

# RNA-Seq analysis of splicing in *Plasmodium falciparum* uncovers new splice junctions, alternative splicing and splicing of antisense transcripts

Katherine Sorber<sup>1</sup>, Michelle T. Dimon<sup>1,2</sup> and Joseph L. DeRisi<sup>1,3,\*</sup>

<sup>1</sup>Department of Biochemistry and Biophysics, <sup>2</sup>Biological and Medical Informatics Program, University of California San Francisco, San Francisco, CA and <sup>3</sup>Howard Hughes Medical Institute, Chevy Chase, MD, USA

Received September 27, 2010; Revised November 5, 2010; Accepted November 10, 2010

## ABSTRACT

Over 50% of genes in *Plasmodium falciparum*, the deadliest human malaria parasite, contain predicted introns, yet experimental characterization of splicing in this organism remains incomplete. We present here a transcriptome-wide characterization of intraerythrocytic splicing events, as captured by RNA-Seq data from four timepoints of a single highly synchronous culture. Gene model-independent analysis of these data in conjunction with publically available RNA-Seq data with HMMSplicer, an in-house developed splice site detection algorithm, revealed a total of 977 new 5' GU-AG 3' and 5 new 5' GC-AG 3' junctions absent from gene models and ESTs (11% increase to the current annotation). In addition, 310 alternative splicing events were detected in 254 (4.5%) genes, most of which truncate open reading frames. Splicing events antisense to gene models were also detected, revealing complex transcriptional arrangements within the parasite's transcriptome. Interestingly, antisense introns overlap sense introns more than would be expected by chance, perhaps indicating a functional relationship between overlapping transcripts or an inherent organizational property of the transcriptome. Independent experimental validation confirmed over 30 new antisense and alternative junctions. Thus, this largest assemblage of new and alternative splicing events to date in *Plasmodium falciparum* provides a more precise, dynamic view of the parasite's transcriptome.

## INTRODUCTION

Close to one million people every year are killed by malaria, an infectious disease caused by protozoan parasites of the genus *Plasmodium* (World Malaria Report 2009 [http://www.who.int/malaria/world\\_malaria\\_report\\_2009/en/index.html](http://www.who.int/malaria/world_malaria_report_2009/en/index.html)), of which *Plasmodium falciparum* is the deadliest. In efforts to understand the parasite's basic biology and discover unique vulnerabilities, several studies have detailed transcriptome-wide RNA expression data during various parasite lifestages (1–3). However, although more than half of the parasite's genes are predicted to contain introns (4), no specific transcriptome-wide analysis of splicing in this organism has been performed to date. Splicing, the mechanism by which intronic sequences are removed and exonic sequences are joined together, not only determines the protein coding or functional RNA sequence of a mature transcript but also the regulatory information included in the transcript. Alternative splicing adds an additional layer of complexity by allowing the generation of different mature transcripts from the same precursor, and is crucial to such diverse biology as *Drosophila* sex determination and *HIV-1* replication (5,6). Thus, a transcriptome-wide picture of splicing and alternative splicing in *P. falciparum* is crucial for recognizing the full regulatory, protein encoding and functional RNA encoding complexities of the transcriptome.

Although the molecular mechanism of RNA splicing remains murky in *P. falciparum*, it has been well studied in model organisms. In the classical pathway, two transesterification steps are catalyzed by the spliceosome, a large complex of small nuclear ribonucleoproteins (snRNPs), each containing an snRNA component and a core set of proteins. snRNAs have been documented in *P. falciparum* (7,8), but only one protein component, a UAP56 homolog, has been definitively identified (9).

\*To whom correspondence should be addressed. Tel: (415) 418 3059; Fax: (415) 514 2073; Email: joe@derisilab.ucsf.edu

As with splicing components, elucidation of the motifs guiding splicing also remains incomplete. Typically these motifs include the 5'-splice site (AG|GUAUGU in yeast, AG|GURAGU in mammals), the branch point sequence (UACU AAC in yeast, YNYURAY in mammals), the poly-pyrimidine tract (variable length in both yeast and mammals) and the 3'-splice site (CAG| in yeast, YAG| in mammals) (10). In *P. falciparum*, EST data have been used to generate putative 5' (AR|GUAANW) and 3' (YAG|) splice site motifs (7). As in most eukaryotes, the first and last 2 nt of the intron (5' GU-AG 3') are the most consistent markers of intronic sequence. In other organisms, a minority of introns are marked by non-canonical splice sites such as 5' GC-AG 3' (recognized by the major U2-type spliceosome) and 5' AU-AC 3' (recognized by the minor U12-type spliceosome) (11). Non-canonical splice sites occur in *P. falciparum* EST data (12,13) and have been incorporated into some gene models, yet no study to date has documented the types of intron boundaries recognized by the parasite.

Alternative splicing, in which the same precursor transcript can give rise to multiple different mature transcripts, also occurs in the parasite. Although relatively little is known about splicing in general in *P. falciparum*, more than 100 alternative splicing events have been reported in *Plasmodium* species since 1991 (14–20). Alternatively spliced isoforms have also been computationally predicted, yet lack experimental validation (21).

Recent analyses have shown that transcriptome complexity in many organisms extends beyond alternative splicing. Dense transcriptional arrangements, such as overlapping protein-coding genes (in parallel or antiparallel orientation) and natural antisense transcripts (22,23), now appear to be commonplace rather than anomalous. Although the functional importance of these arrangements is not yet well understood, some are known to be important in regulatory relationships between the paired genes (24). In current *P. falciparum* gene models, six instances of protein-coding gene overlap are annotated, resulting in one parallel and five antiparallel gene pairs. In addition, RNA polymerase II has been shown to synthesize long antisense transcripts in the parasite (25), and EST data indicate that at least one of these may be spliced (12). Short antisense transcripts have also been described (26).

In this study, RNA-Seq data were generated from four timepoints in the intraerythrocytic transcriptome of *P. falciparum* for the purpose of characterizing splicing in this organism. Unbiased, gene model-independent splice site detection within our data set in conjunction with RNA-Seq data from Otto *et al.* and Sorber *et al.* (14,27) was accomplished using the HMMSplicer algorithm (28), which was specifically developed to handle the challenging RNA-Seq data sets generated from the A/T-rich genome of *P. falciparum*. A total of 977 new 5' GU-AG 3' and 5 new 5' GC-AG 3' junctions never before documented in gene models or ESTs were discovered. Further analysis uncovered alternative splicing events, largely within 254 genes, as well as splicing events antisense to one another. Antisense events, some of which themselves displayed alternative splicing, likely indicate a mix of

overlapping annotated genes transcribed from opposite strands and unannotated transcripts transcribed antisense to gene models. Unexpectedly, antisense introns overlap sense gene introns more than would be anticipated by chance, perhaps indicating some relationship between overlapping transcripts, or an inherent feature of transcriptome organization. Over 30 antisense and alternative splicing events were independently experimentally verified, indicating that the new, alternative and antisense splicing events elucidated here support a larger, more dynamic understanding of the parasite's transcriptome.

## MATERIALS AND METHODS

### Generation of timepoint samples

3D7 Oxford *P. falciparum* parasites were grown at 2% hematocrit in 30 × T150 ml flasks with 50 ml of volume each. Repeat synchronization during peak invasion and again 12 h later over three consecutive lifecycles produced 30 ml of packed blood containing 11% highly synchronized late schizont parasites. This starter culture was allowed to invade 140 ml of unparasitized blood in 830 ml of culture medium in a 5 l dished bottom bioreactor (Applikon Inc, Brauwegg, Netherlands). Bioreactor conditions and culture medium were as in Bozdech *et al.* (1). After 4 h, the culture was diluted to ~5% hematocrit with 3 l of culture medium. Half (50%) of the culture was harvested 11 h after invasion (TP1), pelleted and frozen at –80°C. Thirty-three percent of the culture was harvested 22 h after invasion (TP2), 10% 33 h after invasion (TP3) and 7% 44 h after invasion (TP4). Total RNA was harvested from frozen pellets using Trizol (Invitrogen Corp., Carlsbad, CA, USA), then poly-A selected using the Micro FastTrack 2.0 kit (Invitrogen Corp.).

### Generation of RNA-Seq libraries

Libraries were generated as in Sorber *et al.* (27). Briefly, 1.2–1.6 µg of polyA-selected RNA was reverse transcribed using 6bp-EciI-N<sub>9</sub> (all primers can be found in [Supplementary Table S1](#)), and second strand cDNA synthesis was carried out with 13-bp-ModSolS-N<sub>9</sub>. Five cycles of PCR were done with 6bp-EciI and biotin-short-Mod-SolS (biotin-short-Mod-PE-SolS for TP1 and TP2 libraries), followed by binding to Dynal Dynabeads M-280 (Invitrogen Corp.). Bead-bound material was digested with EciI, then treated with Antarctic Phosphatase (New England Biolabs, Ipswich, MA, USA). Sol-L-NN annealed adapter was ligated onto cut ends. Five final cycles of PCR were performed on one-fourth of bead-bound material using Sol primer 1 and fullModSolS (fullMod-PE-SolS for TP1 and TP2 libraries). Remaining bead-bound material was subjected to three rounds of Long March using GsuI and the Sol-L-NN annealed adapter (27). The additional TP4 library sequenced here derived from a fourth Long March of the thrice-marched library described in Sorber *et al.* annealed to the Sol-L-AC-NN adapter (27). Final PCR on marched sub-libraries was as described for initial libraries.

### Illumina sequencing of RNA-Seq libraries

For TP1-3, the initial library and the thrice-matched sub-library were clustered on an Illumina flow cell in separate lanes (Illumina, Hayward, CA, USA). For single-end libraries and the first read of paired-end libraries, Sol-SeqPrimer was used as the sequencing primer, and PE-SolS-SeqPrimer was used to sequence the second read of paired-end libraries. Up to 60 single base extensions were performed with image capture using an Illumina GA2 sequencer (Illumina; [Supplementary Table S2](#)). The Illumina Pipeline software suite version 0.2.2.6 (Illumina) was utilized for base calling from these images for TP3 and TP4, and versions 1.3.2 and 1.5.0 were used to base call TP1 and TP2 images. All primary sequencing data can be found in the NCBI Short Read Archive under accession number SRA024324.1.

### Analysis pipeline

Raw sequence data from the above timecourse as well as from Otto *et al.* and Sorber *et al.* (14,27) were aggregated and any barcodes were removed. Reads with greater than 12 nt of adapter sequence, a repeat of A, T, C, G, or AT longer than 11 nt, or more than 10 nt with a quality scores  $\leq 5$  were discarded. Identical sequences within a timepoint were compressed to a single sequence read and the reads were filtered to remove human sequences, as detected by BLAST against the human genome with an *E*-value of  $1 \times 10^{-5}$  (29).

To gauge overall coverage, the filtered read set was aligned to the *P. falciparum* genome, PlasmoDB version 6.3 (30), by Bowtie version 0.12.1, using default parameters except that alignment of reads with multiple matches was disallowed (31). Reads unaligned by Bowtie were then aligned using BLAT version 34 with a tile size of 11, a step size of 1, and using an ooc file to filter repetitive sequence (32). Bowtie alignments were combined with BLAT alignments score  $\geq 35$  to yield the final set of aligned reads from which coverage statistics were generated.

