Challenges and solutions for analysing dual RNA-seq data for non-model host–pathogen systems

Kayleigh R. O'Keeffe[1] **Corbin D. Jones**[1,2]

[1]Department of Biology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina

[2]Integrative Program for Biological & Genome Sciences, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina

Correspondence
Kayleigh R. O'Keeffe
Email: kokeeffe@live.unc.edu

Funding information
NSF-USDA joint program in Ecology and Evolution of Infectious Diseases, Grant/Award Number: 2016-67013-25762; University Cancer Research Fund; Triangle Center for Evolutionary Medicine; the National Science Foundation

Handling Editor: Steven Kembel

Abstract

1. Dual RNA-seq simultaneously profiles the transcriptomes of a host and of a pathogen during infection and may reveal the mechanisms underlying host–pathogen interactions. Dual RNA-seq is inherently a mixture of transcripts from at least two species (host and pathogen), so this mixture must be computationally sorted into host and pathogen components. Sorting relies on aligning reads to respective reference genomes, which may be unavailable for both species in non-model host–pathogen pairs. This lack of genomic resources may present challenges to applying dual RNA-seq to non-model systems.
2. We assessed the accuracy of alignments of dual RNA-seq when using the genomic resources of a closely related species to the species of interest by simulating datasets of mixed transcripts from a host and pathogen. Specifically, we compared how different aligners performed across different proportions of pathogen-to-host transcripts and across variation in the genetic distance between the pathogen genome and reference genome. We performed extensive analyses for a host plant with a fungal pathogen, and then, we extended the plant–fungus results by repeating key analyses in vertebrate (human)–fungus and vertebrate–bacterium systems.
3. Aligners that were able to map pathogen transcripts to the reference genome of a species closely related to the pathogen (a 'related reference genome') also mismapped transcripts originating from the host to the pathogen's related reference genome. This resulted in regions where this occurred being quantified as overexpressed. If a host reference genome was available, we show that, to minimize host transcript mismapping and retain the ability to map pathogen transcripts, one could concatenate the host genome with the pathogen's related reference genome, then map transcripts to the concatenated genomes. If a host genome was unavailable, assembling reads de novo prior to aligning substantially decreased host read mismapping, while retaining the ability to map pathogen transcripts to a related reference genome.
4. The application of dual RNA-seq to organisms without reference genomes is currently limited. We propose an analytical workflow that leverages the genomic resources of species closely related to species of interest to facilitate the application of dual RNA-seq to reveal the mechanisms of host–pathogen interactions across a wider array of systems.

KEYWORDS
disease ecology, dual RNA-seq, gene expression, host–pathogen interactions, transcriptomics

1 | INTRODUCTION

Viruses, bacteria and fungi can invade and parasitize eukaryotic host cells. Hosts may respond to infection by upregulating defence pathways. Pathogens, in turn, evade these host immune responses as they infect and cause disease. As this process unfolds and each organism responds to the other, gene expression changes in both the host and the pathogen (Kawahara et al., 2012). Yet, despite the importance of host–pathogen interactions, the genetic mechanisms underlying host–pathogen interactions during infection remain poorly understood. (Westermann, Gorski, & Vogel, 2012).

Gene expression studies have been revolutionized by RNA-seq (Mortazavi, Williams, McCue, Schaeffer, & Wold, 2008). In infection biology, RNA-seq can involve sequencing RNA extracted from pathogen-infected host tissue, an approach Westermann et al. (2012) coined as dual RNA-seq. Dual RNA-seq allows gene expression profiles of the host and pathogen to be characterized simultaneously during their interaction. For example, Teixeira et al. (2014) illustrate that the fungal parasite *Monoliophthora perniciosa* orchestrates the changes in the metabolism of the cacao plant *Theobroma cacao* that increase the availability of nutrients before the pathogen ultimately kills the plant. Yet, despite the power of dual RNA-seq as a tool to identify the genetic mechanisms underlying host–parasite interactions, a number of complications make dual RNA-seq difficult to adapt to non-model systems.

