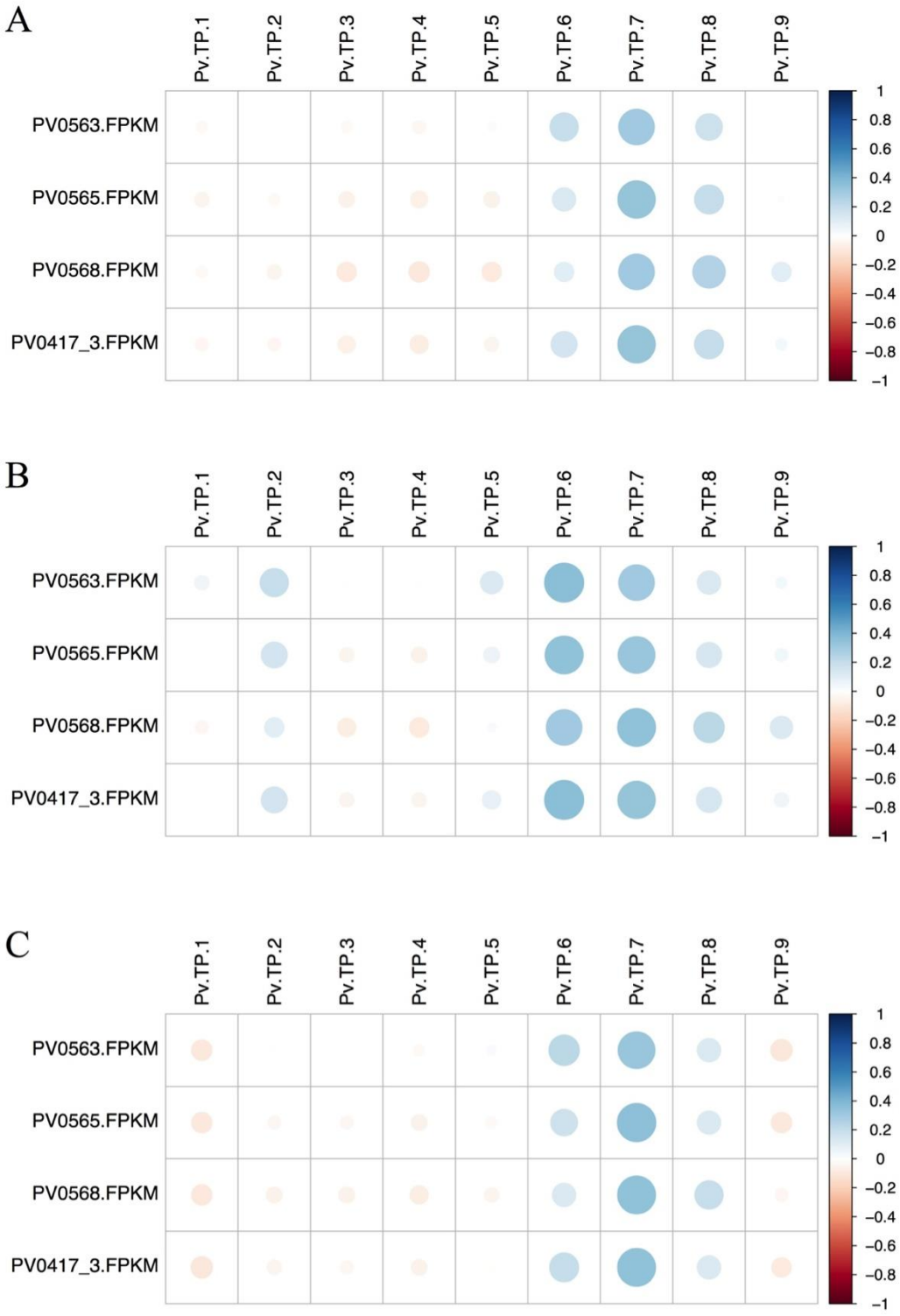

Figure 3.8: *P. vivax* clinical isolates correlate with the early schizont time point from *P. vivax* microarray data

FPKMs from *P. vivax* clinical isolates (PV0563, PV0565, PV0568, PV0417-3) were compared to *P. vivax* microarray data (enrichment values) for 3 isolates (A=SD1, B=SD2, C=SD3) over 9 time points across the 48-hour life cycle using Pearson correlations coefficients. *P. vivax* RNA-Seq data were mapped to *P. vivax* P01 with TopHat, and FPKMs calculated using Cufflinks.

3.2.5 Assessing gametocyte contamination

Thin blood smears made after schizont enrichment established that all samples had some gametocytes present. Significant gametocyte contamination (particularly if very different between samples) could skew the expression analysis, as the gametocyte fraction would diminish the signal of the asexual schizont stage of interest. There are no published RNA-Seq data from *P. vivax* gametocytes, making assessment of the level of gametocyte-specific transcripts challenging. However, RNA-Seq data from several *P. berghei* life stages were available and recently published (Otto et al., 2014). The study contained expression data for 2 biological replicates for the asexual stages (ring, trophozoite, schizont) and gametocytes, and 2 time points for ookinetes. To assess gametocyte contamination, I compared the *P. vivax* isolate's FPKM expression values to all the available *P. berghei* time point FPKMs using Pearson correlation coefficients (Figure 3.9). The *P. vivax* isolates showed the highest correlations with *P. berghei* schizonts, further validating schizont enrichment in the *P. vivax* isolates. All blood-stage comparisons had higher correlations than the gametocyte- and ookinetes-stage comparisons, suggesting that there was no significant gametocyte contamination in the schizont-enriched *P. vivax* samples.