To detect exon-exon spanning reads, HMMSplicer v0.7.0 was run in parallel on the filtered read set against the *P. falciparum* genome, PlasmoDB version 6.3 with a minimum intron size of 5 nt, a maximum intron size of 1000 nt and an anchor size of 6 nt (28). All other parameters were left at default values.

### Operational definitions for data analysis

To avoid confusion, a specific terminology was used to refer to specific parts of individual splice junctions and to classify junctions ([Supplementary Figure S1A–D](#)). For all definitions referencing gene models, a junction maps to a gene model only if at least one inner edge falls within the bounding coordinates of the gene model.

A ‘known junction’ maps to the same pair of inner boundaries as a splice junction found in PlasmoDBv6.3 gene models or in EST data ([Supplementary Figure S1B](#)). A ‘new junction’ maps to a pair of boundaries not seen in PlasmoDBv6.3 gene models or in EST data. ‘Canonical

junctions’ map to 5’ GU-AG 3’ boundaries, while ‘noncanonical junctions’ map to all other possible boundaries. A ‘junction conflict’ occurs when an inner edge of one junction falls within the intronic portion of the other junction such that they must occur in a mutually exclusive manner ([Supplementary Figure S1C](#)). ‘Junction groups’ were built by randomly selecting a nucleating junction, then searching for all relevant conflicting junctions. These junctions were added to the group and the search was iterated until no new junctions were appended. ‘Alternate 5’- and 3’-splice sites’ refers to splice junctions where both the 5’- and 3’-splice sites conflict ([Supplementary Figure S1C](#)). A splice junction that conflicts with two or more junctions that themselves do not conflict is considered a ‘skipped exon’. Although such instances could instead be interpreted as independent alternate 5’- and 3’-splice sites, skipped exon interpretation is consistent not only with our own independent experimental validations, but also frequently with gene models. In an ‘antisense conflict’, two junctions conflict with boundaries on opposite strands ([Supplementary Figure S1D](#)). However, ‘antisense junctions’ must have at least one boundary antisense to a gene model. See [Supplementary Materials and Methods](#) for analysis beyond these categorizations.

### Validation of new splicing events

In determining which new splicing events to assess, only new junctions that conflicted with recovered known junctions were considered, so that each validation had an internal positive control. In addition, the known isoform had to be  $\geq 30$  bp longer than the putative new isoform to ensure a selective restriction digest with size distinguishable final PCR products. The top 20 new junctions for which successful validation schemes could be computationally designed were picked starting with the highest scoring new junction for the group above the threshold, and starting with the highest scoring new junction below 1075 for the group below the threshold.

A biologically independent small-scale timecourse similar to the Bioreactor timecourse was performed using highly synchronous 3D7 Oxford parasites. After total RNA extraction as described above, RNA from each timepoint was reverse transcribed. For each validation, cDNA from the lifecycle stage with the highest representation of the new junction was used in the PCR-restriction digest-PCR scheme depicted in Figure 2A, with products column purified between steps. Appropriate primers and enzymes for each validation are listed in [Supplementary Table S3](#). Size appropriate final PCR bands were gel extracted, then TOPO TA cloned (Invitrogen Corp.). Whole cell PCR products from positive colonies were sequenced using the BigDye Terminator v3.1 Cycle Sequencing Kit on an ABI 3130xl Genetic Analyzer (Life Technologies Corp., Carlsbad, CA, USA). Resulting sequences were trimmed for vector and then aligned to the *P. falciparum* genome (v6.3) using BLAT (32). See [Supplementary Materials and Methods](#) for additional details.

## RESULTS

***Plasmodium falciparum* contains specific orthologs to splicing factors**

In *P. falciparum*, RNA components of the major U2-type spliceosome have been detected (7,8), yet protein components have not been systematically identified. Using reciprocal best hits (RBH) analysis (33) of human and yeast splicing factors, we identified putative homologs to spliceosome and spliceosome-associated protein components (Table 1) (34–36), the majority of which were most similar to their human counterparts. However, homologs of three components of the human spliceosome could not be identified: SFY2, PPIE and PRP2. PRP2, a DEAH/D-box ATPase, is ostensibly the most critical of the three, as it is thought to induce a structural rearrangement

that results in dissociation of the SF3a and b complexes from the branchpoint, rendering the branchpoint competent for nucleophilic attack of the 5'-splice site (37). Interestingly, initial analysis returned PF10\_0294 as the closest match in *P. falciparum* for both human PRP2 and PRP22, though the reciprocal BLAST completing RBH analysis returned PRP22 as a slightly better match for PF10\_0294 within the human genome (Table 1). PRP2 and PRP22 are both DEAH/D-box proteins involved in splicing with a high degree of conservation between their helicase and C-terminal domains. In *Saccharomyces cerevisiae* and other related yeast, PRP2 proteins contain a conserved DC amino acid doublet in their C-terminal domain that distinguishes them from other closely related DEAH/D-box ATPases, such as PRP22 (38). Although RBH analysis points to PF10\_0294 as a PRP22

**Table 1.** Putative *P. falciparum* splicing and non-sense-mediated decay factor homologs identified by reciprocal best hits analysis with human or *S. cerevisiae* sequences

Complex	Human/Yeast	<i>Pf</i> Homolog	Complex	Human/Yeast	<i>Pf</i> Homolog
<b>snRNP core</b> (stability and function of U1, U2, U4 and U5 snRNPs)	SNRPB/SMB1	PF14_0146	<b>U4/U6</b> (catalytic activation of spliceosome)	PRPF3/PRP3	MAL13P1.45
	SNRPD1/SMD1	PF11_0266		NHP2L1/SNU13	PF11_0250
	SNRPD2/SMD2	PFB0865w		PRPF4/PRP4	MAL13P1.385 <sup>a</sup>
	SNRPD3/SMD3	PF10475w		PRPF31/ <i>PRP31</i>	PFD0450c
	SNRPE/SME1	MAL13P1.253		PPIH/-	PF08_0121 <sup>b</sup>
	SNRPF/SMX3	PF11_0280		SART1/SNU66	PFC1060c <sup>a</sup>
	SNRPG/SMX2	MAL8P1.48		USP39/SAD1	PF13_0096 <sup>a</sup>
<b>U6 core</b> (stability and function of U6 snRNP)	LSM2/LSM2	PFE1020w	<b>tri-snRNP</b> (activation of spliceosome)	SNRNP27/-	MAL8P1.71 <sup>a,b</sup>
	LSM3/LSM3	PF08_0049		USP39/SAD1	PF13_0096 <sup>a</sup>
	LSM4/LSM4	PF11_0524		SNRNP27/-	MAL8P1.71 <sup>a,b</sup>
	LSM5/LSM5	PF14_0411		PRPF19/PRP19	PFC0365w
	LSM6/LSM6	PF13_0142 <sup>b</sup>		CRNKL1/CLF1	PFD0180c
	LSM7/LSM7	PFL0460w		CDC5L/CEF1	PF10_0327 <sup>a</sup>
	NAA38/LSM8	MAL8P1.9 <sup>b</sup>		ISY1/ISY1	PF14_0688
	SNRNP70/SNP1	MAL13P1.338		BCAS2/SNT309	PFF0695w <sup>a,b</sup>
<b>U1</b> (initial 5'-ss recognition)	SNRPA/MUD1	MAL13P1.35 <sup>b</sup>	<b>hPrp19/CDC5</b> (specification of U5 and U6 interactions with RNA)	XAB2/SYF1	PFL1735c <sup>a</sup>
	SNRPC/YHC1	PF08_0084		PLRG1/PRP46	PFC0100c <sup>a</sup>
	SNRPA1/LEA1	PF13_0362		SYF2/SYF2	?
<b>U2</b> (BP detection)	SNRPB2/MSL1	PF11695c	<b>Non-snRNP factors</b> (second step factors) (RNA release)	SNW1/ <i>PRP45</i>	PFB0875c <sup>a</sup>
	U2AF1/-	PF11_0200 <sup>b</sup>		BUD31/ <i>BUD31</i>	PFE1140c
<b>U2-related</b> (BP and poly-Y recognition)	U2AF2/MUD2	PF14_0656 <sup>b</sup>	<b>SR and hnRNP</b>	PPIE/-	?
	SF1/MSL5	PFF1135w		CCDC12/-	PF14_0490 <sup>a</sup>
	SF3A1/PRP21	PF14_0713 <sup>a</sup>		AQR/-	PF13_0273 <sup>a,b</sup>
	SF3A2/PRP11	PFF0970w		CWC15/ <i>CWC15</i>	PF07_0091 <sup>b</sup>
	SF3A3/PRP9	PF11215w		PPIL1/-	PFE1430c <sup>b</sup>
<b>SF3a</b> (stability of U2-BaP interaction)	SF3B1/HSH155	PFC0375c	<b>NMD</b> (detection of nonsense transcripts)	DHX16/PRP2	PF10_0294 <sup>a</sup>
	SF3B2/CUS1	PF14_0587		BAT1/SUB2	PFB0445c
	SF3B3/RSE1	PFL1680w		DDX46/PRP5	PFE0430w <sup>a</sup>
	SF3B4/HSH49	PF14_0194		SLU7/SLU7	PFF0500c
	SF3B5/YSF3	PF13_0296		DHX38/PRP16	MAL13P1.322
	PHF5A/RDS3	PF10_0179a		CDC40/CDC40	PFL0970w
	SF3B14/-	PFL1200c <sup>b</sup>		PRPF18/ <i>PRP18</i>	PF11115c
	DDX23/PRP28	PFE0925c		DHX8/PRP22	PF10_0294 <sup>a</sup>
<b>SF3b</b> (stability of U2-BP interaction)	CD2BP2/LIN1	PF10_0310 <sup>a</sup>	<b>SR and hnRNP</b>	UPF1/NAM7	PF10_0057
	EFTUD2/SNU114	PF10_0041 <sup>b</sup>		UPF2/NMD2	PFI1265w <sup>a</sup>
	SNRNP200/ BRR2	PFD1060w		UPF3A/UPF3	?
	TXNL4A/DIB1	PFL1520w		UPF3B/-	PF13_0158 <sup>b</sup>
	PRPF8/PRP8	PFD0265w		SRSF1/-	PFE0865c <sup>b</sup>
	PRPF6/ <i>PRP6</i>	PF11_0108		SF12/-	SFE0160c <sup>b</sup>
	SNRNP40/-	MAL8P1.43 <sup>b</sup>		PTBP2/-	PFF0320c <sup>b</sup>
<b>U5</b> (catalytic activation of spliceosome)			SFRS4/-	PF10_0217 <sup>b</sup>	
			TRA2B/-	PF10_0028 <sup>a,b</sup>	

The human or *S. cerevisiae* factor in bold font represents the best match for the *P. falciparum* homolog. Homologs of spliceosomal and NMD factors not found are denoted with question marks, while SR and hnRNP factors not found are not shown. *Saccharomyces cerevisiae* homologs that do not reside in the same complex as their human counterparts are italicized.

<sup>a</sup>*Plasmodium falciparum* proteins described in PlasmoDB as 'conserved *Plasmodium* protein' or with descriptions that do not reflect involvement in splicing.

<sup>b</sup>Homologs identified only by the human sequence.

homolog, alignment of the C-terminal portion of PF10\_0294 reveals the presence of the DC doublet signature of PRP2 homologs in yeast (Supplementary Figure S2). Without biochemical characterization, it is difficult to determine which role PF10\_0294 might play, and it is possible that it encompasses the activity of both DEAH/D-box ATPases. Thus, while our RBH analysis is helpful as a first step in determining players involved in splicing, careful experimental verification of the exact roles of these putative homologs is still required to fully understand how splicing occurs in *P. falciparum*.

