

UCSF

UC San Francisco Electronic Theses and Dissertations

Title

Nucleic Acid Methods in Yeast and Infectious Disease

Permalink

<https://escholarship.org/uc/item/41m6k08w>

Author

Sorber, Katherine

Publication Date

2010

Peer reviewed|Thesis/dissertation

Nucleic Acid Methods in Yeast and Infectious Disease

by

Katherine Sorber

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biochemistry

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Acknowledgements

The work compiled here would not have been possible without the mentorship of many people including, Dr. Joe DeRisi, Dr. Christine Guthrie, Dr. Sandy Johnson, Dr. Ashwini Jambekar, Dr. Charles Chiu, Dr. Amy Kistler, and many others. Past and present members of the DeRisi lab were instrumental in both experimental advice and in creating an environment of collegiality, intellectual curiosity, and rigor. My closest collaborator, Dr. Michelle Dimon, made many aspects of this work possible with her computational prowess and methodical approach to bioinformatic problems.

I would also like to thank people outside of the lab who helped keep graduate school in healthy perspective for me. My fellow graduate students in the Tetrad program, particularly my classmates, helped me blow off steam when necessary. My family, including my mom, my dad, my brother, and my grandmother, were infinitely supportive and always only a phone call away. Finally, my fiancé Erik Lontok, who was also a classmate, provided balance, laughter, and delicious food. Salamat, mahal ko!

Nucleic Acid Methods in Yeast and Infectious Disease

Katherine Sorber

Abstract:

This graduate work consists of projects concerned with functions of RNA, use of high throughput sequencing technology, and the molecular biology of the malaria parasite *Plasmodium falciparum*. The intricacies of recognition of mRNA molecules by yeast transport machinery were explored in an unbiased manner, leading to identification of both primary sequence and secondary structural motifs. High throughput sequencing was used to recover a new type of bornavirus associated with proventricular dilatation disease in psittacine birds. High throughput sequencing was again used to find leads as to the genetic determinant of artemisinin resistance in *Plasmodium falciparum*, finding an amplification on chromosome 10 as the most promising lead. Parasite response to artemisinin was also studied at the transcriptional level using microarrays to document the entry into and recovery from the ring-like dormant state induced in *Plasmodium falciparum* by artemisinin. Finally, splicing in *Plasmodium falciparum* was investigated by sequencing mRNA from four timepoints in the blood stages of the parasite. Specific software was written to extract all exon-exon junctions within that dataset and higher order analysis revealed previously unknown splice sites, hundreds of alternative splicing events, and the presence of spliced antisense RNAs in the transcriptome.

Table of Contents

| | |
|--|------|
| Title Page | i |
| Acknowledgements..... | iii |
| Abstract..... | iv |
| Table of Contents..... | v |
| List of Tables | vi |
| List of Figures..... | viii |
| Chapter 1: Introduction..... | 1 |
| Chapter 2: Unbiased Selection of Localization Elements Reveals cis-acting Determinants of mRNA Bud-Localization in <i>Saccharomyces cerevisiae</i> | 18 |
| Chapter 3: Recovery of Divergent Avian Bornaviruses from Cases of Proventricular Dilatation Disease: Identification of a Candidate Etiologic Agent..... | 62 |
| Chapter 4: Determination of genetic correlates of resistance to artemisinin in <i>Plasmodium falciparum</i> | 111 |
| Chapter 5: Dihydroartemisinin induces transcriptome arrest in drug susceptible and resistant <i>Plasmodium falciparum</i> | 135 |
| Chapter 6: The long march: a sample preparation technique that enhances contig length and coverage by high-throughput short-read sequencing..... | 180 |
| Chapter 7: HMMSplicer: a tool for efficient and sensitive discovery of known and novel splice junctions in RNA-Seq data..... | 210 |
| Chapter 8: RNA-Seq Analysis of Splicing in <i>Plasmodium falciparum</i> Uncovers New Splice Junctions, Alternative Splicing, and Splicing of Antisense Transcripts..... | 266 |

List of Tables

Chapter 2:

| | |
|---------------|----|
| Table 1 | 38 |
| Table 2 | 39 |
| Table 3 | 41 |
| Table 4 | 42 |
| Table 5 | 43 |
| Table 6 | 44 |
| Table 7 | 45 |
| Table 8 | 46 |

Chapter 3:

| | |
|---------------|----|
| Table 1 | 93 |
| Table 2 | 94 |
| Table 3 | 95 |
| Table 4 | 96 |

Chapter 4:

| | |
|---------------|-----|
| Table 1 | 123 |
| Table 2 | 125 |
| Table 3 | 126 |
| Table 4 | 127 |
| Table 5 | 128 |

Chapter 5:

| | |
|---------------|-----|
| Table 1 | 157 |
|---------------|-----|

| | |
|----------------|-----|
| Table 2 | 162 |
| Table 3 | 166 |
| Chapter 6: | |
| Table 1 | 199 |
| Chapter 7: | |
| Table 1 | 248 |
| Table 2 | 249 |
| Table 3 | 250 |
| Chapter 8: | |
| Table 1 | 300 |
| Table 2 | 302 |
| Table 3 | 303 |
| Table 4 | 304 |
| Table 5 | 306 |
| Table 6 | 307 |
| Table 7 | 309 |
| Table 8 | 310 |
| Table 9 | 311 |
| Table 10 | 313 |

List of Figures

Chapter 2:

| | |
|-----------------|----|
| Figure 1 | 47 |
| Figure 2 | 48 |
| Figure 3 | 49 |
| Figure 4 | 50 |
| Figure 5 | 51 |
| Figure 6 | 52 |
| Figure 7 | 53 |
| Figure 8 | 54 |
| Figure 9 | 55 |
| Figure 10 | 56 |
| Figure 11 | 57 |

Chapter 3:

| | |
|----------------|-----|
| Figure 1 | 97 |
| Figure 2 | 99 |
| Figure 3 | 101 |
| Figure 4 | 102 |
| Figure 5 | 103 |
| Figure 6 | 104 |
| Figure 7 | 105 |

Chapter 4:

| | |
|----------------|-----|
| Figure 1 | 129 |
|----------------|-----|

| | |
|----------------|-----|
| Figure 2 | 130 |
| Figure 3 | 131 |
| Chapter 5: | |
| Figure 1 | 168 |
| Figure 2 | 169 |
| Figure 3 | 170 |
| Figure 4 | 171 |
| Figure 5 | 172 |
| Figure 6 | 173 |
| Chapter 6: | |
| Figure 1 | 200 |
| Figure 2 | 201 |
| Figure 3 | 203 |
| Figure 4 | 204 |
| Chapter 7: | |
| Figure 1 | 251 |
| Figure 2 | 252 |
| Figure 3 | 254 |
| Figure 4 | 255 |
| Figure 5 | 256 |
| Figure 6 | 257 |
| Figure 7 | 258 |
| Figure 8 | 259 |

Chapter 8:

| | |
|----------------|-----|
| Figure 1 | 314 |
| Figure 2 | 315 |
| Figure 3 | 316 |
| Figure 4 | 317 |
| Figure 5 | 319 |
| Figure 6 | 321 |
| Figure 7 | 322 |
| Figure 8 | 323 |
| Figure 9 | 324 |

Chapter 1: Introduction

As a graduate student in Dr. Joseph DeRisi's laboratory, I have had the opportunity to participate in projects touching on at least one of the following three areas of study – 1) RNA function in the cell, 2) the molecular biology of *Plasmodium falciparum*, or 3) high-throughput sequencing technology. Below I present a brief introduction to these three areas of study, as well as introductions to the four projects that comprise subsequent thesis chapters.

Over the last few decades, the variety and importance of roles played by RNA in the cell has gained considerable appreciation. Once viewed as merely the messenger between DNA and proteins, RNA is now understood to encode significant regulatory information and to perform catalytic processes as well. In fact, the ability of RNA to both catalyze chemical reactions and store genetic information has led to the idea that RNA may have pre-dated both proteins and DNA (1). In studying RNA, primary sequence, secondary structure, and expression patterns of the transcript(s) in question are often important in understanding underlying biology, including cell development and response to environmental conditions.

Plasmodium falciparum is the deadliest human malaria parasite and represents a continued threat to public health because of its widespread resistance to available antimalarial drugs. Understanding the basic biology of this parasite is necessary both for developing new treatments and also for anticipating eradication challenges. Interestingly, *Plasmodia* are not only medically relevant organisms, but also occupy a unique position in the tree of life, having diverged before plants and animals diverged from each other (2). From this point of view, molecular quirks unique to the parasite can not only be used

to combat it, but can also inform how general eukaryotic cellular processes may have evolved.

Researchers today routinely investigate biology on transcriptome- or genome-wide levels as a result of the recent technological shift from standard Sanger sequencing to a multitude of high-throughput sequencing platforms. These platforms provide a snapshot of any given nucleic acid sample millions of reads deep. They offer unbiased coverage and improved sensitivity compared to microarray or ChIP-ChIP analysis (3), and allow for rapid sequencing of entire genomes. However, these new platforms also pose significant protocol and data analysis development challenges as the technologies are still emerging.

Yeast She2p-She3p-Myo4p system as a model for understanding mRNA transport:

The basis of spatially segregated protein expression in a cell often lies in translation repression and transport signals in cognate mRNA. The resulting patterns of protein expression can be important during organismal development, as well as for polarized cell function (4, 5). *Saccharomyces cerevisiae*, often used as a model organism to study RNA-related basic biology, contains the She2p-She3p-Myo4p mRNA transport system, which has been used to understand the principles of mRNA localization in a genetically tractable organism.

The She2p-She3p-Myo4p mRNA transport system was discovered while investigating the mechanism of localization of Ash1p, a cell-fate determining protein that inhibits mating-type switching exclusively in daughter cells (6-8). She2p recognizes *cis* elements in the 3' untranslated region (UTR) and coding region of *ASH1* mRNA,

recruiting it to Myo4p myosin motor protein via the adapter protein She3p. Motion of the myosin motor along polarized actin cables localizes the entire ribonucleoprotein (RNP) complex to the bud tip, ensuring high concentrations of ASH1 mRNA, and thus Ash1p protein, in daughter versus mother cells.

Although the sequence elements necessary for localization of ASH1 mRNA were roughly known at the start of our work, *cis* elements necessary for localization of other mRNAs transported by the She2p-She3p-Myo4p system were ill defined. To determine localization elements in an unbiased manner, nonhomologous random recombination (9) was used to generate a random library of fragments from RNAs known to be transported by the She2p-She3p-Myo4p system, including ASH1. This library was used in a yeast 3-hybrid assay (10), allowing only colonies transcribing RNA fragments with sequence recognized by She2p to survive selection. Results from both the ASH1 control experiment and the subsequent experiments with 10 other localized mRNAs were confirmed using the U1A-GFP system (11, 12) for their ability to directly localize GFP-tagged mRNA. MEME and MFOLD were then used to determine common primary and secondary structure elements within these localized fragments, and specific zipcodes were chosen for further mutational analysis. The results of these studies not only identified a short specific sequence element important for recognition by the transport system, but also revealed the fundamentally complex process of recognition by probing context-dependent signals and structural features acting in combination with the primary zipcode sequence.

Isolation and association of avian bornavirus from psittacine proventricular dilatation disease samples:

Proventricular dilatation disease (PDD) of psittacine birds (parrots and macaws) is generally a fatal disease with no known treatment, and therefore poses a significant threat to exotic bird populations both in captivity and in the wild. This disease is marked by an inability of the birds to digest food, generally thought to result from paralysis of the upper digestive tract, as well as a myriad of neurological symptoms (13). At the tissue level, PDD manifests as lymphoplasmacytic infiltrates within myenteric ganglia and nerves, but can also include infiltrates in other nerve and muscle tissue.

PDD has long been suspected of having an infectious cause, as outbreaks occur within bird populations housed in close quarters, such as aviaries (14, 15). Intriguingly, virus-like particles have twice been reported in the tissues of PDD affected birds (16, 17). Several types of viruses have also been observed in the feces of affected birds (13, 18, 19), yet none of these observations were replicated or significantly pursued. As a result, despite clues as to a possible viral etiology, no concrete association had been established between PDD and any particular virus at the start of our study.

To determine in an unbiased manner if any known or closely related unknown virus might be associated with PDD, samples from two different outbreaks from geographically remote locations, as well as controls from birds that died of other causes, were subjected to microarray analysis on the ViroChip (20, 21). A signature for *Bornaviridae* emerged from several PDD samples, but no controls, and was followed up by PCR and sequencing of a portion of the viral genome that corresponded to the array probes. These analyses confirmed the presence of a previously unknown species of

bornavirus termed “avian bornavirus” with limited sequence identity to previously identified mammalian bornaviruses.

To recover the entire genome of this suspected new virus, an iterative strategy of high-throughput Illumina RNA-Seq followed by primer walking was employed on a single ABV-positive sample. Partial sequencing of important regions of the new virus from other ABV-positive samples allowed for phylogenetic analysis and comparison of various isolates. Additionally, statistical analyses confirmed a strong association between the presence of ABV and PDD.

Plasmodium falciparum's genomic and transcriptomic responses to artemisinin:

Artemisinin, a compound isolated from sweet wormwood in the 1970's (22), has proved an extraordinarily effective antimalarial drug. It rapidly clears parasites from the bloodstream and also reduces the transmission efficiency of gametocytes. Despite decades of study however, its mechanism of action remains unknown, though the central endoperoxide bridge is known to be essential for activity (23). Although some researchers believe artemisinin has a specific target in the parasite (24, 25), others believe it has pleiotropic effects, perhaps non-specifically alkylating or oxidizing parasite proteins (26).

Recently, two groups reported the observance of unique “dormant” or “quiescent” parasite forms upon treatment of *P. falciparum* culture-adapted strains with artemisinin (27-29). These strains eventually recrudesced after drug was removed, implying that *P. falciparum* enters this unique state in response to artemisinin, and then eventually emerges to resume normal blood stage growth. Interestingly, these studies may explain

the long-standing observation that patients treated with artemisinin monotherapy have a high rate of recrudescence (22).

Because drug resistance to all previous antimalarials eventually arose in *P. falciparum*, researchers have vigilantly monitored artemisinin treatment in the field. Early reports of longer parasite clearance times in patients treated with artemisinin combination therapies (ATCs) were initially attributed to the failure of partner drugs. However, recent reports demonstrating lingering parasitemia in patients treated with artemisinin monotherapy have demonstrated that resistance to partner drugs is not to blame for the trend (30). These troubling reports raise the specter of parasite resistance to the most widely effective antimalarial drug currently available. In this context, insight into how the drug works and how parasites may have circumvented it is desperately needed.

In conjunction with the laboratory of Dr. Dennis Kyle at the University of South Florida, two related projects involving *P. falciparum* and artemisinin were launched. The first took advantage of a strain they derived to be highly resistant to artemisinin *in vitro* to look for single nucleotide polymorphisms (SNPs), amplifications, and deletions in the genome of the resistant parasite as compared to its sensitive parent strain using high-throughput Illumina sequencing. The goal of this project was to produce new genetic leads correlated to the artemisinin resistance phenotype. Several top-rated SNPs, as well as the only promising amplification, were confirmed using independent experimental techniques. A smaller amplification entirely encompassed within the boundaries of the amplification found in the fully sequenced resistant strain was also found by targeted qPCR in an unrelated *in vitro* derived strain with lower level resistance to artemisinin.

The second project focused on documenting the transcriptional response of both the sensitive parental strain and the derived highly resistant strain to physiological levels of dihydroartemisinin, artemisinin's active metabolite. A large-scale time course over more than 48 hours was performed with resulting samples analyzed by microarray and thin smear. Dormant parasite forms were observed as early as 12 hours after treatment and all treated cultures eventually recrudesced. Microarray data was compared to previous results using the same arrays on the normal developmental cycle of the HB3 strain to determine how artemisinin treatment affected this cycle transcriptionally. Data were also compared between treated and untreated cultures of the same strain to identify potential transcriptional markers of dormancy, and between the two strains as well to identify transcriptional markers of resistance. Interestingly, several genes encompassed by the genomic amplification identified in resistant parasites in the first study were constitutively up-regulated transcriptionally as well in the resistant strain.

Investigating splicing in Plasmodium falciparum using Illumina high-throughput sequencing technology:

Although the cyclical nature of many *Plasmodium falciparum* transcripts during its blood stage lifecycle has been well studied (31-33), the total complement of transcript species produced during this cycle is unknown. Before the rise of high-throughput sequencing technologies, cloning and sequencing of expressed sequence tag (ESTs) libraries was the leading source of transcript sequence information. Despite their limited and shallow coverage of the parasite transcriptome, EST collections in *P. falciparum* have provided useful information about the structure of some transcripts, particularly

with regards to splicing, as intron annotations were frequently inaccurate in early gene models (34).

In general, splicing in *P. falciparum* is vastly understudied. snRNA components of the major U2-type spliceosome have been computationally identified and their expression confirmed by Northern blot (35). However, computational efforts to identify snRNA components of the minor U12-type spliceosome were unsuccessful in rodent *Plasmodia*, perhaps indicating that *Plasmodia* do not possess a minor spliceosome (36). Only one protein component of the major spliceosome, PfUAP56, has been experimentally verified (37) and homologs of other spliceosomal proteins have not been systematically computationally identified. As a first foray into this field, we systematically determined the most likely *P. falciparum* homologs for each spliceosomal or spliceosome-associated protein in humans and yeast.

Work on the transcript signals used by the *P. falciparum* spliceosome to distinguish between introns and exons is similarly incomplete. ESTs confirmed *P. falciparum*'s general use of canonical eukaryotic splice sites (5' GU-AG 3') and U-rich poly-pyrimidine tracks. However, efforts to computationally identify a spliceosome-recognized branchpoint motif by using EST-defined introns were unsuccessful (35). Also, despite gene annotations using 5' GC-AG 3' non-canonical intron boundaries, no study has determined the range of splice sites acceptable to the parasite.

Many studies, ranging from detailed analyses of single genes to small-scale EST or RNA-Seq analysis of several genes, have reported alternatively spliced transcripts in *Plasmodia* (38-40, 34, 41-43). Little is known about the significance of the discovered alternative isoforms or if their production is regulated in some fashion. Since alternative

isoforms have been shown to be important for expansion of protein functionality and even gene regulation in other organisms (44, 45), understanding the prevalence and role(s) of alternative splicing in *P. falciparum* will likely lead to a better understanding of the molecular inner working of this pathogen.

We investigated splicing in *P. falciparum* from the point of view of spliced transcripts by analyzing all splice junctions present in the blood stages using our own RNA-Seq data, as well as published RNA-Seq data. This application of Illumina high-throughput sequencing required both protocol development, as commercial RNA-Seq kits were not yet available, as well as significant development of analysis tools in order to identify and accurately map exon-exon junction reads within the data. Implementation of a tailed random priming strategy to produce double stranded cDNA with the end adapters required for Illumina sequencing produced severely uneven or “jackpotted” coverage of transcripts. This phenomenon was likely caused by lingering RNA structure as well as the known tendency for random priming to actually prime RNA non-randomly (46). This jackpotting problem led us to develop the Long March protocol, which leverages redundancy within the initial cDNA library to create distinct overlapping cDNA molecules.

As with protocols for producing Illumina sequencing libraries for RNA-Seq, analysis techniques for detecting exon-exon junctions within short read data were limited when this project began. The most widely used algorithm, TopHat, determines likely exons in a transcriptome based on read coverage and uses these exons to build an exon-exon junction reference library (47). Unmatched reads are then compared to that library to determine which junctions are present in the data. This approach relies on clear

demarcation of exons within the read coverage, which does not occur when coverage is generally low or uneven, when alternative splicing is prevalent, or when introns or intergenic regions are short. Poor performance of this algorithm on our data led us to develop an alternate algorithm based on directly mapping reads to the genome with gapped alignments.

After identifying junction reads, our secondary analysis focused on identifying intron sequences likely recognized by the *P. falciparum* spliceosome, cataloguing the types of introns recognized in *P. falciparum*, looking for evidence of alternative splicing and its prevalence, and deciphering splice sites found that were antisense to existing gene models. These RNA-Seq datasets and our subsequent analysis of them provide evidence that like many organisms, the *P. falciparum* transcriptome is much more complex in the types of transcripts it produces than previously appreciated.

References:

1. Penny,D., Hoepfner,M.P., Poole,A.M. and Jeffares,D.C. (2009) An Overview of the Introns-First Theory. *J Mol Evol*, **69**, 527-540, 10.1007/s00239-009-9279-5.
2. Cavalier-Smith,T. (1993) Kingdom protozoa and its 18 phyla. *Microbiol Rev*, **57**, 953-994.
3. Liu,S., Lin,L., Jiang,P., Wang,D. and Xing,Y. (2010) A comparison of RNA-Seq and high-density exon array for detecting differential gene expression between closely related species. *Nucleic Acids Res*, 10.1093/nar/gkq817. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/20864445> [Accessed November 17, 2010].
4. Du,T., Schmid,M. and Jansen,R. (2007) Why cells move messages: the biological functions of mRNA localization. *Semin. Cell Dev. Biol*, **18**, 171-177, 10.1016/j.semcdb.2007.01.010.
5. Wang,D.O., Martin,K.C. and Zukin,R.S. (2010) Spatially restricting gene expression by local translation at synapses. *Trends Neurosci*, **33**, 173-182, 10.1016/j.tins.2010.01.005.
6. Bobola,N., Jansen,R., Shin,T.H. and Nasmyth,K. (1996) Asymmetric Accumulation of Ash1p in Postanaphase Nuclei Depends on a Myosin and Restricts Yeast Mating-Type Switching to Mother Cells. *Cell*, **84**, 699-709, 10.1016/S0092-8674(00)81048-X.
7. Long,R.M., Singer,R.H., Meng,X., Gonzalez,I., Nasmyth,K. and Jansen,R.P. (1997) Mating type switching in yeast controlled by asymmetric localization of ASH1 mRNA. *Science*, **277**, 383-387.
8. Takizawa,P.A., Sil,A., Swedlow,J.R., Herskowitz,I. and Vale,R.D. (1997) Actin-

dependent localization of an RNA encoding a cell-fate determinant in yeast.

Nature, **389**, 90-93, 10.1038/38015.

9. Bittker, J.A., Le, B.V. and Liu, D.R. (2002) Nucleic acid evolution and minimization by nonhomologous random recombination. *Nat. Biotechnol.*, **20**, 1024-1029, 10.1038/nbt736.
10. Bernstein, D.S., Buter, N., Stumpf, C. and Wickens, M. (2002) Analyzing mRNA-protein complexes using a yeast three-hybrid system. *Methods*, **26**, 123-141, 10.1016/S1046-2023(02)00015-4.
11. Shepard, K.A., Gerber, A.P., Jambhekar, A., Takizawa, P.A., Brown, P.O., Herschlag, D., DeRisi, J.L. and Vale, R.D. (2003) Widespread cytoplasmic mRNA transport in yeast: identification of 22 bud-localized transcripts using DNA microarray analysis. *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 11429-11434, 10.1073/pnas.2033246100.
12. Takizawa, P.A. and Vale, R.D. (2000) The myosin motor, Myo4p, binds Ash1 mRNA via the adapter protein, She3p. *Proc. Natl. Acad. Sci. U.S.A.*, **97**, 5273-5278, 10.1073/pnas.080585897.
13. Gregory, C., Latimer, K., Niagro, F., Ritchie, B., Campagnoli, R., Norton, T. and Greenacre, C. (1994) A review of proventricular dilatation syndrome. *J Assoc Avian Vet.*, **8**, 69-75.
14. Doneley, R.J.T., Miller, R.I. and Fanning, T.E. (2007) Proventricular dilatation disease: an emerging exotic disease of parrots in Australia. *Aust. Vet. J.*, **85**, 119-123, 10.1111/j.1751-0813.2007.00109.x.
15. Lublin, A., Mechani, S., Farnoushi, I., Perl, S. and Bendheim, U. (2006) An outbreak of

- proventricular dilatation disease in psittacine breeding farm in Israel. *Israel Journal of Veterinary Medicine*, **61**, 16-19.
16. Gough,R.E., Drury,S.E., Harcourt-Brown,N.H. and Higgins,R.J. (1996) Virus-like particles associated with macaw wasting disease. *Vet. Rec*, **139**, 24.
17. Mannl,A., Gerlach,H. and Leipold,R. (1987) Neuropathic gastric dilatation in psittaciformes. *Avian Dis*, **31**, 214-221.
18. Gough,R.E., Drury,S.E., Culver,F., Britton,P. and Cavanagh,D. (2006) Isolation of a coronavirus from a green-cheeked Amazon parrot (*Amazon viridigenalis* Cassin). *Avian Pathol*, **35**, 122-126, 10.1080/03079450600597733.
19. Ritchie,B. (1995) *Avian Viruses: Function and Control* Wingers Publishing, Lake Worth.
20. Chiu,C.Y., Rouskin,S., Koshy,A., Urisman,A., Fischer,K., Yagi,S., Schnurr,D., Eckburg,P.B., Tompkins,L.S., Blackburn,B.G. et al. (2006) Microarray detection of human parainfluenzavirus 4 infection associated with respiratory failure in an immunocompetent adult. *Clin. Infect. Dis*, **43**, e71-76, 10.1086/507896.
21. Chiu,C.Y., Alizadeh,A.A., Rouskin,S., Merker,J.D., Yeh,E., Yagi,S., Schnurr,D., Patterson,B.K., Ganem,D. and DeRisi,J.L. (2007) Diagnosis of a critical respiratory illness caused by human metapneumovirus by use of a pan-virus microarray. *J. Clin. Microbiol*, **45**, 2340-2343, 10.1128/JCM.00364-07.
22. Antimalaria studies on Qinghaosu (1979) *Chin. Med. J*, **92**, 811-816.
23. Mercer,A.E. (2009) The role of bioactivation in the pharmacology and toxicology of the artemisinin-based antimalarials. *Curr Opin Drug Discov Devel*, **12**, 125-132.
24. Eckstein-Ludwig,U., Webb,R.J., Van Goethem,I.D.A., East,J.M., Lee,A.G.,

- Kimura,M., O'Neill,P.M., Bray,P.G., Ward,S.A. and Krishna,S. (2003)
Artemisinin target the SERCA of Plasmodium falciparum. *Nature*, **424**, 957-961,
10.1038/nature01813.
25. Li,W., Mo,W., Shen,D., Sun,L., Wang,J., Lu,S., Gitschier,J.M. and Zhou,B. (2005)
Yeast model uncovers dual roles of mitochondria in action of artemisinin. *PLoS
Genet*, **1**, e36, 10.1371/journal.pgen.0010036.
26. Olliaro,P.L., Haynes,R.K., Meunier,B. and Yuthavong,Y. (2001) Possible modes of
action of the artemisinin-type compounds. *Trends Parasitol*, **17**, 122-126.
27. Chavchich,M., Gerena,L., Peters,J., Chen,N., Cheng,Q. and Kyle,D.E. (2010) Role of
pfmdr1 amplification and expression in induction of resistance to artemisinin
derivatives in Plasmodium falciparum. *Antimicrob. Agents Chemother*, **54**, 2455-
2464, 10.1128/AAC.00947-09.
28. Teuscher,F., Gatton,M.L., Chen,N., Peters,J., Kyle,D.E. and Cheng,Q. (2010)
Artemisinin - induced dormancy in plasmodium falciparum: duration, recovery
rates, and implications in treatment failure. *J. Infect. Dis*, **202**, 1362-1368,
10.1086/656476.
29. Witkowski,B., Lelièvre,J., Barragán,M.J.L., Laurent,V., Su,X., Berry,A. and Benoit-
Vical,F. (2010) Increased tolerance to artemisinin in Plasmodium falciparum is
mediated by a quiescence mechanism. *Antimicrob. Agents Chemother*, **54**, 1872-
1877, 10.1128/AAC.01636-09.
30. Dondorp,A.M., Nosten,F., Yi,P., Das,D., Phyto,A.P., Tarning,J., Lwin,K.M., Arie,F.,
Hanpithakpong,W., Lee,S.J. et al. (2009) Artemisinin resistance in Plasmodium
falciparum malaria. *N. Engl. J. Med*, **361**, 455-467, 10.1056/NEJMoa0808859.

31. Bozdech,Z., Llinás,M., Pulliam,B.L., Wong,E.D., Zhu,J. and DeRisi,J.L. (2003) The Transcriptome of the Intraerythrocytic Developmental Cycle of Plasmodium falciparum. *PLoS Biol*, **1**, e5, 10.1371/journal.pbio.0000005.
32. Le Roch,K.G., Zhou,Y., Blair,P.L., Grainger,M., Moch,J.K., Haynes,J.D., De la Vega,P., Holder,A.A., Batalov,S., Carucci,D.J. et al. (2003) Discovery of Gene Function by Expression Profiling of the Malaria Parasite Life Cycle. *Science*, **301**, 1503-1508, 10.1126/science.1087025.
33. Llinás,M., Bozdech,Z., Wong,E.D., Adai,A.T. and DeRisi,J.L. (2006) Comparative whole genome transcriptome analysis of three Plasmodium falciparum strains. *Nucleic Acids Res*, **34**, 1166-1173, 10.1093/nar/gkj517.
34. Lu,F., Jiang,H., Ding,J., Mu,J., Valenzuela,J., Ribeiro,J. and Su,X. (2007) cDNA sequences reveal considerable gene prediction inaccuracy in the Plasmodium falciparum genome. *BMC Genomics*, **8**, 255, 10.1186/1471-2164-8-255.
35. Chakrabarti,K., Pearson,M., Grate,L., Sterne-Weiler,T., Deans,J., Donohue,J.P. and Ares,M. (2007) Structural RNAs of known and unknown function identified in malaria parasites by comparative genomics and RNA analysis. *RNA*, **13**, 1923-1939, 10.1261/rna.751807.
36. Lopez,M.D., Alm Rosenblad,M. and Samuelsson,T. (2008) Computational screen for spliceosomal RNA genes aids in defining the phylogenetic distribution of major and minor spliceosomal components. *Nucl. Acids Res.*, **36**, 3001-3010, 10.1093/nar/gkn142.
37. Shankar,J., Pradhan,A. and Tuteja,R. (2008) Isolation and characterization of Plasmodium falciparum UAP56 homolog: Evidence for the coupling of RNA

- binding and splicing activity by site-directed mutations. *Archives of Biochemistry and Biophysics*, **478**, 143-153, 10.1016/j.abb.2008.07.027.
38. Bracchi-Ricard,V., Barik,S., Delvecchio,C., Doerig,C., Chakrabarti,R. and Chakrabarti,D. (2000) PfPK6, a novel cyclin-dependent kinase/mitogen-activated protein kinase-related protein kinase from *Plasmodium falciparum*. *Biochem J*, **347**, 255-263.
39. Iriko,H., Jin,L., Kaneko,O., Takeo,S., Han,E., Tachibana,M., Otsuki,H., Torii,M. and Tsuboi,T. (2009) A small-scale systematic analysis of alternative splicing in *Plasmodium falciparum*. *Parasitology International*, **58**, 196-199, 10.1016/j.parint.2009.02.002.
40. Knapp,B., Nau,U., Hundt,E. and Küpper,H.A. (1991) Demonstration of alternative splicing of a pre-mRNA expressed in the blood stage form of *Plasmodium falciparum*. *J. Biol. Chem*, **266**, 7148-7154.
41. Otto,T.D., Wilinski,D., Assefa,S., Keane,T.M., Sarry,L.R., Böhme,U., Lemieux,J., Barrell,B., Pain,A., Berriman,M. et al. (2010) New insights into the blood-stage transcriptome of *Plasmodium falciparum* using RNA-Seq. *Molecular Microbiology*, **76**, 12-24, 10.1111/j.1365-2958.2009.07026.x.
42. Saenz,F.E., Balu,B., Smith,J., Mendonca,S.R. and Adams,J.H. (2008) The Transmembrane Isoform of *Plasmodium falciparum* MAEBL Is Essential for the Invasion of *Anopheles* Salivary Glands. *PLoS ONE*, **3**, e2287, 10.1371/journal.pone.0002287.
43. Wentzinger,L., Bopp,S., Tenor,H., Klar,J., Brun,R., Beck,H.P. and Seebeck,T. (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,

10.1016/j.ijpara.2008.05.016.

44. Barash, Y., Calarco, J.A., Gao, W., Pan, Q., Wang, X., Shai, O., Blencowe, B.J. and

Frey, B.J. (2010) Deciphering the splicing code. *Nature*, **465**, 53-59,

10.1038/nature09000.

45. Sun, S., Zhang, Z., Sinha, R., Karni, R. and Krainer, A.R. (2010) SF2/ASF

autoregulation involves multiple layers of post-transcriptional and translational control. *Nat Struct Mol Biol*, **17**, 306-312, 10.1038/nsmb.1750.

46. Hansen, K.D., Brenner, S.E. and Dudoit, S. (2010) Biases in Illumina transcriptome

sequencing caused by random hexamer priming. *Nucleic Acids Res*, **38**, e131,

10.1093/nar/gkq224.

47. Trapnell, C., Pachter, L. and Salzberg, S.L. (2009) TopHat: discovering splice junctions

with RNA-Seq. *Bioinformatics*, **25**, 1105-1111, 10.1093/bioinformatics/btp120.

Chapter 2: Unbiased Selection of Localization Elements Reveals cis-acting Determinants of mRNA Bud-Localization in *Saccharomyces cerevisiae*

This chapter is a reprint from the following reference:

Ashwini Jambhekar, Kimberly McDermott, Katherine Sorber, Kelly A. Shepard, Ronald D. Vale, Peter A. Takizawa, Joseph L. DeRisi. (2005). Unbiased selection of localization elements reveals cis-acting determinants of mRNA bud-localization in *Saccharomyces cerevisiae*. PNAS **102**(50): 18005-18010.

Copyright © 2005, National Academy of Sciences, U.S.A.

Author contributions:

Ashwini Jambhekar obtained data in figures 1, 2, 3a, b, d, 4, 5, 6, 7, 8, 11 and listed in Tables 1-4, 6-8. Kimberly McDermott and Peter Takizawa obtained data in figures 3c and 9. Katherine Sorber obtained data in figure 11, assisted with the figure 5a experiment, obtained all data in Table 5, and collected 20% of the data listed in Table 4. Kelly Shepard tested 2 of the candidate zipcodes. Ronald D. Vale and Joseph L. DeRisi supervised research.

Joseph L. DeRisi, Thesis Advisor

Abstract:

Cytoplasmic mRNA localization is a mechanism used by many organisms to generate asymmetry and sequester protein activity. In the yeast *Saccharomyces cerevisiae*, mRNA transport to bud tips of dividing cells is mediated by the binding of She2p, She3p, and Myo4p to coding regions of the RNA. To date, twenty-four bud-localized mRNAs have been identified, yet the RNA determinants that mediate localization remain poorly understood. Here, we utilized Nonhomologous Random Recombination (NRR) to generate libraries of sequences that could be selected for their ability to bind She-complex proteins, thereby providing an unbiased approach for minimizing and mapping localization elements in several transported RNAs. Analysis of the derived sequences and predicted secondary structures revealed short sequence motifs that mediate binding to the She-complex as well as RNA localization to the bud tip *in vivo*. A predicted single-stranded core CG dinucleotide appears to be an important component of the RNA-protein interface although other nucleotides contribute in a context-dependent manner. Our findings further our understanding of RNA recognition by the She-complex, and the methods employed here should be applicable for elucidating minimal RNA motifs involved in many other types of interactions.

Introduction:

Localization of mRNA is commonly employed to target proteins to specific regions within a cell. In most cases, this process requires recognition by RNA-binding protein(s) and linkage of the resulting RNA-protein complex directly or indirectly to molecular motors (1). The determinants of recognition, transport factor binding, and

subsequent targeting are cis-acting sequences often found in untranslated regions. Precise characterization of these RNA “zipcodes” has proven to be cumbersome for several reasons. The reported length of the minimal sequence requirements for transport ranges from 50 nucleotides (nt) to several hundred, and this apparent complexity is compounded by functional redundancy among zipcodes and a diversity of cellular recognition components (2-5).

The yeast *Saccharomyces cerevisiae* provides a tractable model system to characterize the determinants of zipcode recognition. To date, twenty-four bud-localized mRNAs have been identified, and coding regions were shown to mediate transport (6). Localization is dependent on the She-complex, which comprises She2p, a putative RNA binding protein, Myo4p, a type V myosin motor, and She3p, which interacts directly with both Myo4p and She2p (7-9).

Independent studies of one transported RNA, *ASH1*, identified three (N, C, U) (10) or four (E1, E2A, E2B, E3) (11) zipcodes based on their ability to mediate localization of a reporter. Only one of the elements lies in the 3' UTR; the remaining are located within the coding region. These elements bear no obvious primary sequence or secondary structural similarity to each other, and mutational analysis suggested that secondary structure was required for activity (10, 11). Recently, Olivier et al. (12) reported that a CGA triplet in a loop, along with a single-stranded cytosine six bases away and opposite to the triplet, was necessary for bud-localization of *ASH1* and two other RNAs. However, these criteria are insufficient to identify zipcodes in other RNAs localized by the She-complex (6).

To extend our understanding of the She-complex-RNA interaction, we employed an unbiased approach to select zipcode-containing fragments from pools of known localized RNAs. The fragments were tested for localization *in vivo* and *bona fide* zipcodes were subjected to further analysis, which revealed a highly-degenerate motif predicted to lie in single-stranded regions and necessary for She-complex-dependent transport. Highlighting the complexity of the She2/3p-RNA interaction, we also found that the precise sequences mediating recognition and transport depend upon the context of the adjacent sequence and structural features in the mRNA.

Materials and Methods:

Nonhomologous Random Recombination (NRR):

NRR was carried out as described (13). Briefly, 1-4 µg of DNA encoding She-complex targets was PCR amplified from S288c genomic DNA or plasmid clones using Pfu DNA polymerase (Stratagene) and digested with 0.1-1 U DNaseI (Gibco) at room temperature for 5 min. DNA fragments 20-200 bp in length were size selected by agarose gel electrophoresis and purified by electroelution. Fragments were blunted with 9U T4 DNA Polymerase (NEB) and 200 µM dNTPs at 12°C for 30 min as described by manufacturer. DNA was purified by phenol-chloroform extraction and sepharose gel filtration. 5-15% of resulting DNA was used for ligation with 30 pmol of 5' phosphorylated T7hairpin (AAACCCTATAGTGAGTCGTATTAGTTTAAACGGC CCGCGCGGGCCGTTTAAACTAATACGACTCACTATAGGGTTT), or with a combination of 15 pmol each of 5' phosphorylated XmaHairpin1 (AAACCCGGGCCTG ACTCCGAAGTCGTTTAAACGGCCGCGCGGCCGTTTAAACGACTTCGGAGTCA

GGCCCGGGTTT) and SphHairpin1 (AAACGCATGCCTGACTCCGAAGTCGTTT AAACGGCCGCGCGGCCGTTTAAACGACTTCGGAGTCAGGCATGCGTTT).

Ligated DNA was digested with *PmeI* to remove hairpin ends and hairpin dimers. 10% of the restriction digest was used for PCR with 1 μ M of XmaT7 primer (TCGACCCGGGTAATACGACTCACTATAGGG) or 1 μ M of NRRprimer1 (AAACGACTTCGGAGTCAGG) using Pfu DNA polymerase (Stratagene). Reactions were denatured at 94°C for 1 min and then cycled 35 to 40 times as follows: 94°C 20 sec, 52°C 30 sec, 68°C 1:30. PCR products were digested with 20 U *XmaI* (XmaT7 products) or 20 U *XmaI* plus 20 U *SphI* (NRRprimer1 products) overnight. NRR was carried out with T7hairpin and *ASH1*, *YLR434c*, *ERG2*, or *MID2* sequences separately. XmaHairpin1 and SphHairpin1 were used for separate NRR reactions with *CPSI*, *DNMI*, *WSC2*, *MMR1*, or *YGR046w*, or with a pool of *ERG2*, *MID2*, and bp 1-1000 and 1500-1761 of *TPO1*. Bp 1000-1500 of *TPO1* were excluded because they do not contain any localization sequences (AJ and JLD, unpublished results). The full coding region of each gene was used for NRR (unless noted), except for *ASH1*, which included the coding region plus 99 bp of downstream sequence which contains a localization signal (10, 11).

3-Hybrid RNA-Expression Library Construction:

NRR products were ligated into the *XmaI* site (for XmaT7 products) or asymmetrically into *XmaI* and *SphI* sites (for NRRprimer1 products) of pIII Δ A/MS2.2. This vector consists of pIII Δ A/MS2.2 (14) with a deletion of the *AatII*-*Tth111*-I fragment encoding *ADE2*. In all cases, the library size was sufficient to ensure that every sequence was represented at least once. Separate libraries were constructed for *YLR434c* and *ASH1*.

Another library contained NRR products derived from a pool of *ERG2*, *MID2*, and bases 1-500 and 1500-1761 of *TPO1*. A fourth library contained *ERG2*, *MID2*, *WSC2*, *DNMI*, *CPS1*, *YGR046W*, *MMR1*, *YMR171C*, *SRL1* NRR products. Each library was screened separately by 3-Hybrid analysis (15).

For randomization experiments, complementary oligonucleotides fully degenerate at the indicated positions were annealed and cloned into the *NotI* and *XhoI* sites of pAJ232, which consists of pIIIΔA/MS2.2 with an insertion of *NotI* and *XhoI* in the *XmaI* site. Plasmid library members were selected at 10mM 3-AT as described below.

3-Hybrid selection:

DNA encoding the carboxyl-terminus of She3 (bp 706-1278) was cloned into *XmaI*/*SacI* sites of Gal4-AD expression vector pACT2. The *ADE2* ORF with 500 bp of upstream sequence and 300 bp downstream sequence was cloned in the *NotI* site of pACT2, providing an additional marker. The resulting plasmid was introduced into the 3-Hybrid L40 coat host strain (15). Where indicated, *SHE2* was deleted in L40 coat as described (16). 12-20 μg of RNA plasmid libraries was transformed into the She3-L40 coat strain using the lithium acetate method (17). Transformants were plated on SD-HIS-URA medium containing 6.67mg/ L adenine and 0, 0.5, 1, 5, 10 or 15mM 3-aminotriazole (3-AT). White transformants represent candidates which require She3 for expression of the *HIS3* reporter. Red transformants represent candidates which have lost the *SHE3* plasmid but express *HIS3* in the absence of any RNA-protein interaction, and accounted for <15% of the transformants selected at or above 5mM 3-AT. 30-80 white colonies growing at the highest 3-AT concentrations were tested for expression of LacZ

by X-gal filter assay (14). All tested candidates expressed LacZ (data not shown).

Plasmids were rescued and inserts fully sequenced from the 5' end.

Quantitative β -galactosidase assays were performed as described (14), except that cells were lysed with Yeast Protein Extraction Reagent (Pierce).

Visualization of RNA:

The U1A-GFP system was used for visualizing RNA localization *in vivo* (6, 8). RNAs longer than 150 nt were cloned directly into the pGAL-U1A vector (6) containing *NotI* and *XhoI* cloning sites. Shorter RNAs were assayed by fusing to the 3' end of the unlocalized *ADHI* gene. For RNAs shorter than 75 nt, a linker containing a 13 bp inverted repeat separated by *NotI* and *XhoI* sites was inserted downstream of *ADHI*. Synthetic oligos (Operon) encoding the target RNA sequences were ligated into the *NotI* and *XhoI* sites, so that the RNA was expressed with flanking inverted repeats which formed a stable helix.

For visualization of RNA, the pGAL-U1A plasmid containing the RNA of interest was introduced into a W303 yeast strain harboring the U1A-GFP plasmid (6, 8). >50 premitotic cells expressing RNA were counted from 2 independent transformants for each RNA as described (6).

RNA structure predictions:

All RNA structure predictions were computed using MFOLD (18, 19).

Protein purification and gel shifts:

She2p-HA contains a single HA epitope at its C-terminus. She2p-HA was overexpressed in *S. cerevisiae* and isolated from cell extracts with anti-HA antibodies coupled to protein A sepharose (Sigma). She2p-HA was eluted from the resin with excess HA peptide, dialyzed to remove free peptide and concentrated in a Microcon YM-10 (Millipore). His-She3p 251-425 contains a His₆ tag at the N-terminus of amino acids 251-425 of She3p. His-She3p 251-425 was expressed in BL21 RIPL (Stratagene) and purified with Ni-NTA agarose (Qiagen) according to the manufacture's instructions. To generate ³²P-labeled RNAs for mobility shifts, annealed oligos containing a T7 promoter followed by a particular zipcode sequence were used as templates in an *in vitro* transcription reaction. The oligo templates were added to a Maxiscript T7 (Ambion) reaction containing UTP-³²P (Amersham). Full length RNAs were gel purified from the reactions. Each gel shift reaction contained 0.5 nM labeled RNA, 0.1 mg/mL tRNA in 25 mM Hepes-KOH pH 7.5, 100 mM KCl, 2 mM MgCl₂, 1 mM DTT. Purified She2p-HA and His-She3 251- 425 were added at varying concentrations. Reactions were incubated at room temperature for 30 min and then run on a 5% acrylamide gel (37.5:1) in TBE at 4°C. The gel was fixed, dried and exposed to film.

Results:

Identification of She-complex dependent localization sequences:

We sought to identify short zipcodes from known transported RNAs in a high-throughput manner without making assumptions about exact zipcode length, orientation, or connectivity. For this reason, we utilized Nonhomologous Random Recombination (NRR) (13) to generate libraries of sequences that could be selected for their ability to

bind to She-complex proteins. We reasoned that the region of overlap of multiple, independently-selected clones would define a short zipcode.

To generate a library by NRR, DNA encoding a target RNA was digested with DNaseI, and 20-200 bp fragments were isolated and ligated in the presence of hairpin linkers to generate products containing 1-3 tandem fragments of various sizes and connectivities flanked by hairpins. The products were PCR amplified with primers complementary to the linker sequence and selected for interaction with the She-complex by 3-Hybrid assay (Fig. 1a). As bait, we used the carboxyl-terminus of She3p, which interacts with She2p (7) and displays proper specificity for RNA targets (9) (vector and IRE controls, Fig. 2). For the two RNAs tested, the 3-Hybrid interaction also required endogenous She2p (*she2* WSC2N and *she2* Umin, Fig. 2), indicating the formation of a tripartite RNA-protein complex.

To validate this approach, we subjected *ASH1* to NRR and 3-Hybrid selection. Sequencing of NRR-generated clones prior to selection revealed fragments derived from various parts of the gene (Fig. 1b). After selection, almost all clones fell within previously identified localization elements (Fig. 1c). Although no sequences were recovered from E2A, this zipcode is active in the 3-Hybrid system (12); therefore its absence in our selection most likely resulted from insufficient sequencing of positive transformants. Only one selected clone did not contain a fragment overlapping known localization elements and was not pursued further. In all cases, the sequences defined by selected overlapping clones were shorter than the zipcodes from which they were derived (10, 11). To verify that the shorter sequences localized *in vivo*, we used the U1A-GFP system (6, 8) to visualize RNA distribution in live cells. Sequences shorter than 150 nt in

length were fused to the 3' end of *ADHI* and assayed for their ability to direct bud localization of the RNA. All *ASHI* sequences defined by the NRR/ 3-Hybrid selection localized to bud tips in >90% of cells (Table 1, Fig. 3 c-e insets).

Ten other genes encoding localized RNAs were screened in this manner individually (*YLR434c*) or in pools (*ERG2*, *MID2*, *TPO1*, *WSC2*, *MMR1*, *SRL1*, *CPS1*, *DNM1*, *YGR046w*), and ten more putative zipcodes were identified ranging from 50 (*YLR434-1*) to 201 (*DNM1N*) nt in length (Table 1). All sequences defined by overlapping clones were tested for localization *in vivo*. Although the control *ADHI* reporter was localized in only 20% of cells, our experience with testing various constructs has revealed that, in rare cases, unlocalized RNAs can produce dim, bud-localized particles in up to 60% of cells in a She2p-independent manner. Thus, we classified any RNA that was localized in fewer than 60% of cells as unlocalized. Only one selected RNA, *CPS1CR*, failed to localize by this criterion. Of the remainder, nine sequences localized in >90% of cells in a She2p-dependent manner (Table 1, data not shown). Two others, *TPO1N* and *DNM1N*, localized less efficiently (in 70-80% of cells). In general, sequences recovered multiple times at high 3-AT concentrations were more likely to localize than those recovered once or only at low 3-AT concentrations (Table 1, 2). Although some zipcodes were recovered numerous times, we failed to recover any zipcodes from *CPS1*, *MID2*, *MMR1*, or *YGR046w*, suggesting that the screen was not saturating.

Identification of a conserved She2/3p-dependent localization motif:

We used MEME analysis, which identifies statistically over-represented sequence motifs within a data set (20), to find any motifs shared by the newly-identified zipcodes. The data set consisted of the nine zipcodes displaying >90% localization activity, including two (WSC2N and YLR434-2) which had been minimized by deletion mapping (Fig. 3). Of several candidates, one degenerate motif (RCGAADA) was present in all input sequences and mapped almost exclusively (in seven out of eight cases) to single-stranded regions of the secondary structures predicted by MFOLD (18, 19) (Fig. 3). One zipcode, WSC2N, displayed two copies of the motif--a more degenerate version in the terminal loop and a consensus sequence in the 3' bulge (Fig. 3a). Additionally, seven zipcodes contained an adenosine six bases upstream of the motif. This sequence pattern was observed in 3 other zipcodes not included in the MEME analysis (E2A in *ASH1* and zipcodes in *IST2* and *YMR171c* (12)).

Five zipcodes were selected for further analysis based on the fact that the 7-base motif could be mutated or deleted in these RNAs without affecting the predicted structure of the remainder of the molecule (Fig. 3a-e, Table 3). Wild-type (WT) zipcodes localized in >90% of budded cells (Fig. 4), and displayed β -galactosidase activities above 200 Miller Units (Fig. 2). All zipcodes required the motif for localization and LacZ expression (Fig. 4a, b, d). Deletions or mutations of the motif in E1min, E2Bmin, and YLR434-2 abolished activity in both assays. Deletion of the motif in Umin also abolished localization, but decreased β -galactosidase activity by only 65% (Fig. 4a, b). WSC2N, which contains two copies of the motif, required mutations in both to abolish localization and β -galactosidase activity (Figs. 4a, b, 5).

The ability of purified She2p and the carboxyl terminus of She3p (251-425) to bind WT and mutant zipcodes directly was also tested by RNA mobility shift. Nanomolar concentrations of She2p and She3p retarded the mobility of all WT zipcodes, indicating that She2/3 bind directly to each zipcode (Fig. 4c, 6). Furthermore, the protein complex displayed sequence-specific binding, as mutations of the motif in Umin, YLR434-2, E2Bmin, and E1min decreased or abolished the shift (Fig. 4c, 6). Although a large amount of WT RNAs remained unbound at the highest protein concentrations, it is unlikely that additional proteins facilitate She-complex binding to RNA *in vivo*, as a limited number of proteins, like She2p (21), are present in both the nucleus and cytoplasm to facilitate bud-localization and 3-Hybrid activity. It is more likely that some of the RNA misfolds and cannot bind She2/3 *in vitro*. Nevertheless, we conclude that the degenerate motif is essential for RNA binding of She2/3, and that activity in localization and β -galactosidase assays reflects binding of the RNA to the She-complex.

In addition to the recognition motif, MEME analysis identified an adenosine six bases upstream in seven zipcodes. Mutation or deletion of this base caused varying effects on 3-Hybrid activity ranging from an increase (YLR434-2min) to a 5-fold reduction (E2Bmin) (Fig 7a). Some base substitutions may be more favorable than others at this position, resulting in the range of phenotypes displayed by the mutations in different zipcodes. Although this adenosine was highly conserved among the zipcodes, its contribution to binding was context-dependent.

While the mutational analyses revealed that the primary sequence of the motif was essential for zipcode activity, they did not address the structural requirements for She2/3 recognition. To determine whether the single-stranded nature of the motif was

necessary for She-complex recognition, the 5' end of YLR434-2 was changed to complement the motif at the 3' end, thus placing the motif in a predicted duplex. The resulting RNA (YLR434-2 double-stranded motif) failed to localize *in vivo* and did not display significant 3-Hybrid β -galactosidase activity (Fig. 4a, b), indicating that the She-complex cannot bind its recognition site in a stable helix. We also observed that the recognition motifs bordered predicted helices in most zipcodes. To test whether this juxtaposition was essential, two nucleotides were inserted between the stems and motifs of four zipcodes. The resulting mutant phenotypes ranged from no decrease in β -galactosidase activity (Umin) to a complete abolition of She-complex interaction (E1min, YLR434-2) (Fig. 7a).

Although the above results implied that the stems of zipcodes were important for She2/3 binding, no primary sequence similarities were observed in these regions. The current models (10-12) proposed that stems play only a structural role in the RNA-protein interaction. In support of the model, compensatory mutations in the stem of YLR434-2 preserved zipcode function (Fig. 8); but similar mutations in E2Bmin abolished 3-Hybrid activity (Fig. 7b). Therefore, each base pair in the E2Bmin stem was individually mutated in order to identify essential bases. Mutation of each of the two base pairs adjacent to the loop decreased β -galactosidase activity 2- to 4-fold, while mutating the pair at the base of the stem had no effect. (The C₁₂₈₄•G₁₃₀₀ pair was not tested because substitutions were predicted to disrupt the entire stem). Surprisingly, no single base-pair mutation decreased activity to the extent that mutation of the entire stem did. These results indicated that the primary sequence of the stem contributes to She2/3 binding in some cases, and that bases in the stem of E2Bmin contribute in an additive manner. Collectively, these results

support the role of the degenerate, single-stranded motif in mediating She-complex recognition; however, the precise sequence and topological requirements appear to be context-dependent.

Analysis of base contributions within a single zipcode:

Because it appeared that conserved bases in the recognition motif as well as other, less-conserved bases contributed to She2/3 binding, we investigated in detail the sequence requirements for She2/3 binding to a single zipcode. Four- to seven-base regions of a further-minimized E2Bmin zipcode were fully randomized, and the resulting sequences were selected for She-complex binding by the 3-Hybrid system.

The contribution of each base in the loop of E2Bmin was determined via two separate, overlapping randomization/ selection experiments. One position in the loop (1288) displayed no base preferences for She-complex recognition (Fig. 9b). In contrast, 6 out of 7 bases in the WT motif were significantly over-represented upon selection (Fig. 9a), but the importance of each base within the motif for She-complex recognition appeared to vary. The 5' A₁₂₉₁CG triplet was highly over-represented in the selected clones, while a lesser bias towards adenosines at the 3' end was detected (Fig. 9a, Table 4). In support of these observations, mutation of the guanosine (G1293C) in the context of a selected E2Bmin clone (A₁₂₉₁CGUUUU → ACCUUUU) decreased activity 10-fold (data not shown). The motif randomization was repeated in zipcode YLR434-2, and although similar results were obtained, the strength of the base preferences varied at some positions (Fig. 10, Table 5). Surprisingly, the strength of the bias for C₁₂₉₂G varied even between the two overlapping E2B experiments (Fig. 9b, Table 6), indicating that the

requirements for She-complex binding are influenced by the variability of the surrounding region. We noticed that most selected sequences were predicted to form the same secondary structure as WT E2Bmin. While the observed sequence biases may have resulted from structural constraints, the recovered clones represented only a small fraction of sequences predicted to form the same structure as the natural zipcode (data not shown), suggesting that secondary structure alone cannot mediate She2/3p recognition.

In addition to the over-representation of bases in the recognition motif, we also detected a bias towards the adenosine at the 5' end of the loop (A₁₂₈₇) and a stronger requirement for the C₁₂₈₉G dinucleotide upstream of the recognition motif (Fig. 9b, Table 6). Olivier et al. recently reported that the C₁₂₈₉GA triplet was essential for She2p binding (12); our results supported the importance of these bases as well as the downstream C₁₂₉₂G. Taken together, our results show that a repeated CG dinucleotide promotes She-complex binding: the consensus sequence, by base frequency, of positions 1289-93 of E2Bmin was CGACG, and CGACGA was most frequently selected in the context of YLR434-2. However, the CG dinucleotide followed by adenosines occurs most frequently in natural zipcodes, and this pattern is sufficient for bud-localization.

The sequence and structural requirements in the stem of E2Bmin were also analyzed by randomization and selection. The bias towards base-pairing was strongest at the second position from the top of the loop, whereas the base of the stem was paired only somewhat more often than was expected at random (Fig. 9c). Although targeted mutagenesis had revealed weak sequence preferences in the two loop-proximal base pairs, no biases were observed by randomization/ selection (Fig. 9c, Table 7), possibly because 3-AT selection does not discriminate between modest differences in 3-Hybrid

activity (22). Surprisingly, we recovered a bias towards the C₁₂₈₃C dinucleotide in the 5' strand of the stem and a weaker bias for G₁₃₀₀ (Fig. 9c, Table 8). The bias towards this guanosine likely results from the need to base-pair with C₁₂₈₄. These results further support our conclusion that stems can contribute both sequence and structural information for She-complex recognition.

Sequence requirements in the 3' tail were also revealed. Although Olivier et al. reported that C₁₃₀₂ was essential for She2 binding (12), only a modest bias towards this cytosine was detected (Fig. 9c, Table 8). 11 out of 12 clones that contained substitutions at this position had a UC dinucleotide immediately upstream, even though this pattern was not observed in native zipcodes lacking an analogous cytosine. It is apparent that the requirements for She-complex recognition are flexible, and that the cytosine described by Olivier et al. is not essential for all zipcodes.

Using the requirements elucidated by the mutational and randomization analyses, we sought to identify zipcodes in other localized RNAs. One candidate zipcode (bases 798-839 of *MID2*), which contains a single-stranded ACGAAAU motif adjacent to a stem and an adenosine 6 bases upstream, was localized above background levels (in 65-70% of budded cells), but less efficiently than other zipcodes isolated by 3-Hybrid assay. Candidate zipcodes in *IST2* and *BROI*, however, failed to be localized above background levels (data not shown). Additionally, WSC2C was the only isolated zipcode that did not contain the recognition motif in a single-stranded region and required two stem-loops for WT activity (Fig. 11). These results suggest that RNA recognition by She2/3 is complex, and that the current knowledge of the binding requirements and/ or the prediction tools are insufficient for accurately identifying new zipcodes.

Discussion:

We have employed a high-throughput selection for mapping She-complex binding sites in RNA targets. This methodology uses NRR to prepare DNA encoding localized RNAs, followed by 3-Hybrid selection to identify small fragments containing binding sites. Unlike other *in vitro* evolution techniques, NRR does not alter WT binding sites, making it easier to deconvolute the sequences after selection. Secondly, NRR covers sequence space efficiently because every starting pool contains a She2/3-binding site, eliminating the need to sample every nucleotide at every position and thus generating positive results from low-complexity libraries. Unlike conventional deletion mapping approaches, NRR samples all orientations and connectivities of input sequences.

By subjecting the NRR-derived pool to an *in vivo* 3-Hybrid selection, we could recover potentially lower-affinity and lower-abundance library members which may be missed by *in vitro* SELEX-style selection or candidate mutagenesis approaches. At the same time, the 3-Hybrid selection resulted in a low rate of false positives, since higher-abundance library members did not have a significant selective advantage. Finally, the *in vivo* selection ensured that the She proteins retained any post-translational modifications that may be necessary for WT activity.

Complex sequence and structural features mediate She2/3 binding:

Initial analysis of the NRR-derived zipcodes revealed a conserved single-stranded, 7-base motif lying proximal to a duplex region. Targeted mutagenesis confirmed that the motif sequence was necessary in different zipcodes for RNA transport and for direct binding to She2/3. The structural context of the motif was also important

for She-complex recognition: positioning the motif in a duplex abolished activity, and increasing the distance between the motif and adjacent stem decreased activity in three out of four zipcodes. A simple sequence motif stabilized by surrounding secondary structure appears to be a common theme of many protein binding sites in mRNAs, e.g. the Smg binding site in *nos* RNA (23). The She2/3 recognition site defined in this work expands on the CGA triplet reported by Olivier et al. (12) by virtue of a larger set of zipcodes which allowed us to identify the more degenerate bases downstream of the triplet as part of the recognition site. An additional single-stranded cytosine defined by Olivier et al. does not appear to be essential for She-complex recognition, since several natural zipcodes do not contain this nucleotide.

Quantitative analysis (by randomization/ selection) of the nucleotide requirements for She-complex binding contributed to a more thorough description of the RNA-protein interaction. Nucleotides at the 5' end of the motif, particularly a CG dinucleotide, were most important for binding, while the 3' adenosines made a weaker contribution. All natural zipcodes contained an adenosine following the CG dinucleotide, and this base was strongly favored in 2 out of 3 randomization experiments, suggesting that it too plays a major role in binding. Bases outside of the conserved motif also facilitated She-complex binding: some bases in the stem and 3' tail of E2Bmin were over-represented in the selected clones even though these sequences were not present in other zipcodes and one zipcode (YLR434-2) did not contain essential stem sequences.

The randomization/ selection experiments revealed an unexpected plasticity in the sequence requirements for She-complex recognition. When the four adenosines at the 3' end of the E2Bmin motif were held constant, there was only a weak bias for the upstream

CG dinucleotide; but when these adenosines were allowed to vary, the CG dinucleotide was strongly required, suggesting that some motif bases can bypass the requirement for others. Surprisingly, the two CG dinucleotides in E2Bmin do not function redundantly, as the requirement for the downstream CG was strongest when the upstream CG was invariable. A second example of sequence flexibility is that a UC dinucleotide can suppress mutations of a downstream cytosine identified by Olivier et al. (12) as essential for She2p binding. Some of these context-dependent effects may result from the RNA adopting a sub-optimal fold upon binding She2/3. The extensive sequence and structural plasticity, however, suggests that the She-complex recognizes a precise three-dimensional structure in its target RNAs—the complex may bind specifically to the key CG dinucleotide, with the surrounding bases simply maintaining the required structure.

One goal of defining a minimal RNA motif is to generate a predictive model whereby zipcodes could be identified in other RNAs *in silico*. We found that the core motif appears in She2/3 targets as well as in other RNAs known not to be localized, confirming that the motif alone does not confer specificity to the RNA-protein interaction. When the motif as well as other accessory features (e.g. an upstream adenosine and/ or a cytosine six nucleotides away from the motif) was used to identify new zipcodes, many localized RNAs did not contain any sequences that fit these criteria. From our analyses of known zipcodes, we conclude that RNA recognition likely involves complex structural features which cannot be appreciated using current tools of searching linear sequences and prediction of secondary structures. Thus, accurate prediction of zipcodes in other localized RNAs awaits a three-dimensional structure of the She-complex bound to a target RNA as well as methods for predicting this structural fold in

other RNAs. Meanwhile, the combination of NRR and 3-Hybrid selection provides a rapid and accurate way to isolate *bona fide* localization signals, and additional minimized zipcodes will aid in elucidating the range of sequences/ structures bound by the She-complex.

Acknowledgements:

We thank M. Wickens for providing yeast strains and plasmids for the 3-Hybrid assay. We also thank David Liu, Josh Bittker, and Jane Liu for advice on NRR, Joel Credle for assistance with sequencing, and members of the DeRisi lab for comments on the manuscript. This work was supported by grants from NSF (AJ), The David and Lucille Packard Foundation (JLD), the Searle Scholars Program (PAT), the Jane Coffin Childs Memorial Fund for Medical Research (KAS), and a National Institutes of Health Grant 38499 (RDV).

