University Library

MINERVA ACCESS

A gateway to Melbourne's research publications

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
Lee, VV;Judd, LM;Jex, AR;Holt, KE;Tonkin, CJ;Ralph, SA

Title:
Direct nanopore sequencing of mrna reveals landscape of transcript isoforms in apicomplexan parasites

Date:
2021-03-01

Citation:
Lee, V. V., Judd, L. M., Jex, A. R., Holt, K. E., Tonkin, C. J. & Ralph, S. A. (2021). Direct nanopore sequencing of mrna reveals landscape of transcript isoforms in apicomplexan parasites. Msystems, 6 (2), https://doi.org/10.1128/MSYSTEMS.01081-20.

Persistent Link:
https://hdl.handle.net/11343/274441

License:
CC BY

AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems®

RESEARCH ARTICLE

Direct Nanopore Sequencing of mRNA Reveals Landscape of Transcript Isoforms in Apicomplexan Parasites

V. Vern Lee,[a,b] Louise M. Judd,[c] Aaron R. Jex,[b,d] Kathryn E. Holt,[c,e] Christopher J. Tonkin,[b,f] Stuart A. Ralph[a]

[a]Department of Biochemistry and Pharmacology, Bio21 Molecular Science and Biotechnology Institute, The University of Melbourne, Melbourne, Victoria, Australia
[b]The Walter and Eliza Hall Institute of Medical Research, Parkville, Melbourne, Victoria, Australia
[c]Department of Infectious Diseases, Central Clinical School, Monash University, Melbourne, Victoria, Australia
[d]Faculty of Veterinary and Agricultural Sciences, The University of Melbourne, Parkville, Victoria, Australia
[e]London School of Hygiene and Tropical Medicine, London, United Kingdom
[f]Department of Medical Biology, The University of Melbourne, Melbourne, Victoria, Australia

ABSTRACT Alternative splicing is a widespread phenomenon in metazoans by which single genes are able to produce multiple isoforms of the gene product. However, this has been poorly characterized in apicomplexans, a major phylum of some of the most important global parasites. Efforts have been hampered by atypical transcriptomic features, such as the high AU content of *Plasmodium* RNA, but also the limitations of short-read sequencing in deciphering complex splicing events. In this study, we utilized the long read direct RNA sequencing platform developed by Oxford Nanopore Technologies to survey the alternative splicing landscape of *Toxoplasma gondii* and *Plasmodium falciparum*. We find that while native RNA sequencing has a reduced throughput, it allows us to obtain full-length or nearly full-length transcripts with comparable quantification to Illumina sequencing. By comparing these data with available gene models, we find widespread alternative splicing, particularly intron retention, in these parasites. Most of these transcripts contain premature stop codons, suggesting that in these parasites, alternative splicing represents a pathway to transcriptomic diversity, rather than expanding proteomic diversity. Moreover, alternative splicing rates are comparable between parasites, suggesting a shared splicing machinery, despite notable transcriptomic differences between the parasites. This study highlights a strategy in using long-read sequencing to understand splicing events at the whole-transcript level and has implications in the future interpretation of transcriptome sequencing studies.

IMPORTANCE We have used a novel nanopore sequencing technology to directly analyze parasite transcriptomes. The very long reads of this technology reveal the full-length genes of the parasites that cause malaria and toxoplasmosis. Gene transcripts must be processed in a process called splicing before they can be translated to protein. Our analysis reveals that these parasites very frequently only partially process their gene products, in a manner that departs dramatically from their human hosts.

KEYWORDS *Plasmodium*, RNA splicing, RNA-seq, *Toxoplasma*, nanopore, transcriptional regulation

Transcriptomic analyses have been central to insights into the biology and pathogenesis of eukaryotic pathogens. The best-characterized eukaryotic pathogen transcriptomes are those of the phylum *Apicomplexa*. This phylum includes some of the most important parasites impacting human and veterinary health, such as *Plasmodium* and *Toxoplasma*. *Plasmodium* is the causative agent of malaria, a devastating parasitic disease infecting over 200 million individuals and killing 400,000 each year (1).

Citation Lee VV, Judd LM, Jex AR, Holt KE, Tonkin CJ, Ralph SA. 2021. Direct Nanopore sequencing of mRNA reveals landscape of transcript isoforms in apicomplexan parasites. mSystems 6:e01081-20. https://doi.org/10.1128/mSystems.01081-20.

Editor Paola Flórez de Sessions, Oxford Nanopore Technologies

Ad Hoc Peer Reviewer Scott Lindner, The Pennsylvania State University

The review history of this article can be read here.

Copyright © 2021 Lee et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International license.

Address correspondence to Stuart A. Ralph, saralph@unimelb.edu.au.

Direct RNA transcriptome analysis of eukaryotic parasites in a single nanopore flow cell.

Received 19 October 2020
Accepted 6 February 2021
Published 9 March 2021

Downloaded from http://msystems.asm.org/ on May 16, 2021 by guest

Toxoplasma causes toxoplasmosis, a widespread zoonoses that primarily impacts immunocompromised, young, and pregnant individuals (2), and is thought to infect a third of the world's population (3). The pathogenesis of apicomplexan infections is intimately linked to the parasites' life cycles. The life cycle of most parasitic apicomplexans is complex, involving multiple differentiated forms and hosts, and this requires reprograming of the parasite transcriptome.

Early transcriptomic experiments sought to utilize techniques such as microarrays and Sanger sequencing of complementary DNA (cDNA) or expressed sequence tag (EST) libraries to understand changes in gene expression that define the pathogenesis of the parasites. These studies reveal that the timing of appearance and abundance of individual mRNAs follow developmentally distinct patterns (4), even for many predicted housekeeping genes. For example, the expression of the actin gene family in *Plasmodium falciparum* is developmentally tuned, with actin I primarily transcribed in asexual intraerythrocytic life stages, while actin II is primarily present in sexual-stage parasites (5, 6). Unusually, however, there is a poor correlation between protein and mRNA expression profiles for many genes in parasitic apicomplexans (7, 8). In one experiment, Foth et al. found widespread discrepancies between temporal expression patterns of proteins and transcripts in *P. falciparum* (9). Such discrepancies suggest that substantial posttranscriptional regulation occurs within these parasites. Indeed, with the advent of transcriptome/RNA sequencing (RNA-seq), more recent studies now show that multiple layers of gene expression regulation are required for parasite life progression, through transcriptional, posttranscriptional, and epigenetic control mechanisms (10–12).

RNA splicing provides one such source of co- and posttranscriptional regulation. In this process, introns are removed from the pre-mRNA and the exons retained to form one contiguous molecule that is then translated by the ribosome. However, for complex mRNAs, alternative splicing either of untranslated regions or the exonic chain can add additional complexity. Through this process, pre-mRNA species can be differentially spliced to create multiple distinct mature mRNAs from a single gene. This can alter regulation of the gene, e.g., by removing small-RNA binding sites (13), or diversify the proteome, as individual genes may encode multiple protein isoforms with altered structure or function (14). Indeed, proteomic analyses have revealed widespread protein isoforms arising from single genes, corresponding with various activity, stability, localization, and posttranslational modifications (15, 16). With advances in genome and transcript sequencing, it has become apparent over the last decade that alternative splicing of pre-mRNA occurs to a great extent. For example, more than 95% of human genes are alternatively spliced, and many transcript isoforms are specific to tissues or cellular states (17). Such observations suggest that RNA diversity is more complex than previously appreciated (18).

Although alternative splicing appears to play a major (though debated) role in posttranscriptional control in metazoans, the process is less understood in apicomplexans. Studies have identified apicomplexan genes with crucial alternative splicing outcomes (19). For example, alternative splicing is required for attaching a protein trafficking presequence onto two adjacent gene coding sequences (20), and normal multiorganellar targeting of the *P. falciparum* cysteinyl tRNA synthetase, which is essential for parasite survival (21). Nonetheless, there are few other studies of alternative splicing in this phylum. Understanding the diversity of parasites transcripts is crucial for drug and vaccine development because certain putative target genes may produce isoforms that escape the intervention. This has been postulated for the *P. falciparum* chloroquine resistance transporter gene (*Pf*CRT) in clinical isolates, though the role of the splice variants remains unclear (22). In other organisms, there is some evidence showing that essential genes are more likely to have alternatively spliced transcripts compared to nonessential genes (23, 24). This has not been explored in apicomplexans but highlights further considerations for investigating drug targets and interventions.

The lack of data for apicomplexan gene isoforms is a major obstacle to dissecting

Downloaded from http://msystems.asm.org/ on May 16, 2021 by guest

the complexity of transcript outcomes. Traditionally, transcriptomic studies employing RNA-seq have relied on short-read technologies such as Illumina, 454, and Ion-Torrent (25). Despite the power of very high sequencing depth and low error rates, the short reads present a limitation in that simultaneously occurring alternative-splicing events within individual transcripts cannot be unambiguously detected or linked. Previously developed computational methods for full-length transcript assembly from short read sequencing data are often computationally intensive and can produce ambiguous or conflicting results between different algorithms (26). In addition, sequencing on cDNA strands amplified by PCR has a propensity to introduce biases in relative transcript abundances and rare isoform identification (27). Hence, it is difficult to draw functional relationships between simultaneous alternative splicing events and observable phenotypes. In apicomplexan parasites, simultaneously occurring alternative splicing events within a specific transcript isoform do occur (28). However, the studies that unearthed these transcript isoforms relied on cDNA probes and reverse transcription-PCR, and the wider extent of this phenomenon is unknown.

Recently developed third generation sequencing platforms, such as those developed by Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio), are capable of producing significantly longer reads at the single-molecule level. These technologies have been used in various applications such as resolving genomic and transcriptional landscapes (29, 30), single-cell transcriptome sequencing (31), and DNA or RNA methylation pattern profiling (32–34). PacBio has recently been used to generate an amplification-free transcriptome from *P. falciparum* cDNA, which has helped to elucidate transcriptional start sites and to improve annotation of the 5′ and 3′ untranslated regions (UTRs) (35). Unlike most other sequencing platforms, a notable characteristic of ONT sequencing is the ability to directly sequence native RNA (36). With this methodology, each read represents a complete molecular transcript, which could thus significantly resolve weaknesses of amplification-based RNA-seq. In particular, each spliced isoform need only be counted as individual reads, as opposed to complex assignment and assembly of multiple spliced reads. Furthermore, due to differences between DNA and RNA molecules, contaminating DNA sequences cannot be correctly base called after sequencing and so are easily discarded (37). Recently, several studies have successfully applied single molecule, long-read sequencing to identify a high number of novel transcript isoforms (29, 38, 39). However, these studies have also identified several caveats, including a reduced throughput and high error rates.

In this study, we evaluate the ability ONT direct RNA sequencing to characterize the alternative splicing landscape of two parasitic apicomplexans, *T. gondii* and *P. falciparum*. Our analyses show that alternative splicing, particularly intron retention, is extensive throughout the transcriptome, with most multi-exon genes having some degree of intron retention and some genes only rarely producing transcripts with all introns removed. The long reads produced from ONT sequencing showed that most of these alternative splicing events are likely nonproductive in protein-coding capacity but may provide an additional layer of gene expression regulation.

(This article was submitted to an online preprint archive [40].)

RESULTS

Direct RNA sequencing of *T. gondii* and *P. falciparum* allows the detection of full-length transcripts. We generated ONT sequencing reads of poly(A)-selected RNA from asynchronous *T. gondii* (Prugniaud/Pru) tachyzoites and *P. falciparum* (3D7) mixed asexual intraerythrocytic-stage parasites. The choice to assay mixed-stage parasites, rather than specific, synchronized stages was dictated by the requirement for significantly more mRNA material than that required for short-read sequencing and the additional loss of yield when purifying full-length poly(A) mRNA. This is a limitation of this approach and means that we achieve a broader survey of transcript diversity present in asexual-stage replicative forms at the expense of stage-specific information. For *T. gondii*, we obtained a total of 310,813 reads corresponding to about 500 million bases (Mb). For *P. falciparum*, we obtained a total of 456,098 reads, corresponding to about

Downloaded from http://msystems.asm.org/ on May 16, 2021 by guest

300 Mb of data. Although the *P. falciparum* sample yielded fewer sequenced bases, we estimate the theoretical gene coverage for both parasite samples to be similar at 25- to 26-fold due to differences in gene number and length. Using minimap2 (41), we successfully mapped 78.90% of the *T. gondii* reads and 44.48% of the *P. falciparum* reads to the parasite genomes. We analyzed the quality of the sequencing reads using FASTQC and found consistently high-quality scores over the length of reads, with no drop-off in quality even at reads of 10 kb (Fig. 1B). This is important because base quality scores generally correlate with read accuracy to the reference sequence. That is indeed the case for our data set (see Fig. S1A in the supplemental material). A few IGV snapshots of mapped reads are shown in Fig. S1B. We used AlignQC to estimate the base-call error rate of the transcript reads based on aligned segments and found, on average, an error rate of 19 to 20% for both parasites, although these numbers rely in part on the accuracy of the reference sequences. The read-length distributions of the mapped and unmapped data (Fig. 1A) show that the mapped reads are predominantly longer than the unmapped reads, with some read lengths exceeding 10 kb. As expected, there was a sharp increase in unmapped read counts at an ~1.35-kb length (Fig. 1A) corresponding to yeast enolase 2, a calibration standard added during the library preparation. We calculated the average mapped read lengths to be ~1.9 and ~1.3 kb for *T. gondii* and *P. falciparum*, respectively, values well within previous range estimates of predicted transcript lengths for both organisms (42).