In other eukaryotes, alternative splicing is guided by the presence or absence of proteins that determine which splice sites are available to the spliceosome (39). To determine if *P. falciparum* has homologs to such proteins, human arginine/serine-rich (SR) and heterogeneous nuclear ribonucleoproteins (hnRNP) proteins with documented roles in alternative splicing were used for best reciprocal hits analysis (40,41). Four SR proteins and one hnRNP protein returned specific homologs (Table 1). These homologs likely represent only a fraction of the proteins that influence splice site selection in *P. falciparum*, as at least 71 additional proteins contain either an RNA recognition motif (RRM) or an RNA binding domain (RBD) according to InterPro (42), and 7 contain an RS domain according to our own analysis. Many proteins involved in splice site selection during alternative splicing utilize one or more of these domains, although they do not guarantee involvement in splicing (40,41). Together these data suggest that alternative splicing could play an important role in *P. falciparum*.

### Overview of *P. falciparum* RNA-Seq data sets

To investigate splicing in *P. falciparum* on a transcriptome-wide scale, we generated short read RNA-Seq data from multiple timepoints of a highly synchronous, large-scale, intraerythrocytic culture and analyzed these data, in conjunction with publically available data sets, for splice junctions. To guarantee adequate representation of distinct blood stages, timepoints were collected from the 3D7 Oxford culture approximately 11 (ring), 22 (trophozoite), 33 (late trophozoite/early schizont) and 44 (late schizont) h post-invasion. After total RNA isolation, poly-A RNA was purified and prepared for Illumina sequencing using the Long March protocol (27). All primary sequencing data can be found in the NCBI Short Read Archive under accession number SRA024324.1. To maximize our transcriptome-wide examination of splicing, we also included two previously published *P. falciparum* RNA-Seq data sets in our analysis: one from seven timepoints within the blood stage of 3D7 parasites by Otto *et al.* (14) and one from the late schizont timepoint of our experiment (27).

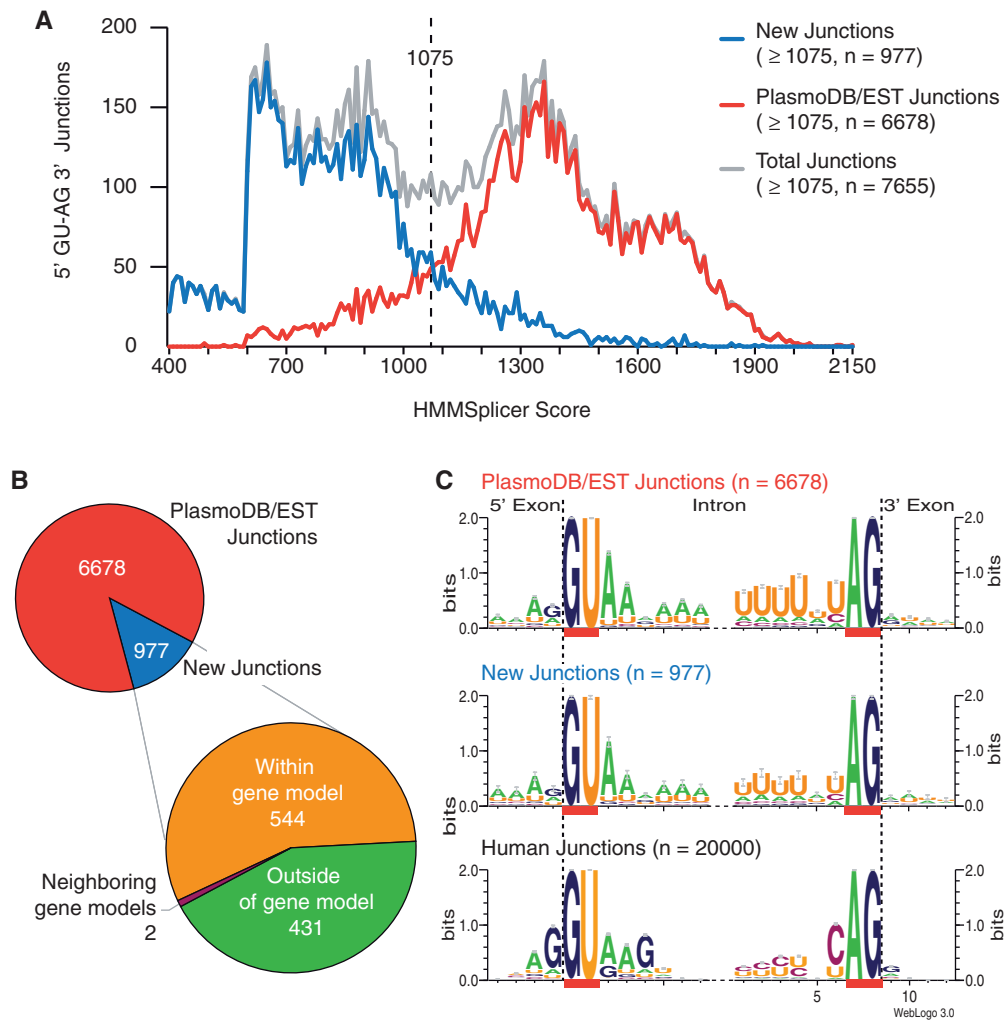
We aggregated these data and after preliminary filtering and sequence collapsing, ran two analyses in parallel: a Bowtie/BLAT (31,32) pipeline to align ungapped reads back to the *P. falciparum* genome (PlasmoDBv6.3) and HMMSplicer v0.7.0 (28) to detect and score exon-exon splice junctions. The Bowtie/BLAT pipeline was able to align between 84 and 194 million bases of sequence to

the *P. falciparum* genome for each independent timepoint (Supplementary Table S2) for a total of over 1.5 billion aligned bases. Discounting antigenic variation gene families (272 *vars*, *rifins* and *stevors*), each exonic nucleotide of a gene was covered by a median of 59 reads. HMMSplicer was also run on the data set with a minimum intron size of 5 bp and a maximum intron size of 1000 bp, covering 99.6% of all annotated *P. falciparum* introns. More than 1.9 million reads in the combined data set were mapped to junctions (Supplementary Table S4).

To gauge the quality of the junctions predicted from the combined data set, we examined the distribution of HMMSplicer scores calculated for all predicted 5' GU-AG 3' junctions (Figure 1A). HMMSplicer scores reflect both the strength of the junction alignment and the cumulative support for that junction within the data set (28). The distribution of HMMSplicer scores for canonical *P. falciparum* junctions is clearly bimodal, perhaps indicating predictions of differing reliability. In this organism, PlasmoDB gene models and ESTs provide a set of previously known splice junctions that are likely to be valid (12,30,43,44) and the distribution of HMMSplicer scores for only those junctions with boundaries matching previously known junctions was found to primarily fall within the higher scoring population. Therefore, an HMMSplicer score of 1075, representing the natural breakpoint in the bimodal distribution, was chosen as an operational threshold for subsequent analysis (Figure 1A). Below this threshold, support for detected junctions decreases rapidly, and thus the false positive rate among these lower-confidence junctions is likely to be higher. However, 13% of all known junctions detected within the combined data set fall below our threshold, indicating the presence of valid junctions with non-ideal coverage, though only 0.1% score below 600 (Figure 1A). While we have enacted an operational threshold, all HMMSplicer junctions regardless of score are accessible for additional analyses (Supplementary Files S1 and S2, upload at <http://plasmodb.org/cgi-bin/gbrowse/plasmodb/>).

### HMMSplicer analysis of RNA-Seq data reveals new canonical splice junctions

HMMSplicer found 7655 5' GU-AG 3' junctions above the operational threshold within the combined RNA-Seq data. More than 88% were supported by reads from both time courses. Of these high scoring junctions, 6678 (87.2%) confirm introns in PlasmoDB gene models or ESTs (Supplementary File S3, upload at <http://plasmodb.org/cgi-bin/gbrowse/plasmodb/>). 977 (12.8%) support new introns, an increase of 11% over the current genome annotation (Figure 1B, Supplementary File S4, upload at <http://plasmodb.org/cgi-bin/gbrowse/plasmodb/>). 431 (43.9%) of these new junctions fall either totally or partially outside of annotated gene models, suggesting splicing in unannotated untranslated regions (UTRs) or in unannotated genes, whereas 544 (55.4%) align within gene models. As discussed below, many of the new junctions discovered within gene models represent alternative



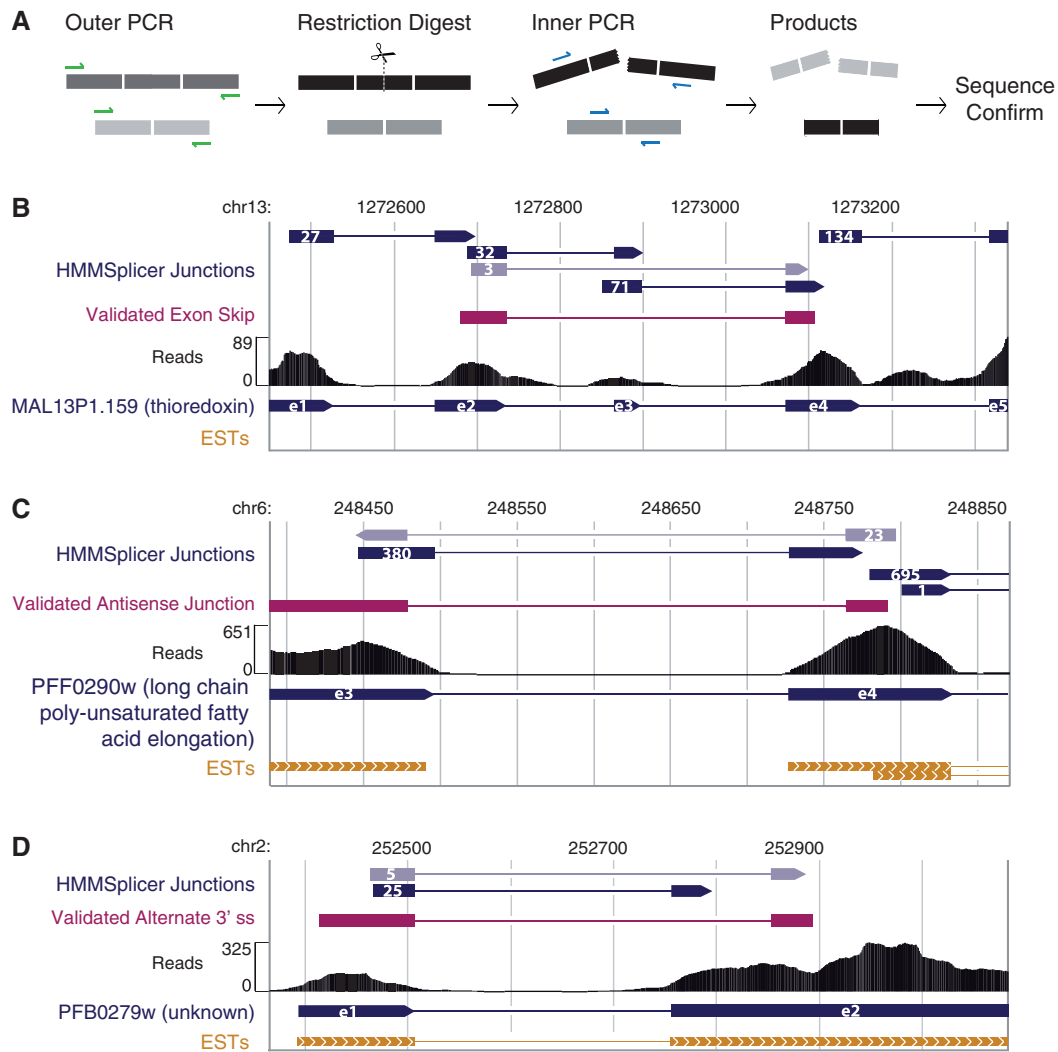
**Figure 1.** (A) Histogram of 5' GU-AG 3' junctions found by HMMSplicer binned by score. Defaults retain all junctions supported by multiple reads scoring above 400 and all junctions supported by single reads scoring above 600. The grey line plots all reported 5' GU-AG 3' junctions, while the red line charts junctions that match previously known junctions in PlasmDB v6.3 gene models or in ESTs. The blue line charts new junctions. The dashed line drawn at 1075 represents the operational score threshold. (B) Breakdown of canonical junctions with scores above 1075, with additional classification of new junctions. 'Outside of gene model' refers to new junctions with at least one inner edge mapped to an intergenic region. 'Within gene model' indicates that both inner edges mapped to the same gene model. 'Neighboring gene models' indicates that the inner edges mapped to neighboring gene models. (C) Comparison of the 5'- and 3'-splice site WebLogos for previously known junctions recovered versus new junctions above 1075. WebLogos calculated for human junctions are included for reference. Red bars indicate the 5' GU-AG 3' boundaries used for inclusion in each set. The height of each letter indicates the preference strength for that nucleotide at each position.