Table 1: Summary of elements identified by NRR/ 3-Hybrid selection. Coordinates indicate the smallest overlapping fragment common to all sequences isolated for each zipcode. Nucleotides are numbered with from the adenosine of the start codon as +1. * sequences derived from *ASH1*. When multiple fragments were contained in one clone, the fragments are listed in 5' to 3' order. Fragments in italics were cloned in the antisense orientation. The length of each clone is given in nucleotides. Activity in the 3-Hybrid assay was assessed by highest 3-AT concentration at which the sequence was recovered. \pm = 1mM, + = 5mM, ++ = 10mM, +++ = 15mM 3-AT. Also shown is the number of recovered clones containing the indicated sequence. %Localized refers to the percent of cells with exclusively bud-localized RNA. N/D: not determined.

| Zipcode | Coordinates | Length (nt) | 3-Hybrid Activity | #times recovered | %Localized |
|----------|------------------|-------------|-------------------|------------------|------------|
| *E1min | 635-683 | 49 | +++ | 8 | >90 |
| *E2Bmin | 1279-1314 | 36 | +++ | 11 | >90 |
| *Umin | 1766-1819 | 54 | + | 2 | >90 |
| *other | 1684-1719R | 36 | ++ | 1 | N/D |
| WSC2N | 418-71 | 54 | ++ | 14 | >90 |
| WSC2C | 1313-84 | 72 | ++ | 6 | >90 |
| ERG2N | 180-250 | 71 | ++ | 24 | >90 |
| DNM1N | 605-805 | 201 | + | 1 | 70-80 |
| DNM1C | 1656-1752 | 97 | + | 1 | >90 |
| SRL1C | 419-596 | 178 | + | 6 | >90 |
| YLR434-1 | [21-55][195-209] | 50 | + | 15 | 70-80 |
| YLR434-2 | [138-186][56-90] | 76 | + | 11 | >90 |
| TPO1N | 2-178 | 177 | \pm | 6 | 70-80 |
| CPS1CR | 1305-1456R | 152 | + | 1 | <60 |

Table 2: Coordinates of all clones isolated by 3-Hybrid selection, and number of times each clone was recovered from independent yeast transformants. Nucleotides are numbered with the adenosine of the start codon as +1. All isolates of SRL1C contained a deletion of base 458. All isolates of TPO1N contained a T80C mutation. “R” indicates that the fragment was recovered in the antisense orientation.

| Gene | Element Name | Coordinates | # times recovered |
|-------------|------------------------------------|------------------------------|-------------------|
| <i>ASH1</i> | E1 | 611-87 | 1 |
| | | 614-739 | 1 |
| | | 620-87 | 2 |
| | | 620-91 | 1 |
| | | 624-87 | 1 |
| | | [1327-42R]-[624-87] | 1 |
| | | 635-83 | 1 |
| | | E2B | 1210-1323 |
| | 1218-1314 | | 1 |
| | [1266-1314]-[354-402R] | | 1 |
| | 1267-1332 | | 1 |
| | 1270-1332 | | 1 |
| | 1273-1328 | | 2 |
| | 1273-1338 | | 1 |
| | [1689-1727]-[746-788R]-[1276-1326] | | 1 |
| | 1279-1323 | | 1 |
| | 1279-1332 | | 1 |
| | U | 1750-1853 | 1 |
| | | [1-5]-[1766-1819]-[859-911R] | 1 |
| | other | 1684-1719R | 1 |
| <i>ERG2</i> | ERG2N | 133-299 | 1 |
| | | 138-267 | 2 |
| | | 139-367 | 1 |
| | | 146-271 | 2 |
| | | 146-328 | 1 |
| | | 158-250 | 2 |
| | | 158-269 | 1 |
| | | 158-271 | 1 |
| | | 158-289 | 1 |

Table 2, continued

| Gene | Element Name | Coordinates | # times recovered |
|---------------|--------------|-------------------------|-------------------|
| <i>ERG2</i> | ERG2N cont'd | 158-291 | 1 |
| | | 158-299 | 4 |
| | | 158-328 | 1 |
| | | 158-355 | 1 |
| | | 160-256 | 1 |
| | | 160-271 | 1 |
| | | 160-303 | 1 |
| | | 171-256 | 1 |
| | | 180-328 | 1 |
| <i>WSC2</i> | WSC2N | 418-520 | 2 |
| | | 412-486 | 1 |
| | | [1121-1161R]-[418-510F] | 6 |
| | | 415-486 | 1 |
| | | 409-486 | 1 |
| | | 415-484 | 1 |
| | | 394-471 | 1 |
| | | 388-487 | 1 |
| | | WSC2C | 1313-1384 |
| | 1278-1384 | | 2 |
| | 1278-1418 | | 1 |
| | 1278-1391 | | 1 |
| | 1354-1512 | | 1 |
| | | | |
| | <i>SRL1</i> | SRL1C | 419-596 |
| 419-599 | | | 1 |
| 419-597 | | | 1 |
| 419-633 | | | 1 |
| 419-598 | | | 1 |
| <i>DNM1</i> | DNM1C | 1656-1752 | 1 |
| | DNM1N | 605-805 | 1 |
| <i>CPS1</i> | CPS1CR | 1305-1456R | 1 |
| <i>TPO1</i> | TPO1N | 2-178 | 6 |
| <i>YLR434</i> | YLR434-1 | [21-55R]-[194-209] | 15 |
| | YLR434-2 | [137-186R]-[56-80R] | 11 |

Table 3: Sequences of all WT and engineered mutant zipcodes analyzed in this work.

Bases identified by MEME analysis are in green, mutations in red, and deletions are indicated by dashes.

| RNA | Sequence |
|-----------------------------------|---|
| E1min (WT) | AAUAC GCGAAGA AGUGGCUCAUUUCAAGCCAUAAGUAUACCC AAACUC |
| E1Δmotif | -----GUGGCUCAUUUCAAGCCAUAAGUAUACCC AAACUC |
| E1shifted motif | AAUAC GCGAAGA CC AGUGGCUCAUUUCAAGCCAUAAGUAUAC CCAAACUC |
| E1mutA | --AAUAC GCGAAGA AGUGGCUCAUUUCAAGCCAUAAGUAUAC CCAAACUC |
| Umin (WT) | GAUACAUGGAUAACUGAAUCUCUUUCAACUAAUAAGAGAC AUUA UC ACGAAACA |
| UΔmotif | GAUACAUGGAUAACUGAAUCUCUUUCAACUAAUAAGAGAC AUUA UC----- |
| Ushifted motif | GAUACAUGGAUAACUGAAUCUCUUUCAACUAAUAAGAGAC AUUA UC AUCGAAACA |
| UmutA | GAUACAUGGAUAACUGAAUCUCUUUCAACUAAUAAGAGAC GUUA UC ACGAAACA |
| WSC2Nmin (WT) | AGUUCAAAA ACGUCCACGAAAU UGGAC ACGAAACU |
| WSC2Nmut1 | AGUUCAAAA ACGUCCACGAAAU UGGAC ACCCGGCU |
| WSC2Nmut2 | AGUUCAAAA ACGUCCACUCUU UGGAC ACGAAACU |
| WSC2Nmut3 | AGUUCAAAA ACGUCCACUCUU UGGAC---- CCCGGCU |
| YLR434-2min (WT) | GAUAUAGAUCCAAAGAAAUCU GCGAAAA AUUUU |
| YLR434-2Δmotif | GAUAUAGAUCCAAAGAAAUCU AUAG ----- |
| YLR434-2mut stem1 | GAUA GUCUA CCAAAG AAUAGAUCGAAAA AUUUU |
| YLR434-2mut stem2 | GAUA GUCUA CCAAAGAAAUCU GCGAAAA AUUUU |
| YLR434-2mut stem3 | GAUAUAGAUCCAAAG AAUAGAUCGAAAA AUUUU |
| YLR434-2 double stranded motif | AAAAUUUUUCGU AGAUCCAAAGAAAUCU GCGAAAA AUUUU |
| YLR434-2shifted motif | GAUAUAGAUCCAAAGAAAUCU CCGCGAAAA AUUUU |
| YLR434-2mutA | GAUAUAGAUCCAAAG U AAUCU GCGAAAA AUUUU |
| E2Bmin (WT) | CCCTCC ACACCGACGAAAA GUGGCAAGAUGAGAUCA |
| E2Bmut motif | CCCTCC ACACCGUGCGUUC GUGGCAAGAUGAGAUCA |
| E2B flip stem | CC GAGGUGA CCGACGAAAA CACCAUC AUGAGAUCA |
| E2B flip bp1 | CCT G ACACCG ACGAAAA GUG CCAAG |
| E2B flip bp3 | CCTCC U ACACCG ACGAAAA ACGCAAG |
| E2B flip bp4 | CCTCC AG ACCG ACGAAAA CUGGCAAG |
| E2B shifted motif | CCTCC ACACCGACGAAAA UAUGGCAAG |
| E2BmutA | CCTCC G ACACCG ACGAAAA GUGGCAAG |

Table 4: Sequences recovered after randomization of bases 1291-97 of E2Bmin and selection for interaction with She3p. Sequences of bases 1291-97 are shown; those recovered more than once are indicated.

| Sequence | # times recovered | Sequence | # times recovered |
|----------|-------------------|----------|-------------------|
| ACGCTAA | x2 | ACGTAGA | |
| AGAGTAC | | ACGCGAT | x2 |
| ACGCATT | | AAGCACT | |
| ACGTACAC | | ACCAGAA | x2 |
| ACGAAGA | x2 | ACGTAAT | x2 |
| ACGCTTT | | ACGCAAA | x2 |
| ACGCAAC | | ACGTGAC | x2 |
| ACCTACG | | ACGCACA | x2 |
| ACGAAAC | | ACGCGTA | |
| AAGTCTT | | ACGAGAA | |
| ATGTGAA | | ACGTCTT | |
| ACGAATG | x2 | AAGTACA | |
| ACGCTTC | x2 | ACGCGAC | |
| ACGTCAA | x2 | AAGTCAA | |
| ACGAAAT | | ACGTATC | |
| ACGTCTC | | ACGCTCA | |
| ACGCAAT | x2 | ACGATTC | |
| ACGCTAT | | AAGCAAA | |
| ACGTATA | | ATACTAA | |
| ACGTATT | x2 | AAGTAAC | |
| AAGTAAA | | ACGCCTT | |
| ACGTTTT | | | |

Table 5: Sequences recovered after randomization of bases 23-28 of YLR434-2 and selection for interaction with She3p. Sequences of bases 23-28 are shown; those recovered more than once are indicated.

| Sequence | # times recovered |
|----------|-------------------|
| CGAATA | |
| CGATAC | |
| CGAAGC | x2 |
| CGATGA | |
| CGAAGT | x2 |
| CGAACT | |
| CGTATC | x2 |
| CGACGC | x2 |
| CGACGA | x4 |
| CAAATC | |
| CTACGT | x2 |
| CGACTT | |
| CGAAGA | x2 |
| CGCGAT | |
| CACATC | |
| ACAGAT | |
| CGAACA | |
| CGAGAT | x2 |
| CACAAT | x2 |
| CTGAAT | |
| CGACGT | |
| CTTAAT | |
| CAACGT | |
| CGAGGC | |
| CGTTAT | |
| CGATAA | |

Table 6: Sequences recovered after randomization of bases 1287-93 of E2Bmin and selection for interaction with She3p. Sequences of bases 1287-93 are shown; sequences recovered more than once are indicated. WT sequence is also indicated.

| Sequence | # times recovered |
|----------|-------------------|
| TCCGATT | x3 |
| AACGACG | x6 |
| GTCGATG | |
| ACCGATT | x2 |
| CCCGACG | x3 |
| ACGGAAT | |
| ACCGACG | x5 WT |
| AGCGAAG | x3 |
| ATCGATG | |
| TGCGATC | |
| TGCGAAT | |
| AACGAAG | |
| GACGATT | |
| AGCGAAA | |
| AACGATG | x4 |
| GTCGAAG | x2 |
| ATCGACG | x3 |
| AGCGGAA | |
| GACGAAT | |
| GACGACG | |
| GACGAAG | |
| AACGTAT | |
| AGCGAAT | |
| TGCGAAC | x2 |
| GACGGCG | |
| AGCGACG | |
| AACGAAT | |
| TCCGACT | |
| ACCGAAA | |
| ATCGAAG | |

Table 7: Sequences recovered after randomization of bases 1285-6 and 1298-9 of E2Bmin and selection for interaction with She3p. Each row represents a single selected clone; clones recovered more than once are indicated. Headings indicate the coordinate of each base; only bases at the randomized positions are shown.

| 1285 | 1286 | 1298 | 1299 | #times recovered |
|------|------|------|------|------------------|
| G | C | A | C | |
| A | T | T | T | |
| T | C | T | G | |
| C | T | A | G | x3 |
| T | T | A | A | x3 |
| C | A | T | G | x2 |
| G | C | G | G | x3 |
| C | A | C | T | |
| C | G | C | G | |
| G | C | G | C | |
| T | T | G | A | x2 |
| C | T | G | T | |
| A | A | T | T | |
| G | G | C | C | |
| T | A | T | A | x2 |
| A | C | A | C | |
| C | A | C | G | |
| C | T | T | G | |
| T | A | A | A | |
| T | A | C | A | |
| C | C | G | G | |
| G | C | G | T | |
| C | T | C | G | |
| A | G | C | T | |
| G | T | A | C | |

Table 8: Sequences recovered after randomization of bases 1283-4 and 1300-03 of E2Bmin and selection for interaction with She3p. Each row represents a single selected clone; clones recovered more than once are indicated. Headings indicate the coordinate of each base; only bases at the randomized positions are shown. WT sequence is indicated.

| 1283 | 1284 | 1300 | 1301 | 1302 | 1303 | #times recovered |
|------|------|------|------|------|------|------------------|
| A | G | C | T | C | A | |
| C | C | T | C | A | T | |
| C | C | T | C | G | C | x2 |
| A | C | G | G | C | A | |
| A | C | G | G | C | T | |
| C | C | T | C | T | G | x2 |
| C | T | A | G | C | A | x2 |
| C | T | A | G | C | T | |
| T | C | G | A | C | A | |
| C | C | T | C | A | G | x2 |
| C | C | T | T | C | T | |
| A | C | G | T | C | A | x2 |
| A | G | G | G | G | G | |
| A | C | T | C | T | C | |
| A | G | A | T | C | G | |
| C | C | G | G | C | A | x4 WT |
| T | C | G | A | C | G | |
| C | C | T | T | C | C | |
| C | T | T | C | T | A | |
| C | C | C | T | C | T | |
| A | C | G | A | C | G | x2 |
| C | C | T | C | T | T | |
| C | C | T | C | T | C | |
| C | C | G | G | C | G | x3 |
| A | T | A | T | C | G | |
| G | A | T | C | C | G | |
| A | T | A | T | C | T | |
| T | T | A | A | C | A | |
| C | T | A | G | C | C | |

Figure 1: 3-Hybrid scheme for selection of She3-interacting RNA fragments. (A) Schematic of 3-Hybrid assay and representation of *ASH1* NRR library members (B) prior to and (C) following 3-Hybrid selection. Each arrow represents a fragment from *ASH1*. The direction of the arrowhead indicates whether the fragment is expressed in the sense (right) or antisense (left) orientation from the 3-Hybrid RNA expression vector. The position of each arrow corresponds to the location of the fragment within the gene, and arrow colors indicate the connectivity of the fragments in the clone. Clones recovered in more than one independent yeast transformant are indicated.

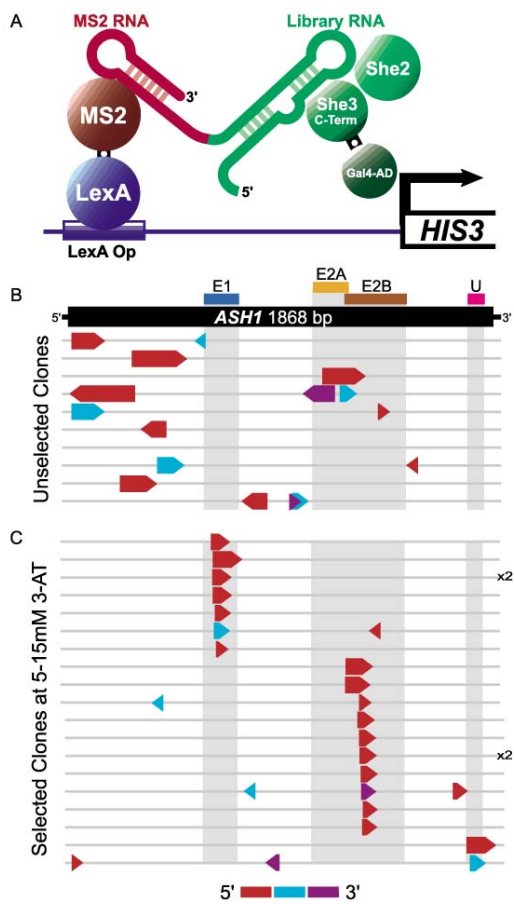


Figure 2: Interaction of RNAs with the carboxyl-terminus of She3p in the 3-Hybrid system. β -galactosidase activities are shown for WT RNAs depicted in Fig. 2, as well as empty vector and IRE controls. *she2* indicates that the *SHE2* ORF was deleted in the 3-Hybrid host strain.

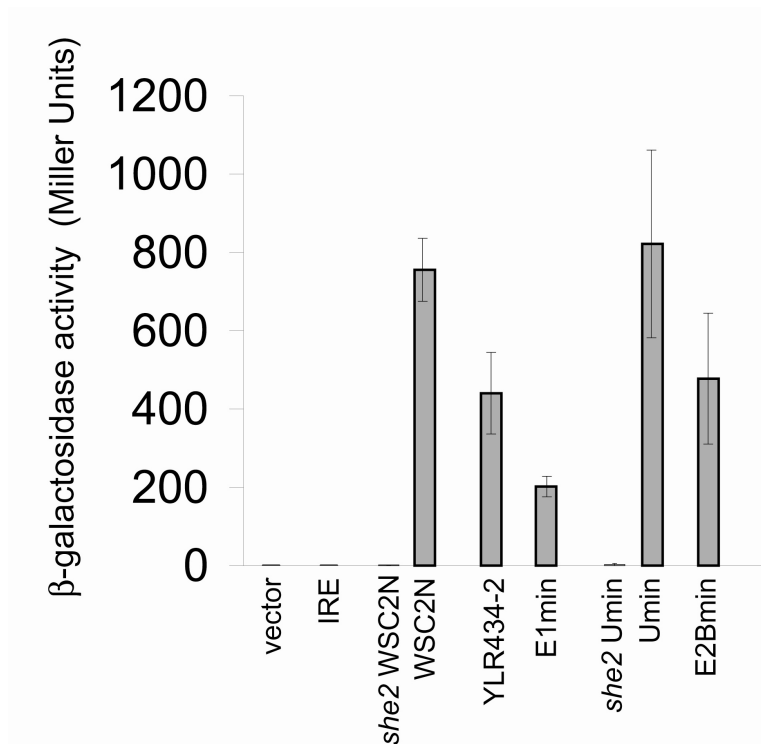


Figure 3: Sequences and predicted structures of fragments shorter than 100 nt isolated by NRR/ 3-Hybrid analysis. Bases identified by MEME analysis are green. (A) WSC2N, (B) YLR434-2, (C) E2Bmin, (D) E1min, (E) Umin, (F) ERG2N, (G) YLR434-1, (H) WSC2C, (I) DNM1C. Bases are numbered with the adenosine of the start codon as +1, with the exception of YLR434-1 and YLR434-2, which are numbered with the 5' base as +1. Insets contain representative GFP-RNA localization images. RNA particles are cytoplasmic; excess, unbound U1A-GFP is sequestered in the nucleus.

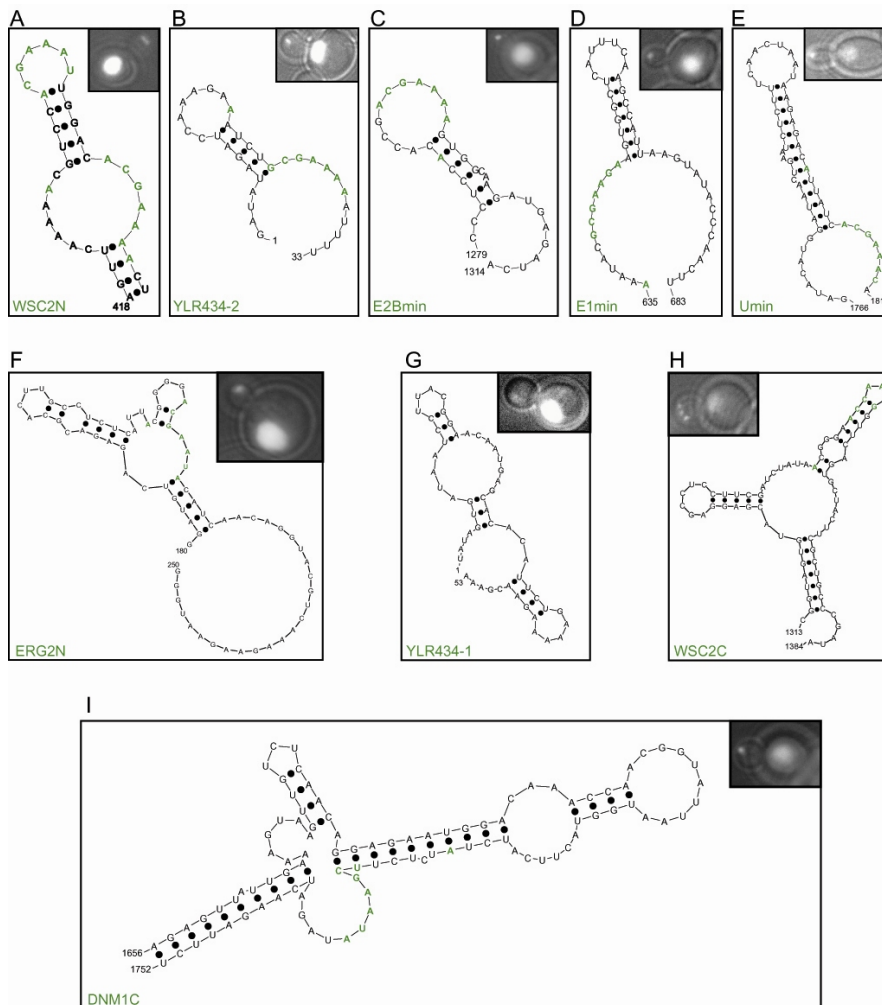


Figure 4: The recognition motif mediates She2/3-dependent localization and

binding. (A) Localization ability and (B) 3-Hybrid β -galactosidase activity of zipcodes in

Fig. 2 containing WT bases or mutations in the recognition motif. “double-stranded motif” indicates that the recognition motif is in an ectopic duplex. Dashed line in (A)

indicates the threshold below which RNAs were considered unlocalized. (C) *In vitro*

binding of She2p and She3p to E2Bmin. RNA mobility shift assay consists of WT or mutant RNA lacking the recognition motif with increasing concentrations of purified

She2p-HA and His-She3p carboxyl terminus. (D) WT and mutant motifs sequences used

in a-c. Motif bases are green and mutations red.

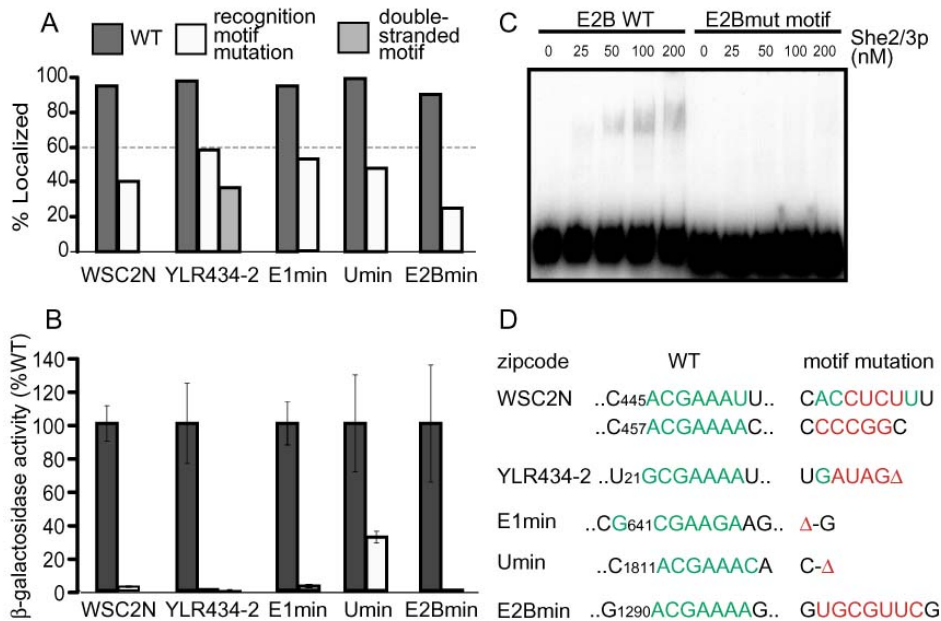


Figure 5: Two copies of the recognition motif in WSC2N are partially redundant.

(A) Sequences and predicted secondary structures for WT and mutant WSC2N RNAs tested for (B) localization and (C) 3-Hybrid β -galactosidase activity. In (A), bases identified by MEME analysis are green and mutations in red. Insets contain representative GFP-RNA localization images. RNA particles are cytoplasmic; excess, unbound GFP is sequestered in the nucleus.

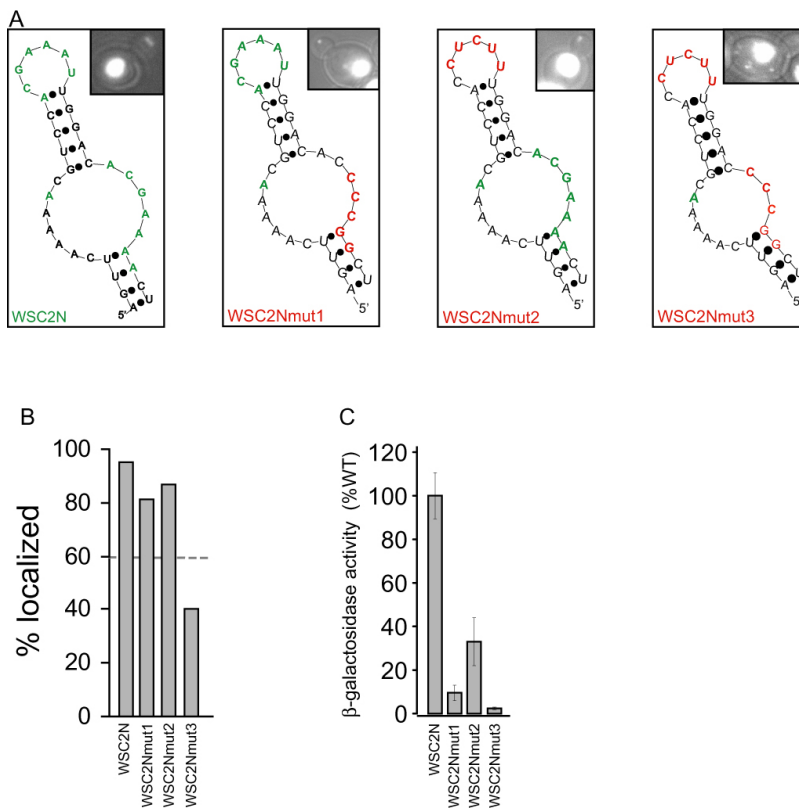


Figure 6: Gel mobility shift assays as described in Figure 6 using WT and recognition motif mutant RNAs. (A) E1min, (B) Umin, (C) YLR434-2.

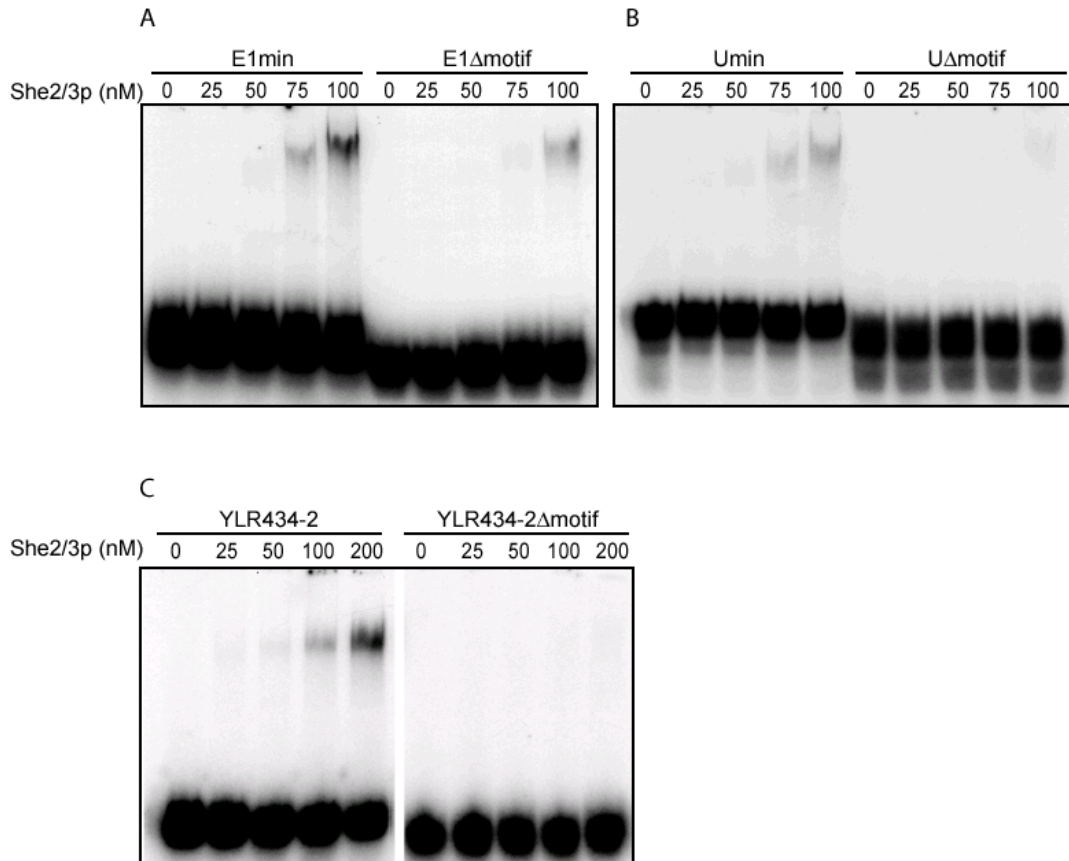


Figure 7: Context-dependency of recognition motif is revealed by mutational analysis. (A) 3-Hybrid β -galactosidase activity of zipcodes bearing mutations in the upstream adenosine (mutA) or 2 nt insertions between the motif and adjacent helix (shifted motif). Mutations are defined for each zipcode. Motif bases are in green, insertions red, and duplex bases are bold. (B) β -galactosidase activity of E2Bmin sequences containing mutations in the stem. “Flip” indicates that each base of the indicated pair was changed to its partner. The “flip stem” RNA contains compensatory mutations along the entire stem. Bases-pairs are numbered with the bottom of the stem as 1.

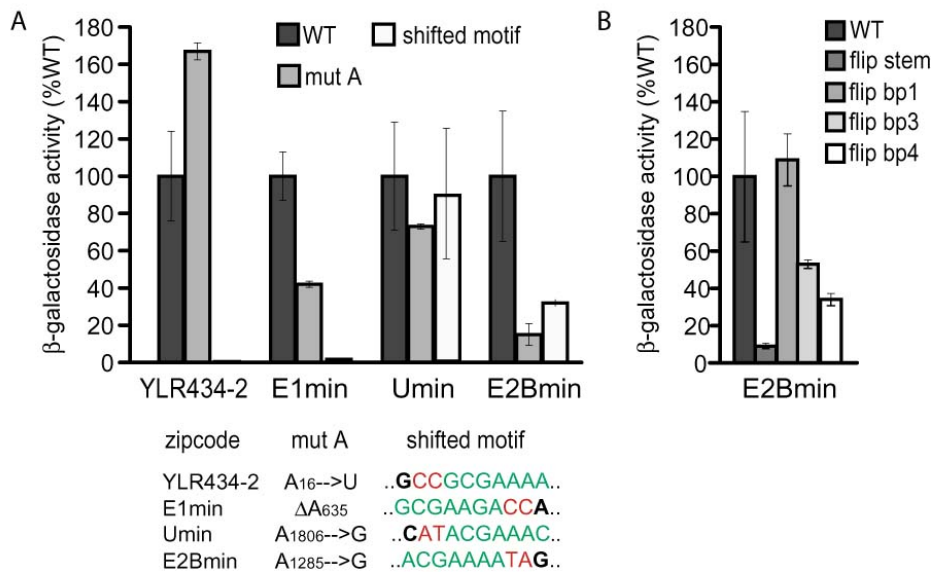


Figure 8: β -galactosidase activities of WT YLR434-2 and sequences containing mutations in the stem. In YLR434-2mut stem1, the sequence of each strand was exchanged with that of the opposite strand, preserving all base pairs. In YLR434-2mut stem2, the 5' strand of the stem was changed to its complement. The 3' strand was similarly mutated in YLR434-2mut stem3 (see Table 3).

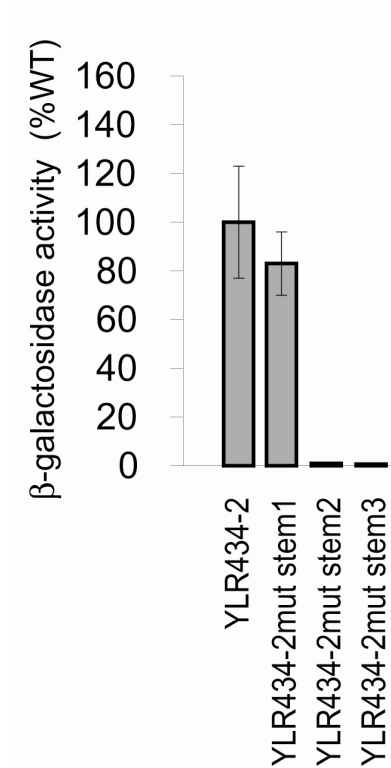


Figure 9: Predicted secondary structure for E2Bmin and sequence logos derived from randomization and 3-Hybrid selection of bases (A) 1291-7, (B) 1287-93, or (C) 1283-6 and 1298-1303. The height of each letter is proportional to the fraction of the observed frequency relative to the expected frequency at each position (24, 25). The color of each dot in (C) indicates the frequency of base-pairing among the selected clones.

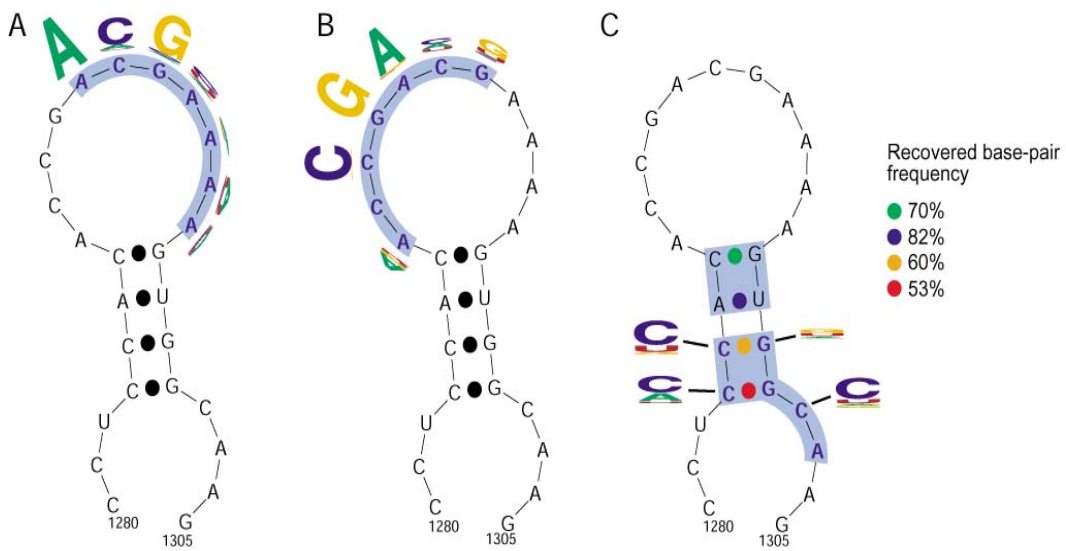


Figure 10: Randomization and 3-hybrid selection of recognition motif in YLR434-2 reveals sequence requirements for She2/3 recognition. Sequence and predicted structure of YLR434-2, and sequence logo derived from randomization and selection of bases 23-28 for interaction with She3p.

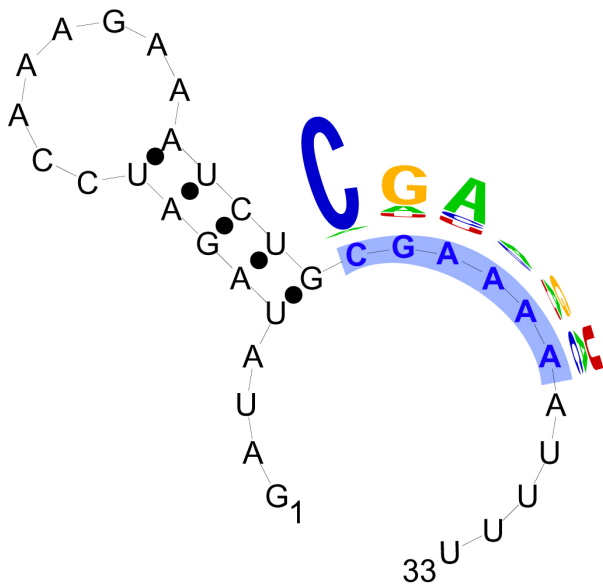
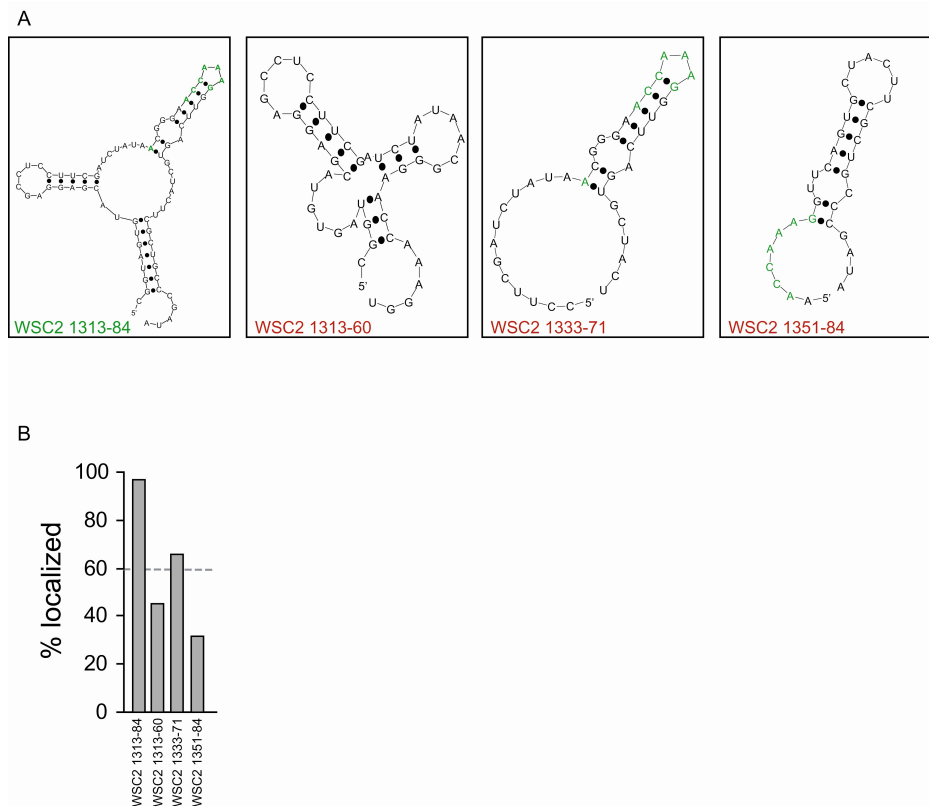


Figure 11: Two stem-loops are required for full activity of zipcode WSC2C. (A)

Sequences and predicted structures of full-length WSC2C and WSC2C fragments tested for (B) localization *in vivo*.



References:

1. Oleynikov, Y. and Singer, R.H. (1998) RNA localization: different zipcodes, same postman? *Trends Cell Biol*, **8**, 381-383.
2. Betley, J.N., Frith, M.C., Graber, J.H., Choo, S. and Deshler, J.O. (2002) A ubiquitous and conserved signal for RNA localization in chordates. *Curr. Biol*, **12**, 1756-1761.
3. Kim-Ha, J., Webster, P.J., Smith, J.L. and Macdonald, P.M. (1993) Multiple RNA regulatory elements mediate distinct steps in localization of oskar mRNA. *Development*, **119**, 169-178.
4. Macdonald, P.M. and Kerr, K. (1998) Mutational analysis of an RNA recognition element that mediates localization of bicoid mRNA. *Mol. Cell. Biol*, **18**, 3788-3795.
5. Gautreau, D., Cote, C.A. and Mowry, K.L. (1997) Two copies of a subelement from the Vg1 RNA localization sequence are sufficient to direct vegetal localization in *Xenopus* oocytes. *Development*, **124**, 5013-5020.
6. Shepard, K.A., Gerber, A.P., Jambhekar, A., Takizawa, P.A., Brown, P.O., Herschlag, D., DeRisi, J.L. and Vale, R.D. (2003) Widespread cytoplasmic mRNA transport in yeast: identification of 22 bud-localized transcripts using DNA microarray analysis. *Proc. Natl. Acad. Sci. U.S.A*, **100**, 11429-11434, 10.1073/pnas.2033246100.
7. Böhl, F., Kruse, C., Frank, A., Ferring, D. and Jansen, R.P. (2000) She2p, a novel RNA-binding protein tethers ASH1 mRNA to the Myo4p myosin motor via She3p. *EMBO J*, **19**, 5514-5524, 10.1093/emboj/19.20.5514.

8. Takizawa,P.A. and Vale,R.D. (2000) The myosin motor, Myo4p, binds Ash1 mRNA via the adapter protein, She3p. *Proc. Natl. Acad. Sci. U.S.A*, **97**, 5273-5278, 10.1073/pnas.080585897.
9. Long,R.M., Gu,W., Lorimer,E., Singer,R.H. and Chartrand,P. (2000) She2p is a novel RNA-binding protein that recruits the Myo4p-She3p complex to ASH1 mRNA. *EMBO J*, **19**, 6592-6601, 10.1093/emboj/19.23.6592.
10. Gonzalez,I., Buonomo,S.B., Nasmyth,K. and von Ahsen,U. (1999) ASH1 mRNA localization in yeast involves multiple secondary structural elements and Ash1 protein translation. *Curr. Biol*, **9**, 337-340.
11. Chartrand,P., Meng,X.H., Singer,R.H. and Long,R.M. (1999) Structural elements required for the localization of ASH1 mRNA and of a green fluorescent protein reporter particle in vivo. *Curr. Biol*, **9**, 333-336.
12. Olivier,C., Poirier,G., Gendron,P., Boisgontier,A., Major,F. and Chartrand,P. (2005) Identification of a conserved RNA motif essential for She2p recognition and mRNA localization to the yeast bud. *Mol. Cell. Biol*, **25**, 4752-4766, 10.1128/MCB.25.11.4752-4766.2005.
13. Bittker,J.A., Le,B.V. and Liu,D.R. (2002) Nucleic acid evolution and minimization by nonhomologous random recombination. *Nat. Biotechnol*, **20**, 1024-1029, 10.1038/nbt736.
14. Bernstein,D.S., Buter,N., Stumpf,C. and Wickens,M. (2002) Analyzing mRNA-protein complexes using a yeast three-hybrid system. *Methods*, **26**, 123-141, 10.1016/S1046-2023(02)00015-4.

15. SenGupta,D.J., Zhang,B., Kraemer,B., Pochart,P., Fields,S. and Wickens,M. (1996)
A three-hybrid system to detect RNA-protein interactions in vivo. *Proc. Natl. Acad. Sci. U.S.A.*, **93**, 8496-8501.
16. Longtine,M.S., McKenzie,A., Demarini,D.J., Shah,N.G., Wach,A., Brachat,A., Philippsen,P. and Pringle,J.R. (1998) Additional modules for versatile and economical PCR-based gene deletion and modification in *Saccharomyces cerevisiae*. *Yeast*, **14**, 953-961, 10.1002/(SICI)1097-0061(199807)14:10<953::AID-YEA293>3.0.CO;2-U.
17. Gietz,R.D. and Woods,R.A. (2002) Transformation of yeast by lithium acetate/single-stranded carrier DNA/polyethylene glycol method. *Meth. Enzymol*, **350**, 87-96.
18. Mathews,D.H., Sabina,J., Zuker,M. and Turner,D.H. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol*, **288**, 911-940, 10.1006/jmbi.1999.2700.
19. Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res*, **31**, 3406-3415.
20. Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol*, **2**, 28-36.
21. Kruse,C., Jaedicke,A., Beaudouin,J., Bohl,F., Ferring,D., Guttler,T., Ellenberg,J. and Jansen,R. (2002) Ribonucleoprotein-dependent localization of the yeast class V myosin Myo4p. *J. Cell Biol*, **159**, 971-982, 10.1083/jcb.200207101.
22. Hook,B., Bernstein,D., Zhang,B. and Wickens,M. (2005) RNA-protein interactions in the yeast three-hybrid system: affinity, sensitivity, and enhanced library

- screening. *RNA*, **11**, 227-233, 10.1261/rna.7202705.
23. Crucs,S., Chatterjee,S. and Gavis,E.R. (2000) Overlapping but distinct RNA elements control repression and activation of nanos translation. *Mol. Cell*, **5**, 457-467.
24. Crooks,G.E., Hon,G., Chandonia,J. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res*, **14**, 1188-1190, 10.1101/gr.849004.
25. Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res*, **18**, 6097-6100.

Chapter 3: Recovery of Divergent Avian Bornaviruses from Cases of Proventricular Dilatation Disease: Identification of a Candidate Etiologic Agent

This chapter is a reprint from the following reference:

Kistler AL, Gancz A, Clubb S, Skewes-Cox P, Fischer K, Sorber K, Chiu CY, Lublin A, Mechani S, Farnoushi Y, Greninger A, Wen CC, Karlene SB, Ganem D, DeRisi JL. (2008). Recovery of divergent avian bornaviruses from cases of proventricular dilatation disease: identification of a candidate etiologic agent. *Virology* 481: 88-98.

Copyright © 2008, BioMed Central, Ltd.

Author contributions:

Amy Kistler participated in the conception, design, and coordination of the study, performed specimen extraction of specimens from Florida case/control study, array analyses for both sets of PDD case/control series, follow-up PCR screening and sequencing of samples and wrote the manuscript; Ady Gancz orchestrated and collected the PDD case/control specimens from Israel and coordinated the clinical and histopathology analyses, and nucleic acid extraction for samples from Israel, and participated in revising the manuscript; Susan Clubb orchestrated and collected the Florida PDD case/control specimens and oversaw the clinical and histopathologic analyses of these samples from Florida, and participated in revising the manuscript; Peter

Skewes-Cox carried out filtering and iterative BLAST analysis of ultra high throughput sequence data for ABV genome sequence recovery, participated in primer design and complete genome sequence recovery, and drafting the manuscript; Kael Fischer participated in array analysis, developed pipeline for ultra high throughput sequence analysis, and participated in design of filtering and iterative BLAST analysis; Katherine Sorber performed modified library preparation for ultra high throughput sequencing and participated in revising the manuscript; Charles Chiu performed ultra high throughput sequencing and participated in revising the manuscript; A Lublin, S Mechani, and Y Farnoushi participated in clinical evaluation, specimen collection and extraction of samples from Israel; Alex Greninger participated in extraction of specimens from Florida and follow-up microarray analysis and high throughput sequencing; Christopher Wen developed additional primers for PCR follow-up studies; SB Karlene assisted in the selection of the PDD case/control specimens from Florida and participated in review of clinical and histological status of cases and controls included in the study; Don Ganem and Joseph L. DeRisi oversaw the overall conception and design of the project and supervised all phases of its execution and the drafting and revision of the manuscript.

Joseph L. DeRisi, Thesis Advisor

Abstract:

Proventricular dilatation disease (PDD) is a fatal disorder threatening domesticated and wild psittacine birds worldwide. It is characterized by lymphoplasmocytic infiltration of the ganglia of the central and peripheral nervous system, leading to central nervous system disorders as well as disordered enteric motility and associated wasting. For almost 40 years, a viral etiology for PDD has been suspected, but to date no candidate etiologic agent has been reproducibly linked to the disease. Analysis of 2 PDD case-control series collected independently on different continents using a pan-viral microarray revealed a bornavirus hybridization signature in 62.5% of the PDD cases (5/8) and none of the controls (0/8). Ultra high throughput sequencing was utilized to recover the complete viral genome sequence from one of the virus-positive PDD cases. This revealed a bornavirus-like genome organization for this agent with a high degree of sequence divergence from all prior bornavirus isolates. We propose the name avian bornavirus (ABV) for this agent. Further specific ABV PCR analysis of an additional set of independently collected PDD cases and controls yielded a significant difference in ABV detection rate among PDD cases (71%, n=7) compared to controls (0%, n=14) ($P=0.01$; Fisher's Exact Test). Partial sequence analysis of a total of 16 ABV isolates we have now recovered from these and an additional set of cases reveals at least 5 distinct ABV genetic subgroups. These studies clearly demonstrate the existence of an avian reservoir of remarkably diverse bornaviruses and provide a compelling candidate in the search for an etiologic agent of PDD.

Introduction:

Proventricular dilatation disease (PDD) is considered by many to be the greatest threat to aviculture of psittacine birds (parrots). This disease has been documented in multiple continents in over 50 different species of psittacines as well as captive and free-ranging species in at least 5 other orders of birds (1-5). Most, if not all major psittacine collections throughout the world have experienced cases of PDD. It has been particularly devastating in countries like Canada and northern areas of the United States where parrots are housed primarily indoors. However, it is also problematic in warmer regions where birds are typically bred in outdoor aviaries. Moreover, captive breeding efforts for at least one psittacine which is thought to be extinct in the wild, the Spix's macaw (*Cyanopsitta spixii*), have been severely impacted by PDD.

PDD is an inflammatory disease of birds, first described in the 1970s as Macaw Wasting Disease during an outbreak among macaws (reviewed in (3)). PDD primarily affects the autonomic nerves of the upper and middle digestive tract, including the esophagus, crop, proventriculus, ventriculus, and duodenum. Microscopically, the disease is recognized by the presence of lymphoplasmacytic infiltrates within myenteric ganglia and nerves. Similar infiltrates may also be present in the brain, spinal cord, peripheral nerves, conductive tissue of the heart, smooth and cardiac muscle, and adrenal glands. Non-suppurative leiomyositis and/or myocarditis may accompany the neural lesions (6-9). Clinically, PDD cases present with GI tract dysfunction (dysphagia, regurgitation, and passage of undigested food in feces), neurologic symptoms (e.g. ataxia, abnormal gait, proprioceptive defects), or both (3). Although the clinical course of the disease can vary, the lesion is generally fatal in untreated animals (3).

The cause of PDD is unknown, but several studies have raised the possibility that PDD may be caused by a viral pathogen. Evidence for an infectious etiology stems from the initial outbreaks of Macaw Wasting Disease, and other subsequent outbreaks of PDD (2, 10). Reports of pleomorphic virus-like particles of variable size (30-250nm) observed in tissues of PDD affected birds (8) led to the proposal that paramyxovirus (PMV) may cause the disease; however, serological data has shown that PDD affected birds lack detectable antibodies against PMV of serotypes 1-4, 6, and 7, as well as against avian herpes viruses, polyomavirus, and avian encephalitis virus (3). Similarly, a proposed role for equine encephalitis virus in PDD has been ruled out (11). Enveloped virus-like particles of approximately 80nm in diameter derived from the feces of affected birds have been shown to produce cytopathic effect in monolayers of macaw embryonic cells (12), but to date no reports confirming these results or identifying this possible agent have been published. Likewise, adeno-like viruses, enteroviruses, coronaviruses and reoviruses have also been sporadically documented in tissues or excretions of affected birds (3, 13, 14) yet in each case, follow-up evidence for reproducible isolation specifically from PDD cases or identification of these candidate agents has not been reported. Thus, the etiology of PDD has remained an open question.

To address this question, we have turned to a comprehensive, high throughput strategy to test for the presence of known or novel viruses in PDD affected birds. We employed the Virochip, a DNA microarray containing representation of all viral taxonomy to interrogate 2 PDD case/control series independently collected on two different continents for the presence of viral pathogens. We report here the detection of a novel bornavirus signature in 62.5% of the PDD cases and none of the controls. These

bornavirus-positive samples were confirmed by virus-specific PCR testing, and the complete genome sequence has been recovered by ultra-high throughput sequencing combined with conventional PCR-based cloning.

Bornaviruses are a family of negative strand RNA viruses whose prototype member is Borna Disease Virus (BDV), an agent of encephalitis whose natural reservoir is primarily horses and sheep (15). Although experimental transmission of BDV to many species (including chicks (16)) has been described, there is little information on natural avian infection, and existing BDV isolates are remarkable for their relative sequence homogeneity. The agent reported here, which we designate avian bornavirus (ABV) is highly diverged from all previously identified members of the *Bornaviridae* family and represents the first full-length bornavirus genome cloned directly from avian tissue. Subsequent PCR screening for similar ABVs confirmed a detection rate of approximately 70% among PDD cases and none among the controls. Sequence analysis of a single complete genome and all of the additional partial sequences that we have recovered directly from the PDD case specimens suggests that the viruses detected in cases of PDD form a new, genetically diverse clade of the *Bornaviridae*.

Methods:

PDD case and control definitions, specimen collection, and RNA extraction for pan-viral microarray screening:

Two independent sets of PDD case and control specimens collected from two distinct geographic locations were independently prepared for pan-viral microarray screening and subsequent PCR screening. Sampling collection and inclusion criteria for

each set are described below. Detailed information on each sample, along with results from histology, microarray, and PCR assays are provided in Additional file 5: Summary of clinical and molecular data for specimens provided in this study.

United States PDD case/control series. Specimen collection: All specimens provided for initial screening were crop tissue biopsies obtained from live psittacine birds to be used as normal controls or multiple tissue samples collected from clinically diseased birds at the time of euthanasia. Specimens were collected from client-owned birds from approximately August 2006 to May 2008 (All samples collected by S. Clubb). All of these samples originated from the southeast region of Florida. Crop biopsy tissue was collected from live birds under isoflurane anaesthesia. Following routine surgical preparation and sterile technique, the skin was incised over the center of the crop. The crop tissue was exposed and a section of tissue removed taking care to include large visible blood vessels. Fresh crop biopsy tissue was trimmed into tissue slices < 5mm thick and submersed in RNAlater (Qiagen, Inc., USA, Valencia, CA) solution immediately upon harvest and frozen within 2 minutes of collection at -20°C to -80°C according to manufacturer's protocol, and held in this manner until shipped. A duplicate sample was fixed in 10% buffered formalin for routine histological examination with H & E stain. Time of frozen storage varied (2 weeks to 12 months) as samples were accumulated prior to shipping frozen. Clinically affected birds submitted as positives were euthanized under isoflurane anaesthesia and mixed tissues (proventriculus, ventriculus, heart, liver, spleen, kidneys, brain) were placed into RNAlater within 1 minute of death and frozen within 2 minutes of death. Duplicate samples were collected for histopathologic diagnosis of PDD.

Inclusion criteria: PDD-positive cases were required to meet the following criteria

- 1) Clinical history of characteristic wasting/malabsorption syndrome with dilation of the proventriculus and/or ventriculus and presence of undigested food in the stool and in most cases, a clinical history of ataxia or other CNS signs consistent with clinical PDD, and
- 2) histopathology confirming the presence of moderate to extensive lymphoplasmacytic ganglioneuritis affecting crop tissue and at least one of the following additional areas: proventriculus, ventriculus, brain, adrenal gland, or myocardium.

PDD-negative controls were required to be from birds with no evidence of lymphoplasmacytic neurogangliitis on histopathology derived either from 1) normal birds with no clinical history of PDD or no known exposure to PDD or 2) birds which died of other causes. Crop biopsies from samples from living birds classified as suspicious cases were also submitted. Suspicious cases were defined histologically as having lymphocytes and plasma cells surrounding neurons but not infiltrating into the neurons. An additional specimen derived from a live bird raised with two necropsy-confirmed PDD birds in Virginia was also collected for analysis. Here, only cloacal swab and blood specimens were available and the lack of histopathological confirmation and crop tissue excluded this specimen from the ABV- PDD association analysis. However, we did perform ABV PCR on these clinically suspicious specimens and the resulting viral sequences isolated were included in the subsequent comparative sequence analyses.

RNA extractions: For RNA extractions, specimens were thawed in RNALater, sliced into 0.5mm x 0.5mm pieces, transferred to 2ml of RNABee solution (Tel-Test, Inc., Friendswood, TX), homogenized with freeze thawing and scapel mincing, then extracted in the presence of chloroform according to manufacturer's instructions.

Resulting RNA was next incubated with DNase (DNA-free, Applied Biosystems/Ambion, Austin, TX) to remove any potential contaminating DNA present in the specimen.

Israeli case/control series. Specimen collection: Tissue samples were obtained from psittacine birds submitted to the Division of Avian and Fish Diseases, Kimron Veterinary Institute (KVI) Bet Dagan, Israel, for diagnostic necropsy between July 2004 and March 2008. A few additional specimens were obtained through private veterinarians. Some tissues were kept for nearly 4 years frozen either at -20°C or -80°C prior to testing, while others were fresh tissues from recent cases. The types of banked frozen tissue varied from case to case, while for some of the older cases only gastrointestinal content had been banked. Clinical histories for these birds were available from the submission forms or through communication with the submitting veterinarians. The results of ancillary tests performed at the KVI were available through the KVI computerized records.

Inclusion criteria: Only cases for which appropriate histological sections were available for inspection were considered for this study. These had to include brain and at least two of the following tissues: crop, proventriculus, ventriculus. The tissue- types examined for each bird for which specimens were provided are listed in Data File S1. PDD-positive cases were required to have evidence of lymphoplasmacytic infiltration of myenteric nerves and/or ganglia within one or more of the upper GI tract tissues mentioned above. These were all derived from birds that had been suspected to have PDD based on their clinical case histories and/or necropsy findings. PDD-negative controls had no detectable lesions and no evidence of non-suppurative encephalitis. For most birds in the PDD-negative group, a cause of death (other than PDD) has been

determined. Two birds that came from a known PDD outbreak, but showed only cerebral lymphoplasmacytic perivascular cuffing, were classified as 'suspicious'. These were excluded from the statistical analysis, as were all other birds for which a PDD status could not be clearly determined and classified as 'inconclusive' (e.g. due to poor tissue preservation, poor section quality, or scarcity of myenteric nerves within the tissues examined).

RNA extraction: When possible, a sample of brain as well as a combined proventricular/ventricular sample was prepared for RNA extraction for each bird. If not available, other tissues and/or gastrointestinal content were used (see Additional file 5: Summary of clinical and molecular data for specimens provided in this study). Frozen samples were allowed to thaw for 1-2 hours at room temperature prior to handling. Then, under a laminar flow biohazard hood and using aseptic technique, approximately 1cm³ of each tissue was macerated by two passages through a 2.5ml sterile syringe and transferred into sterile test tubes containing 4ml nuclease-free PBS. The content of the tubes was mixed by vortex for 30sec, and the tubes were placed overnight at 4°C. RNA extraction was performed on the following day, using either the QIAamp® viral RNA kit (Qiagen, Valencia, CA; batch1&2, specimens 1-8) or the TRI Reagent® kit (Molecular Research Center, Cincinnati, OH; all other specimens), following the manufacturers' instructions. The end product was either provided lyophilized (batches 1 and 2, samples 1-9) as a dry pellet, or re-suspended in 40ul nuclease-free water.

Virus chip hybridization experiments:

Microarray analysis of specimens was carried out as previously described (17).

Briefly, 50-200ng of DNase-treated total RNA from each sample was amplified and labelled using a random-primed amplification protocol and hybridized to the Virochip. Microarrays (NCBI GEO platform GPL3429) were scanned with an Axon 4000B scanner (Axon Instruments). Virochip results were analyzed using E-Predict (18) and vTaxi (K. Fisher et al., in preparation).

PCR primers for detection of avian bornaviruses:

Microarray-based Bornaviridae PCR primers. Initial PCR primers were generated based on two of the 70mer microarray probes with hybridization signal in the *Bornaviridae* positive arrays that localize to positions 3676-3745 and 4201-4270 of the *Bornaviridae* reference sequence [GenBank:NC_001607]. Subsequences within each of these probes (BDV_LconsensusF: 5'-CCTCGCGAGGAGGAGACGCCTC- 3' and BDV_LconsensusR: 5' CTGCTCTTGGCTGTGTCTGCTGC-3'; positions 3710-3729 and 4252-4230, respectively of the NCBI *Bornaviridae* reference sequence) that are 100% conserved across the 12 other fully sequenced bornavirus genome isolates in NCBI (huP2br [GenBank:AB258389], Bo/04w [GenBank:AB246670], No/98 [GenBank:AJ311524], H1766 [GenBank:AJ311523], He/80/FR [GenBank:AJ311522], V/FR [GenBank:AJ311522], virus rescue plasmid pBRT7-HrBDVc [GenBank:AY05791], CRNP5 [GenBank:AY114163], CRP3B [GenBank:AY114162], CRP3A [GenBank:AY114161], He/80 [GenBank:L27077], and V [GenBank:U040608]) were utilized for initial follow-up PCR and sequence confirmation of microarray screening results. Briefly, 1ul of the randomly amplified nucleic acid prepared for microarray hybridization from all specimens was utilized as template for 35 cycles of

PCR, under the following conditions: 94°C, 30 seconds; 50°C, 30 seconds; 72°C, 30 seconds. Resulting PCR products were gel purified, subcloned into the TOPO TA cloning vector pCR2.1 (Invitrogen, USA, Carlsbad CA) and sequenced with M13F and M13R primers.

Generation of ABV consensus PCR primers. Sequences recovered from BDV_LconsensusF and BDV_LconsensusR PCR products were aligned, and an additional set of ABV consensus primers biased towards the ABV sequences were identified: ABV_LconsensusF, 5'-CGCCTCGGAAGGTGGTCGG-3' (maps to positions aligning with residues 3724-3742 of BDV reference genome) and ABV_LconsensusR, 5'-GGCAYCAYCKACTCTTRAYYGTRTCAGC-3' (maps to positions aligning with residues 4233-4257 of BDV reference genome). Using identical PCR cycling conditions as described above for the microarray-based *Bornaviridae* PCR assay, these ABV consensus primers were found to be >100X more sensitive for ABV detection compared to BDV_LconsensusF and BDV_LconsensusR primers, and were thus utilized to re-screen the initial set of PDD case and control samples provided for microarray analysis (no additional positives identified) and all subsequently provided samples. Two additional PCR primers in the N (ABV_NconsensusF: 5'-CCHCATGAGGCTATWGATTGGATTAACG-3' and ABV_NconsensusR: 5'-GCMCGGTAGCCNGCCATTGTDGG-3') and M (ABV_MconsensusF: 5'-GGRCAAGGTAATYGTYCCTGGATGGCC-3' and ABV_PconsensusR: 5'-CCAACACCAATGTTCCGAAGMCG-3') that mapped to conserved sequences shared between the complete ABV genome sequence and the 12 other fully sequenced BDV genomes in the NCBI database were also employed for PCR screening of PDD cases and

controls.

