

UCSF

UC San Francisco Electronic Theses and Dissertations

Title

Genomic approaches to the study of splicing in Plasmodium falciparum and other organisms using high throughput sequencing

Permalink

<https://escholarship.org/uc/item/9zz4q2t9>

Author

Dimon, Michelle Therese

Publication Date

2010

Peer reviewed|Thesis/dissertation

Genomic approaches to the study of splicing in *Plasmodium falciparum* and
other organisms using high throughput sequencing

by

Michelle Dimon

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biological & Medical Informatics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Copyright (2010)

Michelle Dimon

Acknowledgements

I thank my advisor, Dr. Joseph DeRisi, for his support and guidance throughout my graduate work. Joe's enthusiasm for science is infectious. He consistently pushed me to understand every aspect and implication of my research. Also, his Adobe Illustrator skills are legendary.

I would also like to thank the current and former members of the DeRisi lab. Most importantly, I'd like to thank Katherine Sorber, my collaborator in all the work in this thesis. Her enormous skill at the bench is much appreciated, but her discriminating mind and perspective as a biologist has been even more crucial to our work together.

Additionally, I'd like to thank the computational corner of the lab: Sharon Chao, Peter Skewes-Cox, and Graham Ruby for the friendship and useful advice. In addition, I've enjoyed being a member of the Malaria team. The words of encouragement and suggestions have helped me enormously, especially when I was learning how to culture and do wet lab work. Also, I'd like to acknowledge Tara for all her work in keeping the lab running smoothly.

I am also grateful to my other committee members, Andrej Sali and Hao Li, for their advice and suggestions over the years. Their encouragement and perspective has improved my work immensely.

Finally, I wish to acknowledge my family for their unwavering support over the last five years. I would not be here without the encouragement of my parents, who always insisted I should pursue the career that made me happy, whether that was a nuclear physicist or a tollbooth worker. My sister, Kim, can always be counted on to be enthusiastic and proud of my work. My husband, Court, deserves kudos for living with me and supporting me with through the long road that is graduate school. I could not have done this without him; his encouragement and love was indispensable. My daughter, Kate, born during my graduate work, provided infinite joy and a perspective on what is really important in life.

Genomic approaches to the study of splicing in *Plasmodium falciparum* and other organisms using high throughput sequencing

Michelle Dimon

Abstract

In the last five years, high throughput sequencing has revolutionized biological research. The ability to quickly generate millions of short sequence reads enables studies that would have been inconceivable even 10 years ago. This work focuses on RNA-Seq, the application of high throughput sequencing to an organism's transcriptome. We describe a method of library preparation that improves sequence coverage, a new algorithm for detecting splice junctions in the datasets, and finally, application of these techniques to the study of splicing in *Plasmodium falciparum*.

The long march is a technique for Solexa library preparation that increases contig length and target sequence coverage. The long march incorporates a Type IIS restriction enzyme into the sequencing primer adapter. Each round of marching cuts off the initial part of the read and ligates a new adapter downstream, creating overlapping reads. Validation on *P. falciparum* genomic and human hepatitis B virus positive samples showed 39% and 42%, respectively, increases in numbers of bases covered.

Next we developed an algorithm to detect spliced reads crossing exon-exon junctions in RNA-Seq datasets. Our algorithm uses an unbiased approach, relying only on the read dataset and a reference genome, detecting canonical and noncanonical splice junctions. This works by dividing reads in half for initial seeding in the reference genome then using an HMM, trained on the input data, to determine the optimal splice position. Our

algorithm provides a score for each splice junction, which allows researchers to tune the false positive rate to the requirements of their experiment. This approach identifies more splice junctions than currently available algorithms, without a reduction in specificity, when tested on publicly available datasets for *Arabidopsis thaliana*, *Plasmodium falciparum*, and *Homo sapiens*.

Finally, our library preparation technique and splice detection algorithm were used to study splicing in *P. falciparum*. Both our data and publicly available datasets were used to identify splicing events in the blood stages of the parasite. We confirmed 6,678 previously known introns and identified 977 novel introns with canonical splice edges. In addition, we detected 310 alternative splicing events as well as splicing events antisense to known transcripts.

Table of Contents

Chapter 1: Introduction	1
Chapter 2: The long march: a sample preparation technique that enhances contig length and coverage by high-throughput short-read sequencing	17
<i>Introduction</i>	19
<i>Materials and Methods</i>	21
<i>Results</i>	26
<i>Discussion</i>	33
<i>References</i>	37
Chapter 3: HMMSplicer: a tool for efficient and sensitive discovery of known and novel splice junctions in RNA-Seq data	47
<i>Introduction</i>	49
<i>Results</i>	54
<i>Discussion</i>	67
<i>Materials and Methods</i>	70
<i>References</i>	86
Chapter 4: RNA-Seq analysis of splicing in <i>Plasmodium falciparum</i> uncovers novel splice junctions, alternative splicing, and splicing of antisense transcripts	102
<i>Introduction</i>	104
<i>Materials and Methods</i>	108
<i>Results</i>	114
<i>Discussion</i>	127
<i>References</i>	132

List of Tables

Chapter 2

Table 1. Overview of sequencing reads obtained for each sample	46
----------------------------------------------------------------	----

Chapter 3

Table 1. Simulation Results	99
Table 2. Simulation Results	100
Table 3. Datasets	101

Chapter 4

Table 1. Putative <i>Plasmodium falciparum</i> splicing factor orthologs	142
Table 2. Verification of novel junctions in conflict with known junctions.	144
Table 3. Alternative splicing events by type and area in the genome	146

List of Figures

Chapter 2

- Figure 1. Iterative rounds of GsuI digestion and barcoded adapter ligation create nested sub-libraries **Error! Bookmark not defined.**
- Figure 2. The long march produces barcoded, offset reads that aid in contig growth **Error! Bookmark not defined.**
- Figure 3. Marched sub-libraries show significantly increased genome coverage over a wide range of dataset sizes **Error! Bookmark not defined.**
- Figure 4. Theoretical optimization of the long march for *de novo* amplicon assembly **Error! Bookmark not defined.**

Chapter 3

- Figure 1. HMMSplicer pipeline **Error! Bookmark not defined.**
- Figure 2. Algorithm parameters **Error! Bookmark not defined.**
- Figure 3. Simulation results **Error! Bookmark not defined.**
- Figure 4. Overview of HMMSplicer and TopHat results in (a) *A. thaliana*, and (b) *P. falciparum* and (c) *H. sapiens* **Error! Bookmark not defined.**
- Figure 5. Human results compared by transcript abundance **Error! Bookmark not defined.**
- Figure 6. Alternative 5' and 3' splice sites **Error! Bookmark not defined.**
- Figure 7. XBP1 non-canonical intron **Error! Bookmark not defined.**
- Figure 8. Experimental confirmation of predicted *Plasmodium falciparum* splice junctions **Error! Bookmark not defined.**

Chapter 4

- Figure 1. Canonical junctions found by HMMSplicer **Error! Bookmark not defined.**
- Figure 2. Splice site motifs for GC-AG junctions **Error! Bookmark not defined.**
- Figure 3. Validation of conflicting junctions **Error! Bookmark not defined.**

Chapter 1: Introduction

Malaria is an enormous global health burden, responsible for more than 800,000 deaths per year [1]. *Plasmodium falciparum* is one of five causative agents of malaria in humans [2], by far the most deadly. Drug resistance is an enormous problem with malaria. Resistance to chloroquine and sulphadoxine–pyrimethamine caused an increase in childhood mortality due to malaria in Sub-Saharan Africa before the introduction of artemisinin combination therapy (ACT) [3]. The current frontline treatment for uncomplicated malaria is ACT, but resistance has developed in Southeast Asia [4]. New drugs are needed against *P. falciparum*, and a deep understanding of the molecular biology of the organism can help direct drug development efforts.

P. falciparum has a complex lifecycle in two hosts, humans and *Anopholes* mosquito. Infection for a human begins with a bite from an infected mosquito. The sporozoites travel from the mosquito salivary gland to the liver. After 14 days, thousands of merozoites are released into the bloodstream and invade red blood cells, initiating the intraerythrocytic development cycle (IDC). The parasites progress from an initial ring stage into trophozoites then schizonts, then finally each parasite divides into approximately 30 merozoites, which burst from the red blood cell and invade a new red blood cell, continuing the cycle. This lifecycle takes 48 hours and is synchronized; all the parasites burst from their red blood cells at approximately the same time. This is the source of the cyclical fevers characteristic of malaria.

The *P. falciparum* genome was published in 2002 [5]. The genome was challenging to sequence because of the approximately 80% AT content across the genome. There are 5,589 genes in the *P. falciparum* genome and 2,940 of these contain introns. The average exon size is 940 bp (552 bp in genes with introns) and the average intron size is 176 bp, with an average of 1.4 introns per gene. The original genome annotation was done mainly by computer algorithms, with some manual curation, and this annotation has proven quite error prone. It has been estimated that 24% of the genes are incorrectly annotated [6]. There has been a lot of effort in recent years to improve the annotation, but errors remain prevalent.

Initial microarray analyses on the transcriptome of *P. falciparum* have elucidated an intriguing pattern of gene expression through the stages of the IDC [7, 8]. Whereas only about 15% of genes are periodically regulated through the cell cycle in yeast or human HeLa cells, over 80% of *P. falciparum* genes vary in transcript abundance as a function of the life cycle. Most genes have a single peak of transcript abundance, and taken together the transcriptome forms a continuous cascade of gene expression. Subsequent studies have expanded our knowledge further, with studies of RNA decay throughout the lifecycle [9].

One area of *P. falciparum* that is poorly understood is splicing and the regulation of splicing. Although over half the genes in the *P. falciparum* transcriptome are spliced [5], and there are individual examples of genes which have alternative isoforms [10-16], a genome-wide survey of alternative splicing has not been undertaken in this organism. A

natural approach for this study is RNA-Seq [17], applying high-throughput sequencing to an organism's mRNA to analyze the transcriptome at a single nucleotide resolution. To accomplish this study of alternative splicing in *P. falciparum* using RNA-Seq, we developed library prep techniques to improve our coverage of the highly AT-rich genome. Next we developed an algorithm to align RNA-Seq reads back to a *P. falciparum* genome. Finally, we used these techniques to generate and analyze a large library of RNA-Seq *P. falciparum* reads to produce a genome-wide study of alternative splicing in *P. falciparum*.

Background on High Throughput Sequencing

Sanger sequencing has been used since the 1970s to determine the sequence of DNA. The first viral sequence, of phi X174, was sequenced by Sanger in 1977 [18]. The past twenty years have seen an explosion in genome sequencing, with the *Haemophilus influenzae* genome in 1995 [19] and the *Saccharomyces cerevisiae* genome in 1996 [20]. Draft sequences of the human genome were released in 2001 [21, 22] and finished in 2004 [23]. However, the next generation sequencing technologies such as Illumina's Solexa / Genome Analyzer and Roche's 454 sequencing are revolutionizing what can be studied with sequencing. The potential of next generation sequencing technologies has been compared to the early days of PCR [24]. The initial sequence of a human genome took over 10 years and has been reported as costing \$100 million [25]. The genome sequence of James Watson was recently sequenced in 2 months for less than \$1 million using Roche's 454 sequencing technology [25]. Five Illumina Genome Analyzers were used to generate the sequence of an Asian (Chinese) individual in 1-2 months using less

than half a million dollars [26], and a Korean genome was recently published also generated with Illumina sequencing technology [27].

Next generation sequencing technologies, particularly Illumina's Solexa technology, have been applied to the transcriptome of organisms, in a process called RNA-Seq [28]. The idea of studying an organism's transcriptome, or the set of transcripts expressed in a cell, began with microarrays [29-31]. Microarrays allowed researchers to look at gene expression for every gene in an organism quantitatively, thus quickly ascertain what genes were regulated in response to a specific stimulus. EST sequencing, introduced in 1991 provided a way for researchers to sequence many transcripts from an organism [32]. While EST sequencing did not provide the quantitative information or the breadth of microarrays, the ability to look at specific transcripts at a nucleotide resolution complemented microarray technology well. High throughput sequencing provides both the breadth and quantitative data of microarrays and the nucleotide resolution of EST sequencing.

The main challenge presented by high throughput sequencing technologies, for both genome and transcriptome sequencing, is the creation of tools designed to process millions of short read sequences rapidly. The short reads can be assembled, by tools such as Velvet [33], ABySS [34], SSAKE [35], and SOAPdenovo [36]. More frequently, though, the short reads are aligned back to a reference genome. Illumina / Solexa sequencing provides ELAND (unpublished) for this purpose, but faster, more accurate tools have been developed recently. The first generation of alignment tools included

MAQ [37] and SOAP [38]. The second generation relies on the Burrows-Wheeler Transform, yielding significant improvements in speed and accuracy. These tools include BWA [39], SOAPv2 [40], and Bowtie [41].

Background on Splicing

While the alignment tools mentioned can align full-length matches to a reference genome, RNA-Seq reads from organisms with splicing also have reads that cross exon-exon boundaries. When aligned back to the reference genome, these reads appear to show an insert in the genome for the intron sequence, which was spliced out of the mRNA transcript that was sequenced. Splicing is regulated by the spliceosome, which breaks the pre-mRNA transcript at the intron-exon boundaries then joins the two exons together [42]. Motifs within the transcript help guide the splicing, the most important of which are the first few nucleotides of the intron (the 5' splice site) and the last few nucleotides of the intron (the 3' splice site). The consensus motif for the 5' splice site is GTRAGT and the consensus motif for the 3' splice site is YAG, but the motifs are extremely degenerate so only the GT and AG are reliable. These splice edges are found in 98% of human introns [43].

Splicing without GT-AG edges can also occur. The second most frequent splice site edges are GC-AG, which is a degeneracy of GT-AG and processed by the same machinery. Some organisms, including humans, have a minor spliceosome which processes AT-AC splice sites [42]. There are also examples of splicing not mediated by a spliceosome, for example the processing of the *HAC1* transcript in yeast [44]. *HAC1* mRNA is spliced by the protein Ire1p as part of the unfolded protein response [45].

As with many concepts in biology, the simple description of a pre-mRNA transcript being splicing to yield the one version of the final mRNA transcript has proven overly simplistic. Instead, a pre-mRNA transcript can often be spliced in many different ways to yield multiple versions of a gene. The most common types of alternative splicing are alternative 5' edges, alternative 3' edges, and cassette splicing [46]. Alternative 5' and 3' edges occur when an intron can be spliced at multiple different edges. Cassette splicing occurs when an exon is skipped entirely. For example, one version of the transcript may include exons 1, 2, and 3 whereas another version of the transcript includes only exons 1 and 3. Alternative splicing is much more frequent than previously thought. Recent RNA-Seq studies indicate that almost all multi-exonic human transcripts have alternative isoforms [47, 48].

Aligning RNA-Seq reads across exon-exon junctions without bias

Our goal in developing HMMSplicer was to accurately align *P. falciparum* short reads across splice junctions relying on as few assumptions about the biology of splicing as possible. When this project was initiated, there were two realistic tools available: ERANGE [49] and BLAT [50]. BLAT was originally designed to align EST reads, generally 200 to 500 nucleotides long. While it can accurately ‘unsplice’ these longer reads, it generates many false positives with the short reads from high throughput sequencing. Early RNA-Seq experiments using BLAT for alignment required extensive filtering steps. ERANGE had the advantage of being designed for short reads. Unfortunately, it was designed to align 36 bp reads back to a mammalian genome. At this short length, the reads could not be aligned *de novo* back to the genome accurately.

ERANGE took another approach of aligning the reads not to the whole genome but to known exon-exon junctions, thus limiting the search space to improve results.

Neither BLAT nor ERANGE was sufficient for our problem of aligning 40-50 bp sequence reads back to the *P. falciparum* genome. The ERANGE approach was completely opposite our goal of unbiased discovery. The genome annotation in *P. falciparum* is known to have up to 24% incorrect gene models [6]. In addition, the goal of the project was to identify alternative splicing and novel splice junctions. Thus a solution that only found splice junctions already in the genome annotation would not aid us. The basic approach of BLAT provided a better starting point for our project. However, we needed a way to distinguish the BLAT false positives from the true alignments. Previous approaches relied heavily on existing genome annotations. Two studies in yeast used BLAT to find spliced reads. One study divided BLAT results into those matching known annotations and novel splices [51]. In this study, the false positives were manually curated and no true novel splice forms were found (unsurprising in yeast, which has limited and well-characterized splicing). The other study used existing gene models to train a support vector machine (SVM) that was then used to distinguish whether a junction was good [17], again finding no novel splice isoforms in yeast. Another approach used EST sequences and gene models to build an SVM then used spliced reads simulated from a high throughput sequencing set to optimize the alignments [52]. These approaches can detect *de novo* splice junctions, but they rely heavily on existing gene annotations for their training, so they will miss true alignments that differ from known exon-exon junctions. For example, if the splice motif surrounding

a set of alternatively spliced genes is different from the standard splice motif in known genes, these junctions would be missed using this approach.

Rather than using existing gene annotations to score junctions, we wanted to use all the information inherent in the high throughput sequences, without using any biological assumptions. The information in the high throughput sequencing reads includes the sequence and quality score for each nucleotide and the information available from the alignment, including the number of bases on each side of the exon-exon junction and the genomic sequence in the exon and intron boundaries. The biological information that we wanted to avoid including in our scoring algorithm included similarity to known gene models, splice site motifs, and intron lengths. The scoring algorithm included in HMMSplicer, described in detail in Chapter 3, reflects these goals, and is shown to be extremely accurate at judging true and false positive alignments.

After determining our scoring algorithm, we revisited the decision to use BLAT to generate the initial alignments. While BLAT's speed was sufficient in *P. falciparum*, tests on human datasets found it quite slow. Since we were using BLAT only for the initial seeding, and we could make the assumption that the read only crossed a single exon-exon junction, we found we could increase the speed considerably without a loss of sensitivity by dividing the reads in the half and aligning each half with Bowtie. This method does require that read-halves be long enough to match uniquely within the target genome. The original problem solved by ERANGE of aligning 36 bp reads back to a mammalian genome could not use this approach because 18 bases is not long enough for

a unique match within a mammalian genome. But our problem of aligning 40-50 bp reads back to a *P. falciparum* genome is tractable using this approach.

Alongside our goal of limiting biological assumptions, we wanted to make the HMMSplicer algorithm as flexible as possible so users could both change any parameters we used and see what reads were discarded at each filter step. For example, HMMSplicer does filter for canonical splice sites (GT-AG and GC-AG), but the user can adjust which canonical splice sites are used and the junctions with noncanonical splice sites are also provided for analysis.

As a result of our efforts, HMMSplicer sensitively and specifically detects splice junctions in genomes of all sizes with different intron sizes, from *P. falciparum* with a genome of 23 million bases and an average intron size of 300 bases, to the human genome with 3 billion bases and an average intron size of 3413 bases [53]. It can detect noncanonical junctions, as evidenced by the detection of XBP1 junction in the human dataset (the human homolog to the yeast *HAC1* gene [44]).

Since we initiated our project, several tools have been published for detection of spliced reads in RNA-Seq datasets. The most popular tool is TopHat [54]. TopHat, built by the group that developed Bowtie, works by using full-length alignments to build “exon islands”. These islands are then used to create putative gene models, the edges of which are used as the search space for spliced reads in a manner similar to ERANGE. TopHat was designed for mammalian genomes and performs best when the organism has larger

introns and for genes with relatively high coverage. More recently, several algorithms have been published with approaches somewhat comparable to HMMSplicer, including SuperSplat [55], SplitSeek [56], and SpliceMap [57]. Each algorithm has benefits and drawbacks, but none of the algorithms provide the unbiased splice junction alignments that HMMSplicer provides. SuperSplat is unbiased and does detect non-GT-AG alignments, but it requires a perfect match (no errors) and is extremely memory intensive. SplitSeek is similar to HMMSplicer but it is designed for SOLiD reads only and requires at least one read be split evenly across the splice junction, lowering sensitivity. SpliceMap is also similar to HMMSplicer but requires 50 nt long reads. In addition, SpliceMap only detects GT-AG splice junctions without an option to retrieve other splice edges.

Splicing in P. falciparum

With the library preparation techniques and data analysis algorithm completed, we were ready to approach the genome-wide study of splicing in *P. falciparum*. In addition to the Illumina Solexa reads generated in our lab, another group recently published RNA-Seq data on *P. falciparum* [10]. Their analysis included only a cursory look at splicing and alternative splicing, but their data was publicly available, providing us with additional data for our analysis.

As described in Chapter 4, all the datasets were analyzed in the same way, generating both general coverage statistics and HMMSplicer predicted splice junctions. Our score threshold was set fairly conservatively to include many true positives while excluding most false positives. Indeed, experimental validation showed that at least 90.5% of the

splice junctions above the score threshold while 65% of the splice junctions below the score threshold could be confirmed. The splice junctions above the score threshold include known junctions found in PlasmoDB or published ESTs, as well as novel junctions and noncanonical GC-AG junctions. In addition, we find antisense junctions. These junctions contain splice edges that are the reverse complement of the splice edges expected in the region (for example, CT-AC edges in a gene on the forward strand). *P. falciparum* has long been known to have antisense transcripts [58, 59]. These antisense junctions show that these antisense transcripts can be spliced. Our results show that these antisense junctions overlap sense introns more frequently than expected by chance ($p < 0.001$). The purpose of these spliced antisense transcripts will provide fuel for future research.

In summary, this thesis describes a method to improve Solexa library preparation, a method to analyze RNA-Seq data for splice junctions, and the application of these two techniques to the study of splicing in the malaria parasite *P. falciparum*.

References

1. Kappe SHI, Vaughan AM, Boddey JA, Cowman AF (2010) That Was Then But This Is Now: Malaria Research in the Time of an Eradication Agenda. *Science* 328: 862-866.
2. Cox-Singh J, Davis TME, Lee K, Shamsul SSG, Matusop A, et al. (2008) *Plasmodium knowlesi* malaria in humans is widely distributed and potentially life threatening. *Clin. Infect. Dis* 46: 165-171.
3. Korenromp EL, Williams BG, Gouws E, Dye C, Snow RW (2003) Measurement of trends in childhood malaria mortality in Africa: an assessment of progress toward targets based on verbal autopsy. *Lancet Infect Dis* 3: 349-358.
4. Dondorp AM, Yeung S, White L, Nguon C, Day NP, et al. (2010) Artemisinin resistance: current status and scenarios for containment. *Nat Rev Micro* 8: 272-280.
5. Gardner MJ, Hall N, Fung E, White O, Berriman M, et al. (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* 419: 498-511.
6. Lu F, Jiang H, Ding J, Mu J, Valenzuela JG, et al. (2007) cDNA sequences reveal considerable gene prediction inaccuracy in the *Plasmodium falciparum* genome. *BMC Genomics* 8: 255.
7. Bozdech Z, Zhu J, Joachimiak MP, Cohen FE, Pulliam B, et al. (2003) Expression profiling of the schizont and trophozoite stages of *Plasmodium falciparum* with a long-oligonucleotide microarray. *Genome Biol* 4: R9.
8. Bozdech Z, Llinás M, Pulliam BL, Wong ED, Zhu J, et al. (2003) The Transcriptome of the Intraerythrocytic Developmental Cycle of *Plasmodium falciparum*. *PLoS Biol* 1: e5.
9. Llinás M, Bozdech Z, Wong ED, Adai AT, DeRisi JL (2006) Comparative whole genome transcriptome analysis of three *Plasmodium falciparum* strains. *Nucleic Acids Res* 34: 1166-1173.
10. Otto TD, Wilinski D, Assefa S, Keane TM, Sarry LR, et al. (2010) New insights into the blood-stage transcriptome of *Plasmodium falciparum* using RNA-Seq. *Mol. Microbiol* 76: 12-24.
11. Iriko H, Jin L, Kaneko O, Takeo S, Han E, et al. (2009) A small-scale systematic analysis of alternative splicing in *Plasmodium falciparum*. *Parasitol. Int* 58: 196-199.
12. Knapp B, Nau U, Hundt E, Küpper HA (1991) Demonstration of alternative splicing of a pre-mRNA expressed in the blood stage form of *Plasmodium falciparum*. *J. Biol. Chem* 266: 7148-7154.

13. Bracchi-Ricard V, Barik S, Delvecchio C, Doerig C, Chakrabarti R, et al. (2000) PfPK6, a novel cyclin-dependent kinase/mitogen-activated protein kinase-related protein kinase from *Plasmodium falciparum*. *Biochem. J* 347 Pt 1: 255-263.
14. Muhia DK, Swales CA, Eckstein-Ludwig U, Saran S, Polley SD, et al. (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.
15. Saenz FE, Balu B, Smith J, Mendonca SR, Adams JH (2008) The transmembrane isoform of *Plasmodium falciparum* MAEBL is essential for the invasion of *Anopheles* salivary glands. *PLoS ONE* 3: e2287.
16. Wentzinger L, Bopp S, Tenor H, Klar J, Brun R, et al. (2008) Cyclic nucleotide-specific phosphodiesterases of *Plasmodium falciparum*: PfPDEalpha, a non-essential cGMP-specific PDE that is an integral membrane protein. *Int. J. Parasitol* 38: 1625-1637.
17. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, et al. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 320: 1344-1349.
18. Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, et al. (1977) Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* 265: 687-695.
19. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, et al. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269: 496-512.
20. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, et al. (1996) Life with 6000 genes. *Science* 274: 546, 563-567.
21. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. *Nature* 409: 860-921.
22. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. (2001) The sequence of the human genome. *Science* 291: 1304-1351.
23. (2004) Finishing the euchromatic sequence of the human genome. *Nature* 431: 931-945.
24. Metzker ML (2010) Sequencing technologies - the next generation. *Nat. Rev. Genet* 11: 31-46.
25. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, et al. (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature* 452: 872-876.

26. Wang J, Wang W, Li R, Li Y, Tian G, et al. (2008) The diploid genome sequence of an Asian individual. *Nature* 456: 60-65.
27. Ahn S, Kim T, Lee S, Kim D, Ghang H, et al. (2009) The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome Res* 19: 1622-1629.
28. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5: 621-628.
29. Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270: 467-470.
30. DeRisi JL, Iyer VR, Brown PO (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278: 680-686.
31. Lashkari DA, DeRisi JL, McCusker JH, Namath AF, Gentile C, et al. (1997) Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proc. Natl. Acad. Sci. U.S.A* 94: 13057-13062.
32. Morozova O, Hirst M, Marra MA (2009) Applications of new sequencing technologies for transcriptome analysis. *Annu Rev Genomics Hum Genet* 10: 135-151.
33. Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 18: 821-829.
34. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, et al. (2009) ABySS: a parallel assembler for short read sequence data. *Genome Res* 19: 1117-1123.
35. Warren RL, Sutton GG, Jones SJM, Holt RA (2007) Assembling millions of short DNA sequences using SSAKE. *Bioinformatics* 23: 500-501.
36. Li R, Zhu H, Ruan J, Qian W, Fang X, et al. (2010) De novo assembly of human genomes with massively parallel short read sequencing. *Genome Research* 20: 265 - 272.
37. Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* 18: 1851-1858.
38. Li R, Li Y, Kristiansen K, Wang J (2008) SOAP: short oligonucleotide alignment program. *Bioinformatics* 24: 713-714.
39. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25: 1754-1760.

40. Li R, Yu C, Li Y, Lam T, Yiu S, et al. (2009) SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* 25: 1966-1967.
41. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10: R25.
42. Wahl MC, Will CL, Lührmann R (2009) The Spliceosome: Design Principles of a Dynamic RNP Machine. *Cell* 136: 701-718.
43. Stamm S, Riethoven J, Le Texier V, Gopalakrishnan C, Kumanduri V, et al. (2006) ASD: a bioinformatics resource on alternative splicing. *Nucleic Acids Res* 34: D46-55.
44. Yoshida H, Matsui T, Yamamoto A, Okada T, Mori K (2001) XBP1 mRNA is induced by ATF6 and spliced by IRE1 in response to ER stress to produce a highly active transcription factor. *Cell* 107: 881-891.
45. Sidrauski C, Walter P (1997) The Transmembrane Kinase Ire1p Is a Site-Specific Endonuclease That Initiates mRNA Splicing in the Unfolded Protein Response. *Cell* 90: 1031-1039.
46. Nilsen TW, Graveley BR (2010) Expansion of the eukaryotic proteome by alternative splicing. *Nature* 463: 457-463.
47. Wang ET, Sandberg R, Luo S, Khrebtkova I, Zhang L, et al. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature* 456: 470-476.
48. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet* 40: 1413-1415.
49. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5: 621-628.
50. Kent WJ (2002) BLAT--the BLAST-like alignment tool. *Genome Res* 12: 656-664.
51. Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, et al. (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* 453: 1239-1243.
52. De Bona F, Ossowski S, Schneeberger K, Ratsch G (2008) Optimal spliced alignments of short sequence reads. *Bioinformatics* 24: i174-180.
53. Deutsch M, Long M (1999) Intron-exon structures of eukaryotic model organisms. *Nucleic Acids Res* 27: 3219-3228.

54. Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25: 1105-1111.
55. Bryant DW, Shen R, Priest HD, Wong W, Mockler TC (2010) Supersplat -- spliced RNA-seq alignment. *Bioinformatics* Available at: <http://www.ncbi.nlm.nih.gov/pubmed/20410051>. Accessed 7 June 2010.
56. Ameer A, Wetterbom A, Feuk L, Gyllenstein U (2010) Global and unbiased detection of splice junctions from RNA-seq data. *Genome Biol* 11: R34.
57. Au KF, Jiang H, Lin L, Xing Y, Wong WH (2010) Detection of splice junctions from paired-end RNA-seq data by SpliceMap. *Nucleic Acids Res* Available at: <http://www.ncbi.nlm.nih.gov/pubmed/20371516>. Accessed 7 June 2010.
58. Gunasekera AM, Patankar S, Schug J, Eisen G, Kissinger J, et al. (2004) Widespread distribution of antisense transcripts in the *Plasmodium falciparum* genome. *Mol. Biochem. Parasitol* 136: 35-42.
59. Raabe CA, Sanchez CP, Randau G, Robeck T, Skryabin BV, et al. (2010) A global view of the nonprotein-coding transcriptome in *Plasmodium falciparum*. *Nucleic Acids Res* 38: 608-617.