transcript isoforms or splicing of antisense transcripts. Unexpectedly, 2 (0.2%) new junctions map to neighboring genes encoded on opposite strands, suggesting unannotated overlap between these gene pairs.

We sought to lend support to the new 5' GU-AG 3' junctions detected by HMMSplicer by calculating WebLogos (45) from their 5'- and 3'-splice sites. If these new junctions represent true splicing events, they would be predicted to recapitulate the nucleotide preferences found within 5'- and 3'-splice sites of known 5' GU-AG 3' junctions. Indeed, no significant differences were observed between our calculated sequence logos for PlasmDB/EST matching junctions versus new junctions, and both sets of logos closely matched previously published results (Figure 1C) (7). In contrast, logos produced from the

bottom 10% of new junctions below the operational threshold contained little information other than their 5' GU-AG 3' boundaries (Supplementary Figure S3A). Efforts to determine a branchpoint motif from the introns defined by our high-scoring canonical junctions yielded no convincing results, similar to Chakrabarti *et al.*'s efforts to determine a branch point motif from a smaller set of EST introns (7).

The validity of new 5' GU-AG 3' junctions was also independently assessed by experimental validation using an biological replicate of our original blood stage timecourse and the strategy described in Figure 2A. Twelve 5'-alternate splice sites, 2 3'-alternate splice sites, 17 skipped exons and 10 spliced antisense transcripts were tested (Table 2, Supplementary Figure S1C and D). A



**Figure 2.** Validation of new splicing events. (A) Shade indicates the relative abundance of each isoform. Initial outer PCR (green arrows) amplifies both isoforms from cDNA. A restriction enzyme then cuts the known isoform. Nested inner PCR (blue arrows) amplifies only the uncut, new isoform, which is then sequence confirmed. Gbrowse (66) windows depict validation of a skipped exon in MAL13P1.159 (B), an antisense junction in PFF0290w (C), and an alternate 3'-splice site in PFB0279w (D). All HMMSplicer junctions scoring higher than 980 are shown as either dark blue bars (known junctions) or light blue bars (new conflicting junctions). The number of reads supporting each junction is shown in the bars, while the direction of the arrow reflects the direction of the splice sites. Validation sequencing results are shown in magenta. Bowtie coverage for each nucleotide in the window is shown as a histogram. Underneath, the dark blue bars depict PlasmoDB v6.3 gene models with numbers denoting the exons, while the gold bars at the bottom of each window depict ESTs.

total of 19/21 (90.5%) new splicing events ranging in score from 1189.3 to 1544.2 were experimentally confirmed, including a skipped exon in MAL13P1.159 (thioredoxin) and splicing of an antisense transcript mapping to PFF0290w (long chain polyunsaturated fatty acid elongation enzyme) (Figure 2B and C); 13/20 (65%) events below the operational threshold with scores ranging from 984.6 to 1050.5 were also confirmed. Since more than half of these lower scoring events were successfully verified, these validations also confirm that our threshold is conservative—in addition to excluding false positive junctions, it also excludes some true splicing events, such as the 3'-alternate splice site in PFB0279w (conserved *Plasmodium* protein, Figure 2D). Overall, these results indicate that a high percentage of new junctions both above and below the threshold are genuine, although

independent confirmation may be required for lower scoring junctions.

Our results suggest both that a number of true positive junctions exist below our operational threshold and that the nucleotide preferences present at the 5'- and 3'-splice sites of known junctions do not hold for the lowest scoring, least reliable junctions in the data set. Therefore, to attempt recovery of true 5' GU-AG 3' junctions below our operational threshold, an orthogonal score based on position specific scoring matrices (46) of the splice site logos was evaluated. Although this type of motif scoring ultimately lacked sufficient information for large-scale computational rescue (Supplementary Figure S3B), it could potentially be used to prioritize experimental assessment of new junctions (Supplementary File S5).

**Table 2.** Verification of new junctions in conflict with known junctions

Gene name	PlasmoDBv.6.3 description	Score	Validated	Type	Frame-shift?	Isoform difference in bp (aa)	R	T	LT/ES	S
PFL1810w	Conserved <i>Plasmodium</i> protein	1544.2	Yes	5'ss	No	132 (44)	11	8	11	5
PFE0390w	Conserved <i>Plasmodium</i> protein	1283.2	Yes	5'ss	No	219 (73)	3	0	3	1
PF13_0138	MSF-1 like protein	1422.1	Yes	5'ss	No	66 (22)	13	10	8	3
PF10400c	Conserved <i>Plasmodium</i> membrane protein	1372	Yes	5'ss	Yes	56	7	5	3	7
PFF0290w	Conserved <i>Plasmodium</i> membrane protein	1369.8	Yes	Exon skip	No	126 (42)	2	1	19	1
PFF0290w	Long chain polyunsaturated fatty acid elongation enzyme	1291.8	Yes	Antisense	–	–	9	6	14	5
MAL13P1.225	Thioredoxin	1277.3	Yes	Exon skip	Yes	34	2	6	0	0
PFE0055c	Heat shock protein	1275.4	Yes	5'ss	Yes	37	54	14	3	8
MAL8P1.126	Serine protease	1257.4	Yes	5'ss	Yes	110	1	1	11	8
PF10_0025	PF70 protein	1256.8	Yes	5'ss	No	75 (25)	18	2	0	0
PFD1050w	Alpha-tubulin II	1243.2	No	Antisense	–	–	1	0	5	1
MAL13P1.159 <sup>a</sup>	Thioredoxin	1239.9	Yes	Exon skip	No	33 (11)	0	1	0	3
PFC0780w	Cleavage and polyadenylation specific factor	1231.7	No	Antisense	–	–	2	6	27	8
PFD0775c	RNA binding protein	1228.4	Yes	Antisense	–	–	1	6	8	0
PF10_0194	NoOP12-like protein	1219.4	Yes	Exon skip	Yes	41	1	0	0	1
PFL1440c	Conserved <i>Plasmodium</i> protein	1217.6	Yes	Exon skip	No	57 (19)	0	2	0	1
PF11_0291	Conserved <i>Plasmodium</i> protein	1203.5	Yes	5'ss	Yes	40	1	0	0	5
PFC0360w	Activator of HSP90 ATPase homolog 1-like protein	1200.5	Yes	Exon skip	Yes	223	1	1	3	0
PFC0495w	Plasmeprin VI	1192.6	Yes	Antisense	–	–	0	0	8	3
PF14_0394	Conserved <i>Plasmodium</i> protein	1190	Yes	5'ss	No	99 (33)	2	4	5	0
MAL13P1.146	AMP deaminase	1189.3	Yes	antisense	–	–	1	1	0	0
PF11_0379	Conserved <i>Plasmodium</i> protein	1050.5	Yes	Exon skip	No	60 (20)	1	0	0	1
PFL1445w	Conserved <i>Plasmodium</i> protein	1041.3	Yes	Exon skip	Yes	85	0	6	0	0
MAL13P1.16	SNARE protein	1034.7	Yes	Exon skip	No	108 (36)	0	0	0	4
MAL13P1.277	DNAJ-like protein	1034.2	Yes	Exon skip	Yes	146	2	0	3	0
PFF1210w	Phosphatidic acid phosphatase	1032.4	Yes	5'ss	Yes	67	2	0	3	5
PFB0600c	Conserved <i>Plasmodium</i> protein	1026.1	Yes	Antisense	–	–	1	1	3	0
PF14_0128	Ubiquitin conjugating enzyme	1018.5	Yes	Exon skip	Yes	103	0	1	3	0
PF14_0316	DNA topoisomerase II	1011.4	No	5'ss	Yes	460	0	0	3	1
PFB0279w <sup>a</sup>	Conserved <i>Plasmodium</i> protein	1010.9	Yes	3'ss	Yes	98	1	4	3	0
PFL1465c	Heat shock protein hslv	1004.5	No	Exon skip	Yes	39 (13)	0	2	0	4
PF10_0372	Antigen UB05	1004.4	No	antisense	–	–	0	1	0	1
PF11_0182	Conserved <i>Plasmodium</i> protein	1004.1	Yes	Exon skip	Yes	56	0	0	5	0
PFF0365c	G-protein associated signal transduction protein	996.3	No	Exon skip	No	162 (54)	2	0	0	0
PFB0445c	DEAD box helicase, UAP56	995.9	No	3'ss	No	75 (25)	1	0	0	0
PFD0895c	Bet3 transport protein	991	No	Antisense	–	–	0	0	2	0
PF10_0116	Conserved <i>Plasmodium</i> protein	989.9	Yes	5'ss	No	75 (25)	0	0	2	0
PF14_0604	Conserved <i>Plasmodium</i> protein	988.1	Yes	Exon skip	Yes	343	1	0	0	0
PF10560c	Conserved <i>Plasmodium</i> protein	987.7	Yes	Exon skip	Yes	40	1	0	0	2
PFB0550w	Peptide chain release factor subunit 1	985.5	No	Exon skip	Yes	155p	0	0	2	0
PF11_0355	Conserved <i>Plasmodium</i> protein	984.6	Yes	Antisense	–	–	1	1	0	0

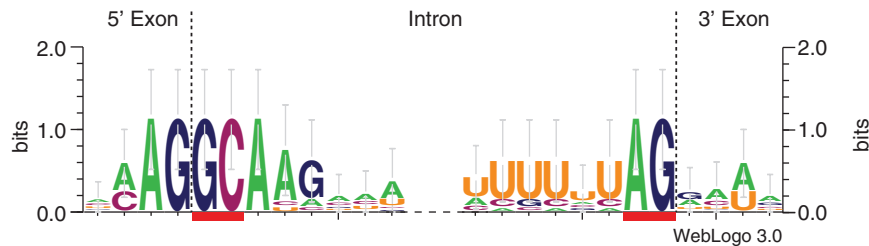
Conflicts are ranked by lowest HMMSplicer score within the pair, and the black line denotes the operating HMMSplicer threshold of 1075.