Ultra high-throughput sequencing:

Sample preparation and sequencing. 500ng of total RNA derived from one of the PDD case specimens was linearly amplified via modification of the MessageAmp aRNA kit (Applied Biosystems/Ambion, Austin, TX). To ensure the amplification of both mRNA and vRNA present in the specimen, T7-tailed random nonamer was mixed in an equimolar ratio with the manufacturer-provided T7-oligo(dT) primer during the 1st strand synthesis step. The resulting aRNA was next used as input for modified version of Genomic DNA sample preparation protocol for ultra high- throughput Solexa sequencing (Illumina, Hayward, CA). 400ng of the input aRNA was reverse-transcribed with reverse transcriptase (Clontech Laboratories, Inc., Mountain View, CA) using a random nonamer tailed with 19bp of the Solexa Long (5'-CACGACGCTCTTCCGATCTNNNNNNNNN-3') primer sequence (Illumina, Hayward CA). Following termination of reaction, first strand cDNA products were purified from the reaction with Qiagen MinElute spin column (Qiagen USA, Valencia CA). To ensure stringent separation from primers, the MinElute eluate was then filtered through a Microcon YM30 centrifugal filter (Millipore Corp., Billerica, MA). The resulting eluate served as template for 2nd strand synthesis in a standard Sequenase 2.0 (USB, Cleveland, OH) reaction primed with a random nonamer tailed with 22bp (5'-GGCATACGA GCTCTTCCGATCTNNNNNNNNN-3') of the Solexa Short primer sequence (Illumina, Hayward CA). Double-stranded DNA products were separated from primers and very short products through a second Qiagen MinElute spin column run followed by a Microcon YM50 centrifugal filter. This eluate was used as

template for 10 cycles of PCR amplification with the full length Solexa L and S primers using KlenTaq LA DNA polymerase mix (Sigma-Aldrich, St. Louis, MO). PCR product was purified from the reaction with a MinElute spin column. Following cluster generation, Solexa sequencing primer was annealed to the flow cell, and 36 cycles of single base pair extensions were performed with image capture using a 1G Genome Analyzer (Illumina, Hayward, CA). The Solexa Pipeline software suite version 0.2.2.6 (Illumina, Hayward, CA) was utilized for base calling from these images. Using software default quality filters, cycles 4-36 were deemed high quality, resulting in a total of 1.4 million 33mer reads for downstream sequence analyses.

Identification of Bornaviridae reads. Reads sharing 100% identity to each other or the Solexa amplification primers were filtered, reducing our initial set of 1.4 million reads to a working set of 600,000 unique reads. In order to quickly assess the homology of this set of reads to different sequence databases, we employed an iterative strategy using ELAND (Efficient Local Alignment of Nucleotide Data) and BLAST analyses. To filter reads from our analysis potentially derived from psittacine host tissue, the working set of reads were aligned to a database of all *Aves* sequences from NCBI (n=918,511) using ELAND, which tolerates no more than 2 base mismatches, and discards both low quality reads and reads with low sequence complexity. Reads that did not align to the *Aves* database by ELAND analysis were next re-aligned to the *Aves* database for high stringency blastn analysis (e=10⁻⁷, word size=11), followed by progressively lower stringencies (down to e=10⁻², word size=8), corresponding to reads containing only 22 nucleotide identities to sequences in the *Aves* database. To identify reads with some homology to *Bornaviridae* sequences in the resulting set of 322,790 host-filtered reads,

we re-implemented the ELAND/iterative blastn analysis strategy (down to > 15 nucleotides identity) using a database of all NCBI BDV sequences (n=207) augmented by our previously recovered ABV sequences (n=5). An additional iterative tblastx analysis was incorporated to capture distantly related reads that shared similarity to the known BDV sequences only at the level of predicted amino acid sequence (down to > 6 amino acid identity).

Complete ABV vRNA genome sequence recovery by RT-PCR:

Initial genome sequence recovery. Sequences from 33mer reads from the deep sequencing with a minimum of 91% sequence identity with known BDV sequences present in the NCBI database were utilized to generate a set of primers for additional cloning and sequence recovery by RT-PCR of both mRNA and vRNA present in the clinical specimen. In this manner, we generated a hybrid assembly derived from multiple overlapping clones and 5' RACE products encompassing the ABV genome sequence.

vRNA genome sequence recovery. To ensure recovery of accurate sequence across the ABV genome, especially at splice junctions and transcription initiation and termination sites, we utilized the sequence from ABV hybrid assembly to design primers for recovery of 3 overlapping products by RT-PCR directed against the vRNA present in the specimen. Aliquots of 500ng of DNase-treated total RNA extracted from the clinical specimen were annealed with 3 primers complementary to the predicted vRNA sequences: ABV1r, 5'-ATGACCAGGACGAGGAGATG-3' (maps to residues 8831-8812 of vRNA), ABV2r, 5'- CCTGTGAATGTCTCGTTTCTG-3' (maps to residues 5754-5733 of vRNA), and ABV3r 5-TTCTTTCAGCAACCACTGACG-3' (maps to

residues 2563-2543 of vRNA). Reverse transcription was carried out at 50°C for 1hr with SuperScriptIII (Invitrogen, Carlsbad CA) according to manufacturer's instructions. Following RNase H treatment, PCR was performed on the resulting cDNA with Phusion polymerase (NEB, Ipswich, MA) with the primers used for reverse transcription and the following primers: ABV1f: 5'-GGATCATTCCTTGATGATGTATTAGC-3', (maps to residues 5567-5589) ABV2f: 5'-CAAATGGAGAGCCTGATTGG-3' (maps to residues 2378-2397) ABV3f: 5'-AATCGGTAAGTCCAGAGTCAAGG-3' (maps to residues 155-177). All products were amplified for 35 cycles under the following conditions: 98°C, 3 minutes; 98°C, 10 seconds, 50°C, 30 seconds, 72°C 3 minutes. Resulting products were gel purified, and subcloned into the TOPO T/A cloning vector pCR2.1 after incubation with Taq polymerase and dATP for 10 minutes at 72°C. For each product, 4 independent transformants were prepared for standard dideoxy sequencing on an ABI3730 sequencer (ElimBio, Hayward CA). Forward and reverse reads spanning each clone were generated using M13F and M13R and additional overlapping primers spaced at 600-800bp intervals across the each of the clones.

5' and 3' RACE to sequence at vRNA termini. vRNA RT-PCR products containing uncapped vRNA termini were captured using the First Choice RLM RACE kit (Ambion, Austin TX) with the following modifications to the standard protocol: 1) tobacco acid phosphatase treatment was omitted, 2) a phosphorylated RNA, RNAligate, 5'-p-GUUAUCACUUUCACCC-3' (gift of J. Shock, DeRisi lab) was substituted for the 3' RNA ligation-mediated RACE primer provided in the kit and ligated to 3' ends as per manufacturer's 5' RACE protocol, and 3) in the 3' RACE reverse transcription reactions, two reverse transcription reactions were performed and carried forward in parallel: one

with random decamers and one with a DNA oligo complementary to oJSmer utilized in the RNA ligation step (ligateRC, 5'-p- GGGTGAAAGTGATAAC-3'). For 5' RACE, a single round of PCR was sufficient to generate a product using the vRNA specific primer ABV5RaceOuter, 5'- CAGTCGGTTCTTGGACTTGAAGTATCTAGG-3' (maps to residues 346-317 of vRNA) and manufacturer provided outer PCR primer. For 3' RACE, nested PCR was required to recover detectable PCR product of expected size using outer PCR primers oJSmerRC and the gene specific primer ABV3RaceOuter, 5'- CCCGTCTACTGTTCTTTTCGCCG-3' (maps to residues 8479-8497 of vRNA), followed by inner PCR using Tailed_RNAligateRC, 5'- AAGCAGTGGTAACAACGCAGAGTACGGGTGAAAGTGATAAC-3' and the gene specific primer, ABV3RaceInner, 5'- GCAATCCAGGAATAAGCAAGCACAAA-3' (maps to residues 8595-8620 of vRNA). Both of the RACE PCR reactions were carried out with Platinum Taq polymerase (Invitrogen, Carlsbad, CA) in 35 cycles of gradient PCR (with varying annealing temperature): 94°C, 30 seconds; 55-58°C, 30 seconds; 72°C, 30 seconds. Resulting PCR products were gel purified and subcloned into TOPO T/A cloning vector pCR 4.0. For the 5' RACE products, 7 independent transformants from 3 independently generated PCR products were subcloned and sequenced with M13F and M13R primers. For the 3' RACE products, 6 independent transformants from 4 independently generated PCR products were subcloned and sequenced with M13F and M13R primers. Terminal sequences reported here reflect the majority consensus sequence obtained from these reads.

Genome sequence assembly. Genome sequence assemblies from both initial genome sequence recovery and vRNA genome sequence recovery were generated using

Consed, version 16.0 software (19). All bases from the resulting vRNA genome sequence assembly are covered at least 4X with a minimum Phred value of 20.

Blinded PCR screening of additional PDD cases and controls:

Beyond the initial set of 16 specimens provided for microarray analysis, specimens from a total of 38 additional PDD cases, PDD controls, and PDD suspicious birds with varied clinical histories were provided to us blinded by our 2 collaborators (see Additional file 5: Summary of clinical and molecular data for specimens provided in this study). *Sample processing:* For specimens provided in tissue form from the US collaborators, total RNA was extracted as described above with RNABee, DNase treated, then reverse-transcribed and PCR-amplified according to our random amplification protocol for microarray sample preparation (Materials and Methods). Specimens provided from Israel in the form of extracted RNA were similarly DNase- treated and amplified prior to PCR screening. *PCR screening:* 1ul of the randomly amplified material generated from these RNA samples was used as input template for ABV consensus PCRs as described above. In parallel, as an independent control for input specimen RNA integrity, PCR for glyceraldehyde 3-phosphate dehydrogenase (GAPDH) mRNA was performed on all specimens using designed based on Friedman-Einat et al (20) and *Gallus gallus* GAPDH sequence: Gg_GAPDHf: 5'-AGTCATCCCTGAGCTSAAYGG*GAAGC- 3' (bp708-733 in Gallus gallus cDNA (NCBI accession NM_204305), * indicates the junction of GAPDH exon 8 and 9 spanned by this primer), Gg_GAPDhr 5'-ACCATCAAGTCCACAACACGG-3' (Spans bp 1037-1017 in Gallus gallus GAPDH cDNA (NCBI accession NM_204305), maps to GAPDH

exon 12). After PCR results were tallied, clinical information on all specimens tested was unmasked. A complete accounting of ABV, GAPDH PCR results, specimen type and clinical status is provided in Additional file 5: Summary of clinical and molecular data for specimens provided in this study. *Sample inclusion for association analysis:* To reduce potential confounding due to differences in viral detection resulting from specimen tissue source, only specimens derived from upper GI tract tissue (crop, proventriculus/ventriculus) that tested positive by GAPDH mRNA PCR were included in association analysis presented in Table 3. This consisted of a total of 21 specimens, 7 of which were derived from histologically confirmed PDD cases and 14 derived from histologically negative control specimens. *Samples excluded from association analysis:* The remaining 17 samples were excluded from the analysis because they were either 1) GAPDH-positive or GAPDH- negative samples derived from specimen other than upper GI tract tissue (GI content, brain, liver, or intestine) or 2) derived from cases that were histologically or clinically ‘suspicious’, but unconfirmed PDD cases. Six additional ABV PCR positives were identified among this set of samples excluded from the statistical analyses: 1 derived from GI content from a confirmed PDD case, and 5 derived from a variety of tissues from the PDD suspicious cases.

Phylogenetic and comparative sequence analysis:

Multiple sequence alignments of complete genome sequences or partial sequences derived from PCR screening studies were generated with ClustalW (21) version 1.83. Resulting alignments were used for scanning pairwise sequence analysis (window size, 100; step size 1 nucleotide steps). Additional ClustalW alignments and neighbor-joining

phylogenetic trees were generated using Mega software, version 4.0.2 (22).

Results:

Microarray-based detection of a Bornaviridae signature in PDD cases:

To identify a possible viral cause of PDD, we applied the Virus chip, a DNA microarray containing 70mer oligonucleotide probes representing all known viral sequences conserved at multiple nodes of the viral taxonomic tree (17, 23) to identify viral signatures unique to histologically confirmed cases of PDD. At the outset of this study, specimens from two independently collected PDD case/control series were -6-available for this investigation (Figure 1, Materials and Methods). The first series (n=8), from samples originating in the United States, consisted of crop biopsy specimens from 3 histologically confirmed PDD cases and 5 controls that were provided for nucleic acid extraction and follow-up Virus chip analysis. The samples from the second series (n=8) originated in Israel, where total RNA and DNA from proventriculus, ventriculus and brain specimens were extracted from 5 PDD cases and 3 controls. For each series, total RNA was reverse-transcribed with random primers, PCR-amplified, and fluorescently labeled and hybridized to the Virus chip microarray as previously described (17).

In these combined PDD case/control series, a *Bornaviridae* signature was detectable in 62.5% of the cases and none of the controls (Table 1). In the US cohort, which contained only GI tract specimens, we detected a bornavirus in 2 of 3 cases. Surprisingly, in samples from the Israeli PDD case/control series for which we had both GI tract and brain specimen RNA for each animal, we detected the *Bornaviridae*

signature in 3 of the cases, but only in samples derived from brain tissue. These signatures were unambiguously confirmed by follow-up PCR and sequence recovery, using primers based on the sequences of the most strongly annealing *Bornaviridae* oligonucleotides on the microarray (Figure 2, Array probes and PCR probes tracks). These analyses revealed the presence of a set of surprisingly divergent avian bornaviruses (ABVs) in the PDD cases; the recovered sequences shared less than 70% sequence identity to any of the previously identified mammalian bornavirus isolates in the NCBI database.

Recovery of complete genome sequence of a divergent avian bornavirus (ABV) from a PDD case via ultra high-throughput sequencing and conventional RT- PCR:

To determine if the sequence fragments we detected among specimens derived from PDD cases corresponded to the presence of a full-length bornavirus, we performed unbiased deep sequencing on a PCR-confirmed bornavirus positive PDD case that contained the highest concentration of RNA. To recover both mRNA and vRNA present in the sample, RNA from this specimen was linearly amplified with both oligo(dT) and random hexamer primers, and then PCR-amplified using a modified random amplification strategy compatible with the Solexa sequencing platform (Materials and Methods). An initial set of 1.4 million 33mer reads was obtained from this template material. Filtering on read quality, insert presence, and sequence complexity reduced this data set to 600,000 unique reads. Additional ELAND and iterative BLAST analyses ((24), Materials and Methods) of these reads against all avian sequences in NCBI (including ESTs, n= 918,511) identified reads in the dataset with at least 22 nucleotides

of sequence identity likely derived from host transcripts randomly amplified during sequencing sample preparation. The 322,790 reads that passed this host filter were next screened for the presence of bornavirus sequence through similar ELAND and iterative BLAST analyses (Materials and Methods) using a database generated from all Borna Disease virus (BDV) sequences present in NCBI (n=207) and the sequences we had recovered from PCR follow-up of the PDD samples that tested positive for bornavirus by Virus chip microarray (n=5). These analyses provided us with 1400 reads with at least a match of 15 or more nucleotides (blastn) or 7 or more predicted amino acids (tblastx) to known BDV sequences.

Mapping these 1400 reads onto their corresponding positions on a consensus sequence for the 14 publicly available BDV genome sequences revealed spikes of high read coverage distributed discontinuously across the entire span of the BDV genome consensus. Reads containing blastn scores >90% identity to known BDV sequences were used as source sequences for primer design for PCR and sequence recovery of additional bornavirus sequence from both mRNA and vRNA templates present in the PDD specimen. Sequences recovered in this manner facilitated subsequent primer design for recovery of complete genome sequence via RT-PCR of 3 large overlapping fragments of the genome and 5'- and 3'-RACE (Figure 2A, vRNA RT-PCR track) directly from negative stranded vRNA present in the total RNA extracted from this clinical specimen. As our initial PCR results suggested, the bornavirus genome sequence we recovered is quite diverged from all known BDV genomes, including the BDV isolate No/98, a divergent isolate sharing only 81% sequence identity with all other BDV genomes (25). Overall, this newly recovered bornavirus genome sequence shares only 64% sequence

identity at the nucleotide level to each of the complete BDV genomes. Scanning pairwise sequence identity analysis indicates this genetic divergence exists across the entire genome (Figure 2A, Sequence identity shared with BDV genomes track). Given this divergence, we re-examined the depth and distribution of the 322,790 reads from this specimen that passed the host filter to determine if we had missed reads derived from the recovered ABV in our initial screen against all BDV sequences. Not surprisingly, this retrospective BLAST analysis revealed an additional 2600 reads from across the recovered bornavirus genome that were missed in the initial BLAST analyses due to the lack of sequence conservation between the ABV sequence and the available BDV sequences (Figure 2A, Solexa reads track). In total, approximately 1% of all the high throughput shotgun reads could be mapped to the recovered bornavirus genome.

Despite this sequence divergence, this avian bornavirus genomic sequence possesses all of the hallmarks of a *Bornaviridae* family member (Figure 2A): six distinct ORFs encoding homologs of the N, X, P, M, G, and L genes are detectable. Likewise, non-coding regulatory sequence elements (the inverted terminal repeat sequences ((26), see Figure 3), the transcription initiation and termination sites ((27), see Figure 4), and each of the signals for pre-mRNA splicing ((28), see Figure 5) are all conserved in sequence and location, with the exception of the splice acceptor site 3 at position 4560 that has been previously found in a subset, but not all BDV genomes (29, 30). Taken together, these data provide evidence that our analysis has uncovered a novel divergent avian bornavirus (ABV) present in cases of PDD.

Phylogenetic and pairwise sequence analyses support this conclusion. Genomic and sub-genomic phylogenetic analyses of nucleotide sequences place the recovered ABV

sequence on a branch distant from representative members of the 4 distinct genetic isolates of BDV for which complete genome sequences are available (Figure 2B, Figure 6). Strikingly, the ABV genome sequence segregates to a position virtually equidistant from both the set of 3 closely related BDV isolates (V/Ref, H1766, and He/80) and the divergent No/98 BDV isolate (Figure 2B). Moreover, in contrast to the previously identified divergent No/98 isolate, which retains a high level of conservation with other BDV isolates at the amino acid level, the ABV isolate also shows significant sequence divergence in the predicted amino acid sequence of every ORF in the genome (Table 2).

PCR screening of additional PDD cases and controls suggests an association between the presence of ABV and PDD:

Recovery of the complete ABV genome sequence confirmed that the microarray hybridization signature we detected accurately reflected the presence of bornaviruses in our PDD specimens. With these results in hand, we designed a set of PCR primers to perform ABV-specific PCR screening of an independent set of PDD case and control specimens to investigate the association between the presence of ABV and clinical signs and symptoms of PDD. An additional set of 21 samples derived from upper GI tract specimens (crop, proventriculus or ventriculus) from PDD cases and controls were screened for ABV sequences in a blinded fashion (Materials and Methods). For this analysis, we targeted three regions of the genome: 1) the L gene region of the genome that we used for PCR confirmation of the microarray results, (Figure 2, PCR probes track), 2) a subregion within the N gene and 3) a subregion within the M gene (Materials and Methods). PCR for glyceraldehyde 3 phosphate dehydrogenase (GAPDH) mRNA

was performed in parallel with the ABV PCR on all specimens to control for integrity of RNA provided from each specimen. Of the 21 specimens analyzed, 5 were positive for ABV by PCR and confirmed by sequence recovery. Unmasking the clinical status of these samples revealed that 7 of the samples were derived from confirmed PDD cases and 14 samples were derived from PDD controls. Among the PDD cases, we found 71% (5/7) to be positive by ABV PCR (Table 3). In contrast, all PDD controls were negative by ABV PCR, and positive only for GAPDH mRNA. This PCR analysis provides an independent test of the statistical significance of the association between the presence of ABV and histologically confirmed PDD ($P=0.01$, Fisher's Exact Test). Although we do not observe ABV in 100% of PDD cases in this series (see Discussion), our results nonetheless indicate a significant association of ABV with PDD.

Additional ABV isolates identified through PCR screening:

Because we applied stringent inclusion criteria for the above-described association analysis study, a number of ABV (+) and ABV (-) samples were excluded. From these materials, six additional ABV isolates were detected—5 derived from cases considered clinically suspicious and a sixth isolate derived from a confirmed PDD case for which only GI content and liver specimens were available. Additional PCR screening of a set of 12 PDD control crop biopsy specimens provided to us unblinded again yielded solely ABV PCR (-) and GAPDH (+) results. These samples were excluded from the association analysis because we knew their clinical status prior to screening. We note that inclusion of these samples in statistical analyses would not diminish the association of ABV with known or suspected PDD.

Sequence analysis of ABV isolates indicates at least 5 divergent isolates in this branch of the Bornaviridae family:

Recovery of partial sequence from additional isolates of ABV (from the above PDD case/control specimens as well as an additional samples derived from known or suspected PDD cases (Materials and Methods)) from 3 distinct regions of the ABV genome provided the opportunity to further investigate the genetic diversity within this new branch of the *Bornaviridae*. Here, our description of results is restricted to comparison with representative members of the 4 major isolates of BDV, but virtually identical results were obtained when all available BDV sequences were analyzed.

As we observed for the complete ABV genome sequence, phylogenetic analysis of the recovered subgenomic ABV sequences revealed that each of the ABV isolates we recovered resides on a branch distant from the BDV isolates (Figure 7). PCR with the L gene consensus primers detected 14 isolates corresponding to 4 genetic subgroups of ABV. Each of these isolates were also detected with at least one of primer sets corresponding to the more highly expressed N gene and more conserved M gene regions of the genome; however, PCR with these two additional primer sets identified 2 additional ABV isolates that segregate to a genetically distinct 5th subgroup among the ABVs (ABV5, Figure 7B and 7C). Although these 5 distinct branches correlate largely according to the geographic origin of the isolates, the genetic diversity we detect cannot be ascribed solely to differences in geographic origin of the isolates, since one of the branches (ABV4) is comprised of isolates derived from both the U.S. and Israel. Likewise, we did not detect an obvious correlation between host species and genetic subgroup of ABV among the recovered isolates.

Pairwise sequence analyses of the nucleotide and predicted amino acid sequence from the L region of the genome provide additional evidence for surprising genetic diversity among the ABV branches compared to that seen among the BDV branches (Table 4). Although derived from coding sequences of one of the more divergent genes of the bornavirus genome (Table 2, L gene), the region of the L gene we have used for PCR screening is relatively conserved among the BDV isolates, ranging from 81-98% at the nucleotide level, and 96-99% at the amino acid level (Table 4). In contrast, the sequence identity shared across this region of the genome among the ABV branches of the tree ranges from 77-83% at the nucleotide level and 86-94% at the amino acid level. Taken together with the phylogenetic analysis described above, these data provide evidence that these ABV isolates form a new, genetically diverse branch of the *Bornaviridae* phylogeny that is significantly diverged from the founder BDV isolates.

Discussion:

It has been almost 40 years since the first description of PDD. Although a viral etiology has long been suspected, a convincing lead for a responsible viral pathogen has been lacking. By combining veterinary clinical investigation with genomics and molecular biology, we have identified a genetically diverse set of novel avian bornaviruses (ABVs) that are likely to play a significant role in this disease. Through microarray analysis and follow-up PCR, we detected ABV sequences in 62.5% of the PDD cases in a set of specimens from two carefully collected PDD case/control series originating from two different continents. We confirmed that these assays faithfully reflect the presence of full-length bornavirus in ABV PCR positive specimens through

cloning of the complete ABV vRNA sequence directly from RNA extracted from one of these ABV PCR positive PDD case specimens. We next found evidence for a significant association between the presence of ABV and clinically confirmed PDD in follow-up blinded PCR screening of a set of additional PDD cases and controls, with ABV was detected in 71% of PDD cases and none of the controls ($P=0.01$, Fisher's Exact Test). Almost all prior sightings of bornaviruses in nature have been among mammals, and the mammalian isolates have been remarkably homogeneous at the sequence level (Table 2 and (15)). The latter is a surprising feature for RNA viruses, whose RNA-dependent RNA polymerases typically have high error rates. By contrast, the ABV isolates reported here are quite diverged from their mammalian counterparts, and show substantial heterogeneity among themselves. We note with interest that a single earlier report suggesting a potential avian reservoir for bornaviruses has been presented (31). In that study, RT-PCR based on mammalian BDV sequences was used to recover partial sequences from stool collected near duck ponds where wild waterfowl congregate. However, the resulting sequences shared ca. 98% amino acid sequence homology to the mammalian BDVs, raising the possibility that these putative avian sequences might have resulted from possible environmental or laboratory contamination (15). Our ABV isolates, which are unequivocally of avian origin, are clearly very different from these sequences; it remains to be seen if other wild birds can indeed harbor BDV-like agents. The expanded sequence diversity of the bornaviruses discovered here should facilitate design of PCR primers that will enable expanded detection of diverse bornaviral types in future epidemiological studies.

The known neurotropism of bornaviruses makes them attractive and biologically

plausible candidate etiologic agents in PDD, since (i) PDD cases have well-described neurological symptoms such as ataxia, proprioceptive defects and motor abnormalities; and (ii) the central GI tract pathology in the disorder results from inflammation and destruction of the myenteric ganglia that control peristaltic activity. However, despite our success in ABV detection in PDD, we did not observe ABV in every PDD case analyzed. There are several possible explanations for this result. First, we do not know the tissue distribution (tropism) of ABV infection, or how viral copy number may vary at different sites as a function of the stage of the disease. By weighting our sample collection towards clinically overt PDD, we may have biased specimen accrual towards advanced disease. At this stage, where destruction of myenteric ganglial elements is often extensive, loss of infected cells may have contributed to detection difficulties (We note with interest in this context that in one of our case collections from Israel, virus detection occurred preferentially in CNS rather than in GI specimens). There are many precedents for such temporal variation in clinical virology – for example, in chronic hepatitis B viral loads typically decline by several orders of magnitude over the long natural history of the infection (32). It is also possible that our detection rate may merely reflect suboptimal selection of PCR primers employed for screening; after all, our consensus primer selection was based on sequences we had recovered (L gene consensus primers) or sequence homology between the first fully sequenced ABV genome we recovered and a set of highly related mammalian BDV genome sequences (N and M gene consensus primers). We now recognize that there is substantial sequence variation within the ABVs (see Fig. 3); as more sequence diversity is recognized, better choices for more highly conserved primers will become apparent and could impact upon these prevalence

estimates.

Finally, there could actually be multiple etiologic agents in PDD, with ABV infection accounting for only ~70% of the cases. Certainly both human and veterinary medicine are replete with examples of multiple agents that can trigger the same clinical syndrome – for example, at least 5 genetically unrelated viruses (hepatitis viruses A-E) are associated with acute hepatitis, and at least 3 of these can be implicated in chronic liver injury; similarly, several agents (RSV, rhinoviruses and occasionally influenza viruses) are implicated in bronchiolitis. To investigate this possibility, further high-throughput sequencing analysis of PDD cases that were negative for bornaviruses by PCR screening is currently underway.

Although ABV is clearly a leading candidate etiologic agent in PDD, formally establishing a causal role for ABV in PDD will require further experimentation. Such experiments could include (i) attempts to satisfy Koch's postulates via cultivation of ABV, followed by experimental transmission of infection and disease in inoculees, (ii) examination of seroprevalence rates in flocks with high and low PDD incidences, (iii) documentation of seroconversion accompanying development of PDD-like illnesses and (iv) examination of PDD cases by immunohistochemistry or in situ hybridization for evidence of colocalization of ABV infection at sites of histopathology. The recovery and characterization of a complete ABV genome and multiple isolates from this diverse new branch of the *Bornaviridae* family now opens the door to such investigations.

By combining clinical veterinary medical investigation with comprehensive pan-viral microarray and high throughput sequence analyses, we have identified a highly diverged set of avian bornaviruses directly from tissues of PDD cases, but not controls.

These results are significant for a number of reasons. First, they provide a compelling lead in the long-standing search for a viral etiology of PDD, and pave the way for further investigations to assess the link between ABV and PDD. Second, these results also unambiguously demonstrate the existence of an avian reservoir of bornaviruses, expanding our understanding of the bornavirus host range. Finally, these results also provide the first evidence that the *Bornaviridae* family is not confined to a set of genetically homogeneous species as was previously thought, but actually encompasses a set of heretofore unanticipated genetically diverse viral species.

Acknowledgements:

Jenny Shock (DeRisi lab, UCSF) for providing RNA oligos for 3' RACE experiments; Prof. Shmuel Perl, head of the Division of Pathology (KVI), for allowing us access to the KVI histopathology specimen collection; Dr. Asaf Berkovich (KVI) for assistance with specimen preparation and retrieval, Dr. Uri Bendheim, Dr. Revital Harari, and Dr. Anthony Poutous for submitting case material from their practices; and the Lahser Interspecies Research Foundation for providing funding for US specimen collection and veterinary care. The remainder of this work was supported by HHMI grants to JLD and DG.

Table 1 - ABV detection in PDD. ^a3 crop biopsies from US source and 5 brain and proventriculus/ventriculus biopsies from Israel source were examined, with ABV detected in 2 of crop specimens and 3 brain specimens. ^b5 crop biopsies from US source and 3 brain and proventriculus/ventriculus biopsies from Israel source were examined.

| | cases ^a | controls ^b | totals |
|-----------------------|--------------------|-----------------------|--------|
| Virochip ⁺ | 5 | 0 | 5 |
| Virochip ⁻ | 3 | 8 | 11 |
| Totals | 8 | 8 | 16 |

Table 2 - Predicted amino acid sequence similarity between ABV, the divergent

BDV-No/98 and other BDV genomes. *Values without parentheses have no deviation in

% pairwise amino acid identity among compared isolates.

| Genome locus | Average % pairwise amino acid identity (min, max)*: | | | |
|--------------------|--|----------------------|-------------------|----------------------|
| | <i>ABV and BDV</i> | <i>ABV and No/98</i> | <i>BDVs</i> | <i>No/98 and BDV</i> |
| N (nucleocapsid) | 72.5 (72.5, 73.0) | 72.0 | 98.9 (97.3, 100) | 97.0 |
| X (p10 protein) | 40.7 (40.0, 41.0) | 45.0 | 96.9 (96.2, 97.8) | 80.6 (80.0, 81.0) |
| P (phosphoprotein) | 59.9 (59.0, 61.0) | 61.0 | 98.9 (98.6, 99.2) | 96.8 (96.0, 97.0) |
| M (matrix) | 84.0 | 84.0 | 98.2 (97.7, 99.4) | 98.4 (93.0, 94.0) |
| G (glycoprotein) | 65.8 (65.0, 66.0) | 66.0 | 98.4 (96.3, 98.9) | 93.4 (93.0, 94.0) |
| L (polymerase) | 68.0 | 68.0 | 98.8 (98.6, 99.0) | 93.0 |

Table 3 - Analysis of significance of ABV detection rate in PDD.

| | cases | controls | totals |
|----------|-------|----------|--------|
| ABV PCR+ | 5 | 0 | 5 |
| ABV PCR- | 2 | 14 | 16 |
| totals | 7 | 14 | 21 |

P=0.01, Fisher's Exact Test

Table 4 - Average pairwise sequence identity shared between ABV and BDV

isolates*. PCR fragment examined corresponds to bp 3735-4263 of antigenomic strand of BDV V/Ref genome isolate [GenBank: NC_001607]. **Bold text**, average % nucleotide identity; plain text, average % predicted amino acid identity. ABV1 isolate [GenBank:EU781953], ABV2 isolates [GenBank: EU781954 and GenBank:EU781962-66], ABV3 isolate [GenBank:EU781955], ABV4 isolates [GenBank:EU781956-61], Ref/V isolates [GeneBank:NC_001607, GenBank:AJ311521, GenBank:U04608], H1766 isolates GenBank:AJ311523, GenBank:AB258389, GenBank:AB246670], He/80 isolates [GenBank:L27077, GenBan:AJ311522, GenBank:AY05791, GenBank:AY114163, GenBank:AY114162, GenBank:AY114161], No/98 isolate [GenBank:AJ311524].

| | ABV1 | ABV2 | ABV3 | ABV4 | Ref/V | H1766 | He/80 | No/98 |
|-------|------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| ABV1 | 100 | 77 | 79 | 79 | 61 | 61 | 61 | 62 |
| ABV2 | 86 | 100 | 80 | 78 | 59 | 59 | 58 | 60 |
| ABV3 | 89 | 89 | 100 | 83 | 59 | 59 | 58 | 58 |
| ABV4 | 87 | 87 | 94 | 100 | 61 | 60 | 60 | 59 |
| Ref/V | 68 | 64 | 64 | 67 | 100 | 98 | 96 | 82 |
| H1766 | 68 | 64 | 64 | 67 | 99 | 100 | 95 | 83 |
| He/80 | 68 | 64 | 64 | 67 | 99 | 99 | 100 | 81 |
| No/98 | 67 | 65 | 63 | 67 | 97 | 96 | 96 | 100 |

Figure 1 - Clinical presentation of proventricular dilatation disease (PDD) cases and controls.

A) Necropsy view of control (left panel) African gray parrot (*Psittacus erithacus*) that died of other causes. The normal-sized proventriculus is not visible in this view as it lies under the left liver lobe (L). Necropsy view of a great green macaw (*Ara ambiguus*) with PDD (right panel). The proventriculus (PV) is markedly distended and extends laterally well beyond the left lobe of L. The heart (H) is marked for orientation.

B) Contrast fluoroscopy view of control (left panel) African gray parrot (*Psittacus erithacus*) 1.5 hours after administration of barium sulfate. The kidney (K) is marked for orientation. The outline of both the PV and V is clearly visible, with normal size and shape. Within the intestinal loops (IL), wider and thinner sections represent active peristalsis. Right panel, representative PDD case, Eclectus parrot (*Eclectus roratus*) 18 hours after administration of barium. The PV is markedly distended and contains most of the contrast material, with less in the V and within the IL. A large filling defect (*) representing impacted food material. The kidney (K) is shown for orientation. These findings are typical for PDD; however PDD was not confirmed by histology in this case.

C) Proventriculus histopathology. Hematoxylin and eosin staining of proventriculus histological sections from a blue and yellow macaw (*Ara ararauna*) with PDD.

Proventricular gland (G) is shown for orientation. Left panel, normal appearing myenteric ganglion detected within the proventriculus of this case (arrow); right panel, marked lymphoplasmacytic infiltration present within a myenteric ganglion (arrows). Right panel inset, higher magnification. **D)** CNS histopathology. Hematoxylin and eosin staining of a cerebral section from a control (left panel) African gray parrot (*Psittacus erithacus*) that died of other causes. Right panel, African gray parrot (*Psittacus erithacus*) with PDD.

Perivascular cuffing is evident around blood vessels (arrows). Inset, higher magnification.

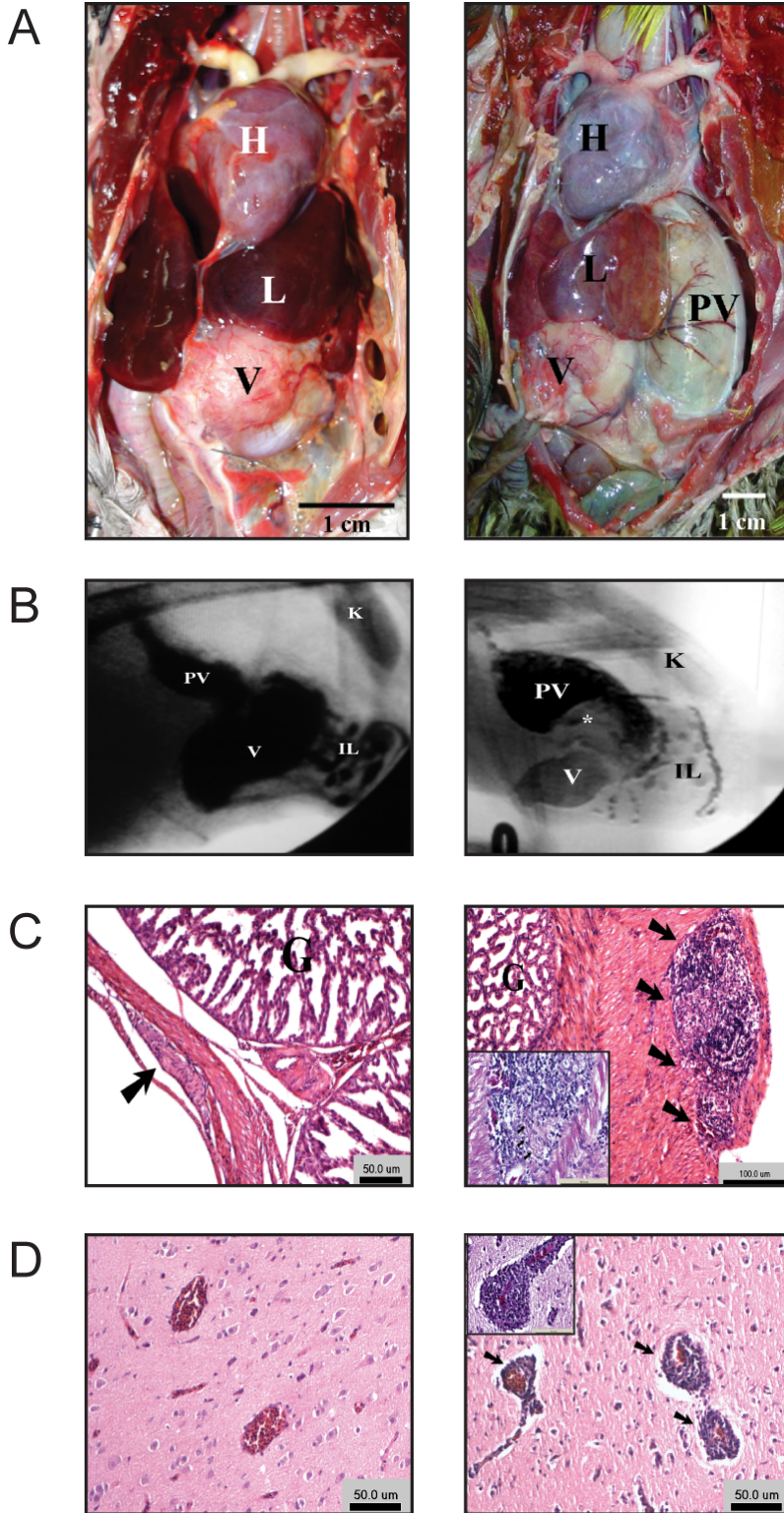


Figure 2 - Avian bornavirus (ABV) genome sequence recovery and comparative analysis to Borna disease virus (BDV) genomes. A) *Bornaviridae* genome schematic. Grey bar at base, non-segmented negative sense viral RNA (vRNA) of *Bornaviridae* genome; coordinates of major sequence landmarks highlighted below. Green bars and dashed lines, transcription initiation sites (TISs); red bars, transcription termination sites. Distinct ORF-encoding transcription products and the gene products they encode are diagrammed above: TIS1 transcripts encoding nucleocapsid (N) gene, pink; TIS2 transcripts encoding phosphoprotein (P) and X genes, green; TIS3 transcripts encoding the matrix (M), glycoprotein (G) and polymerase (large or 'L') gene, blue. Exons, thick solid black lines; introns, thin solid black lines; dashed black lines, 3' ends of transcripts generated transcription termination read-through; shaded boxes, location of ORFs in transcripts; reading frames for ORFs from multiple genes generated from TIS3 indicated at right. Array probes track, *Bornaviridae* oligonucleotide 70mer probes from the Virochip array. PCR primers track, primers generated for PCR follow up and screening of specimens in this study for detection of *Bornaviridae* species with expected product diagrammed below. vRNA RT-PCR track, overlapping vRNA clones and RACE products recovered directly from RNA extracted from crop tissue of a histologically confirmed case of PDD. Solexa reads track shows distribution of 33mer reads with at least 15bp sequence identity to recovered ABV genome sequence. Sequence identity with BDV genomes track shows scanning average pairwise nucleotide sequence identity (window size of 100 nucleotides, advanced in single nucleotide steps) shared between ABV and all BDV genome sequences in NCBI. A dashed line on the graph indicates 50% identity threshold for reference. **B)** Phylogenetic analysis of ABV genome and the 4

representative BDV genome isolates. Neighbor-joining phylogenetic trees based on nucleotide sequences of the ABV genome sequence [GenBank:EU781967] and the following representative BDV genome sequences: H1766 [GenBank:AJ311523], V/Ref [GenBank:NC_001607], He/80 [GenBank:L27077], and No/98 [GenBank:AJ311524])

Scale bar, genetic distance.

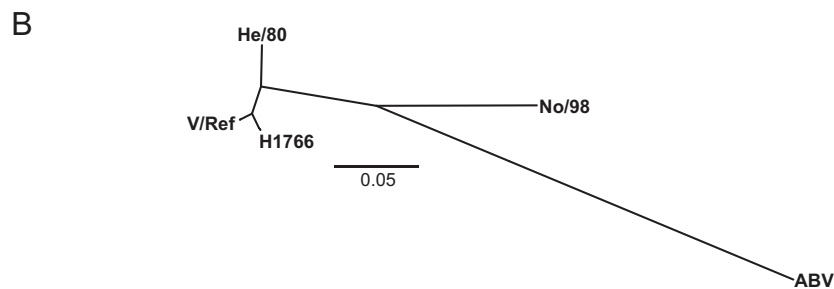
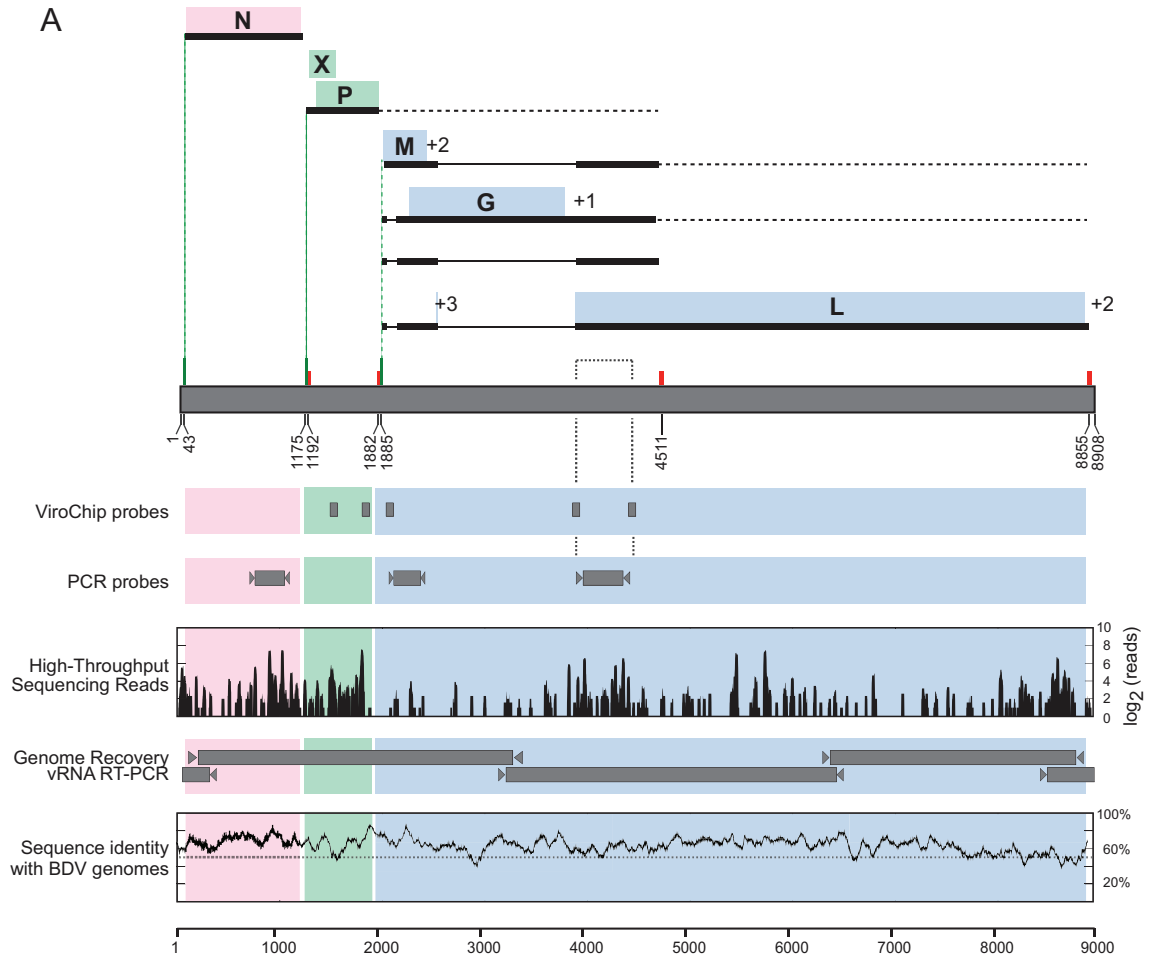


Figure 3 – Alignment of bornavirus genomes 5’ and 3’ termini. Bornavirus genome organization overview diagrammed as in Figure 1. Sequences in alignments shown are complementary to vRNA sequence, genome isolate names shown at left. 3’ end sequence recovered for ABV genome and other BDV genomes is shown in left panel, 5’ end sequence recovered for ABV genome and other BDV genomes is shown in right panel. Accession numbers for genomes aligned: hu2Pbr [GenBank:AB258389], Bo/04w [GenBank:AB246670], H1766 [GenBank:AJ311523], Ref [GenBank:NC_001607], V [GenBank:U04608], V/FR [GenBank:AJ311521], CRNP5 [GenBank:AY114163], CRP3B [GenBank:AY114162], CRP3A [GenBank:AY114161], He/80/FR [GenBank:AJ311522], He/80 [GenBank:L27077], pBRT7-HrBDVc [GenBank:AY705791], No/98 [GenBank:AJ311524], ABV [GenBank:EU781967].

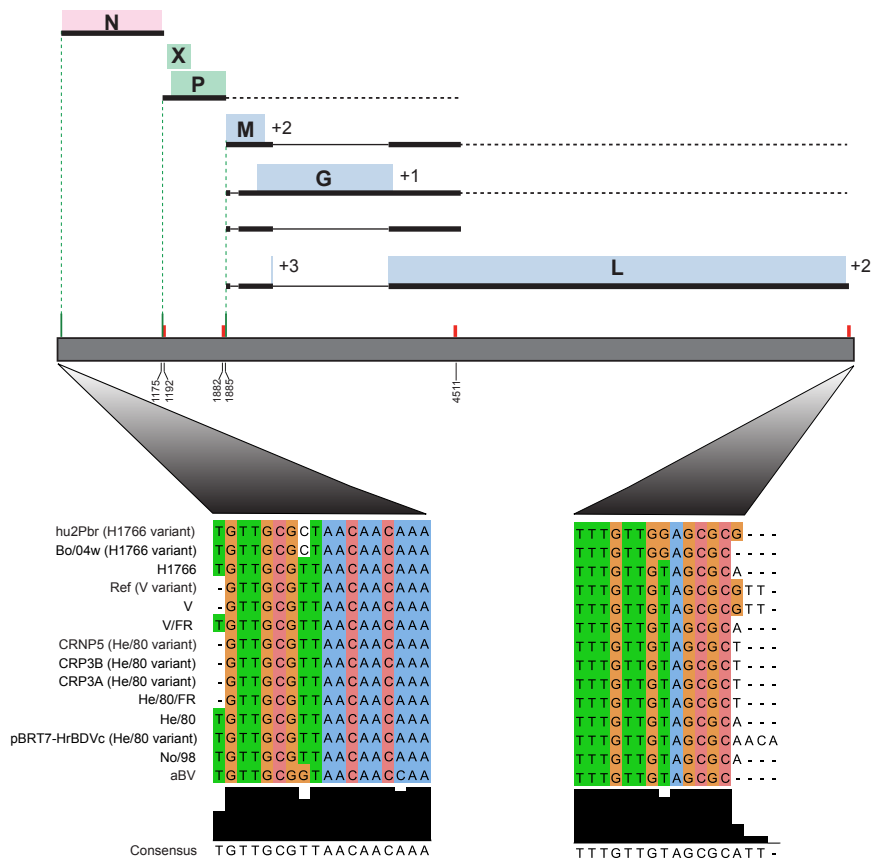


Figure 4 – Alignment of transcription initiation and termination sites in bornavirus genomes. A) Alignment of the 3 bornavirus transcription initiation sites (TIS) and 6 nucleotides of flanking sequences. B) Alignment of the 4 bornavirus transcription termination sites. Source genomes for alignments are shown at left. Black triangles highlight ABV sequences.

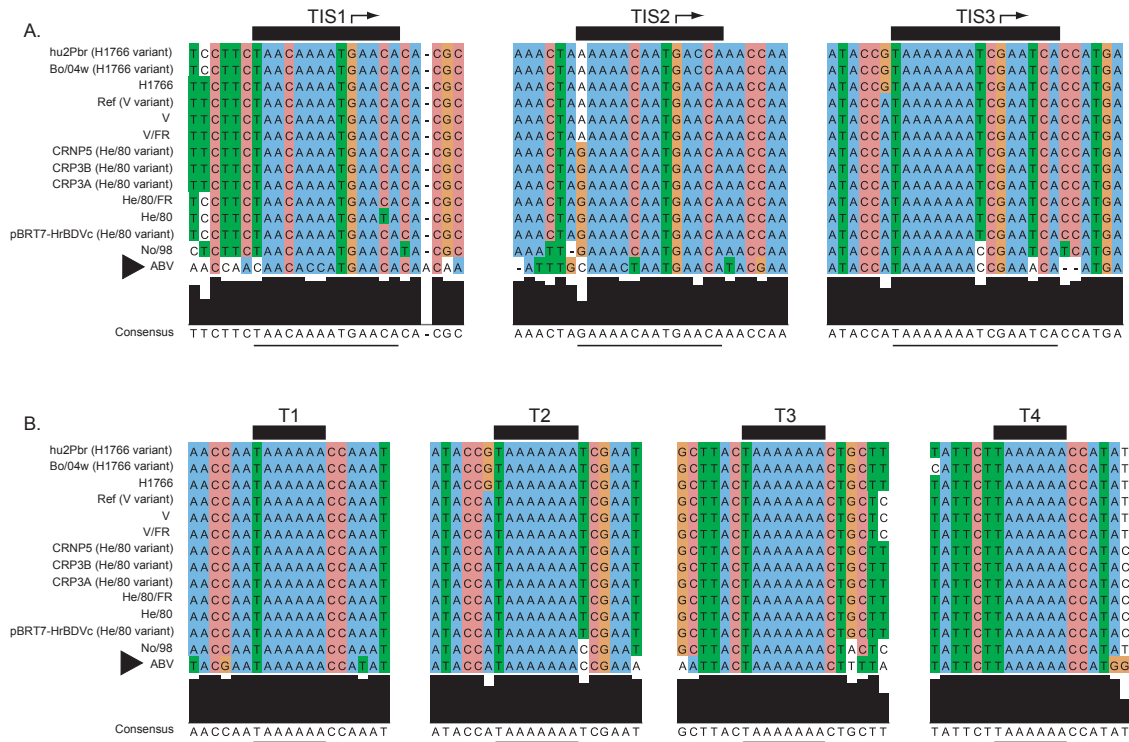


Figure 5 – Alignment of splice donor and splice acceptor sequences in

bornavirus genomes. A) Alignment of splice donor 1 and splice acceptor 1 sequences;

B) Alignment of splice donor 2 and splice acceptor 2 sequences; C) Alignment of

splice acceptor 3 sequences. Source genomes for alignments are shown at left.

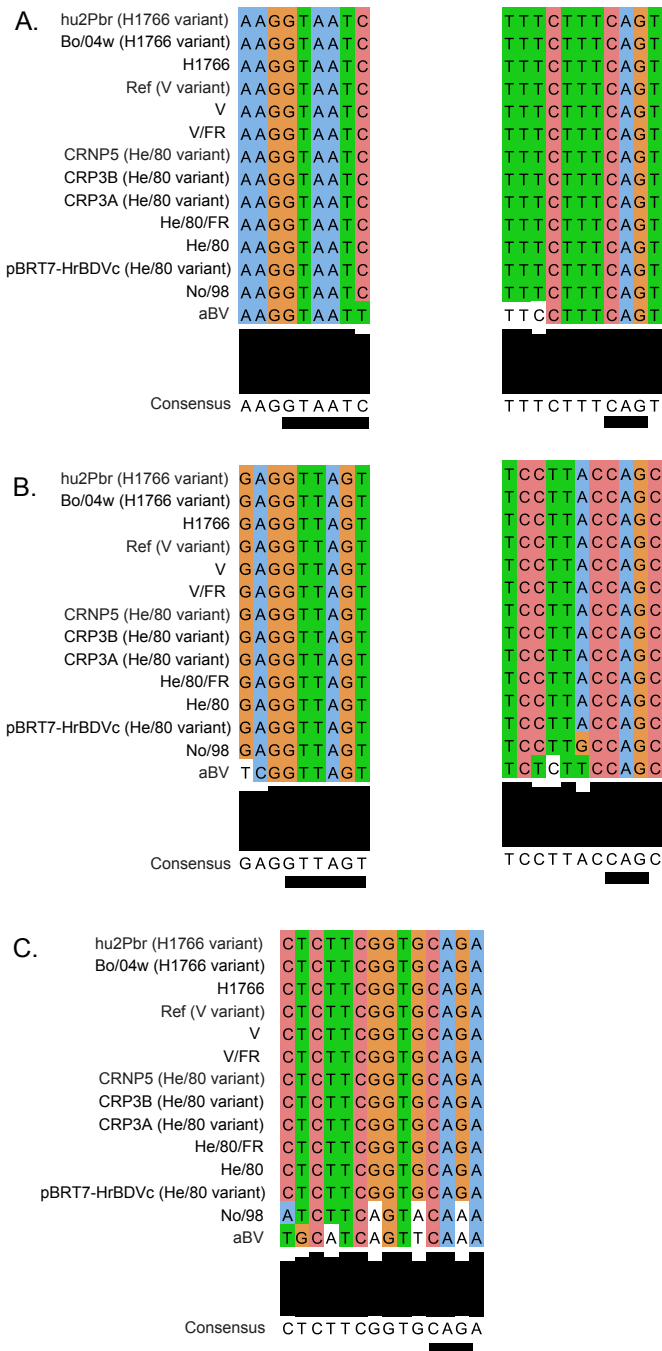


Figure 6 – Phylogenetic relationships between sub-genomic loci of ABV

and representative BDV genomes. Neighbor-joining trees generated for the indicated

nucleotide sequences of ABV and a representative set of BDV genomes are shown for

each ORF in the bornavirus genome. Accession numbers of representative BDV genomes

are: Ref/V [GenBank:NC_001607], H1766 [GenBank:AJ311523], He/80

[GenBank:AY705791], No/98 [GenBank:AJ311524].

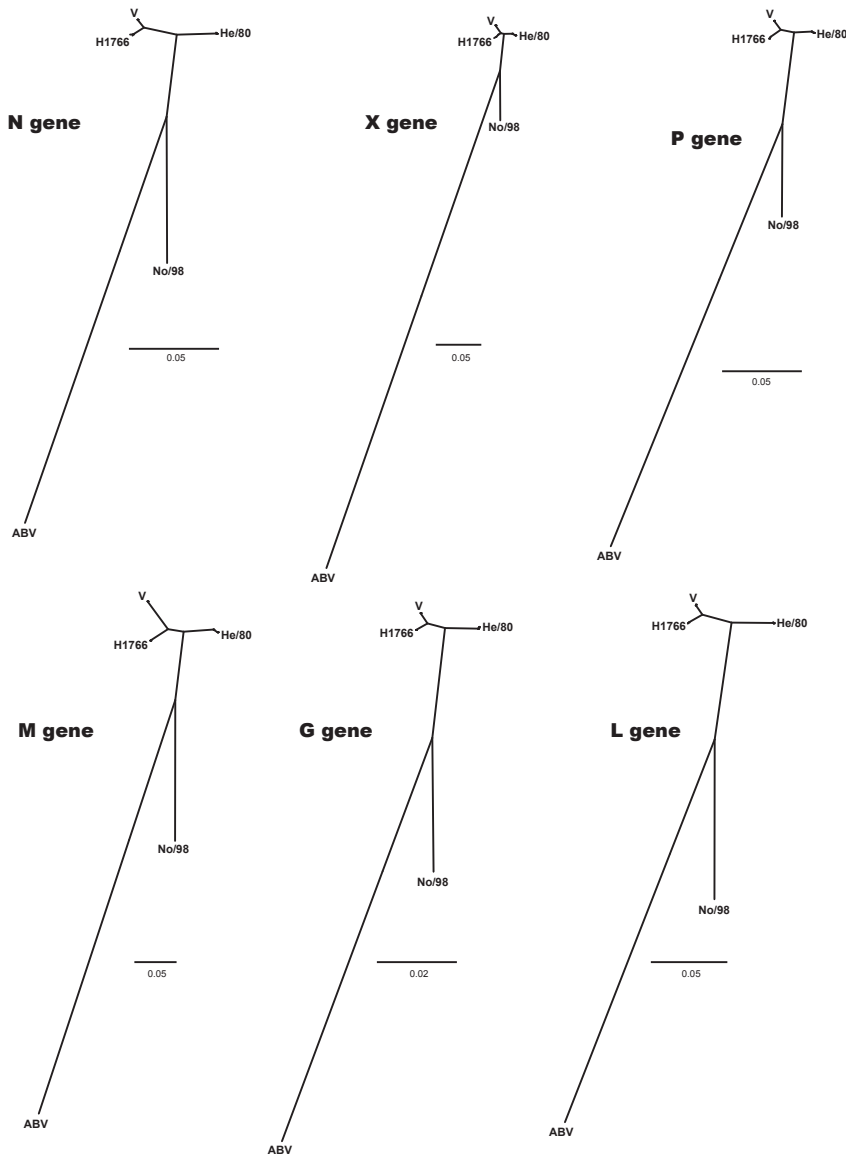
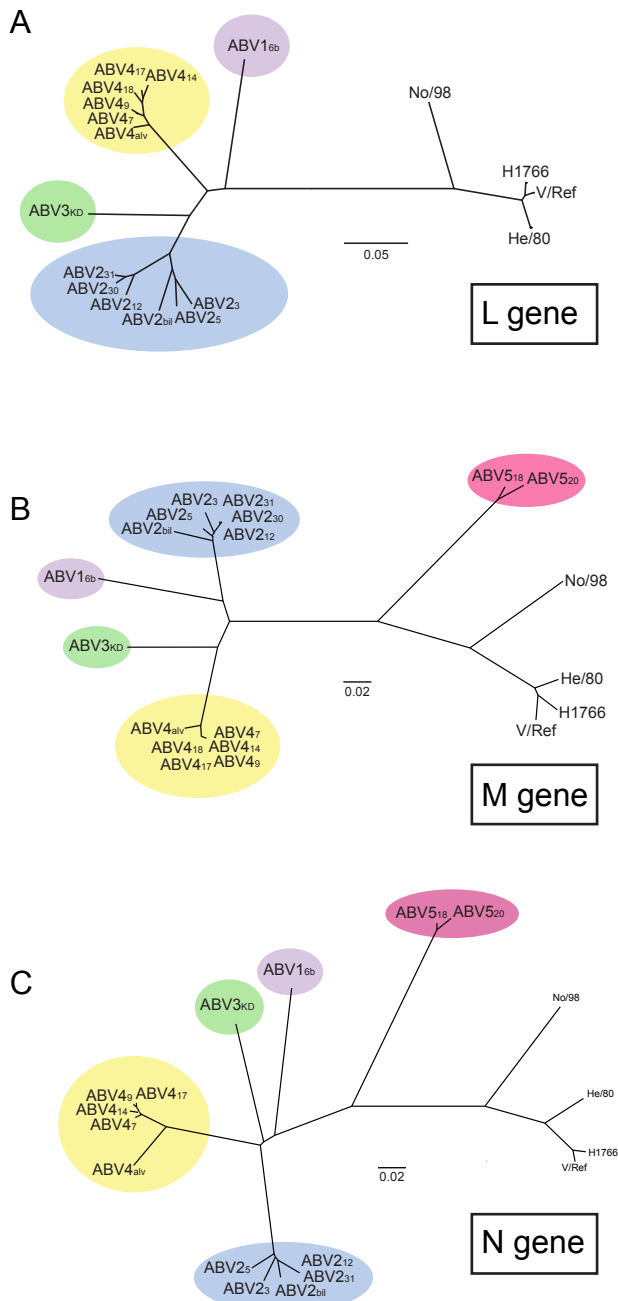


Figure 7 - Comparison of sequences recovered from ABV PCR screening to 4 representative genetic isolates of BDV. Neighbor-joining Phylogenetic tree of ABV nucleotide sequences recovered by PCR screening with ABV consensus primers for subsequences within the L gene (A), the N gene (B), or the M gene (C).



References:

1. Daoust,P.Y., Julian,R.J., Yason,C.V. and Artsob,H. (1991) Proventricular impaction associated with nonsuppurative encephalomyelitis and ganglioneuritis in two Canada geese. *J. Wildl. Dis*, **27**, 513-517.
2. Doneley,R.J.T., Miller,R.I. and Fanning,T.E. (2007) Proventricular dilatation disease: an emerging exotic disease of parrots in Australia. *Aust. Vet. J*, **85**, 119-123, 10.1111/j.1751-0813.2007.00109.x.
3. Gregory,C., Latimer,K., Niagro,F., Ritchie,B., Campagnoli,R., Norton,T. and Greenacre,C. (1994) A review of proventricular dilatation syndrome. *J Assoc Avian Vet*, **8**, 69-75.
4. Perpiñán,D., Fernández-Bellon,H., López,C. and Ramis,A. (2007) Lymphoplasmacytic myenteric, subepicardial, and pulmonary ganglioneuritis in four nonpsittacine birds. *J. Avian Med. Surg*, **21**, 210-214.
5. Sullivan,N.D., Mackie,J.T., Miller,R.I. and Giles,A. (1997) First case of psittacine proventricular dilatation syndrome (macaw wasting disease) in Australia. *Aust. Vet. J*, **75**, 674.
6. Berhane,Y., Smith,D.A., Newman,S., Taylor,M., Nagy,E., Binnington,B. and Hunter,B. (2001) Peripheral neuritis in psittacine birds with proventricular dilatation disease. *Avian Pathol*, **30**, 563-570, 10.1080/03079450120078770.
7. Lutz,M.E. and Wilson,R.B. (1991) Psittacine proventricular dilatation syndrome in an umbrella cockatoo. *J. Am. Vet. Med. Assoc*, **198**, 1962-1964.
8. Mannl,A., Gerlach,H. and Leipold,R. (1987) Neuropathic gastric dilatation in psittaciformes. *Avian Dis*, **31**, 214-221.

9. Vice,C.A. (1992) Myocarditis as a component of psittacine proventricular dilatation syndrome in a Patagonian conure. *Avian Dis*, **36**, 1117-1119.
10. Lublin,A., Mechani,S., Farnoushi,I., Perl,S. and Bendheim,U. (2006) An outbreak of proventricular dilatation disease in psittacine breeding farm in Israel. *Israel Journal of Veterinary Medicine*, **61**, 16-19.
11. Gregory,C., Niagro,F., Roberts,A., Campagnoli,R., Pesti,D., Ritchie,B. and Lukert,P. (1997) Investigations of Eastern Equine Encephalomyelitis Virus as the Causative Agent of Psittacine Proventricular Dilatation Syndrome. *J Avian Medicine and Surgery*, **11**, 187-193.
12. Gough,R.E., Drury,S.E., Harcourt-Brown,N.H. and Higgins,R.J. (1996) Virus-like particles associated with macaw wasting disease. *Vet. Rec*, **139**, 24.
13. Gough,R.E., Drury,S.E., Culver,F., Britton,P. and Cavanagh,D. (2006) Isolation of a coronavirus from a green-cheeked Amazon parrot (*Amazon viridigenalis* Cassin). *Avian Pathol*, **35**, 122-126, 10.1080/03079450600597733.
14. Ritchie,B. (1995) *Avian Viruses: Function and Control* Wingers Publishing, Lake Worth.
15. Dürwald,R., Kolodziejek,J., Muluneh,A., Herzog,S. and Nowotny,N. (2006) Epidemiological pattern of classical Borna disease and regional genetic clustering of Borna disease viruses point towards the existence of to-date unknown endemic reservoir host populations. *Microbes Infect*, **8**, 917-929, 10.1016/j.micinf.2005.08.013.
16. Rott,R. and Becht,H. (1995) Natural and experimental Borna disease in animals. *Curr. Top. Microbiol. Immunol*, **190**, 17-30.