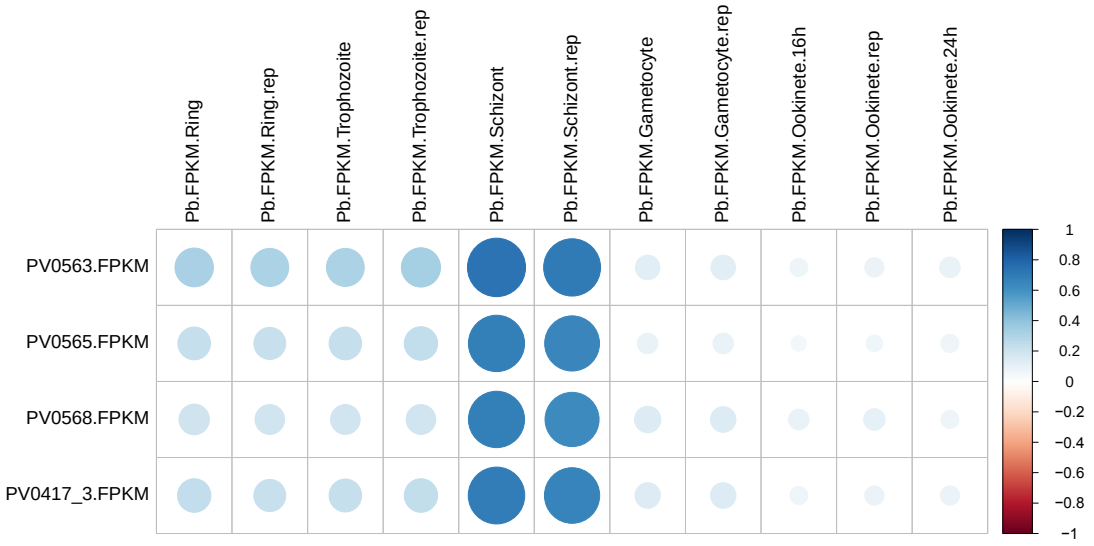

Figure 3.9: *P. vivax* and *P. berghei* schizont RNA-Seq data are highly correlated

FPKMs from *P. vivax* clinical isolates (PV0563, PV0565, PV0568, PV0417-3) were compared to *P. berghei* RNA-Seq data (FPKMs) for 2 replicates (single replicate only for the 24-hour ookinetes stage) over 6 time points during both asexual and sexual stages using Pearson correlation coefficients. *P. vivax* RNA-Seq data were mapped to *P. vivax* P01 with TopHat, and FPKMs calculated using Cufflinks.

3.2.6 Using RNA-Seq data to improve the *P. vivax* reference genome

The *P. vivax* RNA-Seq data was used by Ulrike Böhme in the Parasite Genomics Group at the WTSI to improve 352 gene models in the *P. vivax* Sal 1 reference genome (the only available reference genome at the time of the completion of sequencing). This included detecting 20 novel gene transcripts, merging 7 pairs of genes, splitting 3 genes into 6 separate genes and changing the exons of 350 genes (including adding exons, deleting exons, and changing exon coordinates). No alternative splicing events were detected, however (using the default parameters with Cufflinks tool suite). A detailed list of improvements can be found in Supplementary Table A.

3.2.7 Comparing mapping to *P. vivax* Sal 1 and *P. vivax* P01 reference genomes

Given that 2 reference genomes were available for the analysis phase of this project, I wanted to evaluate the mapping quality to each genome. Running identical assembly parameters, 1% more reads mapped on average to the *P. vivax* Sal 1 reference genome compared to the *P. vivax* P01 reference genome (comparing only the 14 chromosomes) (Table 3.6). Therefore, both reference genomes were fairly similar in terms of overall mapping success. However, the *P. vivax* P01 reference assembly contained more complete chromosomal assemblies (especially in telomere regions) with 24.2 MB assembled compared to *P vivax* Sal 1 with 22.6 MB assembled and linked to chromosomes. The telomeric regions are highly repetitive and this combined with the unique mapping requirement (using TopHat) may partially explain why slightly fewer reads mapped to *P. vivax* P01 than to *P. vivax* Sal 1. Because we were interested in comparing the largest set of annotated genes, the *P. vivax* P01 genome was used for all inter-isolate expression analysis.

Table 3.6: Reads mapping to *P. vivax* Sal 1 vs. *P. vivax* P01

	Reads mapping to chromosomes			
	PV0563	PV0565	PV0568	PV0417-3
P. vivax* Sal 1	47349527	44926798	54397463	55117488
P. vivax* P01	46963376	44669990	53061483	54933162
Total reads	55720594	52651506	63124660	64316506
% Mapped *P. vivax* Sal 1	85	85	86	86
% Mapped *P. vivax* P01	84	85	84	85