To evaluate the capability of ONT sequencing to generate full-length transcript reads, we remapped the reads to the parasites' annotated transcriptome and calculated the fraction of full-length transcripts per gene. We define full-length transcripts as reads that cover more than 95% of the predicted canonical transcript based on the annotation file and only considered genes with at least three mapped reads. As illustrated in Fig. 1C, many transcripts were observed to have full-length reads, particularly for the *P. falciparum* data. In *T. gondii*, 1,117 genes have 75% or more of their corresponding reads that were considered full length. In *P. falciparum*, this number is 1835 genes. The difference can be attributed to the absence of 5′ and 3′ UTR annotations for *P. falciparum*, which results in the underestimation of predicted transcript lengths. In both cases, the fraction of mapped reads corresponding to full-length transcripts fell with increasing transcript length, independent of expression levels, which is consistent with read truncation disproportionately affecting the longest reads. To better understand the overall distribution of transcript coverage, we calculated the average coverage of the transcripts per gene. Again, there is a general decrease in transcript coverage with increasing transcript length (Fig. 1C). However, many of the genes retain a high level of transcript coverage, even when there is a low fraction of full-length reads. For example, in *T. gondii*, genes with predicted transcript lengths of 3 kb or longer only had an estimated 12% of their reads that were considered full length on average, even though the average read coverage for those genes is 50.64%. The overall average transcript coverage is calculated to be more than 60% in both parasites (*T. gondii*, 65.02%; *P. falciparum*, 80.51%). It should be noted that the relatively high number of genes with low expression (<10 reads) results in the overrepresentation of certain decimal values (observed as horizontal stripes on Fig. 1C). Only considering highly expressed genes improves coverage, but not significantly so (<5%). Together, the data indicate a generally high proportion of full-length or nearly full-length transcript reads.

ONT sequencing is comparable with traditional RNA-seq for quantifying gene expression levels. To investigate the utility of ONT data to measure transcript abundance, we computed read count correlations between our ONT data set and reanalyzed published Illumina-based RNA-seq data sets on comparable parasite samples. In theory, ONT RNA sequencing reads directly correspond to complete transcripts, so quantifying the expression of genes can be done by simple counting of the assigned reads. This is dissimilar to traditional short read RNA-Seq, which necessitates a further normalization step (e.g., reads or fragments per kilobase of transcript or transcripts per million) to account for the higher number of reads that would be generated from longer transcripts. For *T. gondii*, we used Illumina data sets from a closely related strain

Downloaded from http://msystems.asm.org/ on May 16, 2021 by guest

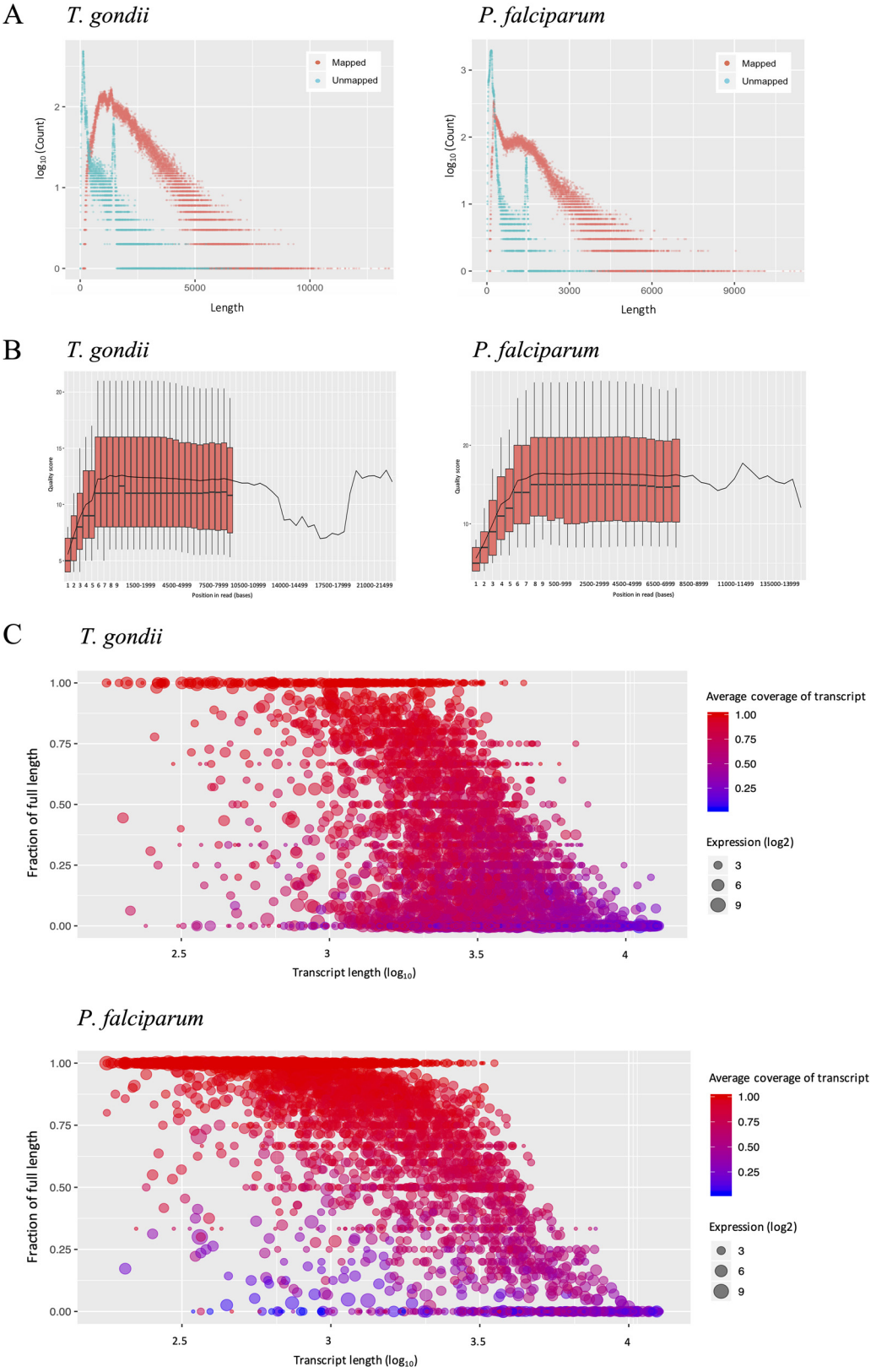


FIG 1 Summary of the ONT direct RNA sequencing data from *T. gondii* and *P. falciparum*. (A) Scatterplot of read length distribution of mapped (red) and unmapped (blue) reads. (B) Boxplot of quality scores across all bases at each position of the mapped sequencing reads. Boxes signify interquartile ranges (25 to 75%), and whiskers represent the 10 and 90% points. The continuous line represents mean quality. (C) Bubble scatter heat plots of the fraction of full-length transcripts against transcript length. Size and color denote expression and average coverage, respectively.

Downloaded from http://msystems.asm.org/ on May 16, 2021 by guest

(ME49) because it has the highest coverage replicated transcriptome data sets and is the closest well characterized strain (<0.01% genetic variation) to Prugniaud (43). Strikingly, we observe strong positive correlations between the ONT and Illumina data sets (Fig. 2), regardless of mapping to the transcriptome or genome (Spearman's rho = 0.81 for transcriptome and 0.87 for genome). For *P. falciparum*, we correlated the mixed-stage ONT data set with individual data sets from three main developmental stages (rings, trophozoites, and schizonts) and a final combined data set. In all cases, we found moderately positive correlations between the data sets. As expected, the correlation is higher in the later stages (see Fig. S2A; Spearman's rho = 0.47 for rings, 0.57 for trophozoites, and 0.63 for schizonts), and the highest with the combined data set (Fig. 2; Spearman's rho = 0.64 for transcriptome and 0.68 for genome), a reflection of mRNA abundance in these different stages.

For both parasites, a higher number of gene transcripts were detected in the Illumina data sets than in the ONT data (see Table S1). We detected at least 1 Illumina read for 8,379 genes in *T. gondii* and 5,639 genes in *P. falciparum*, as opposed to 6,778/4,656 genes for ONT reads. This is expected, given the greater sequencing depth that was obtained from the Illumina runs. The theoretical average fold coverages from our Illumina data are estimated to be 75 and 1,000 for *T. gondii* and *P. falciparum*, respectively (compared to 25- to 26-fold for the ONT data). Most of the genes that had detectable Illumina but not ONT reads were at the lower end of the expression range (see Fig. S2B). Notably however, abundant non-mRNA species, particular rRNA, were detected in the Illumina data sets. This is absent in the ONT data sets, possibly because of differences in poly(A) RNA purification methodologies and the fact that the sequencing adapters for ONT specifically ligate to poly(A) RNA. We further evaluated the ONT transcriptomes for genome coverage completeness, as shown in Fig. 2B. The read coverages from the ONT and Illumina reveal no significant bias toward a particular region of the nuclear genome. Of note, however, there appear to be high numbers of ONT reads mapping to the *P. falciparum* mitochondrion genome, but none to the apicoplast. This is expected and provides additional evidence that we are selectively capturing poly(A) RNA, since *P. falciparum* mitochondrion mRNA is typically polyadenylated (44), whereas apicoplast mRNA is not (45).

Intron retained transcripts are prevalent and generally nonproductive. A major goal of mature full-length transcript sequencing is the identification of splicing isoforms. Alternative splicing can be broadly classed into four types: intron retention, alternative 3′ splice site selection, alternative 5′ splice site selection, and exon skipping (46). Of these, intron retention is the least-studied form of alternative splicing despite the numerous studies implicating the significance of the event (47–49). This is in part due to the limitations of short-read sequencing but also the relatively long and low-complexity introns in metazoan genomes, which impose limitations on sequencing and assembly. For example, the intronic sequence in the human genome is several magnitudes longer than the length of the exonic sequence (50, 51). In contrast, the compact genomes of *Plasmodium* and *Toxoplasma* both have gene models that have similar or longer exon lengths than introns (52, 53), so reads that span multiple entire introns are quite achievable for these organisms.