Chapter 2: The long march: a sample preparation technique that enhances contig length and coverage by high-throughput short-read sequencing

This chapter is a reprint from the following reference:

Sorber K, Chiu C, Webster D, Dimon M, Ruby JG, et al. (2008) The long march: a sample preparation technique that enhances contig length and coverage by high-throughput short-read sequencing. PLoS ONE 3: e3495.

Author contributions:

Katherine Sorber, Charles Chiu, and Armin Hekele performed the experiments. Dale Webster, Michelle Dimon and J. Graham Ruby analyzed the data. Joseph DeRisi supervised the research.

Joseph L. DeRisi, thesis advisor

Abstract

High-throughput short-read technologies have revolutionized DNA sequencing by drastically reducing the cost per base of sequencing information. Despite producing gigabases of sequence per run, these technologies still present obstacles in resequencing and *de novo* assembly applications due to biased or insufficient target sequence coverage. We present here a simple sample preparation method termed the “long march” that increases both contig lengths and target sequence coverage using high-throughput short-read technologies. By incorporating a Type IIS restriction enzyme recognition motif into the sequencing primer adapter, successive rounds of restriction enzyme cleavage and adapter ligation produce a set of nested sub-libraries from the initial amplicon library. Sequence reads from these sub-libraries are offset from each other with enough overlap to aid assembly and contig extension. We demonstrate the utility of the long march in resequencing of the *Plasmodium falciparum* transcriptome, where the number of genomic bases covered was increased by 39%, as well as in metagenomic analysis of a serum sample from a patient with hepatitis B virus (HBV)-related acute liver failure, where the number of HBV bases covered was increased by 42%. We also offer a theoretical optimization of the long march for *de novo* sequence assembly.

Introduction

DNA sequencing technology has benefited from tremendous progress over the past several years, with many platforms routinely producing $>10^9$ nucleotides (nt) of data during a single run [59]. Current generation high-throughput sequencers require a library of amplicons from which reads are generated at random by a variety of different methods, including pyrosequencing [57], reversible chain-terminator extension (3), and ligation [58]. Many of these strategies produce relatively short reads, in the range of 36-70 nt [59], compared to traditional Sanger sequencing which routinely produces reads >800 nt in length (6,7). For some applications, such as microRNA analysis (8), ChIP-Seq (9), or SAGE (Serial Analysis of Gene Expression) (10), short reads are sufficient. However, for resequencing known genomes (5) and *de novo* assembly of unknown sequences (11,12), short reads present a bioinformatics challenge and make sufficient target sequence coverage difficult to achieve.

To date, experimental solutions to these difficulties have focused on two approaches: increasing the number of reads produced from a sample or extending read length. Technical advances such as paired-end reads (13,14) or optimization of sequencing platforms with hardware, software, and / or reagent upgrades can increase the number of reads produced from a sample. Alternatively, additional reads can be produced by simply sequencing a sample multiple times. However, reaching satisfactory coverage of target sequences with these solutions is expensive.

Coverage with short-read technologies can also be increased by directly extending read length, which is achieved by increasing the number of synthesis or ligation cycles performed during sequencing. While lengthening reads does not necessarily incur

additional cost, in practice, the signal to noise ratio of current technologies decreases at each cycle much more rapidly than in traditional Sanger sequencing, effectively limiting the number of bases that can be read with an acceptable degree of accuracy (3,15).

We describe and demonstrate here a simple method for improving high-throughput short-read sequencing results using a cost-effective sample preparation technique. This process, termed the “long march,” utilizes a Type IIS restriction enzyme that cleaves DNA distal to its recognition motif (16,17). By embedding this recognition motif in the sequencing primer adapter of the initial amplicon library, iterative rounds of digestion and ligation produce a nested set of sub-libraries for sequencing. While we demonstrate this method using the Illumina (Solexa) GA2 platform, the long march procedure is applicable to any short-read shotgun sequencing system, including the ABI SOLiD and Helicos. We show that the long march increases contig length and absolute coverage (compared to the same number of reads produced without the procedure) using a cDNA library generated from *Plasmodium falciparum*, the protozoan parasite responsible for the most deadly form of human malaria. In addition, we show that the long march can aid in metagenomic analysis of a complex clinical specimen by increasing coverage of a particular pathogen (in this case hepatitis B virus, or HBV, in a serum sample from a patient with acute liver failure) (18). Finally, we provide a theoretical framework for optimizing the long march for *de novo* genome assembly applications, based on relative enzyme efficiencies as well as starting DNA pool complexity. These results suggest that considerable improvements in absolute base coverage may be achieved through relatively simple and cost-effective modifications of high-throughput sequencing sample preparation protocols. In essence, the long march

technique combines the desirable aspects of both shotgun sequencing and directed primer walking to produce substantially greater coverage within the same number of reads and using the same read length

Materials and Methods

Long marching and barcoding bead-bound cDNA

For *Plasmodium falciparum*, 40 μ L bead-bound cDNA aliquots (see Materials and Methods S1) were digested in 1x Fermentas Buffer B and 0.01 mM S-adenosylmethionine with 5 U GsuI (Fermentas International Inc., Burlington, Ontario) for 1 hour at 30°C, then at 65°C for 20 min. The digestion reactions were dephosphorylated as described in Materials and Methods S1, then washed and ligated to adapter “Sol-L-AA-NN” (short-SolL-GsuI-AANN and Sol-Adapter-L-short-phos-AA annealed). All primer sequences can be found in Table S1. Bead aliquots were again washed and resuspended in ddH₂O. 40 μ L was removed for PCR amplification with fullModSolS and Sol primer 1 for 10 cycles (see Materials and Methods S1 for PCR conditions). The remaining 2 aliquots were digested again with GsuI, dephosphorylated, washed, and ligated to adapter “Sol-L-CC-NN” (short-SolL-GsuI-CCNN and Sol-Adapter-L-short-phos-CC annealed). After ligation, the beads were again washed and resuspended, and 40 μ L was removed for PCR amplification with fullModSolS and Sol primer 1 for 10 cycles, while the remaining beads underwent one more round of GsuI digestion, dephosphorylation, washing, and ligation to adapter “Sol-L-TT-NN” (short-SolL-GsuI-TTNN and Sol-Adapter-L-short-phos-TT annealed). The final aliquot was washed after ligation and PCR amplified with fullModSolS and Sol primer 1 for 10 cycles.

For the HBV sample, the long march and barcoding were carried out in an essentially identical fashion to that of *Plasmodium falciparum* with the following modifications: (1) the HBV sample used the adapters “Sol-L-CC-RR” (short-SolL-GsuI-CCRR and Sol-Adapter-L-short-phos-CC annealed), “Sol-L-GG-RR” (short-SolL-GsuI-GGRR and Sol-Adapter-L-short-phos-GG annealed), and “Sol-L-TT-RR” (short-SolL-GsuI-TTTRR and Sol-Adapter-L-short-phos-TT annealed) for march rounds 1 through 3, and (2) PCR amplification of all marched aliquots was carried out for 15 cycles instead of 10 cycles using the PCR conditions described for the initial HBV library in Materials and Methods S1 .

Solexa sequencing of initial and long marched cDNA

For *Plasmodium falciparum*, the initial library and each marched sub-library were clustered on a Solexa flow cell in a separate lane (Illumina, Hayward, CA). For the HBV sample, the initial library and round 3 marched sub-library were clustered with 15 other barcoded clinical samples in one lane. Following cluster generation, Sol-SeqPrimer was annealed to the clusters on the flow cell, and 48 cycles (*P. falciparum*) or 36 cycles (HBV) of single base pair extensions were performed with image capture using an Illumina (Solexa) GA2 sequencer (Illumina, Hayward, CA). The Solexa Pipeline software suite version 0.2.2.6 (Illumina, Hayward, CA) was utilized for base calling from these images. Base called data can be found at <http://derisilab.ucsf.edu/data/longmarch>.

Analysis of sequence data

Illumina's Solexa software ELAND was used to align reads, with the initial two nt of marched sub-library reads masked, to either *Plasmodium falciparum* genome release 5.4 (23) or to the HBV genome (accession number: NC_003977) (21). Any reads that did not match the genomes in a unique position were not considered for further analysis.

Genome-aligned reads that mapped to the same genomic coordinates were then collapsed into one to determine the redundancy of each library.

The percent of *P. falciparum* reads converted to the destination barcode for each round was determined by examining the initial two barcoded nt of the full reads in each lane. For reads with the correct barcode, if the barcode did not match the two bases directly upstream of the genomic alignment, it was considered “definitely barcoded.” If the barcode did match the two bases directly upstream of the genomic alignment, it was considered “possibly barcoded.” The ratio of “definitely barcoded” reads to total reads was calculated as a conservative estimate of barcoding efficiency for each library. The number of “definitely barcoded” reads, plus the number of “possibly barcoded” reads times the barcoding efficiency, gave the estimated number of correctly barcoded reads due to ligation. This number divided by the total number of reads gave the estimated percent of correctly barcoded reads resulting from ligation.

The offset histogram was calculated by comparing the starting positions of the *P. falciparum* reads in each dataset. For the march round 3 line, the upstream reads were half of the location-collapsed reads with no barcode (NN) from the initial library lane and the downstream dataset was an equal number of location-collapsed reads with a TT barcode from the lane marched three times. For the initial library line, half the location-

collapsed reads with no barcode (NN) from the initial library lane were compared with the other half. The offset was counted as the distance from the start of the upstream read to the start of the downstream read.

Contig length for *P. falciparum* was calculated by counting the length of genomic segments covered by at least one read for 400,000 randomly selected reads from the initial library and the round 3 sub-library. Contig lengths were then averaged independently for each library.

Calculation of genome coverage

For both *P. falciparum* and HBV sample libraries, reads from the initial and the round 3 libraries were chosen at random to fill datasets of various fixed sizes. Each dataset was then mapped back to its respective genome (minus the first 2 nt) and the number of genomic bases covered was determined. In order to account for extremely small dataset sizes, HBV datasets were randomly filled and analyzed 1000 times and the coverage results were averaged.

Simulating optimization of the long march for de novo genome assembly

The theoretical probability of a contig-generating match between two sequences (p_m) was calculated as a function of the overlap length between the sequences (O_L). Equal probability of all four nucleotides at each position was assumed. The p_m value was taken as the number of matching sequences (s_m) divided by the number of total sequences (s_t) of length O_L . When only perfect matches were considered, $s_m = 1$ and $s_t = 4^{O_L}$, so $p_m = 1 / 4^{O_L}$. When mismatches were allowed, s_m equaled the number of sequences

within the allowed mismatch distance, which was calculated as described (24). Given a dataset of S unique sequences, the probability of a sequence being spuriously joined with another to form a contig (p_s) was calculated as $p_s = 1 - (1 - p_m)^S$. The probability of at least one sequence in a dataset of size S being spuriously linked to another (p_{st}) was calculated as $p_{st} = 1 - (1 - p_s)^S$. The assumption of a search for overlap between the 3' end of the given read and the 5' ends of the remaining reads was assumed when calculating p_s . Therefore, the value of p_{st} reflected the application of p_s to an all-against-all search in which each sequence could be connected to all others based on either a 5' overlap, a 3' overlap, or both.

Assembly was simulated *in silico* using an abstract amplicon data class. Each amplicon contained a number of step positions numbered from zero through the number of simulated march rounds. A number of amplicon instances was created equal to the simulated amplicon pool complexity. The number of reads obtained was specified for each simulation. For each read, an amplicon instance was selected randomly (assuming even representation of all amplicons in the pool), and a step number was randomly selected for that amplicon with the probabilities of various steps weighted as specified. The resulting amplicon-step combination (read) was added to a collection, and the contents of that collection were evaluated in terms of the redundancy of its contents and the ability to assemble amplicon sequences. Reads were joined into a contig if they derived from adjacent step positions of the same amplicon instance. Unlinked reads formed contigs of length = 1.

Results

The long march uses a Type IIS restriction enzyme to create a series of nested sub-libraries with reduced read redundancy

The long march approach exploits the ability of certain classes of restriction enzymes (Type IIS and some Type III enzymes) to cleave DNA downstream of their recognition motifs (19). These motifs are engineered into the required library adapters to permit iterative rounds of restriction enzyme cleavage and adapter ligation, which produce a set of nested sub-libraries. One can sequence either the sub-library generated at the final round or a combined pool created by mixing successive sub-libraries, depending on the efficiency of cleavage and ligation during the long march.

To initiate the long march procedure, RNA from *Plasmodium falciparum* was reverse transcribed into double-stranded cDNA, biotinylated, and bound to streptavidin beads (see Materials and Methods S1). In construction of the initial library, the adapter containing the sequencing primer hybridization site (Sol-L) was modified before its NN overhang to incorporate the recognition motif of the Type IIS restriction enzyme GsuI (5'-CTGGAG-3'). Each march round began with digestion of the bead-bound cDNA with GsuI, which cleaves double-stranded DNA 14 nt distal to this motif (Figure 1) (16,17). Digested cDNA was then ligated to barcoded Sol-L adapters, and this digestion and ligation process was repeated iteratively to generate three nested sub-libraries in addition to the initial cDNA library. The initial library contained no barcode while subsequent rounds were barcoded AA, CC, and TT, respectively. After 5-10 cycles of PCR, the initial library and each sub-library was clustered and sequenced in a separate Illumina (Solexa) GA2 flow cell lane.

The resulting 48bp sequence reads were aligned to the *P. falciparum* genome (23Mb) using Illumina's ELAND software (20). This analysis yielded the working dataset of genome-aligned reads presented in Table 1 and all subsequent analysis is based on this dataset unless otherwise noted.

In order to estimate the redundancy of each library, reads aligned to the genome were collapsed by location – that is, reads that mapped to the same genomic coordinates were merged into one. Location collapse was used rather than sequence-based collapse to discount aligned reads with sequencing errors. While the genome-aligned reads from the initial library collapsed to 25.7% of the original dataset (an average of 3.89 reads collapsed into one), the genome-aligned reads from the round 3 sub-library collapsed less, to 38.2% of the original dataset (an average of 2.62 reads collapsed into one) (Table 1). These results indicate that the long march reduced the redundancy of the initial cDNA library.

Marching creates offset overlapping reads and longer average contigs

The first two nucleotides of each read from the three *P. falciparum* sub-libraries were analyzed to determine the fraction of reads in each pool that successfully ligated to the appropriate barcoded adapter (Figure 2A). The first round of digestion and ligation, which should have added an AA barcode to each cDNA molecule, resulted in 91% of sequenced reads possessing an AA barcode. After adjusting for reads beginning with AA by chance instead of by ligation, we estimated that 89% of reads from the first round of marching received a barcoded adapter (see Materials and Methods). The second round of marching resulted in 76% CC barcodes (~76% from barcoded adapter ligation), while the

third round of marching resulted in 75% TT barcodes (~71% from barcoded adapter ligation). The high percentage of correctly barcoded reads from each marched sub-library confirms that significant decreases in digestion and ligation efficiency did not occur over three rounds of the long march procedure.

Successful ligation of the barcoded adapters to each sub-library does not necessarily indicate that amplicons were iteratively marched forward. To assess how well the long march succeeded in producing offset, overlapping reads along library amplicons, the genome locations of successfully barcoded reads from the final round of digestion and ligation and non-barcoded reads from the initial library were compared. In cases where a read from the final round mapped downstream of a read from the initial library, the distance between the 5' termini was measured (Figure 2B). In an ideal long march, where both digestion and ligation efficiency are 100%, this comparison would yield a histogram of alignments with one offset peak at 38bp (14bp+12bp+12bp) corresponding to molecules three steps removed from the original amplicon. While GsuI cuts 14bp into the cDNA (16,17), the portion removed in rounds 2 and 3 contained a two nucleotide barcode that did not match the genome, thus reducing the effective offset to 12bp for those rounds. However, because the efficiency of each round was not 100%, three peaks emerged, representing cDNA that was successfully digested and ligated once, twice, or all three times (Figure 2B). The first (14nt) and second (26nt) offset peaks each displayed a distinct shoulder two nucleotides 5' of the expected peak, because some molecules were not successfully ligated to the unbarcoded adapter initially but were later ligated to barcoded adapters, leading to a first step of 12bp, rather than 14bp. To control for chance offset unrelated to the long march protocol, the same analysis was performed

comparing half of the reads from the initial library to the other half. This analysis yielded no offset peaks, indicating that the long march procedure was responsible for the peaks observed at 14bp, 26bp, and 38bp.

The ability to construct long contigs is important in both resequencing and *de novo* assembly applications. Therefore, the average contig sizes for the initial and the round 3 libraries were calculated using 400,000 reads each. Contigs were defined as continuous stretches of the *P. falciparum* genome covered by at least one read. The long march procedure increased the average contig size from 59 nt to 69 nt. In addition, the long march resulted in more exceptionally long contigs due to its ability to connect shorter contigs by covering previously inaccessible intervening sequence. The final sub-library generated 17 contigs >1000 nt, the longest of which was 4952 nt, whereas the initial library generated only 7 contigs >1000 nt, the longest of which was 1630 nt. Library coverage for PF14_0572 (a “hypothetical protein” gene located on the minus strand of chromosome 14 from nt positions 2,450,143 to 2,450,743) demonstrated the benefit to contig assembly provided by the long march (Figure 2C). Without the series of overlapping marched reads indicated at the bottom, the region from 2,450,594 to 2,450,621 remained unsequenced and the contigs on either side were discontinuous. However, the additional information gained from sequencing these adjacent marched reads covered the previous gap and stitched the two contigs together into a much longer total covered area.

The long march increases sequence coverage

In addition to contig size, the advantage to total genome coverage provided by the long march was examined. Several datasets of randomly sampled genome-aligned reads from the round 3 sub-library and from the initial library were mapped back to the *P. falciparum* genome and the number of genomic bases covered by at least one read was measured for each dataset (Figure 3A). Even with a small dataset of 50,000 reads, the round 3 sub-library covered 35% more genomic bases (898,625 nt) than the initial library (664,114 nt). As the number of reads in each dataset grew, so too did the difference in coverage. At 500,000 reads apiece, the marched sub-library vastly outpaced the initial library by covering an additional 1.1 million bases, an increase in coverage of 39%.

The long march protocol was also applied to RNA extracted from a serum specimen from a patient with HBV-related acute liver failure (“HBV sample”) in order to assess its applicability to metagenomic analysis. 36bp reads from the initial library as well as the round 3 sub-library were aligned to the HBV genome (3.2kb) using ELAND (see Materials and Methods) (21). Sequencing of the round 3 sub-library generated a greater percentage of location-collapsed HBV reads than were generated by sequencing the corresponding initial library (Table 1). This trend translated to enhanced genome coverage of HBV – with a dataset of 300 genome-aligned reads, the round 3 sub-library covered 42% more genomic bases (1828 nt) than the initial library (1284 nt) (Figure 3B). Thus the long march increases coverage of a target genome in both resequencing and metagenomic contexts.

Simulating optimization of the long march for de novo genome assembly

We used theoretical considerations to assess the utility of the long march protocol for *de novo* genome or metagenome assembly as well. For such assembly to be reliable, the length of overlap between any two reads must be sufficient to identify their common origin (22). In the initial *P. falciparum* library, the extent of overlap between reads decayed exponentially (Figure 2B) and therefore included many instances of both insufficient overlap for *de novo* assembly and excess overlap for minimal contig extension. In the long march procedure, a step size can be selected that creates the minimum overlap between adjacent steps necessary for correct assembly given the read length and dataset size. To avoid spurious joining, datasets with many unique sequences required longer overlaps than those with few unique sequences (Figure 4A).

Modeling and simulation of the assembly process revealed amplicon library complexity to be critical to the assembly of marched reads into contigs. The benefit gained from optimization of overlap length requires the sequencing of all steps from a given library amplicon within a reasonable number of reads. With increasing complexity of the template pool, this stipulation becomes less likely. Given a dataset of one million randomly-selected reads and assuming that only adjacent steps have enough overlap to be unambiguously assembled, the majority of reads could not be joined into contigs of ≥ 2 steps until the pool complexity was reduced to $< 200,000$ amplicons (Figure 4B). Reduction of pool complexity also generated higher read redundancy (Figure 4C), the error-correcting potential of which would permit lower mismatch tolerances during assembly, in turn reducing the probability of spurious joining (Figure 4D). Thus, a

balance must be struck with the long march in *de novo* assembly applications between genome coverage and contig assembly.

In the above simulations, equal probability of generating a read from any step along an amplicon was assumed. However, the true distribution of sequencing substrates among march steps reflects the cleavage/ligation efficiency during the long march. In simulated sequencing of a round 3 sub-library, the calculated abundance of reads derived from the Nth step (where N can be 0, 1, 2, or 3) was biased towards high N values when cleavage/ligation efficiencies were high and towards low N values when cleavage/ligation efficiencies were low (Figure 4E). Either of these scenarios negated the benefits of marching because few adjacent steps from the same amplicon were sequenced. The most even distribution of reads along march steps was produced with intermediate cleavage/ligation efficiencies (Figure 4E). Simulation of contig assembly using a cleavage/ligation efficiency of 0.5 resulted in fewer full-length contigs, but also fewer unjoined reads, than was produced given an artificially even distribution of reads across all march steps (Figure 4F; compare to Figure 4B).

The possibility of guiding contig assembly by applying a unique barcode to each round of marching was also considered. Such tagging would reduce the probability of misassembling reads by reducing the number of candidate reads for each step (Figure 4A), but would only be effective if reads with barcodes corresponding to the Nth march round also represented the Nth step. The failure of a molecule to cleave/ligate at one round of marching would result in the Nth step receiving a tag from round N+1 and prevent its proper assembly with reads from the N-1 step. Generally, the use of barcodes to guide assembly was not predicted to be useful due to the low frequency with which

this requirement would be met, especially at the intermediate cleavage/ligation efficiencies yielding the most uniform distribution of reads across steps (Figure 4G).

Discussion

Although the cost per base provided by short-read sequencing technologies, such as Illumina, SOLiD, and Helicos is at present far lower than longer read sequencing technologies, like 454 or Sanger sequencing, shorter read lengths pose significant challenges for resequencing and *de novo* assembly applications. The long march overcomes these challenges by extending the average contig length and significantly increasing the target sequence coverage obtained from high-throughput short-read sequencing technologies without the cost of obtaining more reads per sample or the high error rate of directly extending read lengths. High-throughput sequencing platforms generally require the addition of adapters to the ends of DNA fragments. The long march utilizes repeated cycles of Type IIS restriction enzyme cleavage and adapter ligation to allow extended sequencing of each library amplicon without loss of gene expression information. We have demonstrated the utility of the long march in the context of transcriptome resequencing (*Plasmodium falciparum*), as well as in the context of clinical specimen metagenomics (HBV). We have also provided a theoretical framework for the application of the long march to *de novo* genome assembly.

The long march protocol capitalizes on amplicon library redundancies resulting from biases introduced during sample preparation (in our case, random-primed cDNA synthesis followed by PCR library amplification) (25). These redundancies typically result in wasteful sequencing of multiple identical short reads derived from the ends of

identical amplicons. For the *Plasmodium falciparum* and HBV samples described here, the long march extended the amount of genome coverage within a dataset of a fixed number of reads, even when that dataset was relatively small. This extension in genome coverage stems from narrowing the dynamic range of individual nucleotide coverage, since redundant reads from the initial libraries were distributed over a longer distance after the libraries were marched.

In metagenomic analysis, short-read redundancy can obscure the identities of the organisms present in the sample. Characterization of microbial diversity and function from metagenomic sequence data is dependent on the identification of homology to known biological sequence (26). Longer contigs permit more effective detection of genetic homology to known sequences by use of BLASTN or TBLASTX (27,28). The availability of greater coverage and longer contigs from the long march improves the likelihood of successful alignment and thus discovery of both known and novel organisms in a heterogeneous metagenomic sample.

The ability to assemble overlapping reads into reliable contigs is also crucial for *de novo* genome sequencing applications. With standard amplicon libraries, chance is relied upon to produce reads with sufficient overlap for assembly, and thus short-read datasets pose particular challenges by limiting the amount of overlap obtainable between any two reads. The long march allows read overlaps to be biased toward lengths sufficient for accurate assembly but also conservative enough to promote contig growth. Informed choice of restriction enzyme allows adjustment of the procedure's step size to facilitate accurate assembly of a predicted number of unique sequences. Also, in order to capture the adjacent march steps from a given amplicon necessary for contig assembly,

library complexity, as well as cutting and ligation efficiency, must be taken into account. Reduction of library complexity may be required in order to capture enough adjacent march steps to enhance assembly within a reasonable number of reads. If a high cleavage and ligation efficiency (>80%) is achieved, bias toward sequencing only the last march steps of each amplicon can be counteracted by sequencing a pool of the marched sub-libraries from each round, rather than sequencing only the final round sub-library. However, low cleavage and ligation efficiency (<20%) cannot be overcome so easily. While low efficiencies do result in some gain in target sequence coverage (data not shown), both the restriction and ligation enzymes used for long march should be tested for robust activity before beginning the procedure.

The long march protocol described here was not optimized for a particular application. Because the long march relies only on minor modifications to adapter sequence and an appropriate Type IIS or Type III restriction enzyme, it can be readily customized for a variety of applications. Here, marching was carried out for 3 rounds; the only theoretical limit to the number of iterative rounds is the length of the starting amplicons. Also, the restriction enzyme GsuI (5'-CTGGAG-3'; 16/14) (16,17) was chosen arbitrarily; another restriction endonuclease could be used, such as the Type III restriction enzyme EcoP151, which cleaves at a site much further downstream than GsuI (5'-CAGCAG-3'; 27/25) (29). For these studies, long march rounds were tagged using a 2 nt DNA barcode encoded within the adapter sequence. However, the use of DNA barcodes also has the potential to allow multiple samples to be individually coded, and then sequenced simultaneously without physical separation. This approach is appropriate in applications where only a fixed depth of sequencing is required (e.g. detection of small

nucleotide polymorphisms (SNPs); resequencing of small genomes or genomic subregions; pathogen detection), and / or where multiplexing of samples makes high-throughput sequencing more cost-effective.

Acknowledgements

The HBV sample was graciously provided as part of an ongoing study of etiologies of acute liver failure by Dr. Tim Davern (UCSF). We thank Alexander Greninger and Peter Skewes-Cox for expert technical assistance.

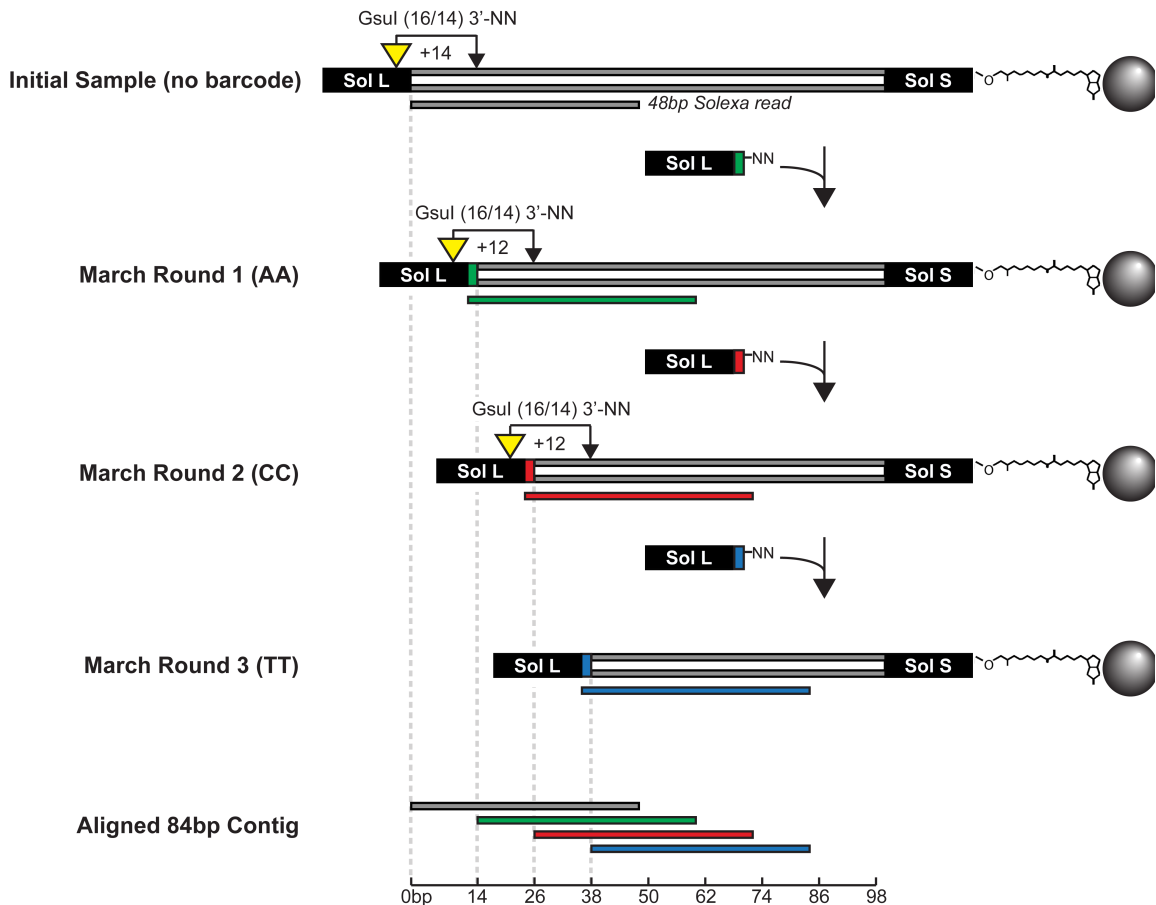
References

1. Holt, R.A. and Jones, S.J. (2008) The new paradigm of flow cell sequencing. *Genome Res*, 18, 839-846.
2. Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlen, M. and Nyren, P. (1996) Real-time DNA sequencing using detection of pyrophosphate release. *Anal Biochem*, 242, 84-89.
3. Seo, T.S., Bai, X., Ruparel, H., Li, Z., Turro, N.J., et al. (2004) Photocleavable fluorescent nucleotides for DNA sequencing on a chip constructed by site-specific coupling chemistry. *Proc Natl Acad Sci U S A*, 101, 5488-5493.
4. Shendure, J., Porreca, G.J., Reppas, N.B., Lin, X., McCutcheon, J.P., et al. (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, 309, 1728-1732.
5. Pop, M. and Salzberg, S.L. (2008) Bioinformatics challenges of new sequencing technology. *Trends Genet*, 24, 142-149.
6. Sanger, F., Nicklen, S. and Coulson, A.R. (1992) DNA sequencing with chain-terminating inhibitors. 1977. *Biotechnology*, 24, 104-108.
7. Salas-Solano, O., Carrilho, E., Kotler, L., Miller, A.W., Goetzinger, W., et al. (1998) Routine DNA sequencing of 1000 bases in less than one hour by capillary electrophoresis with replaceable linear polyacrylamide solutions. *Anal Chem*, 70, 3996-4003.
8. Hafner, M., Landgraf, P., Ludwig, J., Rice, A., Ojo, T., et al. (2008) Identification of microRNAs and other small regulatory RNAs using cDNA library sequencing. *Methods*, 44, 3-12.
9. Mardis, E.R. (2007) ChIP-seq: welcome to the new frontier. *Nat Methods*, 4, 613-614.
10. Wakaguri, H., Yamashita, R., Suzuki, Y., Sugano, S. and Nakai, K. (2008) DBTSS: database of transcription start sites, progress report 2008. *Nucleic Acids Res*, 36, D97-101.
11. Chaisson, M., Pevzner, P. and Tang, H. (2004) Fragment assembly with short reads. *Bioinformatics*, 20, 2067-2074.
12. Whiteford, N., Haslam, N., Weber, G., Prugel-Bennett, A., Essex, J.W., et al. (2005) An analysis of the feasibility of short read sequencing. *Nucleic Acids Res*, 33, e171.
13. Korbel, J.O., Urban, A.E., Affourtit, J.P., Godwin, B., Grubert, F., et al. (2007) Paired-end mapping reveals extensive structural variation in the human genome. *Science*, 318, 420-426.
14. Siegel, A.F., van den Engh, G., Hood, L., Trask, B. and Roach, J.C. (2000) Modeling the feasibility of whole genome shotgun sequencing using a pairwise end strategy. *Genomics*, 68, 237-246.
15. Mashayekhi, F. and Ronaghi, M. (2007) Analysis of read length limiting factors in Pyrosequencing chemistry. *Anal Biochem*, 363, 275-287.
16. Janulaitis, A., Bitinaite, J. and Jaskelviene, B. (1983) A new sequence-specific endonuclease from *Gluconobacter suboxydans*. *FEBS Lett*, 151, 243-247.