*Reads mapped uniquely with TopHat

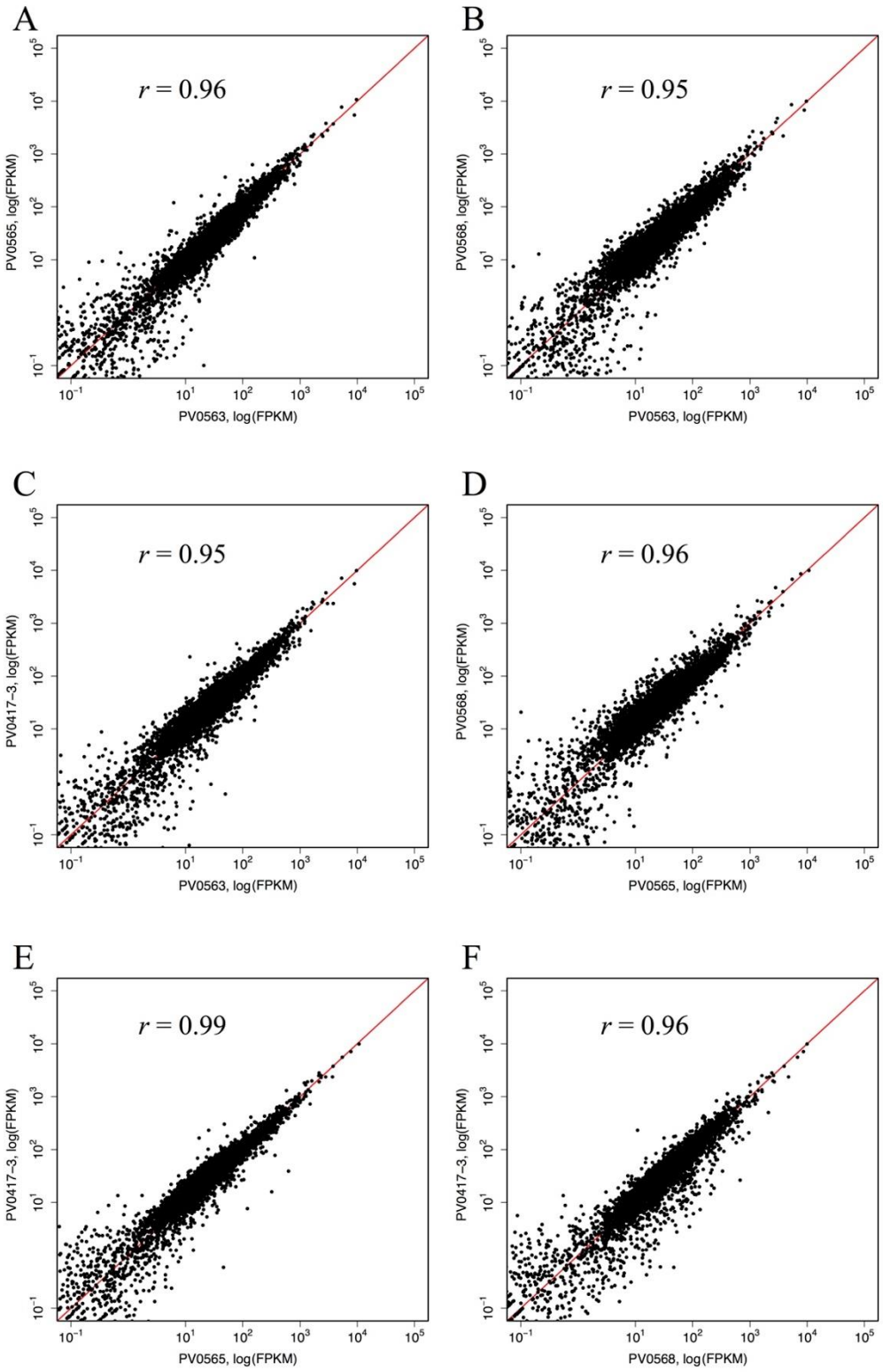

Figure 3.10: Expression in *P. vivax* clinical isolates is highly correlated

(A-F) Pairwise plots of the log(FPKM) for each isolate indicated that the expression profiles between samples were highly correlated with Pearson correlation coefficients (untransformed data) ranging from 0.95 to 0.99. *P. vivax* RNA-Seq data were mapped to *P. vivax* P01 with TopHat, and FPKMs calculated using Cufflinks.

3.2.8 Comparing expression data between clinical isolates at genome scale

In order to understand the how similar (or different) each clinical isolate was to each other, I performed several comparisons. The first was to plot all pair-wise comparisons of the 4 isolates using the log(FPKM) for all genes (Figure 3.10). The Pearson correlation coefficients for each pair (untransformed data) were uniformly high at 0.95 to 0.99, indicating the expression profiles between the isolates were highly similar despite being from separate clinical infections, collected at different times, and subjected to different lengths of *ex vivo* culturing, all of which could potentially impact RNA expression. This high degree of similarity also suggested that the data were not heavily skewed by the gametocyte contamination. This also confirmed that the isolates were matured and collected at very similar time points in the life cycle (despite different lengths of *ex vivo* culture).

Normalising the data to enable clear comparisons was a challenge. While the transcription of both *P. falciparum* and *P. vivax* follows a well-characterized cascade over the 48-hour life cycle with RNA abundance for stage-specific genes peaking and declining in a regular cycle (Bozdech et al., 2003, Bozdech et al., 2008), some transcription for each gene is likely to be detected at every stage with deep sequencing. The ideal way to overcome this problem would have been to sequence control samples that contained all life stages in equal proportion. Such samples would have provided the average RNA abundance for each gene, which could have been used to normalize our data and compare to the enrichment values computed in the *P. falciparum* and *P. vivax* microarray datasets. However, given the difficulty in obtaining even these 4 samples, no such mixed infection control samples could be obtained. This lack of normalization complicates attempts to compare between isolates, and also made it difficult to compare data to the published *Plasmodium* transcriptome datasets. Despite these challenges, some general conclusions could be made, as long as the caveats for comparison are borne in mind.