17. Chiu,C.Y., Rouskin,S., Koshy,A., Urisman,A., Fischer,K., Yagi,S., Schnurr,D., Eckburg,P.B., Tompkins,L.S., Blackburn,B.G. et al. (2006) Microarray detection of human parainfluenzavirus 4 infection associated with respiratory failure in an immunocompetent adult. *Clin. Infect. Dis*, **43**, e71-76, 10.1086/507896.
18. Urisman,A., Fischer,K.F., Chiu,C.Y., Kistler,A.L., Beck,S., Wang,D. and DeRisi,J.L. (2005) E-Predict: a computational strategy for species identification based on observed DNA microarray hybridization patterns. *Genome Biol*, **6**, R78, 10.1186/gb-2005-6-9-r78.
19. Gordon,D., Abajian,C. and Green,P. (1998) Consed: a graphical tool for sequence finishing. *Genome Res*, **8**, 195-202.
20. Friedman-Einat,M., Boswell,T., Horev,G., Girishvarma,G., Dunn,I.C., Talbot,R.T. and Sharp,P.J. (1999) The chicken leptin gene: has it been cloned? *Gen. Comp. Endocrinol*, **115**, 354-363, 10.1006/gcen.1999.7322.
21. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, **22**, 4673-4680.
22. Tamura,K., Dudley,J., Nei,M. and Kumar,S. (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol. Biol. Evol*, **24**, 1596-1599, 10.1093/molbev/msm092.
23. Chiu,C.Y., Alizadeh,A.A., Rouskin,S., Merker,J.D., Yeh,E., Yagi,S., Schnurr,D., Patterson,B.K., Ganem,D. and DeRisi,J.L. (2007) Diagnosis of a critical respiratory illness caused by human metapneumovirus by use of a pan-virus

- microarray. *J. Clin. Microbiol*, **45**, 2340-2343, 10.1128/JCM.00364-07.
24. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, **25**, 3389-3402.
25. Nowotny,N., Kolodziejek,J., Jehle,C.O., Suchy,A., Staeheli,P. and Schwemmler,M. (2000) Isolation and characterization of a new subtype of Borna disease virus. *J. Virol*, **74**, 5655-5658.
26. Schneider,U., Schwemmler,M. and Staeheli,P. (2005) Genome trimming: a unique strategy for replication control employed by Borna disease virus. *Proc. Natl. Acad. Sci. U.S.A*, **102**, 3441-3446, 10.1073/pnas.0405965102.
27. Schneemann,A., Schneider,P.A., Kim,S. and Lipkin,W.I. (1994) Identification of signal sequences that control transcription of borna disease virus, a nonsegmented, negative-strand RNA virus. *J. Virol*, **68**, 6514-6522.
28. Schneider,P.A., Schneemann,A. and Lipkin,W.I. (1994) RNA splicing in Borna disease virus, a nonsegmented, negative-strand RNA virus. *J. Virol*, **68**, 5007-5012.
29. Cubitt,B., Ly,C. and de la Torre,J.C. (2001) Identification and characterization of a new intron in Borna disease virus. *J. Gen. Virol*, **82**, 641-646.
30. Tomonaga,K., Kobayashi,T., Lee,B.J., Watanabe,M., Kamitani,W. and Ikuta,K. (2000) Identification of alternative splicing and negative splicing activity of a nonsegmented negative-strand RNA virus, Borna disease virus. *Proc. Natl. Acad. Sci. U.S.A*, **97**, 12788-12793, 10.1073/pnas.97.23.12788.
31. Berg,M., Johansson,M., Montell,H. and Berg,A.L. (2001) Wild birds as a possible

natural reservoir of Borna disease virus. *Epidemiol. Infect.*, **127**, 173-178.

32. Ganem,D. and Prince,A.M. (2004) Hepatitis B virus infection--natural history and clinical consequences. *N. Engl. J. Med.*, **350**, 1118-1129, 10.1056/NEJMra031087.

Chapter 4: Determination of genetic correlates of resistance to artemisinin in *Plasmodium falciparum*

This chapter is a summary of work done by:

Sorber K, Dimon M, Tucker M, Kyle D, DeRisi JL.

Author contributions:

Matt Tucker derived the artemisinin resistant strain *in vitro* and isolated genomic DNA from the resistant and parent strains. Katherine Sorber performed Illumina sequencing on the genomic DNA and performed all follow-up experiments. Michelle Dimon processed the Illumina sequencing data and analyzed it for SNPs and amplifications. Dennis Kyle and Joseph L. DeRisi conceived of and supervised the project.

Joseph L. DeRisi, Thesis Advisor

Introduction:

Malaria continues to be a significant and deadly human disease with over 1.38 billion people at risk for infection with *Plasmodium falciparum* (the most deadly human malaria parasite) in 2007 (1), and 243 million cases of malaria in 2008 resulting 863,000 deaths (2). Several factors have contributed to the intractability of eradicating malaria, including the emergence of parasite strains resistant to available anti-malarial drugs (3). In the face of reduced efficacy of other anti-malarials, the World Health Organization (WHO) began recommending artemisinin derivatives in combination with other drugs (ACTs) as first line treatments in 2006 (4).

Artemisinin and its derivatives are sesquiterpene lactones that contain an endoperoxide bridge essential for activity (5), making them structurally unique from previously known antimalarial scaffolds. They are fast-acting, kill all blood stages of the parasite, and reduce gametocyte carriage (6-8). However, the mechanism of action of the active metabolite dihydroartemisinin (DHA) is not clear. Several proposed hypotheses include general alkylation or oxidation of parasite proteins due to metabolism of the endoperoxide ring to form either free radicals or hydroperoxides/metalloperoxides (9), specific inhibition of a SERCA-like Ca^{2+} -dependent ATPase, PfATP6 (10), or inhibition of the parasite's mitochondrial electron transport chain (11).

As artemisinin derivatives enter their second or third decade of widespread use in some areas, troubling field research indicates that they may be losing their efficacy. The evidence for emerging parasite resistance to artemisinin includes: 1) longer parasite clearance times in western Cambodia versus the Thai-Burmese border with both combination and monotherapy (12), 2) increase in gametocyte carriage and longer

parasite clearance times since ACTs were introduced in 1995 at the Thai-Burmese border (13), and 3) elevated IC50s in isolates from west to east Asia (14). Although some studies tracking parasite clearance time and *in vitro* IC50 find a correlation (14), others do not (12). Recent analysis does confirm, however, that longer parasite clearance times are heritable among *P. falciparum* parasites, indicating a genetic basis for resistance (15). Although several candidate genes have been proposed, no correlation between candidate gene mutations (or amplifications in the case of PfMDR1) and parasite resistance has been confirmed in field isolates (16), including mutation of PfATP6 or the mitochondrial genome.

Here we present whole-genome sequencing of an *in vitro* selected *Plasmodium falciparum* strain resistant to high levels of artemisinin. By comparing this strain (D6.QHS2400x2) to its parent strain (D6), we find both candidate single nucleotide polymorphisms (SNPs), as well as candidate genome amplifications that correlate with acquisition of resistance. The top 5 most promising SNPs, as well as the largest amplification, on chromosome 10, were validated using independent experimental techniques. The chromosome 10 amplification was not found in other laboratory *P. falciparum* strains or in strains selected for resistance to other drugs, but was detected in an independently derived artemisinin selected strain. Such correlation makes it the most promising lead to result from these experiments.

Materials and Methods:

Selection of D6 with Dihydroartemisinin (DHA):

Parental D6 *Plasmodium falciparum* parasites were first selected for resistance to 80ng/mL of artelinic acid (AL) as described in (17). D6.AL80 was then further selected on dihydroartemisinin (DHA) in a step-wise manner over approximately one year (Figure 1, Tucker unpublished results). Each selection began with mixed stage parasites between 1.4-3.9% parasitemia. Selection steps consisted of adding 80 ng/mL, 120 ng/mL, 160 ng/mL, 200 ng/mL, 240 ng/mL, 280 ng/mL, 300 ng/mL, 340 ng/mL, or 2400 ng/mL of DHA to the culture media for 48 hours, then washing drug out. Selections were allowed to recover until normal parasites (by smear) reached at least 1.4% parasitemia. Recovered cultures were then either subjected to the same step again or to the next step in drug concentration. Each step was applied at least twice and up to 4 times.

Extraction of Genomic DNA:

Genomic DNA was extracted from mixed stage D6 parental and D6.QHS2400x2 resistant parasites (Tucker, unpublished results).

Generation and Sequencing of Illumina Genomic DNA Libraries:

Approximately 1.5 µg of D6 parental genomic DNA was sonicated for three 15 minute cycles on high of 1 minute on, 30 seconds off in a Bioruptor water bath sonicator (Diagenode Inc., Denville, NJ). Resulting fragmented DNA was end polished and A-extended using Illumina's Genomic DNA Preparation kit (Illumina Inc., Hayward, CA). MinElute PCR purification columns (Qiagen Inc., Valencia, CA) were used to purify and concentrate DNA after each step. Extended fragments were ligated to Sol-L-GsuI-T and Sol-S-short-biotin-T (see Table 1 for all primers and adapters) at a 10:1 molar ratio using

reagents from Illumina's Genomic DNA Preparation kit. Agencourt AMPure magnetic beads (Beckman Coulter Inc., Brea, CA) were used to purify ligated DNA, which was then immobilized on Dynabeads M-280 Streptavidin according to the manufacturer's instructions (Invitrogen Corp., Carlsbad, CA). PCR was performed with the beads using 1x KlenTaq LA buffer, 0.1 mM dATP, 0.1 mM dTTP, 0.05 mM dCTP, 0.05 mM dGTP, 0.1 μ M short-PE-Sol-S, 0.1 μ M short- Sol-L-GsuI-T, and 10U KlenTaq LA (Sigma-Aldrich Co., St. Louis, MO). Cycling conditions were 95°C for 2 min, then 5 cycles of 95°C for 30 sec, 52°C for 45 sec, 65°C for 3 min, and a final extension at 65°C for 7 min. The PCR was cleaned up with a MinElute PCR purification column, then products were subjected to a second PCR with the same conditions except 0.1 μ M biotin-short-Sol-L-GsuI was used instead of short- Sol-L-GsuI-T. Products were again purified with a MinElute PCR purification column, then bound to a fresh aliquot of Dynabeads M-280 Streptavidin. Final PCR was performed with the beads as described above except 0.1 μ M full-Sol-L-GsuI and 0.1 μ M full-PE-Sol-S were used. Products were purified with AMPure magnetic beads. The same protocol was used to produce a sequencing library using D6.QHS2400x2 genomic DNA as well.

The D6 genomic library was clustered in 3 lanes of a v3 Illumina paired-end flowcell, while the D6.QHS2400x2 library was clustered in 4 lanes. Following cluster generation, Sol-SeqPrimer was annealed to the clusters on the flow cell, and 41 cycles single base pair extensions were performed with image capture using an Illumina (Solexa) GA2 sequencer (Illumina, Hayward, CA). After conversion of the clusters, sequencing of the second end was performed using PE-SolS-SeqPrimer with 41 cycles of single base pair extensions were performed with image capture. The Solexa Pipeline

software suite version 0.2.2.6 (Illumina, Hayward, CA) was utilized for base calling from these images.

Analysis of Sequencing Data:

Sequencing reads were aligned to the *Plasmodium falciparum* genome (PlasmoDB v5.5, (18)). The aligned reads were used to determine the sequenced base for each base of the genome. If there were no reads, then the base could not be determined. Otherwise, if there was a majority of one base, then that base was called. If there was a tie and one of the top bases was the same as the genome base, then that one was used. In other conditions, the base was left as unknown. Once the sequenced base was determined for each base of the parental and resistant populations, these bases were compared to each other to look for places where the sequenced base was different between the two populations. These SNPs were ranked based on their coverage level and the percent of the reads with the majority base in parental and resistant populations.

To find amplifications, coverage was smoothed in 5 Kb windows using the median number of reads per basepair. A conservative threshold of 100 reads per basepair within a window was set to define an amplification.

Confirmation of SNPs:

The top 5 SNPs ranked by purifying selection in D6.QHS2400x2 versus D6 were subjected to confirmation by PCR (see Table 1 for primer sequences). Each PCR reaction contained 1x Herculase II Fusion buffer, 0.1 mM dTTP, 0.1 mM dATP, 0.05 mM dCTP, 0.05 mM dGTP, 0.05 μ M each primer, 30 ng D6.QHS2400x2 genomic DNA, and 0.4U Herculase II Fusion (Agilent Technologies Inc., Santa Clara, CA). Cycling

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 a final extension at 65°C for 7 min. PCRs were purified with MinElute PCR purification columns (Qiagen Inc., Valencia, CA), and then sequenced by Sequetech using the appropriate F primer (Sequetech, Mountain View, CA).

Confirmation of MAL10 Amplification:

qPCR measuring the relative copy number (19) of select genes in and around the perceived chromosome 10 amplification in D6.QHS2400x2 was performed using the LightCycler 480 SYBR Green I Master mix (Roche Applied Science, Indianapolis, IN) with 0.5 µM of each appropriate primer and 5ng of the appropriate genomic DNA. Each reaction was done in triplicate. Cycling was performed on a DNAEngine Opticon machine (MJ Research Inc., Waltham, MA) using Opticon Monitor Analysis Software v1.4 (MJ Research Inc., Waltham, MA). Cycling conditions were 95°C for 10 min, followed by 40 cycles of 95°C for 30 sec, 52°C for 45 sec, 65°C for 1.5 min, and 68°C for 10 sec. Fluorescence was read following each cycle and a final extension was performed at 65°C for 7 min before melting curve analysis was performed (65°C to 95°C with a 10 sec hold for every 0.5°C followed by a fluorescence read). Only reactions with one clean peak from melting curve analysis were analyzed. C(t) values were defined as the point in the fluorescence curve just before linear growth (usually ~0.01-0.015 fluorescence units) and normalized by subtraction of baseline fluorescence. Genomic DNA from mixed stage 3D7 Oxford parasites served as the control sample, while PFL2510w (chitinase) served as the reference gene.

Results:

Selection and sequencing of *in vitro* selected artemisinin resistant *P. falciparum*:

The *Plasmodium falciparum* laboratory strain D6 was selected for artemisinin resistance in a step-wise fashion as described in Chavchich et al (17). Over the period of approximately one year, step-wise drug selection cycled with periods of recovery led to the selection of a parasite population that could recover from 2400ng/mL of dihydroartemisinin (DHA) within 14 days (see Materials and Methods, Figure 1).

Genomic DNA was isolated from a population of parasites that had twice recovered from treatment with a high dose of DHA (D6.QHS2400x2), as well as from the parental D6 parasite population. This DNA was prepared for paired-end Illumina sequencing, resulting in 29.5 million 41 bp reads for D6.QHS2400x2 and 40.4 million 41 bp reads for D6 (Table 2).

SNP detection and verification:

Using Illumina's ELAND tool and BLAT (20), reads for each strain were aligned to the published *Plasmodium falciparum* genome (21). We reasoned that if a SNP were responsible for conferring artemisinin resistance, it would most likely reside within an exon of a non-antegenic variation, protein-coding gene. Therefore, we looked for exonic positions in the genome with at least three reads of coverage for each strain where the dominant bp differed between the two strains. This list was further narrowed by assuming that a causative SNP was also most likely to cause a non-synonymous rather than synonymous amino acid change in the resulting protein (Table 3). Of the resulting 17 SNPs, 5 appeared in more than 90% of the relevant D6.QHS2400x2 reads and

occurred at positions with little variation in D6. These SNPs were considered top candidates because of apparently highly purifying selection within the resistant population. PCR primers were designed flanking each of the top 5 candidate SNPs (Table 1). PCR products from both strains covering each SNP were sequenced and all 5 SNPs were confirmed in D6.QHS2400x2 (Table 3).

Detection and verification of an amplification on chromosome 10:

Because of uneven coverage in D6.QHS2400x2, programmatic determination of genome amplifications required smoothing of the data. Based on empirical trial-and-error, median coverage over 5000 bp windows was plotted for the entire genome. A stringent cutoff of a median of 100 reads per bp was used to identify potential amplifications (Table 4). The largest potential amplification covered 19 genes of chromosome 10.

To more accurately determine the edges of the chromosome 10 amplification, as well as to verify it using an independent experimental technique, qPCR primers were designed to several genes both within and near the edges of the perceived amplification (Table 1). Relative to D6, D6.QHS2400x2 had an elevated copy number for all genes tested from PF10_0279 to PF10_0299 (Figure 3). Although relative copy numbers in the region ranged from 1.51-1.77, the entire area is most likely duplicated. The genomic DNA tested was from a population rather than a cloned parasite, which could explain such variation in copy number across the region.

Other lab strains were also tested for the chromosome 10 amplification (Table 5) relative to 3D7 Oxf, which was shown to have a copy number of ~1 for all of the genes

amplified in D6.QHS2400x2 relative to D6. Both wildtype W2 (JD) and I55S, a W2 (JD)-derived strain selected for resistance to an aminoquinoline (St. Jude reference number SJ000311327), showed a copy number around 1 for the 6 genes amplified in D6.QHS2400x2 tested. These results indicate that the amplification does not commonly occur in other strains of *Plasmodium falciparum* and is specific to artemisinin drug pressure. W2-5x, which was selected multiple times on low levels of artemisinin (200 ng/mL versus 2400 ng/mL for D6.QHS2400x2), showed evidence of a smaller amplification on chromosome 10, as PF10_0286 and PF10_0299 are not amplified, but PF10_0292 is. The exact edges of this apparent amplification were not determined, but could extend anywhere from PF10_0287 to PF10_0298.

Unexpectedly, amplification of PF10_0292 was also detected in W2 (DK), the wildtype W2 strain used in Dennis Kyle's lab, even though this gene was not amplified in W2 (JD), the wildtype W2 strain used in Joe DeRisi's lab. Further investigation indicated that this strain retained its sensitivity to artemisinin in recovery assays, confirming that it was not mislabeled W2-5x (Emily Wilson, unpublished results). However, the history of this particular strain of W2 is ambiguous and it is possible that the parasites were exposed to artemisinin at some point in time.

The lab-adapted patient isolate TM91c235 was selected twice for resistance to 240 ng/mL artelinic acid, yielding TM91c235.AL240x2. Neither TM91c235 nor TM91c235.AL240x2 showed qPCR evidence of an amplification on chromosome 10, although genes from PF10_0293 to PF10_0298 were not tested.

Discussion:

Here we have developed new genetic leads in the hunt for the element responsible for artemisinin resistance in *Plasmodium falciparum* by sequencing the entire genome of an *in vitro* selected parasite population, D6.QHS2400x2. Classically, pathogen drug resistance can arise by mutation of the drug's target (loss of function), or enhanced metabolism or of efflux the drug (gain of function). These mechanisms are most likely to arise from changes in the amino acid composition of proteins. In addition, the responsible element is rapidly selected for in resistant organisms and usually occurs at very low frequency in wildtype populations, as such resistance elements are often associated with fitness costs when drug is not present. Of all the SNPs discovered in D6.QHS2400x2 relative to D6, the 5 confirmed here represent the most likely to be responsible for artemisinin resistance based on these criteria. However, all of the SNPs in D6.QHS2400x2 remain candidates.

Several potential amplifications were also found in D6.QHS2400x2 (Table 4). However, only the apparent amplification on chromosome 10 encompassed more than 10 Kb of sequence and also had the median coverage well over 100 reads/bp, indicating that it was most likely to be a true amplification. Indeed, this amplification was verified in D6.QHS2400x2, but not in another strain selected for resistance to another class of drugs. However, the link between the chromosome 10 amplification and artemisinin resistance is ambiguous. The unexpected amplification of PF10_0292 in W2 (DK) could indicate that amplification of this portion of chromosome 10 is incidental and unrelated to acquisition of artemisinin resistance, or that W2 (DK) was exposed to artemisinin at some point in its history, gained resistance through amplification of this portion of chromosome 10, and then lost resistance through a secondary, compensatory mutation.

To truly establish a causal relationship between the chromosome 10 amplification and resistance to artemisinin, each encompassed gene will have to be expressed from a plasmid transfected into wildtype parasites, mimicking individual amplification of each. Stable transfectants will then have to be tested for resistance to artemisinin to identify the responsible gene.

Table 1 – Primers for sequencing library preparation, SNP confirming PCRs, and chromosome 10 amplification confirming qPCRs.

| Illumina Library Prep | | |
|---|--------------------------------------|--|
| Adapter | Primer | Primer Sequence |
| Sol-L-GsuI-T | short-SolL-GsuI-T | 5'-TACACGACGCTCTTCTGGAGT-3' |
| | phos-short-SolL-GsuI-revcomp-6Camino | 5'-/5Phos/CTCCAGGAAGAGCGTCGTGTA/3AmM/-3' |
| Sol-S-short-biotin-T | biotin-short-PE-SolS-T | 5'-/5Biosg/CTGCTGAACCGCTCTTCCGATCTT-3' |
| | phos-short-PE-SolS-revcomp | 5'-/5Phos/AGATCGGAAGAGCGGTTTCAGCAG/3AmM/-3' |
| | biotin-short-Sol-L-GsuI | 5'-/5Biosg/TACACGACGCTCTTCTGGAG-3' |
| | full-PE-Sol-S | 5'-CAAGCAGAAGACGGCATAACGAGATCGGTCTCGCATTCTGCTGAACCGCTCTTCCGATCT-3' |
| | full-Sol-L-GsuI | 5'-AATGATACGGCGACCACCGAGATCTACACTCTTCCCTACACGCTCTTCTGGAG-3' |
| | Sol-SeqPrimer | 5'-CACTCTTCCCTACACGACGCTCTTCTGGAG-3' |
| | PE-SolS-SeqPrimer | 5'-CGGTCTCGGCATTCTGCTGAACCGCTCTGCCGCTCT-3' |
| SNP Confirmation by PCR | | |
| Primer | Primer Sequence | |
| MAL13P1.298_F | 5'-TCCGTGTACCGAGCATTGTA-3' | |
| MAL13P1.298_R | 5'-TGGAGCTTCGTCATTGTGTTTC-3' | |
| PF13_0238_F | 5'-TTGTACAATCGTACTCTTTCCATTTC-3' | |
| PF13_0238_R | 5'-TCTCCATCAATTATGAATACCAACA-3' | |
| PFC0320w_F | 5'-TGTTACGTTTCATCTTGTAGGAA-3' | |
| PFC0320w_R | 5'-TGGATGAAATGCTTGACGAA-3' | |
| PFE1155c_F | 5'-AACTTTTCCAACCCCTTGT-3' | |
| PFE1155c_R | 5'-TGTTGAAACATTTCTCACCAAAA-3' | |
| PFF0275c_F | 5'-TCATCCTCAACCATTAATATAGCC-3' | |
| PFF0275c_R | 5'-TGTGTTTACATTTGATGGTCCAA-3' | |
| Chromosome 10 Amplification Confirmation by qPCR | | |
| Primer | Primer Sequence | |
| PF10_0275_rt-2F | 5'-CAAATGGAAAGACGCAATACC-3' | |
| PF10_0275_rt-2R | 5'-CGTTCCAGTTATCCATCCAGA-3' | |
| PF10_0294_rt-2F | 5'-CGTCCTGAATATCCACCTGAA-3' | |
| PF10_0294_rt-2R | 5'-TCACACTCTGCATTTCTGACG-3' | |
| PF10_0277F | 5'-TTTTTGAGGAAGCCTTTCTTTT-3' | |
| PF10_0277R | 5'-GCTGGAAAATAACCGCAA-3' | |
| PF10_0300F | 5'-TCGAAGTTGTGTTGCTTTTAATG-3' | |
| PF10_0300R | 5'-TAATTTGCCACACAGCAA-3' | |
| PF10_0307F | 5'-GGGAAATGTGTGCACAAGAA-3' | |
| PF10_0307R | 5'-CAAGCATGTTGGGGATAAGG-3' | |
| PF10_0286F | 5'-GCCATTCATCCATTTCTGTT-3' | |

| | |
|-------------------|--------------------------------|
| PF10_0286R | 5'-CAACTTGAAGGATTTTCGTTCC-3' |
| PF10_0296F | 5'-AACATTTTCACGCGACTTCC-3' |
| PF10_0296R | 5'-TGTGCGTTTTGCTCCAATAA-3' |
| PF10_0299F | 5'-TTCATTGCATCCTTGATTGG-3' |
| PF10_0299R | 5'-AATGCACCCTCACCAGGATA-3' |
| PF10_0278F | 5'-TTTCACTGAAGACGCCATGA-3' |
| PF10_0278R | 5'-TTCTTGTAGCTTGGGAGGTTG-3' |
| PF10_0279F | 5'-ATCCGGCAAATTCTCACATC-3' |
| PF10_0279R | 5'-GGAAGCGAAAAACCATAACG-3' |
| PF10_0282F | 5'-TGTTGCAATTTCTGGATTCCG-3' |
| PF10_0282R | 5'-GGATGTATAATCCTTCTGGACACA-3' |
| PF10_0285F | 5'-TGAACAAACCGAAAAAGGAA-3' |
| PF10_0285R | 5'-AGGGAGATATGTCCAGAAGGTG-3' |
| OFC288(chitinase) | 5'-TGTTTCCTTCAACCCCTTTT-3' |
| OFC289(chitinase) | 5'-TAATCAAACCCGTCTGCTC-3' |

Table 2 – Statistics for Illumina short read datasets generated for D6.QHS2400x2 and D6.

| | Strains | |
|------------------------|----------------|----------------|
| | D6 | D6.QHS2400x2 |
| Uniquely Matched Reads | 40467419 | 29562729 |
| Mean Coverage | 64.44 reads/bp | 64.11 reads/bp |
| Standard Deviation | 59.14 | 56.99 |
| Median Coverage | 69.86 reads/bp | 49.55 reads/bp |

Table 3 – Non-synonymous single nucleotide polymorphisms (SNPs) in exons of non-antigenic, protein-coding genes. P = D6, while R = D6.QHS2400x2. Highlighted rows indicate most the most promising SNPs that were confirmed by independent PCR.

| <u>Position</u> <u>(Gene ID)</u> | <u>Description</u> | <u>P</u> | <u>R</u> | <u>P Reads</u> | <u>R Reads</u> | <u>AA-</u> <u>change</u> | <u>AA</u> <u>position</u> | <u>R</u> <u>PCR</u> |
|-------------------------------------|---|----------|----------|----------------|----------------|-----------------------------|------------------------------|------------------------|
| 13: 2394027 (MAL13P1.298) | hypothetical protein, conserved | G | A | 68 (97%) | 22 (100%) | Gly->Asp | 573 | A |
| 13: 1726407 (PF13_0238) | kelch domain containing protein | C | T | 76 (100%) | 36 (100%) | Glu->Lys | 207 | T |
| 14: 903806 (PF14_0214) | hypothetical protein | T | C | 7 (57%) | 6 (83%) | Ile->Thr | 1158 | |
| 14: 2760111 (PF14_0644) | hypothetical protein | A | C | 6 (50%) | 19 (84%) | Cys->Gly | 1325 | |
| 3: 323233 (PFC0320w) | hypothetical protein, conserved | A | T | 98 (99%) | 41 (100%) | Asn->Ile | 138 | T |
| 3: 715601 (PFC0770c) | kinesin-related protein, putative | G | T | 7 (57%) | 35 (94%) | Asp->Glu | 817 | |
| 3: 911306 (PFC0965w) | hypothetical protein, conserved | C | G | 11 (63%) | 4 (100%) | His->Asp | 573 | |
| 3: 911304 (PFC0965w) | hypothetical protein, conserved | G | A | 8 (87%) | 4 (100%) | Cys->Tyr | 572 | |
| 3: 1003453 (PFC1075w) | hypothetical protein | T | G | 3 (66%) | 5 (100%) | Leu->Val | 277 | |
| 4: 1134307 (PFD1195c) | conserved protein, pseudogene | T | G | 3 (66%) | 7 (85%) | Lys->Thr | 53 | |
| 5: 963839 (PFE1155c) | mitochondrial processing peptidase alpha subunit, putative | T | C | 83 (98%) | 27 (100%) | Ser->Gly | 331 | C |
| 5: 1106592 (PFE1325w) | hypothetical protein, conserved | C | A | 4 (50%) | 3 (100%) | Pro->Thr | 1798 | |
| 6: 230387 (PFF0275c) | nucleoside diphosphate kinase, putative Plasmodium exported protein | C | G | 82 (97%) | 24 (95%) | Asp->His | 1076 | G |
| 6: 1298105 (PFF1510w) | (PHISTb), unknown function Plasmodium exported protein | G | A | 3 (66%) | 3 (100%) | Asp->Asn | 43 | |
| 8: 62843 (MAL8P1.163) | (PHISTa), unknown function surface-associated | A | T | 7 (57%) | 10 (100%) | Ser->Thr | 98 | |
| 8: 71941 (MAL8P1.162) | interspersed gene 8.3 (SURFIN8.3) surface-associated | A | T | 3 (66%) | 4 (100%) | Phe->Tyr | 269 | |
| 8: 1327999 (MAL8P1.1) | interspersed gene 8.1, (SURFIN8.1) | A | G | 7 (57%) | 11 (81%) | Met->Val | 502 | |

Table 4 – Potential amplifications with median coverage >100 reads/bp in D6.QHS2400x2 after smoothing. The coordinates correspond to the edges of the smoothing window.

| Coordinates | Median Coverage (5 Kb window, reads/bp) |
|------------------------|---|
| Chr4: 350000-355000 | 104.4 |
| Chr7: 245000-250000 | 100.2 |
| Chr7: 525000-535000 | 102.2-102.8 |
| Chr10: 1175000-1245000 | 100.2-132.4 |

Table 5 – qPCR data for chromosome 10 in other lab strains. W2(JD) is wildtype W2 cultured in our lab, while W2 (DK) is wildtype W2 from Dennis Kyle’s lab. I55S was selected for resistance to aminoquinoline SJ000311327 (St. Jude reference number) from a W2 (JD) background. W2-5x was selected on 200 ng/mL of dihydroartemisinin 5 times after initial selection on 240 ng/mL artelinic acid. TM91c235 is a lab-adapted patient isolate from Thailand, and TM91c235.AL240x2 is the same strain recovered from 2 selections with 240 ng/mL artelinic acid.

| Gene | Strain (relative to 3D7) | | | | | |
|-----------|--------------------------|------|---------|-------|----------|------------------|
| | W2 (JD) | I55S | W2 (DK) | W2-5x | TM91c235 | TM91c235.AL240x2 |
| PF10_0282 | - | - | 0.81 | 0.89 | 0.81 | 0.86 |
| PF10_0286 | 0.95 | 0.72 | 0.91 | 0.90 | 0.95 | 0.94 |
| PF10_0292 | 1.00 | 1.01 | 1.82 | 1.51 | 0.65 | 0.89 |
| PF10_0294 | 0.93 | 1.03 | - | - | - | - |
| PF10_0295 | 0.91 | 0.93 | - | - | - | - |
| PF10_0296 | 1.01 | 0.85 | - | - | - | - |
| PF10_0299 | 0.87 | 0.86 | 0.80 | 0.97 | 0.89 | 0.90 |

Figure 1 – Selection of D6.QHS2400x2 from D6. Each point represents recovery of a parasite population from the indicated level of drug selection with dihydroartemisin (DHA). The red arrow indicates the D6.QHS2400x2 population used for sequencing.

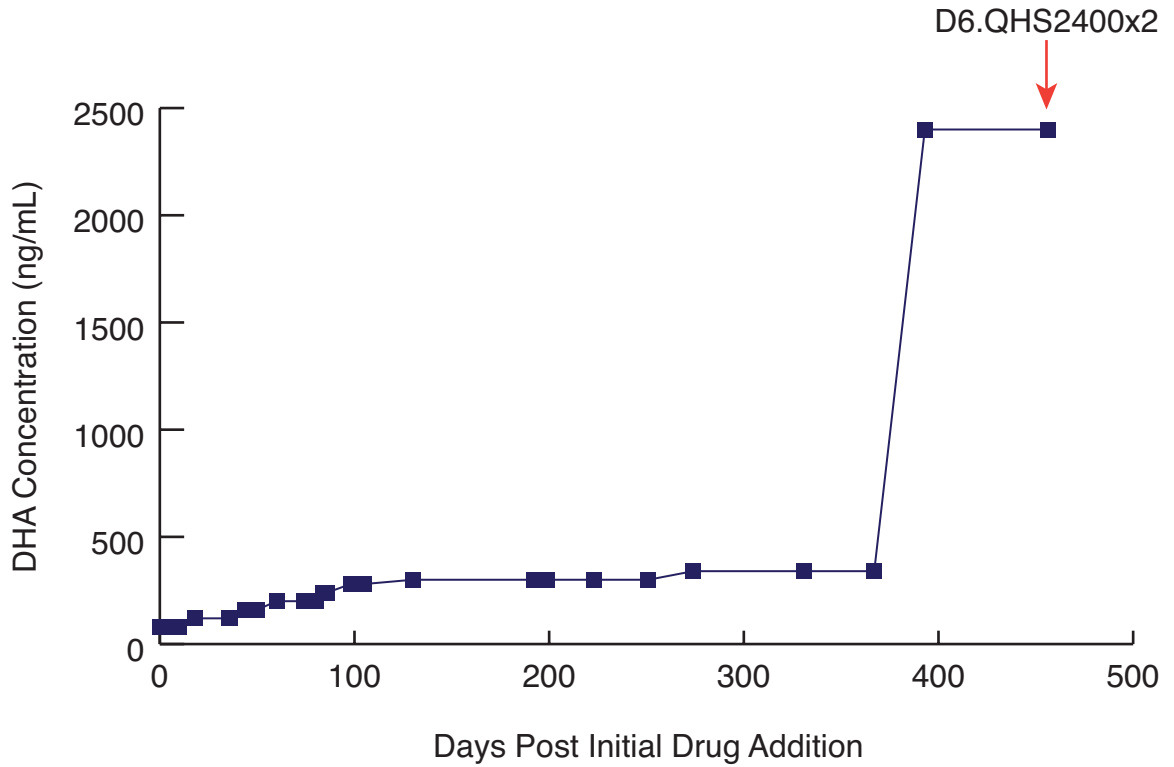


Figure 2 – Distribution of coverage for each bp in the genome for A) D6 and B) D6.QHS2400x2.

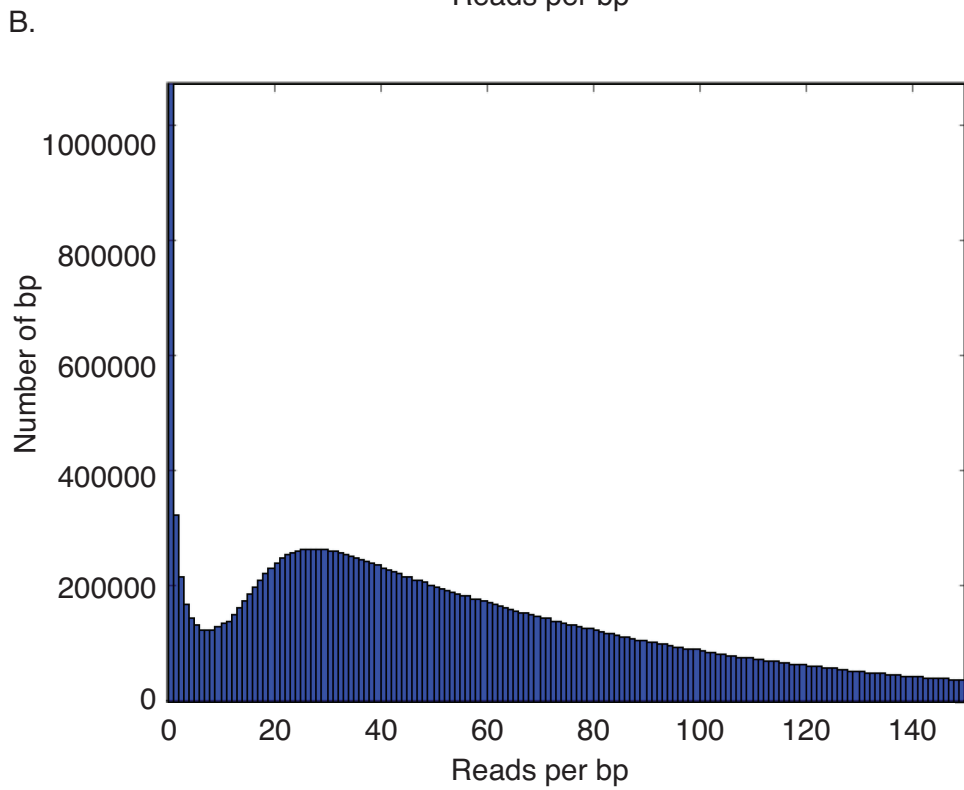
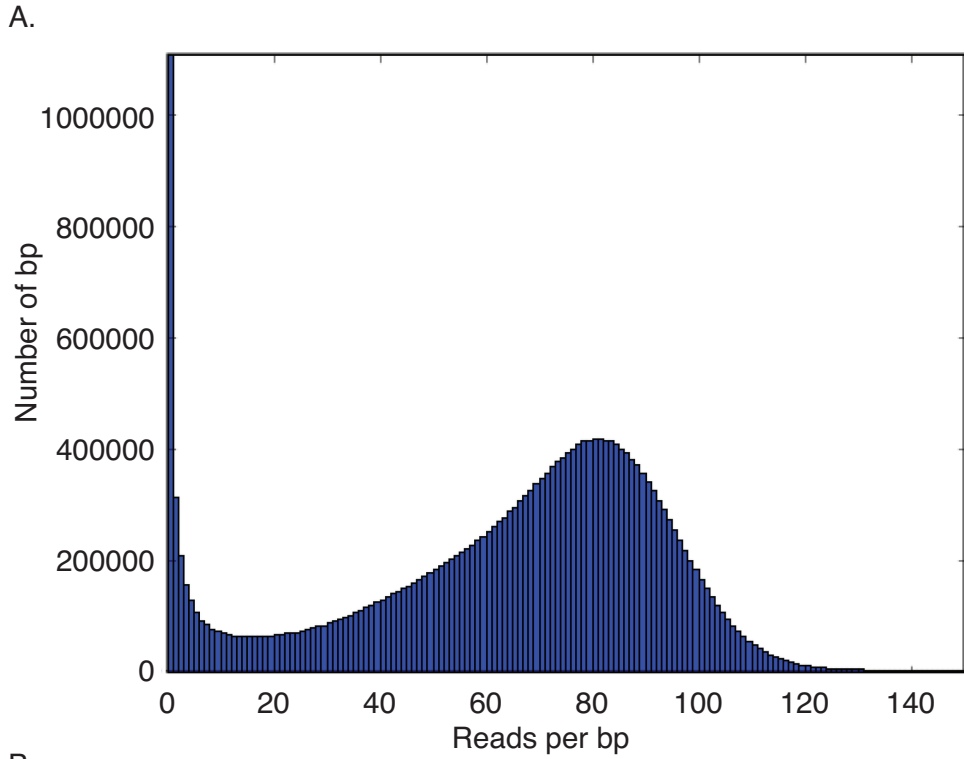
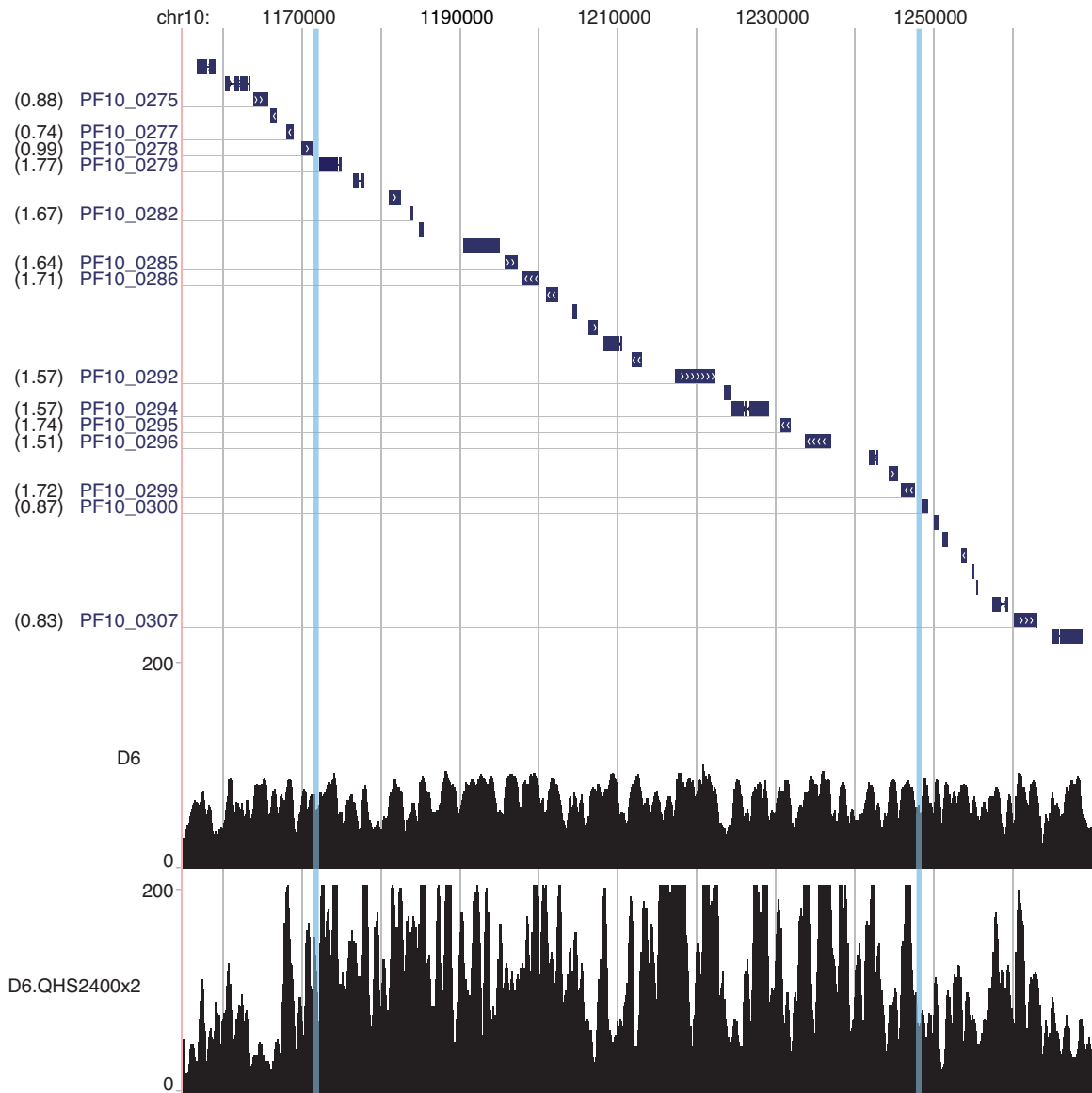


Figure 3 – qPCR verification of the amplification on chromosome 10 for D6.QHS2400x2. Chromosome coordinates are shown at the top, with PlasmoDBv5 gene models represented by dark blue bars below. The numbers in parentheses represent the number of copies of each gene in D6.QHS2400x2 relative to D6. Based on these numbers, the light blue lines represent the boundaries of the amplification. Shown below the gene models are the smoothed histograms of coverage for D6 and D6.QHS2400x2.



References:

1. Hay, S.I., Guerra, C.A., Gething, P.W., Patil, A.P., Tatem, A.J., Noor, A.M., Kabaria, C.W., Manh, B.H., Elyazar, I.R.F., Brooker, S. et al. (2009) A world malaria map: Plasmodium falciparum endemicity in 2007. *PLoS Med*, **6**, e1000048, 10.1371/journal.pmed.1000048.
2. WHO | World Malaria Report 2009 Available at: http://www.who.int/malaria/world_malaria_report_2009/en/index.html [Accessed June 14, 2010].
3. Mackinnon, M.J. and Marsh, K. (2010) The selection landscape of malaria parasites. *Science*, **328**, 866-871, 10.1126/science.1185410.
4. WHO | Guidelines for the treatment of malaria, second edition Available at: <http://www.who.int/malaria/publications/atoz/9789241547925/en/index.html> [Accessed October 7, 2010].
5. Mercer, A.E. (2009) The role of bioactivation in the pharmacology and toxicology of the artemisinin-based antimalarials. *Curr Opin Drug Discov Devel*, **12**, 125-132.
6. ter Kuile, F., White, N.J., Holloway, P., Pasvol, G. and Krishna, S. (1993) Plasmodium falciparum: in vitro studies of the pharmacodynamic properties of drugs used for the treatment of severe malaria. *Exp. Parasitol*, **76**, 85-95.
7. Chen, P.Q., Li, G.Q., Guo, X.B., He, K.R., Fu, Y.X., Fu, L.C. and Song, Y.Z. (1994) The infectivity of gametocytes of Plasmodium falciparum from patients treated with artemisinin. *Chin. Med. J*, **107**, 709-711.
8. Antimalaria studies on Qinghaosu (1979) *Chin. Med. J*, **92**, 811-816.
9. Olliaro, P.L., Haynes, R.K., Meunier, B. and Yuthavong, Y. (2001) Possible modes of

- action of the artemisinin-type compounds. *Trends Parasitol*, **17**, 122-126.
10. Eckstein-Ludwig,U., Webb,R.J., Van Goethem,I.D.A., East,J.M., Lee,A.G., Kimura,M., O'Neill,P.M., Bray,P.G., Ward,S.A. and Krishna,S. (2003) Artemisinins target the SERCA of Plasmodium falciparum. *Nature*, **424**, 957-961, 10.1038/nature01813.
 11. Li,W., Mo,W., Shen,D., Sun,L., Wang,J., Lu,S., Gitschier,J.M. and Zhou,B. (2005) Yeast model uncovers dual roles of mitochondria in action of artemisinin. *PLoS Genet*, **1**, e36, 10.1371/journal.pgen.0010036.
 12. Dondorp,A.M., Nosten,F., Yi,P., Das,D., Phyoo,A.P., Tarning,J., Lwin,K.M., Ariey,F., Hanpithakpong,W., Lee,S.J. et al. (2009) Artemisinin resistance in Plasmodium falciparum malaria. *N. Engl. J. Med*, **361**, 455-467, 10.1056/NEJMoa0808859.
 13. Carrara,V.I., Zwang,J., Ashley,E.A., Price,R.N., Stepniewska,K., Barends,M., Brockman,A., Anderson,T., McGready,R., Phaiphun,L. et al. (2009) Changes in the treatment responses to artesunate-mefloquine on the northwestern border of Thailand during 13 years of continuous deployment. *PLoS ONE*, **4**, e4551, 10.1371/journal.pone.0004551.
 14. Noedl,H., Socheat,D. and Satimai,W. (2009) Artemisinin-resistant malaria in Asia. *N. Engl. J. Med*, **361**, 540-541, 10.1056/NEJMc0900231.
 15. Anderson,T.J.C., Nair,S., Nkhoma,S., Williams,J.T., Imwong,M., Yi,P., Socheat,D., Das,D., Chotivanich,K., Day,N.P.J. et al. (2010) High heritability of malaria parasite clearance rate indicates a genetic basis for artemisinin resistance in western Cambodia. *J. Infect. Dis*, **201**, 1326-1330, 10.1086/651562.
 16. Imwong,M., Dondorp,A.M., Nosten,F., Yi,P., Mungthin,M., Hanchana,S., Das,D.,

- Phyo,A.P., Lwin,K.M., Pukrittayakamee,S. et al. (2010) Exploring the contribution of candidate genes to artemisinin resistance in *Plasmodium falciparum*. *Antimicrob. Agents Chemother*, **54**, 2886-2892, 10.1128/AAC.00032-10.
17. Chavchich,M., Gerena,L., Peters,J., Chen,N., Cheng,Q. and Kyle,D.E. (2010) Role of *pfmdr1* amplification and expression in induction of resistance to artemisinin derivatives in *Plasmodium falciparum*. *Antimicrob. Agents Chemother*, **54**, 2455-2464, 10.1128/AAC.00947-09.
18. PlasmoDB: An integrative database of the *Plasmodium falciparum* genome. Tools for accessing and analyzing finished and unfinished sequence data. The Plasmodium Genome Database Collaborative (2001) *Nucleic Acids Res*, **29**, 66-69.
19. Pfaffl,M.W. (2001) A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res*, **29**, e45.
20. Kent,W.J. (2002) BLAT--the BLAST-like alignment tool. *Genome Res*, **12**, 656-664, 10.1101/gr.229202. Article published online before March 2002.
21. Gardner,M.J., Hall,N., Fung,E., White,O., Berriman,M., Hyman,R.W., Carlton,J.M., Pain,A., Nelson,K.E., Bowman,S. et al. (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature*, **419**, 498-511, 10.1038/nature01097.

Chapter 5: Dihydroartemisinin induces transcriptome arrest in drug susceptible and resistant *Plasmodium falciparum*

This chapter is a summary of work done by:

Nelson C, Sorber K, Tucker M, LaCrue A, Azizan A, Kyle D, DeRisi JL.

Author contributions:

Matt Tucker derived all artemisinin resistant strains *in vitro* and performed all recrudescence assays. Azliyati Azizan performed the pilot drug treatment timecourse and extracted RNA from the samples. Katherine Sorber amplified the samples, and Katherine Sorber and Chris Nelson hybridized the pilot samples to microarrays. Chris Nelson, Katherine Sorber, Matt Tucker, and Alexis LaCrue performed the large drug treatment timecourse, extracted RNA from the samples, and amplified the RNA. Katherine Sorber and Chris Nelson hybridized the amplified RNA to microarrays. Chris Nelson performed all analysis of the pilot and large timecourse microarray data, and re-processed samples where necessary. Dennis Kyle and Joseph L. DeRisi conceived of and supervised the project.

Joseph L. DeRisi, Thesis Advisor

Abstract:

Artemisinin combination therapy (ACT) is required for the treatment of multidrug resistant falciparum malaria. Although artemisinin and its derivatives rapidly clear parasitemia and reduce malaria symptoms, frequent recrudescence is observed when artemisinins are administered alone. Recent evidence of clinical resistance in Asia as well as the established recrudescence problem highlight the urgent need to better understand of the mechanism(s) of action and resistance to the artemisinin drug class. We have conducted experiments to assess the effect of dihydroartemisinin (DHA) on *P. falciparum* and have identified a unique ring-like stage with concomitant transcriptome arrest not previously observed with other antimalarial drugs. The transcriptome of DHA-treated parasites is similar to 10-12 hr rings and coincides with morphological observations of arrest at early ring stage. A set of 113 genes were strongly induced or repressed in this state, and the DHA-induced state occurred in both DHA resistant and susceptible lines. DHA-induced dormancy persisted for three days before transcriptome release and normal growth resumed, with resistant parasites exiting dormancy slightly earlier than the sensitive parental clone. Additional analysis of DHA-treated sensitive and resistant strains revealed substantial baseline and drug-inducible expression differences between these strains that could form the basis of the resistance phenotype. These data demonstrate novel ring-like dormancy and the first evidence for transcriptome arrest following exposure to any antimalarial drug, as well as possible new leads in understanding the artemisinin resistance phenotype.

Introduction:

Malaria is a disease with a large global impact. In 2007, 1.38 billion people, mostly in Africa, central and southeast Asia, and South America, were at risk of infection with *Plasmodium falciparum* (1). In 2008, malaria resulted in 243 million illnesses and over 800,000 deaths (2). During the past four decades, *P. falciparum*, the most lethal malaria parasite infecting humans, has developed resistance to every commonly available antimalarial drug. Many countries in Africa are now faced with high failure rates when using chloroquine or sulfadoxine/pyrimethamine, the former first and second line drugs, respectively. The World Health Organization has reacted to this monumental problem by supporting the widespread use of artemisinin combination therapy (ACT) for malaria in endemic areas.

Artemisinin compounds represent the most rapidly acting antimalarial drugs. They contain an endoperoxide bridge that is essential for activity (3), and all possess remarkable activity against *P. falciparum* isolates that are multiply resistant to other antimalarial drugs. As a class, the artemisinin drugs are active at low nM concentrations, and the primary human metabolite dihydroartemisinin (DHA) is the most active (IC-50 range 2.2-3.9 nM). *In vivo*, artemisinins produce faster parasite and fever clearance times than any other antimalarial drug (4) and effectively reduce transmission of malaria by reducing gametocyte carriage (5, 6). However, these drugs have short half lives (~45min), which may be responsible for the frequent recrudescence observed in patients after monotherapy treatment.

Given the global adoption of ACTs, considerable research has been devoted to their mechanism of action on *Plasmodium spp.*; however, it remains controversial.

Several proposed hypotheses include non-specific alkylation or oxidation of parasite proteins by free radical byproducts of endoperoxide breakdown (7), specific inhibition of a SERCA-like Ca²⁺-dependent ATPase (8), inhibition of the parasite's mitochondrial electron transport chain (9), or drug reduction in a two electron transfer process that is not mediated by heme (10).

Perhaps equally controversial is the debate on the emergence of artemisinin resistance and molecular determinants associated with this phenotype. Increased parasite clearance times in patients treated with ACTs constitute the bulk of evidence for emerging resistance to artemisinins in the field. Historically, the area of Southeast Asia and specifically the Thai-Cambodian border has been a focus of malaria drug resistance (11, 12). A recent report from the same region demonstrates longer parasite clearance times in western Cambodia in comparison with the Thai-Burmese border (13). This study avoided longstanding arguments over whether such data can be attributed to inefficacy of ACT partner drugs by demonstrating longer parasite clearance times for both ACTs and artesunate administered as a monotherapy. In addition, Carrara et al. (14) showed a longitudinal increase in parasite clearance times since ACTs were introduced at the Thai-Burmese border in 1995, along with an increase in gametocyte carriage of treated patients. Unfortunately, these field observations have been difficult to translate into tangible phenotypes such as an increase in 50% inhibitory concentrations (IC₅₀s), a traditional measure of *in vitro* drug susceptibility. A few recent studies observed a trend of slightly elevated IC₅₀s in isolates from west to east Asia (15, 16); however, studies attempting to directly correlate increased parasite clearance time with an increase in IC₅₀ remain inconclusive (13, 15).

Recently, a novel hypothesis has emerged that may explain frequent recrudescence following artemisinin monotherapy treatment. Studies have shown that ring-stage parasites of *Plasmodium falciparum* become dormant following exposure to artemisinin and its derivatives (17, 18). In this study, we conducted transcriptome analysis of synchronous and asynchronous cultures of artemisinin sensitive and *in vitro* derived artemisinin resistant strains of *P. falciparum* after a 6-hour exposure to low, but physiologically relevant levels of DHA. Remarkably, we found both sensitive and resistant parasites enter dormancy and arrest their transcriptomes in a ring-like state. Transcriptional analysis identified genes that are differentially expressed in DHA-induced dormant stages. Additional constitutive and DHA-induced transcriptional differences were identified in the resistant versus sensitive clones. Interestingly, morphological and correlogram analysis suggest resistant parasites may exit dormancy slightly before sensitive parasites. These data provide the first evidence for transcriptome arrest in response to an antimalarial drug.

Materials and Methods:

Pilot study:

Synchronized W2 ring-stage parasites at 4% hematocrit and ~2% parasitemia were exposed to 28.4 ng/mL dihydroartemisinin (DHA) for six hours, after which drug was washed out and parasites were returned to culture. Time points were taken before drug addition (T=0), six hours after treatment (T=6, washout), and 27 hours after treatment (T=27). RNA extraction and amplification, as well as microarray analysis of these samples was carried out as described below.

***In vitro* drug selection and parasite cloning:**

P. falciparum laboratory isolate D6 was subjected to selection first with artemisinin (19), then with dihydroartemisinin (see Chapter 4, Tucker unpublished results). The parent D6 population, as well as the resulting D6.QHS2400x5 population, was cloned by limiting dilution (20). Both the C11 D6 clone and the C9 D6.QHS2400x5 were tested for appropriate sensitive and resistant responses to DHA, respectively.

Cell culture:

Plasmodium falciparum D6 clone C11 and D6.QHS2400x5 clone C9 parasites were cultured at 2% hematocrit in RPMI 1640 (Invitrogen Corp., Carlsbad, CA) containing 10% heat-inactivated human plasma. When parasites were mostly in the ring stage, one half of each culture was sorbitol synchronized – the other half of each culture was maintained as mixed stages for the duration of the experiment. Synchronous cultures were synchronized twice in the subsequent cell cycle 8 hours apart, then 10 hours apart in the next cycle, and finally 12 hours apart in the cycle after that, for a total of 7 synchronizations over a 192-hour period. Following the last synchronization, all cultures (mixed and synchronous) were brought up to 4% hematocrit with uninfected blood. Invasion in the synchronous cultures was monitored by smear every hour starting approximately 26 hours after the last synchronization. Maximum invasion (number of rings = number of schizonts) was recorded and the cultures were synchronized one last time 2 hours after maximum invasion. Both mixed stage and synchronous cultures were split down to 2-3% parasitemia in uninfected blood.

Drug Treatment:

8 hours post-invasion, 6 mL of each culture was harvested as the T=0 sample before drug addition. 44 mL of 37°C 1xPBS was added to the cells and the entire volume was spun down at 500xg for 5 min. The cell pellet was washed with an additional 25 mL warm PBS, then spun down again. After removal of the supernatant, pellets were flash frozen in liquid nitrogen and transferred to -80°C. 200 ng/mL final concentration of dihydroartemisinin (DHA) in DMSO was added to the remaining experimental culture of each strain, while an equal volume of DMSO was added to untreated control culture 8 hours post-invasion (T=0). After 6 hours of drug or DMSO treatment (T=6), another 6 mL aliquot of culture was harvested from each condition. Remaining culture was spun down and drug/DMSO was washed off twice with 37°C RPMI wash (no plasma added). Parasites were returned to culture at 4% hematocrit in 10% plasma RPMI. Subsequent timepoints were taken at 12, 18, 24, 32, 40, and 48 hours after drug addition for each condition, except for mixed stage DMSO control cultures, which were sampled only at 0, 24, and 48 hours after addition. Synchronous DHA-treated cultures were carried out past 48 hours in order to observe recrudescence and timepoints were taken at 56 hours, and then every 12 hours after that until smears showed ~3% normal parasitemia.

RNA extraction and amplification:

Total RNA was harvested from the frozen pellets using 10 mL Trizol (Invitrogen Corp., Carlsbad, CA) and 2 mL chloroform for every 1-2 mL of cell pellet. Aqueous phases were re-extracted with 1 volume of acid phenol (pH4.3) and 1 volume of

chloroform. Aqueous phases were extracted again with 1 volume of chloroform, then precipitated with 1/10th volume of 3M sodium acetate (pH5.2) and 1 volume of isopropanol. Where possible, 110 ng of total RNA was amplified and amino allyl labeled using the Amino Allyl MessageAmp II aRNA Amplification kit (one round, 14 hour IVT, Ambion, Austin, TX). For samples with < 110 ng total RNA yield, as much total RNA as possible was used for amplification. An amplified RNA pool representing transcripts expressed throughout the IDC was compiled as a reference.

Cy dye labeling and microarray hybridization:

2 µg aliquots of amplified pool RNA were coupled to Cy3, while up to 2 µg of each amplified sample RNA was coupled to Cy5 (GE Healthcare, Piscataway, NJ). Cy3 pool and Cy5 sample were competitively hybridized at 65 °C on a printed microarray containing 8,159 70-mer oligos that map to 5,338 ORFs annotated in PlasmoDB release 6.3. Prior to hybridization, microarrays were printed and post-processed as described in Bozdech et al. (21). After a minimum of 18 hours hybridizing, microarrays were washed in 65°C 0.6x SSC, 0.03% SDS, and then in room temperature 0.06x SSC. Spun dry arrays were then scanned on an Axon 4000B scanner using Axon Genepix v 6.0 and 6.1 software (Molecular Devices, Union City, CA).

Array analysis:

Microarrays were manually gridded and Cy3 and Cy5 intensity of each spot was extracted using Genepix software. Arrays were uploaded to Nomad v2.0 (<http://ucsf-nomad.sourceforge.net/>) where the data was normalized in bins of pixel intensity R^2 , and

then filtered to remove spots with “bad” or “missing” manual flags added during gridding and spots with sum of median intensities less than 500. The resulting ratio Cy5/Cy3 intensity tables were \log_2 transformed and re-centered about 0. Several arrays have technical replicates. For subsequent calculations, the results of replicate arrays without large artifacts were averaged together for the following samples: PTS0, PTS6, PTS12, PTS24, PTS32, PTS48, RTM40, RTM48, RTS0, RUS6. Sample abbreviations are P or R (D6 or D6.QHS2400x5), T or U (DHA treated or untreated), S or M (synchronous or mixed stage), and the hour post-treatment.

Var rifin and surfin genes were removed from the data in order to eliminate confounding strain-dependent antigenic variation effects. Re-centered arrays were compared with all timepoints of the HB3 intraerythrocytic developmental cycle by Pearson’s correlation (22).

The lists of genes up-regulated and down-regulated in the 48 hours after drug treatment in treated sensitive and resistant strains were generated by two class significance analysis of microarrays (SAM version 3.0, <http://www-stat.stanford.edu/~tibs/SAM/> Tusher 2006, (23)) after filtering for oligos with 70% of data available. We compared all synchronized time zero arrays to all arrested samples (T=6 to T=48, with the exception of T=18 which had poor quality arrays). The number of arrays from D6 was balanced by an equal number of arrays from D6.QHS2400x5 for each group in order to avoid strain-specific results. Seeking <1 expected false positive, we chose a delta score threshold of 0.665 with an overall false discovery rate (FDR) of 0.335%. Although no direct stipulation was made to ensure that resulting genes were DHA-responsive in the same direction in both D6 and D6.QHS2400x5, our delta-score

threshold was conservative enough that genes in Table 1 were either up-regulated or down-regulated in both strains. ORFs with multiple oligos are represented by the best scoring oligo. In Tables 1-3, fold change is expressed as the geometric mean of the fold change or fold induction from the timepoint(s) in question.

To look for constitutive differences between the resistant strain and parental strain, we compared time-matched data produced from untreated parasites after filtering for oligos with 90% of data available (Table 2). For each time point from 0-48 hr, we subtracted \log_2 transformed parental ratios from resistant ratios. The results of these calculations were grouped and subjected to one class SAM. A delta score threshold of 0.765 was chosen to yield an overall FDR of 0.7% and 0.73 expected false positives.

We also were interested in DHA-induced expression differences between the two strains after filtering for oligos with 70% of data available (Table 3). Array \log_2 ratio data from time points 6, 12, 24, 32 and 48 hr post drug addition were zero transformed by subtracting the time zero \log_2 ratios of the same strain. Again, T=18 was excluded due to poor quality arrays. The difference between zero transformed data for each strain was analyzed for significant outliers via SAM. Seeking <1 expected false positive, we chose at delta score threshold of 1.442, with an overall FDR of 1.17%.

Microscopy and smear counts:

Thin smears for each time point were independently read in a blinded fashion by two (after T=48) to three (T=0-T=48) people. Each person counted between 300-900 total red blood cells per slide. Parasitemia was calculated as the number of parasite-

infected cells per total number of erythrocytes counted, and stage of observed parasites (ring, troph, schizont, or dormant) was recorded.

Results:

DHA induces transcriptome arrest in W2 parasites:

To determine dihydroartemisinin's (DHA) effect on the transcriptome of *P. falciparum* parasites, synchronized rings of wildtype W2 parasites were exposed to 28.4 ng/mL (100nM) DHA for 6 hrs, then washed and returned to culture. Total RNA samples taken 6 hours after treatment (T=6, at the time of washout) as well as 27 hours after treatment (T=27) were analyzed by microarray. Surprisingly, these DHA-treated samples taken 21 hours apart had very similar transcriptomes, with a Pearson's correlation of 0.63. Parasites going through a normal cell cycle would be expected to undergo a huge transcriptome change from ring to trophozoite during the 21-hour separation (24, 21, 22). To simulate the expected result in normal parasites, the Pearson's correlation between the 6 and 27-hour post-invasion transcriptomes of normally growing HB3 parasites (22) was calculated using publically available data from the same microarray platform. This correlation was -0.69, indicating that such time divergent samples of the same culture should actually be anticorrelated rather than correlated.