17. Petrusyte, M., Bitinaite, J., Menkevicius, S., Klimasauskas, S., Butkus, V., et al. (1988) Restriction endonucleases of a new type. *Gene*, 74, 89-91.
18. Wai, C.T., Fontana, R.J., Polson, J., Hussain, M., Shakil, A.O., et al. (2005) Clinical outcome and virological characteristics of hepatitis B-related acute liver failure in the United States. *J Viral Hepat*, 12, 192-198.
19. Roberts, R.J., Belfort, M., Bestor, T., Bhagwat, A.S., Bickle, T.A., et al. (2003) A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. *Nucleic Acids Res*, 31, 1805-1812.
20. Gardner, M.J., Hall, N., Fung, E., White, O., Berriman, M., et al. (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature*, 419, 498-511.
21. Okamoto, H., Imai, M., Shimozaki, M., Hoshi, Y., Iizuka, H., et al. (1986) Nucleotide sequence of a cloned hepatitis B virus genome, subtype ayr: comparison with genomes of the other three subtypes. *J Gen Virol*, 67 (Pt 11), 2305-2314.
22. Warren, R.L., Sutton, G.G., Jones, S.J. and Holt, R.A. (2007) Assembling millions of short DNA sequences using SSAKE. *Bioinformatics*, 23, 500-501.
23. Stoeckert, C.J., Jr., Fischer, S., Kissinger, J.C., Heiges, M., Aurrecochea, C., et al. (2006) PlasmoDB v5: new looks, new genomes. *Trends Parasitol*, 22, 543-546.
24. Knight, R. and Yarus, M. (2003) Analyzing partially randomized nucleic acid pools: straight dope on doping. *Nucleic Acids Res*, 31, e30.
25. Mathieu-Daude, F., Welsh, J., Vogt, T. and McClelland, M. (1996) DNA rehybridization during PCR: the 'Cot effect' and its consequences. *Nucleic Acids Res*, 24, 2080-2086.
26. Wommack, K.E., Bhavsar, J. and Ravel, J. (2008) Metagenomics: read length matters. *Appl Environ Microbiol*, 74, 1453-1463.
27. 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.
28. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25, 3389-3402.
29. Hadi, S.M., Bachi, B., Shepherd, J.C., Yuan, R., Ineichen, K., et al. (1979) DNA recognition and cleavage by the EcoP15 restriction endonuclease. *J Mol Biol*, 134, 655-666.

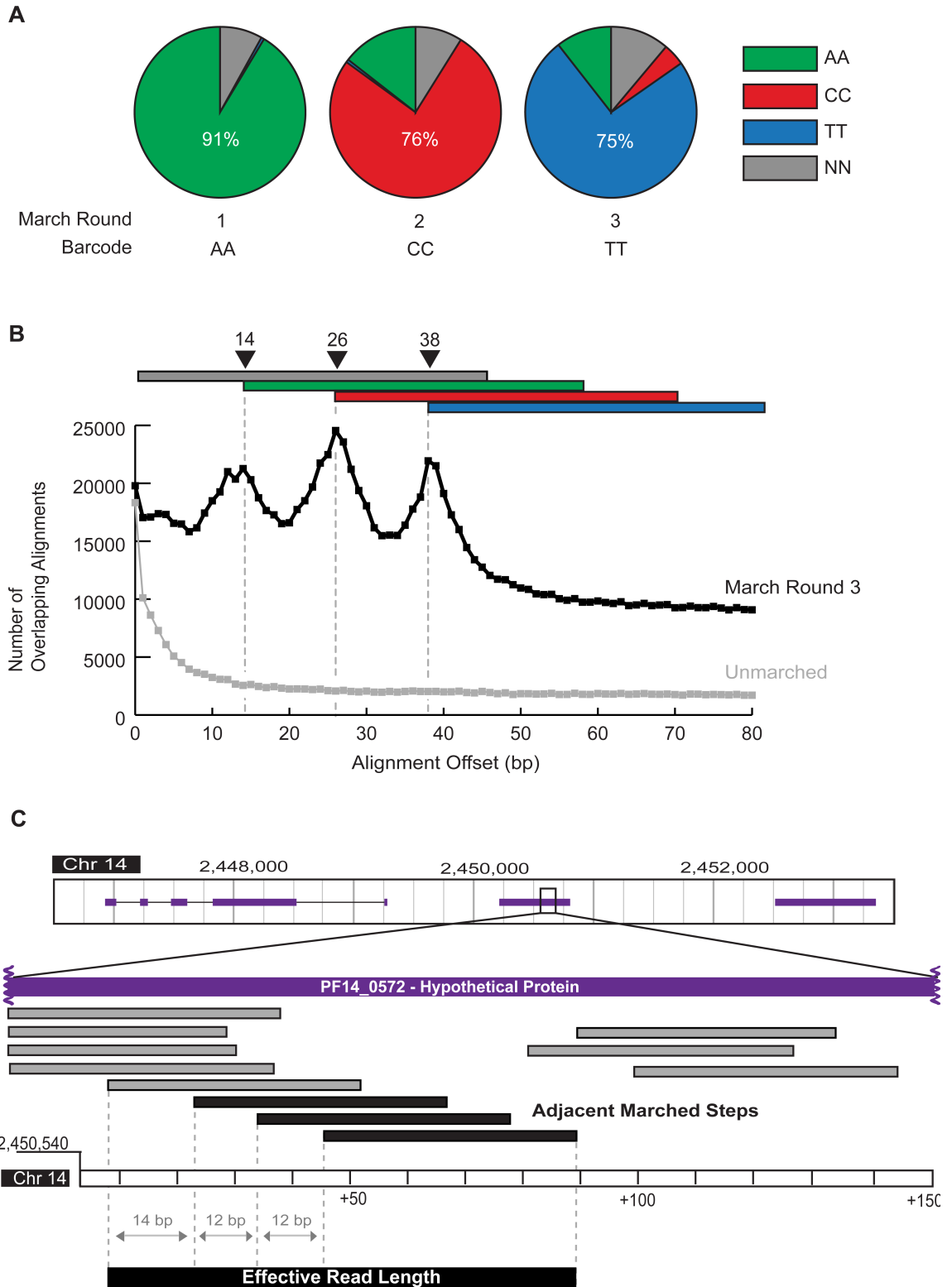
Figures

Figure 1. Iterative rounds of GsuI digestion and barcoded adapter ligation create nested sub-libraries



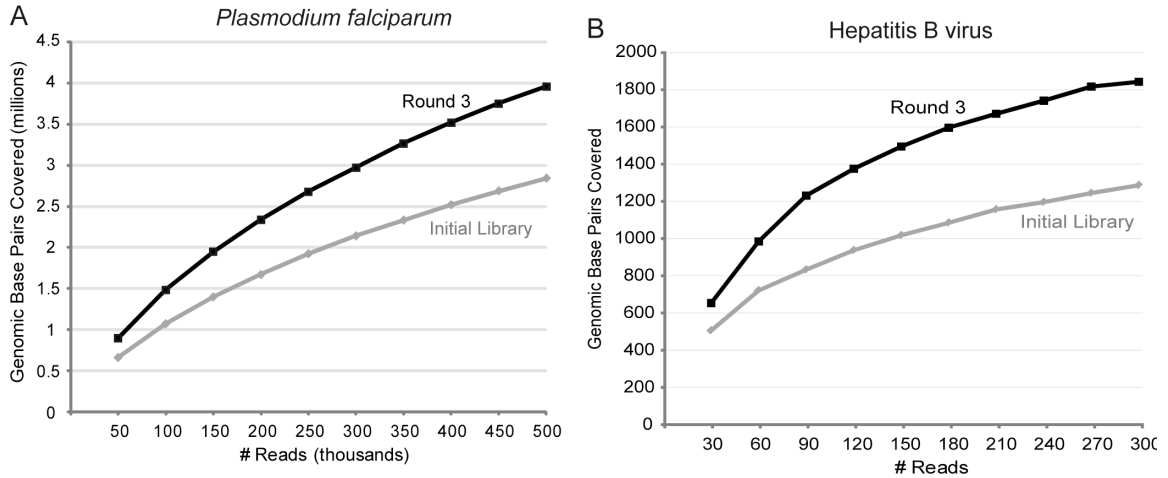
Adapter flanked cDNA molecules are attached to streptavidin beads via biotin modification of the Sol-S adapter. Yellow triangles indicate the GsuI recognition motif engineered into the Sol-L adapter, while the connected black arrow represents the distal cut site. Adapter barcodes and corresponding reads are classified as AA (green), CC (red), or TT (blue). Reads from the initial library and all three long march steps are aligned to form an 84bp contig.

Figure 2. The long march produces barcoded, offset reads that aid in contig growth



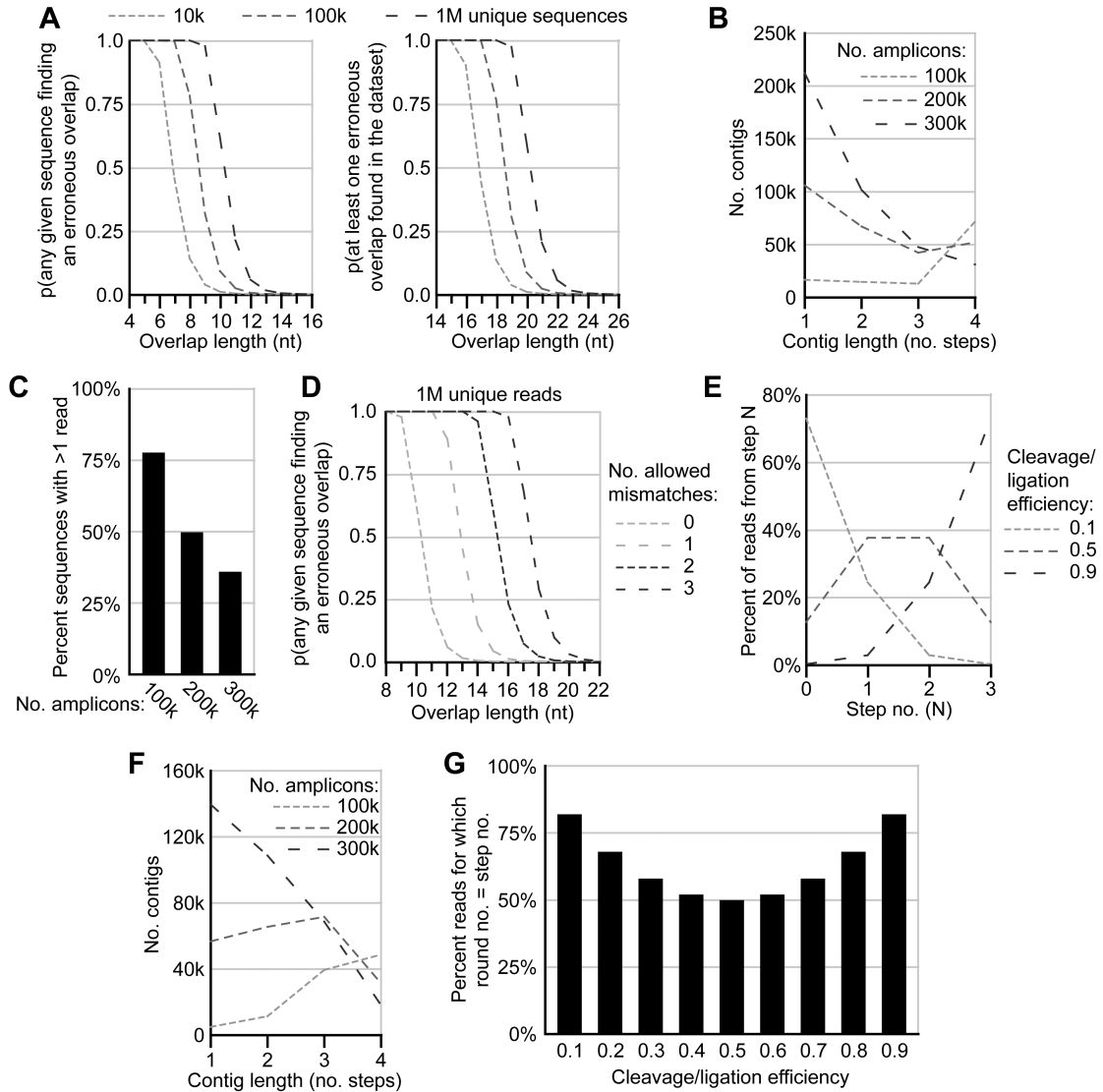
- (A) Barcodes for each round of the long march. The first two bases, masked during genomic alignment, were analyzed for all reads aligning to the *P. falciparum* genome. Barcodes are classified as AA (green), CC (red), TT (blue) and NN (gray), where NN represents any barcode other than AA, CC, or TT. For each round of marching, the dominant barcode was that of the adapter added during that round.
- (B) Histogram of offset, overlapping alignments between 400,000 reads from the round 3 sub-library and 400,000 reads from the initial library. Reads were aligned to the *P. falciparum* genome and the difference between the starting positions of their 5' termini was measured in cases where a round 3 read mapped distal to an initial library read. The resulting three peaks represent reads successfully marched once, twice, or three times. The gray line demonstrates that similar analysis of two pools of 400,000 reads from the initial library show no offset peaks.
- (C) Example of contig joining by adjacent marched reads from the same amplicon. A segment of *P. falciparum* chromosome 14 from 2,450,540 to 2,450,690 (representing a portion of the “hypothetical protein” gene PF14_0572) demonstrates the long march’s utility in increasing contig size. Reads from all four libraries mapping to the area are shown. The four bottom reads derive from the libraries marched zero, one, two, and three times, respectively. While the gray reads cover much of the region shown, the adjacent marched steps from the last gray amplicon, shown in black, are required to cover the entire area and stitch together neighboring contigs.

Figure 3. Marched sub-libraries show significantly increased genome coverage over a wide range of dataset sizes



Identical numbers of genome-aligned reads were randomly sampled from the round 3 sub-libraries and the initial libraries to simulate varying degrees of sequencing depth. The number of genomic base pairs covered by at least one read (y axis) was computed and plotted against the number of randomly selected input reads (x axis) for A) *Plasmodium falciparum* and B) hepatitis B virus (HBV) samples. Because of the small dataset sizes for HBV, each dataset of a given size was randomly filled and analyzed 1000 times; graphed coverage is an average for those datasets.

Figure 4. Theoretical optimization of the long march for *de novo* amplicon assembly



(A) Effect of overlap length on the probability of erroneous assembly of non-overlapping reads. For datasets with the indicated numbers of unique sequences, the probability was calculated of each sequence being erroneously joined to another in the dataset (left) or of at least one read in the dataset being erroneously joined to another (right).

- (B) Effect of initial pool complexity on the length of contigs. For each indicated number of amplicons in the initial pool, a simulation was performed assuming 1 million reads, and contigs were built by joining adjacent reads (see Methods). Each distribution of contig lengths, expressed in number of unique sequences assembled into the contig, was derived from a single simulation.
- (C) Effect of initial pool complexity on dataset redundancy. Simulations were performed as in (B) for each of the indicated amplicon pool complexities, and the fraction of unique sequences that were observed more than once is indicated.
- (D) Effect of allowed mismatches on the probability of erroneous assembly of non-overlapping reads. Probabilities were calculated assuming datasets of 1 million unique sequences. Allowed mismatches were single-nucleotide substitutions in the context of an ungapped alignment.
- (E) Effect of cleavage/ligation efficiency on the distribution of reads across the four steps of a three-round march. “Step 0” refers to unreacted molecules after three rounds of marching, while “Step 1”, “Step 2”, and “Step 3” refer to molecules that have been cleaved/ligated in one, two, or all three of three march rounds, respectively.
- (F) Effect of initial pool complexity on the length of contigs given a non-uniform distribution of reads across four steps. Contig lengths were determined through simulation as in (B), but using the probability of obtaining a read from each step as determined in panel (E) assuming a cleavage/ligation efficiency of 0.5.
- (G) Expected correspondence between round-associated barcode tags and the step no. of tagged reads. For instance, round no. = step no. = 1 if a molecule was

cleaved/ligated in the first round and only the first round and was therefore tagged with the first round barcode and was advanced by one step along the amplicon template.

Tables

Table 1. Overview of sequencing reads obtained for each sample

Sample	Library	Total Reads*	Genome-Aligned Reads (% of Total Reads)	Location-Collapsed Reads (% of Genome-Aligned Reads)
<i>P. falciparum</i>	Initial Library	2,316,937	525,509 (22.7%)	134,912 (25.7%)
	Round 1	4,194,002	968,063 (23.1%)	308,173 (31.8%)
	Round 2	2,747,609	485,034 (17.1%)	200,754 (41.4%)
	Round 3	4,881,843	1,088,583 (22.3%)	415,836 (38.2%)
HBV	Initial Library	294,625	328 (0.1%)	94 (28.7%)
	Round 3	643,611	1291 (0.2%)	416 (32.2%)

**Plasmodium falciparum* reads are 48 bp long, while HBV reads are 36 bp long.

Chapter 3: HMMSplicer: a tool for efficient and sensitive discovery of known and novel splice junctions in RNA-Seq data

This chapter is a manuscript, prepared for publication in a peer-reviewed journal.

Michelle T. Dimon, Katherine Sorber, Joseph L. DeRisi. HMMSplicer: a tool for efficient and sensitive discovery of known and novel splice junctions in RNA-Seq data

Author contributions:

Michelle Dimon developed the algorithm, wrote the code and performed the tests.

Katherine Sorber performed the experimental validation in *P. falciparum*. Michelle Dimon, Katherine Sorber, and Joseph DeRisi conceived the project and guided the implementation.

Joseph L. DeRisi, thesis advisor

Abstract

Background: High-throughput sequencing of an organism's transcriptome, or RNA-Seq, is a valuable and versatile new strategy for capturing snapshots of gene expression.

However, transcriptome sequencing creates a new class of alignment problem: mapping short reads that span exon-exon junctions back to the reference genome, especially in the case where a splice junction is previously unknown.

Methodology/Principal Findings: Here we introduce HMMSplicer, an accurate and efficient algorithm for discovering canonical and non-canonical splice junctions in short read datasets. HMMSplicer identifies more splice junctions than currently available algorithms when tested on publicly available *A. thaliana*, *P. falciparum*, and *H. sapiens* datasets without a reduction in specificity.

Conclusions/Significance: HMMSplicer was found to perform especially well in compact genomes and on genes with low expression levels, alternative splice isoforms, or non-canonical splice junctions. Because HMMSplicer does not rely on pre-built gene models, the products of inexact splicing are also detected. For *H. sapiens*, we find 3.6% of 3' splice sites and 1.4% of 5' splice sites are inexact, typically differing by 3 bases in either direction. In addition, HMMSplicer provides a score for every predicted junction allowing the user to set a threshold to tune false positive rates depending on the needs of the experiment. HMMSplicer is implemented in Python. Code and documentation are freely available at <http://derisilab.ucsf.edu/software/hmmsplicer>.

Introduction

RNA-Seq, which applies high-throughput sequencing technology to an organism's transcriptome, has revolutionized the study of RNA dynamics within a cell [1]. Millions of short read sequences allow both the presence and abundance of transcripts to be ascertained. RNA-Seq has been shown to have a better dynamic range for gene expression levels than microarrays [2] and enables scientists to view the transcriptome at single nucleotide resolution. Thus this technique combines the genome-wide scale of microarrays with the transcript variant detection power of Expressed Sequence Tags (ESTs).

RNA-Seq reads fall into two main classes: reads with full-length alignments to the genome and reads that span exon-exon junctions. Current sequencing runs produce tens of gigabases and it is likely that terabase sequences will be a reality in the near future. This massive output necessitates rapid techniques to analyze the data in a reasonable amount of time. For full-length alignments of sequence reads back to a reference genome, recent tools that rely on the Burrows-Wheeler Transform have yielded significant improvements in speed and accuracy. These include BWA [3], SOAPv2 [4] and Bowtie [5].

The more difficult RNA-Seq challenge is aligning reads that bridge exon-exon junctions since they by definition form gapped alignments to the genome with very short flanking sequence. These exon-exon junction reads reveal the exact location of splicing events, an intricate process wherein the intron in a pre-mRNA transcript is removed and the

flanking exons are joined together. This tightly regulated process is coordinated by the spliceosome, a complex of many small-nuclear ribonucleoproteins (snRNPs) (reviewed in [6]). The spliceosome facilitates nucleophilic attack of the phosphodiester bond at the 5' splice site (5'SS) by the branch point sequence. The 3'-hydroxyl at the 5'SS then reacts with the start of the next exon, the 3' splice site (3'SS), ligating the exons and releasing the intron lariat. The branch point sequence, 5'SS and 3'SS are defined by short motifs within the intron sequence. In metazoans, the consensus splice site motifs are GTRAGT for the first six bp of the intron (5'SS) and YAG as the last 3 bp of the intron (3'SS). However, these motifs are extremely degenerate, leaving just 'GT-AG' as fairly reliable splice sites, found in 98% of known human introns [7]. Although most splicing in eukaryotic cells is performed by the spliceosome, non-spliceosomal splicing occurs and can be essential. One well-characterized example is the splicing of yeast *HAC1* and the homologous *XBPI* in metazoans [8]. In yeast, the transcription factor HAC1p regulates the unfolded protein response. HAC1p is, in turn, regulated by unconventional splicing of *HAC1* mRNA [9]. This splicing is not accomplished by the spliceosome. Instead, the protein Ire1p cleaves the *HAC1* mRNA in two places and the resulting edges are ligated with tRNA ligase [10]. In metazoans, *XBPI* is cleaved in a homologous manner, with the non-canonical splice boundaries CA-AG instead of GT-AG.

During the past decade, there has been a growing appreciation of the importance of alternative splicing as a mechanism for organisms to increase proteomic diversity and regulatory complexity (reviewed in [11] and [12]). The model of static exon and intron definitions yielding a single mRNA transcript and single protein sequence from each

gene has proven overly simplistic. In reality, alternative splicing, the creation of multiple mRNA transcripts from a single pre-mRNA sequence by differential splicing, is extensive in multicellular organisms, increasing with organismal complexity. Recent RNA-Seq studies suggest that virtually all multi-exonic human transcripts have alternative isoforms [13], [14]. The extent of alternative splicing, as well as the balance between types of alternative splicing (*e.g.* alternate 5' SS versus exon-skipping splicing), differs by organism [15]. The regulation of splicing in different tissues and developmental stages, as well as the mechanisms for its regulation, is a subject of ongoing research [11, 16, 17]. Therefore, the ability to detect alternative splice isoforms with accuracy and sensitivity is key to comprehensive RNA-Seq analysis.

Aligning exon-spanning reads to the genome is difficult. Instead of a single full-length alignment, an algorithm must break a short read into two even shorter pieces and align each piece accurately. One early approach to short read splice junction detection was alignment using existing gene annotations, as done by ERANGE [18]. While this approach was necessary to align very short reads (36 nt or less) back to mammalian genomes, it does not address the question of novel junctions and cannot be used for organisms with incomplete or inaccurate genome annotations. Another early approach was to use BLAT [19], a tool developed for the alignment of longer EST sequence. This method can provide good results but requires extensive effort by the researcher to post-process and filter the search results, which could be achieved by the construction and training of a support vector machine specific to the organism and dataset [20]. In addition, BLAT searches on mammalian genomes can be slow.

The current leading algorithm for finding novel junctions in RNA-Seq data is TopHat [21]. TopHat uses full-length read alignments to build a set of exon ‘islands’, then searches for short reads that bridge these exon islands. The strength of this approach is that the resulting set of putative gene models can be used to estimate transcript abundance, as in the recently released Cufflinks software [22]. However, the algorithm must be able to define exon islands, which can be difficult when the coverage is low or uneven or when introns are small. While TopHat can find GT-AG, GC-AG, and AT-AC splice sites under ideal conditions, it does not extract any other splice sites. As a result, TopHat performs best on mammalian transcripts with relatively high abundance, but can stumble in more compact genomes and with non-canonical junctions.

Recently, several algorithms have been published that match reads more directly to the genome, including SplitSeek [23], SuperSplat [24], and SpliceMap [25]. SplitSeek divides the read into two non-overlapping anchors and initially detects junctions as places where the two anchors map to different places on a chromosome (i.e. the two exons with the intron between them), with no requirement for specific splice sites. These initial junctions are further supported by reads where only a single anchor maps to an exon - however, the requirement for at least one read split evenly across the exon-exon boundary reduces sensitivity in low coverage datasets and transcripts. Additionally, SplitSeek only supports ABI SOLiD reads currently. SuperSplat is another algorithm that reports non-canonical junctions (junctions with intron edges other than GT-AG, GC-AG, or AT-AC). However, this algorithm requires both pieces of a read to be exact matches to the

reference sequence so it is not robust against sequencing errors or SNPs. SpliceMap divides reads in half, aligns each read half to the genome, then locates the remaining part of the read downstream within the maximum intron size. However, this algorithm considers only canonical splice junctions and requires read lengths of 50 nt or greater. In addition, although SplitSeek, SuperSplat, and SpliceMap all provide methods to filter the resulting junctions by the number and types of supporting reads, none provide a score that predicts the accuracy of a junction.

Here we introduce HMMSplicer, an accurate and efficient algorithm for finding canonical and non-canonical splice junctions in short-read datasets. The design of HMMSplicer was conceived to circumvent the inherent bias introduced by relying upon previously defined biological information. HMMSplicer begins by dividing each read in half, then seeding the read-halves against the genome and using a Hidden Markov Model to determine the exon boundary. The second piece of the read is then matched downstream. Both canonical and non-canonical junctions are reported. Finally, a score is assigned to each junction, dependent only on the strength of the alignment and the number and quality of bases supporting the splice junction. The scoring algorithm is highly accurate at distinguishing between true and false positives, aiding in novel splice junction discovery for both canonical and non-canonical junctions. HMMSplicer was benchmarked against TopHat and SpliceMap. It outperformed TopHat across a range of genome sizes, but most dramatically in compact genomes and in transcripts with low sequence read coverage. Compared to SpliceMap, it performed similarly in a human dataset and outperformed SpliceMap on an *A. thaliana* dataset.

Results

Algorithm Overview

An overview of the HMMSplicer algorithm is shown in Figure 1. Before the HMMSplicer algorithm begins, full-length alignments to the genome are detected using Bowtie [5] and removed from the dataset. HMMSplicer begins by dividing the remaining reads in half and aligning each half to the genome. All alignments for both read halves are considered autonomously and are not resolved until the final scoring step. Once a read-half is aligned, a Hidden Markov Model (HMM) is used to detect the most probable splice position. The HMM is trained on a subset of read-half alignments to best reflect the quality and base composition of the dataset and genome. Next, the remaining portion of the read is aligned downstream of the exon-intron boundary, completing the junction definition. Finally, identical junctions are collapsed into a single junction and all junctions are scored, filtered by score, and divided by splice-site edges, with canonical (GT-AG and GC-AG) junctions in one result set and non-canonical edges in a second result set.