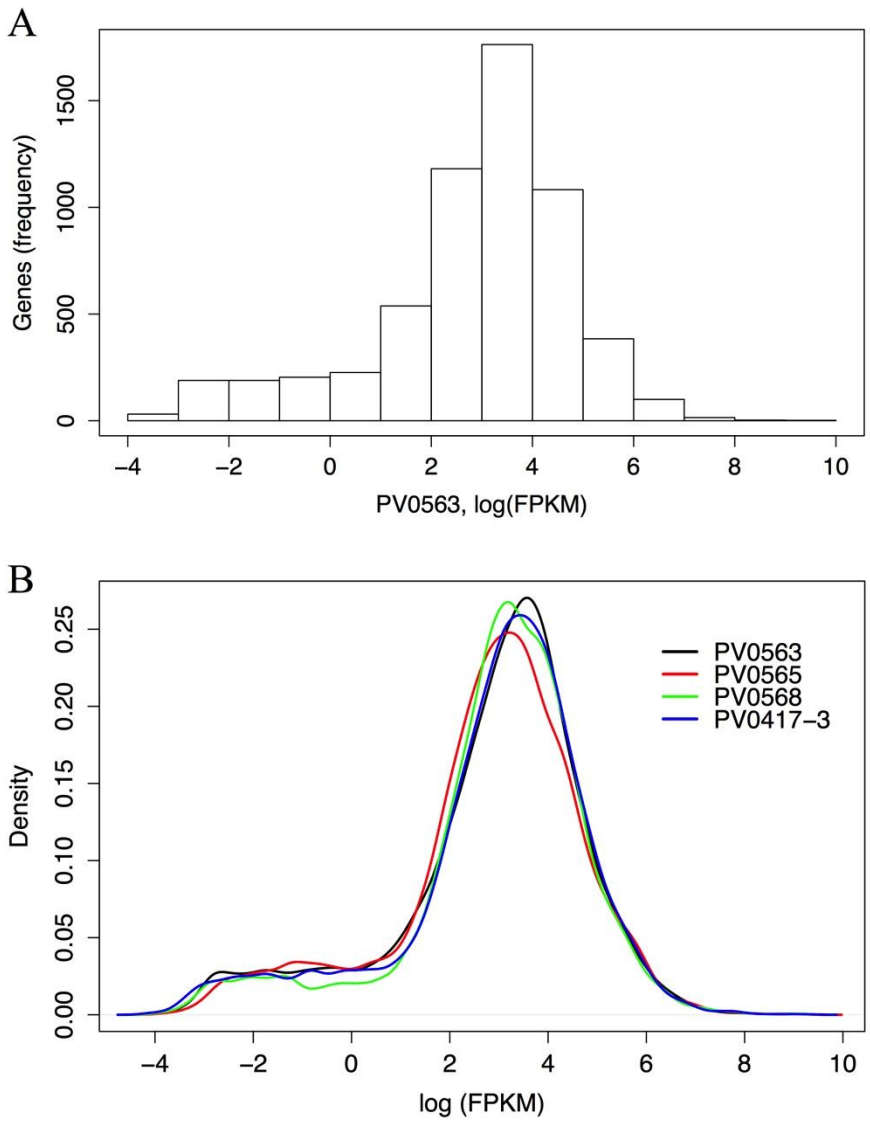

Figure 3.11: FPKM distributions are similar between *P. vivax* clinical isolates

(A) Example histogram with gene frequencies of log(FPKM) for PV0563. (B) Density plots of the log(FPKM) for each of the 4 *P. vivax* clinical isolates (PV0563, PV0565, PV0568, PV0417-3).

The FPKMs for the 4 clinical isolates ranged from 0 to 10,000, with 5290 genes having an FPKM >1. The distribution was very similar in each isolate (Figure 3.11). Over 70% (4704/6672) of the genes had FPKMs ranging from 1 to ~130 (1^e – 6^e). More than 25% of genes had FPKMs of 20 to 43 (3^e – 4^e). The published *P. vivax* microarray data contained expression data (in at least 1 of the 3 isolates) for 4633 genes. This *P. vivax* RNA-Seq dataset contained expression data for over 450-750* additional genes (*minimum FPKM of 10 for inclusion). Gene naming and annotation changes between

the *P. vivax* microarray dataset, the *P. vivax* Sal 1 dataset, and the *P. vivax* P01 reference make reporting an absolute number difficult. 44% of these genes for which transcription data is available for the first time are annotated as conserved hypothetical proteins, for which little is known. Nearly 100 of the complete list of new genes (including both hypotheticals and named genes) have very high expression (FPKMs over 100, or among the top 12% of expressed genes), including high molecular weight rhoptry protein 3, RhopH3, a known invasion-related gene in *P. falciparum* (Sam-Yellowe et al., 1988). The table listing the top-expressed genes unique to the *P. vivax* clinical isolate RNA-Seq data (compared to the published microarray dataset) can be found in Supplementary Table C. These highly expressed genes, for which expression was not previously known, are potentially important for understanding *P. vivax* merozoite invasion of reticulocytes.

3.2.9 Comparing expression between 4 clinical isolates at the individual gene level

While at an overall level the number of expressed genes and FPKM ranges for the 4 clinical isolates was highly similar, I next wanted to evaluate any specific expression differences between the isolates. Standard differential expression analysis compares a set of biological replicates (potentially with technical replicates) to another set of biological replicates after some change in condition (frequently drug pressure) and is commonly performed with tools such as DeSeq or EdgeR (Anders and Huber, 2010, Robinson et al., 2010). In our case we have essentially 4 biological replicates, no technical replicates, and no conditions, as we have no phenotypic categories (e.g., asymptomatic and symptomatic infection) to compare, meaning that applying these approaches is not valid. Looking for differences in our samples is therefore more accurately described as looking at gene expression variability or spread. Before investigating this, I first set a lower FPKM threshold for genes to consider by any method. As discussed above, based on intron coverage indicative of genomic DNA contamination, an initial threshold of 10 FPKMs was set for considering a gene to be expressed. Nearly 64% of genes (4295/6672) have expression over this threshold. However, for this analysis, I was most interested in variable expression among genes that are most relevant in the schizont stage. I therefore considered only the top 50% (3305/6672) of expressed genes, which set a minimum FPKM of 20, in downstream expression variability methods. I then explored variable expression using several different methods.

The first method simply ranked the genes based on how different expression for individual isolates was from the mean expression all 4 isolates, described as the "max fold-change" method. For instance, by comparing all FPKMs to the average FPKM from the group and ranking the genes from highest fold-change to lowest, 56 genes show at least a 2-fold change from the mean in at least 1 isolate. Next I calculated the coefficient of variation (CV) or relative standard deviation (RSD) for the data. This gives a ratio of the standard deviation to the mean for the data, so that we can compare genes with very different means, as is the case with the RNA-Seq data where FPKMs range from 0 to 10000. The third normalized measure of spread is the index of dispersion or variance-to-mean ratio (VMR). This is similar to the first step of the DeSEQ differential expression analysis, which assesses the dispersion (or over-dispersion) of the data for a set of biological replicates, in order to gauge the normal variability or "noisiness" of the biological replicates which will be compared to another condition (Anders and Huber, 2010).