Such initial analysis, as well as previous data (19, 17), led us to postulate that artemisinin could be causing a developmental transcriptome stall in the W2 parasites. To determine roughly where that stall occurred, array data from T=6 and T=27 were independently correlated with transcriptome data from every hour of the normal intraerythrocytic developmental cycle (IDC) of HB3 (Llinás et. al 2006, Figure 1). Our

data correlated best with HB3 12-13 hours post-invasion, suggesting there is an arrest in development most similar to the ring-stage following treatment with DHA.

Overview of timecourse experiment:

To probe DHA-induced dormancy over a prolonged period of time and to address how this arrested state might differ between resistant and sensitive strains, a longer time course with higher timepoint resolution was conducted using D6 and D6.QHS2400x5 clones representing DHA-sensitive and -resistant parasites, respectively. Wildtype D6 was initially selected with 80 ng/mL artelinic acid (AL) to produce resistant parasites (D6.AL80) as described in Chavich et al. (19). Recovered parasites were switched to pressure with DHA, eventually producing D6.QHS2400x5 (resistant to 2400 ng/mL DHA, see Chapter 4, Tucker unpublished results). Clones of the parental D6 and D6.QHS2400x5 were obtained by limiting dilution and used in the experiment illustrated in Figure 2. Both highly synchronous ring-stage cultures (approximately 8 hours after invasion) and asynchronous mixed stage cultures of each strain were treated with 200 ng/mL DHA, while matched controls split from the same starting culture were simultaneously treated with DMSO. Just before treatment, T=0 time points were collected. The second time point was taken 6 hours after treatment, at the same time that drug was washed out (T=6). Subsequent time points were collected at lengthening intervals. Mixed stage cultures were discontinued after 48 hours, while synchronous drug treated cultures were sampled until they began to recover by smear (T=116). RNA was isolated from time course samples and linearly amplified in parallel before hybridization to microarrays printed in-house (21).

Synchronous ring-stage cultures arrest transcriptome in ring-like state after DHA treatment:

Microarray data from the synchronous treated parasites, as well as their controls, was analyzed by comparing each time point array to published array results from every hour of the intraerythrocytic developmental cycle (IDC) of normally growing HB3 parasites (22). In Figure 5, DMSO-treated control parasites present a drifting sine wave of correlation whose peak correlation (0.73 ± 0.05 for parental and 0.64 ± 0.07 for the resistant strain) shifts across HB3's IDC as expected (Figure 5A and 5C). In stark contrast, DHA-treated D6 and D6.QHS2400x5 parasites failed to progress their transcriptomes past a state most closely resembling a normal ring, with peak correlation at 8-11 hours post invasion (hpi) (correl 0.63 ± 0.02) for DHA-sensitive D6 and at 9-11 hpi (0.55 ± 0.02) for DHA-resistant D6.QHS2400x5. These results confirm and significantly extend our observation of a DHA-induced transcriptome arrest in the W2 pilot experiment (Figure 1). Importantly, both parental and resistant clones independently exhibited transcriptome arrest consistent with a ring stage parasite. Also, the transcriptome correlation of DHA-treated parasites to normal IDC rings did not appreciably diminish across the prolonged period of dormancy of several days for either the DHA-sensitive strain D6 or the DHA-resistant strain D6.QHS2400x5. We believe this unique observation of transcriptome stall is related to the morphological observation of persistent small rounded parasites by smear at the same time points (Figures 3 and 4).

Interestingly, the period of transcriptome arrest by microarray lasted until $T=86$ post-drug for the treated D6, but appears to be up to 24 hours shorter ($T=62$ post drug) for D6.QHS2400x5, the resistant clone (Figure 5B and 5D). Such a difference in the

timing of release from DHA arrest could be the result of slight parasitemia inconsistencies between the cultures, or alternatively could represent a legitimate difference in how sensitive and resistant parasites recover from DHA treatment. The time points subsequent to both cultures' exits from arrest have reduced peak correlations in the transcriptome probably due to a loss of synchrony as individual parasites exit dormancy at different times and create a combined asynchronous population. This is consistent with observations on the timing and asynchrony of DHA-recovery by Teutscher et al. (17).

Asynchronous cultures exposed to DHA become dominated by a ring-like transcriptome:

An alternative explanation to the data presented thus far is that signal captured on the microarrays could represent residual intact RNA from dead or dying rings in the synchronized cultures. Therefore, we simultaneously exposed asynchronous cultures of the same D6 and D6.QHS2400x5 parasite clones to 200 ng/mL of DHA and harvested samples for microarray analysis. As expected, DMSO treated controls produced low peak correlations throughout the time course (0.18 ± 0.07 for D6 at 21 hpi and 0.16 ± 0.02 for D6.QHS2400x5 at 9 hpi) when compared to the synchronous HB3 IDC, reflecting dominance of no one parasite stage within these cultures (Fig 5A and C). If the so-called transcriptome arrest were due to dead or dying parasites, we expected that the asynchrony of DHA-treated cultures would also persist. In contrast, we found that DHA induced a convergence in treated mixed stage cultures on an arrested transcriptome that most closely resembles that of normal rings, with a peak correlation at 14-16 hpi of $r=0.40 \pm 0.10$ for parental D6, and a peak correlation at 9-13 hpi of $r=0.37 \pm 0.07$ for resistant D6.QHS2400x5 (Figure 6B and D). The difference between the hpi of peak

correlation during stall might be due to differences in stages represented in each mixed population at time zero. However, as with the synchronous cultures, both mixed stage DHA-sensitive D6 and DHA-resistant D6.QHS2400x5 exhibited this transcriptome arrest phenotype most closely resembling ring stage parasites.

Smears confirm presence of dormant parasites for timepoints exhibiting stalled transcriptome:

Thin smears were also collected for every time point to correlate morphology of parasites to microarray results (Figure 3, untreated not shown). For all treated cultures, there was a predominant shift in morphology from rings to “dormant forms” after DHA treatment, and the proportion of parasites with “normal” morphology (rings, trophozoites, schizonts) dropped dramatically. Dormant forms are small with a regular round outline on a Giemsa stained blood smear and differ from the collapsed nuclei of pyknotic bodies in their retention of a small circle of blue-staining material that is presumably cytoplasm (Figure 4). These forms were not seen in control cultures treated with DMSO (data not shown). Although DHA-treated mixed stage cultures were not carried out long enough to observe recrudescence, normal morphological forms gradually reappeared and finally dominated as part of a full recrudescence in both DHA-treated synchronous cultures.

DHA treatment induces up- or down-regulation of specific genes in both strains:

Specific genes were also reproducibly up-regulated or down-regulated during the putative dormant state following DHA exposure. To eliminate confounding developmental stage factors from this and subsequent individual gene analyses, we used

only the expression data from the treated synchronized cultures of D6 and D6.QHS2400x5. Using two-class significance analysis of microarrays (23), we compared all time zero synchronized data to data from all arrested time points before recrudescence was observed for either strain (T=6 to T=48). There is a slight IDC offset of ~2hrs between these two strains, which we reason is within sampling noise and not enough to flood our results with developmentally regulated genes. Classic antigenic gene families were removed from consideration for this and other analyses to control for strain-dependent effects. With parameters expected to yield less than one false positive, 21 genes that met our criteria were up-regulated in the arrested condition relative to the untreated condition and 134 genes that met our criteria were down-regulated (Table 1).

Interesting genes within this list include thioredoxin reductase, lysophospholipase, Myb domain transcription factors, and a putative cyclin dependent kinase. Thioredoxin reductase (PFI1170c) is up-regulated 9-fold in resistant parasites and 23-fold in the sensitive parasites in the first 48 hr after drug treatment. Thioredoxin reductases are antioxidant proteins complementary to the glutathione system (which the parasite lacks) that are involved in detoxifying different type of peroxides (25, 26). Therefore, it is reasonable to suspect that the *P. falciparum* thioredoxin reductase might be involved in DHA-related peroxide detoxification. Lysophospholipase (PF14_0017), an enzyme involved in catabolizing phosphatidyl choline, is up-regulated 8-fold in both strains in the first 48 hr after DHA treatment. Perhaps this up-regulation reflects a need to differentially remodel the parasite niche in the dormant state or to prepare the parasite for eventual recovery. Conversely, two Myb domain proteins (PF10_0327 and PF13_0088) are down-regulated after DHA treatment. PF13_0088 has a reported peak of protein

expression in the nucleus at the trophozoite stage (27). Interestingly, a putative cyclin dependent kinase (PFD0740w) normally expressed in mid ring stage is also down-regulated in the arrested transcriptome. This down-regulation could play a role in the mechanism of arrest by preventing progression through an as-yet-unknown cell cycle checkpoint.

Constitutive transcription differences between DHA-resistant and sensitive strains:

Use of both a DHA-sensitive and resistant strain in our experiment provided the opportunity to examine how their transcriptomes might differ. The genetic determinant of artemisinin resistance could result in constitutive up-regulation or down-regulation of some transcripts in resistant parasites compared to their sensitive counterparts. Therefore, we queried the microarray data by comparing time-matched data produced from untreated, control cultures. For each time point from 0-48 hr, we subtracted \log_2 transformed D6 ratios from D6.QHS2400x5 ratios. The results of these calculations were grouped and subjected to one class SAM with parameters chosen to produce 0.73 expected false positives. 75 genes were up-regulated in the D6.QHS2400x5 strain compared to parental D6, and 13 genes were down-regulated (Table 2). Many of the exported genes in these lists are likely due to strain-specific antigenic switching effects. Future study is merited into genes on this list as putative markers of artemisinin resistance. Interestingly, several genes clustered together on chromosome 10 appear to be up-regulated in D6.QHS2400x5 compared to D6.

DHA induces differential expression of genes in QHS resistant and sensitive parasites:

Genes differentially expressed between DHA-sensitive and DHA-resistant parasites during dormancy could potentially be involved in DHA resistance if this resistance phenotype is induced only after drug treatment (as opposed to constitutively present). To explore this possibility, the microarray data were analyzed for genes that showed large strain-specific changes from baseline expression after drug treatment. The difference between the zero-transformed resistant transcriptome inductions and zero-transformed parental transcriptome inductions was analyzed for significant outliers via SAM (23). For this purpose we used only data from time points 6, 12, 24, 32 and 48 hr post drug addition, which were more complete arrays. Choosing our list for less than one expected false positive, we arrived at a list of 13 genes more up-regulated in arrested resistant parasites and 31 genes more down-regulated in arrested resistant parasites (Table 3).

Interesting among this list is the relative down-regulation of an elongation factor subunit (PF11_0245) in resistant parasites, suggesting the possibility that kinetics of translation may be altered in the resistant dormant state. The relatively low induction of Pfcyc-2 cyclin-related protein (PFL1330c) in resistant parasites is very interesting in the context of putative cell-cycle arrest, although it is near our cutoff threshold.

Discussion:

Here we have described in molecular detail a long-lasting transcriptome arrest in *Plasmodium falciparum* induced by artemisinins, a drug class of tremendous clinical importance. The timing of this transcriptome pause coincides with the observation of morphological changes within the same cultures, from normal stages into what we term “dormant” parasites. Witkowski et al. observed similar morphological changes and transcriptome arrest upon artemisinin treatment, yet incorrectly concluded that the ability to enter a quiescent state is a hallmark of resistance to artemisinin (18). In contrast, our study demonstrates that dormancy occurs in both wild-type W2 and D6 (artemisinin sensitive strains) as well as in D6.QHS2400x5, a strain selected for high level resistance to artemisinin. Therefore, dormancy itself is not indicative of an artemisinin resistance phenotype, though as discussed later, certain characteristics of dormancy may differ between sensitive and resistant parasites.

Previous studies have shown that treatment of parasites with antimalarial drugs typically does not alter normal transcriptome progression through the IDC as assayed by expression microarrays even as parasites die (Ganesan 2008, Hu 2010). Thus we have shown the first evidence of transcriptome arrest in *P. falciparum* in response to an antimalarial drug. Additionally, Shock et al. demonstrated that *P. falciparum* mRNA decay rates are relatively rapid, with the average half-life of 9.5 minutes for ring-stage transcripts (28). It therefore seems unlikely that residual transcripts from dead parasites explain our transcriptome data, as it is characterized by a complete halt in the progression of the transcriptome, as well as a long-lasting, high correlation to normal HB3 ring-stage parasites. Furthermore, if the transcriptome arrest observed only reflected residual mRNA

from dead parasites, it would be difficult to explain the observed DHA-induced convergence upon a ring-like state in originally asynchronous cultures.

DHA treatment of mixed stage parasites not only determined that the measured dormant transcriptome is unlikely to result from dead or dying parasites, but also confirmed that the observed dormant state does not simply reflect the transcriptomes of parasites before exposure to DHA. Since the arrested transcriptomes from both treated synchronous and mixed stage cultures resemble a normal late ring transcriptome, we suggest that there is a window of time during the ring stage in which parasites can go dormant after DHA treatment. The pause in the transcriptome may be relative to the parasites' current transcriptional state at the time DHA is added, as long as the parasite is within the dormant window. This hypothesis might also explain why the peak transcriptome correlations for the asynchronous cultures are not as high as those of the synchronized dormant parasites.

Although still limited, our current data about the dormant state induced by artemisinin in *P. falciparum* is highly suggestive of a disruption in the cell cycle, perhaps caused by triggering a previously unknown natural checkpoint. The possibility that artemisinins induce cell cycle arrest has precedence in studies with human carcinoma cell lines. Willoughby et al. concluded that artemisinin disrupts transcription at the promoter of CDK4 in prostate cancer (29). In pancreatic cancer, T cells, and hepatoma cells artemisinins have been shown to disrupt cyclin levels and induce G1 arrest (30-32). Additionally Efferth and colleagues suggested that in yeast, artemisinins may induce the DNA repair checkpoint based on yeast mutants with increased artemisinin sensitivity (33). Though *Plasmodium spp* are known to encode cell cycle regulatory genes, there has

been doubt over the existence of any inducible checkpoint, as there is no previous molecular description of any growth arrest state in the cell cycle of the malaria parasite, despite multiple attempts to disturb the IDC with a variety of drugs and perturbations (34-36).

Importantly, our study was conducted with a low, but physiologically relevant concentration of DHA that was not selective between the two cloned strains. As a result, we were able to look for genes induced or suppressed during dormancy in both strains. These genes are likely involved in entry into or maintenance of the dormant state, and implicate oxidative stress, transcriptional, and cell remodeling pathways, among others. Interestingly, expression of a select set of genes during dormancy provides unique insight into cell cycle regulation and may open up new strategies for combination therapy. More investigation into the nature and mechanisms of entry and exit from this arrested state is merited.

Although morphological changes and transcriptome arrest were similar for DHA-sensitive D6 and resistant D6.QHS2400x5 clones, we observed that D6.QHS2400x5 released from transcriptome arrest slightly earlier (~24 hrs) than D6. Although this observation has not yet been repeated and thus must be interpreted with caution, it does invite speculation about the nature of the dormant state as it relates to artemisinin resistance. Should this observation hold for subsequent experiments, an obvious explanation would be that some fundamental difference in the dormant state between sensitive and resistant parasites allows resistant parasites to recover from artemisinin treatment faster. Perhaps resistant parasites maintain a heartier metabolism during dormancy or enter dormancy at an increased rate. Regardless of the mechanisms

involved, our data suggest dormancy and resistance may be linked, yet likely separate phenotypic responses to artemisinin.

In addition to the description of the dormant transcriptome, we describe novel constitutive and DHA-induced expression differences in artemisinin resistant D6.QHS2400x5 in comparison with parental D6. Although many of these genes appear to be strain-specific exported members of large gene families, others may prove interesting given independent validation. Notably missing among our baseline transcriptional differences are previously identified candidate genes postulated to be associated with resistance (e.g., SERCA ATPase), suggesting that if changes in these genes are indeed causative for resistance, they do not exert their effect at the transcript abundance level.

In the broader context, our results describe a time course of induction of dormancy that may prove very clinically relevant for better dosing schedules of ACTs and may also provide kinetic data for epidemiological predictions about the acquisition and geographic spread of resistance. In addition, the differentially expressed genes in dormant parasites and in resistant parasites may serve as novel molecular markers for monitoring drug efficacy and emergence of resistance in the field. Given the recent emergence of clinically relevant artemisinin resistance, our study provides new avenues to understanding the parasite's response to artemisinin and its development of artemisinin resistance, currently the most pressing public health problem for malaria control and elimination efforts.

Table 1 – Genes up- or down-regulated in both D6 and D6.QHS2400x5 after treatment with DHA as compared to T=0. Oligo ID refers to the identifier of the oligo from the microarrays used to determine differential regulation. Score(d) refers to delta score from SAM analysis. Fold change refers to the geometric mean fold difference between T=0 and later time points. Local FDR refers to the false discovery rate for data with the corresponding delta score.

Genes upregulated during transcriptome arrest

| oligo ID | PlasmoDB ID | Description | Score(d) | Fold Change | local fdr(%) |
|----------------|------------------|---|----------|-------------|--------------|
| N145_33 | PF14 0014 | Plasmodium exported protein unknown function | 2.57 | 4.52 | 0.00 |
| N145_22 | PF14 0017 | lysophospholipase putative | 2.55 | 8.44 | 0.00 |
| oPFI17632 | PFI1170c | thioredoxin reductase | 2.23 | 10.63 | 0.19 |
| oPF08_0001_261 | PF08 0001 | Plasmodium exported protein unknown function | 2.22 | 3.87 | 0.20 |
| D56470_2 | PFI1520w | asparagine-rich antigen putative | 2.20 | 8.43 | 0.23 |
| F62396_2 | MAL7P1.144 | Serine/Threonine protein kinase FIKK family | 2.16 | 3.59 | 0.28 |
| N143_54 | PF14 0183 | signal recognition particle RNP putative | 2.13 | 7.01 | 0.31 |
| oPFN0249 | PF14 0010 | glycophorin binding protein family Gbph | 2.12 | 8.82 | 0.32 |
| B587 | PFB0930w | Plasmodium exported protein (hyp9) unknown function | 2.00 | 3.60 | 0.55 |
| oMAL6P1.106_83 | PFF0510w | histone H3 | 1.99 | 7.46 | 0.57 |
| oPFA0395c_496 | PFA0395w | conserved Plasmodium protein unknown function | 1.93 | 3.02 | 0.69 |
| J33_27 | PF10 0013 | Plasmodium exported protein (hyp12) unknown function | 1.92 | 3.59 | 0.71 |
| L1_39 | PFL0055c | RESA-like protein with PHIST and DnaJ domains | 1.90 | 4.61 | 0.76 |
| J33_15 | PF10 0020 | alpha/beta hydrolase putative | 1.87 | 3.08 | 0.83 |
| oPFG0019 | MAL7P1.58 | Pfmc-2TM Maurer's cleft two transmembrane protein | 1.84 | 3.44 | 0.91 |
| N143_57 | PF14 0180 | conserved Plasmodium protein unknown function | 1.78 | 5.81 | 1.04 |
| A26463_4 | PFA0130c | Serine/Threonine protein kinase FIKK family putative | 1.74 | 3.34 | 1.18 |
| oPFL0108 | PFL2175w | ubiquitin conjugating enzyme E2 putative | 1.71 | 3.32 | 1.27 |
| oPFRNA0004 | 28S rRNA Chr7 | | 1.70 | 5.65 | 1.29 |
| oPFC0360w_45 | PFC0360w | Activator of Hsp90 ATPase homolog 1-like protein putative | 1.69 | 6.92 | 1.31 |
| M20186_2 | MAL13P1.470 | Plasmodium exported protein (PHISTa) unknown function | 1.67 | 3.25 | 1.38 |

Genes downregulated during transcriptome arrest

| oligo ID | PlasmoDB ID | Description | Score(d) | Fold Change | local fdr(%) |
|------------------|-------------|--|----------|-------------|--------------|
| oMAL6P1.94_744 | PFF0450c | Zn ²⁺ or Fe ²⁺ permease | -2.21 | -4.81 | 0.15 |
| Ks371_8 | PF11 0291 | conserved Plasmodium protein unknown function | -2.15 | -4.84 | 0.10 |
| M45763_1 | PF13 0198 | reticulocyte binding protein 2 homolog A | -2.12 | -4.64 | 0.08 |
| Ks424_2 | PF11 0252 | neutral-sphingomyelinase activation factor protein putative | -2.12 | -4.10 | 0.07 |
| J326_1 | PF10 0242 | conserved Plasmodium protein unknown function | -2.07 | -4.21 | 0.03 |
| M45727_12 | PF13 0035 | U3 small nucleolar RNA-associated protein 6 putative | -2.05 | -4.03 | 0.01 |
| Ks26_15 | PF11 0115 | conserved Plasmodium protein unknown function | -2.04 | -4.07 | 0.00 |
| J33_11 | PF10 0022 | Plasmodium exported protein (PHISTc) unknown function | -2.04 | -3.64 | 0.00 |
| Ks17_16 | PF11 0526 | conserved Plasmodium protein unknown function | -2.03 | -3.75 | 0.00 |
| oPF08_0035_2217 | PF08 0035 | conserved Plasmodium protein unknown function | -2.01 | -5.70 | 0.00 |
| J33_12 | PF10 0021 | Plasmodium exported protein (PHISTc) unknown function | -2.00 | -4.12 | 0.00 |
| E29792_2 | PFE1270c | WD domain G-beta repeat-containing protein | -1.95 | -3.78 | 0.00 |
| oPF08_0086_1650 | PF08 0086 | RNA binding protein putative | -1.93 | -4.09 | 0.00 |
| oPFD67014 | PFD1145c | reticulocyte-binding protein homologue 5 | -1.91 | -5.02 | 0.00 |
| oMAL6P1.105_1475 | PFF0505c | conserved Plasmodium protein unknown function | -1.90 | -4.47 | 0.00 |
| oPFD67000 | PFD1060w | u5 small nuclear ribonucleoprotein- specific protein putative | -1.90 | -3.79 | 0.00 |
| I14335_1 | PF13 0091 | conserved Plasmodium protein unknown function | -1.90 | -4.05 | 0.00 |
| D49942_9 | PFD0110w | reticulocyte-binding protein homologue 1 | -1.89 | -4.05 | 0.00 |
| oPFI17735 | PFI0820c | RNA binding protein putative | -1.87 | -4.16 | 0.00 |
| oPF14_0753_609 | PF14 0753 | Plasmodium exported protein (hyp13) unknown function | -1.86 | -3.74 | 0.00 |
| oPFF72450 | PFF1295w | conserved Plasmodium protein unknown function | -1.84 | -3.32 | 0.00 |
| L1_42 | PFL0060w | Plasmodium exported protein unknown function | -1.84 | -3.16 | 0.00 |
| oPF08_0127_2717 | PF08 0127 | conserved Plasmodium protein unknown function | -1.84 | -4.47 | 0.00 |
| J109_4 | PF10 0180 | conserved Plasmodium protein unknown function | -1.83 | -3.24 | 0.00 |
| Ks51_9 | PF11 0218 | conserved Plasmodium protein unknown function | -1.80 | -4.10 | 0.00 |
| F64738_5 | PF07 0081 | conserved Plasmodium protein unknown function | -1.80 | -3.00 | 0.00 |
| oPFI1725w_124 | PFI1725w | Plasmodium exported protein unknown function | -1.76 | -3.50 | 0.00 |
| oMAL7P1.22_3735 | MAL7P1.22 | conserved Plasmodium protein unknown function | -1.76 | -3.01 | 0.00 |
| F53854_2 | MAL7P1.19 | ubiquitin transferase putative | -1.75 | -3.84 | 0.00 |
| J132_12 | PF10 0327 | Myb2 protein | -1.75 | -3.31 | 0.00 |
| oPF10_0179_3036 | PF10 0179 | conserved Plasmodium protein unknown function | -1.74 | -3.17 | 0.00 |
| M48367_1 | PF13 0088 | Myb1 protein | -1.74 | -3.20 | 0.00 |
| M42966_1 | PF13 0310 | periribosomal processosome UTP putative | -1.73 | -3.12 | 0.00 |
| M8199_3 | PF13 0278 | ran-binding protein putative | -1.72 | -2.90 | 0.00 |

| | | | | | |
|-------------------|-------------|--|-------|-------|------|
| M33579_4 | MAL13P1.268 | conserved Plasmodium protein unknown function | -1.70 | -3.07 | 0.00 |
| Ks370_11 | PF11 0267 | kelch protein putative | -1.69 | -3.11 | 0.00 |
| Kn3709_1 | PFI1480w | conserved Plasmodium protein unknown function | -1.69 | -3.41 | 0.00 |
| oMAL13P1.155_3254 | MAL13P1.155 | conserved Plasmodium protein unknown function | -1.67 | -2.86 | 0.00 |
| F49582_1 | PFE1400c | beta adaptin protein putative | -1.67 | -3.15 | 0.00 |
| N141_27 | PF14 0224 | serine/threonine protein phosphatase | -1.66 | -3.16 | 0.00 |
| E10660_1 | PF10 0232 | Chromodomain-helicase-DNA- binding protein 1 homolog putative | -1.66 | -3.46 | 0.00 |
| D49942_2 | PFD0095c | Plasmodium exported protein (PHISTb) unknown function | -1.66 | -2.94 | 0.00 |
| D16785_1 | PFD1115c | conserved Plasmodium protein unknown function | -1.66 | -2.98 | 0.00 |
| D56950_2 | PFF0880c | conserved Plasmodium protein unknown function | -1.66 | -2.92 | 0.00 |
| oPFBLOB0101 | PF10 0244 | formin 2 putative | -1.65 | -2.83 | 0.00 |
| oMAL7P1.65_376 | MAL7P1.65 | conserved Plasmodium protein unknown function | -1.64 | -3.33 | 0.00 |
| N143_10 | PF14 0201 | surface protein Pf113 | -1.64 | -3.25 | 0.00 |
| J2975_2 | MAL13P1.179 | conserved Plasmodium protein unknown function | -1.62 | -2.80 | 0.00 |
| oPFI17666 | PFI0685w | pseudouridylate synthase putative | -1.61 | -2.90 | 0.00 |
| F28964_1 | PF13 0161 | conserved Plasmodium protein unknown function | -1.61 | -3.43 | 0.00 |
| Ks539_1 | PF11 0433 | conserved Plasmodium protein unknown function | -1.60 | -2.72 | 0.00 |
| M2610_1 | PF13 0058 | RNA binding protein putative | -1.60 | -2.85 | 0.00 |
| Ks8_1 | PF11 0201 | ubiquitin-protein ligase putative | -1.59 | -2.83 | 0.00 |
| oPFM60468 | MAL13P1.304 | malaria antigen | -1.59 | -2.65 | 0.00 |
| oPFBLOB0118 | PFE0505w | cyclophilin putative | -1.59 | -3.52 | 0.00 |
| J106_6 | PF10 0215a | conserved Plasmodium protein unknown function | -1.59 | -3.78 | 0.00 |
| I2966_1 | PFI0325c | conserved Plasmodium protein unknown function | -1.58 | -2.94 | 0.00 |
| Kn862_1 | PFL1410c | ABC transporter (CT family) | -1.58 | -2.94 | 0.00 |
| F20625_1 | PFD0740w | cyclin-dependent kinase putative | -1.57 | -2.98 | 0.00 |
| D33675_1 | PFL0130c | conserved Plasmodium protein unknown function | -1.57 | -3.13 | 0.00 |
| oPFL0150 | PFL2295w | nucleolar rRNA processing protein putative | -1.57 | -3.26 | 0.00 |
| L2_104 | PFL0405w | conserved Plasmodium protein unknown function | -1.57 | -3.03 | 0.00 |
| J1125_1 | PF11 0111 | asparagine-rich antigen | -1.57 | -3.74 | 0.00 |
| M32529_1 | PF13 0215 | conserved Plasmodium protein unknown function | -1.56 | -3.39 | 0.00 |
| C277 | PFC0430w | conserved Plasmodium protein unknown function | -1.55 | -3.06 | 0.00 |
| E5006_5 | PFE0465c | RNA polymerase I | -1.55 | -2.92 | 0.00 |
| M789_3 | MAL13P1.302 | SUMO ligase putative | -1.55 | -2.75 | 0.00 |
| L2_81 | PFL0380c | tRNA delta(2)- isopentenylpyrophosphate transferase putative | -1.55 | -3.00 | 0.00 |
| M51323_1 | PFL2275c | FK506-binding protein (FKBP)-type peptidyl-propyl isomerase | -1.54 | -3.09 | 0.00 |
| oMAL7P1.141_100 | MAL7P1.141 | conserved Plasmodium protein unknown function | -1.54 | -3.33 | 0.00 |
| Ks72_6 | PF11 0398 | conserved Plasmodium protein unknown function | -1.54 | -3.41 | 0.00 |
| oMAL13P1.34_255 | MAL13P1.34 | RED-like protein putative | -1.54 | -2.60 | 0.00 |
| oPFJ12810 | PF10 0136 | initiation factor 2 subunit family putative | -1.54 | -2.98 | 0.00 |
| M33088_2 | PF13 0233 | myosin A | -1.53 | -3.31 | 0.00 |

| | | | | | |
|----------------|-------------|---|-------|-------|------|
| oPFF72491 | PFF1000w | cleavage stimulation factor subunit 1-like protein putative | -1.52 | -3.23 | 0.00 |
| oPFL0117 | PFL2235w | conserved Plasmodium protein unknown function | -1.51 | -3.03 | 0.00 |
| oPFL2100w_829 | PFL2100w | ubiquitin conjugating enzyme E2 putative | -1.51 | -3.41 | 0.00 |
| M18079_5 | MAL13P1.120 | splicing factor putative | -1.51 | -2.65 | 0.00 |
| A14801_24 | PFA0635c | Plasmodium exported protein (hyp1) unknown function | -1.51 | -2.92 | 0.00 |
| Kn12363_3 | PF08 0137 | Plasmodium exported protein (PHISTc) unknown function | -1.50 | -2.48 | 0.00 |
| E25193_1 | PFF1075w | conserved Plasmodium protein unknown function | -1.50 | -3.03 | 0.00 |
| oPFD67004 | PFD0835c | LETM1-like protein putative | -1.49 | -2.58 | 0.00 |
| Kn434_1 | PFL1505c | conserved Plasmodium protein unknown function | -1.49 | -2.85 | 0.00 |
| oPFD66987 | PFD0960c | 60S ribosomal protein L7Ae/L30e putative | -1.48 | -3.09 | 0.00 |
| D17715_55 | PFD0460c | conserved Plasmodium protein unknown function | -1.48 | -3.01 | 0.00 |
| Kn707_1 | PFL2475w | DEAD/DEAH box helicase putative | -1.45 | -2.62 | 0.00 |
| L1_101 | PFL0175c | conserved Plasmodium protein unknown function | -1.45 | -2.51 | 0.00 |
| N150_50 | PF14 0102 | rhoptry-associated protein 1 RAP1 | -1.45 | -4.21 | 0.00 |
| oPF13_0276_110 | PF13 0276 | membrane-associated histidine rich protein 2 (MARHP2) | -1.45 | -3.71 | 0.00 |
| oPFI17683 | PFI0480w | helicase with Zn-finger motif putative | -1.44 | -2.89 | 0.00 |
| N149_12 | PF14 0487 | conserved Plasmodium protein unknown function | -1.44 | -2.96 | 0.00 |
| M1222_1 | MAL13P1.323 | conserved Plasmodium protein unknown function | -1.43 | -2.91 | 0.00 |
| Ks157_11 | PF11 0509 | ring-infected erythrocyte surface antigen putative | -1.43 | -3.62 | 0.00 |
| oPFBLOB0188 | PFI0495w | conserved Plasmodium protein unknown function | -1.43 | -2.62 | 0.00 |
| oPFBLOB0090 | PF10 0150 | methionine aminopeptidase putative | -1.43 | -2.71 | 0.00 |
| E11202_1 | PFE0570w | RNA pseudouridylation synthase putative | -1.42 | -2.71 | 0.00 |
| F71039_5 | MAL7P1.150 | cysteine desulfurase putative | -1.42 | -2.61 | 0.00 |
| L2_30 | PFL0275w | conserved Plasmodium protein unknown function | -1.41 | -2.53 | 0.00 |
| M53930_1 | PF13 0254 | conserved Plasmodium membrane protein unknown function | -1.40 | -2.54 | 0.00 |
| F45048_1 | MAL8P1.150 | conserved Plasmodium protein unknown function | -1.40 | -2.65 | 0.00 |
| J43_19 | PF10 0054 | conserved protein unknown function | -1.40 | -2.71 | 0.00 |
| F4688_1 | PFI1180w | patatin-like phospholipase putative | -1.40 | -2.52 | 0.00 |
| Ks13_2 | PF11 0402 | conserved Plasmodium protein unknown function | -1.40 | -2.64 | 0.00 |
| F42703_3 | PF08 0121 | peptidyl-prolyl cis-trans isomerase precursor | -1.40 | -2.44 | 0.00 |
| C244 | PFC0380w | protein phosphatase | -1.40 | -2.56 | 0.00 |
| L3_1 | PFL0815w | DNA-binding chaperone putative | -1.40 | -2.59 | 0.00 |
| N138_48 | PF14 0296 | 60S ribosomal protein L14 putative | -1.39 | -3.16 | 0.00 |
| B343 | PFB0490c | conserved Plasmodium protein unknown function | -1.39 | -2.67 | 0.00 |
| M7985_1 | MAL13P1.26 | conserved Plasmodium protein unknown function | -1.38 | -2.77 | 0.00 |
| F55492_1 | MAL7P1.171 | Plasmodium exported protein unknown function | -1.38 | -2.82 | 0.00 |
| oPFI17671 | PFI1470c | conserved Plasmodium protein unknown function | -1.38 | -3.37 | 0.00 |
| Kn945_1 | MAL13P1.171 | transmembrane protein Tmp21 homologue putative | -1.38 | -2.35 | 0.00 |

| | | | | | |
|-----------------|------------|---|-------|-------|------|
| J245_3 | PF10 0233 | conserved Plasmodium protein unknown function | -1.38 | -2.60 | 0.00 |
| M25032_3 | MAL13P1.45 | U4/U6 small nuclear ribonucleoprotein putative | -1.38 | -2.74 | 0.00 |
| L2_40 | PFL0290w | conserved Plasmodium protein unknown function | -1.37 | -2.50 | 0.00 |
| N134_45 | PF14 0614 | phosphatase putative | -1.37 | -2.42 | 0.00 |
| oPF08_0040_1207 | PF08 0040 | clp1-related protein putative | -1.37 | -2.47 | 0.00 |
| oPFL0135 | PFL1235c | conserved Plasmodium protein unknown function | -1.37 | -3.15 | 0.00 |
| oMAL6P1.162_5 | PFF1290c | conserved Plasmodium protein unknown function | -1.36 | -2.45 | 0.00 |

Table 2 – Genes up- or down-regulated in D6.QHS2400x5 relative to D6 during IDC (no DHA treatment). Oligo ID refers to the identifier of the oligo from the microarrays used to determine differential regulation. Score(d) refers to delta score from SAM analysis. Fold change refers to the geometric mean fold difference in expression between D6.QHS2400x5 and D6. Local FDR refers to the false discovery rate for data with the corresponding delta score.

| Gene ID | PlasmoDB_ID | Description | Score(d) | Fold change | local fdr(%) |
|---------------|-------------|---|----------|-------------|--------------|
| A11546_1 | PFB0100c | knob-associated histidine-rich protein | 7.17 | 41.98 | 0.00 |
| B52 | PFB0090c | RESA-like protein with PHIST and DnaJ domains | 6.70 | 22.08 | 0.00 |
| oPFE1600w_785 | PFE1600w | Plasmodium exported protein (PHISTb) unknown function | 6.01 | 51.80 | 0.00 |
| oPFK12891 | PF11_0512 | RESA-like protein with PHIST and DnaJ domains | 5.93 | 6.29 | 0.00 |
| B47 | PFB0080c | Plasmodium exported protein (PHISTb) unknown function | 5.14 | 9.35 | 0.02 |
| B50 | PFB0085c | DNAJ protein putative | 4.93 | 27.18 | 0.03 |
| B603 | PFB0953w | Plasmodium exported protein (hyp15) unknown function | 4.62 | 4.84 | 0.05 |
| F13845_1 | PFE1605w | Plasmodium exported protein (PHISTb) unknown function | 4.31 | 20.01 | 0.06 |
| D10455_2 | PFD1185w | Plasmodium exported protein (PHISTa) unknown function | 3.85 | 3.20 | 0.00 |
| B45 | PFB0075c | Plasmodium exported protein (hyp9) unknown function | 3.62 | 4.12 | 0.00 |
| M32813_2 | MAL13P1.184 | endopeptidase putative | 3.23 | 2.83 | 0.00 |
| Ks115_1 | PF11_0359 | coatomer delta subunit putative | 3.19 | 2.74 | 0.00 |
| N133_46 | PF14_0686 | conserved Plasmodium protein unknown function | 3.11 | 2.15 | 0.00 |
| J63_1 | PF10_0283a | | 2.85 | 2.60 | 0.00 |
| N129_1 | PF14_0745 | probable protein unknown function | 2.64 | 3.14 | 0.00 |
| D57574_1 | PFE1615c | Plasmodium exported protein unknown function | 2.63 | 10.48 | 0.00 |
| F17165_1 | MAL8P1.104 | CAF1 family ribonuclease putative | 2.58 | 2.31 | 0.00 |
| J125_3 | PF10_0294 | RNA helicase putative | 2.53 | 1.81 | 0.00 |
| L2_280 | PFL0795c | male development gene 1 | 2.53 | 3.29 | 0.00 |
| D49942_2 | PFD0095c | Plasmodium exported protein (PHISTb) unknown function | 2.51 | 2.57 | 0.00 |

| | | | | | |
|-----------------|------------|--|------|------|------|
| J383_1 | PF10_0291 | RAP protein putative | 2.50 | 2.62 | 0.00 |
| oPFD1200c_505 | PFD1200c | Plasmodium exported protein (hyp6) unknown function | 2.48 | 2.95 | 0.00 |
| J63_5 | PF10_0282 | conserved Plasmodium protein unknown function | 2.47 | 3.06 | 0.00 |
| E17521_1 | MAL7P1.171 | Plasmodium exported protein unknown function | 2.45 | 2.83 | 0.00 |
| Ks26_16 | PF11_0114a | conserved Plasmodium protein unknown function | 2.44 | 2.47 | 0.00 |
| J232_8 | PF10_0258 | conserved Plasmodium protein unknown function | 2.42 | 2.76 | 0.00 |
| oMAL6P1.106_83 | PFF0510w | histone H3 | 2.41 | 2.47 | 0.00 |
| J564_2 | PF10_0296 | conserved Plasmodium protein unknown function | 2.34 | 2.34 | 0.00 |
| E19231_1 | PFE0570w | RNA pseudouridylylate synthase putative | 2.28 | 2.18 | 0.00 |
| Kn8928_4 | PFL2415w | Hbeta58/Vps26 protein homolog putative | 2.27 | 2.02 | 0.00 |
| J116_13 | PF10_0344 | glutamate-rich protein | 2.27 | 2.58 | 0.00 |
| Km765_1 | PFL1540c | phenylalanyl-tRNA synthetase alpha chain putative | 2.26 | 1.64 | 0.00 |
| oMAL6P1.280_514 | PFF0705c | conserved Plasmodium protein unknown function | 2.19 | 2.49 | 0.00 |
| oMAL6P1.209_513 | PFF1055c | conserved Plasmodium protein unknown function | 2.18 | 2.04 | 0.00 |
| J62_1 | PF10_0281 | merozoite TRAP-like protein MTRAP | 2.16 | 2.27 | 0.00 |
| oPFI17673 | PFI1110w | glutamine synthetase putative | 2.15 | 1.78 | 0.00 |
| F44837_1 | No_ORFs | 500bp from the 5' end of PFE0680w | 2.13 | 1.94 | 0.00 |
| oPFL1750c_2362 | PFL1750c | conserved Plasmodium protein unknown function | 2.10 | 2.69 | 0.01 |
| Ks76_9 | PF11_0423 | conserved Plasmodium protein unknown function | 2.08 | 1.92 | 0.18 |
| J64_2 | PF10_0298 | 26S proteasome subunit putative | 2.08 | 1.68 | 0.26 |
| D53895_3 | PFE0730c | ribose 5-phosphate epimerase, putative | 2.07 | 1.92 | 0.31 |
| C240 | PFC0370w | conserved Plasmodium protein unknown function | 2.06 | 2.04 | 0.44 |
| D33539_1 | No_ORFs | 700bp from the 5' end of PFD0695w | 2.06 | 1.68 | 0.45 |
| E19346_1 | PF10_0188 | conserved Plasmodium membrane protein unknown function | 2.06 | 2.03 | 0.46 |
| F16755_1 | PFF0930w | conserved Plasmodium protein unknown function | 2.02 | 1.91 | 0.88 |
| Ks97_2 | PF11_0192 | histone acetyltransferase putative | 2.01 | 1.68 | 0.94 |
| oPFB0161c_228 | PFB0161c | conserved Plasmodium protein unknown function | 2.01 | 2.24 | 0.99 |
| J151_8 | PF10_0287 | conserved Plasmodium protein unknown function | 2.00 | 1.98 | 1.02 |

| | | | | | |
|----------------|------------------------|--|------|------|------|
| C442 | PFC0675c | mitochondrial ribosomal protein L29/L47 precursor putative | 1.99 | 2.14 | 1.12 |
| F42768_1 | PF10_0232 | Chromodomain-helicase-DNA-binding protein 1 homolog putative | 1.98 | 2.16 | 1.29 |
| F32316_2 | PFE0120c | Merozoite Surface Protein 8 MSP8 | 1.97 | 2.16 | 1.38 |
| M35930_9 | MAL13P1.228 | conserved Plasmodium protein unknown function | 1.97 | 2.15 | 1.41 |
| N134_113 | PF14_0586 | conserved Plasmodium protein unknown function | 1.97 | 2.02 | 1.47 |
| oMAL8P1.27_224 | MAL8P1.27 | translation initiation factor IF-3 putative | 1.96 | 2.25 | 1.49 |
| L2_212 | PFL0660w | dynein light chain 1 putative | 1.96 | 1.84 | 1.54 |
| N133_39 | PF14_0691 | conserved Plasmodium membrane protein unknown function | 1.95 | 1.69 | 1.66 |
| F14111_3 | PFF1100c | transcription factor with AP2 domain(s) putative | 1.95 | 1.74 | 1.73 |
| J73_4 | PF10_0084 | tubulin beta chain putative | 1.94 | 2.15 | 1.80 |
| oPFN0248 | PF14_0102 | rhoptry-associated protein 1 RAP1 | 1.94 | 2.18 | 1.82 |
| F71176_2 | PFE0405c | Longevity-assurance (LAG1) domain protein putative | 1.94 | 1.99 | 1.87 |
| Ks89_8 | PF11_0356 | conserved Plasmodium protein unknown function | 1.92 | 1.79 | 2.08 |
| oPF117676 | PFI0230c | bacterial histone-like protein | 1.89 | 1.84 | 2.53 |
| oPFL1785c_762 | PFL1785c | conserved Plasmodium protein unknown function | 1.88 | 1.99 | 2.71 |
| M3696_2 | MAL13P1.333 | conserved Plasmodium protein unknown function | 1.88 | 1.70 | 2.72 |
| F23846_3 | PF08_0034 | histone acetyltransferase GCN5 putative | 1.87 | 1.53 | 2.81 |
| M35431_1 | PF14_0631 | conserved Plasmodium protein unknown function | 1.87 | 2.23 | 2.85 |
| J106_11 | PF10_0214 | RNA binding protein putative | 1.86 | 1.77 | 3.01 |
| Ks488_10 | Unannotated transcript | 700bp from the 3' end of PF11_0298 | 1.86 | 1.92 | 3.02 |
| N137_29 | PF14_0667 | conserved Plasmodium protein unknown function | 1.86 | 1.99 | 3.03 |
| F67443_1 | PF07_0101 | conserved Plasmodium protein unknown function | 1.86 | 1.91 | 3.06 |
| D27953_2 | PFD1180w | Plasmodium exported protein (PHISTb) unknown function | 1.84 | 1.76 | 3.35 |
| M46928_1 | PF13_0190 | conserved Plasmodium protein unknown function | 1.84 | 1.58 | 3.49 |
| Ks222_1 | MAL13P1.420 | hypothetical protein | 1.83 | 1.83 | 3.53 |
| M951_1 | PF13_0222 | phosphatase putative | 1.83 | 1.76 | 3.63 |
| J461_7 | PF10_0300 | RNA methyltransferase putative | 1.81 | 1.87 | 3.91 |

| Gene ID | PlasmoDB_ID | Description | Score(d) | Fold change | local fdr(%) |
|----------|-------------|---|----------|-------------|--------------|
| N145_38 | No_ORFs | Chr 14 1kb from 3' end of PF14 0013 | -2.86 | -2.92 | 0.00 |
| E6820_1 | PFE0250w | conserved Plasmodium protein unknown function | -2.68 | -2.34 | 0.00 |
| L1_32 | PFL0045c | Plasmodium exported protein (PHISTc) unknown function | -2.67 | -2.42 | 0.00 |
| N171_3 | PF14_0073 | conserved Plasmodium protein unknown function | -2.54 | -1.80 | 0.00 |
| E2283_4 | PFE0130c | Plasmodium protein unknown function | -2.53 | -2.29 | 0.00 |
| F13309_2 | PFE1270c | WD domain G-beta repeat-containing protein | -2.42 | -1.99 | 0.00 |
| F21560_1 | MAL8P1.88 | conserved Plasmodium protein unknown function | -2.41 | -2.20 | 0.00 |
| B306 | PFB0435c | transporter putative | -2.41 | -2.09 | 0.00 |
| F18577_5 | MAL8P1.134 | ferlin like protein putative | -2.34 | -2.14 | 0.00 |
| N187_1 | PF14_0678 | exported protein 2 | -2.30 | -2.28 | 0.00 |
| F38025_1 | PFF1260c | conserved Plasmodium protein unknown function | -2.15 | -1.93 | 0.00 |

Table 3 – Genes up- or down-regulated in D6.QHS2400x5 relative to D6 after treatment with DHA as compared to T=0. Oligo ID refers to the identifier of the oligo from the microarrays used to determine differential regulation. Score(d) refers to delta score from SAM analysis. Fold change refers to the geometric mean fold difference in DHA-induced expression between D6.QHS2400x5 and D6. Local FDR refers to the false discovery rate for data with the corresponding delta score.

| oligo ID | PlasmoDB_ID | Description | Score(d) | Fold Change | local fdr(%) |
|-----------------|------------------------------|---|-----------------|--------------------|---------------------|
| oPFE1020w_182 | PFE1020w | U6 snRNA-associated sm-like protein Lsm2 putative | 5.18 | 2.52 | 0.02 |
| N164_5 | PF14_0229 | conserved Plasmodium protein unknown function | 4.71 | 2.99 | 0.04 |
| F27786_1 | PF07_0086 | conserved Plasmodium membrane protein unknown function | 4.57 | 2.91 | 0.04 |
| C677 | PFC1016w | conserved Plasmodium protein unknown function | 4.52 | 16.59 | 0.04 |
| I11857_2 | PFI0865w | XPA binding protein 1 putative | 4.45 | 2.82 | 0.04 |
| L2_104 | PFL0405w | conserved Plasmodium protein unknown function | 4.28 | 2.38 | 0.04 |
| F7915_1 | PFF0305c | ubiquitin conjugating enzyme E2 putative | 4.14 | 2.33 | 0.03 |
| oPF11_0152_17 | PF11_0152 | GTPase activator putative | 4.08 | 2.16 | 0.02 |
| oPFL0023 | PFL1745c | clustered-asparagine-rich protein | 3.91 | 8.19 | 0.00 |
| N135_22 | Gene absent in PlasmoDB v6.3 | transcript between PF14_0633 and PF14_0634 | 3.91 | 2.43 | 0.00 |
| M38913_1 | PF13_0352 | conserved Plasmodium protein unknown function | 3.82 | 2.17 | 0.00 |
| oPFM60522 | MAL13P1.47 | mitochondrial ATP synthase delta subunit putative | 3.76 | 1.79 | 0.00 |
| oligo ID | PlasmoDB_ID | Description | Score(d) | Fold Change | local fdr(%) |
| Ks259_3 | PF11_0245 | translation elongation factor EF-1 subunit alpha putative | -5.64 | -4.77 | 0.17 |
| N155_25 | PF14_0031a | conserved Plasmodium protein unknown function | -5.29 | -4.09 | 0.16 |
| F27351_1 | PFI0690c | conserved Plasmodium protein unknown function | -5.20 | -6.67 | 0.15 |
| N145_12 | PF14_0020 | choline kinase | -5.17 | -3.75 | 0.15 |
| Ks26_12 | PF11_0116 | conserved Plasmodium protein unknown function | -4.96 | -3.29 | 0.15 |

| | | | | | |
|------------|------------|---|-------|--------|------|
| N143_57 | PF14_0180 | conserved Plasmodium protein unknown function | -4.93 | -10.04 | 0.14 |
| B50 | PFB0085c | DNAJ protein putative | -4.91 | -23.52 | 0.14 |
| I4719_6 | PFI0175w-a | conserved Plasmodium protein unknown function | -4.63 | -19.39 | 0.15 |
| Ks1072_1 | PF11_0168 | moving junction protein | -4.58 | -3.55 | 0.15 |
| oPFRNA0002 | MAL14_5S_1 | 5S rRNA | -4.42 | -33.03 | 0.15 |
| oPFN0262 | PF14_0589 | valine-tRNA ligase putative | -4.26 | -7.35 | 0.15 |
| A8408_2 | PF07_0101 | conserved Plasmodium protein unknown function | -4.14 | -3.36 | 0.15 |
| J269_10 | PF10_0231 | conserved Plasmodium protein unknown function | -4.00 | -2.63 | 0.15 |
| Ks370_2 | PF11_0268 | kelch motif containing protein putative | -3.95 | -4.63 | 0.15 |
| D49176_7 | PFD0230c | protease putative | -3.91 | -4.81 | 0.15 |
| F53897_2 | MAL7P1.119 | conserved Plasmodium protein unknown function | -3.78 | -3.76 | 0.14 |
| Ks510_10 | PF11_0381 | subtilisin-like protease 2 | -3.77 | -3.47 | 0.13 |
| F23699_1 | PFI0265c | RhopH3 | -3.74 | -3.29 | 0.13 |
| B60 | PFB0105c | Plasmodium exported protein (PHISTc) unknown function | -3.72 | -3.31 | 0.13 |
| N141_27 | PF14_0224 | | -3.71 | -4.96 | 0.13 |
| oPFF72487 | PFF0645c | serine/threonine protein phosphatase | -3.64 | -8.84 | 0.11 |
| F30848_1 | MAL8P1.101 | integral membrane protein putative | -3.61 | -3.64 | 0.11 |
| J4379_1 | PFL1330c | RNA binding protein putative | -3.59 | -3.18 | 0.10 |
| Kn5186_3 | PFL1180w | cyclin-related protein Pfcyc-2 | -3.58 | -2.92 | 0.10 |
| B541 | PFB0845w | chromatin assembly protein (ASF1) putative | -3.52 | -2.50 | 0.09 |
| oPFL0022 | PFL2460w | conserved Plasmodium membrane protein unknown function | -3.50 | -3.69 | 0.09 |
| A31914_4 | PF10_0258 | coronin | -3.48 | -4.41 | 0.08 |
| A8109_9 | PFA0295c | conserved Plasmodium protein unknown function | -3.45 | -4.65 | 0.07 |
| Ks54_4 | PF11_0036 | flavoprotein putative | -3.44 | -2.16 | 0.07 |
| C237 | PFC0355c | conserved Plasmodium protein unknown function | -3.43 | -5.23 | 0.07 |

Figure 1 – Pilot experiment of dihydroartemisinin treatment of synchronized wildtype W2 parasites. The yellow line plots the Pearson correlations between T=6 microarray results and each hour of normal HB3 IDC, while the blue line plots correlations between T=27 microarray results and the HB3 IDC. Peak correlations for both W2 time points occur at 12-13 hours post-invasion within the HB3 IDC.

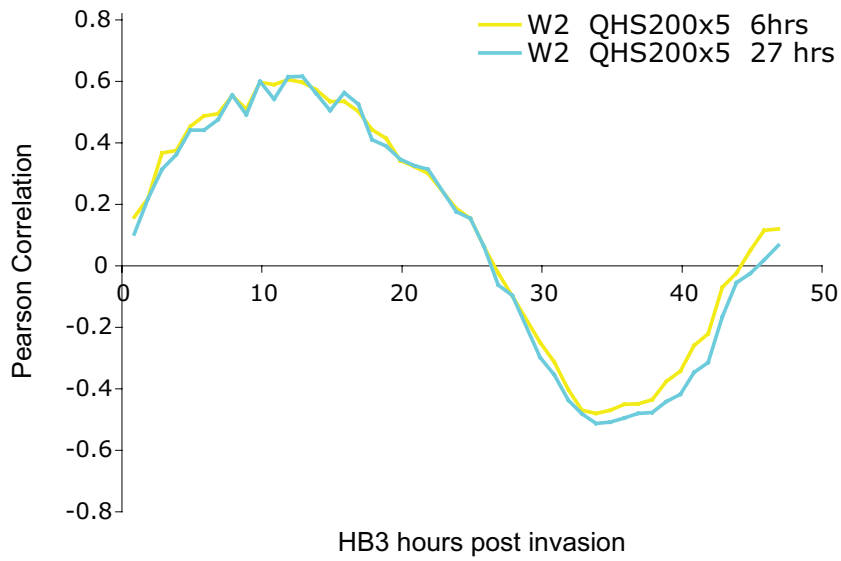


Figure 2 - Dihydroartemisinin (DHA) treatment microarray experimental design.

Clones of the parental strain D6 and the resistant strain D6.QHS2400x5 were expanded into both highly synchronous ring stage cultures as well as asynchronous mixed stage cultures. When the synchronous cultures were approximately 8 hours post-invasion, 200 ng/mL DHA was added to half of each culture (purple arrow). 6 hours later, drug was washed out (orange arrows). Untreated controls were treated with DMSO and it was also washed out 6 hours later (grey arrows). The tick marks for each culture indicate samples hybridized against pool RNA on the microarray. Time course is only shown out to 80 hours after drug addition, but additional samples were taken every 12 hours after that for both treated synchronous cultures until they reached ~3% parasitemia by smear.

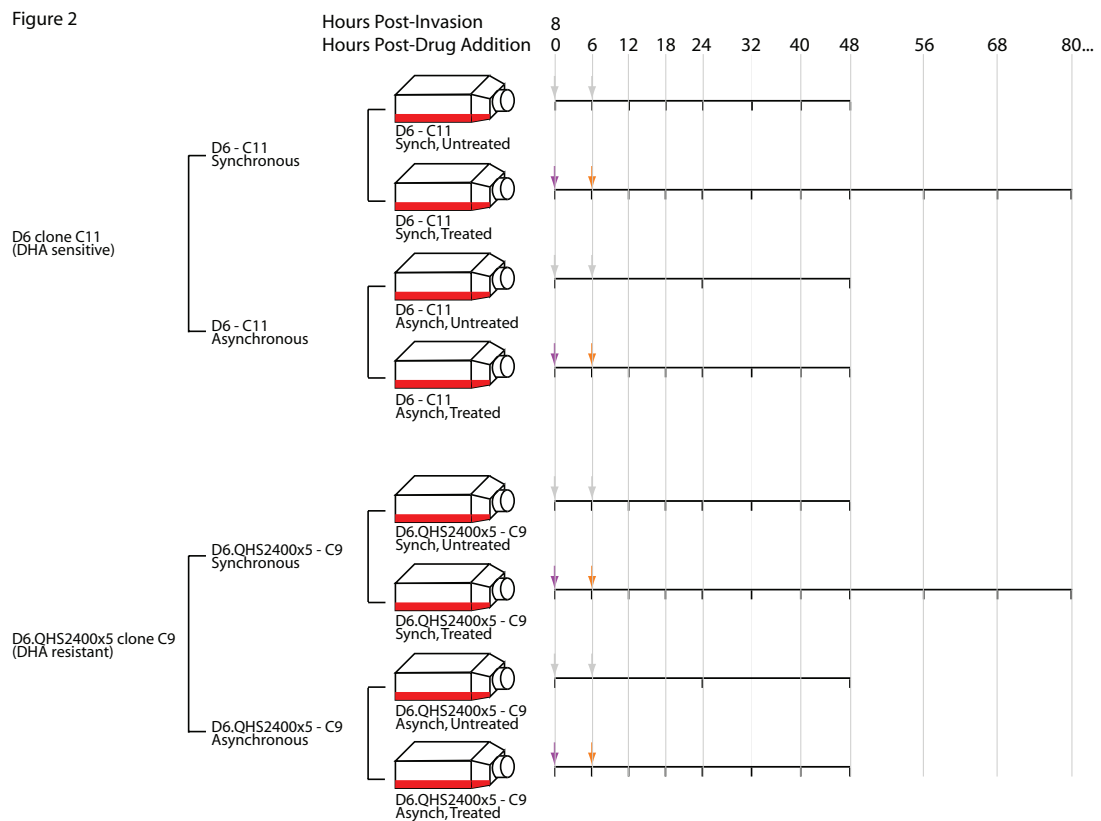


Figure 3 – Smear counts for dihydroartemisinin (DHA)-treated cultures. Black bars represent normal parasites (rings, trophs, and schizonts), while grey bars represent dormant parasites. The average count for each smear is plotted and the standard deviation for each smear is represented by the red lines. **A)** DHA-treated synchronous D6. **B)** DHA-treated synchronous D6.QHS2400x5. **C)** DHA-treated mixed stage D6. **D)** DHA-treated mixed stage D6.QHS2400x5.

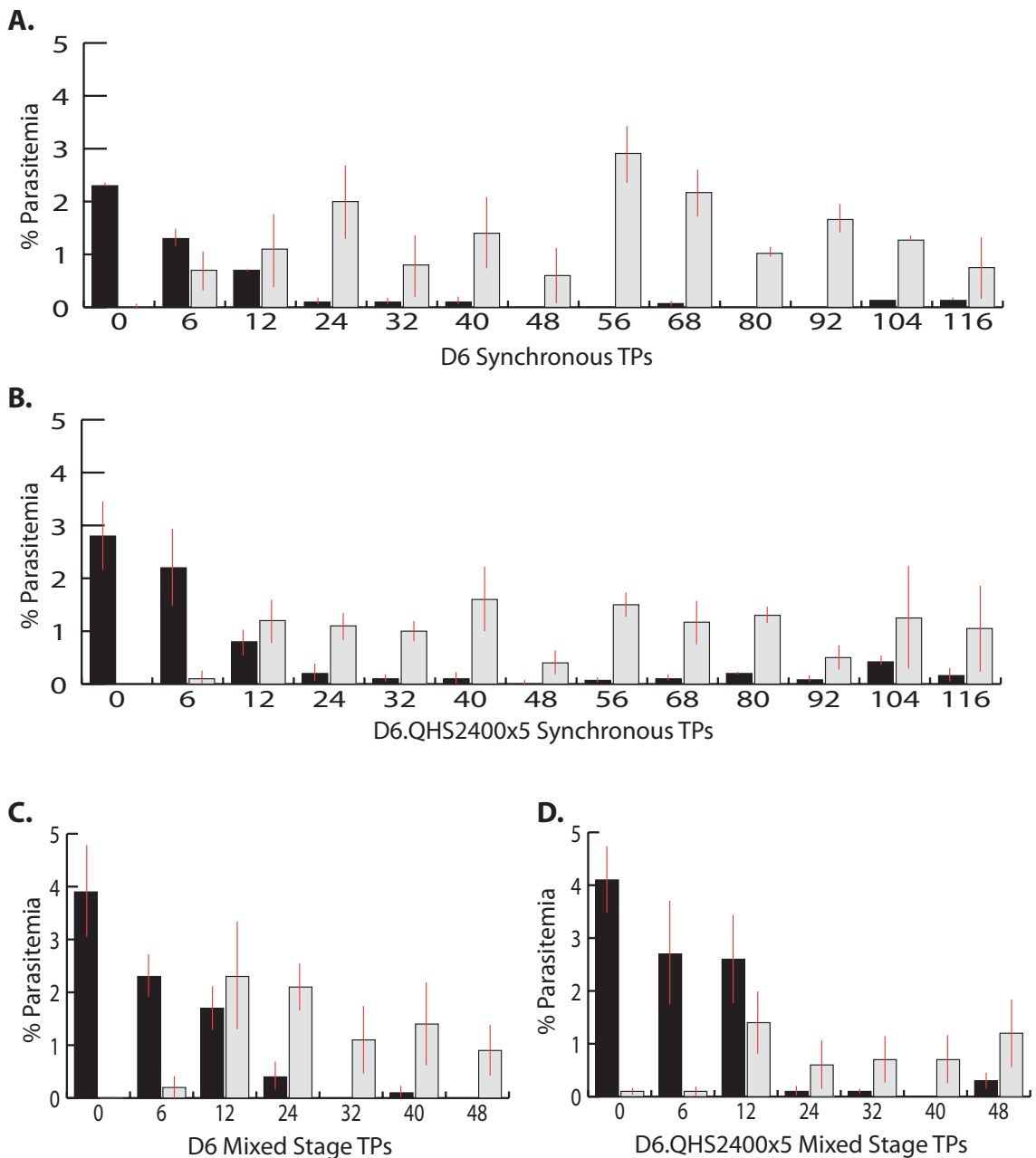


Figure 4 – Representative rings and dormant parasites (T=12) from the treated synchronous cultures.

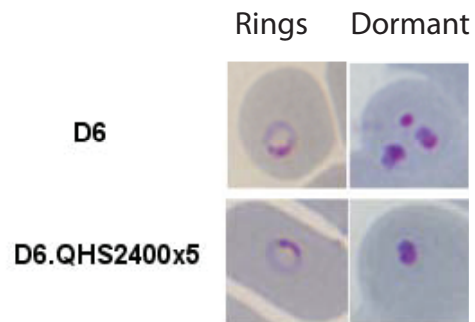


Figure 5 – Microarray results of DMSO and DHA-treated synchronous cultures as compared to the intraerythrocytic developmental cycle of normal HB3 parasites (22). Yellow indicates a positive Pearson correlation, while blue indicates a negative Pearson correlation. A) DMSO-treated synchronous D6 control culture. B) DHA-treated synchronous D6 culture. C) DMSO-treated synchronous D6.QHS2400x5 control culture. D) DHA-treated synchronous D6.QHS2400x5 culture.