Algorithm parameters

Our first step was to test the assumptions underlying HMMSplicer's algorithm by evaluating performance relative to key parameters: the required read length, the robustness of the HMM, and the ability to match the second piece of a read. First we examined the ability of read-halves to seed within a genome by measuring the fraction of

read-halves aligned in the Bowtie read-half alignment step for various read-half sizes and genome sizes (Figure 2a). For the human genome, HMMSplicer performs optimally for reads 45 nt or longer (read-halves of 22 nt or longer), though shorter reads can be used. Simulation results, described below, confirm this assessment, showing a higher false positive rate when aligning 40 nt reads to the human genome. Next, we validated the robustness of the HMM training. An essential feature of HMMSplicer is that the HMM used to determine where the splice occurs within the read is trained from a subset of the input read set by an unsupervised algorithm. For the HMM training to be robust, it must train to similar values for an input read set, regardless of the initial values or the subset of reads used for training. This was validated using the human read set. Training sets ranging from 50 to 50,000 read-half alignments were used to train the HMM with two different sets of initial HMM values. For the first set of initial values, we used completely even values, i.e. a 50/50 probability of a match or mismatch for each quality score. For the second set, we used values close to those we expected as trained HMM values (Table 1). Training for each combination of training set size and initial value was repeated 10 times with different random subsets to measure the mean and standard deviation of the trained values. The results show that the HMM training converges on similar values regardless of training set size and initial values. The two most variable parameters are shown in Figure 2b, all other parameters showed less variability across the conditions (data not shown). Smaller training sets showed more variability so a default training set size of 10,000 was selected for HMMSplicer as sufficient to sample the space. Finally, mapping of sequences of various sizes within an 80 kbp maximum intron was analyzed to determine the optimal anchor size (Figure 2c). In the human genome,

for sequences fewer than 8 nt in length, the most common result was multiple matches, whereas at 8 nt and above, a unique, correct match was the most likely result. Based on these data, the default anchor size was set at 8 nt for the default maximum intron size of 80 kbp. For compact genomes with smaller maximum intron sizes, such as the *P. falciparum* and *A. thaliana* datasets below, a shorter anchor size of 6 nt can be matched uniquely (data not shown).

Benchmark Tests

HMMSplicer's performance was analyzed on simulated reads and three publicly available experimental datasets (Tables 2 and 3). The simulation dataset, generated from human chromosome 20, provides a measurement of the number of junctions detected and the false positive rate at different read lengths and coverage levels. However, simulation results do not model all the complexities found in experimental datasets, such as uneven coverage with a bias towards higher coverage of GC-rich regions, uneven distribution of sequence transversions, and inaccurate quality scores [26]. Three experimental datasets were selected from the NCBI Short Reads Archive (SRA), each representing a real world challenge. The first experimental dataset, ~80 million reads from *Arabidopsis thaliana*, allowed analysis of HMMSplicer's performance using a dataset with low quality reads. The next experimental dataset, ~14 million reads in *Plasmodium falciparum*, was used to assess performance in the context of uneven coverage and high AT content. The final experimental dataset, ~10 million paired-end reads from *Homo sapiens*, was used to test HMMSplicer's performance in a larger metazoan genome. This dataset also provided a

platform for analyzing transcripts with low abundance, alternative splicing and non-canonical splice sites.

HMMSplicer combines high sensitivity with a low false positive rate

HMMSplicer was first tested on simulated read sets to determine its performance in an environment where true and false positive rates could be definitively measured. For the simulation, reads from 503 non-overlapping gene models on human chromosome 20 were generated at varying read lengths and coverage levels. For an accurate quality model, we used the error model from a human dataset [27]. In this read set, the second paired end read was extended to 75 bases, allowing us to simulate longer reads. The program maq was used to generate reads of length 40, 45, 50, 55, 60, 65, 70, and 75 bp at 1x, 5x, 10x, 25x, and 50x coverage [28]. TopHat was run on the same simulated dataset for comparison.

HMMSplicer's false positive rate was low overall, rising with short reads and high coverage (Table 2). The highest false positive rate, 8.3% was seen for 40 bp reads at 50x coverage, re-iterating the conclusion from parameter testing (above) that HMMSplicer performs ideally in the human genome with reads at least 45 bp long. At a length of 45 bp, the false positive rate for 50x coverage was 4.2%, while for reads 50 bp or longer the false positive rate never exceeded 2.5%, with most error rates remaining under 1%.

HMMSplicer was effective at identifying junctions, even at low coverage levels (Figure 3a). With 50 bp reads at 1x coverage, HMMSplicer was able to identify more than 40%

of all the junctions in the set (1701 of 4043). At 5x coverage, more than 90% of the junctions were found (3646 of 4043). Higher coverage levels increase the number of junctions found, and at 50x coverage more than 98% of the junctions are found (3958 of 4043). While TopHat finds similar number of junctions at higher coverage levels, HMMSplicer finds three times as many junctions at 1x coverage with reads less than 70 bp long, and more than 50% more junctions with reads 70 or 75 bp long. Seventy-seven junctions were never detected by either program, even at 50x coverage and 75 bp reads. These junctions either had a homologous region within the genome or encompassed tiny initial or final exons that, because the simulated transcripts did not include UTR regions, had artificially low coverage.

One of HMMSplicer's strengths is that the algorithm provides scores for each junction, indicating the confidence of the prediction. To judge the accuracy of the scoring algorithm, Receiver Operator Characteristic (ROC) curves were generated comparing the true positive and false positive rate (Figure 3b). To measure true and false positive rate, simulation results for all scores were considered. Predicted junctions that aligned to the correct source of the simulated read were considered correct, while predicted junctions that aligned to another location were considered false. The ROC curves show that the HMMSplicer scoring algorithm was highly accurate, with the inflection point for 10x coverage and 50 bp reads including 98.7% of the true junctions and only 6.7% of the false junctions. At the default score threshold, 99.3% of true junctions and only 13.3% of the incorrect junctions were included.

HMMSplicer performs well on datasets with low quality sequence reads

High-throughput sequencing datasets can have high error rates, however there is still useful data to be gleaned from these datasets. The first dataset, ~79 million reads, each 50 bp long, in *Arabidopsis thaliana*, evaluated the performance of HMMSplicer with variable quality sequence reads [29]. *A. thaliana*, a model plant species, has a genome of 125 million base pairs with ~25,500 protein-coding genes [30]. The mean exon and intron sizes are 78 bp and 268 bp, respectively, with an average of 4.5 introns per gene [31].

We analyzed these low-quality reads, using a minimum intron length of 5 bp, a maximum intron length of 6 kbp, and an anchor size of 6 bp. The gene models in the most recent release of The Arabidopsis Information Resource (TAIR9, <http://www.arabidopsis.org>) contain introns from 3 bp to 11,603 bp long with 99.9% of the introns falling between 5 and 6,000 bp. At the default score threshold, HMMSplicer detected 14,982 junctions, with 95% (14,217) of the predicted junctions matching TAIR9 annotations (Figure 4a). The relatively low number of junctions found overall despite the size of the dataset is likely a result of low read quality. The low quality also decreases the HMMSplices scores, causing a sharper decrease in the number of junctions at higher score thresholds compared to other datasets (Figure 4a).

TopHat and SpliceMap were also run on the *A. thaliana* dataset. TopHat, run with a minimum intron size of 5 bp and a maximum intron size of 6 kbp, was able to locate only 6,346 junctions, less than half the number found by HMMSplicer, with 91.7% (5,820) of

these predictions matching TAIR9 annotations (Figure 4a). SpliceMap was run with the same 6 kbp maximum intron size (the minimum intron size is not configurable).

SpliceMap found 9,438 junctions, 92.8% of which match TAIR9 annotations. Although SpliceMap found more junctions than TopHat, HMMSplicer found 50% more junctions than SpliceMap with a higher percentage matching TAIR9 annotations than either competitor.

HMMSplicer performs well in datasets with uneven coverage

The *P. falciparum* genome is fairly compact and AT-rich, containing approximately 5,300 genes in 23 million base pairs [32]. In the latest genome annotation (PlasmoDB 6.3, <http://www.plasmodb.org>), the average exon size is 890 bp and the average intron size is 168 bp with an average of 1.54 introns per gene. Previous research on an earlier release of the genome annotation indicated that approximately 24% of the gene models predicted for *P. falciparum* are incorrect [33]. The malaria research community has focused on improving the genome annotation, and the most recent genome annotation release addresses many incorrect annotations. However, there are still numerous unconfirmed gene models with limited or no EST evidence.

The *P. falciparum* read set was published in the NCBI SRA following work on the Long March technique [34]. The dataset downloaded from NCBI SRA contains 14,139,995 reads, each 46 bp long. This dataset has uneven coverage with coverage varying significantly even within a single transcript. To detect splice junctions in this dataset, HMMSplicer was run with a minimum intron size of 10 bp, a maximum intron size of 1

kbp and an anchor size of 6 bp. This range includes 99.6% of the known introns in the current *P. falciparum* genome annotation. At the default score threshold, HMMSplicer identified 4,323 junctions in this dataset, 85.2% of which overlapped either known gene models or ESTs (Figure 4c). TopHat found 3,138 junctions in this dataset with 77.7% aligning to known gene models or ESTs. By re-running TopHat with more stringent alignment parameters, the percent of confirmed junctions was boosted to 94.8%, but this resulted in a 71% decrease in the number of found junctions (885). In contrast, the output of HMMSplicer can be filtered for more stringent confirmed junction percentages simply by raising the score threshold. SpliceMap could not be tested on this dataset because the reads are less than the minimum 50 nt length required by the algorithm.

HMMSplicer performs well in large metazoan genomes

The *Homo sapiens* genome is large (3.2 billion base pairs with ~25,000 genes), and contains both short exons (~59 bp on average) and large introns (~6,553 bp) [31], creating a significant challenge for identifying splice junctions. However, the human genome is well annotated with abundant EST evidence, allowing evaluation of HMMSplicer's performance on transcripts with low abundance, alternatively spliced junctions, and non-canonical junctions. Although the human genome is well studied, the complications of tissue-specific expression and widespread alternative splicing mean that many splicing events have not yet been detected. For our benchmark tests, we selected a human dataset containing 9,669,944 paired-end reads, each 45 bp long, from a single individual's resting CD4 cells [27]. The version of the genome used for analysis was the February 2009 human reference sequence (GRCh37) produced by the Genome Reference

Consortium. Two reference sets were used to identify known introns. The first set represents known genes and well-studied alternates (genes present in the manually curated RefSeq [35]), while the second set represents a more extensive set of junctions, including many alternative splicing events (RefSeq genes and an additional 8,556,822 mRNAs and ESTs from GenBank [36]).

HMMSplicer was run with a minimum intron length of 5 bp and a maximum intron length of 80,000 bp, covering 99.1% of known introns in the human genome. Because HMMSplicer must match the second piece of the read downstream of the initial exon edge identified, the HMMSplicer algorithm is sensitive to maximum intron size. For efficient and accurate matching in 80 kbp introns, we used an anchor size of 8 nt, instead of the 6 nt anchor used in *A. thaliana*. At the default score threshold, HMMSplicer found 101,664 junctions, 87% of which (88,162) matched known genes or ESTs/mRNAs (Figure 4b). TopHat was run with the default intron size range of 70 to 500,000 bp, which covers 99.9% of known introns in the human genome. TopHat found 72,771 junctions, of which 93.0% (67,664 junctions) matched known genes or ESTs/mRNAs. Increasing the score threshold to 600 for junctions supported by multiple reads (800 for junctions supported by a single read) yields a similar confirmed junction rate of 91.8% and leads HMMSplicer to find 89,130 junctions, 22% more than TopHat.

Because this publicly available 45 nt dataset is too short for analysis by SpliceMap (which requires 50 nt reads), we were unable to directly compare HMMSplicer to SpliceMap on this dataset. Instead, we ran HMMSplicer on the human dataset analyzed in

the SpliceMap publication [25], a set of 23,412,226 paired end reads of 50 nt each from a human brain sample (GEO Accession number GSE19166). SpliceMap is published as finding 175,401 splice junctions in this dataset with 82.96% EST validation. Filtering lowers the number of junctions found while raising the validation rate, so that at a validation rate of 94.5%, SpliceMap detected 121,718 junctions. HMMSplicer was run on the same dataset with default parameters, yielding similar results of 177,890 junctions with 84.2% EST validation at the default score threshold. Raising the score threshold to 800 (1000 for single junctions) we found 131,007 junctions with 94.5% EST validation. Our comparisons suggest that HMMSplicer finds slightly more (7%) junctions than SpliceMap at an equivalent EST validation level (94.5%) in this human dataset.

HMMSplicer identifies many junctions in low abundance transcripts

A recent RNA-Seq study across 24 tissues in humans showed that ~75% of mRNA in a cell is from ubiquitously expressed genes [37]. Furthermore, although transcripts from ~11,000 to ~15,000 genes were detected (depending on the tissue), the 1000 genes with the highest expression levels contributed more than half the mRNA in each tissue. The importance of RNA-Seq in the detection of novel splice junctions is not in these ubiquitous highly expressed genes, which generally have EST coverage, but in the tissue-specific genes with lower transcript abundance.

Therefore, we measured HMMSplicer's capacity for detection of junctions in low-abundance transcripts in the human resting CD4 cell dataset. In RNA-Seq experiments with non-normalized cDNA samples, the coverage level of a gene varies depending on

relative transcript abundance. A convenient measure of read coverage relative to the transcript abundance is Read Per Kilobase per Million reads mapped (RPKM) [18] which counts the number of reads that map to a gene, normalized by the length of the gene in kilobases, per million reads mapped to the genome. Figure 5 shows the number of predicted junctions matching RefSeq-defined introns at different RPKM levels. HMMSplicer identified more junctions than TopHat at all RPKM levels, but the difference is greatest at low values of RPKM. This is relevant to many RNA-seq experiments. In this dataset, 75% of genes had an RPKM of 10 or less.

Sequence-level analysis reveals alternate 5' and 3' splice sites

HMMSplicer's approach allows discrimination of closely spaced alternative splice sites, providing a method to study fundamental questions about the biology of splicing which have not yet been addressed with RNA-Seq experiments. Alternative splicing analysis in RNA-Seq data frequently focuses on quantifying isoform expression level, such as in a recent study measuring isoform abundance based on relative coverage levels of exons [38]. This is an important application, but the sequence-level detail of RNA-Seq data provides the power to examine alternative splicing at a finer level of detail. Analysis within the human resting CD4 cell dataset showed instances where splice sites varied slightly from known intron boundaries, suggesting an inexact splicing event. To investigate these results further, all junctions overlapping RefSeq introns with fewer than 15 bp differences in splice sites were examined and the number of bases added or removed from the exon boundary was counted. Overall, there were 997 instances (1.4% of junctions which match RefSeq) where an intron possessed an alternate 5'SS and 2,577

(3.6% of junctions which match RefSeq) instances of an alternate 3'SS. Alternative splicing which maintained the reading frame (*i.e.* added or removed a multiple of 3 bases from the transcript) was clearly preferred for the 3' splice site (Figure 6). This result is not surprising given that the 3'SS motif, YAG, is shorter and shows more variation than the 5'SS motif, GTRAGT [6]. To investigate this result further, Weblogos [39] were constructed from the sequences at the alternate 3'SS that were off by 3 bases. Analysis of these Weblogos found at the alternate 3'SS shows repetition of the splice motif (*i.e.* YAGYAG).

HMMSplicer identifies non-canonical junctions

We next analyzed the ability of HMMSplicer to identify junctions with splice sites other than GT-AG using the human resting CD4 dataset for analysis. The most common splice sites, GT-AG, GC-AG, and AT-AC, are found in 98.3%, 1.5% and 0.2% of human introns, respectively [7]. By default, HMMSplicer attempts to adjust intron edges to GT-AG, GC-AG or AT-AC but includes only GT-AG and GC-AG introns in the set of canonical junction predictions. The user can alter the splice sites for adjustment and filtering or can eliminate these steps entirely. We examined the splice sites in junctions found by HMMSplicer. Counting only junctions that matched known mRNA/ESTs, HMMSplicer detected 87,245 GT-AG junctions, 791 GC-AG junctions, and 97 AT-AC junctions. This is 99% GT-AG, 0.9% GC-AG, and 0.1% AT-AC, which corresponds well with the published rates. The ratio of junctions that match known junctions is much lower for non-GT-AG junctions (20.3% for GC-AG and 6.5% for AT-AC). To resolve whether HMMSplicer non-canonical junctions are false positives or novel instances,

further experimental validation will be required. Regardless, HMMSplicer provides all junctions and allows the user to filter based on the experiment's objectives.

Although rare, there are also splice junctions that do not have GT-AG, GC-AG or AT-AC splice sites. For example, the *HAC1* mRNA and its metazoan homologue *XBPI* are spliced by Ire1p with the non-canonical splice sites CA-AG, initiating the unfolded protein response [8]. HMMSplicer's non-canonical junction results on the human dataset contained three reads spanning the *XBPI* non-canonical intron with scores ranging from 927 to 971 (Figure 7). The sequence at the beginning of the intron is identical to the initial exon sequence, so the HMM was unable to resolve the exact junction edges correctly. This resulted in two possible predictions, one 2 bp upstream from the actual site and one 4 bp downstream from the actual site. Collapsing identical junctions resulted therefore in two junctions, one with a score of 1024 and one with a score of 1030, which put them in the top 0.5% of the collapsed non-canonical junctions.

HMMSplicer finds true novel junctions in genomes with incomplete annotation

To determine if unconfirmed junctions predicted by HMMSplicer represent true novel junctions or false positive predictions, we experimentally validated four previously unknown junctions predicted from the organism with the least thorough annotation, *P. falciparum* (Figure 8). All four junctions were relatively high scoring but no EST or experimental data exists for comparison, and each case conflicts with the current PlasmoDB gene model. The first junction (score=1300), in PFC0285c (predicted to

encode the beta subunit of the class II chaperonin tailless complex polypeptide 1 ring complex), suggests an additional exon at the 5' end of the gene model, possibly belonging to the 5' untranslated region (UTR). The second junction (score=1198) belongs to PF07_0101, a conserved *Plasmodium* protein of unknown function. This previously unknown junction excises 291bp out of the middle of the first annotated exon, which would result in a protein 97 amino acids (aa) shorter. The third and fourth junctions, with scores of 1261 and 1175, respectively, are in PFD0185c, another gene of unknown function conserved across *Plasmodium* species. One junction lies within the predicted gene, splicing out 85bp and leading to a frameshift near the 3' end, while the other appears to splice together two exons in the 3'UTR. RT-PCR followed by sequence analysis verified all four splice junctions predicted by HMMSplicer (Figure 8), confirming HMMSplicer's ability to predict true novel junctions from RNA-Seq data.

Discussion

HMMSplicer is an efficient and accurate algorithm for finding canonical and non-canonical splice junctions in short read data. Our benchmark tests on simulated data and three publicly available datasets show that HMMSplicer is able to detect junctions in compact and mammalian genomes with high specificity and sensitivity. The real world challenges in these datasets include low quality reads and uneven coverage. Built on Bowtie, HMMSplicer is fast, comparable in CPU time to TopHat. Analysis also demonstrates HMMSplicer's ability to find splice junctions on transcripts with low abundance, alternative splicing, and non-canonical junctions.

Comparisons with TopHat show that HMMSplicer is able to find more junctions with a similar level of specificity in each of these datasets. Comparisons with SpliceMap show that HMMSplicer has similar performance, yielding slightly more (7%) EST matching junctions in paired-end human datasets. However, in the low sequence quality *A.*

thaliana dataset, HMMSplicer significantly outperforms SpliceMap. HMMSplicer was not compared to SplitSeek [23] as this algorithm only processes colorspace reads.

Though the algorithm is similar, we anticipate that HMMSplicer would be more sensitive than SplitSeek, since this algorithm requires at least one read to be split evenly across the splice junction. HMMSplicer, TopHat, and SpliceMap are all free from this constraint.

Finally, the SuperSplat [24] algorithm is the only other currently available algorithm that detects non-canonical junctions to our knowledge. Unfortunately, the current version of SuperSplat does not align reads with any mismatches, and also has large memory requirements (5 – 32 GB to index the *A. thaliana* genome).

A major strength of HMMSplicer is that it is the only software package that provides a score for each junction, reflecting the strength of the junction prediction, which allows tuning of HMMSplicer's results to an experiment. While many splice junction algorithms filter on specific attributes to improve validation rates, for example, SpliceMap has filtering to remove junctions with only a single supporting read, HMMSplicer's score provides a more flexible way to tune true and false positive rates for the experiment. The score is based solely on the number of bases on each side of the junction, the quality of those bases, and the junction's similarity to potential full-length matches. Re-annotation experiments would necessitate a higher threshold to avoid false

positives, but experiments looking for novel junctions could use a lower threshold to include as many true positives as possible. The threshold can also be tuned for non-ideal datasets, such as the low quality *A. thaliana* dataset. The score is highly predictive despite the fact that it does not include biological factors such as splice site or intron length in its calculation, making it ideal for detection of novel splice junctions.

Alternative splicing is an area of intense research where HMMSplicer's approach provides a significant advantage over algorithms that rely on exon islands, such as TopHat. In the case of alternate 5' or 3' splice sites, the major isoform may mask the signal from a minor isoform, especially in genes without high sequence coverage.

HMMSplicer accurately identifies small variations in 5' and 3' splice sites. These small variations in splice sites, most frequently 3 nucleotides added or removed from the transcript at the 3' splice site (1 amino acid added/removed from the translated protein), demonstrate how the repetition of the splice motif can cause inexact splicing.

HMMSplicer's unbiased approach to alignment, combined with the sequence level power of RNA-Seq, has enormous potential for biological inquiry into alternative splicing.

The depth of RNA-Seq and the unbiased approach of HMMSplicer also allow investigation into non-canonical splicing. HMMSplicer allows the researcher to define canonical splice sites, and returns both canonical and non-canonical results. Scores in HMMSplicer's predicted junctions aid the discovery process, as evidenced by the *XBPI* example in the human dataset. In HMMSplicer's results, it was ranked in the top 0.5% of the non-canonical splice results.

In conclusion, HMMSplicer is a valuable addition to the algorithms available for finding splice junctions in RNA-Seq data. The software, documentation and details about the datasets and analysis can be found at <http://derisilab.ucsf.edu/software/hmmsplicer>.

Materials and Methods

The HMMSplicer algorithm has four main steps: seeding reads within the reference genome, finding the splice position, matching the second piece of the read, and scoring/filtering splice junctions. Figure 1 shows an overview diagram of the HMMSplicer pipeline.

As a pre-analysis step, dataset reads are aligned to the reference genome using Bowtie [5]. Reads with full-length alignments to the genome contain no junctions and are therefore removed from consideration. These genome-matching reads may be used to build a coverage track that can be viewed in the UCSC Genome Browser [40] or other applications.

Step 1. Read-half alignment

To determine the read's seed location within the genome, we assume that each read spans at most a single exon-exon junction. Reads are divided in half, rounding down for reads of odd length, and both read-halves are aligned to the genome using Bowtie (current version 0.12.2), although other full-length alignment algorithms may also be used. This approach will locate an alignment for both read halves if the read is somewhat evenly

split across a junction, and these alignments are carried through the algorithm independently until they are resolved during scoring. However, if the read matches unevenly across the junction (*e.g.* if one side of a 45 nt read is 35 nt long and the other side is 10 nt long, referred to as a “35/10 split”), only the longer side will be seeded in this step. A read-half may not align if the larger half falls on another exon-exon junction or if sequencing errors prevent an alignment. Alternatively, a read half may have multiple alignments. As long as the duplicates are below a repeat threshold (50 alignments by default), all seeds are continued through until the filtering part of the algorithm; duplicate junction locations for a read are resolved at that point. For clarity in the text below, the half of the read that seeded will be referred to as the ‘first half’ and will be described as if the initial half of the read matched to the 5’ edge of the intron, with all sequences in the sense direction. In reality, either half of the read could match to either edge of the intron.

Step 2. Determine Splice Site Position

The alignment of a read-half determines an outside edge of the spliced read alignment, but does not determine where the exon-intron boundary occurs. To return to our previous example read with a 35/10 split, the first half of the read, corresponding to 22 bases, will be aligned but it will be unclear that the first side extends to 35 bases. A simplistic approach to this problem would be to extend the seed until a mismatch occurs but this approach ignores both the additional information available in quality scores and the high error rate inherent in many high-throughput sequencing technologies. Continuing from the 35/10 split example, imagine, after the first 22 bases of read-half, there is one

mismatch to the genome at a low quality base and then 12 bases in a row which match the genome. The simplistic approach would be to assume the read stopped aligned after the first mismatch, suggesting the split is 22/23 instead of 35/10, resulting in an incorrect junction alignment. To avoid this type of error, HMMSplicer utilizes a two-state Hidden Markov Model (HMM) to determine the optimal splice position within each read. State 1 describes a read aligning to the genome. In this state, we expect that most bases in the read match their partner in the genome, and that the probability of matching will vary based on the read base quality (high quality bases are less likely to be sequencing errors and thus more likely to match). State 2 is cessation of alignment to the genome. In this state, matches between the read and the genome are essentially random and do not depend on quality. For example, a genome with a GC content of 50% would yield an expected probability of 25% for each base to match the target genome location, regardless of sequence quality score. The most probable transition point from State 1 to State 2 defines the optimal splice position. In the 35/10 split example, the HMM would evaluate the probability of a 22/23 split, with 10 matches in a row in State 2 (where the probability of a match is only 25%) compared to the probability of a 35/10 split where a low quality base causes a single mismatch while remaining in State 1. Assuming the probability of a mismatch in State 1 in a low quality base was about 30% (a typical value), the 35/10 split would be more probable than the 22/23 split. (All other possible splits would also be considered, but these would be low probability compared to the 22/23 and 35/10 split options.)

Within each state of the HMM, the quality is binned into five levels, representing low, medium-low, medium, medium-high, and high quality scores. Using five bins provides the best balance between having sufficient bins to distinguish quality levels, while maintaining enough bases within each quality bin that the HMM can be adequately trained using a random subset of reads. Using a separate bin for each quality score created situations where one or more quality score were under-trained because quality scores are not evenly distributed from zero to forty. Increasing the training subset size can ameliorate this problem, however results with more quality bins were not significantly better than results with five quality bins (data not shown).

The HMM is trained on a randomly selected subset of the input read set. The training is accomplished using the Baum-Welch algorithm [41], an expectation maximization technique that finds the most likely parameters for an HMM given a training set of emissions. For HMMSplicer, emissions are strings of match/mismatch values derived from the alignment of the whole read to the genome at the position of each seed match. By using an unsupervised training method, the HMM values can be trained without additional input from the user, such as known genome annotations. This allows for a more sophisticated approach than the simplistic model described above while maintaining model unbiased by additional information such as known genome annotation. This training allows the values to be optimal for any particular genome and sequencing run. For example, genome specific training can adjust for biases in genomic nucleotide composition. One of the datasets used in our testing is *P. falciparum*, which has a genome that is 80% AT. This reduced complexity makes the probability of a match in

random sequence higher than the 25% that it would be in a genome with balanced nucleotide distributions. In addition, training provides a way to validate the model. The premise behind the model is that in State 1 the probability of a match should increase with the quality of a base, but in State 2 the probability of a match should be independent of the quality score. If this model is accurate then regardless of initial values, the trained HMM should reflect this expectation. The outcome of the training, detailed in the Results section above, confirm the robustness of the model to different initial values. The HMM values for each parameter, before training and after training with each dataset studied, are given in Table 1. For each organism, the model trains as expected. Parameters in State 1 show a higher rate of matches than mismatches, varying by quality score, while parameters in State 2 remain at approximately 25% probability of a match regardless of quality. The only exception is for *P. falciparum*, where the probability of a mismatch in State 2 varies from 37% to 28% depending on quality because of the 80% AT bias in the genome.

After the HMM is trained, it is run for every read-half alignment, yielding the coordinates of the first piece of the read alignment, including the first exon-intron boundary of the splice junction. In the event of multiple equally probable splice positions, the splice position with the shortest second piece is selected. A falsely short second piece may still match within the maximum intron distance and has the potential to be adjusted to the correct splice site in the canonical splice-site adjustment (see below for details). On the other hand, a second piece with false bases added to the beginning will likely not match within the maximum intron distance causing the read to be discarded. If the remaining

part of the read is too short (eight nucleotides or fewer by default), the alignment is set aside. Uncertainty in the precise location of the splice junction and short alignment can be further resolved in a subsequent evaluation process described below.

Step 3. Determine Spliced Exon position

Once the splice position has been determined, the first exon-intron boundary has been identified. To determine the second exon-intron boundary, the remaining part of the read, (the 'second piece'), must be aligned. To reduce search space to a manageable and biologically relevant size, a default of 80 kbp downstream of the initial alignment is considered, although the user may adjust this to the most appropriate value for the organism and experiment. HMMSplicer first determines potential location positions by using the initial eight nucleotides of the second piece as an anchor (this anchor size may also be tuned to the organism and experiment), searching for all locations within the maximum intron size where this anchor matches exactly. To accommodate possible sequencing errors in these initial eight nucleotides, exact matches for the next eight nucleotides (*i.e.* positions 9-16 of the second piece) are found and are added to the set of anchors. For each position where an anchor has an exact match, the entire second piece of the read is compared to the genome and the number of mismatches is counted. The alignment with the fewest mismatches is selected as the best match. In the event of multiple best matches, the read is set aside to be resolved later.

At this point, a preliminary splice junction has been defined. However, the exact splice positions may be offset from the actual intron-exon boundaries by a few nucleotides,

especially in cases where the sequence at the beginning of the intron matches that at the beginning of the second exon. In these cases, sequence alone cannot define the correct edges. To aid in correct splice edge definition, HMMSplicer uses an assumption about the biology of splice sites. The most common splice sites, GT-AG, GC-AG, and AT-AC, are found in 98.3%, 1.5% and 0.2% of human introns, respectively [7]. By default, HMMSplicer uses these three splice sites (in order of their frequency of usage) to adjust intron-exon boundaries, though the sequences can be changed or the feature can be turned off entirely. Given the frequency of these three splice sites compared to other splice sites, the use of splice sites for intron-exon boundary adjustment introduces a conservative assumption and can help resolve small ambiguities in the position of the splice site prediction. To perform the adjustment, both splice edges are moved an equivalent number of nucleotides to reach a canonical splice site, where possible. Junctions already at canonical edges and junctions that cannot be adjusted to canonical edges remain unchanged.