To monitor levels of intron retention, we identified junctions and reads that overlapped annotated intronic regions of a gene based on the annotated coordinates using FeatureCounts (54), and we tallied the proportion of reads mapping to that intron to the total reads for the same gene. Proportion scores are represented using the metric percent intron retention (PIR). Based on this analysis, 17.65% of the mapped reads were considered to have intron overlaps for *T. gondii* versus 4.54% for *P. falciparum*. We further filtered out junctions without a minimum overlap of six bases to exclude artifacts generated by read errors and excluded genes that were not supported by a minimum coverage of three reads. The distribution of PIR scores per gene (Fig. 3A) reveals an overall skew toward low proportions of intron-overlapping reads. This is as expected given the propensity for a dominant canonical transcript (55). However, we

Downloaded from http://msystems.asm.org/ on May 16, 2021 by guest

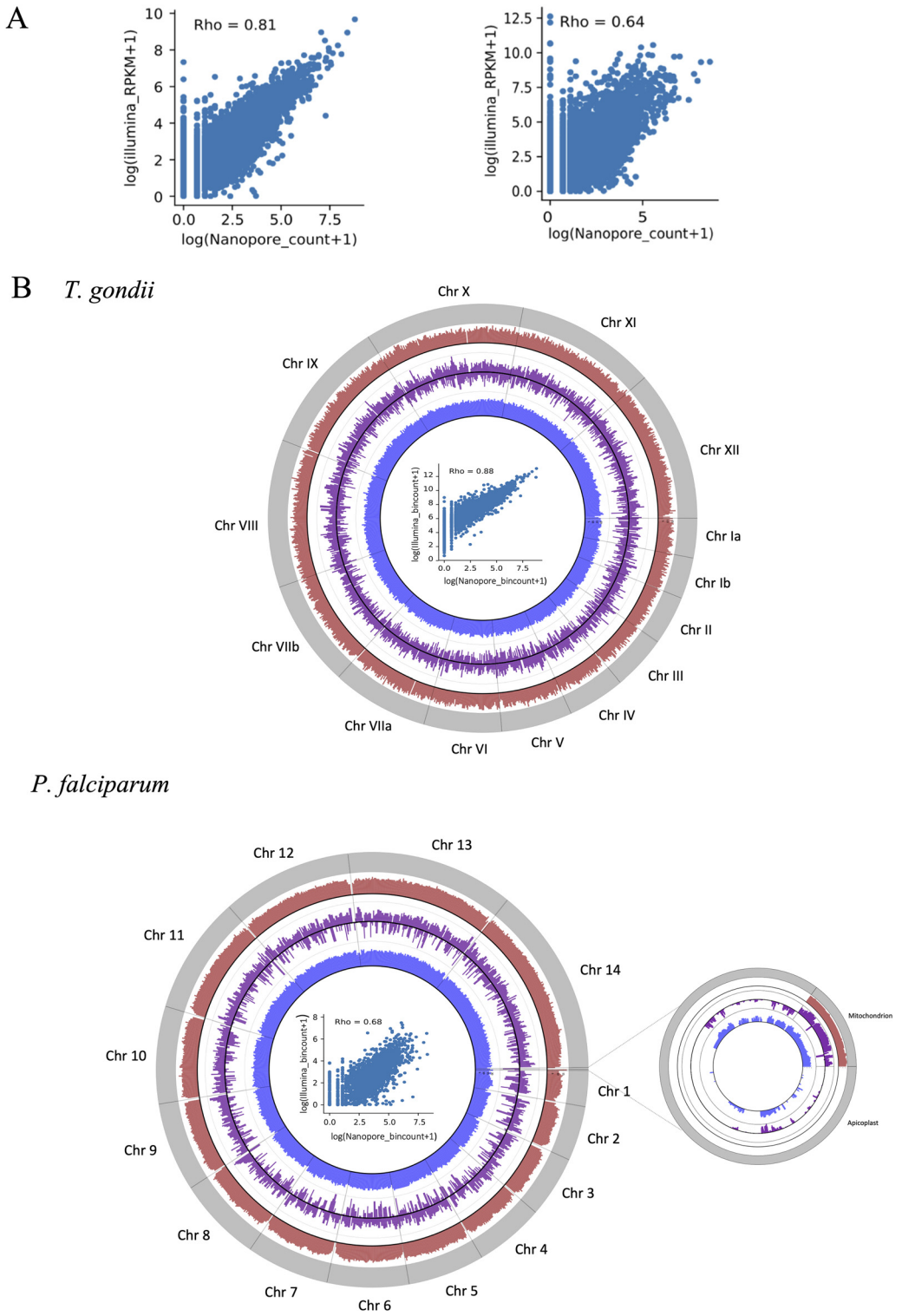


FIG 2 Comparisons between ONT direct RNA sequencing and Illumina data sets. (A) Correlation between transcriptome mapped read counts for *T. gondii* (left) and *P. falciparum* (right), presented as a scatterplot. The Spearman correlation coefficient is shown. (B) Circos plots of genome-mapped reads. The outer band (gray) represents the reference genome/chromosomes. The red and blue bands represent the genome coverage of ONT direct RNA and Illumina reads, respectively. The purple band is the $\log_2$-fold change between the two data sets (*y* axis limit = 20). The scatterplots within the Circos plots show the correlation between the genome-mapped read bin counts.

Downloaded from http://msystems.asm.org/ on May 16, 2021 by guest

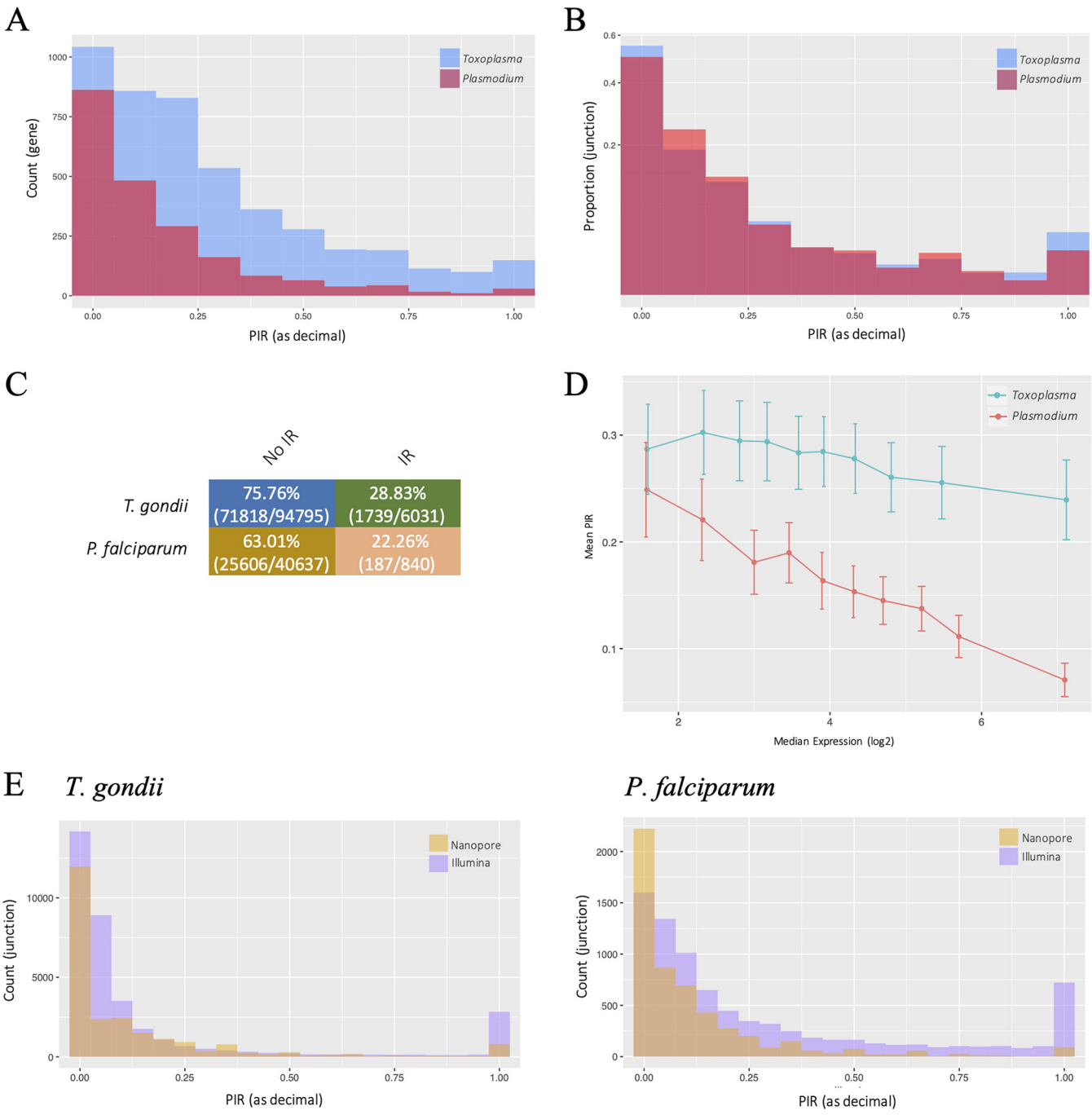


FIG 3 Analysis of intron retention using ONT direct RNA sequencing reads. Levels of intron retention are represented as the percent intron retention (PIR). (A) Distribution of intron retention levels over each gene, represented as total counts. (B) Distribution of intron retention levels over each junction, represented as the proportion of each bin count over the sum all intronic junctions. (C) Table of transcript productivity based on intron retention events as analyzed using the FLAIR pipeline. Numbers in bracket are the transcript counts (D) Relationship between levels of intron retention and gene expression. Genes were classified into 10 bins of equal number based on expression. The median expression of the genes in each bin is used to represent expression for each bin. Error bars represent the 95% confidence intervals. (E) Comparisons of intron retention quantification between ONT and Illumina data sets.

identify a strikingly high number of genes that retain a high level of intronic regions. Using a threshold of 10% PIR, we identified a total of 3,229 of 4,653 (69.40%) expressed genes for *T. gondii* and 978 of 2,090 (46.79%) expressed genes for *P. falciparum* that have intronic reads within their transcripts (see Table S2A). Moreover, for approximately 29.82% (963/3,229) of the *Toxoplasma* genes and 19.63% (192/978) of the *Plasmodium* genes, 50% or more of the reads retain at least one intron.

When we only consider full-length or near-full-length reads (as previously defined),

Downloaded from http://msystems.asm.org/ on May 16, 2021 by guest

the absolute quantification of intron retained genes is expectedly reduced, but the proportions of intron retained genes to expressed genes do not differ significantly (*T. gondii*, 1,325/1,887 [70.22%]; *P. falciparum*, 641/1,446 [44.33%]). Similarly, increasing the filtering threshold to only include genes with a minimum of 10 reads reduces absolute quantification but produces the same proportions (*T. gondii*, 2,085/2,988 [69.78%]; *P. falciparum*, 628/1,443 [43.52%]). This suggests that we are observing an inherent characteristic of the transcriptome rather than an artifact of the threshold that was used.

Unusually, there are a considerable number of genes where none of the transcripts appears to have all of their annotated introns removed (*T. gondii*, 133; *P. falciparum*, 29). We manually investigated these cases further and found major differences between the transcripts and gene model in most cases, suggesting that these highlight genes with an error in the existing gene annotation. A couple of examples are outlined in Fig. S1B. Most of these genes are annotated as hypothetical proteins, highlighting the potential of ONT sequencing to validate gene models.

The most extreme cases of conflict between the junctions we detect and canonical gene models often highlight potential annotation errors, but there are still a strikingly large number of genes where genuine introns are retained in a high (>50%) proportion of transcripts (*T. gondii*, 808 genes; *P. falciparum*, 162 genes). In addition, the differences between the two parasites are striking. In many organisms, the transcripts with the most introns are more likely to retain at least one or more introns (56). This is partially supported in our analysis, where we observed a higher level of overall intron retention for *T. gondii* (which has 4.5 introns per gene on average) than *P. falciparum* (which has only 1.5 introns per gene on average). A possible explanation for this relationship is thus that both organisms have a similar level of intron retention for any given junction, and the higher average intron number in *T. gondii* genes results in more overall intron retention per gene. To examine this, we calculated PIR scores at the individual junction level, rather than per-gene level, and we normalized the count of each PIR value to the proportion of total junctions within each organism. The analysis reveals that, after correction for the intron number, there is virtually no difference in the distribution of intron retention levels between parasites (Fig. 3B). In other words, individual *T. gondii* junctions are no more likely to experience intron retention than *P. falciparum* junctions. We further tested whether intron number was the major predictor of intron retention at the gene level by looking at the correlation between the number of introns per gene and levels of intron retention. Interestingly, we only obtained poor or moderate positive correlations in all data sets (see Fig. S3A; Spearman's rho = 0.28 for *T. gondii*, 0.48 for *P. falciparum*, and 0.39 for pooled samples). This correlation does not significantly improve even when we restrict our analysis to higher (≥10 reads) expressed genes. This suggests that whereas the intron number is associated with increased level of intron retention, it does not strongly influence whether a gene is alternatively spliced.

By taking advantage of the full-length reads made possible by ONT, we are able to predict the protein-coding productivity of the alternate transcripts. We performed productivity analysis on full-length intron-retained reads using the FLAIR (57) pipeline, which corrects and defines unproductive transcripts as transcripts with an annotated start codon and a termination codon that is 55 nucleotides or more upstream of the 3′-most splice junction. The rationale for this definition is based on previous evidence suggesting that only premature terminating transcripts following that 55-nucleotide rule mediate an effect on mRNA turnover (58). This is a conservative estimate of productivity since it does not consider intron retention within the 3′-most splice junction. The Flair method identified >70% of the intron-retained reads to be nonproductive for either parasite (Fig. 3C), suggesting that the high level of observed intron retention only rarely corresponds to alternative protein products, and that most intron-retaining transcripts may instead be targets for nonsense mediated decay. Intron retention is known to fine-tune protein expression through this pathway in mammalian systems

Downloaded from http://msystems.asm.org/ on May 16, 2021 by guest

(59). A related prediction from other studies (47) is that the most highly expressed transcripts should have low levels of intron retention. In our analysis, we do observe a negative relationship between intron retention and gene expression levels, although it is less pronounced for *T. gondii* (Fig. 3D). There is a relatively high variance for this and more for genes with lower expression levels. This is likely due to the limitation in precision for the lower sequencing depths. For example, intron retention occurring in 10% of transcripts for a given gene will not be precisely measured for a gene for which only five reads are available. We circumvented this by classifying the transcripts into bins of equal read number based on expression and quantifying global intron retention levels within each bin. Again, we observe a negative relationship between intron retention and expression levels (see Fig. S3B). To investigate the functional significance of this, we further analyzed the genes for Gene Ontology (GO) enrichment. Here, we only considered genes with a minimum coverage of 10 reads to increase precision. The analyses reveal the consistent enrichment of genes with functions associated with the ribosome when there are lower levels of intron retention across both parasites (see Table S3). This association has been previously observed in other organisms (14), although its basis is unknown. We also tested whether intron retention correlated with essentiality based on previous functional genomic screens (60, 61) and found no significant relationships (see Fig. S3C).