Dual RNA-seq data are inherently a mixture of transcripts from both host and pathogen. Given this mixture, the host and pathogen reads need to be sorted from each other. This sorting is typically done by an alignment algorithm that maps the reads to the two reference genomes, that of the host and that of the pathogen. To increase the accuracy of sorting and limit the potential for reads to mismap to the wrong reference genome (i.e. pathogen reads mapped to the host genome and vice versa), dual RNA-seq studies have employed a variety of analytical approaches such as discarding reads that map to both the host and pathogen reference genomes (e.g. Westermann et al., 2016) or concatenating host and pathogen reference genomes into a composite for genome alignment (Aprianto, Slager, Holsappel, & Veening, 2016). However, both methods of read sorting rely on reference genomes, limiting their application to systems in which the host and/or pathogen species have developed genomic resources. As model species with sequenced and annotated reference genomes comprise a small fraction of total species (https://www.ncbi.nlm.nih.gov/genome), many host–pathogen systems are potentially excluded from dual RNA-seq analysis.

For organisms without complete reference genomes, approaches to analysing RNA-seq data include assembling reads de novo (i.e. assembling many small reads together into fewer longer fragments, Grabherr et al., 2011) and mapping reads to reference genomes of related species (Ekblom & Galindo, 2010). As dual RNA-seq contains a mixture of transcripts from the host and pathogen species, the implications of the choice of one of these methods are complex, as accurate quantification levels of gene expression require reads to be separated correctly, and can influence the accuracy of the biological insight gleaned from the data. Reference-based methods, especially when using the reference genomes of a related species, could result in reads not mapping if the genetic distance between the target species and the reference species is too high. Alternatively, aligners that allow more mismatches between reads and the reference genome could mismap reads from the wrong species, which could lead to spurious inference of how gene expression is affected by infection. De novo assembly can alleviate some of these problems, by creating larger fragments with which to map, but de novo assembly may result in reads from the two species assembling into longer fragments, which would lead to inaccuracies in subsequent mapping. Furthermore, dual RNA-seq datasets vary in the proportion of pathogen reads to host reads in the datasets, depending on the system of interest (Baddal et al., 2015; Choi, Aliota, Mayhew, Erickson, & Christensen, 2014; Hayden et al., 2014), and this variation may affect the accuracy and biological interpretation of various analytical methods.

Despite the variety of analytical approaches and their potential influence on the interpretation of dual RNA-seq data, the accuracy of these methods has not been systematically assessed. Here, we determined if and how dual RNA-seq can be utilized in non-model host–pathogen systems in which genomic resources are limited. Specifically, we investigated if using the genomic resources of species closely related to the pathogen species of interest can facilitate the use of dual RNA-seq in non-model systems. To do this, we simulated dual RNA-seq datasets. Simulations allowed us to manipulate dataset characteristics, like the proportion of pathogen to host reads, as well as facilitated downstream assessments of the accuracy of various analytical approaches. To investigate if using the genomic resources of a species closely related to a species of interest to analyse dual RNA-seq is appropriate, we assessed how genetic divergence between the genome of the pathogen of interest (from here on referred to as the target genome) and the reference genome affected mapping accuracy. Additionally, we assessed how this effect was mediated by different alignment methods and the fraction of pathogen reads in the sample (Figure 1). We explored four different approaches: 1. aligning raw reads to the reference genome of the pathogen or a related species, 2. aligning reads to the host genome first and then mapping those reads that did not map to the host genome to the reference genome of the pathogen or a related species (referred to as the sequential approach), 3. aligning reads to