In order to compare how similar the rankings were between the several methods, I arbitrarily compared the top 300 genes (or around 10% of the total included genes) ranked by each method. The 'max fold-change' and CV calculations were very similar, as evidenced by the 2 sets overlapping for 256/344 genes (76%) (Figure 3.12). The VMR ranking differed most from the other 2 methods, and overall 130/495 genes (26%) were ranked in the top 300 by all 3 methods. These 130 genes could be considered the most reliable set of variably-expressed genes for further analysis.

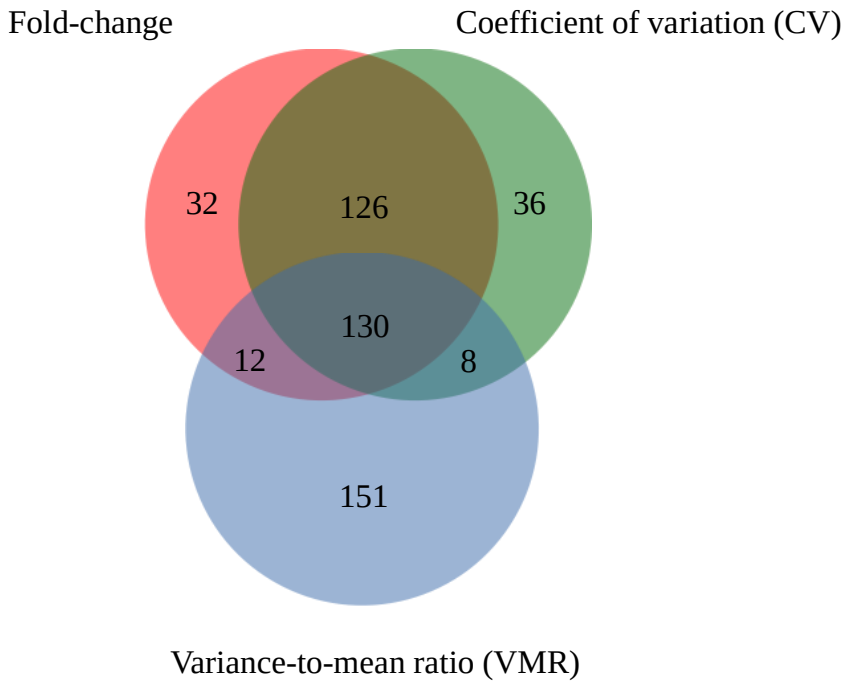

Figure 3.12: Three methods for ranking expression variability in *P. vivax* clinical isolates

Evaluating expression variability between *P. vivax* clinical isolates using fold-change from the mean and 2 normalized measurements of dispersion (coefficient of variation, CV; variance-to-mean ratio, VMR), showed modest agreement, with 26% of genes (130/495) overlapping between the top 300-ranked genes in each approach (setting a minimum average FPKM of 20). The fold-change and CV calculations intersected for 76% of genes (256/344), and the VMR ranking had the least overlap with 31% (142/459) and 30% (138/463) when compared to fold-change and CV, respectively. *P. vivax* RNA-Seq data were mapped to *P. vivax* P01 with TopHat, and FPKMs calculated using Cufflinks.

The 130 genes ranked in the top 300 by each of these 3 methods are summarized in Supplementary Table B. At least 26% of the genes were surface-expressed or invasion-related and/or members of large families, including AMA1, DBP, GAMA, 7 MSP3s, 7 RBPs, and 3 MSP7s. Another 40% (51/129), were either "conserved *Plasmodium* protein, unknown function," "hypothetical protein," or "*Plasmodium* exported protein, unknown function," in keeping with the large number of these genes in the *P. vivax* genome. None of the top 20 differentially expressed *P. berghei* gametocyte genes were present in the top 130 variably-expressed gene set, although 2 variably-expressed genes (PVP01_1258000, gamete egress and sporozoite traversal protein; PVP01_0115300, gamete release protein, putative) did appear to be related to the sexual stage. At least 1 gene from the *P. falciparum* gametocyte expression study (Young et al., 2005), a putative oxidoreductase,

PVP01_1229400 (PF3D7_1325200) was also in the variably-expressed gene set, possibly indicating there were slightly different proportions of gametocytes in the samples. I analysed the list for gene ontology (GO) enrichment using the 77 of these 130 genes that had direct *P. falciparum* 3D7 one-to-one homologs (and therefore for which GO terms were available) and found significant enrichment for invasion and host interacting genes (Table 3.7 below). This is an underestimate of the actual enrichment, as several large *P. vivax* families in the top 130 variably-expressed genes related to the merozoite surface (MSP3 and MSP7) and/or invasion (RBPs) had no direct one-to-one orthologs with *P. falciparum*, and thus had no GO terms.