HMMSplicer provides a score for each predicted junction that does not rely on any biological information or assumptions about splicing machinery beyond the user-configurable adjustment to canonical splice sites, leaving the user free to apply the appropriate data processing filters for the experiment. The goal of the scoring approach is to use available information maximally while minimizing assumptions. For example, a score that incorporated the intron size distribution of the organism could have been more accurate, but would have introduced a strong bias toward typical intron sizes. Similarly, a scoring algorithm that penalizes non-canonical junction edges would have introduced a

bias towards canonical splice sites. Instead, HMMSplicer's score uses information only about the genome sequence, read sequence, read quality, and splice position to derive a score. The researcher can introduce further filtering to the result set, based on the needs of the experiment, but the score is free from these biases.

To accomplish this goal, we chose an information-based approach to the score algorithm, akin to a BLAST bit score rather than the probability-based E-value [39]. The initial step of the scoring algorithm is to measure the amount of information in the alignment of one side of the junction read. Assuming each possible nucleotide is equally likely and the reported read nucleotide was certain (no sequencing errors) there would be four equally possible nucleotides at each position of the read, resulting in 2 bits per position ($\log_2(4)$). However, the reported nucleotide is not certain, and this uncertainty is encoded by the quality of the nucleotide. To scale for this, we multiply the 2 bits by the probability that the nucleotide call is correct, given the quality score. The sum of the information in each matching position of the read piece alignment is then used as the score for that read piece.

Given:

g_i = genome nucleotide at position i
 r_i = reported nucleotide at position i
 r'_i = reference nucleotide at position i
 q_i = quality score at position i

Score for one side is calculated as:

$$h = \sum_{i=0}^j \begin{matrix} g_i = r_i : P(r_i = r'_i | q_i) * 2 \text{ bits} \\ g_i \neq r_i : 0 \end{matrix} \quad (1)$$

Both sides of the junction are scored using equation (1). To combine the scores for the individual read pieces, they are multiplied, giving a strong bias to evenly split reads. The score increase is greater for evenly split reads than for reads with uneven piece sizes. For example, comparing 50 nt reads and 70 nt reads, a 10/40 split compared to a 10/60 split will, under ideal conditions, raise the score from 400 to 600. By contrast, a 25/25 split compared to a 35/35 split will raise the score from 625 to 1225, a much more dramatic increase. This increase reflects the fact that a 10/40 to 10/60 split does not increase the information available as much as a 25/25 split to a 35/35 split.

Next, the score is corrected for the similarity to a full-length alignment. For each junction, if we hypothesize that the junction may actually be a full-length alignment, there are two possible positions for this alignment, either the left side is correct and the right side should be moved left adjacent to it, or the reverse. Both these possible full-length alignments are scored and the better alignment is kept. Half of this score is subtracted from the initial junction information as follows:

$$s = h_a h_b - F * \max \begin{cases} h_{a'} h_b \\ h_a h_{b'} \end{cases} \quad (2)$$

Where h_a is the score for the left side, h_b is the score for the right side, $h_{a'}$ is the score for the left side when moved adjacent to the right side and $h_{b'}$ is the score for the right side when moved adjacent to the left side. F is defined as 0.5, an empirically derived value that gives the best score results when tested on the human dataset (data not shown).

As a final step, the scores are normalized to the range 0-1200, with most scores less than 1000 in practice. This is simply for easy visualization in the UCSC Genome Browser. The BED file output from HMMSplicer can be uploaded directly to the UCSC Genome Browser, which uses grey-scale to represent scores from 0-1000. To perform this scaling, the multiplier is 1200 divided by the theoretical maximum score for a read of the given length. When calculating the theoretical maximum, equation (1) reduces to the length of the read piece times 2 bits. Thus, if the read length is even, the multiplier is:

$$m_{even} = \frac{1200}{\left(\frac{l}{2} * 2 \text{ bits}\right)^2} \quad (3a)$$

If the read length is odd, the multiplier is:

$$m_{odd} = \frac{1200}{\left(\frac{l-1}{2} * 2 \text{ bits}\right)\left(\frac{l+1}{2} * 2 \text{ bits}\right)} \quad (3b)$$

All together, the full equation for the score value is:

$$final \ score = s * m \quad (4)$$

Once splice junctions have been detected and scored, HMMSplicer resolves instances where both halves of a read were aligned independently, as well as instances where one or both read halves created multiple alignments. For reads where independent read half alignments converged on the same junction position, a single copy of the junction is saved. For reads where the read halves had multiple seed positions, if one position has a score much higher than the other(s), that position is retained. If a read matches in

multiple positions and all positions have close scores (by default, scores with differences less than 20, but this is user configurable), reads are saved in a separate set of output results reserved for duplicates.

Step 4: Rescue

Reads that cannot be matched uniquely can be used to lend support to a junction previously identified in the dataset. HMMSplicer attempts to rescue matches where the location of the first piece of the read is uniquely identified, but the location of the second piece is not. There are two sources of such reads: 1) reads with a second piece fewer than eight bases long and 2) reads where the second piece matched equally well to multiple locations within the maximum intron size. In both cases, HMMSplicer can apply the information from mapping the initial part of the read to rescue the read using other junctions found in the dataset. If another read ends at the same point as this read (i.e. has the same junction edge on the known side), the algorithm examines the other side of the junction to determine if the initial bases of the exon sequence match the second piece of this read. If so, this junction is assumed to be the source of the read.

Step 5: Filter and Collapse

Finally, initial junction-spanning reads are filtered and collapsed to yield a final set of predicted junctions. Splice junctions are divided into populations that do and do not match the most frequent splice sites ('GT-AG' and 'GC-AG' by default). Regardless of whether the user chooses to impose these splice site position sequences into the search, nonconforming junctions are saved and ranked separately. All reads creating the same

intron are collapsed into a single junction with the score for these reads increased in relation to number of additionally covered bases. Distinct reads covering the same junction add significantly to a its potential to be real, but two identical reads may be from the same source, such as PCR amplification artifacts. To follow the previous example, a 35/10 split (35 bp on the first exon, 10 bp on the second exon) combined with another 35/10 split would not increase the score, but the 35/10 split plus a 10/35 split would yield a substantial boost to the score because the covered bases would now be now 35/35. To be exact, imagine the 10/35 junction read has a score of 800 and the 35/10 junction read has a score of 600. The higher score read is considered first, then the second read is collapsed onto it. In this case, the new junction adds 25 bases out of a total of, now, 70 bases covered, so a value of $(25 / 70) * 600$ is added to the original score of 800, yielding a collapsed score of 1214.2.

Collapsed junction predictions are then filtered by score. Multiple error-free reads spanning the same splice junction align to the correct splice site, facilitating determination of splice boundaries. In contrast, because sequencing errors are distributed throughout the read with three possible wrong base substitutions, reads with errors that create false positive junctions tend to be scattered as single, incorrect alignments. Previous studies concur that true junctions are more likely than false junctions to be covered by more than one read [42]. Therefore, junctions covered by a single read are evaluated more stringently than junctions covered by multiple reads, with a higher score threshold set for junctions covered by a single read. The default score thresholds for HMMSplicer are 600 for junctions covered by a single read and 400 for

junctions covered by multiple reads. These score thresholds were optimal for the benchmark datasets, but ultimately the score threshold will depend on the number of reads used in the experiment (datasets with more reads may require higher score thresholds) and the purpose of the experiment (re-annotation studies will require higher score thresholds than studies looking for novel junctions).

Benchmark Methods

For the benchmark tests, all analysis was performed on an 8-core Mac Pro with 16 GB of RAM. HMMSplicer was run with default parameters unless otherwise noted. TopHat version 1.0.12 was used. TopHat was run with the best parameters for the dataset/organism, though the only parameter found to have a large effect on results was segment length. For reads shorter than 50 nt, segment lengths of half the read length were used for TopHat, as it was found to dramatically increase the number of splice junctions found (i.e. 30,381 junctions identified for the default segment length of 25 versus 68,946 junctions identified with a reduced segment length of 22 in the human dataset). For the simulation dataset, TopHat was run with the default parameters, except with a segment length of 20 and 22 for reads 40 and 44 nt long. The *A. thaliana* dataset was run with default parameters except for a minimum intron size of 5 and a maximum intron size of 6000. The *H. sapiens* dataset was run with a segment length of 22 using the butterfly search and microexon search parameters. The *H. sapiens* dataset is paired end and, based on information in the publication [27], an inner mate distance of 210 was used. SpliceMap was run on the *A. thaliana* dataset by the SpliceMap first author using a 6 kbp maximum intron size (personal communication).

For most of the analysis, canonical splice junction results from HMMSplicer were used (*i.e.* GT-AG and GC-AG splice sites), as they are most comparable to results from other algorithms. Table 3 contains general characteristics of the datasets downloaded from NCBI SRA, including accession numbers.

Experimental Validation in *Plasmodium falciparum*

Cell culture, RNA preparation, and poly-A selection

Plasmodium falciparum 3D7 Oxford parasites were sorbitol synchronized in early ring stage, then synchronized again 24 and 32 hours later for a total of 3 synchronizations during 2 consecutive cell cycles. Culture conditions were as in Bozdech et al, 2003. Post-synchronization, maximum invasion (number of schizonts = number of rings) was observed by smear and 50mL of 2% hematocrit, 10% parasitemia culture was harvested 44 hours post-invasion (late schizogeny). Harvested cells were centrifuged at 1,500 g for 5 min, washed in phosphate-buffered saline (PBS), and pelleted at 1,500 g for 5 min. The cell pellet was rapidly frozen in liquid nitrogen and stored at -80°C. Total RNA was harvested from the frozen pellet using 10mL Trizol (Invitrogen Corp., Carlsbad, CA). 238ug of total RNA was poly-A selected using the Micro Fasttrack 2.0 kit (Invitrogen Corp., Carlsbad, CA).

DNase treatment, reverse transcription, PCR, and sequencing

3.6ug of poly-A selected RNA was treated twice with 2uL of TURBO DNase according to the manufacturer's instructions for the TURBO DNase-free kit (Applied

Biosystems/Ambion, Austin, TX). Treated RNA tested negative for residual genomic DNA by PCR amplification in the following mix: 1x Herculase II Fusion buffer, 0.25mM dATP, 0.25mM dTTP, 0.0625mM dCTP, 0.0625mM dGTP, 0.25uM PF11_0062-F primer (5'-ACTGGTCCAGATGGAAAGA AAAA-3'), 0.25uM PF11_0062-R (5'-GGAGGTAAATTTTGTTACAGCTTTGGTTCC-3'), and 0.4uL of Herculase II Fusion polymerase (Stratagene, La Jolla, CA). PCR conditions were 95°C for 2 min, then 40 cycles of 95°C for 30 sec, 52°C for 45 sec, 65°C for 3 min, and finally 65°C for 7 min. 2ug of DNased RNA was melted at 65°C for 5 minutes in the presence of 817.5ng random hexamer, and then cooled at room temperature for 5 minutes. To reverse transcribe cDNA, 0.25mM dATP, 0.25mM dTTP, 0.0625mM dCTP, 0.0625mM dGTP, 1x First Strand buffer, 10M DDT, and 1090U Superscript III reverse transcriptase (Invitrogen, Carlsbad, CA) were added and the reaction was incubated at 42°C for 1.5 hours. 1uL of this reverse transcription mix was used for each junction confirming PCR using the previous described mix and cycling conditions, with the following changes: 0.5uM of the appropriate forward and reverse primers were used and 30 cycles of PCR were performed. PCR reactions were cleaned up with Zymo-5 DNA columns (Zymo Research Corp., Orange, CA). 100ng of each PCR product was a-extended by incubation at 37°C for 30 minutes in the presence of 16.7mM dATP, 1x NEB buffer 2, and 5U Klenow exo⁻ (New England Biolabs, Ipswich, MA). Extended products were then TOPO TA cloned and transformed into chemically competent TOP10 cells (Invitrogen, Carlsbad, CA). Transformations were plated on LB + ampicillin plates spread with 100uL 40mg/mL Xgal. After 16 hours of growth, colony PCR was performed on white colonies with the following PCR mix: 1x Taq buffer, 2mM MgCl₂, 0.5uM M13F, 0.5uM

M13R, 0.25mM dATP, 0.25mM dTTP, 0.0625mM dCTP, 0.0625mM dGTP, and 0.5U Taq polymerase (Invitrogen, Carlsbad, CA). PCR conditions were 95°C for 2 min, then 30 cycles of 95°C for 30 sec, 52°C for 45 sec, 65°C for 3 min, and finally 65°C for 7 min. Following precipitation with 3 volumes of isopropanol, 1/4 of each PCR product was primer extended in Sanger sequencing reactions in the presence of 1uM M13F, 1x sequencing buffer, and 0.5uL BigDye Terminator (Applied Biosystems Inc., Foster City, CA). Cycling conditions were 94°C for 2 min, then 60 cycles of 94°C for 30 sec, 50°C for 1 min, 60°C for 1 min, and finally 60°C for 7 min. Sequencing reactions were precipitated with 1/4 volume 125mM EDTA and 1 volume 100% ethanol, and then resuspended in HiDi formamide and run on a 3130xl Genetic Analyzer (Applied Biosystems Inc., Foster City, CA).

Primer Sequences

Name	Sequence
PF07_0101 F	TGGGTTATCTGATCATCAAGGA
PF07_0101 R	TTTTATGAGTGTCTCCCTTTT
PFD0185c F1	CGCACTACCATATTTATGCCTCT
PFD0185c R1	AGTAGAAGGAGGGAGGAGCA
PFD0185c F2	TTCGCGTGATGAAGAAGATG
PFD0185c R2	CAAGCCCACATATAAATCAAGGA
PFC0285c F	TATCTTCTTGGGCCCTTCT
PFC0285c R	TGTGAATGCGTGAAGGATTT

Author Contributions

MTD, KS and JLD conceived and designed the algorithm. MTD developed the algorithm, wrote the software, and performed the analysis. KS performed the experimental confirmation. MTD, KS, and JLD wrote the paper and produced the figures. All authors read and approved the final manuscript.

Acknowledgements

We thank Polly Fordyce, Victoria Newman, J. Graham Ruby and Peter Skewes-Cox for critical reading of the manuscript and Michael Cary for insightful discussions on the algorithm. We also thank Kin Fai Au and John C. Mu for running the SpliceMap program on the *A. thaliana* dataset.

References

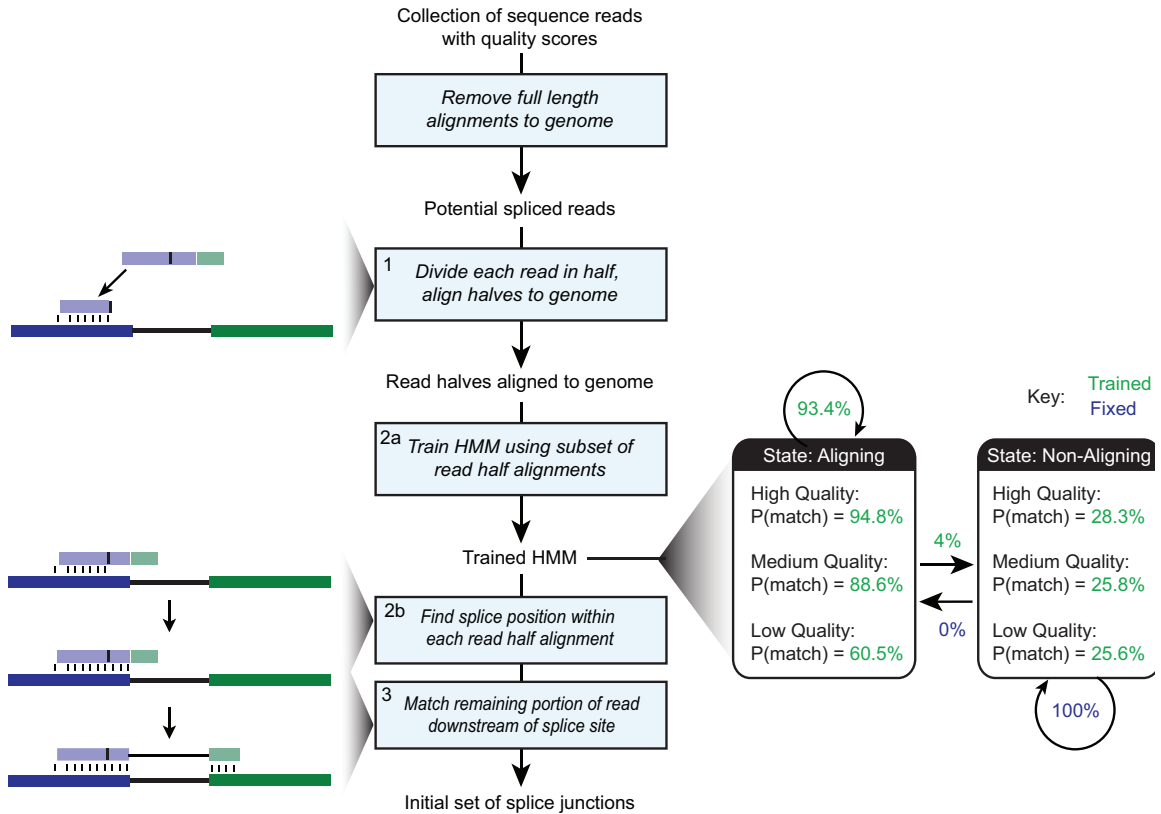
1. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet* 10: 57-63.
2. Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, et al. (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* 453: 1239-1243.
3. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25: 1754-1760.
4. Li R, Yu C, Li Y, Lam T, Yiu S, et al. (2009) SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* 25: 1966-1967.
5. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10: R25.
6. Wahl MC, Will CL, Lührmann R (2009) The Spliceosome: Design Principles of a Dynamic RNP Machine. *Cell* 136: 701-718
7. Stamm S, Riethoven J, Le Texier V, Gopalakrishnan C, Kumanduri V, et al. (2006) ASD: a bioinformatics resource on alternative splicing. *Nucleic Acids Res* 34: D46-55.
8. Yoshida H, Matsui T, Yamamoto A, Okada T, Mori K (2001) XBP1 mRNA is induced by ATF6 and spliced by IRE1 in response to ER stress to produce a highly active transcription factor. *Cell* 107: 881-891.
9. Cox JS, Walter P (1996) A Novel Mechanism for Regulating Activity of a Transcription Factor That Controls the Unfolded Protein Response. *Cell* 87: 391-404.
10. Sidrauski C, Cox JS, Walter P (1996) tRNA ligase is required for regulated mRNA splicing in the unfolded protein response. *Cell* 87: 405-413.
11. Nilsen TW, Graveley BR (2010) Expansion of the eukaryotic proteome by alternative splicing. *Nature* 463: 457-463
12. Shepard PJ, Hertel KJ (2009) The SR protein family. *Genome Biol* 10: 242.
13. Wang ET, Sandberg R, Luo S, Khrebtkova I, Zhang L, et al. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature* 456: 470-476.

14. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet* 40: 1413-1415.
15. Nagasaki H, Arita M, Nishizawa T, Suwa M, Gotoh O (2005) Species-specific variation of alternative splicing and transcriptional initiation in six eukaryotes. *Gene* 364: 53-62.
16. Sen S, Talukdar I, Liu Y, Tam J, Reddy S, et al. (2010) Muscleblind-like 1 (Mbnl1) promotes insulin receptor exon 11 inclusion via binding to a downstream evolutionarily conserved intronic enhancer. *J Biol Chem*
17. Yano M, Hayakawa-Yano Y, Mele A, Darnell RB (2010) Nova2 Regulates Neuronal Migration through an RNA Switch in Disabled-1 Signaling. *Neuron* 66: 848-858.
18. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5: 621-628.
19. Kent WJ (2002) BLAT--the BLAST-like alignment tool. *Genome Res* 12: 656-664.
20. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, et al. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 320: 1344-1349.
21. Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25: 1105-1111.
22. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, et al. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol* 28: 511-515.
23. Ameer A, Wetterbom A, Feuk L, Gyllenstein U (2010) Global and unbiased detection of splice junctions from RNA-seq data. *Genome Biol* 11: R34.
24. Bryant DW, Shen R, Priest HD, Wong W, Mockler TC (2010) Supersplat -- spliced RNA-seq alignment. *Bioinformatics* Available at: <http://www.ncbi.nlm.nih.gov/pubmed/20410051>. Accessed 7 June 2010.
25. Au KF, Jiang H, Lin L, Xing Y, Wong WH (2010) Detection of splice junctions from paired-end RNA-seq data by SpliceMap. *Nucleic Acids Res* Available at: <http://www.ncbi.nlm.nih.gov/pubmed/20371516>. Accessed 7 June 2010.
26. Dohm JC, Lottaz C, Borodina T, Himmelbauer H (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res* 36: e105.
27. Heap GA, Yang JHM, Downes K, Healy BC, Hunt KA, et al. (2010) Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing. *Hum. Mol. Genet* 19: 122-134.
28. Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* 18: 1851-1858,
29. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, et al. (2008) Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* 133: 523-536.

30. (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408: 796-815.
31. Deutsch M, Long M (1999) Intron-exon structures of eukaryotic model organisms. *Nucleic Acids Res* 27: 3219-3228.
32. Gardner MJ, Hall N, Fung E, White O, Berriman M, et al. (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* 419: 498-511.
33. Lu F, Jiang H, Ding J, Mu J, Valenzuela JG, et al. (2007) cDNA sequences reveal considerable gene prediction inaccuracy in the *Plasmodium falciparum* genome. *BMC Genomics* 8: 255.
34. Sorber K, Chiu C, Webster D, Dimon M, Ruby JG, et al. (2008) The long march: a sample preparation technique that enhances contig length and coverage by high-throughput short-read sequencing. *PLoS ONE* 3: e3495.
35. Pruitt KD, Tatusova T, Maglott DR (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucl. Acids Res.* 33: D501-504.
36. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2004) GenBank: update. *Nucl. Acids Res.* 32: D23-26.
37. Ramsköld D, Wang ET, Burge CB, Sandberg R (2009) An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput. Biol* 5: e1000598.
38. Richard H, Schulz MH, Sultan M, Nürnberger A, Schrunner S, et al. (2010) Prediction of alternative isoforms from exon expression levels in RNA-Seq experiments. *Nucleic Acids Res* Available at: <http://www.ncbi.nlm.nih.gov/pubmed/20150413>. Accessed 1 March 2010.
39. Crooks GE, Hon G, Chandonia J, Brenner SE (2004) WebLogo: a sequence logo generator. *Genome Res* 14: 1188-1190.
40. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, et al. (2002) The Human Genome Browser at UCSC. *Genome Research* 12: 996-1006.
41. Baum L, Petrie T, Soules G, Weiss N (1970) A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. *The Annals of Mathematical Statistics* 41: 164-171.
42. De Bona F, Ossowski S, Schneeberger K, Rättsch G (2008) Optimal spliced alignments of short sequence reads. *Bioinformatics* 24: i174-180.

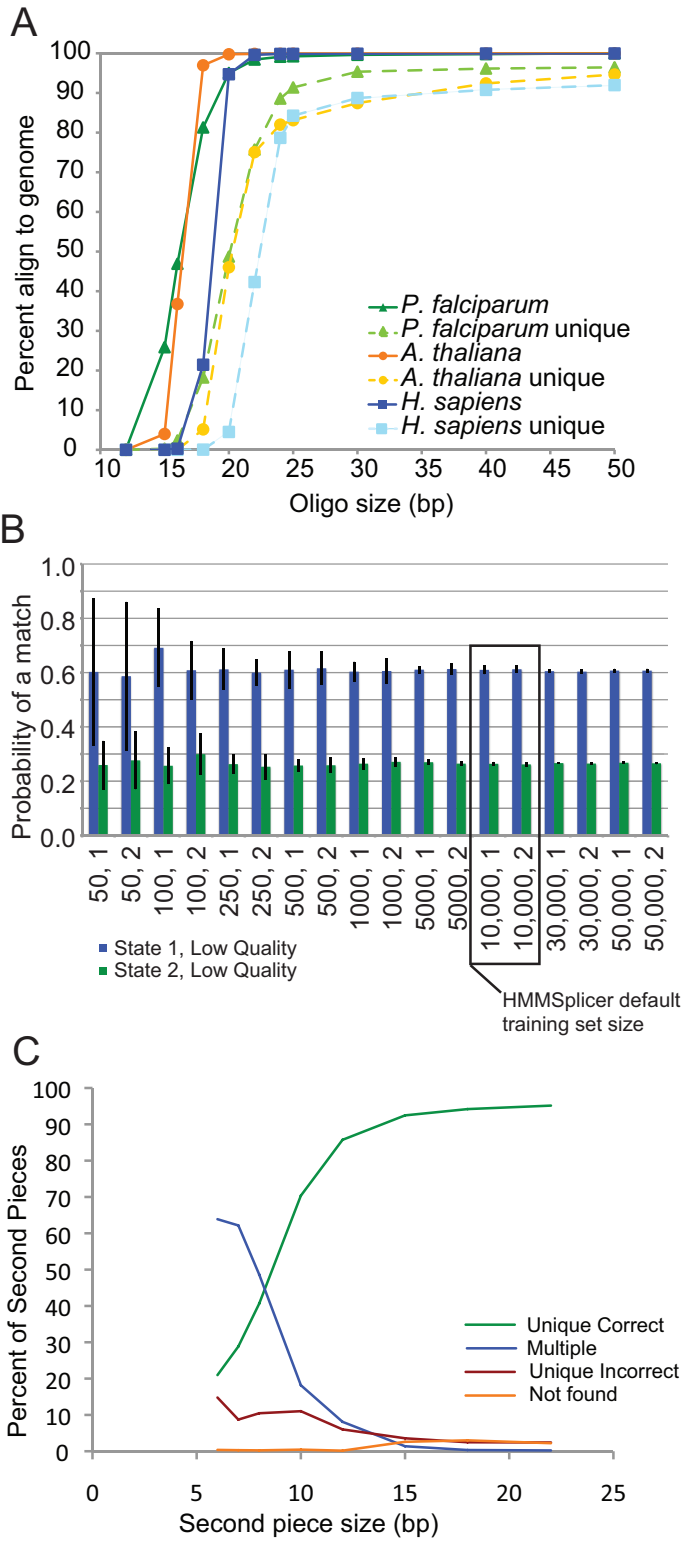
Figures

Figure 1. HMMSplicer pipeline



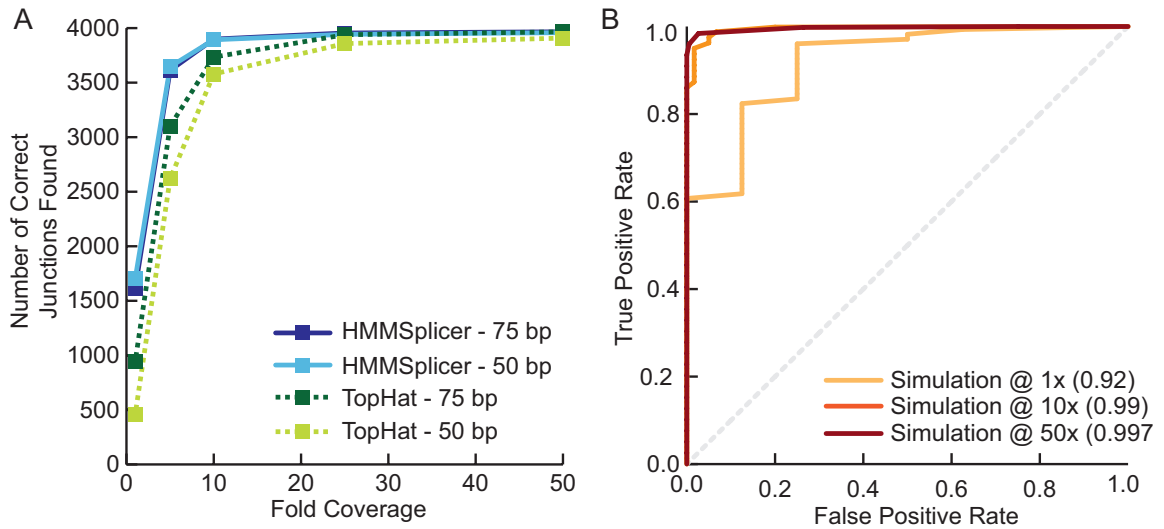
After removing reads that have full-length alignments to the genome, reads are divided in half and aligned to the genome (step 1 as defined in the Materials and Methods). The HMM is trained using a subset of the read-half alignments (step 2a). The HMM bins quality scores into five levels. Although only three levels are shown in this overview for simplification, the values for all five levels can be found in Table 1. The trained HMM is then used to determine the splice position within each read-half alignment (step 2b). The remaining second piece of the read is then matched downstream to find the other intron edge (step 3). The initial set of splice junctions then proceed to rescue (step 4) and filter and collapse (step 5) to generate the final set of splice junctions.