To validate the identification of intron retention events, we looked at whether retained introns apparent in the ONT data were directly supported by Illumina RNA-seq data. We normalized read counts by junction length and only considered intronic data that spanned the full junction. Based on the analysis, 77.88% for *T. gondii* and 87.37% for *P. falciparum* of the intronic junction reads flagged from the ONT data sets were supported by Illumina reads. However, we also noted that some alternative splicing events, particularly the lower frequency ones, failed to be captured by ONT sequencing compared to the Illumina data set (Fig. 3E). When we applied the previously used threshold, we identified 1,272 intron retention events corresponding to 856 genes in *T. gondii*, and 2,008 intron retention events corresponding to 1,117 genes in *P. falciparum*, in the Illumina data set but not in the ONT data set. This is again likely due to the limitation in read depth in the ONT data set. Considering the theoretical fold coverage of the Illumina data set is around 4 times that of the ONT data set for *T. gondii* and 40 times *for P. falciparum*, we may expect to achieve increased levels of isoform quantification with increased number of sequencing runs/flow cells, although with diminishing returns.

Based on our sequencing of poly(A) tailed material combined with the sequencing chemistry of ONT direct RNA sequencing, which in theory only captures RNA with a poly(A) tail, and previous kinetic studies on RNA processing (62, 63), we do not expect the intron retained transcripts to simply be unprocessed transcripts. To confirm this, we looked for evidence that each transcript had at least been partially processed. On average, 92.76% of multi-intron genes identified as having intron retention within their transcripts had at least one junction that was canonically spliced in all the transcripts, demonstrating that our findings cannot be attributed to sequencing of pre-mRNA. To exclude the possibility that the efficacy of poly(A) tail selection of RNA in *P. falciparum* was reduced due to the high AU content throughout the RNA, we looked at the read coverage over each gene body, scaled to 100 nucleotides. This analysis reveals a strong 3′ end bias of reads (see Fig. S4), indicating a high efficacy of poly(A) tail RNA capture.

Alternate junction splicing is often proximal and nonproductive. Having previously identified intron-retained, read junctions using an annotated gene model approach, we used RSeQC (see Materials and Methods) to identify and quantify the other three classes of alternative splicing read junctions (exon skipping and 5′- and 3′-splice site changes) based on a similar methodology. The levels of alternative spliced junctions are calculated as the proportion of alternate junction over the total junction reads and are represented using the metric percent spliced (PS) value. Here, we filtered out junctions unsupported by a minimum coverage of three reads. Using the same

Downloaded from http://msystems.asm.org/ on May 16, 2021 by guest

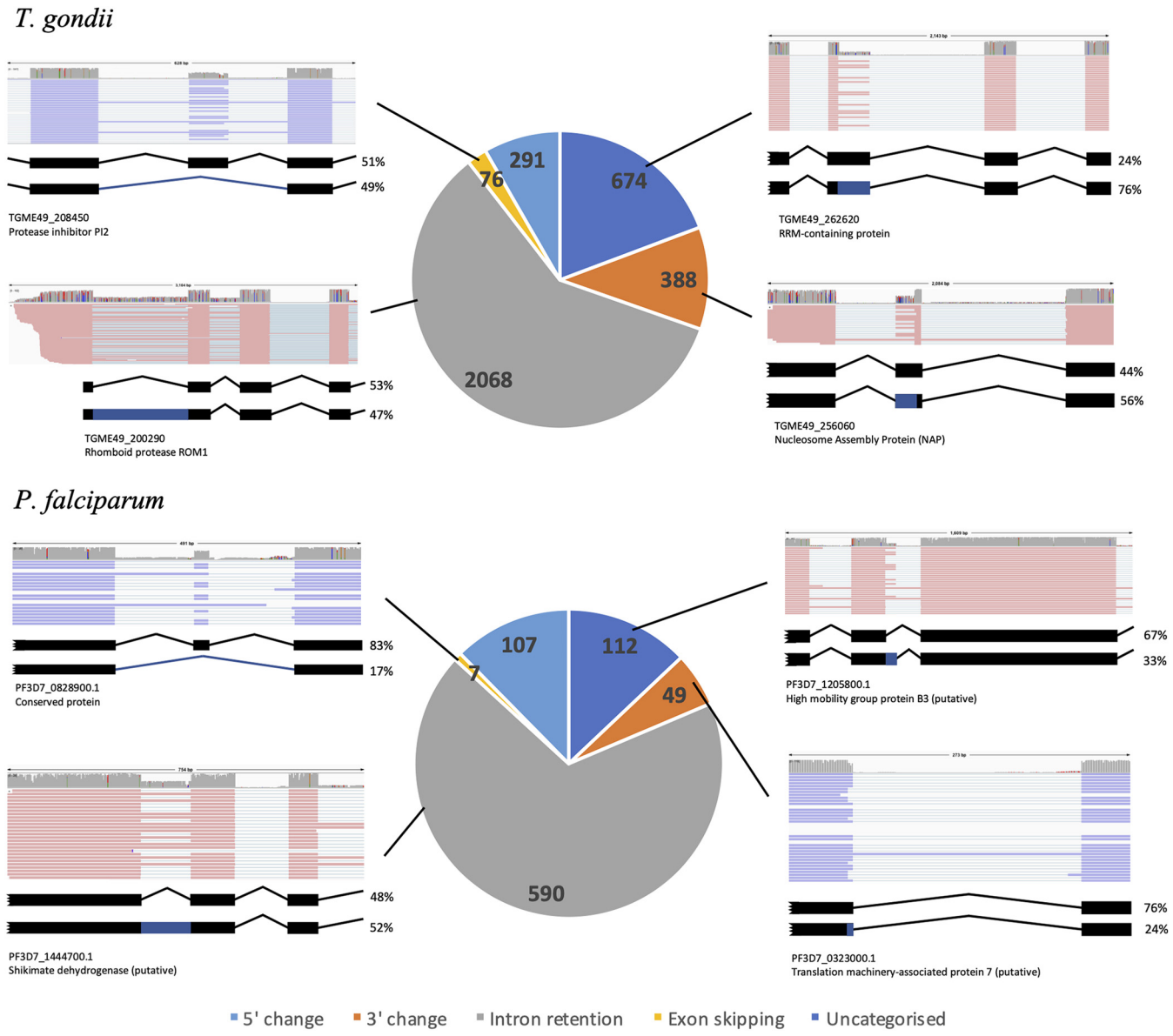


FIG 4 Summary analysis of alternative splicing from the ONT direct RNA sequencing data sets. Pie charts show the number, proportion, and categorization of genes with alternatively spliced transcripts equaling or exceeding 10% of its total transcript. An example of each event is presented. Red and blue represents sense and antisense transcripts, respectively. Uncategorized genes represent genes where there are major mismatches between the RNAseq data and the annotated gene model.

threshold as before (≥10%), we identified a total of 1,138 genes for *T. gondii* and 168 genes for *P. falciparum*, where one or more of their junctions exhibited alternative 5′/3′-splice site selection or exon skipping (see Table S2B to D). Remarkably, these aggregate numbers are lower than those we calculated for intron retention alone. In total, we identified 5,205 splicing events for *T. gondii* and 1,112 for *P. falciparum*, and yet intron retention accounts for 60 to 68% of the alternatively spliced genes identified, alternate 5′ junction and 3′ junction splicing for 13 to 19% and 6 to 11%, respectively, and exon skipping for <3% (Fig. 4). The rest of the junctions flagged in the analysis defy easy categorization due to major mismatches between the RNA-seq data and the annotated gene model. Subsets of genes were also found to have multiple alternative splicing type events within their transcripts as observed in Fig. S5, although there does not appear to be a particular functional trend.

Having identified junctions subject to alternative splicing, we then quantified what

Downloaded from http://msystems.asm.org/ on May 16, 2021 by guest

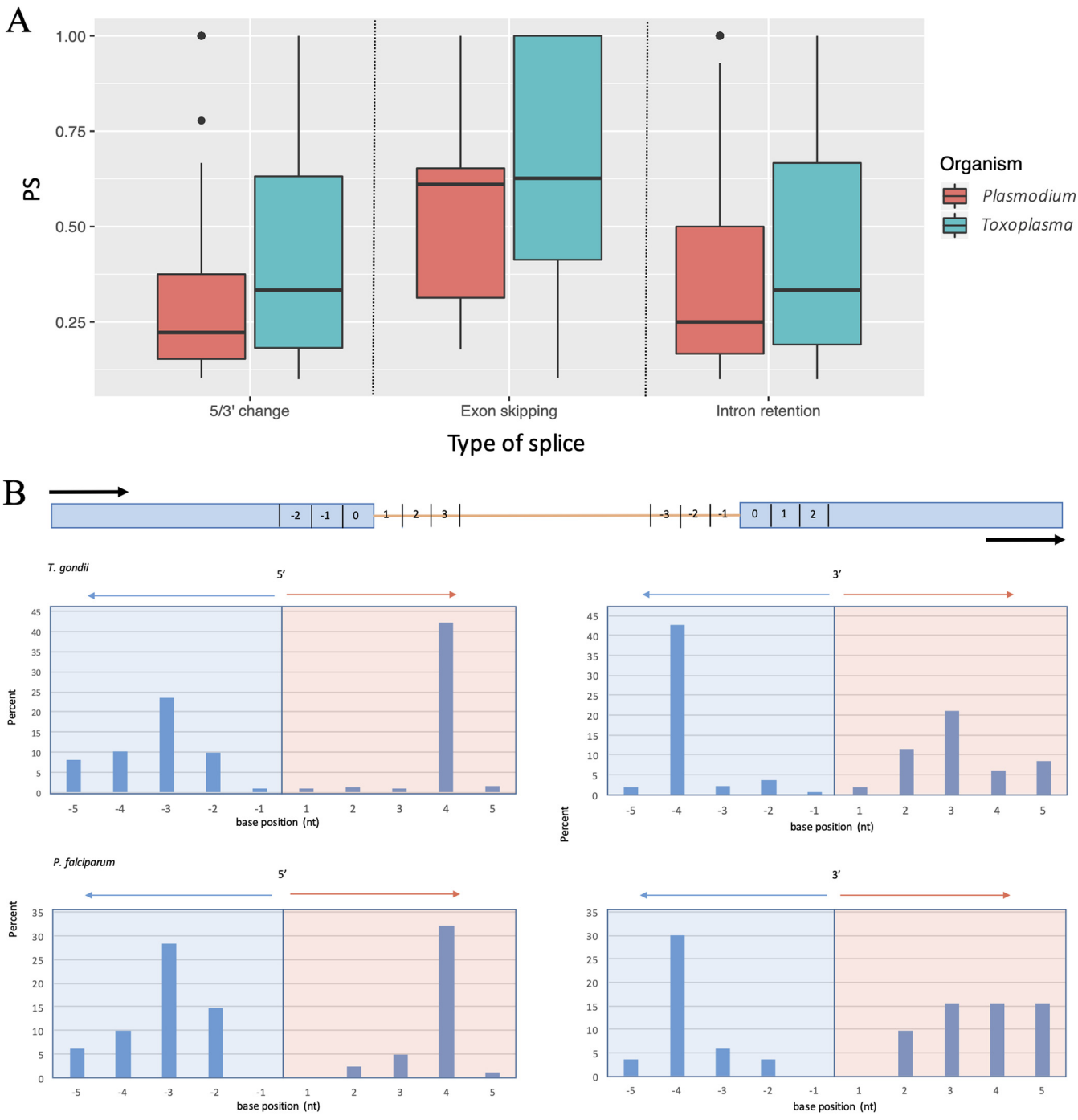


FIG 5 (A) Levels of each major alternative splice type, represented as the percent splice (PS). (B) Distribution of alternate 5′/3′ splice positions within five bases to the dominant splice site.

proportion of the transcripts produced at those junctions represented the noncanonical isoform. For alternative splicing involving 5′ or 3′ changes and for intron retention, substantial proportions of isoforms were represented by the alternative transcript, but the median abundance remained below 50% (data shown in Fig. 5A). However, while exon skipping was a relatively rare event across the genome (Fig. 4), for genes where exon skipping did occur, it represented a higher proportion (approximately two-thirds) of transcript isoforms for those genes than observed for the other forms of alternative splicing (Fig. 5A). We had previously published a list of genes with alternative splicing excluding intron retention in *T. gondii* (RH) identified using Illumina sequencing and

Downloaded from http://msystems.asm.org/ on May 16, 2021 by guest

our in-house-built computational tool JunctionJuror (64). A limitation of the tool over our current methodology is that the information about junction types is ultimately lost. Despite the differences in methodologies and parasite strain, we compared our list of genes with alternative splicing excluding intron with the previous list and found a moderately high degree of overlap (57.46%) (see Table S4). This intersect is double that is expected by chance alone and may represent genes with a steady-state alternative splicing of higher confidence. Similarly, we compared our data set with a previously published Illumina sequencing analysis which identified 310 alternative splicing events in blood-stage *P. falciparum* (65). There is an overlap of 49.75% alternative spliced genes between the data sets.