<sup>a</sup>Validations shown in more detail in Figure 2. For all conflict types except antisense, the new junction was evaluated for maintenance of the ORF—nucleotide and amino acid (if applicable) differences between new and known isoforms are listed. Read counts for new junctions (normalized by the number of reads mapped by Bowtie for each timepoint) are listed for ring [R, (TP1, TP0, TP8)], troph [T, (TP2, TP16, TP24)], late troph/early schizont [LT/ES, (TP3, TP32)] and schizont [S, (TP4, TP40, TP48)] timepoints.

### Inspection of non-canonical splice junctions reveals new 5' GC-AG 3' junctions

In many eukaryotes, splicing occasionally occurs at non-5' GU-AG 3' boundaries, sometimes via the major U2-type spliceosome as with 5' GC-AG 3' introns (47), or via the minor U12-type spliceosome as with 5' AT-AC 3' introns (48), or spliceosome-independently as with the 5' CA-AG 3' intron in yeast HAC1 (49). The presence of 5' GC-AG 3' junctions in *P. falciparum* ESTs and gene models (12,13) suggests that the parasite uses 5' GC non-canonical splice

sites, yet this likelihood has never been examined in detail. Of the 984 non-canonical junctions above our operational threshold (Supplementary File S6, upload at <http://plasmodb.org/cgi-bin/gbrowse/plasmodb/>), 12 map to 5' GC-AG 3' boundaries (Supplementary Table S5). Of these, 7 were supported by either EST evidence or annotated PlasmoDB gene models, and 5 were completely new. We used WebLogo v3.0 to construct 5'- and 3'-splice site sequence logos from all 12 5' GC-AG 3' junctions (45) (Figure 3). The 3'-splice site logo was very similar to the



**Figure 3.** WebLogo 5'- and 3'-splice site motifs for high scoring 5' GC-AG 3' HMMSplicer junctions ( $n = 12$ ). Red bars indicate the boundaries used for inclusion in the set. The height of each letter indicates the information content for that nucleotide at each position. The large error bars derive from the small size of the input set.

canonical 3'-splice site logo. However, several clear differences distinguished the 5'-splice site logo for 5' GC-AG 3' junctions from that of canonical junctions. Whereas canonical *P. falciparum* 5'-splice sites have a slight preference for AG as the last two bases of the 5'-exon, all 12 5' GC-AG 3' examples contain AG in these positions, and all 12 also contain an A at the third position of the intron. These same 3 nt are also present in the two PlasmoDB 5' GC-AG 3' junctions with HMMSplicer scores below our operational threshold. Both the fourth and fifth positions of 5' GC-AG 3' introns also appear to have strong, although not absolute, nucleotide preferences. Though stronger contextual sequence may simply reflect the small number of input sequences, stronger consensus 5'-splice site motifs have been documented for 5' GC-AG 3' introns in other organisms as well (11). As with 5' GU-AG 3' introns, efforts to determine a branchpoint motif from these introns failed to produce any convincing results.

We also considered the possibility that the parasite might employ splice sites other than 5' GU-AG 3' and 5' GC-AG 3'. However, preliminary manual inspection of the remaining non-canonical junctions revealed that many of them were likely to be false positives caused by read errors. Polymerase slipping, template switching, and single base pair substitutions are well-documented phenomena (50–52) that can occur during both the reverse transcription and PCR steps of library preparation. These upstream errors have no associated cost in sequence quality, and therefore may explain the origins of high scoring, erroneous junction reads. Since the probability of an erroneous read mapping to non-canonical boundaries is much greater than the probability of it mapping to canonical boundaries, it is not surprising that the false positive rate within the non-canonical junctions is higher than within the canonical junctions.

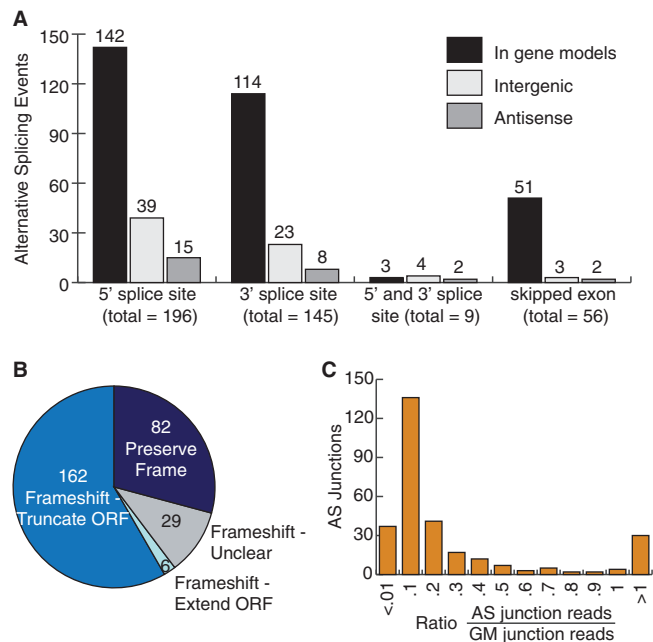
Two additional filters designed to eliminate false positive junctions while retaining any potential true non-canonical junctions were applied to the non-canonical junctions (Supplementary Materials and Methods). Since HMMSplicer is more sensitive to errors the closer they are to the true breakpoint of a junction read (28), the first filter eliminated non-canonical junctions with single base substitutions within 15 bp of either inner edge that caused miscalling of the junction breakpoint (343 of 972 junctions). The second filter removed non-canonical junctions in very highly covered regions since the probability of

error creation during preparation and sequencing increases as the copy number for a given sequence increases (356 of 629 remaining junctions). As an internal check, neither filter eliminated any of the 12 5' GC-AG 3' junctions previously identified.

Manual inspection of the remaining 273 non-canonical junctions yielded no additional, credible non-canonical splice junctions. Although 5' AT-AC 3' splice sites have been observed in introns excised by the U12 minor spliceosome (48), failure to detect any in the RNA-Seq data is consistent with our failure to find *P. falciparum* homologs to proteins specific to the human U12-type spliceosome (53). Similarly, a previous search by Lopez *et al.* for all snRNAs in a variety of eukaryotes returned no minor spliceosome snRNAs in any *Apicomplexa*, including the two rodent *Plasmodium* species examined (54). Together, these results indicate that *P. falciparum* is unlikely to possess a minor U12-type spliceosome.

### Genome-wide characterization of alternative splicing

Alternative splicing increases transcriptome complexity by generating multiple isoforms from the same precursor that differ in single 5'- or 3'-splice sites or in whole exons and introns. To find alternative splicing within the combined data set in an unbiased manner, independent of gene models, high scoring canonical and 5' GC-AG 3' junctions were compared to each other in a pair wise manner. To be considered 'conflicting junctions', one of the inner edges of a junction must have aligned within the intronic area of the other junction (Supplementary Figure S1B). Since direct counting of these occurrences would over-inflate the number of alternative splicing events (for example, a single skipped exon event would count as two pair-wise conflicts), conflicts were further aggregated into junction groups (Supplementary Figure S1B), which were then divided by strand orientation where applicable (Supplementary File S7). In total, 196 (48.3%) alternate 5'-splice sites, 145 (35.7%) alternate 3'-splice sites, 8 (2.0%) mutually exclusive alternate 5'- and 3'-splice sites and 56 (13.8%) skipped exons were tallied (Figure 4A). The majority of alternative splicing events occurred in gene models in the sense direction, though some also occurred outside of gene models. These intergenic events most likely indicate alternative splicing in unannotated *P. falciparum* UTRs or in unannotated genes. Interestingly, all four types of alternative splicing were also seen in antisense junction groups. Further



**Figure 4.** Breakdown of alternative splicing events detected transcriptome-wide. (A) Alternative splicing events both by type and area in the genome. Events ‘In gene models’ belong to junction groups in which at least one junction maps within a gene model in the sense direction. ‘Intergenic’ events belong to junction groups with no junctions mapping to gene models. ‘Antisense’ events occur in junction groups with at least one junction within a gene model in the antisense direction. (B) Breakdown of the 279 alternative splicing events that have the potential to change the gene model’s coding sequence. ‘Frameshift-unclear’ could not be analyzed for ORF extension or truncation without assuming which downstream junction(s) co-occur in a given isoform. (C) Histogram of alternative splicing (AS) junctions ( $n = 296$ ) by ratio of AS junction reads to recovered gene model (GM) junction reads. In cases of conflict with more than one GM junction, the GM junction with the most reads was chosen as the denominator.

analysis of antisense splicing events is discussed in the next section.

Because the combined RNA-Seq data are comprised of short reads rather than full length mRNAs, the collection of splice junctions that compose a given isoform is difficult to resolve, and thus the exact number of isoforms encoded by the alternative splicing events described here could not be determined. However, transcriptome-wide, the combined data set supports the existence of between 279 and 369 alternative isoforms (533 and 623 total isoforms) for the 254 genes in which conflicting junctions were detected (Supplementary File S7). Alternative splicing events for most genes maximally support between 2 and 4 isoforms. However, a handful of genes [PF14\_0338 (conserved *Plasmodium* protein), PFF0630c (conserved *Plasmodium* protein), PFL1440c (conserved *Plasmodium* protein), PFC0495w (plasmepsin VI), and PFC0912w (signal peptidase)] could encode up to 8–16 different isoforms. In addition to supporting up to 8 sense isoforms, an overlapping antisense junction was also validated for PFC0495w (plasmepsin VI), making it particularly interesting (Table 2). Gene ontology (GO) analysis of alternatively spliced genes did not reveal any

functional patterns (Supplementary Materials and Methods).

The transcriptome complexity afforded by alternative splicing often increases the number of distinct proteins encoded by an organism. Of the 310 *P. falciparum* alternative splicing events mapped to gene models in the sense direction, 10% are predicted to produce altered UTRs, while the remaining 279 (90%) are predicted to produce distinct coding sequences. Of these, close to one third maintain coding frame, either adding or removing amino acids from the predicted protein (Figure 4B). In contrast, the majority of alternative splicing events result in frame-shifts, most of which introduce premature termination codons within the gene model’s predicted coding sequence.

One explanation for the abundance of protein truncating alternative splicing events in *P. falciparum* is that these transcripts may not be translated, but instead could be intermediates bound for non-sense-mediated decay (NMD). Regulated splicing controlling the ratio of NMD-targeted to protein-coding isoform produced from certain genes is a mechanism of post-transcriptional regulation in other organisms (39,55). However, NMD has not been shown to exist in the parasite. Using human and yeast sequences for the core conserved NMD surveillance proteins, UPF1, UPF2 and UPF3 (paralogs UPF3a and UPF3b in humans) (56), best reciprocal hits analysis was able to find homologs to all three in *P. falciparum*, suggesting the NMD pathway exists in this parasite (Table 1). While it is unclear what the trigger for NMD may be in *P. falciparum*, 119 (73%) of the 162 truncating events do so >50 bp upstream of the last splice junction, rendering them eligible for NMD in mammalian systems (56). Regardless, our results suggest that the majority of alternative splicing events in the blood stages of *P. falciparum* either produce truncated protein isoforms or tune gene expression post-transcriptionally.