Figure 2. Algorithm parameters



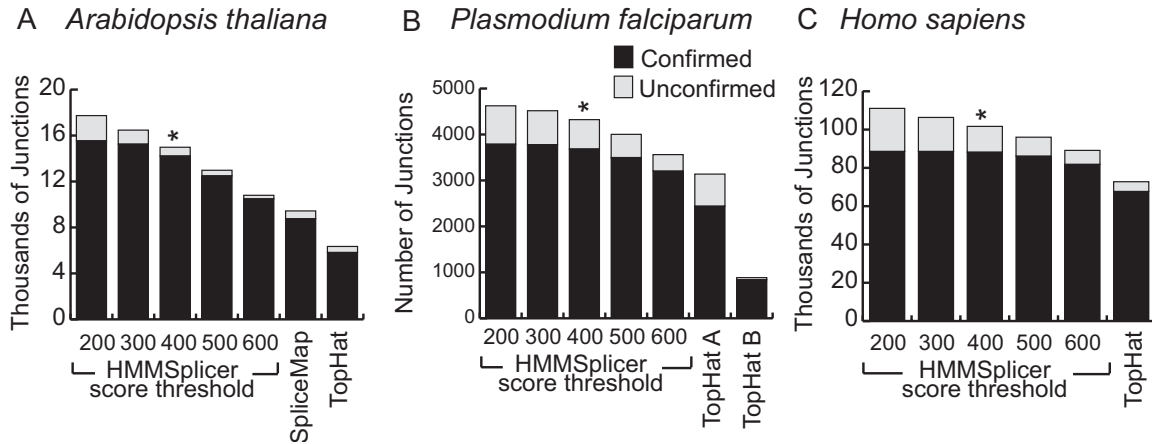
a) Percent of oligos able to map within a genome as a function of oligo size. The solid lines show the percentages if oligos are able to map up to 50 times within the genome (the value used in HMMSplicer seeding). The dashed lines show the percentages if a unique match is required. **b)** HMM training. The values for the two most variable parameters of the HMM are shown here, with the x-axis representing different training set sizes and initial HMM parameters. The error bars show the standard deviation of ten repetitions of training. HMMSplicer uses a training subset size of 10,000. **c)** Effect of size, in bases, for the second piece of the read. The percent of second pieces uniquely mapping within 80 kbp of the first piece increases as the size of the second piece increases, while the percent of second pieces mapping to multiple locations decreases.

Figure 3. Simulation results



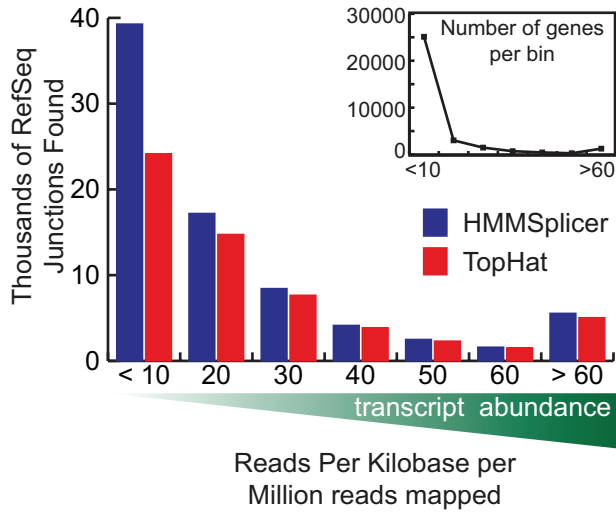
(a) Results for HMMSplicer and TopHat for 50 and 75 bp reads. Although values are similar at higher coverage levels, HMMSplicer exhibits substantial increases in sensitivity at lower coverage levels. (b) ROC curve for the 50 bp simulation results at 1x, 10x, and 50x coverage demonstrates that HMMSplicer's scoring algorithm accurately discriminates between true and false junctions. The number in parentheses is the area under the curve for each coverage level.

Figure 4. Overview of HMMSplicer and TopHat results in (a) *A. thaliana*, and (b) *P. falciparum* and (c) *H. sapiens*



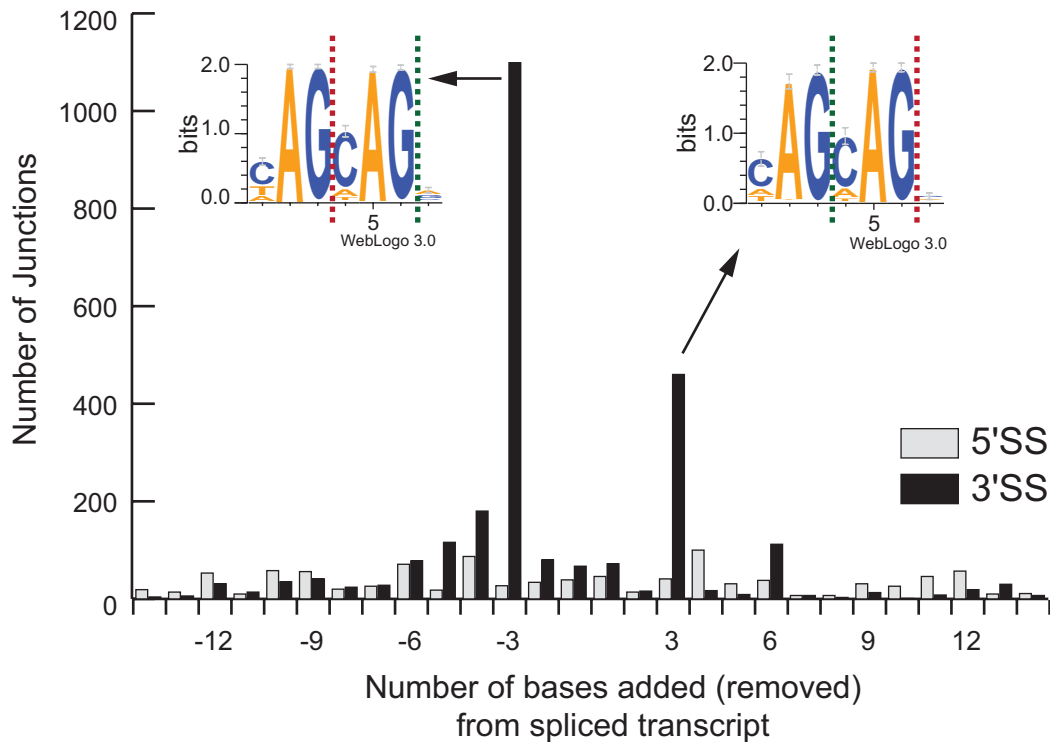
For each dataset, HMMSplicer results are shown at five different score thresholds. The numbers on the bottom axis (200 to 600) are the thresholds for junctions with multiple reads; the threshold was set 200 points higher for junctions with a single read. The * indicates HMMSplicer's default score threshold. SpliceMap results are shown for the *A. thaliana* dataset only, as SpliceMap can not be run datasets with reads less than 50 nt long. For *P. falciparum*, TopHat was run with two different parameter sets. TopHat A was run with a segment length of 23 resulting in more junctions but a lower specificity whereas TopHat B used the default segment length of 25 resulting in fewer junctions with more specificity.

Figure 5. Human results compared by transcript abundance



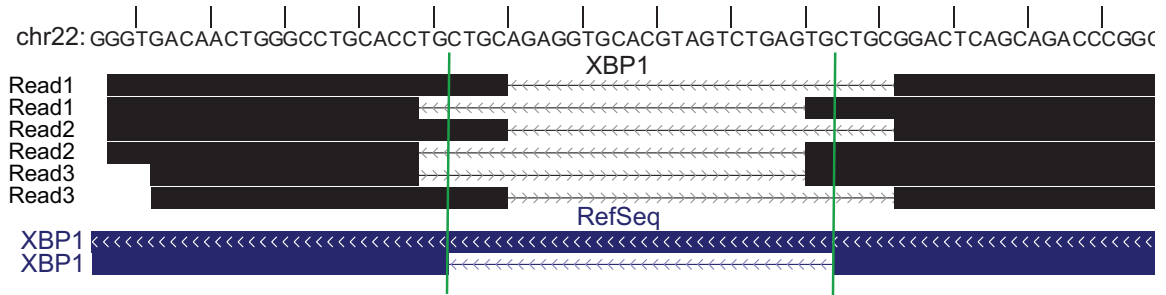
Transcript abundance was measured as Reads Per Kilobase per Million reads mapped (RPKM) and the genes were binned by RPKM to show the number of RefSeq junctions found at different levels of transcript abundance. For genes with an RPKM less than 10, HMMSplicer found 76.2% more junctions, whereas for genes with an RPKM above 50, HMMSplicer found only 6.7% more junctions. While a smaller number of highly expressed genes dominate the mRNA population, 74.8% of genes have RPKM values less than 10.

Figure 6. Alternative 5' and 3' splice sites



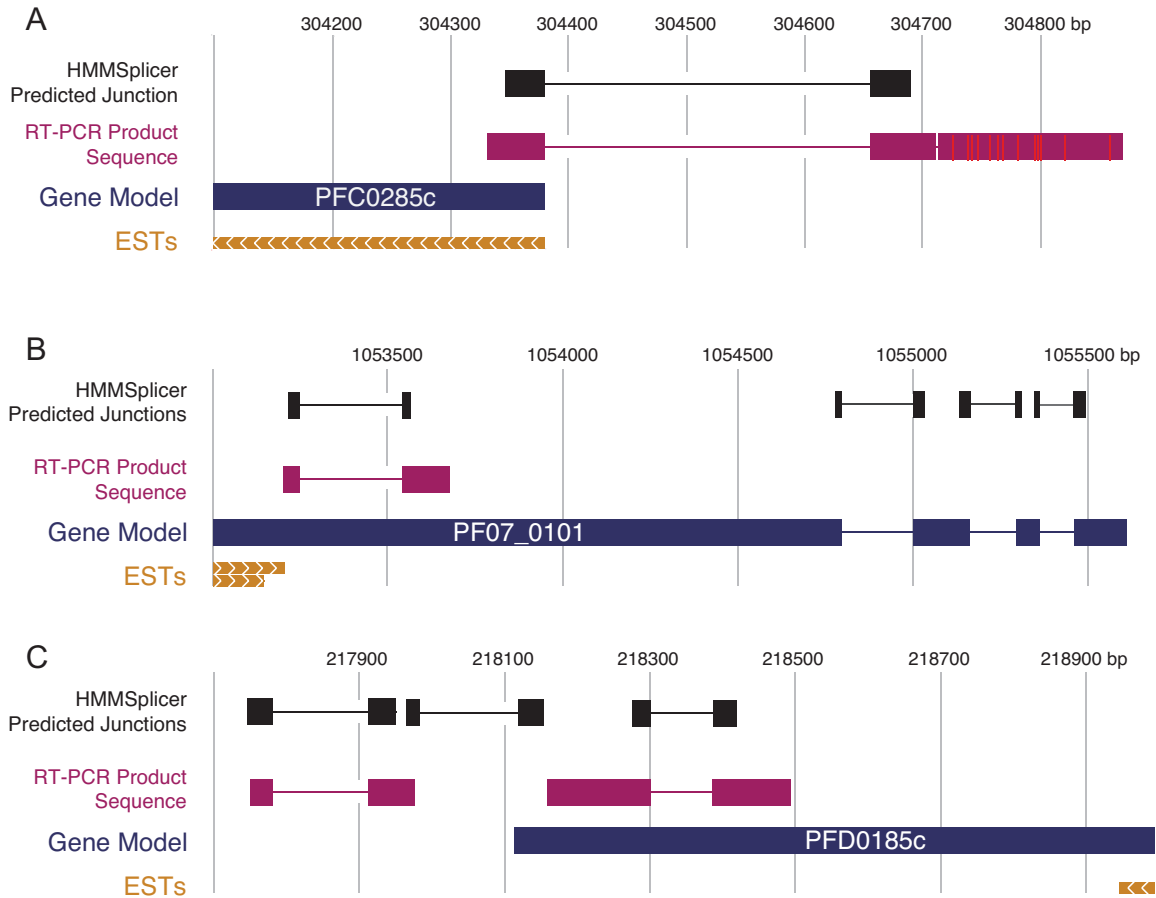
HMMSplicer results within 15 bp of RefSeq introns were analyzed to measure the number of bases added or removed from the spliced transcript. There were 997 instances where the intron had an alternate 5' splice site (5'SS, shown in grey) and 2,577 instances of an alternate 3' splice site (3'SS site, shown in black). The most common alternative splice was 3 bases removed or added to the exon at the 3'SS. TopHat results showed a similar pattern, though only 875 alternates (262 5'SS alternates and 613 3'SS alternates) are found, less than a quarter of the HMMSplicer results. Weblogos were constructed from the sequences at the 1,099 alternate 3'SS with three bases removed from the transcript and the 460 alternate 3'SS with three bases added to the transcript. For these, the green dashed line shows the alternate splice site while the red dashed line shows the canonical splice site. In both cases, a repetition of the YAG splice motif is evident.

Figure 7. *XBPI* non-canonical intron



HMMSplicer discovers the non-canonical *XBPI* intron. HMMSplicer identifies three reads containing the non-canonical CA-AG splice site in *XBPI*. Because the reads are fairly evenly split, both read-halves aligned to the genome. The edges identified by HMMSplicer are 2 and 4 bp off from the actual splice site because the sequence at the beginning of the intron repeats the sequence at the beginning of the subsequent exon. When identical junctions are collapsed, there are two junctions, one with a score of 1024 and one with a score of 1030, which puts them in the top 0.5% of the collapsed non-canonical junctions.

Figure 8. Experimental confirmation of predicted *Plasmodium falciparum* splice junctions



Schematics of the predicted splice junctions and sequenced RT-PCR products for a) PFC0285c, b) PF07_0101, and c) PFD0185c. For PFC0285c, the verified junction likely splices an additional exon in the 5'UTR to the coding region of the gene. The confirmed junction in PF07_0101 splices out 291 nt (97 aa) from the first exon, which could represent an alternative protein-coding isoform, or an error in the gene model. The demonstrated junctions in PFD0185c excise 85bp near the 3' end of the gene, causing a frameshift, and appear to splice two exons within the 3'UTR of the gene together. Again, the junction within the gene model may represent an alternative splicing event or an error

in the gene model. ESTs near all three areas are included to provide the direction of the genes.

Tables

Table 1. Simulation Results

	1 -> 2	2 -> 1	1: high	1: med - high	1: med ium	1: med -low	1: low	2: high	2: med - high	2: med ium	2: med -low	2: low
Initial Value Set 1	0.5	0	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
Initial Value Set 2	0.5	0	0.7	0.7	0.7	0.5	0.4	0.3	0.3	0.3	0.3	0.3
A. thaliana Trained Values	0.916	0	0.983	0.976	0.971	0.949	0.832	0.271	0.261	0.259	0.259	0.276
P. falciparum Trained Values	0.938	0	0.942	0.879	0.853	0.786	0.669	0.368	0.334	0.333	0.317	0.281
H. sapiens Trained Values	0.934	0	0.948	0.925	0.886	0.791	0.605	0.283	0.264	0.258	0.261	0.256

The initial and trained values for the HMM. The first two columns (“1 -> 2” and “2 -> 1”) show the probability of transitioning from State 1 to State 2 and the reverse. The probability of transitioning from State 2 to State 1 is fixed at 0 (indicating a 100% probability of remaining in State 2). For each state, the probability of a match at each quality bin is reported. The initial values were used to validate the HMM. HMMSplicer uses Initial Value Set 2, though the initial values do not impact the final trained values (see Figure 2b). The trained values are shown for each dataset analyzed. The Human values are the same as those shown in Figure 1, though in more detail.

Table 2. Simulation Results

Read Length	Coverage Level	HMMSplicer		TopHat	
		# True Positives	% False Positives	# True Positives	% False Positives
40 bp	1	1484	0.7	451	1.1
	5	3478	1.3	1858	1.1
	10	3835	2.8	2825	1.2
	25	3908	4.7	3490	2.0
	50	3928	8.3	3630	3.3
45 bp	1	1630	0.2	503	0.2
	5	3634	0.8	2422	0.9
	10	3861	1.0	3458	1.3
	25	3928	2.2	3849	2.0
	50	3947	4.1	3901	3.8
50 bp	1	1701	0.2	457	0.7
	5	3646	0.3	2619	0.8
	10	3893	0.5	3579	1.1
	25	3943	1.1	3858	2.2
	50	3958	1.6	3908	3.1
55 bp	1	1711	0.3	390	0.8
	5	3677	0.5	2697	0.7
	10	3898	0.5	3581	1.1
	25	3948	1.1	3870	1.8
	50	3965	2.5	3915	3.1
60 bp	1	1684	0.1	433	0.9
	5	3671	0.3	2629	0.7
	10	3906	0.4	3581	0.8
	25	3951	0.9	3869	1.5
	50	3966	1.0	3930	2.9
65 bp	1	1698	0.1	405	0.7
	5	3684	0.4	2609	0.6
	10	3904	0.5	3525	0.8
	25	3945	1.0	3838	1.8
	50	3966	1.3	3928	2.4
70 bp	1	1629	0.1	1038	0.7
	5	3626	0.2	3297	1.6
	10	3893	0.5	3785	2.2
	25	3951	0.7	3931	6.5
	50	3960	1.2	3958	12.9
75 bp	1	1613	0.2	943	0.5
	5	3613	0.4	3101	0.5
	10	3899	0.5	3734	0.8
	25	3955	0.6	3939	1.5
	50	3966	1.2	3966	2.4

HMMSplicer and TopHat were run on read sets from 40 to 75 bp long at coverage levels from 1x to 50x on 503 non-overlapping gene transcripts from Human Chr20.

Table 3. Datasets

	Accession Number	Number of Reads	Read Length	HMMSplicer time (min)	TopHat time (min)
<i>H. sapiens</i>	SRX011552 (used for quality model)	N/A	75	N/A	N/A
<i>A. thaliana</i>	SRX002554	79,106,696	50	326	1162
<i>H. sapiens</i>	SRX011550	9,669,944 paired end	45	880	645 (or 271)
<i>P. falciparum</i>	SRX001454 SRX001455 SRX001456 SRX001457	14,139,995	46*	108	188 (or 45)

* The 48-bp reads in the NCBI SRA set have a 2 bp initial barcode that was trimmed, resulting in 46 bp reads.

Datasets used for benchmark tests. For *H. sapiens* and *P. falciparum*, two times are given for TopHat. For *H. sapiens*, the longer time is with more sensitive settings, but the shorter time resulted in less than 5% fewer junctions at a similar specificity. For *P. falciparum*, the longer time is with more sensitive but less stringent settings whereas the shorter time is for the more stringent settings that resulted in significantly fewer junctions but with a much higher specificity.

Chapter 4: RNA-Seq analysis of splicing in *Plasmodium falciparum* uncovers novel splice junctions, alternative splicing, and splicing of antisense transcripts

This chapter is a manuscript, prepared for publication in a peer-reviewed journal.

Katherine Sorber, Michelle Dimon and Joseph L. DeRisi. RNA-Seq Analysis of Splicing in *Plasmodium falciparum* Uncovers Novel Splice Junctions, Alternative Splicing, and Splicing of Antisense Transcripts.

Author contributions:

Katherine Sorber performed the experiments and wrote the manuscript. Katherine Sorber and Michelle Dimon analyzed the data. Katherine Sorber, Michelle Dimon, and Joseph DeRisi conceived the project and guided the implementation.

Joseph L. DeRisi, thesis advisor

Abstract

Over 50% of genes in *Plasmodium falciparum*, the most deadly of the human malaria parasites, contain introns, yet little is known about splicing in the parasite. Using publically available RNA-Seq data combined with our own RNA-Seq dataset as input for the in-house developed HMMSplicer algorithm, we have captured the splicing events that occur during red blood cell development on a transcriptome-wide level. In addition to confirming 6,678 previously known gene model or EST splice junctions, 977 novel 5' GU-AG 3' and 5 novel 5' GC-AG 3' junctions were discovered. We detected 310 alternative splicing events in 254 genes, as well as additional events in intergenic regions, thus compiling the largest set of alternative splicing events in the parasite to date. The data also reveal splicing events antisense to gene models, some of which are attributable to overlap between annotated genes on opposite strands, while most are likely due to gene overlap with unannotated antisense transcripts. Some of these antisense transcripts also appear to be subject to alternative splicing. Overall, this effort has advanced our knowledge of not only splicing, but also of transcriptome organization in the human parasite *Plasmodium falciparum*.

Introduction

Close to one million people every year are killed by malaria, an infectious disease caused by protozoan parasites of the genus *Plasmodium* [1]. While five species of *Plasmodium* are known to infect humans [2], *Plasmodium falciparum* is the deadliest. In efforts to understand the parasite's basic biology and thus discover unique vulnerabilities, its transcriptome has been probed extensively for RNA expression data [3-5]. However, information on other transcriptome features, such as the structure of expressed transcripts, has lagged behind. Analysis of recent expressed sequence tag (EST) and RNA-Seq datasets has focused mainly on improving gene annotations [6-9]. Thus, while more than half of the parasite's genes are predicted to contain introns [10], a deeper analysis of splicing in this organism has not been undertaken.

Splicing, the mechanism by which intronic sequences are removed from transcripts and exonic sequences are joined to form mature RNA species, has been well studied in model organisms. The two required transesterification steps are catalyzed by the spliceosome, a large complex of small nuclear ribonucleoproteins (snRNPs). Each snRNP contains an snRNA component, as well as a core set of proteins. In *P. falciparum*, RNA components of the major U2-type spliceosome have been detected [11, 12], yet protein components have not been systematically identified. Using reciprocal best hits (RBH) analysis [13] of human and yeast factors involved in splicing, we identified putative orthologs to spliceosome and spliceosome-associated protein components (Table 1) [14-16], the majority of which were most similar to their human counterparts. However, three components of the human spliceosome could not be identified. Two of these factors,

PPIE and SYF2, belong to the hPrp19/Cdc5 complex, which stabilizes U5 and U6 interactions with the RNA after U4 has been released [17], although their individual contributions to this function remain unknown. Surprisingly, we were also unable to detect an ortholog of PRP2. This DEAH/D-box ATPase is thought to activate the pre-spliceosome by inducing a structural rearrangement that results in disassociation of the SF3a and b complexes from the branchpoint, rendering it competent for nucleophilic attack of the 5' splice site [18]. Such an indispensable function in splicing is unlikely to be absent in the parasite and may be performed by either a very weak PRP2 ortholog or another DEAH/D-box ATPase in the parasite's proteome.

The splicing process is also guided by conserved motifs within the intronic sequence to be removed. In order across the intron, these motifs include the 5' splice site (AG|GUAAUGU in yeast, AG|GURAGU in mammals), the branch point sequence (UACUAAC 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) [19]. In *P. falciparum*, introns identified from EST data generated putative 5' (AR|GUAAANW) and 3' (YAG|) splice site motifs [12]. As in most eukaryotes, these motifs are quite degenerate, leaving the first and last two nucleotides of the intron (5' GU-AG 3') as the most consistent markers of intronic sequence. In other organisms, a minority of introns are marked by noncanonical splice sites such as 5' GC-AG 3', which is recognized by the major U2-type spliceosome, and 5' AU-AC 3', which is recognized by the minor U12-type spliceosome [20]. Noncanonical splice sites have been mapped in

P. falciparum EST data [9, 21], yet all studies to date have excluded explicit analysis of these events and thus their contribution to splicing in the parasite remains unknown.

Alternative splicing adds an additional layer of complexity to RNA processing by generating several different mature transcripts from the same precursor. This process is crucial to such diverse biology as *Drosophila* sex determination and *HIV-1* replication [22, 23]. In 1991, the first instance of alternative splicing in *P. falciparum* was reported in the 41-3 antigen gene [24]. Since then, more than 100 alternative splicing events have been catalogued in *Plasmodium* species [6, 25-29], ranging from intron retention to alternate splice site usage to exon skipping. However, a dedicated transcriptome-wide characterization of alternative splicing in *P. falciparum* has not been undertaken.

Often, alternative splicing is guided by the presence or absence of proteins that determine which splice sites are available to the spliceosome [30]. In humans, a number of arginine/serine-rich proteins (SR) or heterogeneous nuclear ribonucleoproteins (hnRNPs) act as alternative splicing factors, in addition to their roles in constitutive splicing. To determine if *Plasmodium falciparum* has orthologs to these proteins, human SR and hnRNP proteins with documented roles in alternative splicing were used for best reciprocal hits analysis [31, 32]. Four SR proteins and one hnRNP protein returned specific orthologs (Table 1, factors without clear orthologs not shown). These orthologs likely represent only a fraction of those proteins that may influence splice site selection in *P. falciparum*. 71 additional *P. falciparum* proteins contain either an RNA recognition motif (RRM) or an RNA binding domain (RBD) according to InterPro [33], and 7 contain

an RS domain according to our own analysis (see Materials and Methods). While these domains do not guarantee involvement in splicing, many proteins involved in splice site selection during alternative splicing contain some combination of them [31, 32].

Recent analyses have shown that transcriptome complexity in many organisms extends beyond alternative splicing. The presence of dense transcriptional arrangements, such as overlapping protein-coding genes (in parallel or antiparallel orientation) and natural antisense transcripts [34, 35], appears to be commonplace. While the function and evolutionary importance of these complex arrangements is not yet well understood, some are known to be important in regulatory relationships between the paired genes [36]. In the current *P. falciparum* genome annotation, six instances of protein-coding gene overlap are documented, resulting in 1 parallel gene pair and 5 antiparallel gene pairs. In addition, both short and long antisense transcripts have been described in the parasite, and EST data indicates that some antisense transcripts may be spliced [9, 37, 38]. However, the full extent of these complex transcriptional arrangements remains unclear.

Although splicing, alternative splicing, gene overlap, and antisense transcription have all been observed in *Plasmodium falciparum*, none has been thoroughly studied. By analyzing the splicing events that occur in the parasite's blood stages using both our own RNA-Seq data and RNA-Seq data from Otto et al. and Sorber et al. [6, 39], we gained insight into these transcriptome features. 977 novel 5' GU-AG 3' splice junctions and 5 novel 5' GC-AG 3' splice junctions were found in the combined dataset using the in-house developed splice site detection algorithm HMMSplicer (see Chapter 3). Probing the data

for conflicting splicing events produced an extensive, ranked list of possible alternative splicing events, as well as splicing events antisense to one another. Many novel alternative and antisense splicing events above and below the score threshold were successfully verified using a modified nested RT-PCR strategy. Alternate 5' and 3' splice sites were the most common type of alternative splicing encountered, while skipped exons comprised 14% of observed events. The majority of events were documented within 254 gene models, but alternative splicing was also observed in intergenic regions and in antisense transcripts. We believe that the antisense splicing events observed here signify additions to the growing list of antisense transcripts produced in *P. falciparum* and fall into two classes – those belonging to neighboring genes and thus representing gene overlap between already annotated genes, and those belonging to unannotated transcripts antisense to known gene models. Thus, this analysis of splice junctions in *Plasmodium falciparum* has provided insight into several understudied transcriptome features of the parasite.

Materials and Methods

Best reciprocal hits analysis

A list of *S. cerevisiae* and *H. sapiens* splicing factors was assembled from the literature [14-16, 31, 32, 40-42] and protein sequences were obtained from the Saccharomyces Genome Database or the human International Protein Index ("Saccharomyces Genome Database" <http://downloads.yeastgenome.org/> (06-18-2010), "Human International Protein Index" <http://www.ebi.ac.uk/IPI/IPIhuman.html> (06-23-2010)). Each protein sequence was used in a BLASTp search of the *P. falciparum* proteome, and the top resulting hit was recorded

and then used as the query sequence for a reciprocal BLASTp of the appropriate transcriptome [13, 43].

Prediction of RS domains in Plasmodium falciparum proteins

Prediction of *Plasmodium falciparum* RS domain proteins was adapted from Boucher et al. [44]. BLASTp was performed on the *P. falciparum* proteome (PlasmoDB v6.3) using an artificial domain of 30 RS amino acid repeats (60 total amino acids). Hits with e-value < 0.05 had to have RSRS or SRSR exactly in their amino acid sequence to be retained.

Generation of timepoint samples

3D7 Oxford *Plasmodium falciparum* parasites were grown at 2% hematocrit in 30 x T150mL flasks with 50mL of volume each. Repeat synchronization during peak invasion (number of rings = number of schizonts), and again 12 hours later, over 3 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. [3]. 4 hours later, the culture was diluted to approximately 5% hematocrit with 3 L of culture medium. 50% of the culture was harvested 11 hours after invasion, pelleted, and frozen at -80°C. 33% of the culture was harvested 22 hours after invasion, 10% 33 hours after invasion, and 7% 44 hours after invasion. Total RNA was harvested from

frozen pellets using Trizol (Invitrogen Corp., Carlsbad, CA), then poly-A selected using the Micro FastTrack 2.0 kit (Invitrogen Corp., Carlsbad, CA).

Generation of RNA-Seq libraries

Libraries were generated as in Sorber et al. [39]. Briefly, 1.2-1.6 μ g of polyA-selected RNA was reverse transcribed using 6bp-EciI-N₉ (all primers can be found in Supplementary Table 1), and second strand cDNA synthesis was carried out with 13bp-ModSolS-N₉. 5 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., Carlsbad, CA). Bead-bound material was digested with EciI, then treated with Antarctic Phosphatase (New England Biolabs, Ipswich, MA). Sol-L-NN annealed adapter was ligated onto cut ends. 5 final cycles of PCR were performed on ¼ 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. 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 [39]. 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-marched sub-library were clustered on an Illumina flow cell in separate lanes (Illumina, Hayward, CA). For single-end libraries and the first read of paired-end libraries, Sol-SeqPrimer was used as the sequencing

primer, while PE-SolS-SeqPrimer was used to sequence the second read of paired-end libraries. Up to 60 single base pair extensions were performed with image capture using an Illumina GA2 sequencer (Illumina, Hayward, CA; see Supplementary Table 2). The Illumina Pipeline software suite version 0.2.2.6 (Illumina, Hayward, CA) was utilized for base calling from these images for TP3 and TP4, while versions 1.3.2 and 1.5.0 were used to base call TP1 and TP2 images.

Analysis pipeline

Raw sequence data from the above timecourse as well as from Otto et al. and Sorber et al. was subjected to the following analysis pipeline [6, 39]. First, barcodes were removed if necessary. Next, reads were discarded if they contained greater than 12 nt of primer sequence, a repeat of A, T, C, G, or AT longer than 11 nt, or if they had more than 10 nt with a quality scores ≤ 5 . Identical sequences within a lane were compressed to a single sequence read. The reads were then filtered to remove human sequences, as detected by blast against the human genome with an E-value of 0.00001 [43].