To further explore consequences of alternate 5′ or 3′ junction splicing events, we investigated the length distribution of the change in intron length. We found a surprisingly high proportion (~50%) of alternate 5′/3′ splicing to occur proximal (<6 bases) to the expected canonical site. We graphed the distribution of splicing change positions in Fig. 5B and found a substantial spike of splicing changes occurring at the position four bases inside the canonical intron boundary. We tested these junction reads for productivity for a subset of 50 of these "near-miss" alternate splice events and found all reads to prematurely terminate (unsurprisingly, given the necessary frameshift). This striking over-representation of isoforms departing from the canonical model by specifically four bases has been previously identified in metazoans (66). Very small movements in splice site usage have been described as junction wobbling, and this has been proposed as minor splicing noise, or alternatively, as an additional mechanism of regulation through the NMD pathway (66), although the reason for the specific peak of AS four bases away from the canonical junction is unknown.

DISCUSSION

Despite the apparent significance of alternative splicing in metazoans, very little is known about the process in apicomplexans. A number of targeted experiments highlight single splicing events and their impact on parasite survival, but global splicing networks have been poorly described. Untargeted RNA-seq experiments have mainly focused on whole transcriptome assembly and/or gene expression. Those studies that do monitor alternative splicing reveal its occurrence in multiple genes and stages, but the extent of these events and their phenotypic significance remains unknown. The lack of a robust methodology in defining transcript isoforms from short-read data are a particular challenge in dissecting whole-transcriptome splicing. In this study, we investigated whether ONT direct RNA sequencing could be used to explore the splicing landscape of two apicomplexan parasites: *T. gondii* and *P. falciparum*.

To our knowledge, ONT direct RNA sequencing has not been previously described in apicomplexans and so our first objective was to evaluate the capability of ONT sequencing in generating sequencing reads from these parasites. We successfully obtained high-quality sequencing reads for both parasites that were comparable to that previously described in the literature for other organisms (29, 67). In particular, we obtained read lengths that exceeded 1 kb on average, many of them predicted to represent full-length or near-full-length transcripts. Interestingly, although we obtained a higher number of sequencing reads for *P. falciparum*, the mapping of the reads was suboptimal compared to that of *T. gondii*. Repetitive DNA sequence motifs are characteristic of many large eukaryotic genomes, and this has been known to complicate the mapping of reads that cannot be confidently assigned to these particular regions (68). In theory, long-read sequencing mitigates this problem because long enough reads should unambiguously match a unique site on the genome, irrespective of low complexity or repeat sequences. However, the genome of *P. falciparum* is particularly AT-rich (~82%), with numerous regions of extreme low complexity (69). Thus, we may expect some reads, particularly the shorter ones, to fail the mapping parameters. Indeed, as indicated above, reads that fail to map to the genome tended to be shorter than reads that do. This is further exacerbated by the high error rate produced from

Downloaded from http://msystems.asm.org/ on May 16, 2021 by guest

ONT sequencing. Based on previous experiments, the per-base error rate of direct RNA sequencing using ONT is 10 to 20% (29, 57). In our data set, we estimated the error rate to be around 20%. This may have further contributed to the poorer mapping, though we do not expect the high error rate to significantly impact our study because the main analysis is focused on splice connectivity rather than on base sequences. As has happened for ONT DNA sequencing, we are likely to see significant improvements in read and mapping accuracy of RNA sequences as improvements are made to the flow cell and base caller. A study carried out by Runtuwe et al. provides an elegant example, wherein ONT DNA sequencing on targeted *P. falciparum* genes yielded a mapping percentage that improved from 57.86 to 92.46% with improved chemistry of the flow cell and upgrades to the base-calling algorithms (71).

The quantification of gene expression is one important goal of RNA-seq. Traditional, short-read sequencing requires the generation and amplification of cDNA, which can introduce artifacts and biases. Transcriptional amplification or repression is a commonly overlooked bias (72), where the levels of global mRNA, rather than specific mRNA, may vary between different samples. Thus, using a standard amount of total RNA, as is commonly done, can mask actual detection of specific mRNA levels, even after normalization (72). Direct RNA sequencing allows these caveats to be bypassed because a standard amount of isolated mRNA instead is used as the sequencing material. However, because there is no amplification step, direct RNA sequencing is limited by the amount of mRNA that can be practically obtained and used in the sequencing process. Without sufficient sequencing material, it can be difficult to achieve the high levels of sequencing depth that is needed to analyze gene expression (73). In line with the literature, our analysis shows that the current protocol for ONT direct RNA sequencing is comparable to Illumina for quantifying gene expression in the organisms we analyzed. It can be noted, however, that sequencing depth remains the main limitation of ONT sequencing in our study. The reduced throughput and sequencing depth from ONT sequencing compared to Illumina sequencing means that genes or transcript isoforms with low expression may not be captured.

Several previous analyses have reported differences in alternative splicing types and levels among different organisms (56, 74, 75). Notably, the increase in intron number and its retention correlates strongly with multicellular complexity (as defined by numbers of distinct cell types) (76, 77). In apicomplexans, the splicing machinery appears to be largely conserved, but features of gene structure such as intron number, length, and distribution can be highly variable (19). In our study, the difference in intron number between *T. gondii* and *P. falciparum* is a relevant example. Despite the differences, we found that alternative splicing for any given junction occurs at similar rates between the two parasites. This further supports the notion that the parasites share similar splicing processes. Intron number of genes is predicted to be positively associated with alternative splicing events (56, 78). This is not simply a stochastic effect but rather related to the general decrease in 5′ splice site strength with increasing intron number in many organisms (78). As described above, however, there is only a weak or moderately positive correlation between intron number and intron retention level of genes in the two parasites studied. Schmitz et al. (56) reported that other features such AT content and splice site entropy are important modulators of intron retention. This may also be the case in our study, given the observation that certain splice junctions are predisposed to retain their intron over others.

In addition to the difference in alternative splicing levels, there are differences in the composition of alternative splicing types between different organisms as reported by Kim et al. (75) and McGuire et al. (79). For example, exon skipping is the predominant form of alternative splicing in metazoans, and intron retention the rarest (75). Our analysis reveals that the opposite occurs in *T. gondii* and *P. falciparum*, with intron retention being the predominant event to occur and exon skipping the rarest. This composition of alternative splicing type is similar to that observed for plants and fungi (75, 79, 80), although the reason is unclear. More recent studies find that intron

Downloaded from http://msystems.asm.org/ on May 16, 2021 by guest

retention has been previously under detected in metazoans due to methodology limitations or confounding variables, but the high levels of exon skipping has been mostly undisputed (81). Kim et al. (75) speculated that intron retention emerged as the earliest form of alternative splicing, before other mechanisms of complex splicing events evolved. There is some evidence for this, including the apparent shift toward increased exon skipping frequencies in early branching animals (81). This is associated with the preservation of coding frames, suggesting a role of exon skipping in expanding proteome diversity (81). In contrast to this, we found that the majority of the splicing events in *P. falciparum* and *T. gondii*, particularly intron retention, results in nonproductive transcripts. Our results thus indicate that alternative splicing rarely contributes to generating diversity of protein sequence in these parasites and may relate instead to transcriptomic complexity that impacts protein abundance. If that is true, a previous analysis that showed alternative splicing to be essential for *Plasmodium*-stage differentiation (12) may possibly be explained by a requirement for modulation of abundance for specific proteins rather than generation of protein sequences.

Consistent with our observation of alternative splicing playing a minor role in generating true proteome diversity in apicomplexans, many splicing events in other eukaryotes contribute little to the protein isoform repertoire. In particular, many transcripts contain premature termination codons (PTCs), at least in humans and yeast (82, 83). Often, PTC transcripts are the result of the retention of intronic sequences that contain PTCs (84), but translational frameshifts from active splicing events such as alternative splice site selections have been similarly implicated (83, 85). PTC transcripts are not normally translated but rather targeted for degradation through the nonsense-mediated decay (NMD) pathway (86, 87). This is vital because the transcripts encode altered or truncated proteins which may exhibit deleterious activity (88). Some studies postulate that the predicted alternative splice events are the result of either experimental or transcriptional noise (89) or that a substantial portion of such transcripts are contaminating pre-mRNA molecules and so do not represent true alternative splicing (59, 90). Nevertheless, many RNA-seq-based analyses operate on the assumption that PTC transcripts are biologically significant or relevant (91). Congruously, studies focusing on mature mRNA isoforms in other organisms suggest that nonproductive transcripts mediate an additional layer of posttranscriptional regulation, through downstream RNA processing changes such as mRNA turnover, export, and microRNA silencing (47, 92, 93). Alternative splicing in apicomplexans may also play a role in these processes. Strikingly, unique PTC transcript signatures are associated with distinct cell lineages (48, 94, 95) in multicellular eukaryotes, which may be analogous to the essential role of alternative splicing observed in stage differentiation observed in *Plasmodium* (12).

Nonproductive transcripts are typically degraded though the NMD pathway, and this has been shown to regulate gene expression at the posttranscriptional level (96). However, it is difficult to conclusively define the function of the nonproductive transcripts without experimentally testing the proteomic fates of these transcripts. In metazoans, nonproductive transcripts often highlight genes that were downregulated following a transition of cellular states (95). Our study is consistent with this, given the observation that the number of nonproductive transcripts generally decreased with increasing transcript number. This, in association with NMD, has been shown to be crucial to the maintenance and differentiation of many cell types (97, 98). In contrast, in organisms such *Paramecium tetraurelia*, nonproductive transcripts appear to be the result of splicing error rather than function (99). Regardless, because gene expression as measured by transcript levels do not necessarily translate to protein expression levels, our findings have potential implications for the interpretation of RNA-seq studies in these parasites. Several studies have already demonstrated the poor correlation between protein and mRNA expression profiles in apicomplexans (7, 9, 100). Our results highlight that for many genes raw quantifications of transcript abundance will correlate poorly with the number of copies of productive isoforms and provide one source of mismatch between transcriptional initiation and protein abundance.

Downloaded from http://msystems.asm.org/ on May 16, 2021 by guest

Genome annotation is a crucial element of RNA-seq data analysis. For *T. gondii* and *P. falciparum*, the task is a widely accomplished manual effort from experts in the research community. Although genome annotation was not the main focus of the study, the ONT data sets are able to reveal the structure of full-length transcripts. This is crucial in validating gene models. Our data are viewable through the ToxoDB and PlasmoDB web resources (101), and raw data are available at the Sequence Read Archive (SRA; PRJNA606986), which may aid the research community to further curate and validate the current annotations.

Conclusions. In this study, we have performed the first direct transcriptomic analyses on *T. gondii* and *P. falciparum*. We show that ONT direct RNA sequencing enables the quantification of gene expression despite a reduced throughput. In combination with the increased requirement for starting material, this means that the cost and time per base pair sequenced remains higher than that of second-generation sequencing platforms. Nevertheless, because ONT direct RNA sequencing enables the detection of full-length transcripts without amplification, the tool remains promising for resolving the limitations of second-generation sequencing.

We demonstrated that alternative splicing is widespread in the two parasites, particularly intron retention. ONT direct RNA sequencing enabled us to determine the productivity of these transcripts without complex computational methodologies, and we show that most the transcripts are premature terminating. This has implications for the quantification of gene expression, since it is highly unlikely for the wealth of transcript diversity that we identified to directly translate to protein isoforms.