Table 3.7: Host and invasion genes enriched in top 130 variably-expressed genes from *P. vivax* schizont-stage clinical isolates

GO molecular function	Pf. ref (5159)	Var set (77)	Exp'd	+/-	Fold Enrich.	*P* value
intramolecular oxidoreductase activity (GO:0016860)	2	2	0.03	+	> 5	2.89E-02
host cell surface binding (GO:0046812)	19	5	0.28	+	> 5	7.19E-04
protein binding (GO:0005515)	146	12	2.18	+	> 5	1.17E-04
binding (GO:0005488)	199	15	2.97	+	> 5	1.52E-05
molecular_function (GO:0003674)	391	18	5.84	+	3.08	9.50E-04
Unclassified (UNCLASSIFIED)	4768	60	71.16	-	0.84	0.00E+00
GO biological process						
interaction with host (GO:0051701)	47	5	0.7	+	> 5	4.75E-02
biological_process (GO:0008150)	313	13	4.67	+	2.78	4.67E-02
Unclassified (UNCLASSIFIED)	4846	65	72.33	-	0.9	0.00E+00
GO cellular component						
microneme (GO:0020009)	20	6	0.3	+	> 5	4.26E-05
pellicle (GO:0020039)	18	5	0.27	+	> 5	5.55E-04
inner membrane complex (GO:0070258)	18	5	0.27	+	> 5	5.55E-04
apical complex (GO:0020007)	57	10	0.85	+	> 5	1.02E-06
apical part of cell (GO:0045177)	59	10	0.88	+	> 5	1.40E-06
membrane (GO:0016020)	89	8	1.33	+	> 5	3.84E-03
cell (GO:0005623)	560	20	8.36	+	2.39	1.07E-02
cell part (GO:0044464)	559	19	8.34	+	2.28	3.07E-02
cellular_component (GO:0005575)	733	23	10.94	+	2.1	2.10E-02
Unclassified (UNCLASSIFIED)	4426	55	66.06	-	0.83	0.00E+00

Gene Ontology (GO) ids from experimental evidence published in scientific literature. Pf. Ref = *P. falciparum* 3D7. Var set (77) = one-to-one orthologs between *P. falciparum* and *P. vivax* P01 for a set of most variably-expressed genes during the schizont stage for 4 clinical isolates. Exp'd=Expected fraction. Fold enrich=Fold enrichment. *P* values <0.05.

3.2.10 Assessing the impact of diversity and mapping on expression data

A number of the genes identified as variably expressed were members of highly-diverse large gene families, including 7 members of the MSP3 family. MSP3 family members are known to be under heavy diversifying selection (Neafsey et al., 2012, Rice et al., 2014). It was highly possible that such diversity meant that the "expression variability" was due to mapping differences between the isolates rather than actual expression differences.

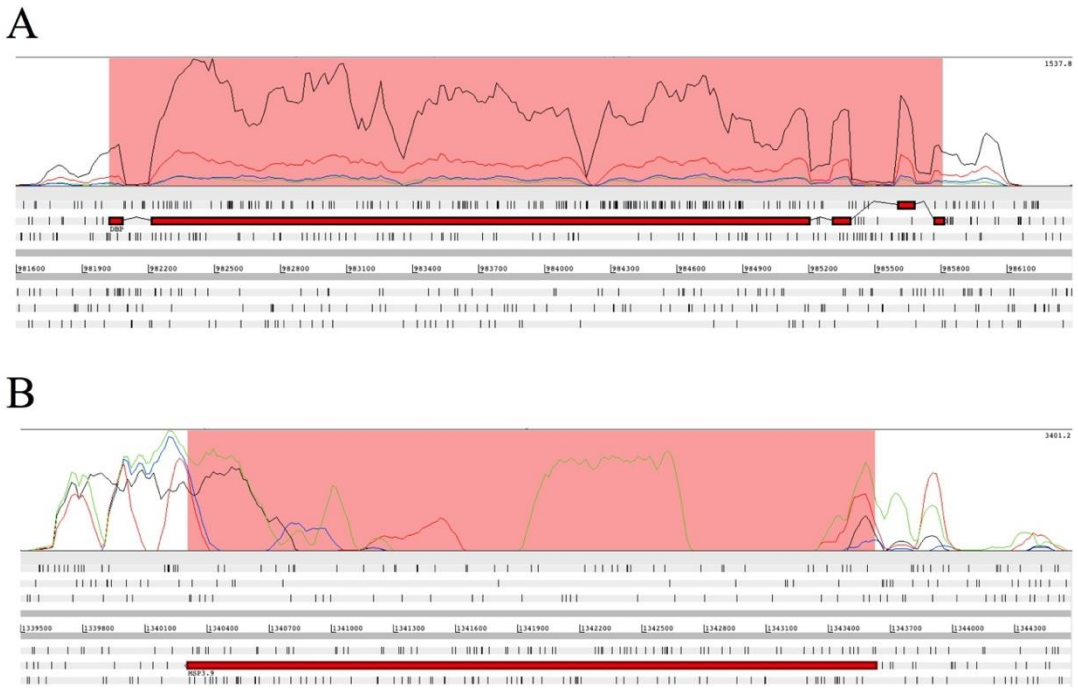

Figure 3.13: Variable expression in *P. vivax* clinical isolates

(A,B) Artemis screenshots showing 2 examples of genes ranked as variably-expressed. RNA-Seq read coverage spans the top half of each panel, with each isolate represented by a different colored line (PV0563=black, PV0565=red, PV0568=blue, PV0417-3=green). The bottom half of each panel shows the 3 forward reading frames (above the chromosomal coordinates) and the 3 reverse reading frames (below the chromosomal coordinates), with stop codons denoted by black vertical bars and gene exons in large red boxes. (A) PVPVP01_0623800, the Duffy Binding Protein (DBP), has mostly even mapping coverage throughout the gene for all 4 isolates, while PV0563 (black) has significantly higher coverage. (B) PVP01_1031200, MSP3.9, has large regions with no coverage in one or more isolates indicating either repetitive regions to which reads could not be uniquely mapped and/or divergence from the *P. vivax* P01 reference.