The filtered read set was aligned to the *Plasmodium falciparum* genome, PlasmoDB version 6.3 [45], by Bowtie version 0.12.1, using default parameters except that alignment of reads with multiple matches was not allowed [46]. The reads unaligned by Bowtie were then aligned using Blat for a less stringent alignment [47]. Blat version 34 was run with a tile size of 11, a step size of 1, using an ooc file to filter repetitive sequence. The Bowtie alignments were combined with the high-scoring Blat alignments to yield a final set of aligned reads from which coverage statistics were generated.

HMMSplicer was also run 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 (see Chapter 3). All other parameters were left at default values.

Operational definitions for data analysis

See Supplementary Figure 1 for a description of terminology used to refer to specific parts of individual splice junctions. For all definitions referencing gene models, a junction was established as mapping to a gene model only if an inner edge of the junction fell within the coordinates of the gene model.

Known junction – maps to the same boundaries as a splice junction found in
PlasmoDBv6.3 gene models or in EST data

Novel junction – maps to boundaries not seen in PlasmoDBv6.3 gene models or in EST
data

Canonical junction – maps to 5' GU-AG 3' boundaries

Noncanonical junction – maps to all other boundaries

Conflicting junctions – a pair of splice junctions where 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 in different transcript isoforms

Antisense conflict – two junctions that conflict, with boundaries on opposite strands

Antisense junction – a splice junction that maps to boundaries antisense to a gene model, including junctions where only one boundary falls within the gene model

Mutually exclusive alternate 5' and 3' splice sites – a pair of splice junctions where both the 5' and 3' splice site conflict

Skipped exon – a splice junction that conflicts with two or more other junctions that themselves do not conflict

Gene overlap – two annotated genes on opposite strands connected by either an antisense junction or by an EST

Validation of conflicting splicing events

A small-scale timecourse similar to the Bioreactor timecourse was performed using highly synchronous 3D7 Oxford parasites. After invasion, samples were taken at 11, 22, 33, and 44 hours and processed for total RNA as described above. For each timepoint, 1.5µg of total RNA was reverse transcribed at 42°C for 1.5 hours using 1.9µM random hexamer with Superscript III (Invitrogen Corp., Carlsbad, CA). For each validation, 1µL of crude cDNA from the lifecycle stage with the highest representation of the novel junction by RNA-Seq was used in the outer PCR reaction with Herculase II Fusion polymerase (Agilent Technologies, Inc., Santa Clara, CA) and the appropriate outer primers (see Supplementary Table 3). PCR conditions were 95°C for 2min, followed by 20 cycles of 95°C for 30s, 52°C for 45s, and 65°C for 3min. Outer PCRs were then purified using Zymo DNA-5 Clean and Concentrator columns (Zymo Research Corp., Orange, CA). 1/20th of the purified PCR product was used in a restriction digest reaction with 1-5U of the appropriate enzyme (Supplementary Table 3). Digest conditions were

as recommended by the individual manufacturers and all digests were allowed to proceed for 1 hour, then were purified as above. Inner PCRs used 1/10th of the purified digestion reaction with the same PCR conditions, except 30 cycles were performed and the appropriate inner primers were used (Supplementary Table 3). Size appropriate bands were gel extracted from a 2% agarose gel using Promega's Wizard SV Gel Extraction and PCR Clean-Up System (Promega Corp., Madison, WI), then TOPO TA cloned (Invitrogen Corp., Carlsbad, CA), and whole cell PCR of positive colonies was performed with M13F and R primers and Taq polymerase (Invitrogen Corp., Carlsbad, CA). Whole cell PCR products were sequenced using the BigDye Terminator v3.1 Cycle Sequencing Kit on an ABI 3130xl Genetic Analyzer (Life Technologies Corp., Carlsbad, CA). Resulting sequences were trimmed for vector, then aligned to the *Plasmodium falciparum* genome (v6.3) using BLAT [47].

Results

Overview of Plasmodium falciparum RNA-Seq datasets

To investigate splicing in *Plasmodium falciparum* on a transcriptome-wide scale, we catalogued splice junctions from the parasite's blood stages using short read RNA-Seq data. To guarantee adequate representation of the different parasite blood stages, timepoints were collected from a highly synchronous 3D7 Oxford culture approximately 11, 22, 33, and 44 hours post-invasion, representing ring, trophozoite, late trophozoite/early schizont, and late schizont parasites. After total RNA isolation, poly-A RNA was purified and prepared for Illumina sequencing using the Long March protocol [39]. Two other *Plasmodium falciparum* RNA-Seq datasets have been published

previously, one from 7 timepoints within the blood stage of 3D7 parasites by Otto et al. and one from the late schizont timepoint of our experiment [6, 39]. To maximize our transcriptome-wide examination of splicing, we included all sets of raw data in our analysis.

The combined data was first filtered to remove reads with adapter sequence, long low complexity stretches, or an abundance of low quality bases. After sequence compression by timepoint, the data was further filtered for human sequence. This set of filtered reads was then used in two parallel analyses - the short read aligner Bowtie [46] and BLAT [47] were used to align reads back to the genome full length, while HMMSplicer was used to identify reads that crossed exon-exon splice junctions (see Chapter 2). The Bowtie/BLAT pipeline was able to align approximately 84-194 million bases of sequence to the *P. falciparum* genome for each independent timepoint (Supplementary Table 2). Discounting antigenic variation gene families (*272 vars*, *rifins*, and *stevors*), the combined dataset gave us a median coverage of 59 reads per coding nucleotide per gene. 93% of annotated genes had a median coverage of at least 5 reads per nucleotide. HMMSplicer was also run on the filtered dataset, with a minimum intron size of 5bp and a maximum intron size of 1000bp, covering 99.6% of all annotated *P. falciparum* introns. Over 1.9 million reads in the combined dataset were mapped to junctions by the algorithm (Supplementary Table 4).

After identification of splice junctions, HMMSplicer assigns each one a score based on the strength of the supporting data and alignment. A score is initially calculated for each

read, and reflects the number and sequence quality of bases supporting each side of the junction, as well as the validity of a gapped versus ungapped alignment. As reads that support the same junction are collapsed, its score is increased for each new exonic base added, thereby indirectly rewarding junctions covered by multiple, unique reads. Such scoring is highly accurate in discriminating true splice junctions from false positive splice junctions (see Chapter 3). However, the algorithm reports all detected splice junctions, allowing the user to determine the appropriate results set for further analysis.

For *Plasmodium falciparum*, PlasmoDB gene models and ESTs provide a set of previously known splice junctions that are likely to be legitimate [7-9, 45]. Therefore, to choose a set of HMMSplicer results for accurate transcriptome-wide characterization of splicing, 5' GU-AG 3' junctions found by HMMSplicer were compared to known 5' GU-AG 3' junctions (Fig 1A), as the majority of *P. falciparum* splice junctions adhere to canonical eukaryotic intron boundaries [9, 48]. This comparison emphasized the bimodal distribution of the HMMSplicer results, with the majority of known junctions segregating by score from the majority of novel junctions. Even though junctions with scores as low as 489.4 matched previously known junctions, the score threshold was set at 1075 since novel junctions outnumber known junctions below this score (Fig 1A). Although most junctions above 1075 are predicted by more than one read, evenly split, high quality single reads with unambiguous alignments can theoretically score 1200.

Canonical Splicing

HMMSplicer finds 7,655 5' GU-AG 3' junctions scoring at least 1075 within the RNA-Seq data. Of these, 6,678 (87.2%) confirm introns in PlasmoDB gene models or ESTs. 977 (12.8%) support novel introns, with the highest-ranking novel junction scoring 1920 (Fig 1B). 431 (43.9%) of these novel junctions fall either totally or partially outside of annotated gene models, while 544 (55.4%) align within gene models. Interestingly, 2 (0.2%) novel junctions directly link genes encoded on opposite strands and suggest gene overlap.

To confirm the validity of the novel 5' GU-AG 3' junctions, MEME motifs were calculated for their 5' and 3' splice sites [49]. These motifs were also calculated from the set of previously known junctions for comparison. The novel junctions were expected to produce the same motifs as the known set if they represent true splicing events. Indeed, no significant differences were observed between our calculated motifs for PlasmoDB/EST matching junctions versus novel junctions, and both sets of motifs closely matched previously published results (Fig 1C) [12]. In contrast, motifs produced from the bottom 10% of all novel junctions (score < 579) contained little information other than their 5' GU-AG 3' boundaries (Supplementary Figure 2).

Noncanonical Splice Junctions

The presence of 5' GC-AG 3' junctions in mapped *P. falciparum* ESTs [9, 21] suggests that the parasite possesses genuine noncanonical splice junctions, yet this likelihood has never been explicitly studied. Since HMMSplicer reports all junctions regardless of boundaries, we explored this possibility. However, three lines of evidence led us to

suspect that many of the 984 non-5' GU-AG 3' junctions with scores ≥ 1075 were false positives, and that discovery of credible noncanonical junctions would require additional curation. First, 11.3% of all predictions above 1075 were bound by noncanonical splice sites, a significantly higher percentage of noncanonical junctions than typically found in other organisms [50, 51]. Second, the most common noncanonical boundaries represented in the list were not those that have been previously described (i.e. 5' GC-AG 3' or 5' AU-AC 3') [50]. Finally, manual inspection revealed that many of them could be reassigned to canonical boundaries if read errors (such as base substitutions near the true splice site or small deletions) were invoked. In fact, given that the probability of an erroneous read mapping to noncanonical boundaries (i.e. all possible boundary combinations except 5' GU-AG 3') is much greater than the probability of it mapping to canonical boundaries, it is not surprising that false positive rate within the noncanonical junctions list is higher than in the canonical junctions list. However, this high false positive rate caused the HMMSplicer score alone to be unable to distinguish the few true junctions from the noise created by erroneous reads.

To aid in discovery of true junctions, two additional filters were applied to the noncanonical list. Predictions where two canonical boundaries matching a previously known junction or one of our novel junctions could be reached by moving up to 15bp either to or from the other side of the read were filtered out first. This filter was very similar to the sequence-dependent "wobble" step of HMMSplicer, except that it allowed for more bps to be wobbled in a sequence-independent manner. However, a significant number of noncanonical junctions remained. Many aligned in areas of very high

coverage. High coverage areas tend to generate more false positive junctions because a higher number of total reads results in a higher number of reads with errors. Therefore, we imposed another filter eliminating junctions where coverage exceeded 1000 reads/bp within 100bp of the junction's outer edges. This filter lowered the number of noncanonical junctions to 273 with scores of at least 1075.

The filtered predictions were then manually inspected. Many were still explainable by read errors, such as missing or inserted bases. However, 12 credible 5' GC-AG 3' junctions were found within this set. Seven were supported by either EST evidence or annotated PlasmoDB gene models, while five novel 5' GC-AG 3' junctions corresponded well with full-length Bowtie coverage in the area (i.e. higher coverage in the proposed exonic versus intronic regions).

We used MEME to construct 5' and 3' splice site motifs from the 12 5' GC-AG 3' junctions [49] (Fig 2). The 3' splice site motif was very similar to the 3' splice site motif for canonical *P. falciparum* junctions. However, there were several clear differences between the 5' splice site motifs for 5' GC-AG 3' junctions versus canonical junctions. While canonical *P. falciparum* 5' splice sites have a slight preference for AG as the last two bases of the 5' exon, this preference is much stronger for 5' GC-AG 3' junctions. In fact, all 12 examples contained AG in these positions. In addition, the preference for A at the third position of the intron is also absolute in all 12. Perhaps most interesting is the strong preference for G at the fifth position in the intron. While G is strongly preferred at this position in human canonical introns [12], and mutation of this G to other bases

reduces splicing fidelity in yeast [52], *P. falciparum* canonical introns have almost no base preference at this position (Fig 1C) [12]. These more rigid sequence preferences could indicate a stronger reliance on surrounding context for recognition of 5' GC splice sites. Stronger consensus 5' splice site motifs have been documented for 5' GC-AG 3' introns in other organisms as well [51].

5' AT-AC 3' splice sites have also been observed in introns excised by the U12 minor spliceosome [53]. Consistent with our failure to find any *P. falciparum* orthologs to proteins specific to the human U12-type spliceosome (data not shown) [42], no credible 5' AT-AC 3' junctions were observed in the filtered noncanonical junction set.

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].

Some PlasmoDB junctions are not encountered

Of the 8,435 predicted junctions in PlasmoDB gene models, 1,103 were not seen in the combined dataset even below the threshold. Most of these belong to genes that were not substantially expressed during the experiments analyzed here. However, we have junctions for 65 of them that do not match the boundaries indicated by the gene model, signifying either the presence of an isoform in the blood stages that differs from the gene model or a gene model error.

Conflicting splice junctions represent alternative and antisense splicing

Alternative splicing increases transcriptome complexity by generating multiple isoforms from the same precursor. Isoforms can differ in single 5' or 3' splice sites or in whole exons and introns. To find alternative splicing within the combined dataset, including events between novel junctions, we constructed a “conflicting junctions” list. In a pair wise manner, all canonical junctions found by HMMSplicer, as well as all 12 putative 5' GC-AG 3' junctions, were compared to each other. To be counted as a conflicting pair, one of the inner edges of a junction must have aligned within the intronic area of the other junction. Such instances were then ranked by the HMMSplicer score of the lowest scoring junction of the pair.

While constructing our conflicting junctions list, we noticed a class of conflicts in which one junction contained intron boundaries in a given direction while the other mapped to intron boundaries in the opposite direction. Although none of the datasets analyzed here were derived from a directional library, the orientation of intron boundaries has been used in the past to assign direction to ESTs [55]. In addition, widespread antisense transcription has been previously noted in *P. falciparum* [37], and ESTs antisense to gene models have also been documented [9]. Therefore, we believe it likely that these “antisense conflicts” derive from overlap of two spliced transcripts transcribed in opposite directions.

To experimentally confirm our threshold for conflicting splice events, 21 events ranked at the top of the list, as well as 20 events below 1075, were tested using the strategy described in Figure 3A. This set included twelve 5' alternate splice sites, two 3' alternate

splice sites, seventeen skipped exons, and ten spliced antisense transcripts (Table 2). For simplicity, only conflicts between a previously known junction and a novel junction were tested. For each validation, outer PCR primers flanking the area of interest were used to amplify the region from reverse transcribed total RNA from the appropriate timepoint of an independent 3D7 Oxford timecourse. Outer PCR reactions were then digested with a restriction enzyme that cut only the known isoform, leaving the novel isoform intact. Nested inner PCR primers were then used to amplify any uncut material to detectable levels. These PCR products were sequenced to confirm the identity of the resulting bands.

Above the threshold, 19 of 21 (90.5%) conflicting splice events with scores ranging from 1189.3 to 1544.2 were confirmed, including a skipped exon in MAL13P1.159 (thioredoxin) and an antisense conflict mapping to PFF0290w (long chain polyunsaturated fatty acid elongation enzyme) (Fig 3B, C). As expected, the confirmation rate was lower for the 20 conflicting events tested below the score threshold. However, 13 (65%) of these events with scores ranging from 984.6 to 1050.5 were also confirmed, verifying 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, Fig 3D). The lowest scoring event tested, an antisense conflict mapping to PF11_0355 (conserved *Plasmodium* protein), was verified, while higher scoring conflicts were not. Differences in primer or restriction enzyme efficiencies could account for this inconsistency, as validation schemes were not individually optimized.

Genome-wide characterization of alternative splicing

Conflicting splicing events scoring at least 1075 were organized into clusters to characterize alternative splicing. Clusters were built by iteratively searching the conflicting junctions list for other pair wise conflicts involving any of the junctions already populating the cluster. These initial clusters were then divided into sense and antisense clusters where applicable. 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 (Table 3). The majority of these events occurred in gene models in the sense direction, though some also occurred in intergenic areas outside of gene models. While these intergenic regions could not be definitively assigned to gene models, they most likely indicate alternative splicing in unannotated *P. falciparum* UTRs. Interestingly, all four types of alternative splicing were also seen in antisense clusters. Further analysis of antisense splicing is discussed in the next section.

Of the 310 alternative splicing events mapped to gene models in the sense direction, 279 (90%) have the potential to change the predicted coding sequence. Of those, 82 (29.4%) maintain coding frame, with 13 adding between one and twelve amino acids, and 68 removing between one and 148 amino acids from the predicted protein. In contrast, the majority of alternative splicing events (197 or 70.6%) result in frameshifts. Six of these events lead to elongation of the coding region, 162 introduce premature termination codons, and 29 could not be analyzed without assuming which downstream junction(s) co-occur in a given isoform.

In other organisms, protein-coding genes have been shown to produce frameshifted alternate transcripts with premature stop codons that are marked for degradation by the nonsense-mediated decay (NMD) pathway as a mechanism of post-transcriptional regulation [56, 57]. Thus, one possible explanation for the abundance of frameshifting alternative splicing events is that these transcripts represent intermediates destined for NMD, although this pathway has never been shown to exist in *P. falciparum*. The core components of the surveillance complex, UPF1, UPF2, and UPF3, are conserved from yeast to humans (although humans contain two UPF3 paralogs, UPF3a and UPF3b) [58]. Using human and yeast sequences for the core surveillance proteins, best reciprocal hits analysis was able to find orthologs to all three necessary NMD components in *Plasmodium falciparum* (Table 1). In lower eukaryotes, distance from the normal context of a stop codon (such as the poly-A tail) triggers NMD. While it is unclear what the criteria may be in *P. falciparum*, 161 of 162 events introduce premature termination codons more than 100bp upstream of the “normal” stop codon. In addition, 119 (73%) of the 162 alternative splicing events that introduce premature termination codons do so more than 50bp upstream of the last splice junction, which would render them eligible for NMD in mammalian systems [58].

Because RNA-Seq data are comprised of short reads, the combination of splice sites used in a given isoform cannot be resolved, and thus the exact number of isoforms encoded by the alternative splicing events discerned here could not be determined. However, transcriptome-wide, the combined dataset 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. The alternative splicing events for most genes maximally supports between 2 and 4 isoforms. However, a handful of genes (PF14_0338, PFF0630c, PFL1440c, PFC0495w, and PFC0912w) have the potential to encode up to 16 different isoforms. PFC0495w (plasmepsin VI) is particularly interesting because, in addition to supporting up to 8 sense isoforms, an overlapping antisense junction was also validated for it (Table 2).

Genome-wide characterization of antisense splicing

Because the initial discovery of antisense splice junctions in the RNA-Seq data was made by looking for conflicts between junctions, we expanded the antisense junctions list to include all junctions with at least one intron boundary antisense to an annotated gene model. Using the score threshold of 1075, this list contains 199 antisense junctions mapping to 148 gene models. In addition, antisense junctions overlapping 16 of these genes appear to undergo alternative splicing to produce between 38 and 59 different isoforms in total.

Interestingly, over 86% of antisense junctions map to intron-containing genes, though only slightly more than half of genes in *Plasmodium falciparum* contain introns. This bias is significant, with a binomial probability of $\sim 3e^{-24}$. One explanation for this trend is that introns in antisense transcripts tend to overlap introns in sense transcripts. Indeed, 54% of bases encompassed in the intronic portions of antisense junctions are annotated as gene model introns, while only 26% of the bases within intron-containing genes are

annotated as intronic and only 10% of bases are annotated as intronic in the transcriptome as a whole. This number increases to 65% when only antisense junctions with both inner edges mapping within the same gene model are considered (161 of 200 total antisense junctions). Overlap with sense introns is highly statistically significant (p-value of Chi squared test < 0.001) when compared to random placement of antisense junctions within the genes they map to. Thus antisense introns appear to not only overlap intron-containing genes significantly more often than expected, but also overlap the intron portions of sense genes significantly more than expected.

The occurrence of novel junctions connecting neighboring gene models on opposite strands in the dataset (Figure 1B) suggested that previously unrecognized overlap between annotated genes could explain at least some antisense junctions. 23 antisense junctions could be attributed to overlap between 15 pairs of annotated genes based on junctions and ESTs that link neighboring gene models on opposite strands (Supplementary Table 5). Only 1 gene pair (PFE1425c/PFE1420w) is annotated as overlapping, while 9 have prior EST evidence of overlap. The remaining 5 pairs have no prior evidence of transcript overlap. Twelve pairs appear to be arranged in a tail-to-tail (overlapping 3' ends) fashion, while 3 are arranged in a head-to-head (overlapping 5' ends) fashion. A few studies have reported a similar bias toward tail-to-tail overlaps in mammalian genomes [59], while others refute this assertion [35].

Overlap between annotated genes, however, could not explain all antisense junctions seen 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 noncoding 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 occur in an ORF greater than 100 amino acids long. For comparison, the average exon size in *P. falciparum* intron-containing genes is 184 amino acids. It is possible that these antisense junctions connect shorter than average exons, or that they occur in UTR regions of unannotated genes. It is also possible that many of them belong to noncoding transcripts. Further elucidation of the structure of these antisense transcripts is necessary to determine if their primary function is to encode proteins or not.

Discussion

Using our own RNA-Seq data, as well as data from Otto et al. and Sorber et al. [6, 39], we investigated splicing transcriptome-wide during the blood-stage development of the malaria parasite *Plasmodium falciparum*. In the past, mRNA and EST data have been used for such analysis. However, no mRNA data exists for *P. falciparum* and while the several EST libraries published from the parasite provided useful supporting information during our analysis, many gene models are still not supported by ESTs while those that are are often incompletely covered. RNA-Seq provides the advantage of capturing an entire transcriptome at once in millions of short reads. However, one disadvantage of

RNA-Seq is that a single transcript is not captured from beginning to end, as the sequencing reads are short. Despite this limitation, the orders-of-magnitude increase in throughput over EST libraries expanded the repertoire of splice junctions known in the parasite by more than 10% in the present study.

The ability to accurately map junction reads within the RNA-Seq dataset proved crucial to our analysis. For this purpose, we used the splice site detection algorithm HMMSplicer (see Chapter 3). The algorithm has two main strengths – first, junctions are found by direct gapped alignment of reads to the reference genome with no reliance on gene models or ungapped read coverage. Second, the algorithm does not filter its output junction set – rather, it scores each splice junction on the strength of its supporting data and allows the user to determine the appropriate set of results for further analysis. We chose to set a score threshold, which in essence defines an acceptable false positive rate within the chosen junction set (see Chapter 3). Because a low false positive rate was desired for accurate characterization of splicing in *P. falciparum*, our threshold was determined based on representation of already known canonical splice junctions. However, setting a threshold held the disadvantage of excluding some known junctions, and therefore excluding some true novel junctions as well. Indeed, 13 novel junctions below 1075 were experimentally validated out of 20 attempted (Table 2). However, the validity of splice junctions below 1075 could not be relied on without such verification and thus they were excluded from downstream analysis.

Unfortunately, the score threshold alone proved too simplistic a filter for noncanonical junctions, as these were plagued by false positive splice junctions created by reads with library preparation or sequencing errors. Given that the expected number of true noncanonical versus canonical junctions was low and that the probability of an erroneous read mapping to noncanonical versus canonical boundaries is quite high, a high false positive rate in the noncanonical set was not surprising. In fact the threshold did filter the set quite well – in the dataset as a whole, noncanonical junctions outnumber canonical junctions, while above the threshold, canonical junctions outnumber noncanonical by more than 8 to 1. The remaining high scoring false positive junctions likely arise from errors created during library preparation that are not reflected by poor sequence quality and thus were not penalized during HMMSplicer’s scoring step. Surprisingly, some of the highest scoring noncanonical false positive junctions were supported by multiple reads from both analyzed timecourses, indicating reproducible errors. Regardless, additional filters along with manual inspection were able to recover 12 credible 5’ GC-AG 3’ junctions within the noncanonical junctions list, 5 of which were novel. It is unclear how complete this list is given that percent of splice junctions mapping to 5’ GC-AG 3’ splice sites remains several fold lower in *P. falciparum* than in other organisms [50].

Pair-wise comparison between junctions uncovered alternative splicing in 254 genes (8.7% of *P. falciparum* genes with introns), as well as in intergenic regions that could not be assigned to a particular gene. This analysis also uncovered the presence of splice junctions antisense to gene models. These junctions likely derive from overlapping annotated genes and unannotated antisense transcripts, and themselves are sometimes

alternatively spliced. Disregarding gene models during discovery of alternative splicing afforded the analysis several advantages. Ambiguous cases where a junction distinct from that in the gene model was found, yet the gene model junction was never recovered were not considered alternative splicing. Although these instances could actually represent gene model errors, making it inappropriate to classify them without additional data. Conversely, areas of the transcriptome with no gene model, but for which multiple junctions that could not possibly exist within the same transcript were found are clear examples of alternative splicing that would be missed by reliance on a gene model reference. In particular, the presence of alternative splicing in antisense transcripts would have been missed.

One limitation of our analysis is the inability to detect intron retention in the *P. falciparum* RNA-Seq data. Fundamentally, identifying retained introns presents a different problem from identifying other types of alternative splicing. Rather than correctly aligning splice junctions, unusually high intron coverage must be detected. Therefore, unless the intron-retained isoform comprises a significant proportion of the total transcript, its signal is difficult to distinguish from the typical low-level coverage associated with most transcribed introns. The problem is compounded in *Plasmodium falciparum* by the fact that intron sequence is extremely A/T-rich (~90%) and thus much less unique than exon sequence [10]. Thus both alignment difficulties and coverage noise prevented analysis of intron retention in this dataset.

Elucidation of alternative splicing events is just the first step in fully understanding a complex transcriptome. As we have shown, some of the alternative splicing events captured here appear to leave the coding frame intact, albeit with added or subtracted codons. The functional implications of these potential protein isoforms must be further examined on a case-by-case basis. Likewise, events that disrupt the coding frame of a transcript could produce functional protein isoforms (truncated in most cases) or could represent intermediates in gene regulation coupling alternative splicing to nonsense-mediated decay. Genetic disruption of the surveillance complex could determine which if any of these isoforms are normally subjected to NMD.

Our analysis uncovered not only constitutive and alternative splicing in *P. falciparum*, but also complex transcriptional arrangements in the parasite. Antisense junctions in the data indicate overlap between annotated sense genes and antisense transcripts, some of which appear to be extensions of neighboring annotated genes, while most likely arise from unannotated genes. For unknown reasons, antisense splice junctions tend to encompass sense introns more than would be expected by chance. It is unknown if this is a general property of antisense junctions or if this phenomenon is specific to *P.*

falciparum antisense splice junctions. Perhaps antisense introns must be spliced out in approximately the same area as sense introns so that the transcripts can physically interact with one another. Another explanation is that the low complexity sequence that comprises *P. falciparum* introns is not useful to a transcript transcribed from either strand and thus must be removed. Alternatively, the degenerate nature of splice signals in *P. falciparum* introns could cause some areas to be recognized in both directions by the

splicing machinery. Further inquiry is necessary to determine which of these hypotheses holds true.

References

1. WHO | World Malaria Report 2009. Available at: http://www.who.int/malaria/world_malaria_report_2009/en/index.html. Accessed 14 June 2010.
2. Cox-Singh J, Davis TME, Lee K, Shamsul SSG, Matusop A, et al. (2008) *Plasmodium knowlesi* Malaria in Humans Is Widely Distributed and Potentially Life Threatening. *Clinical Infectious Diseases* 46: 165-171.
3. Bozdech Z, Llinás M, Pulliam BL, Wong ED, Zhu J, et al. (2003) The Transcriptome of the Intraerythrocytic Developmental Cycle of *Plasmodium falciparum*. *PLoS Biol* 1: e5.
4. Le Roch KG, Zhou Y, Blair PL, Grainger M, Moch JK, et al. (2003) Discovery of Gene Function by Expression Profiling of the Malaria Parasite Life Cycle. *Science* 301: 1503-1508.
5. Silvestrini F, Bozdech Z, Lanfrancotti A, Di Giulio E, Bultrini E, et al. (2005) Genome-wide identification of genes upregulated at the onset of gametocytogenesis in *Plasmodium falciparum*. *Mol. Biochem. Parasitol* 143: 100-110.
6. Otto TD, Wilinski D, Assefa S, Keane TM, Sarry LR, et al. (2010) New insights into the blood-stage transcriptome of *Plasmodium falciparum* using RNA-Seq. *Molecular Microbiology* 76: 12-24.
7. Watanabe J, Wakaguri H, Sasaki M, Suzuki Y, Sugano S (2007) Comparasite: a database for comparative study of transcriptomes of parasites defined by full-length cDNAs. *Nucl. Acids Res.* 35: D431-438.
8. Florent I, Porcel B, Guillaume E, Da Silva C, Artiguenave F, et al. (2009) A *Plasmodium falciparum* FcB1-schizont-EST collection providing clues to schizont specific gene structure and polymorphism. *BMC Genomics* 10: 235.
9. Lu F, Jiang H, Ding J, Mu J, Valenzuela J, et al. (2007) cDNA sequences reveal considerable gene prediction inaccuracy in the *Plasmodium falciparum* genome. *BMC Genomics* 8: 255.
10. Gardner MJ, Hall N, Fung E, White O, Berriman M, et al. (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* 419: 498-511.