MATERIALS AND METHODS

Cell culture and RNA extraction. *Toxoplasma gondii* tachyzoites (Pru Δ*ku80*) were cultured on human foreskin fibroblasts in Dulbecco modified Eagle medium supplemented with 1% (vol/vol) fetal calf serum and 1% (vol/vol) GlutaMAX. Freshly egressed tachyzoites were washed, filter purified (5 μm), and collected for RNA extraction. *P. falciparum* (3D7) were cultured in complete media consisting of human erythrocytes (O+, 2% hematocrit), RPMI-HEPES, 5% (wt/vol) Albumax, and 3.6% (wt/vol) sodium bicarbonate. We collected mixed-stage parasite, purified from host red blood cells via lysis with 0.05% (wt/vol) saponin, for RNA extraction. To obtain the 500 ng of mRNA recommended for the library preparation, we used TRI Reagent (Sigma) for extraction of total RNA followed by the Dynabeads mRNA purification kit for polyadenylated [poly(A)] mRNA (Thermo Scientific) purification according to the manufacturer's protocol. Purity and quantification of mRNA were determined via NanoDrop (Thermo Scientific) and a Qubit RNA HS assay kit (Thermo Scientific).

Library preparation and Nanopore sequencing. Libraries for the direct RNA sequencing were generated using the recommended protocol for the SQK-RNA001 kit (Oxford Nanopore Technologies). We loaded and sequenced the libraries on MinION R9.4 flow cells (Oxford Nanopore Technologies) for 48 h. Base calling was performed concurrent with sequencing using Albacore (v 2.0), which was integrated within the MinION software (MinKNOW, v1.10.23). Only "pass" reads as designated by the tool were used for subsequent analyses.

Mapping and qualitative analysis. ONT sequencing data were first checked for quality with FastQC (v.0.11.7) (102). We then utilized Minimap2 (v2.1) (41) to map raw reads to the parasite genome and transcriptome from ToxoDB and PlasmoDB (r. 39), using the recommended preset commands except that intron length thresholds were set at 5,000 and 1,500 bases for *T. gondii* and *P. falciparum*, respectively. Previously published Illumina data sets (SRR350746, ERR174301, ERR185969, ERR185970, and ERR185971) (103, 104) were mapped using HISAT2 (105) using the preset commands. We checked for mapping quality with SAMtools (v.1.7) (106), Picard (v.2.18.2) (107), and AlignQC (v.1.2) (108). Further postprocessing of data was performed using the command-line interface, and graphical elements were drawn using ggplot (109) on RStudio. We verified and illustrated subsets of mapped reads via IGV (110).

We correlated the ONT sequencing data with the Illumina data sets as previously described (36) using the wub package (v.0.2) (111). The genome coverage of sequencing data sets were generated using bedtools genome coverage (v2.27) (112) and visualized via J-Circos (v1) (113). Log_2-fold ratios were calculated using deepTools bamCompare (v.2.5.1) (114). Gene body coverage was analyzed with the geneBody_coverage.py script of RSeQC (v.2.6.4) (115) using the default settings. For additional details and command lines used in these analyses, see Text S1 in the supplemental material.

Alternative splicing analysis. We applied two approaches to analyzing alternative splicing. We first identified intron retained junctions and transcripts using featureCounts (v.1.6.2) (54) on the genome mapped reads. featureCounts matches features specified in an annotation file (gff) to mapped reads. The annotation files used in the analyses were obtained from ToxoDB and PlasmoDB (r.39), and preprocessed via ToolShed (v.1.0) (116) to specifically extract intron coordinates and gene IDs. We set a minimum threshold requiring mapping to at least six bases of the intron feature and a minimum threshold of three reads mapping to the junction/transcript to be considered for further analysis. PIR scores were calculated as the proportion of alternative splicing events to the sum of reads for each junction/gene.

Downloaded from http://msystems.asm.org/ on May 16, 2021 by guest

Productivity of full-length transcripts was analyzed using the Flair pipeline (57) using default parameters.

For the second approach, we used the junction_annotation.py script of RSeQC (v.2.6.4) (115) to identify novel or partial-novel junctions from the genome mapped reads based on the unmodified annotation file. Again, we filtered out junctions that had fewer than three supporting reads. The junctions were summarized into a table based on coordinate matching to the 5′ and/or 3′ of the expected canonical junction. We identified alternative 5′/3′ splicing and exon skipping based on the coordinates and strandedness of junctions identified by RSeQC that were either consistent with or conflicted with the annotated junctions. We manually validated the data, matched junctions to available gene IDs, and again calculated the proportion of alternative splicing events to the sum of reads for each junction. Using the final data set, we re-curated the list of intron-retained junctions to exclude for alternate 5′/3′ splice changes. Proportional Venn diagrams were drawn using BioVenn (117).

Gene set enrichment analyses were carried out by ranking the genes based on their alternative splicing levels and using the first and third quartiles of the ranked list as input for GO enrichment analysis via ToxoDB/PlasmoDB based on curated and computed assigned associations. We required the adjusted *P* value to be <0.05 and the false discovery rate q-value to be <0.25. This approach was validated using GSEA via WebGestalt (118).

Data availability. Our data are viewable through the ToxoDB and PlasmoDB web resources (101), and raw data are available from the SRA (PRJNA606986).

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

TEXT S1, DOCX file, 0.02 MB.
FIG S1, TIF file, 2.5 MB.
FIG S2, TIF file, 2.3 MB.
FIG S3, TIF file, 2.6 MB.
FIG S4, TIF file, 1 MB.
FIG S5, TIF file, 0.9 MB.
TABLE S1, XLSX file, 0.6 MB.
TABLE S2, XLSX file, 0.7 MB.
TABLE S3, XLSX file, 0.1 MB.
TABLE S4, XLSX file, 0.1 MB.

ACKNOWLEDGMENTS

This research was supported by a grant from the Australian Research Council (DP160100389) and the Australian National Health and Medical Research Council (Project Grant 1165354). The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

REFERENCES

1. World Health Organization. 2019. World malaria report 2019. World Health Organization, Geneva, Switzerland.
2. Torgerson PR, Mastroiacovo P. 2013. The global burden of congenital toxoplasmosis: a systematic review. Bull World Health Organ 91:501–508. https://doi.org/10.2471/BLT.12.111732.
3. Furtado JM, Smith JR, Belfort R, Gattey D, Winthrop KL. 2011. Toxoplasmosis: a global threat. J Glob Infect Dis 3:281–284. https://doi.org/10.4103/0974-777X.83536.
4. Horrocks P, Wong E, Russell K, Emes RD. 2009. Control of gene expression in *Plasmodium falciparum*: ten years on. Mol Biochem Parasitol 164:9–25. https://doi.org/10.1016/j.molbiopara.2008.11.010.
5. Wesseling JG, Snijders PJ, van Someren P, Jansen J, Smits MA, Schoenmakers JG. 1989. Stage-specific expression and genomic organization of the actin genes of the malaria parasite *Plasmodium falciparum*. Mol Biochem Parasitol 35:167–176. https://doi.org/10.1016/0166-6851(89)90119-9.
6. Lopez-Barragan MJ, Lemieux J, Quinones M, Williamson KC, Molina-Cruz A, Cui K, Barillas-Mury C, Zhao K, Su XZ. 2011. Directional gene expression and antisense transcripts in sexual and asexual stages of *Plasmodium falciparum*. BMC Genomics 12:587. https://doi.org/10.1186/1471-2164-12-587.
7. Le Roch KG, Johnson JR, Florens L, Zhou Y, Santrosyan A, Grainger M, Yan SF, Williamson KC, Holder AA, Carucci DJ, Yates JR, III, Winzeler EA. 2004. Global analysis of transcript and protein levels across the *Plasmodium falciparum* life cycle. Genome Res 14:2308–2318. https://doi.org/10.1101/gr.2523904.
8. Wastling JM, Xia D, Sohal A, Chaussepied M, Pain A, Langsley G. 2009. Proteomes and transcriptomes of the *Apicomplexa*: where's the message? Int J Parasitol 39:135–143. https://doi.org/10.1016/j.ijpara.2008.10.003.
9. Foth BJ, Zhang N, Mok S, Preiser PR, Bozdech Z. 2008. Quantitative protein expression profiling reveals extensive posttranscriptional regulation and posttranslational modifications in schizont-stage malaria parasites. Genome Biol 9:R177. https://doi.org/10.1186/gb-2008-9-12-r177.
10. Toenhake CG, Bártfai R. 2019. What functional genomics has taught us about transcriptional regulation in malaria parasites. Brief Funct Genomics 18:290–301. https://doi.org/10.1093/bfgp/elz004.
11. Llora-Batlle O, Tinto-Font E, Cortes A. 2019. Transcriptional variation in malaria parasites: why and how. Brief Funct Genomics 18:329–341. https://doi.org/10.1093/bfgp/elz009.
12. Yeoh LM, Goodman CD, Mollard V, McHugh E, Lee VV, Sturm A, Cozijnsen A, McFadden GI, Ralph SA. 2019. Alternative splicing is required for stage differentiation in malaria parasites. Genome Biol 20:151. https://doi.org/10.1186/s13059-019-1756-6.
13. Mockenhaupt S, Makeyev EV. 2015. Non-coding functions of alternative pre-mRNA splicing in development. Semin Cell Dev Biol 47-48:32–39. https://doi.org/10.1016/j.semcdb.2015.10.018.