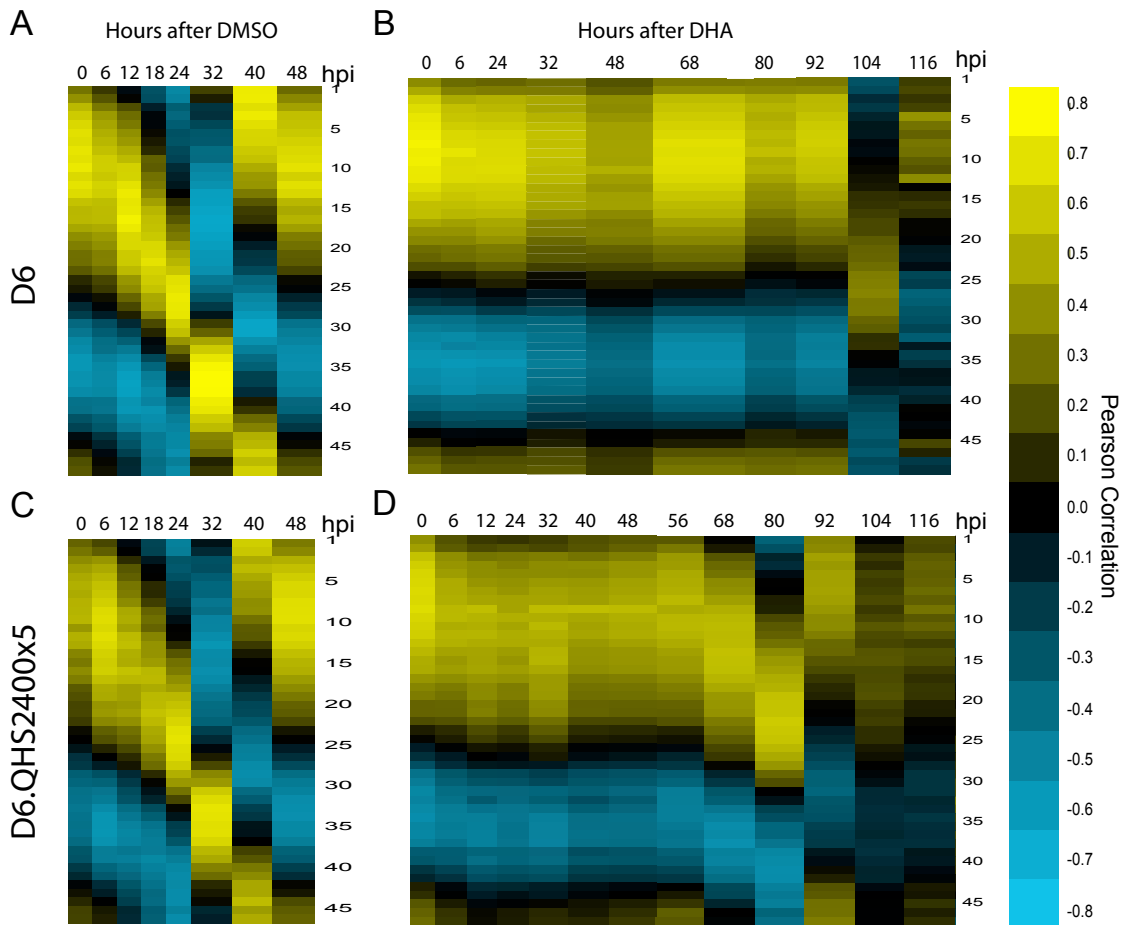
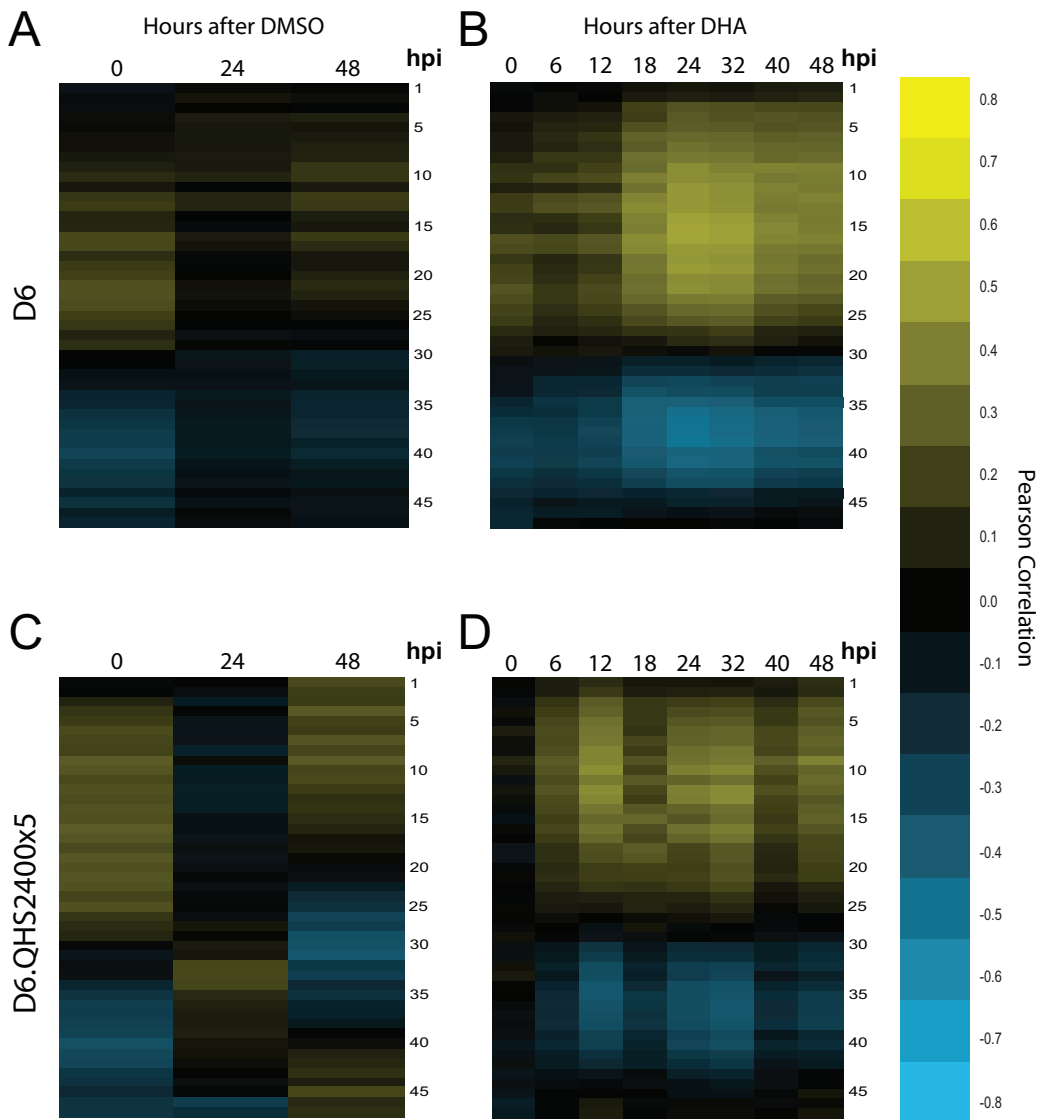


Figure 6 – Microarray results of DMSO and DHA-treated mixed stage cultures as compared to the intraerythrocytic developmental cycle of normal HB3 parasites (22). Yellow indicates a positive Pearson correlation, while blue indicates a negative Pearson correlation. A) DMSO-treated mixed stage D6 control culture. B) DHA-treated mixed stage D6 culture. C) DMSO-treated mixed stage D6.QHS2400x5 control culture. D) DHA-treated mixed stage D6.QHS2400x5 culture.



References:

1. Hay, S.I., Guerra, C.A., Gething, P.W., Patil, A.P., Tatem, A.J., Noor, A.M., Kabaria, C.W., Manh, B.H., Elyazar, I.R.F., Brooker, S. et al. (2009) A world malaria map: Plasmodium falciparum endemicity in 2007. *PLoS Med*, **6**, e1000048, 10.1371/journal.pmed.1000048.
2. WHO | World Malaria Report 2009 Available at: http://www.who.int/malaria/world_malaria_report_2009/en/index.html [Accessed June 14, 2010].
3. Golenser, J., Waknine, J.H., Krugliak, M., Hunt, N.H. and Grau, G.E. (2006) Current perspectives on the mechanism of action of artemisinin. *Int. J. Parasitol*, **36**, 1427-1441, 10.1016/j.ijpara.2006.07.011.
4. Nosten, F. and White, N.J. (2007) Artemisinin-Based Combination Treatment of Falciparum Malaria. *Am J Trop Med Hyg*, **77**, 181-192.
5. Chen, P.Q., Li, G.Q., Guo, X.B., He, K.R., Fu, Y.X., Fu, L.C. and Song, Y.Z. (1994) The infectivity of gametocytes of Plasmodium falciparum from patients treated with artemisinin. *Chin. Med. J*, **107**, 709-711.
6. Price, R.N., Nosten, F., Luxemburger, C., ter Kuile, F.O., Paiphun, L., Chongsuphajaisiddhi, T. and White, N.J. (1996) Effects of artemisinin derivatives on malaria transmissibility. *Lancet*, **347**, 1654-1658.
7. Olliaro, P.L., Haynes, R.K., Meunier, B. and Yuthavong, Y. (2001) Possible modes of action of the artemisinin-type compounds. *Trends Parasitol*, **17**, 122-126.
8. Eckstein-Ludwig, U., Webb, R.J., Van Goethem, I.D.A., East, J.M., Lee, A.G., Kimura, M., O'Neill, P.M., Bray, P.G., Ward, S.A. and Krishna, S. (2003)

- Artemisinin target the SERCA of Plasmodium falciparum. *Nature*, **424**, 957-961, 10.1038/nature01813.
9. Li,W., Mo,W., Shen,D., Sun,L., Wang,J., Lu,S., Gitschier,J.M. and Zhou,B. (2005) Yeast model uncovers dual roles of mitochondria in action of artemisinin. *PLoS Genet*, **1**, e36, 10.1371/journal.pgen.0010036.
10. Haynes,R.K., Chan,W., Wong,H., Li,K., Wu,W., Fan,K., Sung,H.H.Y., Williams,I.D., Prosperi,D., Melato,S. et al. (2010) Facile oxidation of leucomethylene blue and dihydroflavins by artemisinins: relationship with flavoenzyme function and antimalarial mechanism of action. *ChemMedChem*, **5**, 1282-1299, 10.1002/cmdc.201000225.
11. Wernsdorfer,W.H. (1994) Epidemiology of drug resistance in malaria. *Acta Trop*, **56**, 143-156.
12. Wongsrichanalai,C., Sirichaisinthop,J., Karwacki,J.J., Congpuong,K., Miller,R.S., Pang,L. and Thimasarn,K. (2001) Drug resistant malaria on the Thai-Myanmar and Thai-Cambodian borders. *Southeast Asian J. Trop. Med. Public Health*, **32**, 41-49.
13. Dondorp,A.M., Nosten,F., Yi,P., Das,D., Phyto,A.P., Tarning,J., Lwin,K.M., Arie,F., Hanpithakpong,W., Lee,S.J. et al. (2009) Artemisinin resistance in Plasmodium falciparum malaria. *N. Engl. J. Med*, **361**, 455-467, 10.1056/NEJMoa0808859.
14. Carrara,V.I., Zwang,J., Ashley,E.A., Price,R.N., Stepniewska,K., Barends,M., Brockman,A., Anderson,T., McGready,R., Phaiphun,L. et al. (2009) Changes in the treatment responses to artesunate-mefloquine on the northwestern border of Thailand during 13 years of continuous deployment. *PLoS ONE*, **4**, e4551,

- 10.1371/journal.pone.0004551.
15. Noedl,H., Socheat,D. and Satimai,W. (2009) Artemisinin-resistant malaria in Asia. *N. Engl. J. Med*, **361**, 540-541, 10.1056/NEJMc0900231.
16. Lim,P., Wongsrichanalai,C., Chim,P., Khim,N., Kim,S., Chy,S., Sem,R., Nhem,S., Yi,P., Duong,S. et al. (2010) Decreased in vitro susceptibility of Plasmodium falciparum isolates to artesunate, mefloquine, chloroquine, and quinine in Cambodia from 2001 to 2007. *Antimicrob. Agents Chemother*, **54**, 2135-2142, 10.1128/AAC.01304-09.
17. Teuscher,F., Gatton,M.L., Chen,N., Peters,J., Kyle,D.E. and Cheng,Q. (2010) Artemisinin-induced dormancy in plasmodium falciparum: duration, recovery rates, and implications in treatment failure. *J. Infect. Dis*, **202**, 1362-1368, 10.1086/656476.
18. Witkowski,B., Lelièvre,J., Barragán,M.J.L., Laurent,V., Su,X., Berry,A. and Benoit-Vical,F. (2010) Increased tolerance to artemisinin in Plasmodium falciparum is mediated by a quiescence mechanism. *Antimicrob. Agents Chemother*, **54**, 1872-1877, 10.1128/AAC.01636-09.
19. Chavchich,M., Gerena,L., Peters,J., Chen,N., Cheng,Q. and Kyle,D.E. (2010) Role of pfmdr1 amplification and expression in induction of resistance to artemisinin derivatives in Plasmodium falciparum. *Antimicrob. Agents Chemother*, **54**, 2455-2464, 10.1128/AAC.00947-09.
20. François,G., Hendrix,L. and Wery,M. (1994) A highly efficient in vitro cloning procedure for asexual erythrocytic forms of the human malaria parasite Plasmodium falciparum. *Ann Soc Belg Med Trop*, **74**, 177-185.

21. Bozdech,Z., Llinás,M., Pulliam,B.L., Wong,E.D., Zhu,J. and DeRisi,J.L. (2003) The Transcriptome of the Intraerythrocytic Developmental Cycle of Plasmodium falciparum. *PLoS Biol*, **1**, e5, 10.1371/journal.pbio.0000005.
22. Llinás,M., Bozdech,Z., Wong,E.D., Adai,A.T. and DeRisi,J.L. (2006) Comparative whole genome transcriptome analysis of three Plasmodium falciparum strains. *Nucleic Acids Res*, **34**, 1166-1173, 10.1093/nar/gkj517.
23. Tusher,V.G., Tibshirani,R. and Chu,G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. U.S.A*, **98**, 5116-5121, 10.1073/pnas.091062498.
24. Le Roch,K.G., Zhou,Y., Blair,P.L., Grainger,M., Moch,J.K., Haynes,J.D., De la Vega,P., Holder,A.A., Batalov,S., Carucci,D.J. et al. (2003) Discovery of Gene Function by Expression Profiling of the Malaria Parasite Life Cycle. *Science*, **301**, 1503-1508, 10.1126/science.1087025.
25. Björnstedt,M., Hamberg,M., Kumar,S., Xue,J. and Holmgren,A. (1995) Human thioredoxin reductase directly reduces lipid hydroperoxides by NADPH and selenocystine strongly stimulates the reaction via catalytically generated selenols. *J. Biol. Chem*, **270**, 11761-11764.
26. Nickel,C., Rahlfs,S., Deponete,M., Koncarevic,S. and Becker,K. (2006) Thioredoxin networks in the malarial parasite Plasmodium falciparum. *Antioxid. Redox Signal*, **8**, 1227-1239, 10.1089/ars.2006.8.1227.
27. Boschet,C., Gissot,M., Briquet,S., Hamid,Z., Claudel-Renard,C. and Vaquero,C. (2004) Characterization of PfMyb1 transcription factor during erythrocytic development of 3D7 and F12 Plasmodium falciparum clones. *Molecular and*

- Biochemical Parasitology*, **138**, 159-163, 10.1016/j.molbiopara.2004.07.011.
28. Shock,J.L., Fischer,K.F. and DeRisi,J.L. (2007) Whole-genome analysis of mRNA decay in Plasmodium falciparum reveals a global lengthening of mRNA half-life during the intra-erythrocytic development cycle. *Genome Biol*, **8**, R134, 10.1186/gb-2007-8-7-r134.
29. Willoughby,J.A., Sundar,S.N., Cheung,M., Tin,A.S., Modiano,J. and Firestone,G.L. (2009) Artemisinin blocks prostate cancer growth and cell cycle progression by disrupting Sp1 interactions with the cyclin-dependent kinase-4 (CDK4) promoter and inhibiting CDK4 gene expression. *J. Biol. Chem*, **284**, 2203-2213, 10.1074/jbc.M804491200.
30. Chen,H., Sun,B., Wang,S., Pan,S., Gao,Y., Bai,X. and Xue,D. (2010) Growth inhibitory effects of dihydroartemisinin on pancreatic cancer cells: involvement of cell cycle arrest and inactivation of nuclear factor-kappaB. *J. Cancer Res. Clin. Oncol*, **136**, 897-903, 10.1007/s00432-009-0731-0.
31. Wang,J., Tang,W., Shi,L., Wan,J., Zhou,R., Ni,J., Fu,Y., Yang,Y., Li,Y. and Zuo,J. (2007) Investigation of the immunosuppressive activity of artemether on T-cell activation and proliferation. *Br. J. Pharmacol*, **150**, 652-661, 10.1038/sj.bjp.0707137.
32. Hou,J., Wang,D., Zhang,R. and Wang,H. (2008) Experimental therapy of hepatoma with artemisinin and its derivatives: in vitro and in vivo activity, chemosensitization, and mechanisms of action. *Clin. Cancer Res*, **14**, 5519-5530, 10.1158/1078-0432.CCR-08-0197.
33. Efferth,T., Dunstan,H., Sauerbrey,A., Miyachi,H. and Chitambar,C.R. (2001) The

- anti-malarial artesunate is also active against cancer. *Int. J. Oncol*, **18**, 767-773.
34. Hu,G., Cabrera,A., Kono,M., Mok,S., Chaal,B.K., Haase,S., Engelberg,K., Cheemadan,S., Spielmann,T., Preiser,P.R. et al. (2010) Transcriptional profiling of growth perturbations of the human malaria parasite *Plasmodium falciparum*. *Nat. Biotechnol*, **28**, 91-98, 10.1038/nbt.1597.
35. Gunasekera,A.M., Myrick,A., Le Roch,K., Winzeler,E. and Wirth,D.F. (2007) *Plasmodium falciparum*: genome wide perturbations in transcript profiles among mixed stage cultures after chloroquine treatment. *Exp. Parasitol*, **117**, 87-92, 10.1016/j.exppara.2007.03.001.
36. Ganesan,K., Ponmee,N., Jiang,L., Fowble,J.W., White,J., Kamchonwongpaisan,S., Yuthavong,Y., Wilairat,P. and Rathod,P.K. (2008) A genetically hard-wired metabolic transcriptome in *Plasmodium falciparum* fails to mount protective responses to lethal antifolates. *PLoS Pathog*, **4**, e1000214, 10.1371/journal.ppat.1000214.

Chapter 6: 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.

Copyright © 2008, Public Library of Science, U.S.A.

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 (1). 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 (2), reversible chain-terminator extension (3), and ligation (4). Many of these strategies produce relatively short reads, in the range of 36-70 nt (5), 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 (19) or to the HBV genome (accession number: NC_003977) (20). 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 (21). 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 (22). 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 (23). 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) (20). 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 (21). 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) (24). 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 (25). Longer contigs permit more effective detection of

genetic homology to known sequences by use of BLASTN or TBLASTX (26, 27). 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) (28). 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.

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 | 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 | 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.

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.

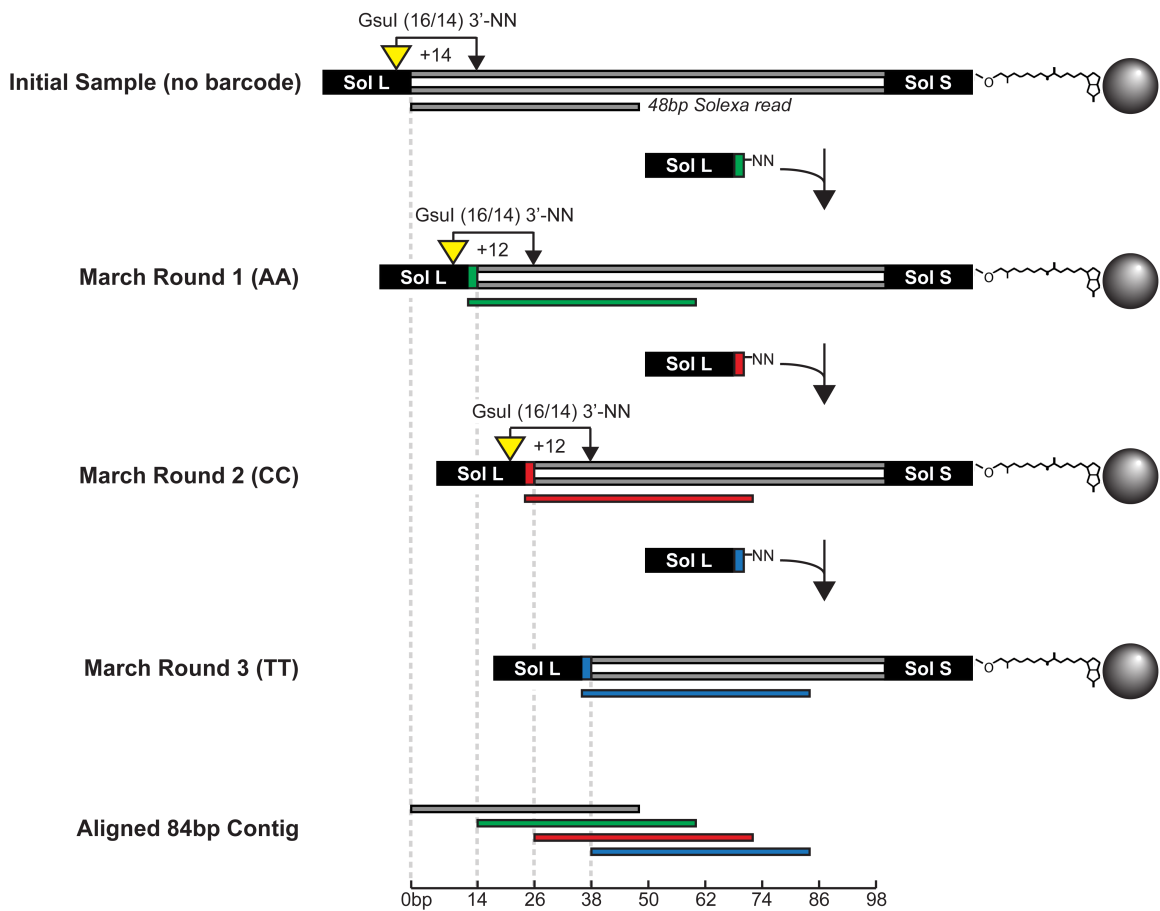


Figure 2. The long march produces barcoded, offset reads that aid in contig growth.

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.

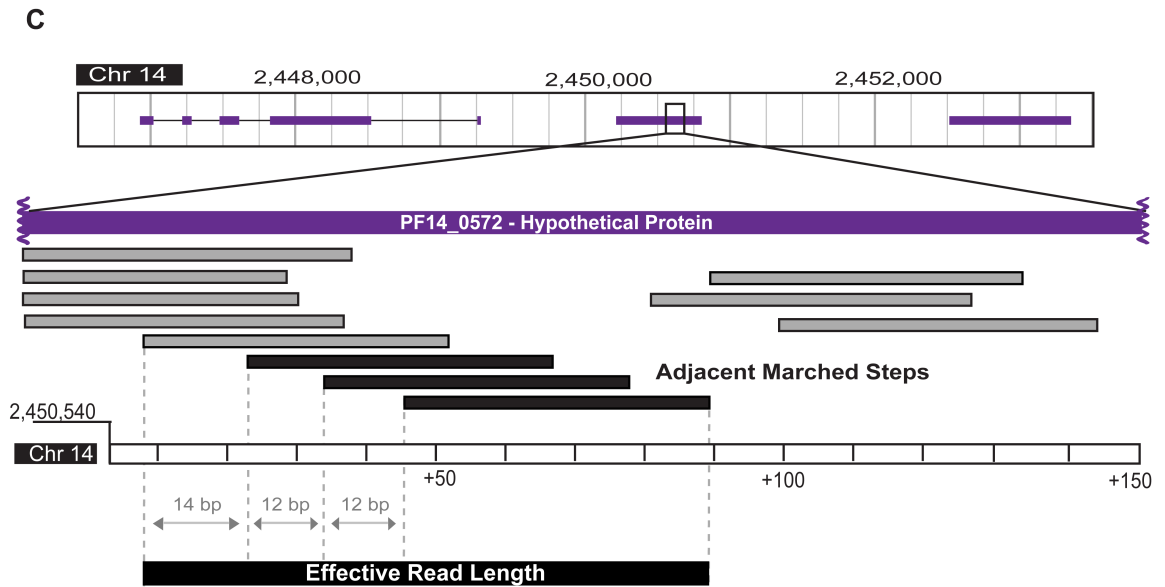
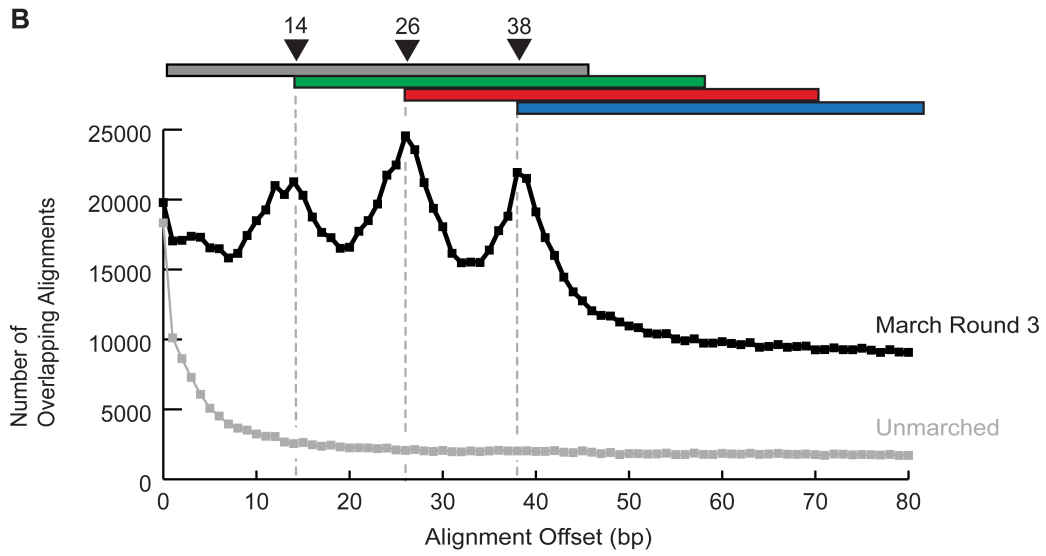
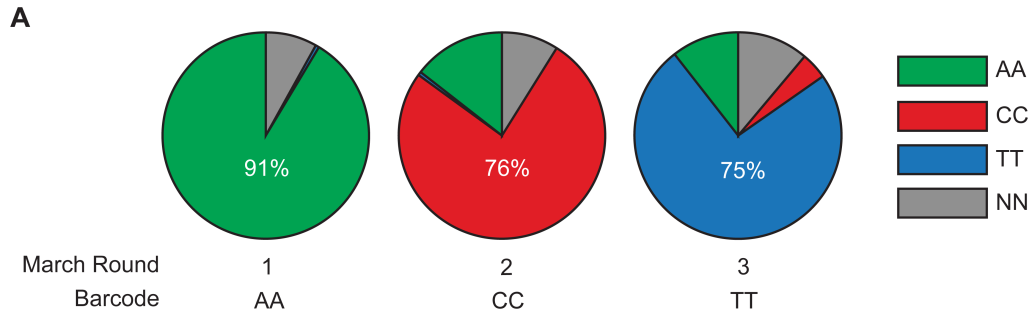


Figure 3. Marched sub-libraries show significantly increased genome coverage over a wide range of dataset sizes. Identical numbers of genome-aligned reads were randomly sampled from the round 3 sub-libraries and the initial libraries to simulate varying degrees of sequencing depth. The number of genomic 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.

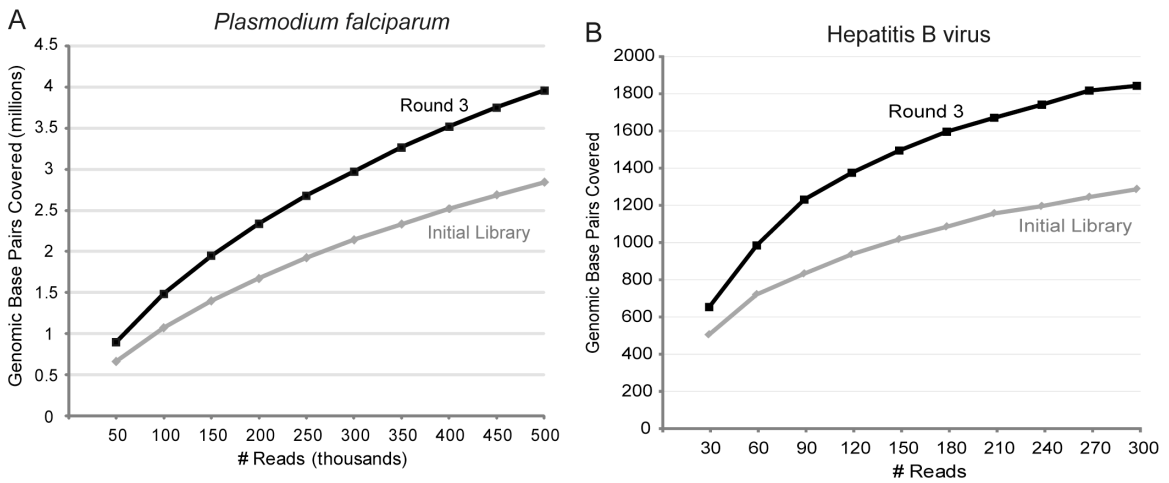


Figure 4. Theoretical optimization of the long march for *de novo* amplicon assembly.

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.

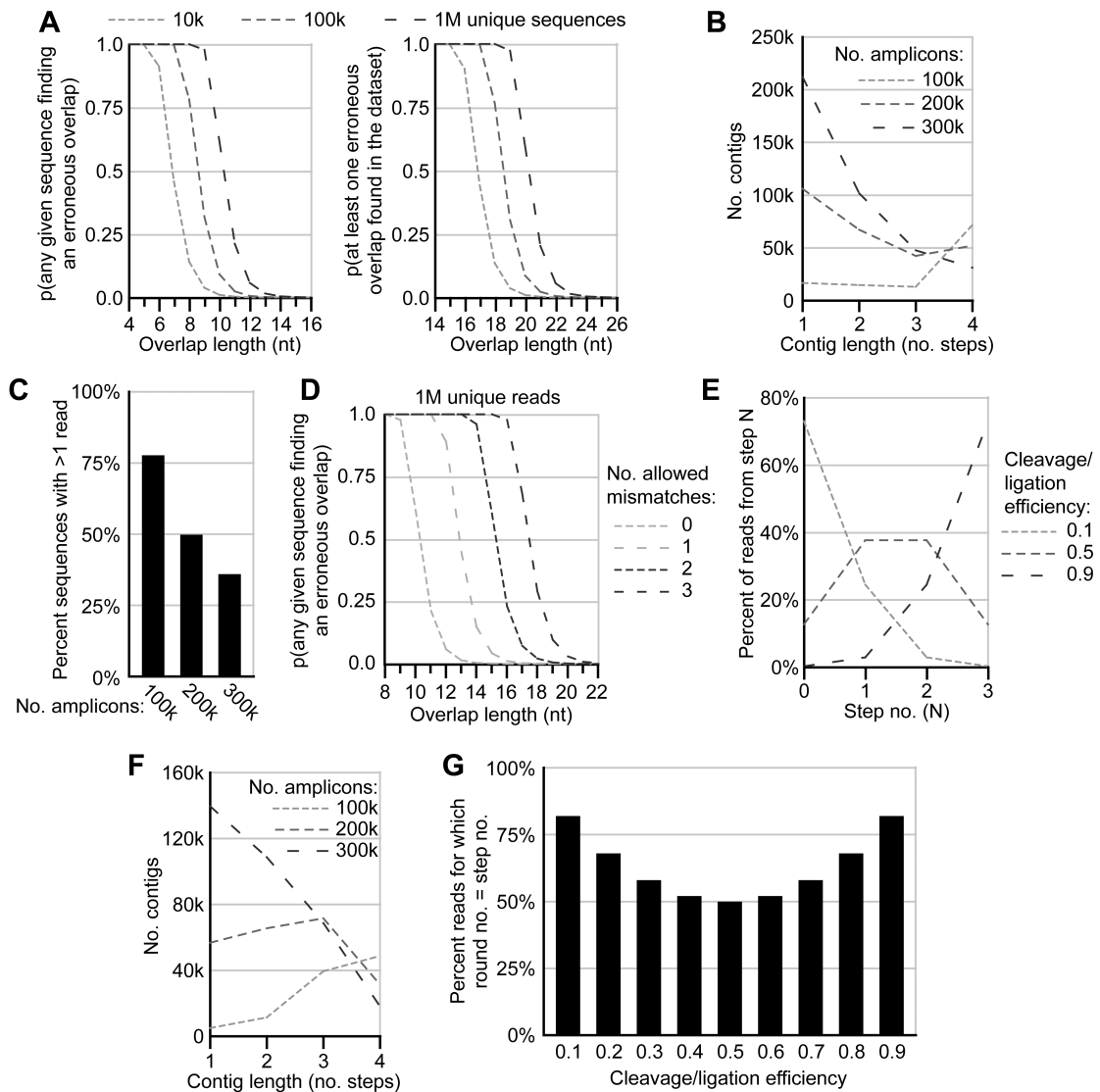
C) Effect of initial pool complexity on dataset redundancy. Simulations were performed as in (B) for each of the indicated amplicon pool complexities, and the fraction of unique sequences that were observed more than once is indicated.

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.



References:

1. Holt,R.A. and Jones,S.J.M. (2008) The new paradigm of flow cell sequencing.
Genome Res, **18**, 839-846, 10.1101/gr.073262.107.
2. Ronaghi,M., Karamohamed,S., Pettersson,B., Uhlén,M. and Nyrén,P. (1996) Real-time DNA sequencing using detection of pyrophosphate release. *Anal. Biochem*, **242**, 84-89, 10.1006/abio.1996.0432.
3. Seo,T.S., Bai,X., Ruparel,H., Li,Z., Turro,N.J. and Ju,J. (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, 10.1073/pnas.0401138101.
4. Shendure,J., Porreca,G.J., Reppas,N.B., Lin,X., McCutcheon,J.P., Rosenbaum,A.M., Wang,M.D., Zhang,K., Mitra,R.D. and Church,G.M. (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, **309**, 1728-1732, 10.1126/science.1117389.
5. Pop,M. and Salzberg,S.L. (2008) Bioinformatics challenges of new sequencing technology. *Trends Genet*, **24**, 142-149, 10.1016/j.tig.2007.12.006.
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., Sobic,Z. and Karger,B.L. (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., Lin,C., Holoch,D., Lim,C. and

- Tuschl,T. (2008) Identification of microRNAs and other small regulatory RNAs using cDNA library sequencing. *Methods*, **44**, 3-12, 10.1016/j.ymeth.2007.09.009.
9. Mardis,E.R. (2007) ChIP-seq: welcome to the new frontier. *Nat. Methods*, **4**, 613-614, 10.1038/nmeth0807-613.
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, 10.1093/nar/gkm901.
11. Chaisson,M., Pevzner,P. and Tang,H. (2004) Fragment assembly with short reads. *Bioinformatics*, **20**, 2067-2074, 10.1093/bioinformatics/bth205.
12. Whiteford,N., Haslam,N., Weber,G., Prügel-Bennett,A., Essex,J.W., Roach,P.L., Bradley,M. and Neylon,C. (2005) An analysis of the feasibility of short read sequencing. *Nucleic Acids Res*, **33**, e171, 10.1093/nar/gni170.
13. Korbel,J.O., Urban,A.E., Affourtit,J.P., Godwin,B., Grubert,F., Simons,J.F., Kim,P.M., Palejev,D., Carriero,N.J., Du,L. et al. (2007) Paired-end mapping reveals extensive structural variation in the human genome. *Science*, **318**, 420-426, 10.1126/science.1149504.
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, 10.1006/geno.2000.6303.
15. Mashayekhi,F. and Ronaghi,M. (2007) Analysis of read length limiting factors in Pyrosequencing chemistry. *Anal. Biochem*, **363**, 275-287, 10.1016/j.ab.2007.02.002.

16. Janulaitis,A., Bitinaite,J. and Jaskeleviciene,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. and Janulaitis,A. (1988) Restriction endonucleases of a new type. *Gene*, **74**, 89-91.
18. Wai,C., Fontana,R.J., Polson,J., Hussain,M., Shakil,A.O., Han,S., Davern,T.J., Lee,W.M. and Lok,A.S. (2005) Clinical outcome and virological characteristics of hepatitis B-related acute liver failure in the United States. *J. Viral Hepat*, **12**, 192-198, 10.1111/j.1365-2893.2005.00581.x.
19. Stoeckert,C.J., Fischer,S., Kissinger,J.C., Heiges,M., Aurrecochea,C., Gajria,B. and Roos,D.S. (2006) PlasmoDB v5: new looks, new genomes. *Trends Parasitol*, **22**, 543-546, 10.1016/j.pt.2006.09.005.
20. Okamoto,H., Imai,M., Shimozaki,M., Hoshi,Y., Iizuka,H., Gotanda,T., Tsuda,F., Miyakawa,Y. and Mayumi,M. (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.
21. Knight,R. and Yarus,M. (2003) Analyzing partially randomized nucleic acid pools: straight dope on doping. *Nucleic Acids Res*, **31**, e30.
22. Roberts,R.J., Belfort,M., Bestor,T., Bhagwat,A.S., Bickle,T.A., Bitinaite,J., Blumenthal,R.M., Degtyarev,S.K., Dryden,D.T.F., Dybvig,K. et al. (2003) A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. *Nucleic Acids Res*, **31**, 1805-1812.
23. Gardner,M.J., Hall,N., Fung,E., White,O., Berriman,M., Hyman,R.W., Carlton,J.M., Pain,A., Nelson,K.E., Bowman,S. et al. (2002) Genome sequence of the human

- malaria parasite *Plasmodium falciparum*. *Nature*, **419**, 498-511,
10.1038/nature01097.
24. Mathieu-Daudé,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.
25. Wommack,K.E., Bhavsar,J. and Ravel,J. (2008) Metagenomics: read length matters. *Appl. Environ. Microbiol*, **74**, 1453-1463, 10.1128/AEM.02181-07.
26. 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, 10.1006/jmbi.1990.9999.
27. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, **25**, 3389-3402.
28. Hadi,S.M., Bächli,B., Shepherd,J.C., Yuan,R., Ineichen,K. and Bickle,T.A. (1979) DNA recognition and cleavage by the EcoP15 restriction endonuclease. *J. Mol. Biol*, **134**, 655-666.

Chapter 7: HMMSplicer: a tool for efficient and sensitive discovery of known and novel splice junctions in RNA-Seq data

This chapter is a reprint from the following reference:

Michelle T. Dimon, Katherine Sorber, Joseph L. DeRisi. (2010) HMMSplicer: a tool for efficient and sensitive discovery of known and novel splice junctions in RNA-Seq data. PLoS ONE. 5(11): e13875.

Copyright © 2010, Public Library of Science, U.S.A.

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 DeRisi supervised the research.

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.

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 (26) 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 (27), 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 (28). 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.

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

Score for one side is calculated as:

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

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:

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

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 (29). 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 (30), 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.

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 schizogony). 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 | TTTTATGAGTGTCGTCCCTTTT |
| PFD0185c F1 | CGCACTACCATATTTATGCCTCT |
| PFD0185c R1 | AGTAGAAGGAGGGAGGAGCA |
| PFD0185c F2 | TTCGCGTGATGAAGAAGATG |
| PFD0185c R2 | CAAGCCCACATATAAATCAAGGA |
| PFC0285c F | TATCTTCTTGGGCCCTTCT |
| PFC0285c R | TGTGAATGCGTGAAGGATTT |

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 (31). 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 (30). 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 (32). 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 (33). *A. thaliana*, a model plant species, has a

genome of 125 million base pairs with ~25,500 protein-coding genes (34). The mean exon and intron sizes are 78 bp and 268 bp, respectively, with an average of 4.5 introns per gene (35).

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 (36). 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 (37). 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 (38). 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) (35), 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 (30). 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 (39)), 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 (40)).

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 (41). 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 (42). 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 (28) 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>.

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.

Table 1. Simulation Results. 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.

| | 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 |

Table 2. Simulation Results. 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.

| 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 |

Table 3. Datasets. 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. * The 48-bp reads in the NCBI SRA set have a 2 bp initial barcode that was trimmed, resulting in 46 bp reads.

| | 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) |

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.

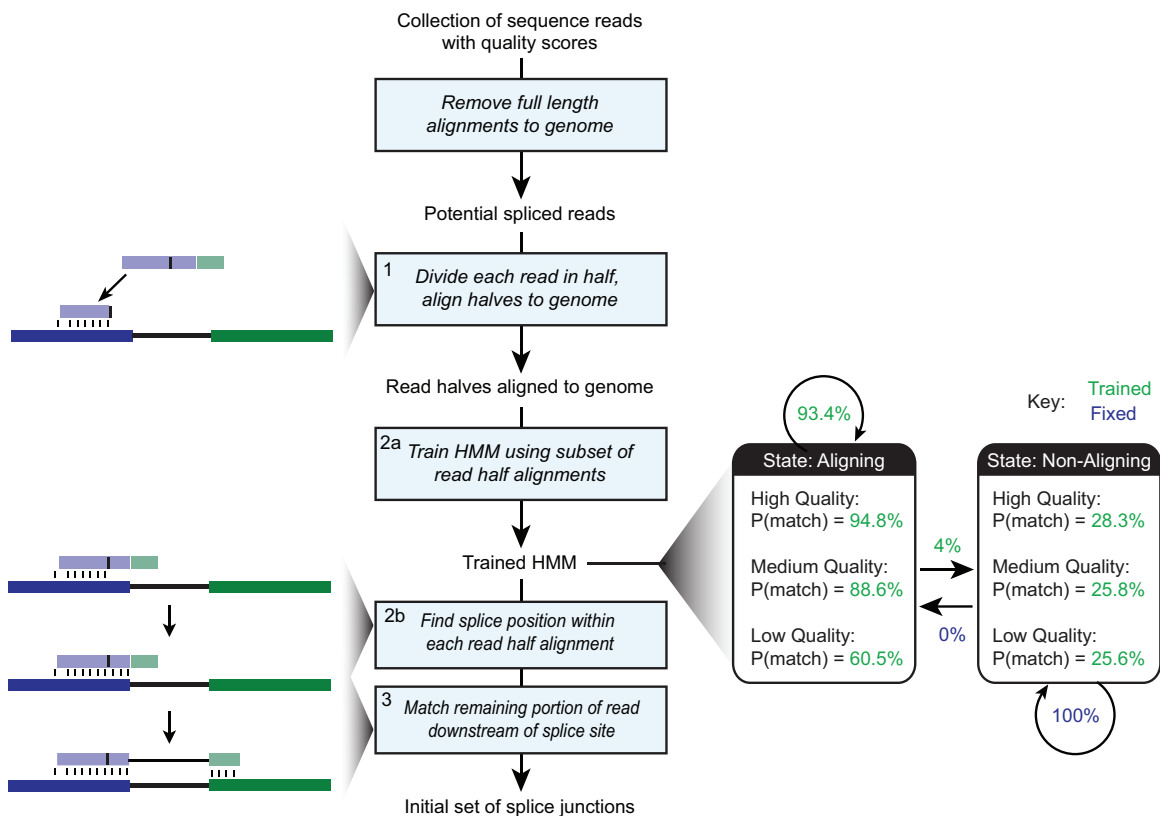


Figure 2. Algorithm parameters. **A)** Percent of oligos able to map within a genome as a function of oligo size. The solid lines show the percentages if oligos are able to map up to 50 times within the genome (the value used in HMMSplicer seeding). The dashed lines show the percentages if a unique match 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.

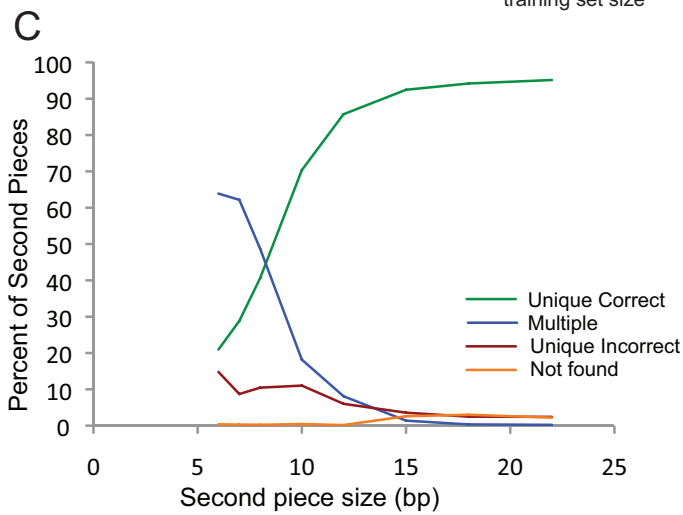
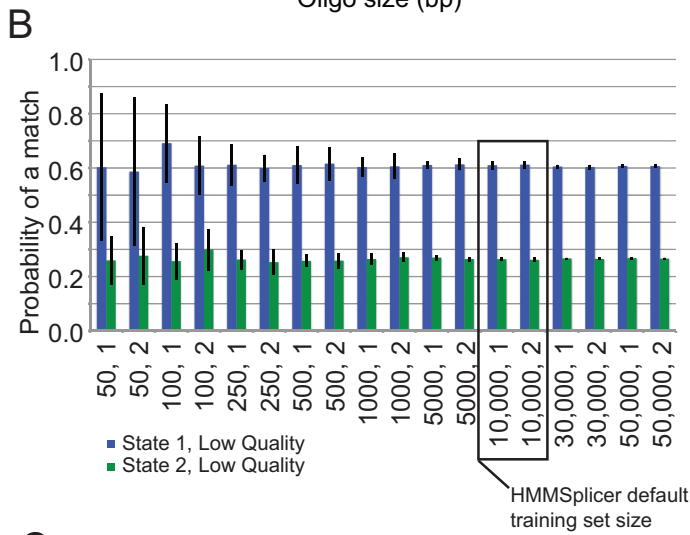
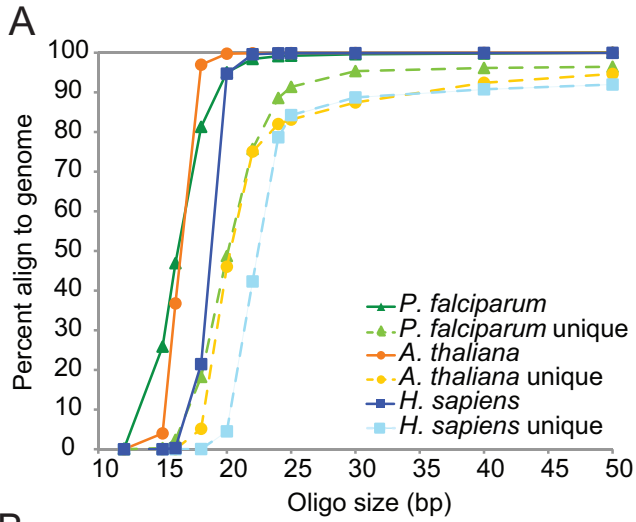


Figure 3. Simulation results. **A)** Results for HMMSplicer and TopHat for 50 and 75 bp reads. Although values are similar at higher coverage levels, HMMSplicer exhibits substantial increases in sensitivity at lower coverage levels. **B)** ROC curve for the 50 bp simulation results at 1x, 10x, and 50x 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.

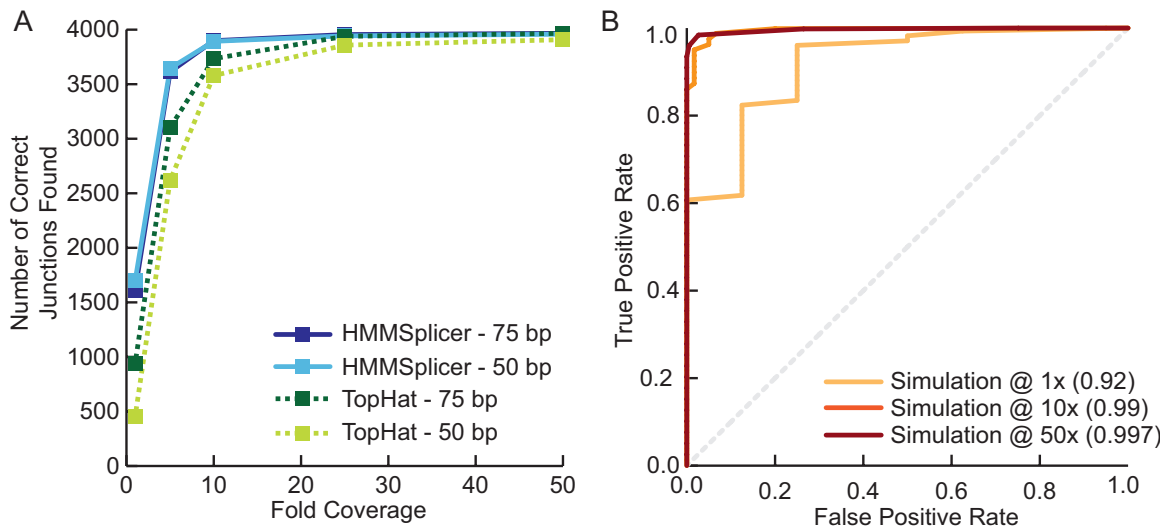


Figure 4. Overview of HMMSplicer and TopHat results in (a) *A. thaliana*, and (b) *P. falciparum* and (c) *H. sapiens*. For each dataset, HMMSplicer results are shown at five different score thresholds. The numbers on the bottom axis (200 to 600) are the thresholds for junctions with multiple reads; 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.

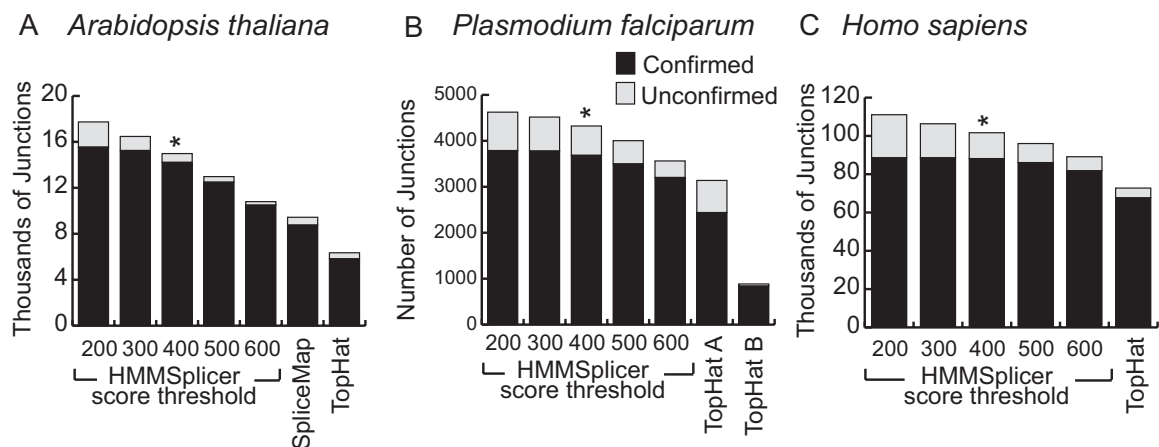


Figure 5. Human results compared by transcript abundance. Transcript abundance was measured as Reads Per Kilobase per Million reads mapped (RPKM) and the genes were binned by RPKM to show the number of RefSeq junctions found at different levels of transcript abundance. For genes with an RPKM less than 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.

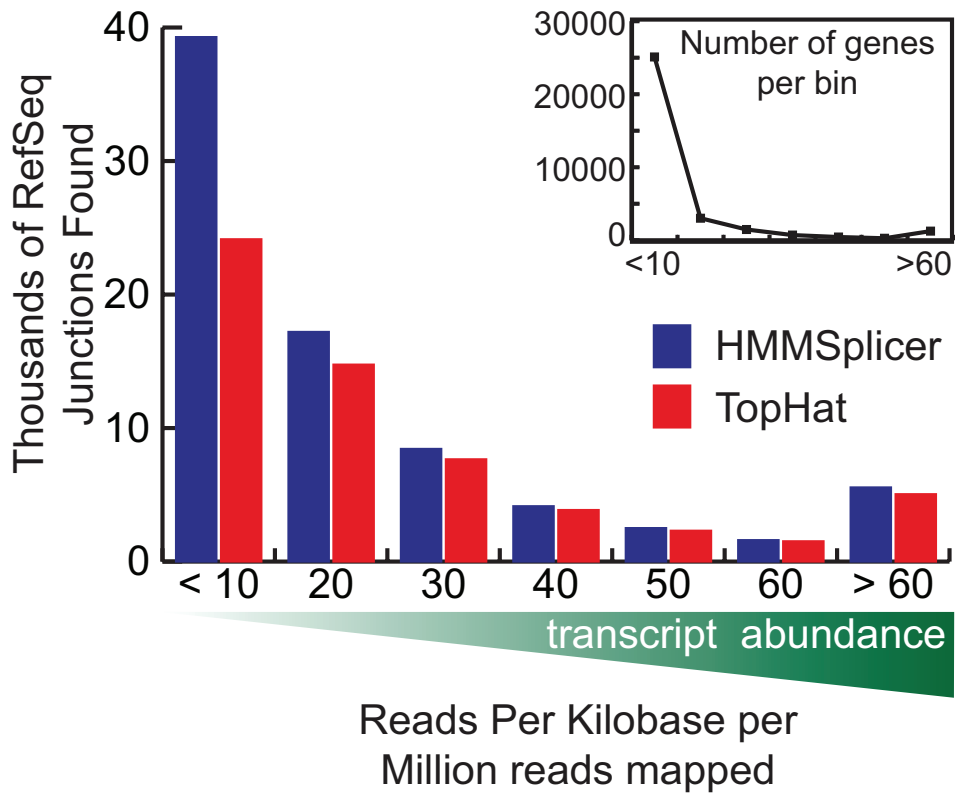


Figure 6. Alternative 5' and 3' splice sites. HMMSplicer results within 15 bp of RefSeq introns were analyzed to measure the number of bases added or removed from the spliced transcript. There were 997 instances where the intron had an alternate 5' splice site (5'SS, shown in grey) and 2,577 instances 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.

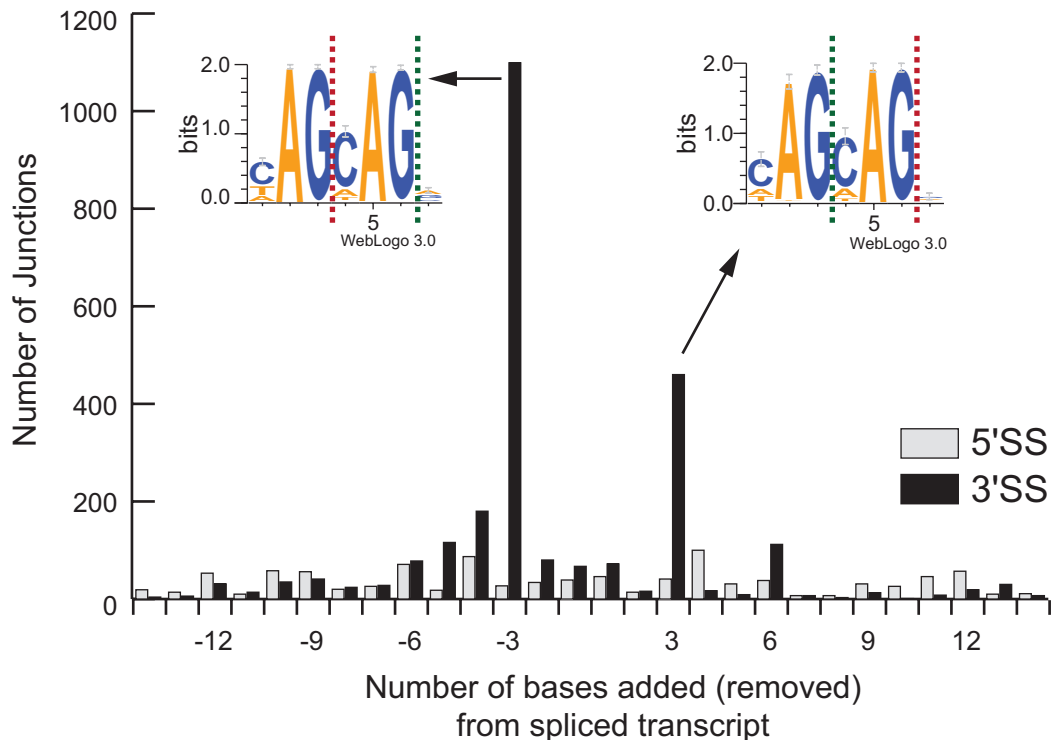


Figure 7. XBP1 non-canonical intron. HMMSplicer discovers the non-canonical *XBP1* intron. HMMSplicer identifies three reads containing the non-canonical CA-AG splice site in *XBP1*. Because the reads are fairly evenly split, both read-halves aligned to the genome. The edges identified by HMMSplicer 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.

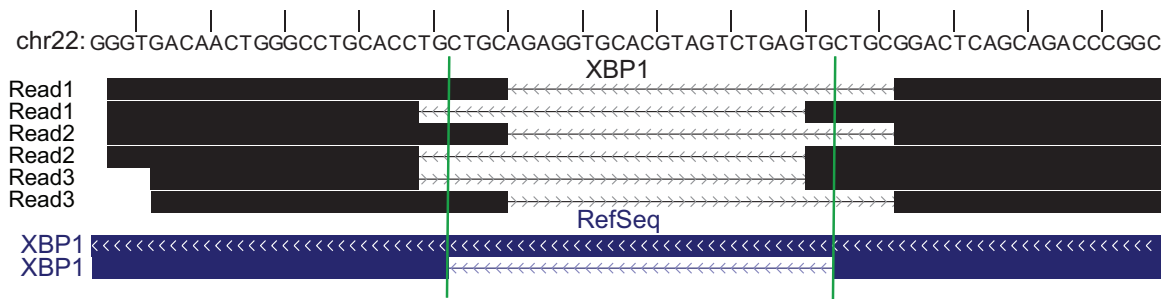
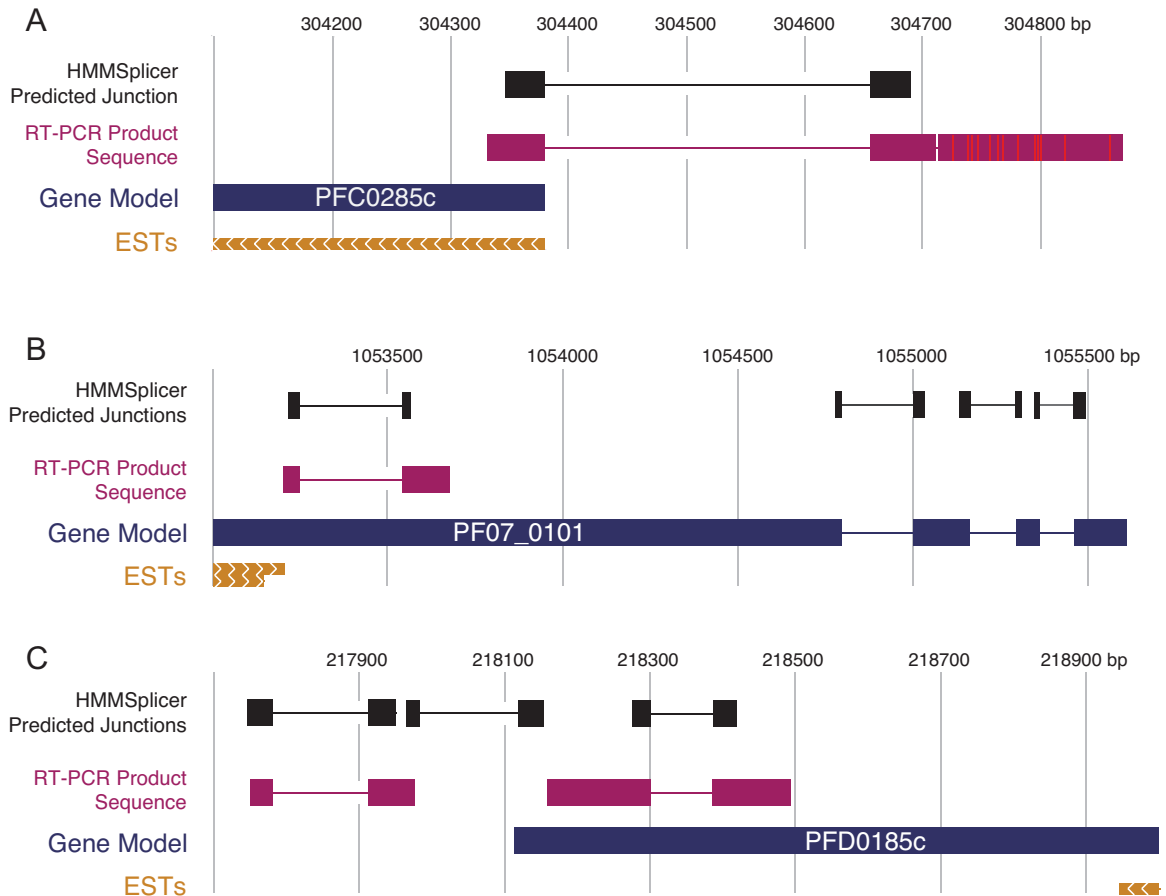


Figure 8. Experimental confirmation of predicted *Plasmodium falciparum* splice junctions. Schematics of the predicted splice junctions and sequenced RT-PCR products for **A) PFC0285c**, **B) PF07_0101**, and **C) PFD0185c**. For PFC0285c, the verified junction likely splices an additional exon in the 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.



References:

1. Wang,Z., Gerstein,M. and Snyder,M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, **10**, 57-63, 10.1038/nrg2484.
2. Wilhelm,B.T., Marguerat,S., Watt,S., Schubert,F., Wood,V., Goodhead,I., Penkett,C.J., Rogers,J. and Bähler,J. (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*, **453**, 1239-1243, 10.1038/nature07002.
3. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754-1760, 10.1093/bioinformatics/btp324.
4. Li,R., Yu,C., Li,Y., Lam,T., Yiu,S., Kristiansen,K. and Wang,J. (2009) SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, **25**, 1966-1967, 10.1093/bioinformatics/btp336.
5. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25, 10.1186/gb-2009-10-3-r25.
6. Wahl,M.C., Will,C.L. and Lührmann,R. (2009) The spliceosome: design principles of a dynamic RNP machine. *Cell*, **136**, 701-718, 10.1016/j.cell.2009.02.009.
7. Stamm,S., Riethoven,J., Le Texier,V., Gopalakrishnan,C., Kumanduri,V., Tang,Y., Barbosa-Morais,N.L. and Thanaraj,T.A. (2006) ASD: a bioinformatics resource on alternative splicing. *Nucleic Acids Res*, **34**, D46-55, 10.1093/nar/gkj031.
8. Yoshida,H., Matsui,T., Yamamoto,A., Okada,T. and 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,J.S. and 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,J.S. and Walter,P. (1996) tRNA ligase is required for regulated mRNA splicing in the unfolded protein response. *Cell*, **87**, 405-413.
 11. Nilsen,T.W. and Graveley,B.R. (2010) Expansion of the eukaryotic proteome by alternative splicing. *Nature*, **463**, 457-463, 10.1038/nature08909.
 12. Shepard,P.J. and Hertel,K.J. (2009) The SR protein family. *Genome Biol*, **10**, 242, 10.1186/gb-2009-10-10-242.
 13. Wang,E.T., Sandberg,R., Luo,S., Khrebtkova,I., Zhang,L., Mayr,C., Kingsmore,S.F., Schroth,G.P. and Burge,C.B. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470-476, 10.1038/nature07509.
 14. Pan,Q., Shai,O., Lee,L.J., Frey,B.J. and Blencowe,B.J. (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet*, **40**, 1413-1415, 10.1038/ng.259.
 15. Nagasaki,H., Arita,M., Nishizawa,T., Suwa,M. and Gotoh,O. (2005) Species-specific variation of alternative splicing and transcriptional initiation in six eukaryotes. *Gene*, **364**, 53-62, 10.1016/j.gene.2005.07.027.
 16. Sen,S., Talukdar,I., Liu,Y., Tam,J., Reddy,S. and Webster,N.J.G. (2010) Muscleblind-like 1 (Mbnl1) promotes insulin receptor exon 11 inclusion via binding to a downstream evolutionarily conserved intronic enhancer. *J. Biol. Chem*, **285**, 25426-25437, 10.1074/jbc.M109.095224.