We also looked at the relative abundance of alternate junctions in comparison to their recovered gene model counterparts (Figure 4C). Many occurred at <10% of the frequency of the conflicting gene model junction within the combined data sets, and may correspond to isoforms either targeted for non-sense-mediated decay or of minimal use in the blood stages. Interestingly, 33 alternative junctions occurred at  $\geq 100\%$  of the frequency of their conflicting gene model counterparts, indicating that the gene model isoform of the transcript may not be the dominant blood stage isoform (Supplementary Table S6).

#### A minority of introns are poorly spliced in *P. falciparum*

Previous reports of alternative splicing in *P. falciparum* have noted instances of transcripts with retained introns (12), and regulated splicing efficiency can control such important biology as onset of meiosis in *S. cerevisiae* (57). Therefore, to gauge general splicing efficiency as well as to discover poorly spliced outlier introns, we calculated the ratio of junction reads to the average number of reads covering both cognate exon–intron borders (58). Only recovered gene model junctions in genes without mapped antisense junctions were considered to avoid complicating

factors. Because the data sets analyzed here were not generated specifically for the purpose of analyzing splicing efficiency, calculations could be made for only a subset of splice junctions in which the read counts covering both exon–intron borders were relatively similar (Supplementary Materials and Methods). For the 779 introns analyzed, junction reads were recovered a median of five times more often than exon–intron reads (Supplementary File S8). However, 44 (5.6%) analyzed introns appear to be very poorly spliced in the blood stages as they are retained in at least 50% of the transcripts sampled here.

### A subset of new junctions within genes challenge their corresponding gene models

Although gene models were not consulted during detection of junctions or alternative splicing, we assessed how thoroughly they were encompassed by our results. Of the 8435 predicted splice junctions in PlasmoDB v6.3 gene models, 1103 were not observed in the combined data set, even below our operational threshold. Gene models with unrecovered splice junctions had a median coverage of six reads per coding nucleotide, indicating that in general, these genes were not substantially expressed during the blood stages. However, for 50 unrecovered known junctions, new junctions above the operational threshold were observed that did not match the boundaries indicated by the gene model (Supplementary Table S7). Although it is possible that the gene model isoforms are not expressed in the blood stages in these cases and that the new junctions represent blood stage-specific alternate isoforms, it is more likely that the corresponding gene models are incorrect.

### Genome-wide characterization of antisense splicing

While probing for conflicting junctions, we noticed a class of conflicts in which one junction contained intron boundaries on a given strand while the other mapped to intron boundaries on the opposite strand (Supplementary Figure S1D). Although none of the data sets analyzed here were derived from a directional library, the orientation of intron boundaries has been used in the past to assign direction to ESTs (59). In addition, antisense transcription has been previously noted in *P. falciparum* (60), and ESTs antisense to gene models have also been documented (12). Therefore, it is likely that these ‘antisense conflicts’ derive from overlap of two spliced transcripts transcribed in opposite directions.

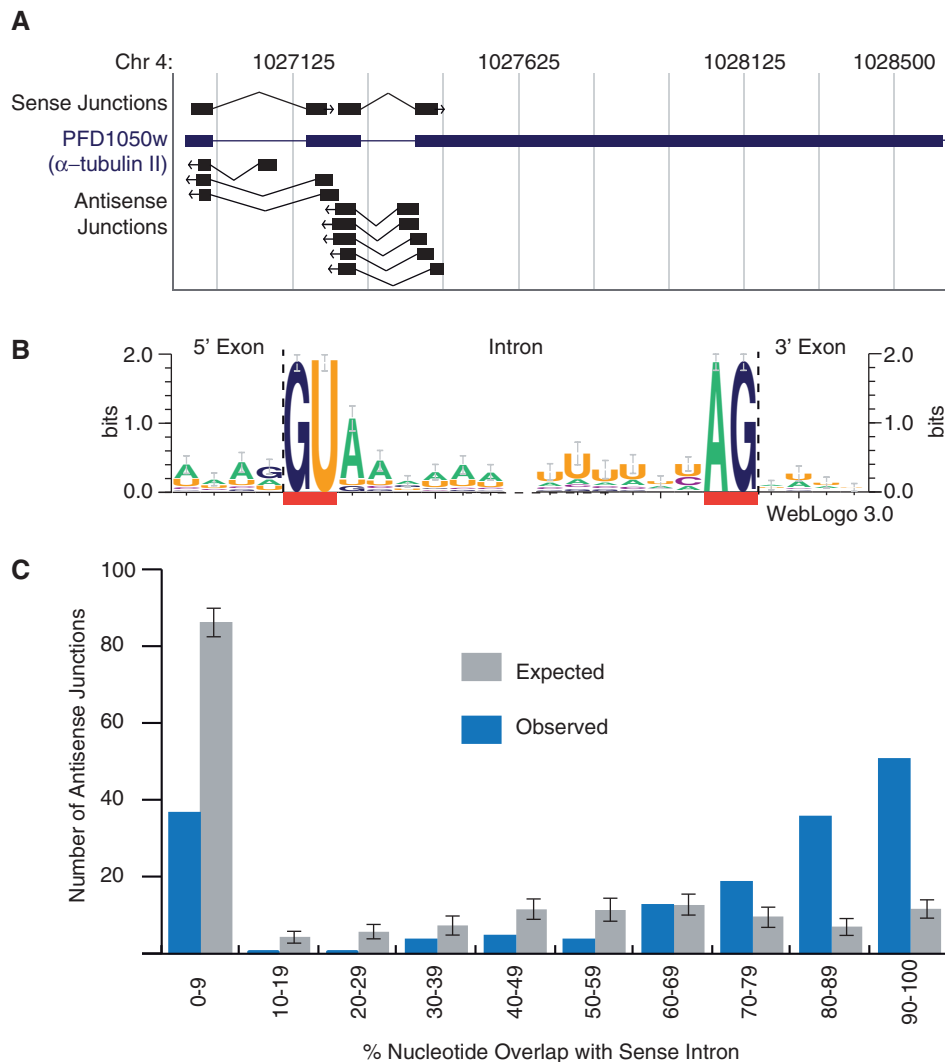
To expand on the initial discovery of antisense conflicts, we searched for all high scoring junctions with at least one intron boundary antisense to an annotated gene model. This analysis differed from the conflicting junctions analysis in two important ways. First, it incorporated antisense junctions that do not conflict with any sense introns (Supplementary Figure S1D). Second, it excluded antisense conflicts between junctions in which neither mapped to a gene model (four independent conflicts), and thus neither could be deemed ‘sense’ or ‘antisense’. In total, this list contains 200 antisense junctions mapping to 149 gene models (Supplementary File S9).

In addition, antisense junctions overlapping 16 of these genes appear to undergo alternative splicing to produce between 38 and 59 different isoforms (example shown in Figure 5A, Supplementary File S7). Weblogos of the 5′- and 3′-splice sites of antisense junctions revealed no significant differences compared to known junctions (Figure 5B), suggesting that antisense junctions arise from the same mechanism as other splice junctions in the transcriptome. No GO terms were significantly enriched within genes with mapped antisense junctions (Supplementary Materials and Methods).

Antisense junctions could derive from either overlap between neighboring gene models encoded on opposite strands of the genome or from unannotated transcripts antisense to gene models. Indeed, 23 antisense junctions could be attributed to overlap between 15 pairs of neighboring annotated genes on opposite strands based on linking junctions or ESTs (Supplementary Table S8). Only one gene pair (PFE1425c/PFE1420w) is annotated as overlapping, while nine had prior EST evidence of overlap. The remaining five pairs had no prior evidence of overlap. Twelve pairs were arranged in a tail-to-tail (overlapping 3′-ends) fashion, while three were arranged in a head-to-head (overlapping 5′-ends) fashion. Several studies have reported a bias toward tail-to-tail overlaps in mammalian genomes (61), although others refute this assertion (23).

Overlap between annotated genes, however, could not explain all antisense junctions observed in the RNA-Seq data. Of the 177 antisense junctions without direct evidence of neighboring gene overlap, 49 map to genes where neighbors on either side are on the same strand. This observation argues strongly for the presence of unannotated transcripts overlapping annotated genes in an antisense manner. We further investigated whether these 177 antisense junctions might belong to coding or non-coding transcripts. Genomic sequence 300 nt upstream and downstream of each junction was merged and translated in all three frames, and the length of the longest open reading frame (ORF) that crossed the junction was assessed. Of 177 junctions, only 16 occurred in an ORF greater than 300 bases long (average exon size in intron-containing genes is 552 bases). It is possible that these antisense junctions connect shorter than average exons, or occur in UTR regions of unannotated genes. It is also possible that many of them belong to non-coding transcripts. Further elucidation of the structure of these antisense transcripts is necessary to determine their primary function.

Interestingly, over 86% of antisense junctions map to intron-containing genes, though only slightly more than half of genes in *P. falciparum* contain introns. This bias is significant, with a binomial probability of  $\sim 3e^{-24}$ . A similar bias was seen in *Arabidopsis thaliana* in tail-to-tail overlapping transcripts (22), and could be explained by preferential overlap between introns in antisense transcripts and introns in sense transcripts (Supplementary Figure S1D). In some cases, multiple antisense introns overlap extensively with multiple sense gene introns in the same gene model, but not with more expansive exon regions (Figure 5A). The observed distribution of overlap



**Figure 5.** Characterization of antisense splice junctions. (A) Schematic of all sense and antisense junctions recovered for PFD1050w ( $\alpha$ -tubulin II). (B) WebLogos of the 5'- and 3'-splice sites of antisense junctions. The height of each letter indicates the preference strength for that nucleotide at each position. (C) Observed and expected distributions of antisense intron overlap with sense introns. The expected distribution was calculated by first determining the probability of encountering a GU (5'-splice site) or an AG (3'-splice site) on the opposite strand of introns versus exons in the genes with mapped antisense junctions. These probabilities guided otherwise random re-placement of each antisense junction within its corresponding gene model. This re-placement was iterated 100 times, with the mean percent of nucleotide overlap with sense introns  $\pm$  standard deviation shown.

with sense introns is highly statistically significant ( $P$ -value of chi-squared test  $<0.001$ ) when compared to the expected distribution from random re-placement of antisense junctions within their associated genes (Figure 5C). This expected distribution was calculated by first determining the probability of encountering a GU (5'-splice site) or an AG (3'-splice site) on the opposite strand of introns versus exons within the group of genes with mapped antisense junctions. These probabilities then guided otherwise random re-placement of each antisense junction within its corresponding gene model, keeping the original length of the antisense intron intact. After re-placement, the distribution of overlaps for simulated antisense introns with sense introns was tallied. This re-placement was iterated 100 times, with the mean percent of nucleotide overlap with sense introns  $\pm$  standard deviation shown. Thus antisense introns appear to not

only overlap intron-containing sense genes significantly more often than expected, but also overlap the intron portions of sense genes significantly more than expected.

## DISCUSSION

Completion and preliminary annotation of the *P. falciparum* genome in 2002 facilitated a series of large-scale experiments designed to illuminate the parasite's biology on a genomic-, transcriptomic- or proteomic-wide level. In pursuit of a thorough understanding of *P. falciparum* blood stage genetic regulation, steady-state gene expression experiments captured its unique, cascade-like transcriptome (1,2). Subsequent genome-wide RNA decay experiments revealed global rapid turn over of RNA in the early hours post-invasion, and then progressively longer transcript half-lives during the remainder of the

blood stage cycle (62). The new splicing events described here reveal additional complexities within the transcriptome not captured by these previous studies, such as alternative splicing, gene overlap and spliced antisense transcripts, and thus, the present study fits into a larger, more dynamic understanding of the transcriptome of *P. falciparum*.