Downloaded from http://msystems.asm.org/ on May 16, 2021 by guest

14. Neverov AD, Artamonova II, Nurtdinov RN, Frishman D, Gelfand MS, Mironov AA. 2005. Alternative splicing and protein function. BMC Bioinformatics 6:266. https://doi.org/10.1186/1471-2105-6-266.
15. Kelemen O, Convertini P, Zhang Z, Wen Y, Shen M, Falaleeva M, Stamm S. 2013. Function of alternative splicing. Gene 514:1–30. https://doi.org/10.1016/j.gene.2012.07.083.
16. Birzele F, Csaba G, Zimmer R. 2008. Alternative splicing and protein structure evolution. Nucleic Acids Res 36:550–558. https://doi.org/10.1093/nar/gkm1054.
17. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. 2008. Alternative isoform regulation in human tissue transcriptomes. Nature 456:470–476. https://doi.org/10.1038/nature07509.
18. Licatalosi DD, Darnell RB. 2010. RNA processing and its regulation: global insights into biological networks. Nat Rev Genet 11:75–87. https://doi.org/10.1038/nrg2673.
19. Yeoh LM, Lee VV, McFadden GI, Ralph SA. 2019. Alternative splicing in apicomplexan parasites. mBio 10:e02866-18. https://doi.org/10.1128/mBio.02866-18.
20. van Dooren GG, Su V, D'Ombrain MC, McFadden GI. 2002. Processing of an apicoplast leader sequence in *Plasmodium falciparum*, and the identification of a putative leader cleavage enzyme. J Biol Chem 277:23612–23619. https://doi.org/10.1074/jbc.M201748200.
21. Pham JS, Sakaguchi R, Yeoh LM, De Silva NS, McFadden GI, Hou YM, Ralph SA. 2014. A dual-targeted aminoacyl-tRNA synthetase in *Plasmodium falciparum* charges cytosolic and apicoplast tRNACys. Biochem J 458:513–523. https://doi.org/10.1042/BJ20131451.
22. Gadalla NB, Malmberg M, Adam I, Oguike MC, Beshir K, Elzaki S-E, Mukhtar I, Gadalla AA, Warhurst DC, Ngasala B, Mårtensson A, El-Sayed BB, Gil JP, Sutherland CJ. 2015. Alternatively spliced transcripts and novel pseudogenes of the *Plasmodium falciparum* resistance-associated locus *pfcrt* detected in East African malaria patients. J Antimicrob Chemother 70:116–123. https://doi.org/10.1093/jac/dku358.
23. Jex AR, Nejsum P, Schwarz EM, Hu L, Young ND, Hall RS, Korhonen PK, Liao S, Thamsborg S, Xia J, Xu P, Wang S, Scheerlinck J-PY, Hofmann A, Sternberg PW, Wang J, Gasser RB. 2014. Genome and transcriptome of the porcine whipworm *Trichuris suis*. Nat Genet 46:701–706. https://doi.org/10.1038/ng.3012.
24. Lees JG, Ranea JA, Orengo CA. 2015. Identifying and characterising key alternative splicing events in *Drosophila* development. BMC Genomics 16:608–608. https://doi.org/10.1186/s12864-015-1674-2.
25. Goodwin S, McPherson JD, McCombie WR. 2016. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 17:333–351. https://doi.org/10.1038/nrg.2016.49.
26. Salzberg SL, Phillippy AM, Zimin A, Puiu D, Magoc T, Koren S, Treangen TJ, Schatz MC, Delcher AL, Roberts M, Marçais G, Pop M, Yorke JA. 2012. GAGE: a critical evaluation of genome assemblies and assembly algorithms. Genome Res 22:557–567. https://doi.org/10.1101/gr.131383.111.
27. Kozarewa I, Ning Z, Quail MA, Sanders MJ, Berriman M, Turner DJ. 2009. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. Nat Methods 6:291–295. https://doi.org/10.1038/nmeth.1311.
28. Chaudhary K, Donald RG, Nishi M, Carter D, Ullman B, Roos DS. 2005. Differential localization of alternatively spliced hypoxanthine-xanthine-guanine phosphoribosyltransferase isoforms in *Toxoplasma gondii*. J Biol Chem 280:22053–22059. https://doi.org/10.1074/jbc.M503178200.
29. Jenjaroenpun P, Wongsurawat T, Pereira R, Patumcharoenpol P, Ussery DW, Nielsen J, Nookaew I. 2018. Complete genomic and transcriptional landscape analysis using third-generation sequencing: a case study of *Saccharomyces cerevisiae* CEN.PK113-7D. Nucleic Acids Res 46:e38. https://doi.org/10.1093/nar/gky014.
30. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes IT, Malla S, Marriott H, Nieto T, O'Grady J, Olsen HE, Pedersen BS, Rhie A, Richardson H, Quinlan AR, Snutch TP, Tee L, Paten B, Phillippy AM, Simpson JT, Loman NJ, Loose M. 2018. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol 36:338–345. https://doi.org/10.1038/nbt.4060.
31. Byrne A, Beaudin AE, Olsen HE, Jain M, Cole C, Palmer T, DuBois RM, Forsberg EC, Akeson M, Vollmers C. 2017. Nanopore long-read RNA-seq reveals widespread transcriptional variation among the surface receptors of individual B cells. Nat Commun 8:16027. https://doi.org/10.1038/ncomms16027.
32. Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, Timp W. 2017. Detecting DNA cytosine methylation using nanopore sequencing. Nat Methods 14:407–410. https://doi.org/10.1038/nmeth.4184.
33. Liu Q, Fang L, Yu G, Wang D, Xiao C-L, Wang K. 2019. Detection of DNA base modifications by deep recurrent neural network on Oxford Nanopore sequencing data. Nat Commun 10:2449. https://doi.org/10.1038/s41467-019-10168-2.
34. Workman RE, Tang AD, Tang PS, Jain M, Tyson JR, Razaghi R, Zuzarte PC, Gilpatrick T, Payne A, Quick J, Sadowski N, Holmes N, de Jesus JG, Jones KL, Soulette CM, Snutch TP, Loman N, Paten B, Loose M, Simpson JT, Olsen HE, Brooks AN, Akeson M, Timp W. 2019. Nanopore native RNA sequencing of a human poly(A) transcriptome. Nat Methods 16:1297–1305. https://doi.org/10.1038/s41592-019-0617-2.
35. Chappell L, Ross P, Orchard L, Otto TD, Berriman M, Rayner JC, Llinás M. 2019. Refining the transcriptome of the human malaria parasite *Plasmodium falciparum* using amplification-free RNA-seq. bioRxiv https://doi.org/10.1101/852038:852038.
36. Garalde DR, Snell EA, Jachimowicz D, Sipos B, Lloyd JH, Bruce M, Pantic N, Admassu T, James P, Warland A, Jordan M, Ciccone J, Serra S, Keenan J, Martin S, McNeill L, Wallace EJ, Jayasinghe L, Wright C, Blasco J, Young S, Brocklebank D, Juul S, Clarke J, Heron AJ, Turner DJ. 2018. Highly parallel direct RNA sequencing on an array of nanopores. Nat Methods 15:201–206. https://doi.org/10.1038/nmeth.4577.
37. Viehweger A, Krautwurst S, Lamkiewicz K, Madhugiri R, Ziebuhr J, Hölzer M, Marz M. 2019. Direct RNA nanopore sequencing of full-length coronavirus genomes provides novel insights into structural variants and enables modification analysis. Genome Res 29:1545–1554. https://doi.org/10.1101/gr.247064.118.
38. Sharon D, Tilgner H, Grubert F, Snyder M. 2013. A single-molecule long-read survey of the human transcriptome. Nat Biotechnol 31:1009–1014. https://doi.org/10.1038/nbt.2705.
39. Soneson C, Yao Y, Bratus-Neuenschwander A, Patrignani A, Robinson MD, Hussain S. 2019. A comprehensive examination of Nanopore native RNA sequencing for characterization of complex transcriptomes. Nat Commun 10:3359. https://doi.org/10.1038/s41467-019-11272-z.
40. Lee VV, Judd LM, Jex AR, Holt KE, Tonkin CJ, Ralph SA. 2020. Direct nanopore sequencing of mRNA reveals landscape of transcript isoforms in apicomplexan parasites. bioRxiv https://doi.org/10.1101/2020.02.16.946699:2020.02.16.946699.
41. Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34:3094–3100. https://doi.org/10.1093/bioinformatics/bty191.
42. Shaw PJ, Kaewprommal P, Piriyapongsa J, Wongsombat C, Yuthavong Y, Kamchonwongpaisan S. 2016. Estimating mRNA lengths from *Plasmodium falciparum* genes by virtual Northern RNA-seq analysis. Int J Parasitol 46:7–12. https://doi.org/10.1016/j.ijpara.2015.09.007.
43. Minot S, Melo MB, Li F, Lu D, Niedelman W, Levine SS, Saeij JP. 2012. Admixture and recombination among *Toxoplasma gondii* lineages explain global genome diversity. Proc Natl Acad Sci U S A 109:13458–13463. https://doi.org/10.1073/pnas.1117047109.
44. Rehkopf DH, Gillespie DE, Harrell MI, Feagin JE. 2000. Transcriptional mapping and RNA processing of the *Plasmodium falciparum* mitochondrial mRNAs. Mol Biochem Parasitol 105:91–103. https://doi.org/10.1016/s0166-6851(99)00170-x.
45. Nisbet RER, McKenzie JL. 2016. Transcription of the apicoplast genome. Mol Biochem Parasitol 210:5–9. https://doi.org/10.1016/j.molbiopara.2016.07.004.
46. Nilsen TW, Graveley BR. 2010. Expansion of the eukaryotic proteome by alternative splicing. Nature 463:457–463. https://doi.org/10.1038/nature08909.
47. Braunschweig U, Barbosa-Morais NL, Pan Q, Nachman EN, Alipanahi B, Gonatopoulos-Pournatzis T, Frey B, Irimia M, Blencowe BJ. 2014. Widespread intron retention in mammals functionally tunes transcriptomes. Genome Res 24:1774–1786. https://doi.org/10.1101/gr.177790.114.
48. Wong JJ, Ritchie W, Ebner OA, Selbach M, Wong JW, Huang Y, Gao D, Pinello N, Gonzalez M, Baidya K, Thoeng A, Khoo TL, Bailey CG, Holst J, Rasko JE. 2013. Orchestrated intron retention regulates normal granulocyte differentiation. Cell 154:583–595. https://doi.org/10.1016/j.cell.2013.06.052.
49. Boutz PL, Bhutkar A, Sharp PA. 2015. Detained introns are a novel, widespread class of posttranscriptionally spliced introns. Genes Dev 29:63–80. https://doi.org/10.1101/gad.247361.114.
50. Sakharkar MK, Chow VT, Kangueane P. 2004. Distributions of exons and introns in the human genome. In Silico Biol 4:387–393.

Downloaded from http://msystems.asm.org/ on May 16, 2021 by guest

51. Ivashchenko AT, Khaĭlenko VA, Atambaeva SA. 2009. Variations of the length of exons and introns in human genome genes. Genetika 45:22–29.
52. Lau YL, Lee WC, Gudimella R, Zhang G, Ching XT, Razali R, Aziz F, Anwar A, Fong MY. 2016. Deciphering the draft genome of *Toxoplasma gondii* RH strain. PLoS One 11:e0157901. https://doi.org/10.1371/journal.pone.0157901.
53. Hall N, Pain A, Berriman M, Churcher C, Harris B, Harris D, Mungall K, Bowman S, Atkin R, Baker S, Barron A, Brooks K, Buckee CO, Burrows C, Cherevach I, Chillingworth C, Chillingworth T, Christodoulou Z, Clark L, Clark R, Corton C, Cronin A, Davies R, Davis P, et al. 2002. Sequence of *Plasmodium falciparum* chromosomes 1, 3 to 9, and 13. Nature 419:527–531. https://doi.org/10.1038/nature01095.
54. Liao Y, Smyth GK, Shi W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30:923–930. https://doi.org/10.1093/bioinformatics/btt656.
55. Gonzalez-Porta M, Frankish A, Rung J, Harrow J, Brazma A. 2013. Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. Genome Biol 14:R70. https://doi.org/10.1186/gb-2013-14-7-r70.
56. Schmitz U, Pinello N, Jia F, Alasmari S, Ritchie W, Keightley M-C, Shini S, Lieschke GJ, Wong JJL, Rasko JEJ. 2017. Intron retention enhances gene regulatory complexity in vertebrates. Genome Biol 18:216. https://doi.org/10.1186/s13059-017-1339-3.
57. Tang AD, Soulette CM, van Baren MJ, Hart K, Hrabeta-Robinson E, Wu CJ, Brooks AN. 2020. Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. Nat Commun 11:1438. https://doi.org/10.1038/s41467-020-15171-6.
58. Nagy E, Maquat LE. 1998. A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. Trends Biochem Sci 23:198–199. https://doi.org/10.1016/s0968-0004(98)01208-0.
59. Ge Y, Porse BT. 2014. The functional consequences of intron retention: alternative splicing coupled to NMD as a regulator of gene expression. Bioessays 36:236–243. https://doi.org/10.1002/bies.201300156.
60. Sidik SM, Huet D, Ganesan SM, Huynh MH, Wang T, Nasamu AS, Thiru P, Saeij JPJ, Carruthers VB, Niles JC, Lourido S. 2016. A Genome-wide CRISPR screen in *Toxoplasma* identifies essential apicomplexan genes. Cell 166:1423–1435. https://doi.org/10.1016/j.cell.2016.08.019.
61. Zhang M, Wang C, Otto TD, Oberstaller J, Liao X, Adapa SR, Udenze K, Bronner IF, Casandra D, Mayho M, Brown J, Li S, Swanson J, Rayner JC, Jiang RHY, Adams JH. 2018. Uncovering the essential genes of the human malaria parasite *Plasmodium falciparum* by saturation mutagenesis. Science 360:eaap7847. https://doi.org/10.1126/science.aap7847.
62. Pai AA, Henriques T, McCue K, Burkholder A, Adelman K, Burge CB. 2017. The kinetics of pre-mRNA splicing in the *Drosophila* genome and the influence of gene architecture. Elife 6:e32537. https://doi.org/10.7554/eLife.32537.
63. Skinner SO, Xu H, Nagarkar-Jaiswal S, Freire PR, Zwaka TP, Golding I. 2016. Single-cell analysis of transcription kinetics across the cell cycle. Elife 5:e12175. https://doi.org/10.7554/eLife.12175.
64. Yeoh LM, Goodman CD, Hall NE, van Dooren GG, McFadden GI, Ralph SA. 2015. A serine–arginine-rich (SR) splicing factor modulates alternative splicing of over a thousand genes in *Toxoplasma gondii*. Nucleic Acids Res 43:4661–4675. https://doi.org/10.1093/nar/gkv311.
65. Sorber K, Dimon MT, DeRisi JL. 2011. RNA-Seq analysis of splicing in *Plasmodium falciparum* uncovers new splice junctions, alternative splicing and splicing of antisense transcripts. Nucleic Acids Res 39:3820–3835. https://doi.org/10.1093/nar/gkq1223.
66. Tsai K-W, Chan W-C, Hsu C-N, Lin W-C. 2010. Sequence features involved in the mechanism of 3′ splice junction wobbling. BMC Mol Biol 11:34–34. https://doi.org/10.1186/1471-2199-11-34.
67. Moldován N, Tombácz D, Szűcs A, Csabai Z, Balázs Z, Kis E, Molnár J, Boldogkői Z. 2018. Third-generation sequencing reveals extensive polycistronism and transcriptional overlapping in a baculovirus. Sci Rep 8:8604. https://doi.org/10.1038/s41598-018-26955-8.
68. Dozmorov MG, Adrianto I, Giles CB, Glass E, Glenn SB, Montgomery C, Sivils KL, Olson LE, Iwayama T, Freeman WM, Lessard CJ, Wren JD. 2015. Detrimental effects of duplicate reads and low complexity regions on RNA- and ChIP-seq data. BMC Bioinformatics 16(Suppl 13):S10. https://doi.org/10.1186/1471-2105-16-S13-S10.
69. Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S, Paulsen IT, James K, Eisen JA, Rutherford K, Salzberg SL, Craig A, Kyes S, Chan MS, Nene V, Shallom SJ, Suh B, Peterson J, Angiuoli S, Pertea M, Allen J, Selengut J, Haft D, Mather MW, Vaidya AB, Martin DM, Fairlamb AH, Fraunholz MJ, Roos DS, Ralph SA, McFadden GI, Cummings LM, Subramanian GM, Mungall C, Venter JC, Carucci DJ, Hoffman SL, Newbold C, Davis RW, Fraser CM, Barrell B. 2002. Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature 419:498–511. https://doi.org/10.1038/nature01097.
70. Reference deleted.
71. Runtuwene LR, Tuda JSB, Mongan AE, Makalowski W, Frith MC, Imwong M, Srisutham S, Nguyen Thi LA, Tuan NN, Eshita Y, Maeda R, Yamagishi J, Suzuki Y. 2018. Nanopore sequencing of drug resistance-associated genes in malaria parasites, *Plasmodium falciparum*. Sci Rep 8:8286. https://doi.org/10.1038/s41598-018-26334-3.
72. Loven J, Orlando DA, Sigova AA, Lin CY, Rahl PB, Burge CB, Levens DL, Lee TI, Young RA. 2012. Revisiting global gene expression analysis. Cell 151:476–482. https://doi.org/10.1016/j.cell.2012.10.012.
73. Tarazona S, Garcia-Alcalde F, Dopazo J, Ferrer A, Conesa A. 2011. Differential expression in RNA-seq: a matter of depth. Genome Res 21:2213–2223. https://doi.org/10.1101/gr.124321.111.
74. Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, Slobodeniuc V, Kutter C, Watt S, Çolak R, Kim T, Misquitta-Ali CM, Wilson MD, Kim PM, Odom DT, Frey BJ, Blencowe BJ. 2012. The evolutionary landscape of alternative splicing in vertebrate species. Science 338:1587–1593. https://doi.org/10.1126/science.1230612.
75. Kim E, Magen A, Ast G. 2007. Different levels of alternative splicing among eukaryotes. Nucleic Acids Res 35:125–131. https://doi.org/10.1093/nar/gkl924.
76. Lynch M, Conery JS. 2003. The origins of genome complexity. Science 302:1401–1404. https://doi.org/10.1126/science.1089370.
77. Chen L, Bush SJ, Tovar-Corona JM, Castillo-Morales A, Urrutia AO. 2014. Correcting for differential transcript coverage reveals a strong relationship between alternative splicing and organism complexity. Mol Biol Evol 31:1402–1413. https://doi.org/10.1093/molbev/msu083.
78. Irimia M, Penny D, Roy SW. 2007. Coevolution of genomic intron number and splice sites. Trends Genet 23:321–325. https://doi.org/10.1016/j.tig.2007.04.001.
79. McGuire AM, Pearson MD, Neafsey DE, Galagan JE. 2008. Cross-kingdom patterns of alternative splicing and splice recognition. Genome Biol 9:R50. https://doi.org/10.1186/gb-2008-9-3-r50.
80. Sieber P, Voigt K, Kämmer P, Brunke S, Schuster S, Linde J. 2018. Comparative study on alternative splicing in human fungal pathogens suggests its involvement during host invasion. Front Microbiol 9:2313. https://doi.org/10.3389/fmicb.2018.02313.
81. Grau-Bové X, Ruiz-Trillo I, Irimia M. 2018. Origin of exon skipping-rich transcriptomes in animals driven by evolution of gene architecture. Genome Biol 19:135–135. https://doi.org/10.1186/s13059-018-1499-9.
82. Kawashima T, Douglass S, Gabunilas J, Pellegrini M, Chanfreau GF. 2014. Widespread use of non-productive alternative splice sites in *Saccharomyces cerevisiae*. PLoS Genet 10:e1004249. https://doi.org/10.1371/journal.pgen.1004249.
83. Lewis BP, Green RE, Brenner SE. 2003. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. Proc Natl Acad Sci U S A 100:189–192. https://doi.org/10.1073/pnas.0136770100.
84. Sayani S, Janis M, Lee CY, Toesca I, Chanfreau GF. 2008. Widespread impact of nonsense-mediated mRNA decay on the yeast intronome. Mol Cell 31:360–370. https://doi.org/10.1016/j.molcel.2008.07.005.
85. Plass M, Codony-Servat C, Ferreira PG, Vilardell J, Eyras E. 2012. RNA secondary structure mediates alternative 3′ss selection in *Saccharomyces cerevisiae*. RNA 18:1103–1115. https://doi.org/10.1261/rna.030767.111.
86. Weischenfeldt J, Damgaard I, Bryder D, Theilgaard-Mönch K, Thoren LA, Nielsen FC, Jacobsen SEW, Nerlov C, Porse BT. 2008. NMD is essential for hematopoietic stem and progenitor cells and for eliminating by-products of programmed DNA rearrangements. Genes Dev 22:1381–1396. https://doi.org/10.1101/gad.468808.
87. Hwang J, Maquat LE. 2011. Nonsense-mediated mRNA decay (NMD) in animal embryogenesis: to die or not to die, that is the question. Curr Opin Genet Dev 21:422–430. https://doi.org/10.1016/j.gde.2011.03.008.
88. Shi M, Zhang H, Wang L, Zhu C, Sheng K, Du Y, Wang K, Dias A, Chen S, Whitman M, Wang E, Reed R, Cheng H. 2015. Premature termination codons are recognized in the nucleus in a reading-frame dependent manner. Cell Discov 1. https://doi.org/10.1038/celldisc.2015.1.
89. Harati S, Phan JH, Wang MD. 2014. Investigation of factors affecting RNA-seq gene expression calls. Annu Int Conf IEEE Eng Med Biol Soc 2014:5232–5235.