17. Yano,M., Hayakawa-Yano,Y., Mele,A. and Darnell,R.B. (2010) Nova2 regulates neuronal migration through an RNA switch in disabled-1 signaling. *Neuron*, **66**, 848-858, 10.1016/j.neuron.2010.05.007.
18. Mortazavi,A., Williams,B.A., McCue,K., Schaeffer,L. and Wold,B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621-628, 10.1038/nmeth.1226.
19. Kent,W.J. (2002) BLAT--the BLAST-like alignment tool. *Genome Res*, **12**, 656-664, 10.1101/gr.229202. Article published online before March 2002.
20. Nagalakshmi,U., Wang,Z., Waern,K., Shou,C., Raha,D., Gerstein,M. and Snyder,M. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, **320**, 1344-1349, 10.1126/science.1158441.
21. Trapnell,C., Pachter,L. and Salzberg,S.L. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105-1111, 10.1093/bioinformatics/btp120.
22. Trapnell,C., Williams,B.A., Pertea,G., Mortazavi,A., Kwan,G., van Baren,M.J., Salzberg,S.L., Wold,B.J. and Pachter,L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol*, **28**, 511-515, 10.1038/nbt.1621.
23. Ameer,A., Wetterbom,A., Feuk,L. and Gyllensten,U. (2010) Global and unbiased detection of splice junctions from RNA-seq data. *Genome Biol*, **11**, R34, 10.1186/gb-2010-11-3-r34.
24. Bryant,D.W., Shen,R., Priest,H.D., Wong,W. and Mockler,T.C. (2010) Supersplat--spliced RNA-seq alignment. *Bioinformatics*, **26**, 1500-1505, 10.1093/bioinformatics/btq206.

25. Au, K.F., Jiang, H., Lin, L., Xing, Y. and Wong, W.H. (2010) Detection of splice junctions from paired-end RNA-seq data by SpliceMap. *Nucleic Acids Res*, **38**, 4570-4578, 10.1093/nar/gkq211.
26. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M. and Haussler, D. (2002) The human genome browser at UCSC. *Genome Res*, **12**, 996-1006, 10.1101/gr.229102. Article published online before print in May 2002.
27. Baum, L., Petrie, T., Soules, G. and Weiss, N. A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. *The Annals of Mathematical Statistics*, **41**, 164-171.
28. Crooks, G.E., Hon, G., Chandonia, J. and Brenner, S.E. (2004) WebLogo: a sequence logo generator. *Genome Res*, **14**, 1188-1190, 10.1101/gr.849004.
29. De Bona, F., Ossowski, S., Schneeberger, K. and Ratsch, G. (2008) Optimal spliced alignments of short sequence reads. *Bioinformatics*, **24**, i174-180, 10.1093/bioinformatics/btn300.
30. Heap, G.A., Yang, J.H.M., Downes, K., Healy, B.C., Hunt, K.A., Bockett, N., Franke, L., Dubois, P.C., Mein, C.A., Dobson, R.J. 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, 10.1093/hmg/ddp473.
31. Dohm, J.C., Lottaz, C., Borodina, T. and Himmelbauer, H. (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res*, **36**, e105, 10.1093/nar/gkn425.
32. Li, H., Ruan, J. and Durbin, R. (2008) Mapping short DNA sequencing reads and

- calling variants using mapping quality scores. *Genome Res*, **18**, 1851-1858, 10.1101/gr.078212.108.
33. Lister,R., O'Malley,R.C., Tonti-Filippini,J., Gregory,B.D., Berry,C.C., Millar,A.H. and Ecker,J.R. (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell*, **133**, 523-536, 10.1016/j.cell.2008.03.029.
34. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana (2000) *Nature*, **408**, 796-815, 10.1038/35048692.
35. Deutsch,M. and Long,M. (1999) Intron-exon structures of eukaryotic model organisms. *Nucleic Acids Res*, **27**, 3219-3228.
36. Gardner,M.J., Hall,N., Fung,E., White,O., Berriman,M., Hyman,R.W., Carlton,J.M., Pain,A., Nelson,K.E., Bowman,S. et al. (2002) Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature*, **419**, 498-511, 10.1038/nature01097.
37. Lu,F., Jiang,H., Ding,J., Mu,J., Valenzuela,J.G., Ribeiro,J.M.C. and Su,X. (2007) cDNA sequences reveal considerable gene prediction inaccuracy in the Plasmodium falciparum genome. *BMC Genomics*, **8**, 255, 10.1186/1471-2164-8-255.
38. Sorber,K., Chiu,C., Webster,D., Dimon,M., Ruby,J.G., Hekele,A. and DeRisi,J.L. (2008) The long march: a sample preparation technique that enhances contig length and coverage by high-throughput short-read sequencing. *PLoS ONE*, **3**, e3495, 10.1371/journal.pone.0003495.
39. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and

- proteins. *Nucleic Acids Res*, **33**, D501-504, 10.1093/nar/gki025.
40. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2004) GenBank: update. *Nucleic Acids Res*, **32**, D23-26, 10.1093/nar/gkh045.
41. Ramsköld,D., Wang,E.T., Burge,C.B. and Sandberg,R. (2009) An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput. Biol.*, **5**, e1000598, 10.1371/journal.pcbi.1000598.
42. Richard,H., Schulz,M.H., Sultan,M., Nürnberger,A., Schrinner,S., Balzereit,D., Dagand,E., Rasche,A., Lehrach,H., Vingron,M. et al. (2010) Prediction of alternative isoforms from exon expression levels in RNA-Seq experiments. *Nucleic Acids Res*, **38**, e112, 10.1093/nar/gkq041.

**Chapter 8: RNA-Seq Analysis of Splicing in *Plasmodium falciparum* Uncovers
New Splice Junctions, Alternative Splicing, and Splicing of Antisense Transcripts**

This chapter is a reprint from the following reference:

Sorber K, Dimon MT, DeRisi JL. (2010). RNA-Seq Analysis of Splicing in *Plasmodium falciparum* Uncovers New Splice Junctions, Alternative Splicing, and Splicing of Antisense Transcripts. *Nucleic Acids Research*. In Press.

Copyright © 2010, Oxford Journals.

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

Introduction:

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

Although the molecular mechanism of RNA splicing remains murky in *P. falciparum*, it has been well studied in model organisms. In the classical pathway, two transesterification steps are catalyzed by the spliceosome, a large complex of small nuclear ribonucleoproteins (snRNPs), each containing an snRNA component and a core set of proteins. In *P. falciparum*, RNA components of the major U2-type spliceosome

have been detected ^(7, 8), but only one protein component, a UAP56 homolog, has been definitively identified ⁽⁹⁾. As with splicing components, elucidation of the motifs guiding splicing also remains incomplete. Typically these motifs include the 5' splice site (AG|GUAAUGU in yeast, AG|GURAGU in mammals), the branch point sequence (UACU AAC in yeast, YNYURAY in mammals), the poly-pyrimidine tract (variable length in both yeast and mammals), and the 3' splice site (CAG| in yeast, YAG| in mammals) ⁽¹⁰⁾. In *P. falciparum*, EST data has been used to generate putative 5' (AR|GUAANW) and 3' (YAG|) splice site motifs ⁽⁸⁾. As in most eukaryotes, the first and last two nucleotides of the intron (5' GU-AG 3') are the most consistent markers of intronic sequence. In other organisms, a minority of introns are marked by noncanonical splice sites such as 5' GC-AG 3' (recognized by the major U2-type spliceosome) and 5' AU-AC 3' (recognized by the minor U12-type spliceosome) ⁽¹¹⁾. Noncanonical splice sites occur in *P. falciparum* EST data ^(12, 13) and have been incorporated into some gene models, yet no study to date has documented the types of intron boundaries recognized by the parasite.

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

Recent analyses have shown that transcriptome complexity in many organisms extends beyond alternative splicing. Dense transcriptional arrangements, such as

overlapping protein-coding genes (in parallel or antiparallel orientation) and natural antisense transcripts^(22, 23), now appear to be commonplace rather than anomalous. Although the functional importance of these arrangements is not yet well understood, some are known to be important in regulatory relationships between the paired genes⁽²⁴⁾. In current *P. falciparum* gene models, six instances of protein-coding gene overlap are annotated, resulting in 1 parallel and 5 antiparallel gene pairs. In addition, RNA polymerase II has been shown to synthesize long antisense transcripts in the parasite⁽²⁵⁾, and EST data indicates that at least one of these may be spliced⁽¹²⁾. Short antisense transcripts have also been described⁽²⁶⁾.

In this study, RNA-Seq data was generated from four timepoints in the intraerythrocytic transcriptome of *Plasmodium falciparum* for the purpose of characterizing splicing in this organism. Unbiased, gene model-independent splice site detection within our dataset in conjunction with RNA-Seq data from Otto et al. and Sorber et al.^(14, 27) was accomplished using the HMMSplicer algorithm⁽²⁸⁾, which was specifically developed to handle the challenging RNA-Seq datasets generated from the A/T-rich genome of *P. falciparum*. 977 new 5' GU-AG 3' and 5 new 5' GC-AG 3' junctions never before documented in gene models or ESTs were discovered. Further analysis uncovered alternative splicing events, largely within 254 genes, as well as splicing events antisense to one another. Antisense events, some of which themselves displayed alternative splicing, likely indicate a mix of overlapping annotated genes transcribed from opposite strands and unannotated transcripts transcribed antisense to gene models. Unexpectedly, antisense introns overlap sense gene introns more than would be anticipated by chance, perhaps indicating some relationship between

overlapping transcripts, or an inherent feature of transcriptome organization. Over 30 antisense and alternative splicing events were independently experimentally verified, indicating that the new, alternative, and antisense splicing events elucidated here support a larger, more dynamic understanding of the parasite's transcriptome.

Materials and Methods:

Best reciprocal hits analysis and prediction of RS domain proteins:

S. cerevisiae and *H. sapiens* splicing factor 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^(29, 30). Prediction of *Plasmodium falciparum* RS domain proteins was adapted from Boucher et al.⁽³¹⁾. 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 T150 mL flasks with 50 mL of volume each. Repeat synchronization during peak

invasion 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. ⁽¹⁾. 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 (TP1), pelleted, and frozen at -80°C. 33% of the culture was harvested 22 hours after invasion (TP2), 10% 33 hours after invasion (TP3), and 7% 44 hours after invasion (TP4). 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. ⁽²⁷⁾. Briefly, 1.2-1.6µg of polyA-selected RNA was reverse transcribed using 6bp-EciI-N₉ (all primers can be found in Table 2), 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⁽²⁷⁾. The additional TP4 library sequenced here derived from a fourth Long March of the thrice-marched library described in Sorber et al. annealed to the Sol-L-AC-NN adapter⁽²⁷⁾. Final PCR on marched sub-libraries was as described for initial libraries.

Illumina sequencing of RNA-Seq libraries:

For TP1-3, the initial library and the thrice-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, and PE-SolS-SeqPrimer was used to sequence the second read of paired-end libraries. Up to 60 single base extensions were performed with image capture using an Illumina GA2 sequencer (Illumina, Hayward, CA; see Table 3). 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, and versions 1.3.2 and 1.5.0 were used to base call TP1 and TP2 images. All primary sequencing data can be found in the NCBI Short Read Archive under accession number SRA024324.1.

Analysis pipeline:

Raw sequence data from the above timecourse as well as from Otto et al. and Sorber et al.^(14, 27) were aggregated and any barcodes were removed. Reads with greater than 12 nt of adapter sequence, a repeat of A, T, C, G, or AT longer than 11 nt, or more than 10 nt with a quality scores ≤ 5 were discarded. Identical sequences within a timepoint were compressed to a single sequence read and the reads were filtered to remove human

sequences, as detected by BLAST against the human genome with an E-value of 1×10^{-5} (30).

To gauge overall coverage, the filtered read set was aligned to the *Plasmodium falciparum* genome, PlasmoDB version 6.3 (32), by Bowtie version 0.12.1, using default parameters except that alignment of reads with multiple matches was disallowed (33). Reads unaligned by Bowtie were then aligned using BLAT version 34 with a tile size of 11, a step size of 1, and using an ooc file to filter repetitive sequence (34). Bowtie alignments were combined with BLAT alignments score ≥ 35 to yield the final set of aligned reads from which coverage statistics were generated.

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

Constuction of human splice site WebLogos:

WebLogos were created for human 5' and 3' splice sites using a random sampling of 20,000 genes from the human Consensus CDS (<http://www.ncbi.nlm.nih.gov/projects/CCDS/CcdsBrowse.cgi>, accessed October 2009).

Operational definitions for data analysis:

To avoid confusion, a specific terminology was used to refer to specific parts of individual splice junctions and to classify junctions (Figure 1A-D). For all definitions

referencing gene models, a junction maps to a gene model only if at least one inner edge falls within the bounding coordinates of the gene model.

A “known junction” maps to the same pair of inner boundaries as a splice junction found in PlasmoDBv6.3 gene models or in EST data (Figure 1B). A “new junction” maps to a pair of boundaries not seen in PlasmoDBv6.3 gene models or in EST data. “Canonical junctions” map to 5' GU-AG 3' boundaries, while “noncanonical junctions” map to all other possible boundaries. A “junction conflict” occurs when an inner edge of one junction falls within the intronic portion of the other junction such that they must occur in a mutually exclusive manner (Figure 1C). “Junction groups” were built by randomly selecting a nucleating junction, then searching for all relevant conflicting junctions. These junctions were added to the group and the search was iterated until no new junctions were appended. “Alternate 5' and 3' splice sites” refers to splice junctions where both the 5' and 3' splice sites conflict (Figure 1C). A splice junction that conflicts with two or more junctions that themselves do not conflict is considered a “skipped exon.” Although such instances could instead be interpreted as independent alternate 5' and 3' splice sites, skipped exon interpretation is consistent not only with our own independent experimental validations, but also frequently with gene models. In an “antisense conflict,” two junctions conflict with boundaries on opposite strands (Figure 1D). However, “antisense junctions” must have at least one boundary antisense to a gene model.

Additional filters applied to noncanonical junctions:

If two canonical boundaries matching a previously known junction or a

documented novel junction could be reached by adjusting the breakpoint of the junction up to 15bp in either direction, it was filtered out. The coverage filter eliminated junctions where coverage exceeded 1000 reads/bp within 100bp of the junction's outer edges (Figure 2). Parameters were empirically determined using known junctions as true positives and noncanonical junctions eliminated in the previous filter as false positives. The selected parameters retain more than 95% of true junctions while eliminating more than 50% of false positives.

Validation of conflicting splicing events:

A biologically independent 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, 1.5 µg of total RNA was reverse transcribed at 42°C for 1.5 hours using 1.9 µM random hexamer with Superscript III (Invitrogen Corp., Carlsbad, CA). For each validation, 1 µL 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 Table 4). PCR conditions were 95°C for 2 min, followed by 20 cycles of 95°C for 30 s, 52°C for 45 s, and 65°C for 3 min. 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 1-5 U of the appropriate enzyme (Table 4). Digest conditions were as recommended by the individual manufacturers and all digests were allowed to proceed

for 1 hour before being 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 (Table 4). 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 and then aligned to the *Plasmodium falciparum* genome (v6.3) using BLAT⁽³⁴⁾.

Gene Ontology (GO) analysis:

For alternative splicing, all 254 genes with evidence of alternative splicing from HMMSplicer junctions scoring ≥ 1075 were analyzed with DAVID and Gostat^(35, 36). Parameters were minimal length of considered GO paths = 5, maximal *P* value = 0.01, cluster GOs = -1, and correct for multiple testing = none. Both Benjamini and Bonferroni false discovery rate statistics were applied. The same analysis was done for the 148 genes with mapped antisense junctions.

Calculating minimum and maximum number of isoforms:

The number of isoforms for each conflict group was counted as either 2, for exon skip groups, or the total number of alternate edges for groups with alternate edges. If a

conflict group contained both exon skip and alternative edges, the number of isoforms was determined manually. Across a whole gene, the minimum number of isoforms was the number of isoforms in the group with the most isoforms, and the maximum number of isoforms was the product of the number of isoforms for every conflict group in the gene.

Determining intron splicing efficiency:

To measure splicing efficiency, we calculated the ratio of the number of exon-exon junction reads (EE) to the average number of exon-intron unspliced reads (EI) on either side of the intron in question⁽³⁷⁾. EI for each boundary was the number of reads aligned by Bowtie that spanned the exon-intron boundary with at least 6 bp on either side of the boundary. The EE:EI ratio was measured for every junction, then the results were filtered to remove junctions with low coverage or other confounding factors. Junctions in genes with any antisense were removed as the antisense transcript could alter the EI values. Only PlasmoDB v6.3 junctions were maintained, as alternatively spliced junctions could have EI values affected by the canonical transcript levels. For coverage, the junctions were required to have an average of at least 16 exon-intron reads across the two junction edges. In addition, the two EI values were required to be relatively similar by requiring a z-score, derived from an assumption of a Poisson distribution, of less than 2.

Determining if antisense junctions result from overlapping genes:

Genes were determined to overlap if an EST or a high scoring junction from the combined data mapped to neighboring genes simultaneously. Antisense junctions in such

overlapping genes were determined to be antisense as result of the overlap if the direction of the overlap predicted the gene that contained the antisense junction.

Results:

Plasmodium falciparum contains specific orthologs to splicing factors:

In *P. falciparum*, RNA components of the major U2-type spliceosome have been detected^(8, 7), yet protein components have not been systematically identified. Using reciprocal best hits (RBH) analysis⁽²⁹⁾ of human and yeast splicing factors, we identified putative homologs to spliceosome and spliceosome-associated protein components (Table 1)⁽³⁸⁻⁴⁰⁾, the majority of which were most similar to their human counterparts. However, homologs of three components of the human spliceosome could not be identified: SFY2, PPIE, and PRP2. PRP2, a DEAH/D-box ATPase, is ostensibly the most critical of the three, as it is thought to induce a structural rearrangement that results in dissociation of the SF3a and b complexes from the branchpoint, rendering the branchpoint competent for nucleophilic attack of the 5' splice site⁽⁴¹⁾. Interestingly, initial analysis returned PF10_0294 as the closest match in *P. falciparum* for both human PRP2 and PRP22, though the reciprocal BLAST completing RBH analysis returned PRP22 as a slightly better match for PF10_0294 within the human genome (Table 1). PRP2 and PRP22 are both DEAH/D-box proteins involved in splicing with a high degree of conservation between their helicase and C-terminal domains. In *S. cerevisiae* and other related yeast, PRP2 proteins contain a conserved DC amino acid doublet in their C-terminal domain that is distinguishes them from other closely related DEAH/D-box ATPases, such as PRP22⁽⁴²⁾. Although RBH analysis points to PF10_0294 as a PRP22 homolog,

alignment of the C-terminal portion of PF10_0294 reveals the presence of the DC doublet signature of PRP2 homologs in yeast (Figure 2). Without biochemical characterization, it is difficult to determine which role PF10_0294 might play, and it is possible that it encompasses the activity of both DEAH/D-box ATPases. Thus, while our RBH analysis is helpful as a first step in determining players involved in splicing, careful experimental verification of the exact roles of these putative homologs is still required to fully understand how splicing occurs in *P. falciparum*.

In other eukaryotes, alternative splicing is guided by the presence or absence of proteins that determine which splice sites are available to the spliceosome⁽⁴³⁾. To determine if *Plasmodium falciparum* has homologs to such proteins, human arginine/serine-rich (SR) and heterogeneous nuclear ribonucleoproteins (hnRNP) proteins with documented roles in alternative splicing were used for best reciprocal hits analysis^(44, 45). Four SR proteins and one hnRNP protein returned specific homologs (Table 1). These homologs likely represent only a fraction of the proteins that influence splice site selection in *P. falciparum*, as at least 71 additional proteins contain either an RNA recognition motif (RRM) or an RNA binding domain (RBD) according to InterPro⁽⁴⁶⁾, and 7 contain an RS domain according to our own analysis. Many proteins involved in splice site selection during alternative splicing utilize one or more of these domains, although they do not guarantee involvement in splicing^(44, 45). Together these data suggest that alternative splicing could play an important role in *P. falciparum*.

Overview of Plasmodium falciparum RNA-Seq datasets:

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

We aggregated these data and after preliminary filtering and sequence collapsing, ran two analyses in parallel: a Bowtie/BLAT^(33, 34) pipeline to align ungapped reads back to the *P. falciparum* genome (PlasmoDBv6.3) and HMMSplicer v0.7.0⁽²⁸⁾ to detect and score exon-exon splice junctions. The Bowtie/BLAT pipeline was able to align between 84 and 194 million bases of sequence to the *P. falciparum* genome for each independent timepoint (Table 3) for a total of over 1.5 billion aligned bases. Discounting antigenic variation gene families (272 *vars*, *rifins*, and *stevors*), each exonic nucleotide of a gene was covered by a median of 59 reads. HMMSplicer was also run on the dataset with a minimum intron size of 5 bp and a maximum intron size of 1000 bp, covering 99.6% of

all annotated *P. falciparum* introns. More than 1.9 million reads in the combined dataset were mapped to junctions (Table 5).

To gauge the quality of the junctions predicted from the combined dataset, we examined the distribution of HMMSplicer scores calculated for all predicted 5'GU-AG 3' junctions (Figure 4A). HMMSplicer scores reflect both the strength of the junction alignment and the cumulative support for that junction within the dataset⁽²⁸⁾. The distribution of HMMSplicer scores for canonical *P. falciparum* junctions is clearly bimodal, perhaps indicating predictions of differing reliability. In this organism, PlasmoDB gene models and ESTs provide a set of previously known splice junctions that are likely to be valid^(12, 32, 47, 48) and the distribution of HMMSplicer scores for only those junctions with boundaries matching previously known junctions was found to primarily fall within the higher scoring population. Therefore, an HMMSplicer score of 1075, representing the natural breakpoint in the bimodal distribution, was chosen as an operational threshold for subsequent analysis (Figure 4A). Below this threshold, support for detected junctions decreases rapidly, and thus the false positive rate among these lower-confidence junctions is likely to be higher. However, 13% of all known junctions detected within the combined dataset fall below our threshold, indicating the presence of valid junctions with non-ideal coverage, though only 0.1% score below 600 (Figure 4A). While we have enacted an operational threshold, all HMMSplicer junctions regardless of score are accessible for additional analyses (Supplementary Files 1 and 2, upload at <http://plasmodb.org/cgi-bin/gbrowse/plasmodb/>).

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

HMMSplicer found 7,655 5' GU-AG 3' junctions above the operational threshold within the combined RNA-Seq data. More than 88% were supported by reads from both timecourses. Of these high scoring junctions, 6,678 (87.2%) confirm introns in PlasmoDB gene models or ESTs (Supplementary File 3, upload at <http://plasmodb.org/cgi-bin/gbrowse/plasmodb/>). 977 (12.8%) support new introns, an increase of 11% over the current genome annotation (Figure 4B, Supplementary File 4, upload at <http://plasmodb.org/cgi-bin/gbrowse/plasmodb/>). 431 (43.9%) of these new junctions fall either totally or partially outside of annotated gene models, suggesting splicing in unannotated untranslated regions (UTRs) or in unannotated genes, whereas 544 (55.4%) align within gene models. As discussed below, many of the new junctions discovered within gene models represent alternative transcript isoforms or splicing of antisense transcripts. Unexpectedly, 2 (0.2%) new junctions map to neighboring genes encoded on opposite strands, suggesting unannotated overlap between these gene pairs.

We sought to lend support to the new 5' GU-AG 3' junctions detected by HMMSplicer by calculating WebLogos⁽⁴⁹⁾ from their 5' and 3' splice sites. If these new junctions represent true splicing events, they would be predicted to recapitulate the nucleotide preferences found within 5' and 3' splice sites of known 5' GU-AG 3' junctions. Indeed, no significant differences were observed between our calculated sequence logos for PlasmoDB/ EST matching junctions versus new junctions, and both sets of logos closely matched previously published results (Figure 4C)⁽⁸⁾. In contrast, logos produced from the bottom 10% of new junctions below the operational threshold contained little information other than their 5' GU-AG 3' boundaries (Figure 6A). Efforts to determine a branchpoint motif from the introns defined by our high-scoring canonical

junctions yielded no convincing results, similar to Chakrabarti et al.'s efforts to determine a branchpoint motif from a smaller set of EST introns ⁽⁸⁾.

The validity of new 5' GU-AG 3' junctions was also independently assessed by experimental validation using an biological replicate of our original blood stage timecourse and the strategy described in Figure 5A. Twelve 5' alternate splice sites, two 3' alternate splice sites, seventeen skipped exons, and ten spliced antisense transcripts were tested (Table 6, Figure 1 C, D). 19/21 (90.5%) new splicing events ranging in score from 1189.3 to 1544.2 were experimentally confirmed, including a skipped exon in MAL13P1.159 (thioredoxin) and splicing of an antisense transcript mapping to PFF0290w (long chain polyunsaturated fatty acid elongation enzyme) (Figure 5B, C). 13/20 (65%) events below the operational threshold with scores ranging from 984.6 to 1050.5 were also confirmed. Since more than half of these lower scoring events were successfully verified, these validations also confirm that our threshold is conservative - in addition to excluding false positive junctions, it also excludes some true splicing events, such as the 3' alternate splice site in PFB0279w (conserved *Plasmodium* protein, Figure 5D). Overall, these results indicate that a high percentage of new junctions both above and below the threshold are genuine, although independent confirmation may be required for lower scoring junctions.

Our results suggest both that a number of true positive junctions exist below our operational threshold and that the nucleotide preferences present at the 5' and 3' splice sites of known junctions do not hold for the lowest scoring, least reliable junctions in the dataset. Therefore, to attempt recovery of true 5' GU-AG 3' junctions below our operational threshold, an orthogonal score based on position specific scoring matrices ⁽⁵⁰⁾

of the splice site logos was evaluated. Although this type of motif scoring ultimately lacked sufficient information for large-scale computational rescue (Figure 6B), it could potentially be used to prioritize experimental assessment of new junctions (Supplementary File 5).

Inspection of noncanonical splice junctions reveals new 5' GC-AG 3' junctions:

In many eukaryotes, splicing occasionally occurs at non 5' GU-AG 3' boundaries, sometimes via the major U2-type spliceosome as with 5' GC-AG 3' introns⁽⁵¹⁾, or via the minor U12-type spliceosome as with 5' AT-AC 3' introns⁽⁵²⁾, or spliceosome-independently as with the 5' CA-AG 3' intron in yeast HAC1⁽⁵³⁾. The presence of 5' GC-AG 3' junctions in *P. falciparum* ESTs and gene models^(12, 13) suggests that the parasite uses 5' GC noncanonical splice sites, yet this likelihood has never been examined in detail. Of the 984 noncanonical junctions above our operational threshold (Supplementary File 6, upload at <http://plasmodb.org/cgi-bin/gbrowse/plasmodb/>), 12 map to 5' GC-AG 3' boundaries (Table 7). Of these, seven were supported by either EST evidence or annotated PlasmoDB gene models, and five were completely new. We used WebLogo v3.0 to construct 5' and 3' splice site sequence logos from all 12 5' GC-AG 3' junctions⁽⁴⁹⁾ (Figure 7). The 3' splice site logo was very similar to the canonical 3' splice site logo. However, several clear differences distinguished the 5' splice site logo for 5' GC-AG 3' junctions from that of canonical junctions. Whereas canonical *P. falciparum* 5' splice sites have a slight preference for AG as the last two bases of the 5' exon, all 12 5' GC-AG 3' examples contain AG in these positions, and all 12 also contain an A at the third position of the intron. These same three nucleotides are also present in the two

PlasmoDB 5' GC-AG 3' junctions with HMMSplicer scores below our operational threshold. Both the 4th and 5th positions of 5' GC-AG 3' introns also appear to have strong, although not absolute, nucleotide preferences. Though stronger contextual sequence may simply reflect the small number of input sequences, stronger consensus 5' splice site motifs have been documented for 5' GC-AG 3' introns in other organisms as well ⁽¹¹⁾. As with 5' GU-AG 3' introns, efforts to determine a branchpoint motif from these introns failed to produce any convincing results.

We also considered the possibility that the parasite might employ splice sites other than 5' GU-AG 3' and 5' GC-AG 3'. However, preliminary manual inspection of the remaining noncanonical junctions revealed that many of them were likely to be false positives caused by read errors. Polymerase slipping, template switching, and single basepair substitutions are well-documented phenomena ⁽⁵⁴⁻⁵⁶⁾ that can occur during both the reverse transcription and PCR steps of library preparation. These upstream errors have no associated cost in sequence quality, and therefore may explain the origins of high scoring, erroneous junction reads. Since the probability of an erroneous read mapping to noncanonical boundaries is much greater than the probability of it mapping to canonical boundaries, it is not surprising that the false positive rate within the noncanonical junctions is higher than within the canonical junctions.

Two additional filters designed to eliminate false positive junctions while retaining any potential true noncanonical junctions were applied to the noncanonical junctions. Since HMMSplicer is more sensitive to errors the closer they are to the true breakpoint of a junction read ⁽²⁸⁾, the first filter eliminated noncanonical junctions with single base substitutions within 15 bp of either inner edge that caused miscalling of the

junction breakpoint (343 of 972 junctions). The second filter removed noncanonical junctions in very highly covered regions since the probability of error creation during preparation and sequencing increases as the copy number for a given sequence increases (356 of 629 remaining junctions). As an internal check, neither filter eliminated any of the 12 5' GC-AG 3' junctions previously identified.

Manual inspection of the remaining 273 noncanonical junctions yielded no additional, credible noncanonical splice junctions. Although 5' AT-AC 3' splice sites have been observed in introns excised by the U12 minor spliceosome⁽⁵²⁾, failure to detect any in the RNA-Seq data is consistent with our failure to find *P. falciparum* homologs to proteins specific to the human U12-type spliceosome⁽⁵⁷⁾. 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⁽⁵⁸⁾. Together, these results indicate that *P. falciparum* is unlikely to possess a minor U12-type spliceosome.

Genome-wide characterization of alternative splicing:

Alternative splicing increases transcriptome complexity by generating multiple isoforms from the same precursor that differ in single 5' or 3' splice sites or in whole exons and introns. To find alternative splicing within the combined dataset in an unbiased manner, independent of gene models, high scoring canonical and 5' GC-AG 3' junctions were compared to each other in a pair wise manner. To be considered “conflicting junctions”, one of the inner edges of a junction must have aligned within the intronic area of the other junction (Figure 1B). Since direct counting of these

occurrences would over-inflate the number of alternative splicing events (for example, a single skipped exon event would count as two pair wise conflicts), conflicts were further aggregated into junction groups (Figure 1B), which were then divided by strand orientation where applicable (Supplementary File 7). In total, 196 (48.3%) alternate 5' splice sites, 145 (35.7%) alternate 3' splice sites, 8 (2.0%) mutually exclusive alternate 5' and 3' splice sites, and 56 (13.8%) skipped exons were tallied (Figure 8A). The majority of alternative splicing events occurred in gene models in the sense direction, though some also occurred outside of gene models. These intergenic events most likely indicate alternative splicing in unannotated *P. falciparum* UTRs or in unannotated genes. Interestingly, all four types of alternative splicing were also seen in antisense junction groups. Further analysis of antisense splicing events is discussed in the next section.

Because combined RNA-Seq data are comprised of short reads rather than full length mRNAs, the collection of splice junctions that compose a given isoform is difficult to resolve, and thus the exact number of isoforms encoded by the alternative splicing events described here could not be determined. However, transcriptome-wide, the combined 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 (Supplementary File 7). Alternative splicing events for most genes maximally support between 2 and 4 isoforms. However, a handful of genes (PF14_0338 (conserved *Plasmodium* protein), PFF0630c (conserved *Plasmodium* protein), PFL1440c (conserved *Plasmodium* protein), PFC0495w (plasmepsin VI), and PFC0912w (signal peptidase)) could encode up to 8-16 different isoforms. In addition to supporting up to 8 sense isoforms, an overlapping antisense junction was also validated for PFC0495w

(plasmepsin VI), making it particularly interesting (Table 6). Gene ontology (GO) analysis of alternatively spliced genes did not reveal any functional patterns.

The transcriptome complexity afforded by alternative splicing often increases the number of distinct proteins encoded by an organism. Of the 310 *P. falciparum* alternative splicing events mapped to gene models in the sense direction, 10% are predicted to produce altered UTRs, while the remaining 279 (90%) are predicted to produce distinct coding sequences. Of these, close to one third maintain coding frame, either adding or removing amino acids from the predicted protein (Figure 8B). In contrast, the majority of alternative splicing events result in frameshifts, most of which introduce premature termination codons within the gene model's predicted coding sequence.

One explanation for the abundance of protein truncating alternative splicing events in *P. falciparum* is that these transcripts may not be translated, but instead could be intermediates bound for nonsense-mediated decay (NMD). Regulated splicing controlling the ratio of NMD-targeted to protein-coding isoform produced from certain genes is a mechanism of post-transcriptional regulation in other organisms^(43, 59). However, NMD has not been shown to exist in the parasite. Using human and yeast sequences for the core conserved NMD surveillance proteins, UPF1, UPF2, and UPF3 (paralogs UPF3a and UPF3b in humans)⁽⁶⁰⁾, best reciprocal hits analysis was able to find homologs to all three in *Plasmodium falciparum*, suggesting the NMD pathway exists in this parasite (Table 1). While it is unclear what the trigger for NMD may be in *P. falciparum*, 119 (73%) of the 162 truncating events do so more than 50 bp upstream of the last splice junction, rendering them eligible for NMD in mammalian systems⁽⁶⁰⁾.

Regardless, our results suggest that the majority of alternative splicing events in the blood stages of *P. falciparum* either produce truncated protein isoforms or tune gene expression post-transcriptionally.

We also looked at the relative abundance of alternate junctions in comparison to their recovered gene model counterparts (Figure 8C). Many occurred at < 10% of the frequency of the conflicting gene model junction within the combined datasets, and may correspond to isoforms either targeted for nonsense-mediated decay or of minimal use in the blood stages. Interestingly, 33 alternative junctions occurred at $\geq 100\%$ of the frequency of their conflicting gene model counterparts, indicating that the gene model isoform of the transcript may not be the dominant blood stage isoform (Table 8).

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

Previous reports of alternative splicing in *P. falciparum* have noted instances of transcripts with retained introns⁽¹²⁾, and regulated splicing efficiency can control such important biology as onset of meiosis in *S. cerevisiae*⁽⁶¹⁾. Therefore, to gauge general splicing efficiency as well as to discover poorly spliced outlier introns, we calculated the ratio of junction reads to the average number of reads covering both cognate exon-intron borders⁽³⁷⁾. Only recovered gene model junctions in genes without mapped antisense junctions were considered to avoid complicating factors. Because the datasets analyzed here were not generated specifically for the purpose of analyzing splicing efficiency, calculations could be made for only a subset of splice junctions in which the read counts covering both exon-intron borders were relatively similar. For the 779 introns analyzed, junction reads were recovered a median of 5 times more often than exon-intron reads

(Supplementary File 8). However, 44 (5.6%) analyzed introns appear to be very poorly spliced in the blood stages as they are retained in at least 50% of the transcripts sampled here.

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

Although gene models were not consulted during detection of junctions or alternative splicing, we assessed how thoroughly they were encompassed by our results. Of the 8,435 predicted splice junctions in PlasmoDB v6.3 gene models, 1,103 were not observed in the combined dataset, even below our operational threshold. Gene models with unrecovered splice junctions had a median coverage of 6 reads per coding nucleotide, indicating that in general, these genes were not substantially expressed during the blood stages. However, for 50 unrecovered known junctions, new junctions above the operational threshold were observed that did not match the boundaries indicated by the gene model (Table 9). Although it is possible that the gene model isoforms are not expressed in the blood stages in these cases and that the new junctions represent blood stage-specific alternate isoforms, it is more likely that the corresponding gene models are incorrect.

Genome-wide characterization of antisense splicing:

While probing for conflicting junctions, we noticed a class of conflicts in which one junction contained intron boundaries on a given strand while the other mapped to intron boundaries on the opposite strand (Figure 1D). 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 ⁽⁶²⁾. In addition, antisense transcription has been previously noted in *P. falciparum* ⁽⁶³⁾, and ESTs antisense to gene models have also been documented ⁽¹²⁾. Therefore, it is likely that these “antisense conflicts” derive from overlap of two spliced transcripts transcribed in opposite directions.

To expand on the initial discovery of antisense conflicts, we searched for all high scoring junctions with at least one intron boundary antisense to an annotated gene model. This analysis differed from the conflicting junctions analysis in two important ways. First, it incorporated antisense junctions that do not conflict with any sense introns (Figure 1D). Second, it excluded antisense conflicts between junctions in which neither mapped to a gene model (4 independent conflicts), and thus neither could be deemed “sense” or “antisense”. In total, this list contains 200 antisense junctions mapping to 149 gene models (Supplementary File 9). In addition, antisense junctions overlapping 16 of these genes appear to undergo alternative splicing to produce between 38 and 59 different isoforms (example shown in Figure 9A, Supplementary File 7). Weblogos of the 5' and 3' splice sites of antisense junctions revealed no significant differences compared to known junctions (Figure 9B), suggesting that antisense junctions arise from the same mechanism as other splice junctions in the transcriptome. No GO terms were significantly enriched within genes with mapped antisense junctions.

Antisense junctions could derive from either overlap between neighboring gene models encoded on opposite strands of the genome or from unannotated transcripts antisense to gene models. Indeed, 23 antisense junctions could be attributed to overlap between 15 pairs of neighboring annotated genes on opposite strands based on linking

junctions or ESTs (Table 10). Only 1 gene pair (PFE1425c/PFE1420w) is annotated as overlapping, while 9 had prior EST evidence of overlap. The remaining 5 pairs had no prior evidence of overlap. Twelve pairs were arranged in a tail-to-tail (overlapping 3' ends) fashion, while 3 were arranged in a head-to-head (overlapping 5' ends) fashion. Several studies have reported a bias toward tail-to-tail overlaps in mammalian genomes⁽⁶⁴⁾, although others refute this assertion⁽²³⁾.

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

Interestingly, over 86% of antisense junctions map to intron-containing genes, though only slightly more than half of genes in *Plasmodium falciparum* contain introns. This bias is significant, with a binomial probability of $\sim 3e^{-24}$. A similar bias was seen in

Arabidopsis thaliana in tail-to-tail overlapping transcripts ⁽²²⁾, and could be explained by preferential overlap between introns in antisense transcripts and introns in sense transcripts (Figure 1D). In some cases, multiple antisense introns overlap extensively with multiple sense gene introns in the same gene model, but not with more expansive exon regions (Figure 9A). The observed distribution of overlap with sense introns is highly statistically significant (p-value of Chi squared test < 0.001) when compared to the expected distribution from random re-placement of antisense junctions within their associated genes (Figure 9C). This expected distribution was calculated by first determining the probability of encountering a GU (5' splice site) or an AG (3' splice site) on the opposite strand of introns versus exons within the group of genes with mapped antisense junctions. These probabilities then guided otherwise random re-placement of each antisense junction within its corresponding gene model, keeping the original length of the antisense intron intact. After re-placement, the distribution of overlaps for simulated antisense introns with sense introns was tallied. This re-placement was iterated 100 times, with the mean percent of nucleotide overlap with sense introns +/- standard deviation shown. Thus antisense introns appear to not only overlap intron-containing sense genes significantly more often than expected, but also overlap the intron portions of sense genes significantly more than expected.

Discussion:

Completion and preliminary annotation of the *Plasmodium falciparum* genome in 2002 facilitated a series of large-scale experiments designed to illuminate the parasite's biology on a genomic-, transcriptomic-, or proteomic-wide level. In pursuit of a thorough

understanding of *P. falciparum* blood stage genetic regulation, steady state gene expression experiments captured its unique, cascade-like transcriptome^(1,2). Subsequent genome-wide RNA decay experiments revealed global rapid turn over of RNA in the early hours post-invasion, and then progressively longer transcript half-lives during the remainder of the blood stage cycle⁽⁶⁵⁾. The new splicing events described here reveal additional complexities within the transcriptome not captured by these previous studies, such as alternative splicing, gene overlap, and spliced antisense transcripts, and thus, the present study fits into a larger, more dynamic understanding of the transcriptome of *P. falciparum*.

Traditionally, full-length cDNA and EST data have been used for analysis of transcript structure and variants. EST collections in *P. falciparum* have indeed produced increasingly accurate gene models^(12, 13, 47, 48). However, no full-length cDNA sequences have been published for *P. falciparum*, and many gene models lack or are incompletely covered by ESTs. RNA-Seq provides the advantage of capturing an entire transcriptome at great depth, enabling detection of low copy number transcripts and variants. However, in its current form, RNA-Seq cannot capture a single transcript molecule from beginning to end. Despite this limitation, the orders-of-magnitude increase in throughput over EST libraries expanded the repertoire of splice junctions known in the parasite by more than 11% in the present study.

The ability to accurately and sensitively map junction reads from the RNA-Seq datasets proved crucial to our analysis. For this purpose, we used HMMSplicer, an algorithm we developed specifically to overcome the challenges presented by RNA-Seq data and the inherent biases within the *P. falciparum* genome⁽²⁸⁾. In contrast to previous

RNA-Seq studies in *P. falciparum* and other organisms^(14, 66), we relied only on alignment of junction reads within the genome to detect splice junctions, rather than depending on gene models or ungapped read coverage. Also, HMMSplicer does not use additional assumptions to filter its output junction set, instead scoring each splice junction on the strength of supporting reads. Because a low false positive rate was desired for accurate characterization of splicing in *P. falciparum*, we established an operational HMMSplicer score threshold based on the bimodal distribution of known versus new canonical splice junctions. However, setting this threshold held the disadvantage of excluding some known junctions, and therefore some true new junctions as well. Indeed, our biologically independent validation experiments demonstrated that even lower scoring junctions were more likely than not to represent true splicing events. Although these lower scoring junctions were excluded from downstream analysis in the present study, they remain accessible in the HMMSplicer results (Supplementary Files 1 and 2, upload at <http://plasmodb.org/cgi-bin/gbrowse/plasmodb/>).

We also did not rely on gene models during discovery of alternative splicing. This decision was prompted by several observations within the data. First, there were ambiguous instances in which a junction conflicted with a gene model, but the gene model junction was not recovered within the dataset. These instances could potentially represent gene model errors, making it inappropriate to classify them as alternative splicing without additional data. Conversely, areas of the transcriptome with no gene model contained multiple junctions that could not possibly exist within the same transcript (Figure 8A). These intergenic junction groups clearly exhibit alternative splicing and would have been missed by reliance on gene models. Thus our unbiased

approach allowed for more accurate and sensitive detection of alternative splicing events based solely on experimental observation of the conflicting junctions themselves.

Although *Plasmodium falciparum* ESTs and even some gene models include noncanonical 5' GC-AG 3' splice junctions, to our knowledge, no study has attempted to identify or characterize noncanonical junctions in *P. falciparum*. We found 12 high scoring 5' GC-AG 3' junctions within the noncanonical junctions results, 5 of which were new. As in other organisms, the 5' splice site for these junctions has a remarkably high information content compared to the 5' splice site for canonical *P. falciparum* junctions, perhaps indicating greater reliance on sequence context for recognition of 5' GC splice sites. In particular, the strong preference for G at the fifth position in 5' GC-AG 3' introns is interesting. Although G is strongly preferred at this position in human canonical introns⁽¹¹⁾, and mutation of this G to other bases reduces splicing fidelity in yeast⁽⁶⁷⁾, *P. falciparum* canonical introns have almost no base preference at this position (Figure 4C)⁽⁸⁾. At present, it is unclear how complete the list of 5' GC-AG 3' junctions is, given that the percent of splice junctions mapping to those splice sites (~0.1%) remains several fold lower in *P. falciparum* than in other organisms⁽¹¹⁾. In addition, although filters designed to aid discovery of any additional noncanonical junctions were implemented, manual inspection found no convincing examples. It is possible that despite efforts to limit bias, the filters inadvertently eliminated true positive junctions or manual inspection failed to detect credible non- 5' GU-AG 3' or 5' GC-AG 3' junctions within the data.

Our analysis uncovered not only constitutive and alternative splicing in *P. falciparum*, but also complex transcriptional arrangements in the parasite. Independent validation of new junctions antisense to sense junctions indicates that these are not

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

The larger impact of the transcriptome features revealed by the new canonical and 5' GC-AG 3' junctions captured here remains unknown. Consistent with reports correlating alternative splicing prevalence with organismal complexity⁽⁶⁸⁾, alternative splicing events do not appear to be widespread in *P. falciparum* blood stages, affecting 8.6% of intron-containing genes. Although relatively scarce, alternative splicing events in *P. falciparum* may expand important protein functionalities in the organism and may also contribute to crucial post-transcriptional gene regulation – however, it is possible that their impact on parasite biology is minimal. Interestingly, these events appear to occur with almost no pressure to preserve ORFs, as only 1/3 are predicted to do so, the same proportion expected by chance. We have suggested that alternative splicing events predicted to result in truncated ORFs may be linked to a nonsense-mediated decay

(NMD) system in the parasite as a means of gene regulation. It would be interesting to determine if such isoforms decay faster than their corresponding protein-coding isoforms, although current RNA decay data in *P. falciparum* does not allow for discrimination between the decay rates of isoforms of the same transcript ⁽⁶⁵⁾.

The observed overlap between sense and antisense transcript pairs of may also contribute to important gene regulation by a variety of mechanisms ⁽²⁴⁾. In addition, unannotated antisense transcripts could perform a variety of as-yet-unknown functions that may or may not be restricted to regulation of their sense partners. Unraveling these possibilities in both the symptomatic blood stages of *Plasmodium falciparum* as well as in the organism's larger lifecycle will provide an unprecedented understanding of a deadly human pathogen.

Acknowledgements:

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

Table 1 – Putative *Plasmodium falciparum* splicing and nonsense-mediated decay factor homologs 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* homolog. Spliceosomal and NMD factors not found are in red and their homologs are denoted with question marks, while SR and hnRNP factors not found are not shown. *S. cerevisiae* homologs that do not reside in the same complex as their human counterparts are italicized. * Homologs 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.

| Complex | Human / Yeast | <i>Pf</i> Homolog | Complex | Human / Yeast | <i>Pf</i> Homolog |
|---|-----------------|------------------------|--|-----------------------|--------------------------|
| 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 / PRP6 | PF11_0108 | | SFRS4 / - | PF10_0217* |
| | SNRNP40 / - | MAL8P1.43* | | TRA2B / - | PF10_0028 ^{†*} |

Table 2 - Primer sequences for initial library preparation and the Long March.

| Adapter | Primer | Primer Sequence |
|----------------|-------------------------------|---|
| | 6bp-EciI-N ₉ | 5'-GACGCTGGCGGANNNNNNNNN-3' |
| | 13bp-ModSolS-N ₉ | 5'-GCTCTGCCGCTCTNNNNNNNNN-3' |
| | biotin-short-ModSolS | 5'-/5Biosg/GGCATACGAGCTCTGCCGCTCT-3' |
| | 6bp-EciI | 5'-GACGCTGGCGGA-3' |
| | biotin-shortMod-PE-SolS | 5'-/5Biosg/TGCTGAACCGCTCTGCCGCTCT-3' |
| | fullMod-PE-SolS | 5'- CAAGCAGAAGACGGCATAACGAGATCGGTCTCG GCATTCTGCTGAACCGCTCTGCCGCTCT-3' |
| | fullModSolS | 5'- CAAGCAGAAGACGGCATAACGAGGCATACGAG CTCTGCCGCTCT-3' |
| | Sol primer 1 | 5'- AATGATACGGCGACCACCGACTCTTCCCTA CACGACGCTCTTCTGGAG-3' |
| | Sol-SeqPrimer | 5'- CACTCTTCCCTACACGACGCTCTTCTGGAG-3' |
| | PE-SolS-SeqPrimer | 5'- CGGTCTCGGCATTCTGCTGAACCGCTCTGCCG CTCT-3' |
| Sol-L-NN | 24bp-SolL-GsuI-NN | 5'-CCCTACACGACGCTCTTCTGGAGNN-3' |
| | P-recomp24bpSolL-GsuI-6Camino | 5'-/5Phos/CTCCAGGAAGAGCGTCGTGTAGGG /3AmM /-3' |
| Sol-L-AC-NN | short-SolL-GsuI-ACNN | 5'-CACGACGCTCTTCTGGAGACNN-3' |
| | Sol-Adapter-L-short-phos-AC | 5'-/5Phos/GTCTCCAGGAAGAGCGTCGTG/3AmM/- 3' |

Table 3 – Number of total bases aligned to *Plasmodium falciparum* genome, version 6.3 for each independent timepoint, as well as the read type (single or paired-end) and length.

| | <u>Timepoint</u> | <u>Number of Aligned Bases</u> | <u>Read Type and Length</u> |
|--|-------------------------|---------------------------------------|------------------------------------|
| <u>Present</u> <u>Time-</u> <u>course</u> | TP1 | 125,768,514 | Single end, 42bp or 60bp |
| | TP2 | 86,641,641 | Single end, 42bp or 60bp |
| | TP3 | 103,049,333 | Single end, 42bp |
| | TP4 | 136,426,689 | Single end, 44bp or 46bp |
| | TP0 | 194,190,837 | Paired-end, 54bp |
| | TP8 | 186,072,654 | Paired-end, 54bp |
| <u>Otto et al.</u> <u>Timecourse</u> | TP16 | 151,574,541 | Paired-end, 54bp |
| | TP24 | 187,630,935 | Paired-end, 54bp |
| | TP32 | 84,330,852 | Paired-end, 54bp |
| | TP40 | 117,132,306 | Paired-end, 54bp |
| | TP48 | 130,135,714 | Paired-end, 54bp |

Table 4 – Outer primers, restriction enzymes (RE), and inner primers used for each conflicting splice event validation. All restriction enzymes were obtained from New England Biolabs (Ipswich, MA) except MaeIII (Roche Applied Science, Indianapolis, IN) and AgsI (SibEnzyme, West Roxbury, MA).

| <u>Gene</u> | <u>Outer Primers</u> | <u>RE</u> | <u>Inner Primers</u> |
|-----------------|---|------------------|---|
| PFL1810w | CCGGTTCGTCTTCTCCATA AGAGCGTCTGTATGGCTGTG | NdeI | GGTTCGTCTTCTCCATACCC GGCTTATGAGAACAGAAATGCAG |
| PFE0390w | GCAGTAAGCCATGTAAAAACAGC TGCAGGTAATATTTCGGCAAA | PsiI | CAGTAAGCCATGTAAAAACAGCA TAGTCGTGTGGTCAGGTTTCG |
| PF13_0138 | TCAGCCTTTTGGTTAAAATATCC AACAAAAGGGTTGAGCGTATAA | DraI | GCCTTTTGGTTAAAATATCCCAAT GGCTAGTCCAGAACCATCCA |
| PFI0400c | CTCTTTTCTTTGCTCGGTTGA GAGCAAAGGACTTGAAGAAA | PsiI | AGGTATAAATAATCAATATGGACGGGAC A AGAGCAAAGGACTTGAAGAAA |
| PFF0290w | TGGAACCAGTTTACCTTCCTACA ATACGTTACGCATGCCTTTG | BsrGI | GGAACCAGTTTACCTTCCTACA GGTACTTTGAATTTTACGCTTG |
| MAL13P1.2 25 | TGAGACAAAAGTGATCTTTTCTTGC GGTGTTAATATGAAAAATACAGTCATTG | BsrGI | GCAAGTGATTCATTTTGTCTAGG GGTGTTAATATGAAAAATACAGTCATTG |
| PFE0055c | TTCGTTCCCTCACTCCATCC TTCCATCTTTCAAGGTTTCG | EcoRI | GACCCACAAACCTTTTCTTG TTCCATCTTTCAAGGTTTCG |
| MAL8P1.12 6 | GGAATATTGTGAGCATTGGTT CAAAATGGCCACAAAGAAGAA | DraI | CATTGGTTATAATTAGATGACCCTCA GATATCCGAGGGGAAGAAGG |
| PF10_0025 | TCTTGATCACTACAACCCATTG GAAGAATACGAATGGGGGAAA | MaeIII -Roche | TCAAATTCTACAGCAGGTTTCG GAAGAATACGAATGGGGGAAA |
| PFD1050w | TGTTTTGCCTTGAACATGGA TGCGAAAATTATTGCTGCAT | DraIII | TTTTGCCTTGAACATGGAAT AGGGTGAAACAGCTGACGAT |
| MAL13P1.1 59 | CACATGGGTCATGCAGAAAA GCCTTCTTACCACCGATAA | SspI | TCCTATGATAACAGATTGTCTCCTG GCCTTCTTACCACCGATAA |
| PFC0780w | TCCAAGTTTAAATATAATGATGAGGA GCAAAACTGGAAAAGGGAATC | PsiI | CCAAGTTTAAATATAATGATGAGGATTC CCTCACTCGAGGCACATAAA |
| PFD0775c | GTCGGCTCCTAACACATTCC CCCAAGACAGCATCCAAAAT | HinfI | CCTAACACATTCCCCACACA TGGGTTTCCGAATTATCCAA |
| PF10_0194 | TATGGCCTTTTTACAGCAC GGCACGGTATACGAAGGGTA | SspI | GGCCTTTTTACAGCACTTC TTTGGAGATAGCAAAATCGTTCA |
| PFL1440c | ACATTTATCATACTTCACACGTATT TGTC AAGCTTTCTCAGAAGATACC | MaeIII -Roche | CGAATATTATTTTCTTTCGGGTA TGTC AAGCTTTCTCAGAAGATACCTA |
| PF11_0291 | TGGAAGGATATAAAGGATGCTCA CGGATATGCAACCATTGAAA | HinfI | AAAGGATGCTCATAACTTCTGGA TCCTTCATGGGATTTTCCAA |
| PFC0360w | TTATGTTCCCCCTGAGGTTTT GTCCATCCATTTTGCACCT | PsiI | TTTAACGCTTTTACCGATGC GTCCATCCATTTTGCACCT |
| PFC0495w | AACGTCGAAGGGAAGGAGAT GCTCAATTAGACCTATGAGTTTATGA | EcoNI | ACGTCGAAGGGAAGGAGATT TCATTATCAAAAATGGTGTAATATTTTCTT |
| PF14_0394 | TTCTCGTTAGTCAGTGGCTCAT GGAAATTAACGATTTTATTTTATCGAA | HinfI | TCTCGTTAGTCAGTGGCTCAT CACGCAACTTATAAAAATAGCAAAAA |
| MAL13P1.1 46 | TCAAATTGTTGGATGGGACA ACCGATTTACCACAATGAG | NcoI | GAGGGGAGCTACCAACACCT ACCGATTTACCACAATGAG |
| PF11_0379 | CGAAAGTGAAAGCAGTGAGG GAGGTTTCCATTGAAAATTGCT | MaeIII -Roche | TGAAGATGAAAGAACAGTCCA GAGGTTTCCATTGAAAATTGCT |

| | | | |
|-----------------|---|-------------------------|---|
| PFL1445w | ATGACCCCCGAAATTTATGC GAAACAGGGGTCTGTCGTTC | AgsI - sibEnz yme | CCGAAATTTATGCGTCCTCA AGATACTTTCCAAAAGCCCATA |
| MAL13P1.1 6 | CAGGTTTAAATTCATCCAGTG GCAGAATATACCGAATATGGAGGA | PsiI | AGGTTTAAATTCATCCAGTGA CCACCACATACATCCAGGAA |
| MAL13P1.2 77 | GGCCAAGGTTTTACAAACGA TTAATATAGGCTTTTCTTTAAAATGACTT | PsiI | GTGATTTTCATCGTGAATTTTATG TTAATATAGGCTTTTCTTTAAAATGACTT |
| PFF1210w | TTCATTCAACGATGAAAAGTACAAA AGCACAAACCAAACGCCTTA | DdeI | TCATTCAACGATGAAAAGTACAAA TTTTCTTGTCTTCAAATAATGG |
| PFB0600c | TCTGGACAAATGTGAAGGTGA TGAGAGGACCTTTACCAACAGA | HinI | TCCCAGTAGGAAGATTGTAAAGGA GTGAGAGGACCTTTACCAACAGA |
| PF14_0128 | ATTGAAAACCCCTTCGAGAGC CCGTAGGTAAGTGGGCAGAA | DdeI | TTGAAAACCCCTTCGAGAGCTA TTGCAGTGCCTGTTTCAAAG |
| PF14_0316 | TTGCTGGGTGTTCTTTTTCC AAATTCAAAGTAGCTAGAAAACAAGG | Scal | GCTGGGTGTTCTTTTTCTCTG TCTAGTGATGAAGAATCTGAAGGA |
| PFB0279w | TTTTGCTTATGATGCATTGGA TTTTTGAAACATTGGTCTCTTCA | SacI | TTTGCTTATGATGCATTGGA TTACTATAACGCTGAGAAGTTTTGA |
| PFL1465c | AACAATCCGCTGTAGCTCCT GCAAGTTGATAATTCCTCGTCA | MboI | TAGCTCCTGCGAAAACCCATT CCATATTATGTGTTAGGAAAAAATGAA |
| PF10_0372 | TGCCATGATAATATCGGCATC TACCTTGGGGTTTTCTGCTG | DdeI | CCATGATAATATCGGCATCCTT TCCTTTAATCCATATTTGCTGCT CGATTTAATGTGGATGATGCT CATATTTTATAGTTAGATATTTGTGAAGA CG |
| PF11_0182 | CCGATTTAATGTGGATGATGC CAAGAATATGTTTAAAGTTTGTCAGCA | DdeI | |
| PFF0365c | TTTCAATAATCCCCAATCACAA TGAAAACAAAGGAACCAGCA | SspI | AAGTTGATCCAATAACATAAGAACAAA TGAAAACAAAGGAACCAGCA |
| PFB0445c | GCCAACTCTCGAGTATGTGCT TGAAAGTGGTTTTGAGCATCC | EcoRV | CGAGTATGTGCTAAACCAAGACA AAGAAACTATTCGCCGAGCA |
| PFD0895c | TTCCAAAAATCGCTTAAAGG GGGATTCAACATAGGCACAAG | DdeI | CACAAGTAACTGTTCCACTTATTC GGGATTCAACATAGGCACAAA |
| PF10_0116 | GGCATTATGAAGACTAGCCAAAA GCTTCTAGCACATGCTTATGTATT | SspI | GCATTATGAAGACTAGCCAAAA TTTCCACTTGCAAAAAGAATC |
| PF14_0604 | CAAACATTTGGGACGAAAA GCAACATTTTCTTCGCTTTGA | EcoRI | AAACCATTTGGGACGAAAA AAATTGGGGAATCAAAGGAGA CCCAAGCAATTATCATCCAT TCAGAAAAATATAGAAAACGTCAAATTA A |
| PFI0560c | CCCCAAGCAATTATCATCCA CGACACACAAATAATGACGTG | MboI | |
| PFB0550w | ATGGAAGACCACGATGCTAA ATTGGCCTGAATGGTTCAAA | DdeI | TGGAAGACCACGATGCTAAT TGGACATTTTCTTTCTTTCC |
| PF11_0355 | AGGGGCTCTTAGCAAAATC ATTTAACTGGCCCCAGAAG | EcoRV | CAAAACAGTTTCAGAGGCAATTT ATTTAACTGGCCCCAGAAG |

Table 5 – Number of uncollapsed reads aligned as junctions by HMMSplicer by timepoint and category. Below each number is the percent of uncollapsed junction reads

for that timepoint that fall into that junction category.

| | Time- pt. | Known ≥ 1075 | Known < 1075 | Novel ≥ 1075 | Novel < 1075 | Noncan. ≥ 1075 | Noncan. < 1075 | Total |
|------------------------|--------------|-------------------|-----------------|-----------------|-----------------|-------------------|-------------------|--------|
| Present Timecourse | TP1 | 23990 (63.3%) | 170 (0.4%) | 905 (2.4%) | 1089 (2.9%) | 931 (2.5%) | 10807 (28.5%) | 37892 |
| | TP2 | 37674 (70.8%) | 172 (0.3%) | 996 (1.9%) | 846 (1.6%) | 1222 (2.3%) | 12291 (23.1%) | 53201 |
| | TP3 | 128409 (55.6%) | 324 (0.1%) | 2444 (1.1%) | 1207 (0.5%) | 24495 (10.6%) | 74235 (32.1%) | 231114 |
| | TP4 | 101509 (55.9%) | 319 (0.2%) | 1657 (0.9%) | 1480 (0.8%) | 6110 (3.4%) | 70671 (38.9%) | 181746 |
| Otto et al. Timecourse | TP0 | 114399 (50.8%) | 548 (0.2%) | 1996 (0.9%) | 2114 (0.9%) | 2186 (1.0%) | 103799 (46.1%) | 225042 |
| | TP8 | 98503 (44.6%) | 405 (0.2%) | 1639 (0.7%) | 2173 (1.0%) | 3501 (1.6%) | 114711 (51.9%) | 220932 |
| | TP16 | 83500 (40.0%) | 276 (0.1%) | 1074 (0.5%) | 2008 (1.0%) | 8491 (4.1%) | 113312 (54.3%) | 208661 |
| | TP24 | 139073 (62.6%) | 436 (0.2%) | 2534 (1.1%) | 1770 (0.8%) | 2397 (1.1%) | 75862 (34.2%) | 222072 |
| | TP32 | 74091 (34.0%) | 326 (0.1%) | 1346 (0.6%) | 2168 (1.0%) | 736 (0.3%) | 139290 (63.9%) | 217957 |
| | TP40 | 93374 (60.0%) | 373 (0.2%) | 1786 (1.1%) | 1396 (0.9%) | 1252 (0.8%) | 57454 (36.9%) | 155635 |
| | TP48 | 103913 (59.6%) | 444 (0.3%) | 1920 (1.1%) | 1413 (0.8%) | 1450 (0.8%) | 65354 (37.5%) | 174494 |

Table 6 – Verification of new junctions in conflict with known junctions. Conflicts are ranked by lowest HMMSplicer score within the pair, and the black line denotes the operating HMMSplicer threshold of 1075. * Indicates validations shown in more detail in Figure 2. For all conflict types except antisense, the new junction was evaluated for maintenance of open reading frame - nucleotide and amino acid (if applicable) differences between new and known isoforms are listed.

| <u>Gene Name</u> | <u>PlasmoDBv6.3 Description</u> | <u>Score</u> | <u>Validated</u> | <u>Type</u> | <u>Frame -shift?</u> | <u>Isoform Difference</u> |
|------------------|---|--------------|------------------|-------------|----------------------|---------------------------|
| PFL1810w | conserved Plasmodium protein | 1544.2 | Y | 5'ss | N | 132bp (44aa) |
| | | 1283.2 | Y | 5'ss | N | 219bp (73aa) |
| PFE0390w | conserved Plasmodium protein | 1422.1 | Y | 5'ss | N | 66bp (22aa) |
| PF13_0138 | MSF-1 like protein | 1372 | Y | 5'ss | Y | 56bp |
| PFI0400c | conserved Plasmodium membrane protein | 1369.8 | Y | exon skip | N | 126bp (42aa) |
| PFF0290w | long chain polyunsaturated fatty acid elongation enzyme | 1291.8 | Y | antisense | - | N/A |
| MAL13P1.225 | thioredoxin | 1277.3 | Y | exon skip | Y | 34bp |
| PFE0055c | heat shock protein | 1275.4 | Y | 5'ss | Y | 37bp |
| MAL8P1.126 | serine protease | 1257.4 | Y | 5'ss | Y | 110bp |
| PF10_0025 | PF70 protein | 1256.8 | Y | 5'ss | N | 75bp (25aa) |
| PFD1050w | alpha-tubulin II | 1243.2 | - | antisense | - | N/A |
| MAL13P1.159* | thioredoxin cleavage and polyadenylation specific factor | 1239.9 | Y | exon skip | N | 33bp (11aa) |
| PFC0780w | RNA binding protein | 1231.7 | - | antisense | - | N/A |
| PFD0775c | RNA binding protein | 1228.4 | Y | antisense | - | N/A |
| PF10_0194 | NOP12-like protein | 1219.4 | Y | exon skip | Y | 41bp |
| PFL1440c | conserved Plasmodium protein | 1217.6 | Y | exon skip | N | 57bp (19aa) |
| PF11_0291 | conserved Plasmodium protein activator of HSP90 ATPase homolog 1-like protein | 1203.5 | Y | 5'ss | Y | 40bp |
| PFC0360w | plasmepsin VI | 1200.5 | Y | exon skip | Y | 223bp |
| PFC0495w | plasmepsin VI | 1192.6 | Y | antisense | - | N/A |
| PF14_0394 | conserved Plasmodium protein | 1190 | Y | 5'ss | N | 99bp (33aa) |
| MAL13P1.146 | AMP deaminase | 1189.3 | Y | antisense | - | N/A |
| PF11_0379 | conserved Plasmodium protein | 1050.5 | Y | exon skip | N | 60bp (20aa) |
| PFL1445w | conserved Plasmodium protein | 1041.3 | Y | exon skip | Y | 85bp |
| MAL13P1.16 | SNARE protein | 1034.7 | Y | exon skip | N | 108bp (36aa) |
| MAL13P1.277 | DNAJ-like protein | 1034.2 | Y | exon skip | Y | 146bp |
| PFF1210w | phosphatidic acid phosphatase | 1032.4 | Y | 5'ss | Y | 67bp |
| PFB0600c | conserved Plasmodium protein | 1026.1 | Y | antisense | - | N/A |
| PF14_0128 | ubiquitin conjugating enzyme | 1018.5 | Y | exon skip | Y | 103bp |
| PF14_0316 | DNA topoisomerase II | 1011.4 | - | 5'ss | Y | 460bp |
| PFB0279w* | conserved Plasmodium protein | 1010.9 | Y | 3'ss | Y | 98bp |
| PFL1465c | heat shock protein hslv | 1004.5 | - | exon skip | Y | 39bp (13aa) |
| PF10_0372 | antigen UB05 | 1004.4 | - | antisense | - | N/A |
| PF11_0182 | conserved Plasmodium protein | 1004.1 | Y | exon skip | Y | 56bp |

| | | | | | | |
|-----------|--|-------|---|-----------|---|--------------|
| PFF0365c | G-protein associated signal transduction protein | 996.3 | - | exon skip | N | 162bp (54aa) |
| PFB0445c | DEAD box helicase, UAP56 | 995.9 | - | 3'ss | N | 75bp (25aa) |
| PFD0895c | Bet3 transport protein | 991 | - | antisense | - | N/A |
| PF10_0116 | conserved Plasmodium protein | 989.9 | Y | 5'ss | N | 75bp (25aa) |
| PF14_0604 | conserved Plasmodium protein | 988.1 | Y | exon skip | Y | 343bp |
| PFI0560c | conserved Plasmodium protein | 987.7 | Y | exon skip | Y | 40bp |
| PFB0550w | peptide chain release factor subunit 1 | 985.5 | - | exon skip | Y | 155bp |
| PF11_0355 | conserved Plasmodium protein | 984.6 | Y | antisense | - | N/A |

Table 7 – 5' GC-AG 3' junctions above the 1075 operational HMMSplicer score threshold.

| Junction ID | Score | In PlasmoDB/ ESTs? | Gene | Gene Description |
|-----------------------|--------------|-------------------------------|--------------------------|---|
| chr3:442788-443027 | 1456.5 | Y | PFC0430w | conserved Plasmodium protein |
| chr4:1026986-1027209 | 1195.7 | N | antisense to PFD1050w | alpha tubulin II |
| chr6:987905-988084 | 1492.8 | Y | PFF1170w | conserved Plasmodium protein |
| chr7:815046-815184 | 1500 | N | MAL7P1.87 | conserved Plasmodium protein |
| chr9:395942-396045 | 1298.9 | N | PFI0410c | conserved Plasmodium protein Plasmodium exported protein |
| chr10:671490-671648 | 1113 | Y | PF10_0162 | (PHISTc) |
| chr10:1027894-1028157 | 1472.4 | Y | PF10_0240 | conserved Plasmodium protein |
| chr10:1302659-1302805 | 1302.5 | Y | PF10_0316 | N-acetylglucosaminyl- phosphatidylinositol biosynthetic protein |
| chr11:429839-430026 | 1326 | N | PF11_0114 | actin-like protein homolog |
| chr12:1969285-1969513 | 1203 | Y | PFL2290w | preprocathepsin c precursor |
| chr14:2291528-2291619 | 1392.6 | Y | PF14_0791 | dfg10 like protein |
| chr14:2972968-2973666 | 1242.1 | N | PF14_0697 | dihydroorotase |

Table 8 – Alternative splicing (AS) junctions with more read counts than their conflicting gene model (GM) counterpart(s).

| Gene | AS Junction | GM Junction | Ratio (AS/GM) |
|-------------|-----------------------|--|---------------|
| PF11_0149 | chr11:539968-540165 | chr11:539968-540159 | 2.25 |
| PF14_0338 | chr14:1446627-1446999 | chr14:1446627-1446712 | 3.73 |
| PFF0630c | chr6:533969-534617 | chr6:534249-534617, chr6:533969-534215 | 3.47 |
| PFF0630c | chr6:533966-534617 | chr6:534249-534617, chr6:533969-534215 | 1.47 |
| PFF0630c | chr6:533969-534598 | chr6:534249-534617, chr6:533969-534215 | 1.53 |
| PFD0440w | chr4:433500-433725 | chr4:433506-433725 | 1.64 |
| PF10_0117 | chr10:464972-465324 | chr10:465172-465324, chr10:464972-465111 | 1.04 |
| PF14_0338 | chr14:1446211-1446526 | chr14:1446211-1446367, chr14:1446427-1446526 | 2.37 |
| MAL8P1.106 | chr8:553020-553216 | chr8:553024-553216 | 1.07 |
| PF10_0015 | chr10:68515-68658 | chr10:68515-68728 | 2.27 |
| PFF0920c | chr6:794834-794958 | chr6:794831-794958 | 1.50 |
| MAL8P1.126 | chr8:398687-398974 | chr8:398687-398865 | 2.40 |
| MAL13P1.240 | chr13:1920062-1920318 | chr13:1920062-1920173, chr13:1920206-1920318 | 1.50 |
| MAL7P1.111 | chr7:956990-957159 | chr7:957002-957159 | 1.08 |
| PFD0872w | chr4:805530-805702 | chr4:805407-805702 | 1.42 |
| PFD0872w | chr4:805578-805702 | chr4:805407-805702 | 6.60 |
| PFE1190c | chr5:993243-993442 | chr5:993262-993442 | 1.67 |
| PFL0130c | chr12:149368-149776 | chr12:149368-149519 | 3.28 |
| MAL8P1.93 | chr8:683443-683767 | chr8:683660-683767 | 2.40 |
| PFL1440c | chr12:1229392-1229639 | chr12:1229535-1229639 | 1.00 |
| PFI1470c | chr9:1197006-1197283 | chr9:1197006-1197310 | 1.75 |
| MAL13P1.22 | chr13:224559-225294 | chr13:224559-224725 | 1.25 |
| PFE1420w | chr5:1176398-1176706 | chr5:1176398-1176491 | 2.00 |
| PFC0460w | chr3:466138-466427 | chr3:466213-466427 | 1.00 |
| PFB0125c | chr2:129687-129803 | chr2:129690-129803 | 2.00 |
| PFL0825c | chr12:675551-675933 | chr12:675818-675933 | 4.67 |
| PF14_0581 | chr14:2481102-2481201 | chr14:2481102-2481219 | 23.00 |
| MAL7P1.95 | chr7:885578-885710 | chr7:885608-885710 | 1.00 |
| PF14_0692 | chr14:2957052-2957645 | chr14:2957484-2957645 | 1.50 |
| PFI0280c | chr9:285483-285586 | chr9:285642-285807, chr9:285483-285604 | 4.75 |
| PF10_0149 | chr10:616830-617521 | chr10:616833-617521 | 2.50 |
| PFI0885w | chr9:747034-747344 | chr9:747171-747344 | 2.48 |
| PFL2425w | chr12:2070537-2070668 | chr12:2070537-2070680 | 20.40 |

Table 9 – Genes for which a novel junction conflicted with an unrecovered junction.