11. Upadhyay R, Bawankar P, Malhotra D, Patankar S (2005) A screen for conserved sequences with biased base composition identifies noncoding RNAs in the A-T rich genome of *Plasmodium falciparum*. *Molecular and Biochemical Parasitology* 144: 149-158.
12. Chakrabarti K, Pearson M, Grate L, Sterne-Weiler T, Deans J, et al. (2007) Structural RNAs of known and unknown function identified in malaria parasites by comparative genomics and RNA analysis. *RNA* 13: 1923-1939.
13. Moreno-Hagelsieb G, Latimer K (2008) Choosing BLAST options for better detection of orthologs as reciprocal best hits. *Bioinformatics* 24: 319-324.
14. Kaufer NF, Potashkin J (2000) SURVEY AND SUMMARY: Analysis of the splicing machinery in fission yeast: a comparison with budding yeast and mammals. *Nucl. Acids Res.* 28: 3003-3010.
15. Stevens SW, Barta I, Ge HY, Moore RE, Young MK, et al. (2001) Biochemical and genetic analyses of the U5, U6, and U4/U6 x U5 small nuclear ribonucleoproteins from *Saccharomyces cerevisiae*. *RNA* 7: 1543-1553.
16. Bessonov S, Anokhina M, Will CL, Urlaub H, Luhrmann R (2008) Isolation of an active step I spliceosome and composition of its RNP core. *Nature* 452: 846-850.
17. Chan S, Cheng S (2005) The Prp19-associated Complex Is Required for Specifying Interactions of U5 and U6 with Pre-mRNA during Spliceosome Activation. *Journal of Biological Chemistry* 280:: 31190 -31199.
18. Lardelli RM, Thompson JX, Yates JR, Stevens SW (2010) Release of SF3 from the intron branchpoint activates the first step of pre-mRNA splicing. *RNA* 16: 516-528.
19. Lamond AI (1993) The spliceosome. *Bioessays* 15: 595-603.
20. Clark F, Thanaraj TA (2002) Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human. *Hum. Mol. Genet.* 11: 451-464.
21. Li L, Brunk BP, Kissinger JC, Pape D, Tang K, et al. (2003) Gene Discovery in the Apicomplexa as Revealed by EST Sequencing and Assembly of a Comparative Gene Database. *Genome Research* 13: 443-454.
22. Burtis KC (1993) The regulation of sex determination and sexually dimorphic differentiation in *Drosophila*. *Current Opinion in Cell Biology* 5: 1006-1014.
23. Madsen J, Stoltzfus CM (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.

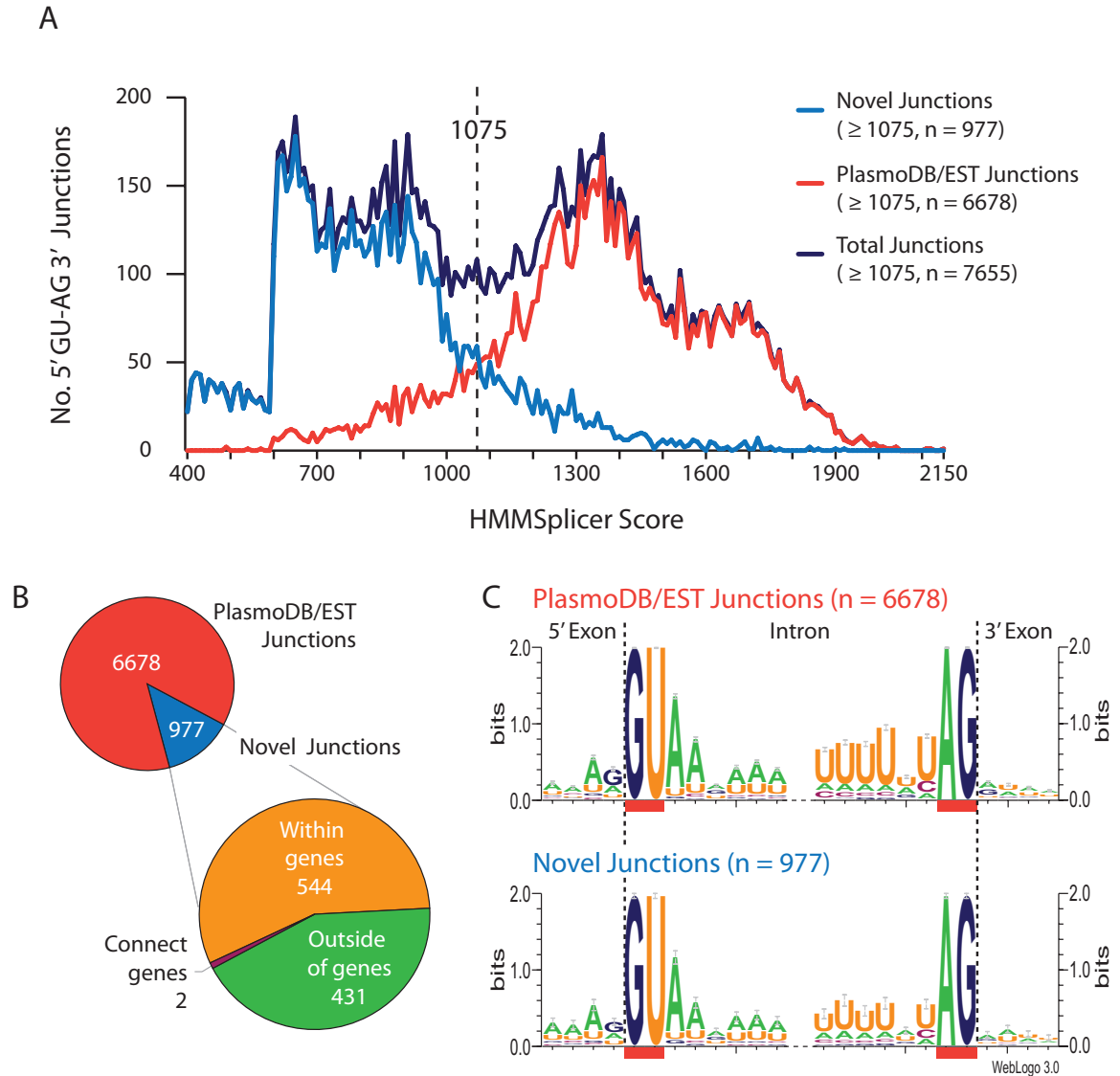
24. Knapp B, Nau U, Hundt E, Küpper HA (1991) Demonstration of alternative splicing of a pre-mRNA expressed in the blood stage form of *Plasmodium falciparum*. *J. Biol. Chem* 266: 7148-7154.
25. Bracchi-Ricard V, BARIK S, DELVECCHIO C, DOERIG C, CHAKRABARTI R, et al. (2000) PfPK6, a novel cyclin-dependent kinase/mitogen-activated protein kinase-related protein kinase from *Plasmodium falciparum*. Available at: <http://www.biochemj.org/bj/347/0255/bj3470255.htm>. Accessed 14 May 2010.
26. Muhia DK, Swales CA, Eckstein-Ludwig U, Saran S, Polley SD, et al. (2003) Multiple Splice Variants Encode a Novel Adenylyl Cyclase of Possible Plastid Origin Expressed in the Sexual Stage of the Malaria Parasite *Plasmodium falciparum*. *Journal of Biological Chemistry* 278: 22014-22022.
27. Saenz FE, Balu B, Smith J, Mendonca SR, Adams JH (2008) The Transmembrane Isoform of *Plasmodium falciparum* MAEBL Is Essential for the Invasion of *Anopheles* Salivary Glands. *PLoS ONE* 3: e2287.
28. Wentzinger L, Bopp S, Tenor H, Klar J, Brun R, et al. (2008) Cyclic nucleotide-specific phosphodiesterases of *Plasmodium falciparum*: PfPDE[alpha], a non-essential cGMP-specific PDE that is an integral membrane protein. *International Journal for Parasitology* 38: 1625-1637.
29. Iriko H, Jin L, Kaneko O, Takeo S, Han E, et al. (2009) A small-scale systematic analysis of alternative splicing in *Plasmodium falciparum*. *Parasitology International* 58: 196-199.
30. Licatalosi DD, Darnell RB (2010) RNA processing and its regulation: global insights into biological networks. *Nat Rev Genet* 11: 75-87.
31. Venables JP, Koh C, Froehlich U, Lapointe E, Couture S, et al. (2008) Multiple and Specific mRNA Processing Targets for the Major Human hnRNP Proteins. *Mol. Cell. Biol.* 28: 6033-6043.
32. Long J, Caceres J (2009) The SR protein family of splicing factors: master regulators of gene expression. *Biochem. J.* 417: 15.
33. Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, et al. (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucl. Acids Res.* 29: 37-40.
34. Jen C, Michalopoulos I, Westhead D, 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 Biology* 6: R51.

35. 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, et al. (2005) Antisense Transcription in the Mammalian Transcriptome. *Science* 309: 1564-1566.
36. Faghihi MA, Wahlestedt C (2009) Regulatory roles of natural antisense transcripts. *Nat Rev Mol Cell Biol* 10: 637-643.
37. Gunasekera AM, Patankar S, Schug J, Eisen G, Kissinger J, et al. (2004) Widespread distribution of antisense transcripts in the *Plasmodium falciparum* genome. *Molecular and Biochemical Parasitology* 136: 35-42.
38. Raabe CA, Sanchez CP, Randau G, Robeck T, Skryabin BV, et al. (2010) A global view of the nonprotein-coding transcriptome in *Plasmodium falciparum*. *Nucl. Acids Res.* 38: 608-617.
39. Sorber K, Chiu C, Webster D, Dimon M, Ruby JG, et al. (2008) The Long March: A Sample Preparation Technique that Enhances Contig Length and Coverage by High-Throughput Short-Read Sequencing. *PLoS ONE* 3: e3495.
40. Golas MM, Sander B, Will CL, Luhrmann R, Stark H (2003) Molecular Architecture of the Multiprotein Splicing Factor SF3b. *Science* 300: 980-984.
41. Tanackovic G, Kramer A (2005) Human Splicing Factor SF3a, but Not SF1, Is Essential for Pre-mRNA Splicing In Vivo. *Mol. Biol. Cell* 16: 1366-1377.
42. WILL CL, SCHNEIDER C, HOSSBACH M, URLAUB H, RAUHUT R, et al. (2004) The human 18S U11/U12 snRNP contains a set of novel proteins not found in the U2-dependent spliceosome. *RNA* 10: 929-941.
43. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology* 215: 403-410.
44. Boucher L, Ouzounis CA, Enright AJ, Blencowe BJ (2001) A genome-wide survey of RS domain proteins. *RNA* 7: 1693-1701.
45. TPGDC (2001) PlasmoDB: An integrative database of the *Plasmodium falciparum* genome. Tools for accessing and analyzing finished and unfinished sequence data. *Nucl. Acids Res.* 29: 66-69.
46. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10: R25.
47. Kent WJ (2002) BLAT—The BLAST-Like Alignment Tool. *Genome Research* 12: 656-664.

48. Weber JL (1988) Molecular biology of malaria parasites. *Experimental Parasitology* 66: 143-170.
49. Bailey TL, Elkan C (1995) Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning* 21: 51-80.
50. Burset M, Seledtsov IA, Solovyev VV (2000) Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucl. Acids Res.* 28: 4364-4375.
51. Kitamura-Abe S, Itoh H, Washio T, Tsutsumi A, Tomita M (2004) Characterization of the splice sites in GT-AG and GC-AG introns in higher eukaryotes using full-length cDNAs. *J Bioinform Comput Biol* 2: 309-331.
52. Fouser LA, Friesen JD (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.
53. Tarn W, Steitz JA (1996) A Novel Spliceosome Containing U11, U12, and U5 snRNPs Excises a Minor Class (AT-AC) Intron In Vitro. *Cell* 84: 801-811.
54. Lopez MD, Alm Rosenblad M, Samuelsson T (2008) Computational screen for spliceosomal RNA genes aids in defining the phylogenetic distribution of major and minor spliceosomal components. *Nucl. Acids Res.* 36: 3001-3010.
55. Zhang Y, Liu XS, Liu Q, Wei L (2006) Genome-wide in silico identification and analysis of cis natural antisense transcripts (cis-NATs) in ten species. *Nucl. Acids Res.* 34: 3465-3475.
56. Barash Y, Calarco JA, Gao W, Pan Q, Wang X, et al. (2010) Deciphering the splicing code. *Nature* 465: 53-59.
57. Sun S, Zhang Z, Sinha R, Karni R, Krainer AR (2010) SF2/ASF autoregulation involves multiple layers of post-transcriptional and translational control. *Nat Struct Mol Biol* 17: 306-312.
58. Conti E, Izaurralde E (2005) Nonsense-mediated mRNA decay: molecular insights and mechanistic variations across species. *Current Opinion in Cell Biology* 17: 316-325.
59. Sun M, Hurst LD, Carmichael GG, Chen J (2005) Evidence for a preferential targeting of 3'-UTRs by cis-encoded natural antisense transcripts. *Nucl. Acids Res.* 33: 5533-5543.
60. Stein LD, Mungall C, Shu S, Caudy M, Mangone M, et al. (2002) The Generic Genome Browser: A Building Block for a Model Organism System Database. *Genome Research* 12: 1599-1610.

Figures

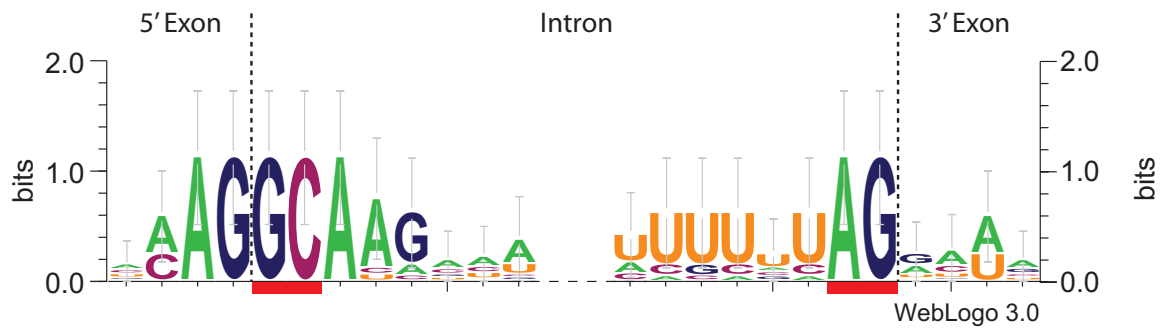
Figure 1. Canonical junctions found by HMMSplicer



A) Histogram of canonical (5' GU-AG 3') junctions found by HMMSplicer in the combined RNA-Seq data binned by score. The dark blue line plots all junctions reported by HMMSplicer, while the red line charts HMMSplicer junctions that match previously known junctions in PlasmDB v6.3 gene models or in ESTs. The light blue line charts

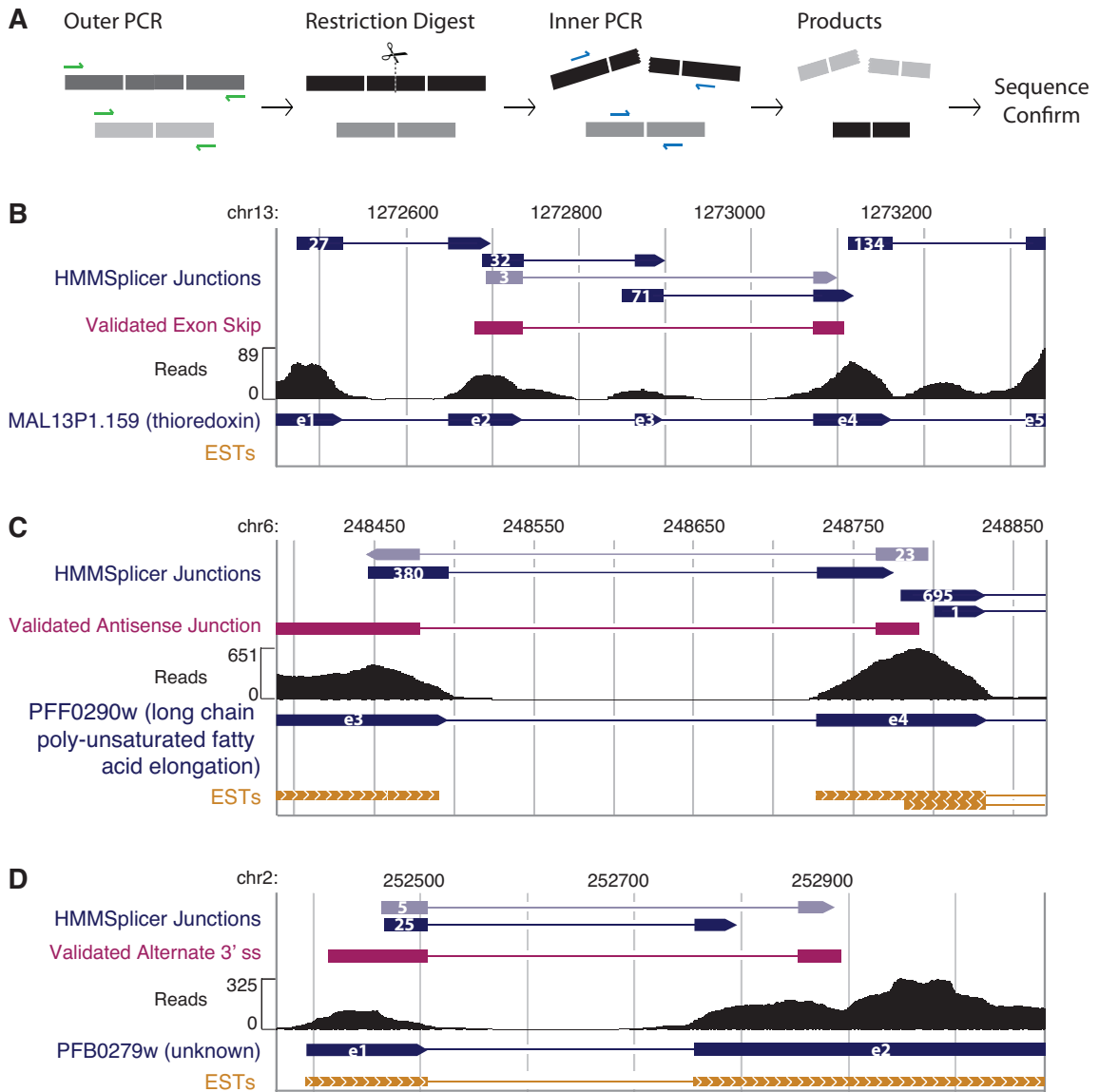
novel HMMSplicer junctions. The dashed line drawn at 1075 represents the score threshold. B) Breakdown of canonical junctions above 1075, with additional classification of novel junctions. “Outside of genes” refers to novel junctions where at least one side did not map within a gene model. “Within genes” indicates that both sides mapped to the same gene model. “Connect genes” indicates that the sides mapped to neighboring gene models. C) Comparison of the 5’ and 3’ splice site MEME motifs for previously known junctions versus novel junctions above 1075. Red bars indicate the 5’ GU-AG 3’ boundaries used for inclusion in the set. The height of each letter indicates the preference strength for that nucleotide at each position.

Figure 2. Splice site motifs for GC-AG junctions



MEME 5' and 3' splice site motifs for manually curated 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 preference strength for that nucleotide at each position. The large error bars derive from the small size of the input set.

Figure 3. Validation of conflicting junctions



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, novel isoform, which is then sequence confirmed. Gbrowse 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) [60]. All HMMSplicer junctions scoring higher than 980 are shown as

either dark blue bars (known junctions) or light blue bars (novel 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.

Tables

Table 1. Putative *Plasmodium falciparum* splicing factor orthologs

Complex	Human / Yeast	<i>Pf</i> Ortholog	Complex	Human / Yeast	<i>Pf</i> Ortholog
snRNP core	SNRPB / SMB1	PF14_0146	U4/U6	PRPF3 / PRP3	MAL13P1.45
(stability and function of U1, U2, U4, and U5 snRNPs)	SNRPD1 / SMD1	PF11_0266	(catalytic activation of spliceosome)	NHP2L1 / SNU13	PF11_0250
	SNRPD2 / SMD2	PFB0865w		PRPF4 / PRP4	MAL13P1.385 [†]
	SNRPD3 / SMD3	PFI0475w		PRPF31 / <i>PRP31</i>	PFD0450c
	SNRPE / SME1	MAL13P1.253		PPIH / -	PF08_0121*
	SNRPF / SMX3	PF11_0280	tri-snRNP	SART1 / SNU66	PFC1060c [†]
	SNRPG / SMX2	MAL8P1.48	(activation of spliceosome)	USP39 / SAD1	PF13_0096 [†]
U6 core	LSM2 / LSM2	PFE1020w		SNRNP27 / -	MAL8P1.71 ^{†*}
(stability and function of U6 snRNP)	LSM3 / LSM3	PF08_0049	hPrp19/CDC5	PRPF19 / PRP19	PFC0365w
	LSM4 / LSM4	PF11_0524	(specification of U5 and U6 interactions with RNA)	CRNKL1 / CLF1	PFD0180c
	LSM5 / LSM5	PF14_0411		CDC5L / CEF1	PF10_0327 [†]
	LSM6 / LSM6	PF13_0142*		ISY1 / ISY1	PF14_0688
	LSM7 / LSM7	PFL0460w		BCAS2 / SNT309	PFF0695w ^{†*}
	NAA38 / LSM8	MAL8P1.9*		XAB2 / SYF1	PFL1735c [†]
U1	SNRNP70 / SNP1	MAL13P1.338		PLRG1 / PRP46	PFC0100c [†]
(initial 5' ss recognition)	SNRPA / MUD1	MAL13P1.35*		SYF2 / SYF2	?
	SNRPC / YHC1	PF08_0084		SNW1 / <i>PRP45</i>	PFB0875c [†]
U2	SNRPA1 / LEA1	PF13_0362		BUD31 / <i>BUD31</i>	PFE1140c
(BP detection)	SNRPB2 / MSL1	PFI1695c		PPIE / -	?
U2-related	U2AF1 / -	PF11_0200*		CCDC12 / -	PF14_0490 [†]
(BP & poly-Y recognition)	U2AF2 / MUD2	PF14_0656*		AQR / -	PF13_0273 ^{†*}
	SF1 / MSL5	PFF1135w		CWC15 / <i>CWC15</i>	PF07_0091*
SF3a	SF3A1 / PRP21	PF14_0713 [†]		PPIL1 / -	PFE1430c*
(stability of U2-BP interaction)	SF3A2 / PRP11	PFF0970w	Non-snRNP factors	DHX16 / PRP2	?
	SF3A3 / PRP9	PFI1215w		BAT1 / SUB2	PFB0445c
SF3b	SF3B1 / HSH155	PFC0375c	(second step factors)	DDX46 / PRP5	PFE0430w [†]
(stability of U2-BP interaction)	SF3B2 / CUS1	PF14_0587		SLU7 / SLU7	PFF0500c
	SF3B3 / RSE1	PFL1680w		DHX38 / PRP16	MAL13P1.322
	SF3B4 / HSH49	PF14_0194		CDC40 / CDC40	PFL0970w
	SF3B5 / YSF3	PF13_0296		PRPF18 / <i>PRP18</i>	PFI1115c
	PHF5A / RDS3	PF10_0179a	(RNA release)	DHX8 / PRP22	PF10_0294 [†]
	SF3B14 / -	PFL1200c*	NMD	UPF1 / NAM7	PF10_0057
U5	DDX23 / PRP28	PFE0925c	(detection of nonsense transcripts)	UPF2 / NMD2	PFI1265w [†]
(catalytic activation of spliceosome)	CD2BP2 / LIN1	PF10_0310 [†]		UPF3A / UPF3	?
	EFTUD2 / SNU114	PF10_0041*		UPF3B / -	PF13_0158*
	SNRNP200 / BRR2	PFD1060w	SR & hnRNP	SRSF1 / -	PFE0865c*
	TXNL4A / DIB1	PFL1520w		SRSF12 / -	PFE0160c*
	PRPF8 / PRP8	PFD0265w		PTBP2 / -	PFF0320c*
	PRPF6 / <i>PRP6</i>	PF11_0108		SFRS4 / -	PF10_0217*
	SNRNP40 / -	MAL8P1.43*		TRA2B / -	PF10_0028 ^{†*}

Putative *Plasmodium falciparum* splicing and nonsense-mediated decay factor orthologs identified by reciprocal best hits analysis with human or *Saccharomyces cerevisiae* sequences. The human or *S. cerevisiae* factor in large font represents the best match for the *P. falciparum* ortholog. Spliceosomal and NMD factors not found are in red and are denoted with question marks, while SR and hnRNP factors not found are not shown. *S. cerevisiae* orthologs that do not reside in the same complex as their human counterparts are italicized. * Orthologs identified only by the human sequence. † *P. falciparum* proteins described in PlasmoDB as “conserved *Plasmodium* protein” or with descriptions that do not reflect involvement in splicing.

Table 2. Verification of novel junctions in conflict with known junctions.

Gene Name	PlasmoDBv6.3 Description	Score	Validated	Type	Frame -shift?	Isoform Difference	R	T	LT /ES	S
PFL1810w	conserved Plasmodium protein	1544.2	Y	5'ss	N	132bp (44aa)	11	8	8	5
		1283.2	Y	5'ss	N	219bp (73aa)	3	0	2	1
PFE0390w	conserved Plasmodium protein	1422.1	Y	5'ss	N	66bp (22aa)	13	10	6	2
PF13_0138	MSF-1 like protein	1372	Y	5'ss	Y	56bp	7	5	2	6
PFI0400c	conserved Plasmodium membrane protein	1369.8	Y	exon skip	N	126bp (42aa)	2	1	14	1
PFF0290w	long chain polyunsaturated fatty acid elongation enzyme	1291.8	Y	antisense	-	N/A	9	6	10	5
MAL13P1.225	thioredoxin	1277.3	Y	exon skip	Y	34bp	2	6	0	0
PFE0055c	heat shock protein	1275.4	Y	5'ss	Y	37bp	54	14	2	7
MAL8P1.126	serine protease	1257.4	Y	5'ss	Y	110bp	1	1	8	7
PF10_0025	PF70 protein	1256.8	Y	5'ss	N	75bp (25aa)	18	2	0	0
PFD1050w	alpha-tubulin II	1243.2	-	antisense	-	N/A	1	0	4	1
MAL13P1.159*	thioredoxin cleavage and polyadenylation specific factor	1239.9	Y	exon skip	N	33bp (11aa)	0	1	0	2
PFC0780w	RNA binding protein	1231.7	-	antisense	-	N/A	2	6	23	7
PFD0775c	RNA binding protein	1228.4	Y	antisense	-	N/A	1	7	6	0
PF10_0194	NOP12-like protein	1219.4	Y	exon skip	Y	41bp	1	0	0	1
PFL1440c	conserved Plasmodium protein	1217.6	Y	exon skip	N	57bp (19aa)	0	2	0	1
PF11_0291	conserved Plasmodium protein	1203.5	Y	5'ss	Y	40bp	1	0	0	5
PFC0360w	activator of HSP90 ATPase homolog 1-like protein	1200.5	Y	exon skip	Y	223bp	1	1	2	0
PFC0495w	plasmepsin VI	1192.6	Y	antisense	-	N/A	0	0	6	2
PF14_0394	conserved Plasmodium protein	1190	Y	5'ss	N	99bp (33aa)	2	4	4	0
MAL13P1.146	AMP deaminase	1189.3	Y	antisense	-	N/A	1	1	0	0
PF11_0379	conserved Plasmodium protein	1050.5	Y	exon skip	N	60bp (20aa)	1	0	0	1
PFL1445w	conserved Plasmodium protein	1041.3	Y	exon skip	Y	85bp	0	7	0	0
MAL13P1.16	SNARE protein	1034.7	Y	exon skip	N	108bp (36aa)	0	0	0	4
MAL13P1.277	DNAJ-like protein	1034.2	Y	exon skip	Y	146bp	2	0	2	0
PFF1210w	phosphatidic acid phosphatase	1032.4	Y	5'ss	Y	67bp	2	0	2	5
PFB0600c	conserved Plasmodium protein	1026.1	Y	antisense	-	N/A	1	1	2	0
PF14_0128	ubiquitin conjugating enzyme	1018.5	Y	exon skip	Y	103bp	0	1	2	0
PF14_0316	DNA topoisomerase II	1011.4	-	5'ss	Y	460bp	0	0	2	1
PFB0279w*	conserved Plasmodium protein	1010.9	Y	3'ss	Y	98bp	1	4	2	0
PFL1465c	heat shock protein hslv	1004.5	-	exon skip	Y	39bp (13aa)	0	2	0	4
PF10_0372	antigen UB05	1004.4	-	antisense	-	N/A	0	1	0	1
PF11_0182	conserved Plasmodium protein	1004.1	Y	exon skip	Y	56bp	0	0	4	0
PFF0365c	G-protein associated signal transduction protein	996.3	-	exon skip	N	162bp (54aa)	2	0	0	0
PFB0445c	DEAD box helicase, UAP56	995.9	-	3'ss	N	75bp (25aa)	1	0	0	0
PFD0895c	Bet3 transport protein	991	-	antisense	-	N/A	0	0	2	0
PF10_0116	conserved Plasmodium protein	989.9	Y	5'ss	N	75bp (25aa)	0	0	2	0
PF14_0604	conserved Plasmodium protein	988.1	Y	exon skip	Y	343bp	1	0	0	0
PFI0560c	conserved Plasmodium protein	987.7	Y	exon skip	Y	40bp	1	0	0	2
PFB0550w	peptide chain release factor subunit 1	985.5	-	exon skip	Y	155bp	0	0	2	0
PF11_0355	conserved Plasmodium protein	984.6	Y	antisense	-	N/A	1	1	0	0

Conflicts are ranked by score. * Indicates validations shown in more detail in Figures 3.

For all conflict types except antisense, the novel junction was evaluated for maintenance of open reading frame - nucleotide and amino acid (if applicable) differences between novel and known isoforms are listed. Read counts for novel junctions (normalized by the number of reads mapped full-length 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.


Table 3. Alternative splicing events by type and area in the genome

	In genes	Intergenic	Antisense	Total
5' splice site	142	39	15	196
3' splice site	114	23	8	145
5' and 3' splice site	3	4	2	9
Skipped exon	51	3	2	56
Total	310	69	27	406

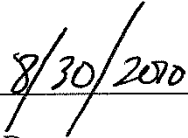
Publishing Agreement

It is the policy of the University to encourage the distribution of all theses and dissertations. Copies of all UCSF theses and dissertations will be routed to the library via the Graduate Division. The library will make all theses and dissertations accessible to the public and will preserve these to the best of their abilities, in perpetuity.

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis or dissertation to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.



Author Signature



Date