Downloaded from http://msystems.asm.org/ on May 16, 2021 by guest

90. Zhao S, Zhang Y, Gamini R, Zhang B, von Schack D. 2018. Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: poly(A)$^+$ selection versus rRNA depletion. Sci Rep 8:4781–4781. https://doi.org/10.1038/s41598-018-23226-4.
91. Zhang AY, Su S, Ng AP, Holik AZ, Asselin-Labat M-L, Ritchie ME, Law CW. 2018. A data-driven approach to characterizing intron signal in RNA-seq data. bioRxiv https://doi.org/10.1101/352823:352823.
92. Yap K, Makeyev EV. 2013. Regulation of gene expression in mammalian nervous system through alternative pre-mRNA splicing coupled with RNA quality control mechanisms. Mol Cell Neurosci 56:420–428. https://doi.org/10.1016/j.mcn.2013.01.003.
93. Wang M, Branco AT, Lemos B. 2018. The Y chromosome modulates splicing and sex-biased intron retention rates in *Drosophila*. Genetics 208:1057–1067. https://doi.org/10.1534/genetics.117.300637.
94. Edwards CR, Ritchie W, Wong JJL, Schmitz U, Middleton R, An X, Mohandas N, Rasko JEJ, Blobel GA. 2016. A dynamic intron retention program in the mammalian megakaryocyte and erythrocyte lineages. Blood 127:e24–e34. https://doi.org/10.1182/blood-2016-01-692764.
95. Middleton R, Gao D, Thomas A, Singh B, Au A, Wong JJL, Bomane A, Cosson B, Eyras E, Rasko JEJ, Ritchie W. 2017. IRFinder: assessing the impact of intron retention on mammalian gene expression. Genome Biol 18:51. https://doi.org/10.1186/s13059-017-1184-4.
96. Nickless A, Bailis JM, You Z. 2017. Control of gene expression through the nonsense-mediated RNA decay pathway. Cell Biosci 7:26–26. https://doi.org/10.1186/s13578-017-0153-7.
97. Han X, Wei Y, Wang H, Wang F, Ju Z, Li T. 2018. Nonsense-mediated mRNA decay: a "nonsense" pathway makes sense in stem cell biology. Nucleic Acids Res 46:1038–1051. https://doi.org/10.1093/nar/gkx1272.
98. Vanichkina DP, Schmitz U, Wong JJL, Rasko JEJ. 2018. Challenges in defining the role of intron retention in normal biology and disease. Semin Cell Dev Biol 75:40–49. https://doi.org/10.1016/j.semcdb.2017.07.030.
99. Saudemont B, Popa A, Parmley JL, Rocher V, Blugeon C, Necsulea A, Meyer E, Duret L. 2017. The fitness cost of mis-splicing is the main determinant of alternative splicing patterns. Genome Biol 18:208–208. https://doi.org/10.1186/s13059-017-1344-6.
100. Foth BJ, Zhang N, Chaal BK, Sze SK, Preiser PR, Bozdech Z. 2011. Quantitative time-course profiling of parasite and host cell proteins in the human malaria parasite *Plasmodium falciparum*. Mol Cell Proteomics 10:M110.006411. https://doi.org/10.1074/mcp.M110.006411.
101. Aurrecoechea C, Barreto A, Basenko EY, Brestelli J, Brunk BP, Cade S, Crouch K, Doherty R, Falke D, Fischer S, Gajria B, Harb OS, Heiges M, Hertz-Fowler C, Hu S, Iodice J, Kissinger JC, Lawrence C, Li W, Pinney DF, Pulman JA, Roos DS, Shanmugasundram A, Silva-Franco F, Steinbiss S, Stoeckert CJ, Jr, Spruill D, Wang H, Warrenfeltz S, Zheng J. 2017. EuPathDB: the eukaryotic pathogen genomics database resource. Nucleic Acids Res 45:D581–D591. https://doi.org/10.1093/nar/gkw1105.
102. Andrews S. 2010. FastQC: a quality control tool for high-throughput sequence data, http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
103. Lorenzi H, Khan A, Behnke MS, Namasivayam S, Swapna LS, Hadjithomas M, Karamycheva S, Pinney D, Brunk BP, Ajioka JW, Ajzenberg D, Boothroyd JC, Boyle JP, Dardé ML, Diaz-Miranda MA, Dubey JP, Fritz HM, Gennari SM, Gregory BD, Kim K, Saeij JP, Su C, White MW, Zhu XQ, Howe DK, Rosenthal BM, Grigg ME, Parkinson J, Liu L, Kissinger JC, Roos DS, Sibley LD. 2016. Local admixture of amplified and diversified secreted pathogenesis determinants shapes mosaic *Toxoplasma gondii* genomes. Nat Commun 7:10147. https://doi.org/10.1038/ncomms10147.
104. Siegel TN, Hon C-C, Zhang Q, Lopez-Rubio J-J, Scheidig-Benatar C, Martins RM, Sismeiro O, Coppée J-Y, Scherf A. 2014. Strand-specific RNA-Seq reveals widespread and developmentally regulated transcription of natural antisense transcripts in *Plasmodium falciparum*. BMC Genomics 15:150. https://doi.org/10.1186/1471-2164-15-150.
105. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol 37:907–915. https://doi.org/10.1038/s41587-019-0201-4.
106. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079. https://doi.org/10.1093/bioinformatics/btp352.
107. Broad Institute. 2019. Picard toolkit. Broad Institute, Cambridge, MA. http://broadinstitute.github.io/picard/.
108. Weirather JL, de Cesare M, Wang Y, Piazza P, Sebastiano V, Wang X-J, Buck D, Au KF. 2017. Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. F1000Res 6:100–100. https://doi.org/10.12688/f1000research.10571.2.
109. Wickham H. 2016. ggplot2: elegant graphics for data analysis. Springer-Verlag, New York, NY.
110. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. Nat Biotechnol 29:24–26. https://doi.org/10.1038/nbt.1754.
111. Oxford Nanopore Technologies, Ltd. 2016. Wub. https://github.com/nanoporetech/wub.
112. Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842. https://doi.org/10.1093/bioinformatics/btq033.
113. An J, Lai J, Sajjanhar A, Batra J, Wang C, Nelson CC. 2015. J-Circos: an interactive Circos plotter. Bioinformatics 31:1463–1465. https://doi.org/10.1093/bioinformatics/btu842.
114. Ramírez F, Dündar F, Diehl S, Grüning BA, Manke T. 2014. deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res 42:W187–W191. https://doi.org/10.1093/nar/gku365.
115. Wang L, Wang S, Li W. 2012. RSeQC: quality control of RNA-seq experiments. Bioinformatics 28:2184–2185. https://doi.org/10.1093/bioinformatics/bts356.
116. Blankenberg D, Von Kuster G, Bouvier E, Baker D, Afgan E, Stoler N, Taylor J, Nekrutenko A, Galaxy Team. 2014. Dissemination of scientific software with Galaxy ToolShed. Genome Biol 15:403. https://doi.org/10.1186/gb4161.
117. Hulsen T, de Vlieg J, Alkema W. 2008. BioVenn: a web application for the comparison and visualization of biological lists using area-proportional Venn diagrams. BMC Genomics 9:488. https://doi.org/10.1186/1471-2164-9-488.
118. Liao Y, Wang J, Jaehnig EJ, Shi Z, Zhang B. 2019. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. Nucleic Acids Res 47:W199–W205. https://doi.org/10.1093/nar/gkz401.

Downloaded from http://msystems.asm.org/ on May 16, 2021 by guest