I reviewed the mapping results for each of the top 130 variably-expressed genes visually using Artemis (Carver et al., 2012, Rutherford et al., 2000) and found that 87% (113/130) contained coverage across the entire gene in all 4 samples. Only 4% (5/130) contained a drop in coverage shared by all 4 samples, and 9% (12/130) had large regions lacking coverage in 1 or more of the isolates, which is indicative of a region of divergence (example shown in Figure 3.13B and noted in Supplementary Table B). Relatively few of the variably-expressed genes can therefore be ascribed to mapping issues, and almost all the mapping issues were from the MSP3 and MSP7 gene families. One RBP showed large gaps in alignment, whereas the other 6 variably-expressed RBPs showed coverage throughout. None of the genes with potential mapping issues had one-to-one orthologs in *P. falciparum*, indicating that our results for GO term enrichment were still valid.

3.3 Discussion

In this chapter, I described the progress toward the goal of studying transcription just prior to *P. vivax* merozoite invasion of erythrocytes with the generation of RNA-Seq data from the schizont stage of 4 Cambodian *P. vivax* clinical isolates. This relied on obtaining a large quantity of high-quality RNA from *P. vivax* samples enriched for schizonts. Starting levels of parasites in all 4 patients were very low at 0.3-0.4%, suggesting that our relatively large blood draws were important for success. The RNA yields after extraction and 2 rounds of DNA digestion were very high, ranging from 10 to 33 µg. At the time the Illumina library protocol required 5 µg of starting RNA, and our high RNA yields meant that multiple attempts could be made from each sample, if needed. The starting material required for library construction has declined 10-fold since we made our libraries, which will make future studies possible with much smaller volumes of patient blood (a limiting factor for many field site protocols). This also raises the possibility of collecting multiple time points from *ex vivo* cultured samples, enabling us to collect transcriptional data throughout the IDC for Cambodian isolates.

RNA-Seq data quantity and quality was similarly high for the 4 isolates. Our schizont-stage sequencing is much deeper than the *P. falciparum* RNA-Seq analogous time point, with over 5 times as many reads mapped in our study compared to the *P. falciparum* RNA-Seq IDC study [Table S2, (Otto et al., 2010)]. The samples contained very little *P. falciparum* or human contamination (under 1% each), validating that these were *P. vivax*

mono-infections and that the leucocyte depletions by CF11 columns were highly effective. It also indicated that our enrichment for schizont-infected erythrocytes by Percoll® largely removed uninfected erythrocytes, which are known to contain high quantities of RNA (Kabanova et al., 2009). The mRNA selection using oligo(dT) magnetic beads was also very effective at removing ribosomal RNA, which accounted for ~1% of the overall data. Thus, our overall laboratory processes were very effective for isolating high quality RNA that was nearly free of contamination and targeted mRNA almost completely. This process will serve as a useful guide for future field studies both within and outside our group.

The average coverage over exons was consistently about 5 times that of introns (209x-271x for exons compared to 42x-52x for introns). The coverage of introns may represent sequencing of incompletely spliced transcripts and/or the presence of contaminating genomic DNA. I investigated this further by comparing the exon and intron coverage for 1% of genes with the highest, middle, and lowest expression. The results indicated that genomic DNA contamination accounted for 15-20x of the overall coverage in each isolate, but that coverage of introns above this threshold was primarily due to incomplete splicing of transcripts. Based on the sequencing depth of each isolate, I set a conservative lower limit of 10 FPKMs for inclusion as an "expressed" gene in downstream analyses. Overall, the similarity in the ratio between exon and intron coverage across all samples suggested that our inter-isolate comparisons were not likely to be affected, even without setting a conservative FPKM threshold.

Comparisons to the published microarray IDC studies for both *P. falciparum* and *P. vivax* confirmed that our samples correlated best to the early schizont stage of the IDC. While the comparison between our samples and the *P. falciparum* RNA-Seq data did not produce a best-matching time point, this may be due to technical issues rather than biological differences. It is noted by the authors that the percentage of uniquely mapping reads was lower at the schizont stage than for any other stage, potentially indicating a loss of transcripts due to high A-T content and low complexity sequence, and no Pearson correlation coefficient comparing the *P. falciparum* microarray and RNA-Seq is reported for the likely best, 40-hour time point (Otto et al., 2010). The *P. falciparum* experiment also used a combination of exonuclease and specific oligos for depletion of rRNAs compared to our oligo(dT) selection of mRNA. It is possible that a bias in sequencing at the schizont stage combined with the differing rRNA depletion methods have reduced the

correlations between the 2 RNA-Seq datasets. In contrast, the comparison with *P. berghei* RNA-Seq for several asexual and sexual stages represented the strongest evidence (with highest Pearson correlation coefficients) that the *P. vivax* clinical isolates represent the schizont stage of the IDC with very little gametocyte contamination.

The RNA-Seq data mapped very similarly to the 2 available reference genomes, *P. vivax* Sal 1 and *P. vivax* P01. On average, 1% more reads mapped to the *P. vivax* Sal 1 genome, which contains about 1.6 fewer Mb of chromosome compared to *P. vivax* P01. One therefore might have expected more data to map to the genome with a longer assembly. This loss in mapping percentage may relate to the stringent mapping parameters I used (only unique placements allowed), as the newest reference is greatly expanded in the repetitive telomeric regions. A more detailed sequence comparison, looking at SNPs between each isolate and each reference, for example, would further define which reference genome is best suited to the Cambodian RNA-Seq dataset. Despite the slightly lower overall mapping, given that the *P. vivax* P01 genome contained over 1000 additionally annotated genes, it was deemed the best reference for subsequent inter-isolate comparisons.

The overall expression of the 4 isolates was highly correlated; pairwise isolate comparisons computed Pearson correlation coefficients ranging from 0.95 to 0.99. This is striking as the samples infected hosts of different ages with potentially different immune statuses, were cultured *ex vivo* for different lengths of time, and underwent a series of laboratory processes from RNA extraction through Illumina library construction, all of which might have affected the resulting comparisons. This overall high similarity in expression values between samples suggests that the lab processing of the samples was very similar and introduced limited technical noise. Critically, the samples were cultured *ex vivo* to the nearly identical stage, early schizonts, before Percoll® enrichment. These high correlations also provided additional support that intron coverage and gametocyte contamination did not greatly skew expression results. It also supports that the transcription during the schizont stage appears to be very similar among different circulating isolates in Cambodia.