“Left” indicates the nt difference between the left edges of the PlasmoDB and RNA-Seq, while “Right” indicates the nt difference between the right edges.

| Gene | PlasmoDB Junction | RNA-Seq Junction | Score | Left | Right |
|-------------|--------------------------|-------------------------|--------------|-------------|--------------|
| PFA0355w | chr1:297957-298132 | chr1:297982-298112 | 1178.0 | 25 | 20 |
| PFA0570w | chr1:449013-449177 | chr1:449054-449177 | 1337.9 | 41 | 0 |
| PFC0150w | chr3:166222-166464 | chr3:166248-166464 | 1656.7 | 26 | 0 |
| PFC0615w | chr3:602208-602569 | chr3:602355-602569 | 1488.6 | 147 | 0 |
| PFD0075w | chr4:107202-107474 | chr4:107202-107297 | 1287.7 | 0 | 177 |
| PFD1140w | chr4:1085469-1085668 | chr4:1085481-1085665 | 1085.8 | 12 | 3 |
| PFD0100c | chr4:138336-138555 | chr4:138461-138555 | 1167.9 | 125 | 0 |
| PFD0330w | chr4:357191-357321 | chr4:357203-357321 | 1297.6 | 12 | 0 |
| PFD0700c | chr4:663522-663655 | chr4:663531-663649 | 1226.3 | 9 | 6 |
| PFD0872w | chr4:805156-805321 | chr4:805156-805288 | 1235.9 | 0 | 33 |
| PFF1190c | chr6:1010724-1010950 | chr6:1010724-1010881 | 1142.0 | 0 | 69 |
| PFF1377w | chr6:1177947-1178041 | chr6:1177963-1178041 | 1266.4 | 16 | 0 |
| PFF0925w | chr6:796991-797193 | chr6:796991-797131 | 1398.4 | 0 | 62 |
| MAL7P1.160 | chr7:1314328-1314521 | chr7:1314328-1314515 | 1551.7 | 0 | 6 |
| MAL7P1.339 | chr7:343513-343726 | chr7:343526-343721 | 1486.2 | 13 | 5 |
| MAL7P1.66 | chr7:675378-675598 | chr7:675378-675559 | 1125.6 | 0 | 39 |
| MAL7P1.82 | chr7:794576-794751 | chr7:794616-794751 | 1302.8 | 40 | 0 |
| PF07_0086 | chr7:962680-962919 | chr7:962770-962919 | 1341.0 | 90 | 0 |
| MAL8P1.99 | chr8:615646-616358 | chr8:615646-615840 | 1107.8 | 0 | 518 |
| MAL8P1.83 | chr8:732576-732757 | chr8:732576-732751 | 1521.8 | 0 | 6 |
| MAL8P1.160 | chr8:84319-84460 | chr8:84320-84460 | 1173.4 | 1 | 0 |
| PFI1220w | chr9:1011147-1011251 | chr9:1011156-1011251 | 1461.4 | 9 | 0 |
| PFI1400c | chr9:1146381-1146526 | chr9:1146381-1146520 | 1639.8 | 0 | 6 |
| PFI1690c | chr9:1364768-1364904 | chr9:1364780-1364904 | 1367.8 | 12 | 0 |
| PFI0175w | chr9:169300-169702 | chr9:169300-169379 | 1521.4 | 0 | 323 |
| PFI0790w | chr9:674191-674340 | chr9:674191-674337 | 1317.6 | 0 | 3 |
| PFI0800c | chr9:681777-681873 | chr9:681781-681873 | 1189.7 | 4 | 0 |
| PF10_0254 | chr10:1092995-1093173 | chr10:1092995-1093119 | 1246.6 | 0 | 54 |
| PF10_0359 | chr10:1446641-1446725 | chr10:1446641-1446722 | 1548.4 | 0 | 3 |
| PF10_0362 | chr10:1462730-1463026 | chr10:1462730-1462862 | 1499.9 | 0 | 164 |
| PF10_0365 | chr10:1476403-1476603 | chr10:1476442-1476603 | 1218.7 | 39 | 0 |
| PF10_0188 | chr10:788206-788552 | chr10:788249-788463 | 1548.5 | 43 | 89 |
| PF10_0208 | chr10:866155-866292 | chr10:866169-866292 | 1569.5 | 14 | 0 |
| PF11_0530 | chr11:1499568-1500174 | chr11:1499575-1499754 | 1699.2 | 7 | 420 |
| PF11_0423 | chr11:1657666-1657814 | chr11:1657678-1657814 | 1294.6 | 12 | 0 |
| PF11_0560 | chr11:1756632-1756808 | chr11:1756676-1756808 | 1686.1 | 44 | 0 |
| PF11_0263 | chr11:989901-990239 | chr11:990107-990239 | 1374.7 | 206 | 0 |
| PFL1440c | chr12:1231376-1231539 | chr12:1231382-1231539 | 1316.6 | 6 | 0 |

| | | | | | |
|-------------|-----------------------|-----------------------|--------|----|-----|
| PFL2255w | chr12:1948468-1948653 | chr12:1948474-1948653 | 1305.7 | 6 | 0 |
| PFL2525c | chr12:2146243-2146457 | chr12:2146243-2146409 | 1184.1 | 0 | 48 |
| PFL0395c | chr12:359055-359333 | chr12:359055-359289 | 1267.9 | 0 | 44 |
| PFL0700w | chr12:612891-613131 | chr12:612891-613090 | 1183.3 | 0 | 41 |
| PFL1015w | chr12:848566-848695 | chr12:848566-848671 | 1173.6 | 0 | 24 |
| MAL13P1.216 | chr13:1716025-1716158 | chr13:1716025-1716140 | 1315.5 | 0 | 18 |
| MAL13P1.252 | chr13:2001119-2001299 | chr13:2001128-2001299 | 1328.9 | 9 | 0 |
| MAL13P1.268 | chr13:2112106-2112324 | chr13:2112140-2112295 | 1573.6 | 34 | 29 |
| MAL13P1.84 | chr13:670135-670472 | chr13:670135-670217 | 1162.6 | 0 | 255 |
| PF14_0726 | chr14:3121364-3121442 | chr14:3121376-3121442 | 1129.0 | 12 | 0 |
| PF14_0739a | chr14:3162436-3162610 | chr14:3162447-3162610 | 1358.9 | 11 | 0 |
| PF14_0153 | chr14:624589-624991 | chr14:624589-624788 | 1389.9 | 0 | 203 |

Table 10 – Genes with antisense junctions for which ESTs or junctions provide evidence of overlap with a neighbor.

| <u>Gene 1</u> <u>Gene 2</u> | <u>Gene Descriptions</u> | <u>In PlasmoDB/ ESTs?</u> | <u>Type</u> | <u>Antisense Junction ID(s)</u> | <u>Score</u> |
|--------------------------------|---|-------------------------------|--------------|--|--------------------------------------|
| PFB0295w PFB0300c | adenylosuccinate lyase merozoite surface protein 2 precursor | N / Y | Tail-to-tail | chr2:272978-273765 chr2:272978-273224 | 1516.1 1283.2 |
| PFD0330w PFD0335c | conserved Plasmodium protein conserved Plasmodium protein | N / N | Tail-to-tail | chr4:358235-358447 | 1294.4 |
| PFE1425c PFE1420w | conserved Plasmodium protein F-actin capping protein, alpha subunit | Y / Y | Tail-to-tail | chr5:1176803-1177056 chr5:1177122-1177247 chr5:1177117-1177247 | 1097.8 1234.1 1189.5 |
| PFE1530c PFE1525w | XAP-5 DNA binding protein conserved Plasmodium membrane protein | N / Y | Tail-to-tail | chr5:1248492-1248783 | 1295.4 |
| PFF1160w PFF1165c | conserved Plasmodium protein conserved Plasmodium protein | N / Y | Tail-to-tail | chr6:985100-985367 | 1135 |
| PFI0425w PFI0430c | Transporter conserved Plasmodium protein | N / Y | Tail-to-tail | chr9:406262-406377 chr9:405899-406215 | 1383.4 1085.2 |
| PF10_0070 PF10_0070a | conserved Plasmodium membrane protein conserved Plasmodium protein | N / N | Tail-to-tail | chr10:286265-286493 | 1163.2 |
| PF10_0227 PF10_0226 | HORMA domain protein conserved Plasmodium protein | N / Y | Head-to-head | chr10:985100-985247 chr10:985100-985363 | 1104.4 1146.5 |
| PF11_0426 PF11_0427 | conserved Plasmodium protein dolichol phosphate mannose synthase | N / Y | Tail-to-tail | chr11:1668191-1668311 | 1081.3 |
| PFL0746c PFL0750w | conserved Plasmodium protein conserved Plasmodium protein | N / Y | Head-to-head | chr12:634427-635023 | 1415.1 |
| PFL0980w PFL0985c | conserved Plasmodium protein conserved protein | N / N | Tail-to-tail | chr12:818465-818764 | 1189.9 |
| PFL1355w PFL1360c | conserved Plasmodium protein conserved Plasmodium protein | N / Y | Tail-to-tail | chr12:1141426-1141590 | 1368.4 |
| PFL2055w PFL2060c | 40S ribosomal protein S17 rabGDI protein | N / N | Tail-to-tail | chr12:1804508-1805371 | 1414.8 |
| PFL2105c PFL2100w | conserved protein ubiquitin conjugating enzyme E2 | N / N | Tail-to-tail | chr12:1831710-1831854 chr12:1831697-1831854 chr12:1831028-1831631 chr12:1831710-1831830 | 1709.4 1686.1 1306.3 1167.2 |
| PF14_0666 PF14_0667 | conserved Plasmodium protein conserved Plasmodium protein | N / Y | Head-to-head | chr14:2866269-2866721 | 1438.5 |

Figure 1 – Operational definitions of terms referencing A) specific portions of junctions, B) relationships of junctions to gene models and C) each other, and D) relationships of junctions on opposite strands to gene models and each other.

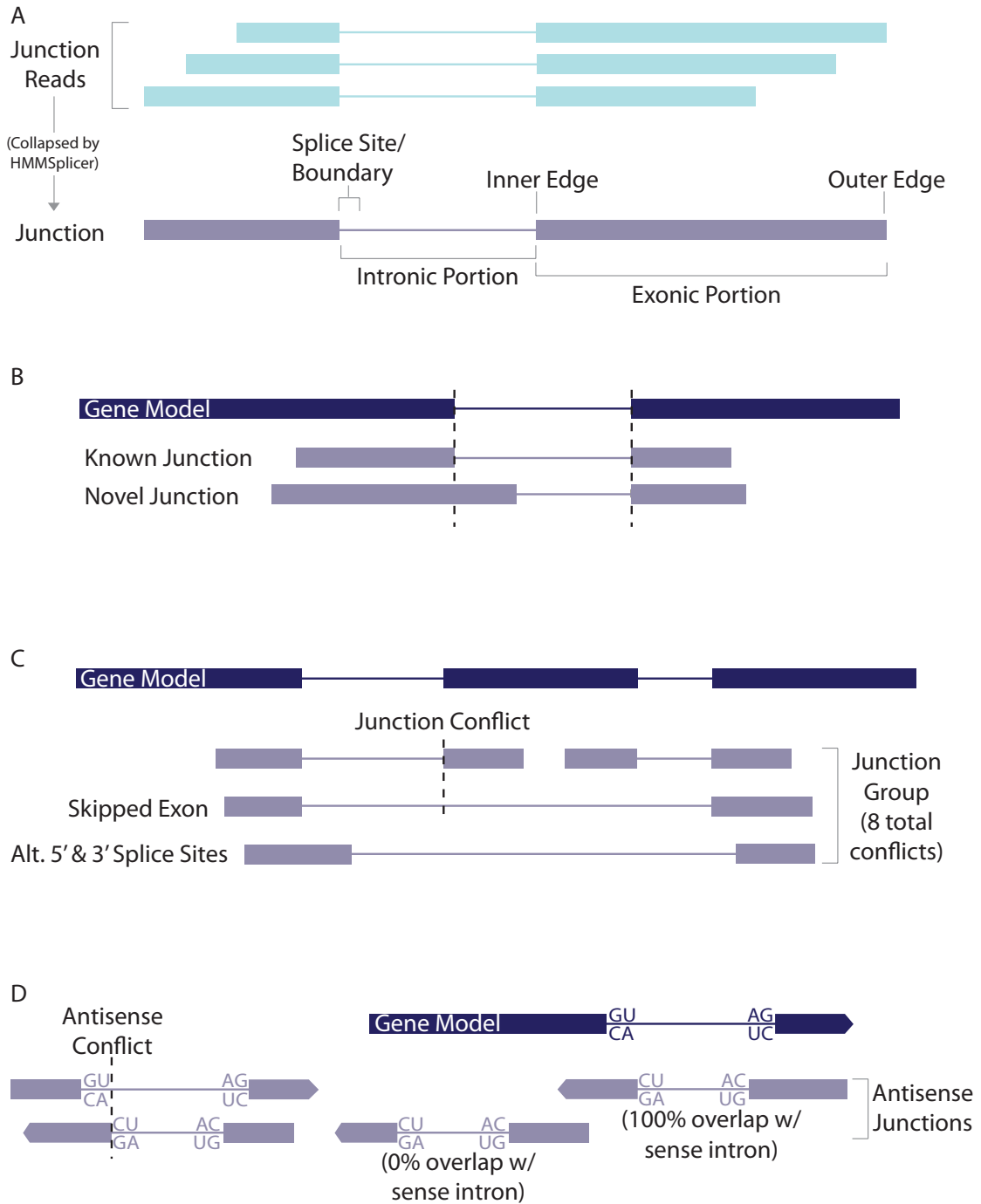


Figure 2 – Receiver operating characteristic (ROC) curve used to determine optimal parameters for the coverage filter implemented within the noncanonical junctions list. Known junctions served as true positives, while junctions filtered out because their breakpoint could be adjusted to match a previously seen junction (n=355) served as false positives. Coverage was measured as both a median (red line) and maximum (all others) per nucleotide within 0-200bp of the outer edges of each junction. The chosen parameters (maximum coverage of 1000 reads/nt within 100bp of the outer junction edges) are indicated by the dashed line and include more than 95% of true positives, while excluding more than 50% of false positives.

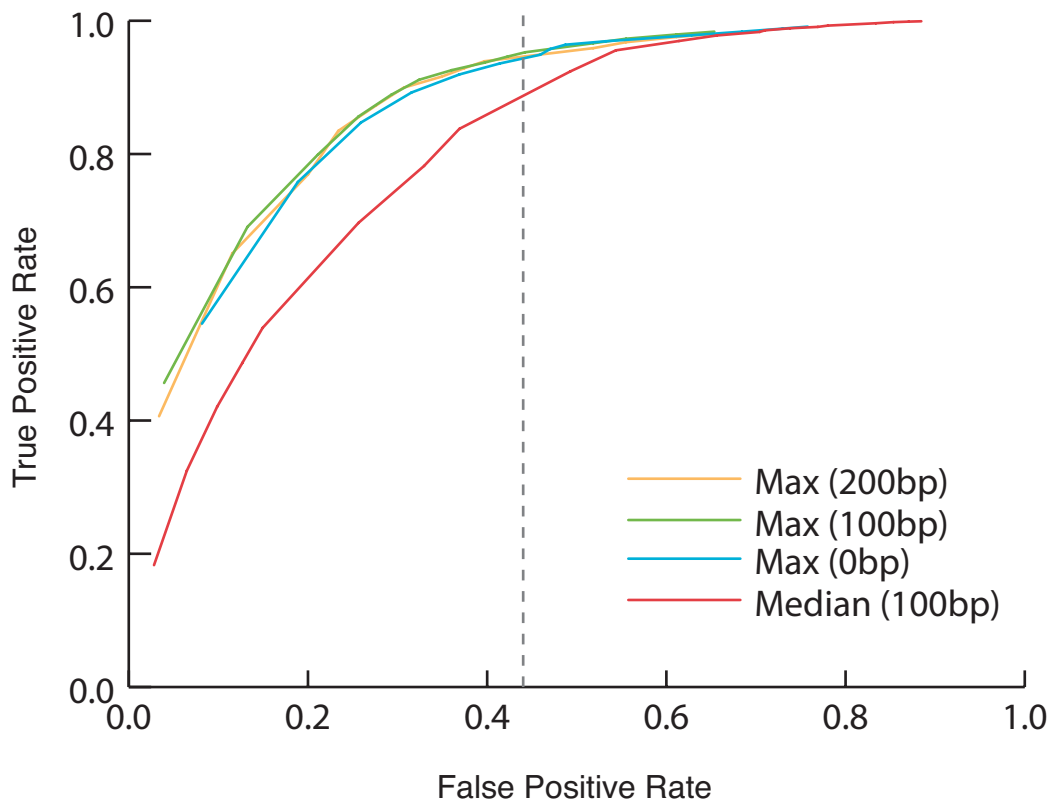


Figure 3 – Multiple alignment of C-termini of *S. cerevisiae* and *H. sapiens* PRP22 and PRP2 with PF10_0294. Highlighted amino acids represent DC amino acid doublet present in PRP22 proteins in yeast species.

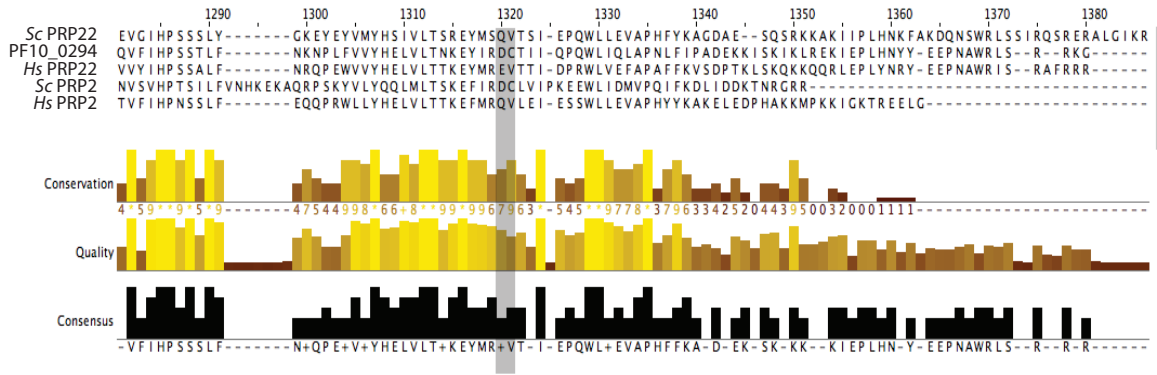


Figure 4 – A) Histogram of 5' GU-AG 3' junctions found by HMMSplicer binned by score. Defaults retain all junctions supported by multiple reads scoring above 400 and all junctions supported by single reads scoring above 600. The grey line plots all reported 5' GU-AG 3' junctions, while the red line charts junctions that match previously known junctions in PlasmoDB v6.3 gene models or in ESTs. The blue line charts new junctions. The dashed line drawn at 1075 represents the operational score threshold. **B)** Breakdown of canonical junctions with scores above 1075, with additional classification of new junctions. “Outside of gene model” refers to new junctions with at least one inner edge mapped to an intergenic region. “Within gene model” indicates that both inner edges mapped to the same gene model. “Neighboring gene models” indicates that the inner edges mapped to neighboring gene models. **C)** Comparison of the 5' and 3' splice site WebLogos for previously known junctions recovered versus new junctions above 1075. WebLogos calculated for human junctions are included for reference. Red bars indicate the 5' GU-AG 3' boundaries used for inclusion in each set. The height of each letter indicates the preference strength for that nucleotide at each position.

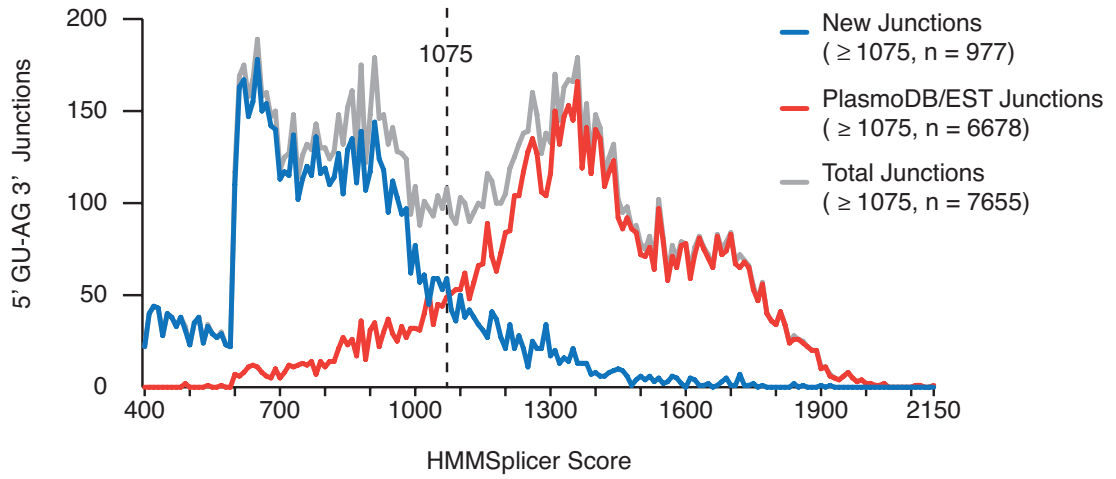
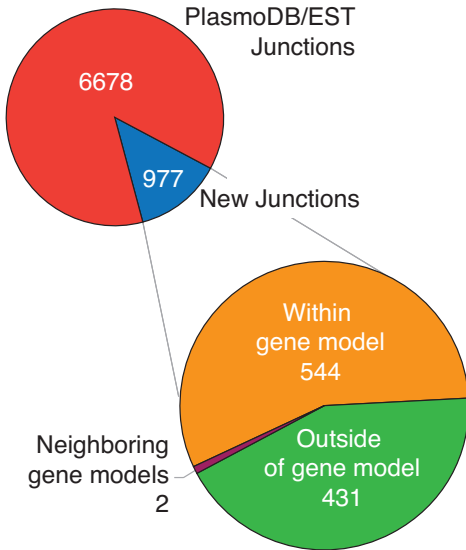
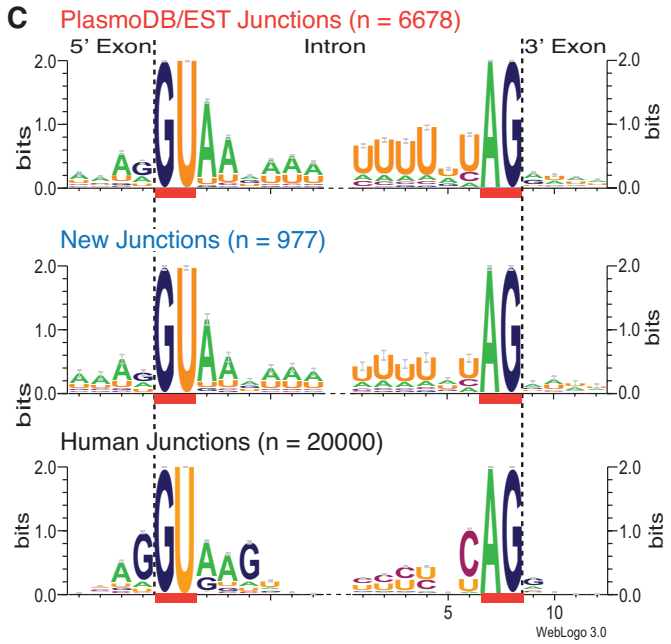
A**B****C**

Figure 5 – Validation of new splicing events. **A)** Shade indicates the relative abundance of each isoform. Initial outer PCR (green arrows) amplifies both isoforms from cDNA. A restriction enzyme then cuts the known isoform. Nested inner PCR (blue arrows) amplifies only the uncut, new isoform, which is then sequence confirmed. Gbrowse⁽⁶⁹⁾ windows depict validation of a skipped exon in MAL13P1.159 (**B**), an antisense junction in PFF0290w (**C**), and an alternate 3'splice site in PFB0279w (**D**). All HMMSplicer junctions scoring higher than 980 are shown as either dark blue bars (known junctions) or light blue bars (new conflicting junctions). The number of reads supporting each junction is shown in the bars, while the direction of the arrow reflects the direction of the splice sites. Validation sequencing results are shown in magenta. Bowtie coverage for each nucleotide in the window is shown as a histogram. Underneath, the dark blue bars depict PlasmoDB v6.3 gene models with numbers denoting the exons, while the gold bars at the bottom of each window depict ESTs.

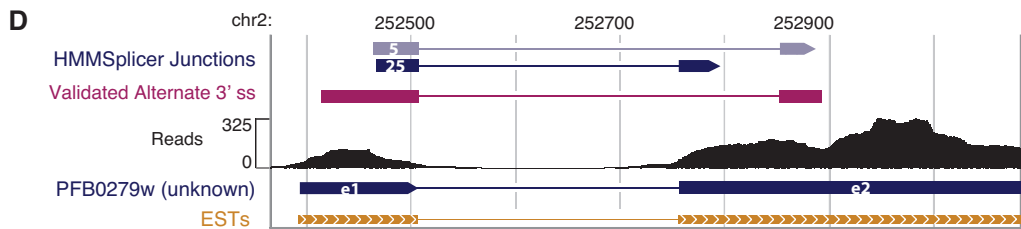
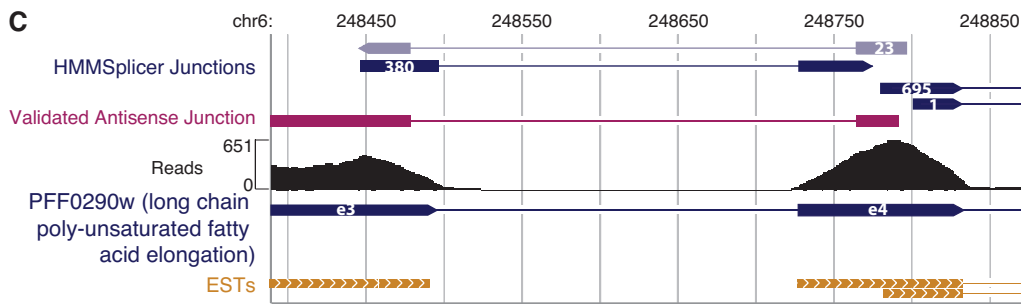
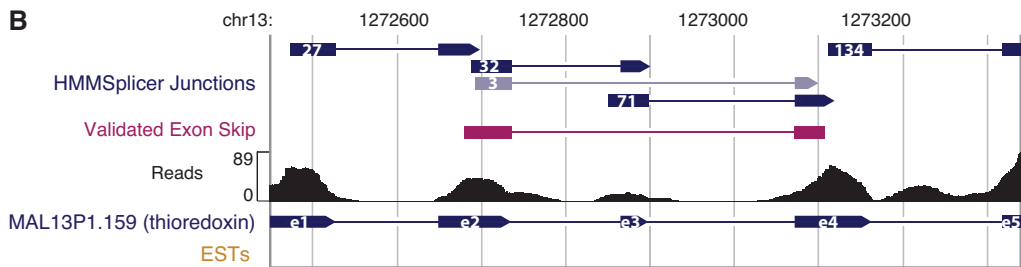
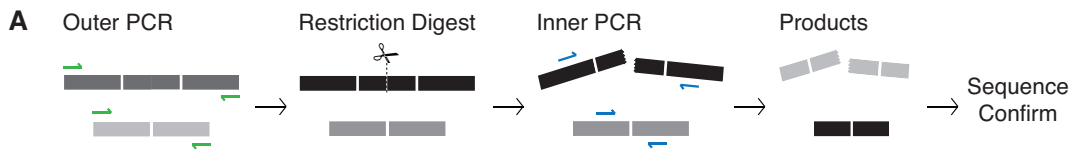


Figure 6 – A) WebLogo motifs for the 5' and 3' splice sites of the bottom 10% of novel 5' GU-AG 3' junctions scoring under 1075. Dashed lines indicate exonintron boundaries, while red lines indicate sequence selected for inclusion in the set. **B)** Distribution, for known (red line) and novel (blue line) junctions with HMMSplicer scores < 1075, of motif scores generated from position-specific scoring matrices of the known splice site motifs in Figure 4C.

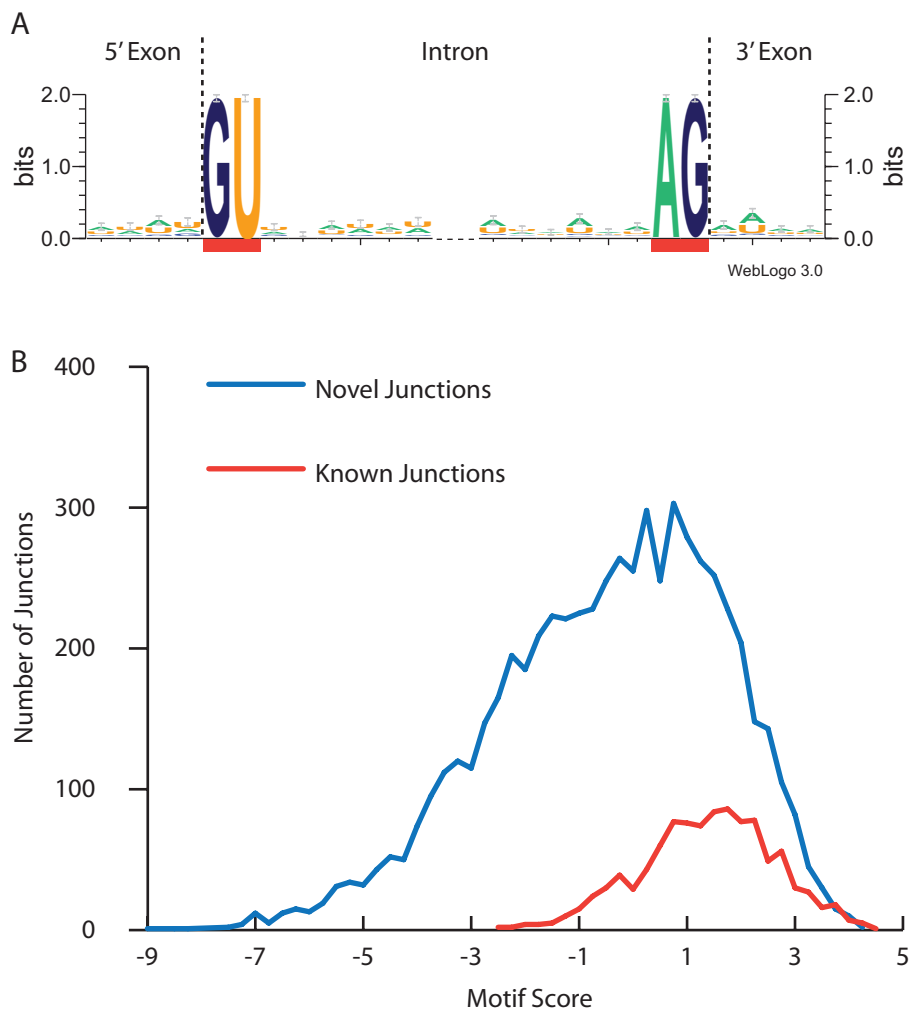


Figure 7 – WebLogo 5' and 3' splice site motifs for manually curated 5' GC-AG 3' HMMSplicer junctions (n = 12). Red bars indicate the boundaries used for inclusion in the set. The height of each letter indicates the information content for that nucleotide at each position. The large error bars derive from the small size of the input set.

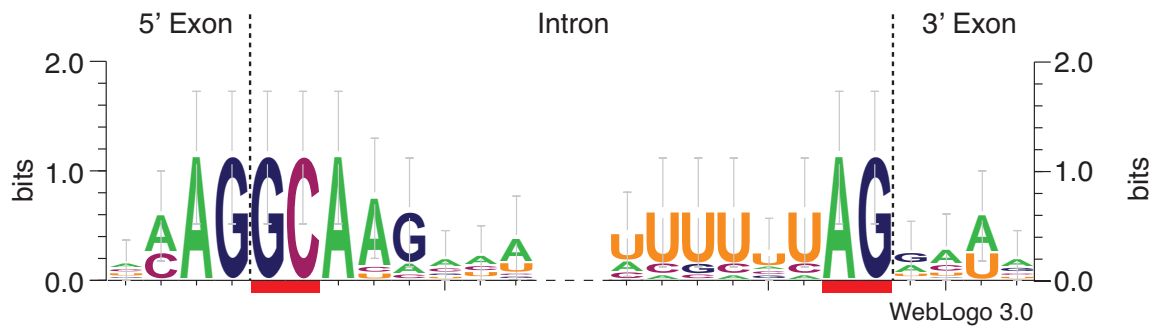


Figure 8 – Breakdown of alternative splicing events detected transcriptome-wide.

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

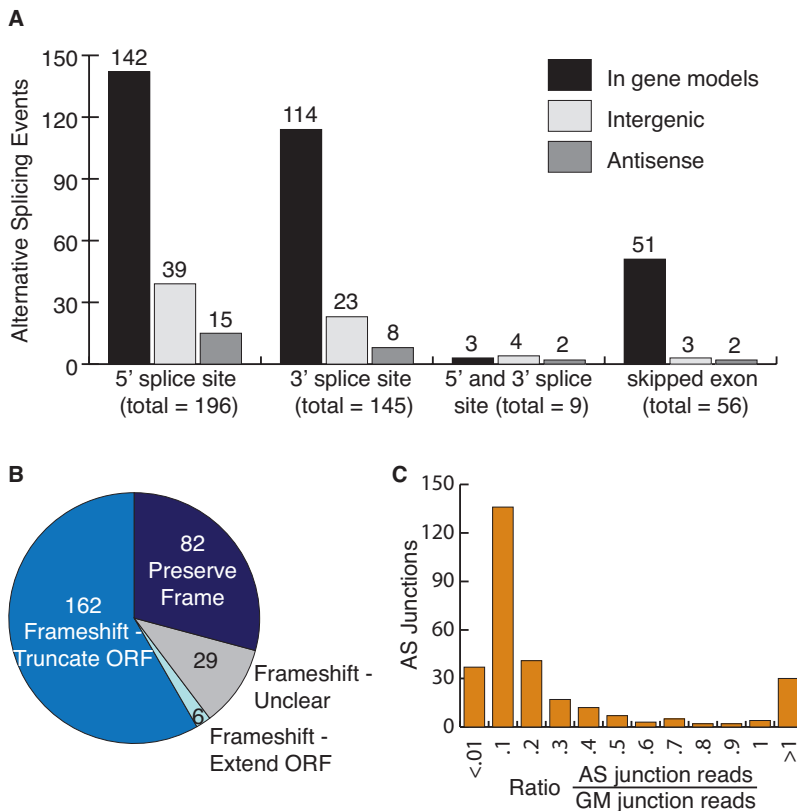
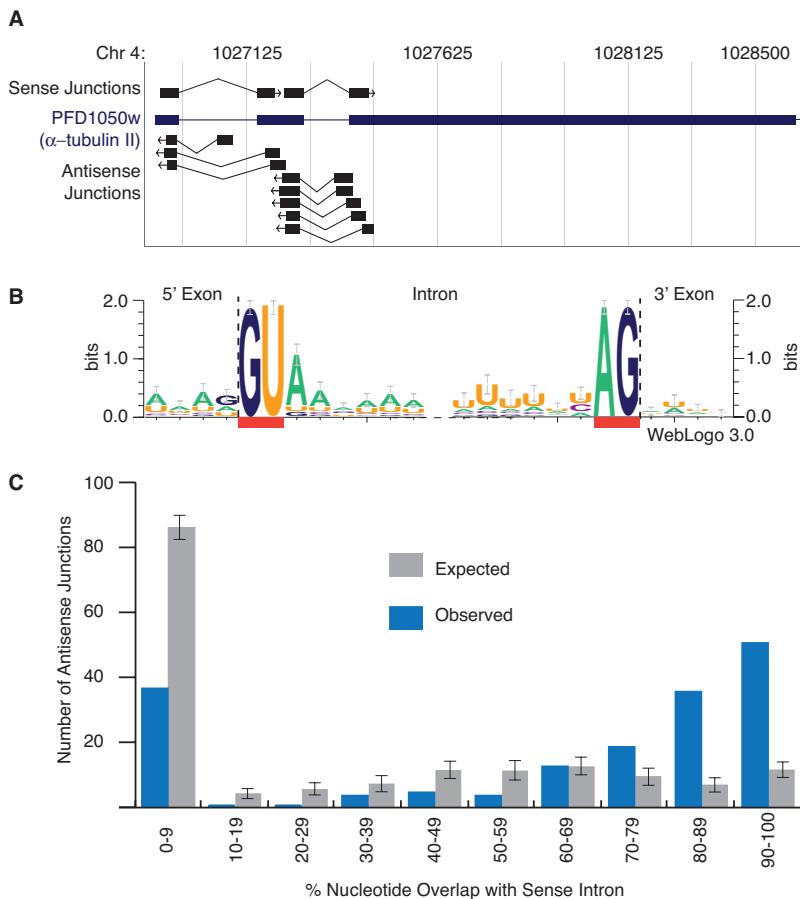


Figure 9 – Characterization of antisense splice junctions. **A)** Schematic of all sense and antisense junctions recovered for PFD1050w (α -tubulin II). **B)** WebLogos of the 5' and 3' splice sites of antisense junctions. The height of each letter indicates the preference strength for that nucleotide at each position. **C)** Observed and expected distributions of antisense intron overlap with sense introns. The expected distribution was calculated by first determining the probability of encountering a GT (5' splice site) or an AG (3' splice site) on the opposite strand of introns versus exons in the genes with mapped antisense junctions. These probabilities guided otherwise random re-placement of each antisense junction within its corresponding gene model. This re-placement was iterated 100 times, with the mean percent of nucleotide overlap with sense introns +/- standard deviation shown.



References:

1. Bozdech,Z., Llinás,M., Pulliam,B.L., Wong,E.D., Zhu,J. and DeRisi,J.L. (2003) The Transcriptome of the Intraerythrocytic Developmental Cycle of *Plasmodium falciparum*. *PLoS Biol*, **1**, e5, 10.1371/journal.pbio.0000005.
2. Le Roch,K.G., Zhou,Y., Blair,P.L., Grainger,M., Moch,J.K., Haynes,J.D., De la Vega,P., Holder,A.A., Batalov,S., Carucci,D.J. et al. (2003) Discovery of Gene Function by Expression Profiling of the Malaria Parasite Life Cycle. *Science*, **301**, 1503-1508, 10.1126/science.1087025.
3. Silvestrini,F., Bozdech,Z., Lanfrancotti,A., Di Giulio,E., Bultrini,E., Picci,L., Derisi,J.L., Pizzi,E. and Alano,P. (2005) Genome-wide identification of genes upregulated at the onset of gametocytogenesis in *Plasmodium falciparum*. *Mol. Biochem. Parasitol*, **143**, 100-110, 10.1016/j.molbiopara.2005.04.015.
4. Gardner,M.J., Hall,N., Fung,E., White,O., Berriman,M., Hyman,R.W., Carlton,J.M., Pain,A., Nelson,K.E., Bowman,S. et al. (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature*, **419**, 498-511, 10.1038/nature01097.
5. Burtis,K.C. (1993) The regulation of sex determination and sexually dimorphic differentiation in *Drosophila*. *Current Opinion in Cell Biology*, **5**, 1006-1014, 10.1016/0955-0674(93)90085-5.
6. Madsen,J. and Stoltzfus,C.M. (2006) A suboptimal 5' splice site downstream of HIV-1 splice site A1 is required for unspliced viral mRNA accumulation and efficient virus replication. *Retrovirology*, **3**, 10, 10.1186/1742-4690-3-10.
7. Upadhyay,R., Bawankar,P., Malhotra,D. and Patankar,S. (2005) A screen for

- conserved sequences with biased base composition identifies noncoding RNAs in the A-T rich genome of *Plasmodium falciparum*. *Molecular and Biochemical Parasitology*, **144**, 149-158, 10.1016/j.molbiopara.2005.08.012.
8. Chakrabarti,K., Pearson,M., Grate,L., Sterne-Weiler,T., Deans,J., Donohue,J.P. and Ares,M. (2007) Structural RNAs of known and unknown function identified in malaria parasites by comparative genomics and RNA analysis. *RNA*, **13**, 1923-1939, 10.1261/rna.751807.
 9. Shankar,J., Pradhan,A. and Tuteja,R. (2008) Isolation and characterization of *Plasmodium falciparum* UAP56 homolog: Evidence for the coupling of RNA binding and splicing activity by site-directed mutations. *Archives of Biochemistry and Biophysics*, **478**, 143-153, 10.1016/j.abb.2008.07.027.
 10. Lamond,A.I. (1993) The spliceosome. *Bioessays*, **15**, 595-603, 10.1002/bies.950150905.
 11. Burset,M., Seledtsov,I.A. and Solovyev,V.V. (2000) Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucl. Acids Res.*, **28**, 4364-4375, 10.1093/nar/28.21.4364.
 12. Lu,F., Jiang,H., Ding,J., Mu,J., Valenzuela,J., Ribeiro,J. and Su,X. (2007) cDNA sequences reveal considerable gene prediction inaccuracy in the *Plasmodium falciparum* genome. *BMC Genomics*, **8**, 255, 10.1186/1471-2164-8-255.
 13. Li,L., Brunk,B.P., Kissinger,J.C., Pape,D., Tang,K., Cole,R.H., Martin,J., Wylie,T., Dante,M., Fogarty,S.J. et al. (2003) Gene Discovery in the Apicomplexa as Revealed by EST Sequencing and Assembly of a Comparative Gene Database. *Genome Research*, **13**, 443-454, 10.1101/gr.693203.

14. Otto,T.D., Wilinski,D., Assefa,S., Keane,T.M., Sarry,L.R., Böhme,U., Lemieux,J., Barrell,B., Pain,A., Berriman,M. et al. (2010) New insights into the blood-stage transcriptome of Plasmodium falciparum using RNA-Seq. *Molecular Microbiology*, **76**, 12-24, 10.1111/j.1365-2958.2009.07026.x.
15. Bracchi-Ricard,V., Barik,S., Delvecchio,C., Doerig,C., Chakrabarti,R. and Chakrabarti,D. (2000) PfPK6, a novel cyclin-dependent kinase/mitogen-activated protein kinase-related protein kinase from Plasmodium falciparum. *Biochem J*, **347**, 255-263.
16. Muhia,D.K., Swales,C.A., Eckstein-Ludwig,U., Saran,S., Polley,S.D., Kelly,J.M., Schaap,P., Krishna,S. and Baker,D.A. (2003) Multiple Splice Variants Encode a Novel Adenylyl Cyclase of Possible Plastid Origin Expressed in the Sexual Stage of the Malaria Parasite Plasmodium falciparum. *Journal of Biological Chemistry*, **278**, 22014-22022, 10.1074/jbc.M301639200.
17. Saenz,F.E., Balu,B., Smith,J., Mendonca,S.R. and Adams,J.H. (2008) The Transmembrane Isoform of Plasmodium falciparum MAEBL Is Essential for the Invasion of Anopheles Salivary Glands. *PLoS ONE*, **3**, e2287, 10.1371/journal.pone.0002287.
18. Wentzinger,L., Bopp,S., Tenor,H., Klar,J., Brun,R., Beck,H.P. and Seebeck,T. (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, 10.1016/j.ijpara.2008.05.016.
19. Iriko,H., Jin,L., Kaneko,O., Takeo,S., Han,E., Tachibana,M., Otsuki,H., Torii,M. and

- Tsuboi,T. (2009) A small-scale systematic analysis of alternative splicing in *Plasmodium falciparum*. *Parasitology International*, **58**, 196-199, 10.1016/j.parint.2009.02.002.
20. Knapp,B., Nau,U., Hundt,E. and Küpper,H.A. (1991) Demonstration of alternative splicing of a pre-mRNA expressed in the blood stage form of *Plasmodium falciparum*. *J. Biol. Chem*, **266**, 7148-7154.
21. Liu,Q., Mackey,A.J., Roos,D.S. and Pereira,F.C.N. (2008) Evigan: a hidden variable model for integrating gene evidence for eukaryotic gene prediction. *Bioinformatics*, **24**, 597-605, 10.1093/bioinformatics/btn004.
22. Jen,C., Michalopoulos,I., Westhead,D. and Meyer,P. (2005) Natural antisense transcripts with coding capacity in *Arabidopsis* may have a regulatory role that is not linked to double-stranded RNA degradation. *Genome Biology*, **6**, R51, 10.1186/gb-2005-6-6-r51.
23. RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group) and the FANTOM Consortium, Katayama,S., Tomaru,Y., Kasukawa,T., Waki,K., Nakanishi,M., Nakamura,M., Nishida,H., Yap,C.C., Suzuki,M. et al. (2005) Antisense Transcription in the Mammalian Transcriptome. *Science*, **309**, 1564-1566, 10.1126/science.1112009.
24. Faghihi,M.A. and Wahlestedt,C. (2009) Regulatory roles of natural antisense transcripts. *Nat Rev Mol Cell Biol*, **10**, 637-643, 10.1038/nrm2738.
25. Militello,K.T., Patel,V., Chessler,A., Fisher,J.K., Kasper,J.M., Gunasekera,A. and Wirth,D.F. (2005) RNA polymerase II synthesizes antisense RNA in *Plasmodium falciparum*. *RNA*, **11**, 365-370, 10.1261/rna.7940705.

26. Raabe,C.A., Sanchez,C.P., Randau,G., Robeck,T., Skryabin,B.V., Chinni,S.V., Kube,M., Reinhardt,R., Ng,G.H., Manickam,R. et al. (2010) A global view of the nonprotein-coding transcriptome in Plasmodium falciparum. *Nucl. Acids Res.*, **38**, 608-617, 10.1093/nar/gkp895.
27. Sorber,K., Chiu,C., Webster,D., Dimon,M., Ruby,J.G., Hekele,A. and DeRisi,J.L. (2008) The Long March: A Sample Preparation Technique that Enhances Contig Length and Coverage by High-Throughput Short-Read Sequencing. *PLoS ONE*, **3**, e3495, 10.1371/journal.pone.0003495.
28. Dimon,M., Sorber,K. and DeRisi,J.L. (2010) HMMSplicer: a tool for efficient and sensitive discovery of known and novel splice junctions in RNA-Seq data. *PLoS ONE*, **accepted**.
29. Moreno-Hagelsieb,G. and Latimer,K. (2008) Choosing BLAST options for better detection of orthologs as reciprocal best hits. *Bioinformatics*, **24**, 319-324, 10.1093/bioinformatics/btm585.
30. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403-410, 10.1016/S0022-2836(05)80360-2.
31. Boucher,L., Ouzounis,C.A., Enright,A.J. and Blencowe,B.J. (2001) A genome-wide survey of RS domain proteins. *RNA*, **7**, 1693-1701.
32. PlasmoDB: An integrative database of the Plasmodium falciparum genome. Tools for accessing and analyzing finished and unfinished sequence data (2001) *Nucl. Acids Res.*, **29**, 66-69, 10.1093/nar/29.1.66.
33. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-

- efficient alignment of short DNA sequences to the human genome. *Genome Biol*, **10**, R25, 10.1186/gb-2009-10-3-r25.
34. Kent, W.J. (2002) BLAT—The BLAST-Like Alignment Tool. *Genome Research*, **12**, 656-664, 10.1101/gr.229202.
35. Huang, D.W., Sherman, B.T. and Lempicki, R.A. (2008) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protocols*, **4**, 44-57, 10.1038/nprot.2008.211.
36. Beissbarth, T. and Speed, T.P. (2004) GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics*, **20**, 1464-1465, 10.1093/bioinformatics/bth088.
37. Wilhelm, B.T., Marguerat, S., Watt, S., Schubert, F., Wood, V., Goodhead, I., Penkett, C.J., Rogers, J. and Bahler, J. (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*, **453**, 1239-1243, 10.1038/nature07002.
38. Kaufer, N.F. and Potashkin, J. (2000) Analysis of the splicing machinery in fission yeast: a comparison with budding yeast and mammals. *Nucl. Acids Res.*, **28**, 3003-3010, 10.1093/nar/28.16.3003.
39. Stevens, S.W., Barta, I., Ge, H.Y., Moore, R.E., Young, M.K., Lee, T.D. and Abelson, J. (2001) Biochemical and genetic analyses of the U5, U6, and U4/U6 x U5 small nuclear ribonucleoproteins from *Saccharomyces cerevisiae*. *RNA*, **7**, 1543-1553.
40. Bessonov, S., Anokhina, M., Will, C.L., Urlaub, H. and Luhrmann, R. (2008) Isolation of an active step I spliceosome and composition of its RNP core. *Nature*, **452**, 846-850, 10.1038/nature06842.

41. Lardelli,R.M., Thompson,J.X., Yates,J.R. and Stevens,S.W. (2010) Release of SF3 from the intron branchpoint activates the first step of pre-mRNA splicing. *RNA*, **16**, 516-528, 10.1261/rna.2030510.
42. Edwalds-Gilbert,G., Kim,D., Silverman,E. and Lin,R. (2004) Definition of a spliceosome interaction domain in yeast Prp2 ATPase. *RNA*, **10**, 210-220, 10.1261/rna.5151404.
43. Barash,Y., Calarco,J.A., Gao,W., Pan,Q., Wang,X., Shai,O., Blencowe,B.J. and Frey,B.J. (2010) Deciphering the splicing code. *Nature*, **465**, 53-59, 10.1038/nature09000.
44. Venables,J.P., Koh,C., Froehlich,U., Lapointe,E., Couture,S., Inkel,L., Bramard,A., Paquet,E.R., Watier,V., Durand,M. et al. (2008) Multiple and Specific mRNA Processing Targets for the Major Human hnRNP Proteins. *Mol. Cell. Biol.*, **28**, 6033-6043, 10.1128/MCB.00726-08.
45. Long,J. and Caceres,J. (2009) The SR protein family of splicing factors: master regulators of gene expression. *Biochem. J.*, **417**, 15, 10.1042/BJ20081501.
46. Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Birney,E., Biswas,M., Bucher,P., Cerutti,L., Corpet,F., Croning,M.D.R. et al. (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucl. Acids Res.*, **29**, 37-40, 10.1093/nar/29.1.37.
47. Watanabe,J., Wakaguri,H., Sasaki,M., Suzuki,Y. and Sugano,S. (2007) Comparasite: a database for comparative study of transcriptomes of parasites defined by full-length cDNAs. *Nucl. Acids Res.*, **35**, D431-438, 10.1093/nar/gkl1039.
48. Florent,I., Porcel,B., Guillaume,E., Da Silva,C., Artiguenave,F., Marechal,E.,

- Brehelin,L., Gascuel,O., Charneau,S., Wincker,P. et al. (2009) A Plasmodium falciparum FcB1-schizont-EST collection providing clues to schizont specific gene structure and polymorphism. *BMC Genomics*, **10**, 235, 10.1186/1471-2164-10-235.
49. Crooks,G.E., Hon,G., Chandonia,J. and Brenner,S.E. (2004) WebLogo: A Sequence Logo Generator. *Genome Research*, **14**, 1188-1190, 10.1101/gr.849004.
50. D'haeseleer,P. (2006) What are DNA sequence motifs? *Nat Biotech*, **24**, 423-425, 10.1038/nbt0406-423.
51. Clark,F. and Thanaraj,T.A. (2002) Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human. *Hum. Mol. Genet.*, **11**, 451-464, 10.1093/hmg/11.4.451.
52. Tarn,W. and Steitz,J.A. (1996) A Novel Spliceosome Containing U11, U12, and U5 snRNPs Excises a Minor Class (AT-AC) Intron In Vitro. *Cell*, **84**, 801-811, 10.1016/S0092-8674(00)81057-0.
53. Sidrauski,C., Cox,J.S. and Walter,P. (1996) tRNA Ligase Is Required for Regulated mRNA Splicing in the Unfolded Protein Response. *Cell*, **87**, 405-413, 10.1016/S0092-8674(00)81361-6.
54. Eckert,K.A. and Kunkel,T.A. (1991) DNA polymerase fidelity and the polymerase chain reaction. *PCR Methods Appl*, **1**, 17-24.
55. Cocquet,J., Chong,A., Zhang,G. and Veitia,R.A. (2006) Reverse transcriptase template switching and false alternative transcripts. *Genomics*, **88**, 127-131, 10.1016/j.ygeno.2005.12.013.
56. Shinde,D., Lai,Y., Sun,F. and Arnheim,N. (2003) Taq DNA polymerase slippage

- mutation rates measured by PCR and quasi-likelihood analysis: (CA/GT)_n and (A/T)_n microsatellites. *Nucl. Acids Res.*, **31**, 974-980, 10.1093/nar/gkg178.
57. Will,C.L., Schneider,C., Hossbach,M., Urlaub,H., Rauhut,R., Elbashir,S., Tuschl,T. and Luhrmann,R. (2004) The human 18S U11/U12 snRNP contains a set of novel proteins not found in the U2-dependent spliceosome. *RNA*, **10**, 929-941, 10.1261/rna.7320604.
58. Lopez,M.D., Alm Rosenblad,M. and Samuelsson,T. (2008) Computational screen for spliceosomal RNA genes aids in defining the phylogenetic distribution of major and minor spliceosomal components. *Nucl. Acids Res.*, **36**, 3001-3010, 10.1093/nar/gkn142.
59. Sun,S., Zhang,Z., Sinha,R., Karni,R. and Krainer,A.R. (2010) SF2/ASF autoregulation involves multiple layers of post-transcriptional and translational control. *Nat Struct Mol Biol*, **17**, 306-312, 10.1038/nsmb.1750.
60. Conti,E. and Izaurralde,E. (2005) Nonsense-mediated mRNA decay: molecular insights and mechanistic variations across species. *Current Opinion in Cell Biology*, **17**, 316-325, 10.1016/j.ceb.2005.04.005.
61. Engebrecht,J.A., Voelkel-Meiman,K. and Roeder,G.S. (1991) Meiosis-specific RNA splicing in yeast. *Cell*, **66**, 1257-1268.
62. Zhang,Y., Liu,X.S., Liu,Q. and Wei,L. (2006) Genome-wide in silico identification and analysis of cis natural antisense transcripts (cis-NATs) in ten species. *Nucl. Acids Res.*, **34**, 3465-3475, 10.1093/nar/gkl473.
63. Gunasekera,A.M., Patankar,S., Schug,J., Eisen,G., Kissinger,J., Roos,D. and Wirth,D.F. (2004) Widespread distribution of antisense transcripts in the

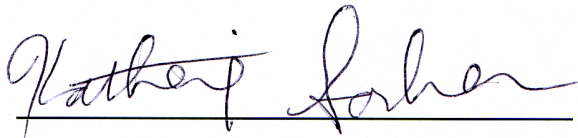
- Plasmodium falciparum genome. *Molecular and Biochemical Parasitology*, **136**, 35-42, 10.1016/j.molbiopara.2004.02.007.
64. Sun,M., Hurst,L.D., Carmichael,G.G. and Chen,J. (2005) Evidence for a preferential targeting of 3'-UTRs by cis-encoded natural antisense transcripts. *Nucl. Acids Res.*, **33**, 5533-5543, 10.1093/nar/gki852.
65. Shock,J.L., Fischer,K.F. and DeRisi,J.L. (2007) Whole-genome analysis of mRNA decay in Plasmodium falciparum reveals a global lengthening of mRNA half-life during the intra-erythrocytic development cycle. *Genome Biol*, **8**, R134, 10.1186/gb-2007-8-7-r134.
66. Trapnell,C., Pachter,L. and Salzberg,S.L. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105 -1111, 10.1093/bioinformatics/btp120.
67. Fouser,L.A. and Friesen,J.D. (1986) Mutations in a yeast intron demonstrate the importance of specific conserved nucleotides for the two stages of nuclear mRNA splicing. *Cell*, **45**, 81-93, 10.1016/0092-8674(86)90540-4.
68. Kim,E., Magen,A. and Ast,G. (2006) Different levels of alternative splicing among eukaryotes. *Nucleic Acids Research*, **35**, 125 -131, 10.1093/nar/gkl924.
69. Stein,L.D., Mungall,C., Shu,S., Caudy,M., Mangone,M., Day,A., Nickerson,E., Stajich,J.E., Harris,T.W., Arva,A. et al. (2002) The Generic Genome Browser: A Building Block for a Model Organism System Database. *Genome Research*, **12**, 1599-1610, 10.1101/gr.403602.

UCSF Library Release Publishing Agreement

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

Please sign the following statement:

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



Author Signature

12-15-2010

Date