Traditionally, full-length cDNA and EST data have been used for analysis of transcript structure and variants. EST collections in *P. falciparum* have indeed produced increasingly accurate gene models (12,13,43,44). However, no full-length cDNA sequences have been published for *P. falciparum*, and many gene models lack or are incompletely covered by ESTs. RNA-Seq provides the advantage of capturing an entire transcriptome at great depth, enabling detection of low copy number transcripts and variants. However, in its current form, RNA-Seq cannot capture a single transcript molecule from beginning to end. Despite this limitation, the orders-of-magnitude increase in throughput over EST libraries expanded the repertoire of splice junctions known in the parasite by >11% in the present study.

The ability to accurately and sensitively map junction reads from the RNA-Seq data sets proved crucial to our analysis. For this purpose, we used HMMSplicer, an algorithm we developed specifically to overcome the challenges presented by RNA-Seq data and the inherent biases within the *P. falciparum* genome (28). In contrast to previous RNA-Seq studies in *P. falciparum* and other organisms (14,63), we relied only on alignment of junction reads within the genome to detect splice junctions, rather than depending on gene models or ungapped read coverage. Also, HMMSplicer does not use additional assumptions to filter its output junction set, instead scoring each splice junction on the strength of supporting reads. Because a low false positive rate was desired for accurate characterization of splicing in *P. falciparum*, we established an operational HMMSplicer score threshold based on the bimodal distribution of known versus new canonical splice junctions. However, setting this threshold held the disadvantage of excluding some known junctions, and therefore some true new junctions as well. Indeed, our biologically independent validation experiments demonstrated that even lower scoring junctions were more likely than not to represent true splicing events. Although these lower scoring junctions were excluded from downstream analysis in the present study, they remain accessible in the HMMSplicer results (Supplementary Files S1 and S2, upload at <http://plasmodb.org/cgi-bin/gbrowse/plasmodb/>).

We also did not rely on gene models during discovery of alternative splicing. This decision was prompted by several observations within the data. First, there were ambiguous instances in which a junction conflicted with a gene model, but the gene model junction was not recovered within the data set. These instances could potentially represent gene model errors, making it inappropriate to classify them as alternative splicing without additional data. Conversely, areas of the transcriptome with no gene model contained multiple junctions that could not possibly exist within the same transcript (Figure 4A). These intergenic junction

groups clearly exhibit alternative splicing and would have been missed by reliance on gene models. Thus our unbiased approach allowed for more accurate and sensitive detection of alternative splicing events based solely on experimental observation of the conflicting junctions themselves.

Although *P. falciparum* ESTs and even some gene models include non-canonical 5' GC-AG 3' splice junctions, to our knowledge, no study has attempted to identify or characterize non-canonical junctions in *P. falciparum*. We found 12 high scoring 5' GC-AG 3' junctions within the non-canonical junctions results, 5 of which were new. As in other organisms, the 5'-splice site for these junctions has a remarkably high information content compared to the 5'-splice site for canonical *P. falciparum* junctions, perhaps indicating greater reliance on sequence context for recognition of 5' GC splice sites. In particular, the strong preference for G at the fifth position in 5' GC-AG 3' introns is interesting. Although G is strongly preferred at this position in human canonical introns (11), and mutation of this G to other bases reduces splicing fidelity in yeast (64), *P. falciparum* canonical introns have almost no base preference at this position (Figure 1C) (7). At present, it is unclear how complete the list of 5' GC-AG 3' junctions is, given that the percent of splice junctions mapping to those splice sites (~0.1%) remains several fold lower in *P. falciparum* than in other organisms (11). In addition, although filters designed to aid discovery of any additional non-canonical junctions were implemented, manual inspection found no convincing examples. It is possible that despite efforts to limit bias, the filters inadvertently eliminated true positive junctions or manual inspection failed to detect credible non- 5' GU-AG 3' or 5' GC-AG 3' junctions within the data.

Our analysis uncovered not only constitutive and alternative splicing in *P. falciparum*, but also complex transcriptional arrangements in the parasite. Independent validation of new junctions antisense to sense junctions indicates that these are not artifacts of the RNA-Seq technique. Rather, antisense junctions in the data suggest overlap between annotated sense genes and antisense transcripts, some of which appear to be extensions of neighboring annotated genes, while others are likely unannotated. For unknown reasons, antisense splice junctions tend to encompass sense introns more than would be expected by chance. It is unknown if this phenomenon is specific to *P. falciparum* antisense splice junctions, as it has not been explored in other organisms to our knowledge. Perhaps antisense introns must be spliced out in approximately the same area as sense introns to allow transcript pairs to physically interact with one another. Conversely, if the low complexity sequence that comprises *P. falciparum* introns generally does not encode useful information on either strand, it would have to be removed from both sense and antisense transcripts to preserve function. Further inquiry is necessary to distinguish between these hypotheses.

The larger impact of the transcriptome features revealed by the new canonical and 5' GC-AG 3' junctions captured here remains unknown. Consistent with reports

correlating alternative splicing prevalence with organismal complexity (65), alternative splicing events do not appear to be widespread in *P. falciparum* blood stages, affecting 8.6% of intron-containing genes. Although relatively scarce, alternative splicing events in *P. falciparum* may expand important protein functionalities in the organism and may also contribute to crucial post-transcriptional gene regulation—however, it is possible that their impact on parasite biology is minimal. Interestingly, these events appear to occur with almost no pressure to preserve ORFs, as only one-third are predicted to do so, the same proportion expected by chance. We have suggested that alternative splicing events predicted to result in truncated ORFs may be linked to a NMD system in the parasite as a means of gene regulation. It would be interesting to determine if such isoforms decay faster than their corresponding protein-coding isoforms, perhaps by extending previous methods for determining RNA decay rates in *P. falciparum* using high-throughput sequencing. Unfortunately, current RNA decay data in *P. falciparum* does not allow for discrimination between the decay rates of isoforms of the same transcript (62). The observed overlap between sense and antisense transcript pairs of may also contribute to important gene regulation in the parasite by a variety of mechanisms (24). In addition, unannotated antisense transcripts could perform a variety of as-yet-unknown functions that may or may not be restricted to regulation of their sense partners. Unraveling these possibilities in both the symptomatic blood stages of *P. falciparum* as well as in the organism's larger lifecycle will provide an unprecedented understanding of a deadly human pathogen.

## ACCESSION NUMBER

SRA024324.1.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGEMENTS

The authors would like to thank Dr Quinn Mitrovich for technical expertise in devising the junction validation strategy. We would also like to thank Alex Plocik for helpful discussions during article preparation, Dr Polly Fordyce for insightful reorganization of the article and Dr Steven Brenner for analysis advice and suggestions.

## FUNDING

Howard Hughes Medical Institute; National Science Foundation (DGE-0648991). Funding for open access charge: Howard Hughes Medical Institute.

*Conflict of interest statement.* None declared.

## REFERENCES

- Bozdech,Z., Llinás,M., Pulliam,B.L., Wong,E.D., Zhu,J. and DeRisi,J.L. (2003) The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*. *PLoS Biol.*, **1**, e5.
- Le Roch,K.G., Zhou,Y., Blair,P.L., Grainger,M., Moch,J.K., Haynes,J.D., De la Vega,P., Holder,A.A., Batalov,S., Carucci,D.J. *et al.* (2003) Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science*, **301**, 1503–1508.
- Silvestrini,F., Bozdech,Z., Lanfrancotti,A., Di Giulio,E., Bultrini,E., Picci,L., Derisi,J.L., Pizzi,E. and Alano,P. (2005) Genome-wide identification of genes upregulated at the onset of gametocytogenesis in *Plasmodium falciparum*. *Mol. Biochem. Parasitol.*, **143**, 100–110.
- Gardner,M.J., Hall,N., Fung,E., White,O., Berriman,M., Hyman,R.W., Carlton,J.M., Pain,A., Nelson,K.E., Bowman,S. *et al.* (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature*, **419**, 498–511.
- Burtis,K.C. (1993) The regulation of sex determination and sexually dimorphic differentiation in *Drosophila*. *Curr. Opin. Cell Biol.*, **5**, 1006–1014.
- Madsen,J. and Stoltzfus,C.M. (2006) A suboptimal 5' splice site downstream of HIV-1 splice site A1 is required for unspliced viral mRNA accumulation and efficient virus replication. *Retrovirology*, **3**, 10.
- Chakrabarti,K., Pearson,M., Grate,L., Sterne-Weiler,T., Deans,J., Donohue,J.P. and Ares,M. (2007) Structural RNAs of known and unknown function identified in malaria parasites by comparative genomics and RNA analysis. *RNA*, **13**, 1923–1939.
- Upadhyay,R., Bawankar,P., Malhotra,D. and Patankar,S. (2005) A screen for conserved sequences with biased base composition identifies noncoding RNAs in the A–T rich genome of *Plasmodium falciparum*. *Mol. Biochem. Parasitol.*, **144**, 149–158.
- Shankar,J., Pradhan,A. and Tuteja,R. (2008) Isolation and characterization of *Plasmodium falciparum* UAP56 homolog: Evidence for the coupling of RNA binding and splicing activity by site-directed mutations. *Arch. Biochem. Biophys.*, **478**, 143–153.
- Lamond,A.I. (1993) The spliceosome. *Bioessays*, **15**, 595–603.
- Burset,M., Seledtsov,I.A. and Solovyev,V.V. (2000) Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res.*, **28**, 4364–4375.
- Lu,F., Jiang,H., Ding,J., Mu,J., Valenzuela,J., Ribeiro,J. and Su,X. (2007) cDNA sequences reveal considerable gene prediction inaccuracy in the *Plasmodium falciparum* genome. *BMC Genomics*, **8**, 255.
- Li,L., Brunk,B.P., Kissinger,J.C., Pape,D., Tang,K., Cole,R.H., Martin,J., Wylie,T., Dante,M., Fogarty,S.J. *et al.* (2003) Gene discovery in the apicomplexa as revealed by EST sequencing and assembly of a comparative gene database. *Genome Res.*, **13**, 443–454.
- Otto,T.D., Wilinski,D., Assefa,S., Keane,T.M., Sarry,L.R., Böhme,U., Lemieux,J., Barrell,B., Pain,A., Berriman,M. *et al.* (2010) New insights into the blood-stage transcriptome of *Plasmodium falciparum* using RNA-Seq. *Mol. Microbiol.*, **76**, 12–24.
- Bracchi-Ricard,V., Barik,S., Delvecchio,C., Doerig,C., Chakrabarti,R. and Chakrabarti,D. (2000) PpPK6, a novel cyclin-dependent kinase/mitogen-activated protein kinase-related protein kinase from *Plasmodium falciparum*. *Biochem. J.*, **347**, 255–263.
- Muhia,D.K., Swales,C.A., Eckstein-Ludwig,U., Saran,S., Polley,S.D., Kelly,J.M., Schaap,P., Krishna,S. and Baker,D.A. (2003) Multiple splice variants encode a novel adenylyl cyclase of possible plastid origin expressed in the sexual stage of the malaria parasite *Plasmodium falciparum*. *J. Biol. Chem.*, **278**, 22014–22022.
- Saenz,F.E., Balu,B., Smith,J., Mendonca,S.R. and Adams,J.H. (2008) The transmembrane isoform of *Plasmodium falciparum* MAEBL is essential for the invasion of *Anopheles* salivary glands. *PLoS ONE*, **3**, e2287.
- Wentzinger,L., Bopp,S., Tenor,H., Klar,J., Brun,R., Beck,H.P. and Seebeck,T. (2008) Cyclic nucleotide-specific phosphodiesterases of *Plasmodium falciparum*: PpPDE[alpha], a non-essential cGMP-specific PDE that is an integral membrane protein. *Inter. J. Parasitol.*, **38**, 1625–1637.