I searched for genes with the most dispersed or variable expression using 3 methods and considered in detail those genes that ranked in the top 300 most variably-expressed genes of all 3 methods. Such a cut-off was arbitrary, but was useful to evaluate whether the

genes with the highest variability appeared to be enriched in any way. Assessing GO enrichment using GO terms based on *P. falciparum* one-to-one orthologs, showed a greater than 5-fold enrichment with functions relating to intramolecular oxidoreductase activity, host cell surface binding, protein binding and cellular locations relating to the microneme, pellicle, inner membrane complex, apical complex or apical part of cell, and membrane; almost all of these GO terms relate to invasion in some way. The most variably-expressed genes appeared to be enriched for host interacting genes and locate to the apical end or in invasion-related organelles of merozoites, which might indicate that the isolates were responding to the host environment through the modulation of expression of genes involved in parasite invasion of reticulocytes. Given that several of these variably-expressed genes were members of multi-gene families, such as the RBPs, modulating expression of family members might enable a parasite to evade the immune system and/or improve invasion efficiency. This requires much more study and consideration, however, as it is also likely such genes are overrepresented in general in the schizont stage compared to the entire genome, and may just reflect the genes most likely to be in the schizont transcriptome. Performing a GO enrichment analysis using a curated set of genes combining those expressed throughout the IDC and with peak schizont-stage expression will help to clarify this. One might also hypothesize that the expression variability relates to the efficiency with which each isolate produced merozoites, such that a higher merozoite to schizont ratio might appear as variable expression. Given that the expression for the vast majority of the genes were highly similar, and that there was no pattern of expression ranking (i.e. no single isolate with highest expression throughout the "variably-expressed" list), this seems unlikely.

The vast majority of genes identified as variably-expressed had reads mapping for the entire length of the gene in all 4 isolates, including all of the genes included in the GO enrichment analysis, establishing that differences in expression were not due to mapping artefacts. However, all the variably-expressed members of the MSP3 and MSP7 gene families, as well as a single RBP gene, contained large gaps in alignment for at least 1 underlying isolate indicating either the region was repetitive and reads could not be uniquely mapped or that the reference was highly divergent from the clinical isolates in these areas. *De novo* assemblies will be needed to further investigate expression for these gene families.

3.3.1 Limitations and future work

The biggest limitation of our dataset was having access to only a single life stage of the IDC. In order to understand the full relevance of transcription in our samples, we need a baseline level of transcription for each gene with which to compare the schizont-stage results. This can likely be addressed computationally to some degree requesting the raw intensity values from the *P. vivax* microarray dataset (Bozdech et al., 2008), but would ideally be done through RNA sequencing of a sample with a relatively even mixture of IDC stages at a minimum. The best comparison would be made through completing a study of the complete IDC in Cambodian field isolates. The recent significant reductions in the amount of material needed for Illumina library construction make this highly feasible in future studies.

Shortly before the submission of this dissertation, an RNA-Seq study describing the IDC of *P. vivax* was published (Zhu et al., 2016). The study produced non-stranded RNA-Seq data for 2 Southeast Asian *P. vivax* clinical isolates over 9 time points of the IDC, from the same samples published in the original *P. vivax* IDC microarray study (Bozdech et al., 2008). The study was able to expand on the microarray conclusions to comment on the IDC transcription using the most current *P. vivax* Sal 1 reference annotation as well as non-coding RNAs, long UTRs, and more. These published *P. vivax* RNA-Seq data will be a valuable resource to which the data described in this chapter can be compared. It can be utilized for normalizing the observed expression to best understand which genes are enriched during the schizont stage in naturally-circulating Cambodian clones. As we prepare these data for publication, we will compare and contrast expression between the 2 Thai isolates in the paper and the 4 Cambodian clinical isolates in our dataset. There are also significant features of the Cambodian analysis, which are not captured in the published Thai study. First, we compare mapping and expression data using both *P. vivax* Sal 1 and the new *P. vivax* P01 reference genomes, the latter of which contains over 1000 new annotated genes and was not used in the Thai study. Secondly, our RNA libraries are strand-specific, allowing identification of 5’ and 3’ UTRs and both sense and antisense transcripts. These findings will therefore add significantly to our understanding of *P. vivax* transcription, even if these data will not now be the first *P. vivax* RNA-Seq data published.

3.4 Conclusion

To characterize gene expression in the schizont stage of *P. vivax* parasites, which contain invasive merozoites and are therefore the most likely source of new blood-stage vaccine targets, we sequenced the transcriptome of 4 *P. vivax* clinical isolates from Pursat Province, Cambodia. We tested 3 RNA extraction methods prior to extracting RNA from merozoite-containing schizonts that were purified using a Percoll® gradient and stored in RNAlater® under field conditions. Strand-specific libraries using only 8 cycles of PCR were generated for Illumina sequencing and mapped to the *P. vivax* Sal 1 and *P. vivax* P01 reference genomes. RNA-Seq data from the clinical isolates correlated most closely with schizont stages from published microarray data from *P. vivax* and *P. falciparum*. This study produced genome-wide unbiased transcript abundance data and enabled the correction of more than 300 gene models. The data showed that expression between the clinical isolates was highly correlated, and the few genes with variable expression were enriched for merozoite surface/invasion genes and or members of large gene families, potentially pointing to differential transcription of genes in response to the host environment. Overall, the production of *P. vivax* schizont stage RNA-Seq data provided a valuable resource to consult for the next stage of my project, the creation of a *P. vivax* merozoite protein library for functional studies.