19. Iriko,H., Jin,L., Kaneko,O., Takeo,S., Han,E., Tachibana,M., Otsuki,H., Torii,M. and Tsuboi,T. (2009) A small-scale systematic analysis of alternative splicing in *Plasmodium falciparum*. *Parasitology Inter.*, **58**, 196–199.
20. Knapp,B., Nau,U., Hundt,E. and Küpper,H.A. (1991) Demonstration of alternative splicing of a pre-mRNA expressed in the blood stage form of *Plasmodium falciparum*. *J. Biol. Chem.*, **266**, 7148–7154.
21. Liu,Q., Mackey,A.J., Roos,D.S. and Pereira,F.C.N. (2008) Evigan: a hidden variable model for integrating gene evidence for eukaryotic gene prediction. *Bioinformatics*, **24**, 597–605.
22. Jen,C., Michalopoulos,I., Westhead,D. and Meyer,P. (2005) Natural antisense transcripts with coding capacity in *Arabidopsis* may have a regulatory role that is not linked to double-stranded RNA degradation. *Genome Biol.*, **6**, R51.
23. RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group) and the FANTOM Consortium. Katayama,S., Tomaru,Y., Kasukawa,T., Waki,K., Nakanishi,M., Nakamura,M., Nishida,H., Yap,C.C., Suzuki,M. *et al.* (2005) Antisense transcription in the mammalian transcriptome. *Science*, **309**, 1564–1566.
24. Faghihi,M.A. and Wahlestedt,C. (2009) Regulatory roles of natural antisense transcripts. *Nat. Rev. Mol. Cell Biol.*, **10**, 637–643.
25. Militello,K.T., Patel,V., Chessler,A., Fisher,J.K., Kasper,J.M., Gunasekera,A. and Wirth,D.F. (2005) RNA polymerase II synthesizes antisense RNA in *Plasmodium falciparum*. *RNA*, **11**, 365–370.
26. Raabe,C.A., Sanchez,C.P., Randau,G., Robeck,T., Skryabin,B.V., Chinni,S.V., Kube,M., Reinhardt,R., Ng,G.H., Manickam,R. *et al.* (2010) A global view of the nonprotein-coding transcriptome in *Plasmodium falciparum*. *Nucleic Acids Res.*, **38**, 608–617.
27. Sorber,K., Chiu,C., Webster,D., Dimon,M., Ruby,J.G., Hekele,A. and DeRisi,J.L. (2008) The Long March: a sample preparation technique that enhances contig length and coverage by high-throughput short-read sequencing. *PLoS ONE*, **3**, e3495.
28. Dimon,M., Sorber,K. and DeRisi,J.L. (2010) HMMSplicer: a tool for efficient and sensitive discovery of known and novel splice junctions in RNA-Seq data. *PLoS ONE*, **5**, e13875.
29. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
30. The Plasmodium Genome Database Collective. (2001) PlasmoDB: an integrative database of the *Plasmodium falciparum* genome. Tools for accessing and analyzing finished and unfinished sequence data. *Nucleic Acids Res.*, **29**, 66–69.
31. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
32. Kent,W.J. (2002) BLAT—The BLAST-Like Alignment Tool. *Genome Res.*, **12**, 656–664.
33. Moreno-Hagelsieb,G. and Latimer,K. (2008) Choosing BLAST options for better detection of orthologs as reciprocal best hits. *Bioinformatics*, **24**, 319–324.
34. Kaufer,N.F. and Potashkin,J. (2000) Analysis of the splicing machinery in fission yeast: a comparison with budding yeast and mammals. *Nucleic Acids Res.*, **28**, 3003–3010.
35. Stevens,S.W., Barta,I., Ge,H.Y., Moore,R.E., Young,M.K., Lee,T.D. and Abelson,J. (2001) Biochemical and genetic analyses of the U5, U6, and U4/U6 x U5 small nuclear ribonucleoproteins from *Saccharomyces cerevisiae*. *RNA*, **7**, 1543–1553.
36. Bessonov,S., Anokhina,M., Will,C.L., Urlaub,H. and Luhrmann,R. (2008) Isolation of an active step I spliceosome and composition of its RNP core. *Nature*, **452**, 846–850.
37. Lardelli,R.M., Thompson,J.X., Yates,J.R. and Stevens,S.W. (2010) Release of SF3 from the intron branchpoint activates the first step of pre-mRNA splicing. *RNA*, **16**, 516–528.
38. Edwards-Gilbert,G., Kim,D., Silverman,E. and Lin,R. (2004) Definition of a spliceosome interaction domain in yeast Prp2 ATPase. *RNA*, **10**, 210–220.
39. Barash,Y., Calarco,J.A., Gao,W., Pan,Q., Wang,X., Shai,O., Blencowe,B.J. and Frey,B.J. (2010) Deciphering the splicing code. *Nature*, **465**, 53–59.
40. Venables,J.P., Koh,C., Froehlich,U., Lapointe,E., Couture,S., Inkel,L., Bramard,A., Paquet,E.R., Watier,V., Durand,M. *et al.* (2008) Multiple and specific mRNA processing targets for the major human hnRNP proteins. *Mol. Cell Biol.*, **28**, 6033–6043.
41. Long,J. and Caceres,J. (2009) The SR protein family of splicing factors: master regulators of gene expression. *Biochem. J.*, **417**, 15.
42. Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Birney,E., Biswas,M., Bucher,P., Cerutti,L., Corpet,F., Croning,M.D.R. *et al.* (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.*, **29**, 37–40.
43. Watanabe,J., Wakaguri,H., Sasaki,M., Suzuki,Y. and Sugano,S. (2007) Comparasite: a database for comparative study of transcriptomes of parasites defined by full-length cDNAs. *Nucleic Acids Res.*, **35**, D431–D438.
44. Florent,I., Porcel,B., Guillaume,E., Da Silva,C., Artiguenave,F., Marechal,E., Brehelin,L., Gascuel,O., Charneau,S., Wincker,P. *et al.* (2009) A *Plasmodium falciparum* FcB1-schizont-EST collection providing clues to schizont specific gene structure and polymorphism. *BMC Genomics*, **10**, 235.
45. Crooks,G.E., Hon,G., Chandonia,J. and Brenner,S.E. (2004) WebLogo: A sequence logo generator. *Genome Res.*, **14**, 1188–1190.
46. D'haeseleer,P. (2006) What are DNA sequence motifs? *Nat. Biotech.*, **24**, 423–425.
47. Clark,F. and Thanaraj,T.A. (2002) Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human. *Hum. Mol. Genet.*, **11**, 451–464.
48. Tarn,W. and Steitz,J.A. (1996) A novel spliceosome containing U11, U12 and U5 snRNPs excises a minor class (AT–AC) intron *in vitro*. *Cell*, **84**, 801–811.
49. Sidrauski,C., Cox,J.S. and Walter,P. (1996) tRNA ligase is required for regulated mRNA splicing in the unfolded protein response. *Cell*, **87**, 405–413.
50. Eckert,K.A. and Kunkel,T.A. (1991) DNA polymerase fidelity and the polymerase chain reaction. *PCR Methods Appl.*, **1**, 17–24.
51. Cocquet,J., Chong,A., Zhang,G. and Veitia,R.A. (2006) Reverse transcriptase template switching and false alternative transcripts. *Genomics*, **88**, 127–131.
52. Shinde,D., Lai,Y., Sun,F. and Arnheim,N. (2003) Taq DNA polymerase slippage mutation rates measured by PCR and quasi-likelihood analysis: (CA/GT)<sub>n</sub> and (A/T)<sub>n</sub> microsatellites. *Nucleic Acids Res.*, **31**, 974–980.
53. Will,C.L., Schneider,C., Hossbach,M., Urlaub,H., Rauhut,R., Elbashir,S., Tuschl,T. and Luhrmann,R. (2004) The human 18S U11/U12 snRNP contains a set of novel proteins not found in the U2-dependent spliceosome. *RNA*, **10**, 929–941.
54. Lopez,M.D., Alm Rosenblad,M. and Samuelsson,T. (2008) Computational screen for spliceosomal RNA genes aids in defining the phylogenetic distribution of major and minor spliceosomal components. *Nucleic Acids Res.*, **36**, 3001–3010.
55. Sun,S., Zhang,Z., Sinha,R., Karni,R. and Krainer,A.R. (2010) SF2/ASF autoregulation involves multiple layers of post-transcriptional and translational control. *Nat. Struct. Mol. Biol.*, **17**, 306–312.
56. Conti,E. and Izaurralde,E. (2005) Nonsense-mediated mRNA decay: molecular insights and mechanistic variations across species. *Curr. Opin. Cell Biol.*, **17**, 316–325.
57. Engebrecht,J.A., Voelkel-Meiman,K. and Roeder,G.S. (1991) Meiosis-specific RNA splicing in yeast. *Cell*, **66**, 1257–1268.
58. Wilhelm,B.T., Marguerat,S., Watt,S., Schubert,F., Wood,V., Goodhead,I., Penkett,C.J., Rogers,J. and Bahler,J. (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*, **453**, 1239–1243.
59. Zhang,Y., Liu,X.S., Liu,Q. and Wei,L. (2006) Genome-wide *in silico* identification and analysis of cis natural antisense transcripts (cis-NATs) in ten species. *Nucleic Acids Res.*, **34**, 3465–3475.
60. Gunasekera,A.M., Patankar,S., Schug,J., Eisen,G., Kissinger,J., Roos,D. and Wirth,D.F. (2004) Widespread distribution of antisense transcripts in the *Plasmodium falciparum* genome. *Mol. Biochem. Parasitol.*, **136**, 35–42.

61. Sun,M., Hurst,L.D., Carmichael,G.G. and Chen,J. (2005) Evidence for a preferential targeting of 3'-UTRs by cis-encoded natural antisense transcripts. *Nucleic Acids Res.*, **33**, 5533–5543.
62. Shock,J.L., Fischer,K.F. and DeRisi,J.L. (2007) Whole-genome analysis of mRNA decay in *Plasmodium falciparum* reveals a global lengthening of mRNA half-life during the intra-erythrocytic development cycle. *Genome Biol.*, **8**, R134.
63. Trapnell,C., Pachter,L. and Salzberg,S.L. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.
64. Fouser,L.A. and Friesen,J.D. (1986) Mutations in a yeast intron demonstrate the importance of specific conserved nucleotides for the two stages of nuclear mRNA splicing. *Cell*, **45**, 81–93.
65. Kim,E., Magen,A. and Ast,G. (2006) Different levels of alternative splicing among eukaryotes. *Nucleic Acids Res.*, **35**, 125–131.
66. Stein,L.D., Mungall,C., Shu,S., Caudy,M., Mangone,M., Day,A., Nickerson,E., Stajich,J.E., Harris,T.W., Arva,A. *et al.* (2002) The Generic Genome Browser: a Building Block for a Model Organism System Database. *Genome Res.*, **12**, 1599–1610.