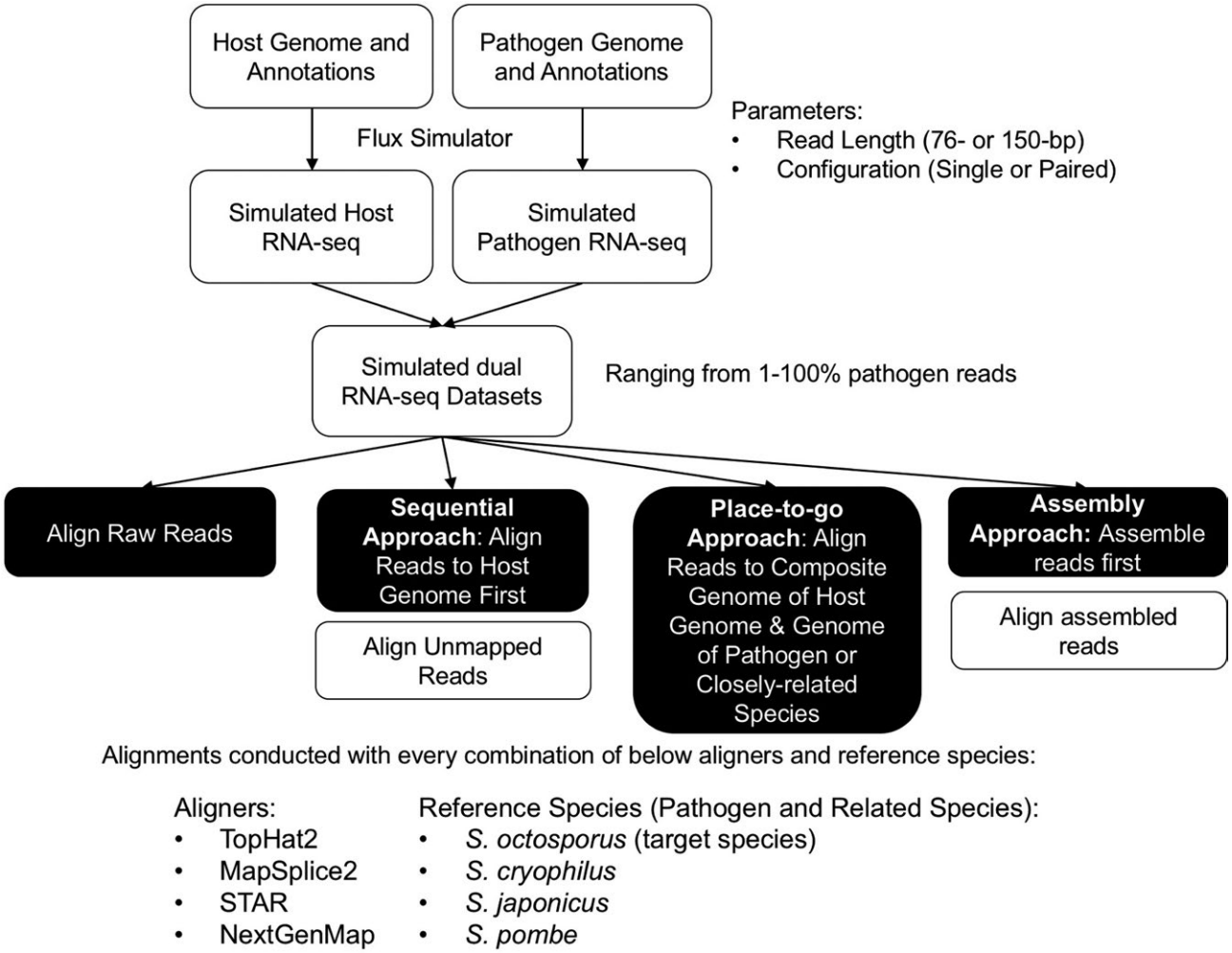

FIGURE 1 Dual RNA-seq Simulation Study Workflow. We outline our steps to investigate best approaches for analysing dual RNA-seq datasets of non-model systems

a composite genome comprised of the genomes of the host species and pathogen or a related species (referred to as the place-to-go approach), and 4. assembling reads de novo prior to aligning (referred to as the assembly approach). This assessment provides guidance as to how to approach dual RNA-seq when studying organisms that do not have fully sequenced genomes.

2 | MATERIALS AND METHODS

2.1 | Study system

Simulating RNA-seq relies on a reference genome and annotation file as inputs; therefore, we model dual RNA-seq using well-characterized genomes for both host and pathogen. Additionally, to investigate the effect of genetic distance between the pathogen of interest and the reference genome used for aligning reads, we needed to model a pathogen species for which closely related sister species were also fully sequenced. First, we used *Arabidopsis thaliana* and *Schizosaccharomyces octosporus* to represent host and pathogen species, respectively. While this is not a naturally occurring host–pathogen system, or symbiosis for that matter, as model organisms, these species have fully sequenced and well-annotated genomes, which were ideal for our approach. Additionally, most species within the *Schizosaccharomyces* genus have sequenced and annotated genomes (Table S1). Using the genomes of the other *Schizosaccharomyces* species as references for read mapping allowed us to assess if dual RNA-seq data could be analysed by using the genomic information of a related species as a reference when studying a species without a sequenced genome. Therefore, *A. thaliana* and *S. octosporus* allowed us to quantify how sensitive (or robust) different potential analysis methods were to increasing genetic distances between the focal pathogen and the reference genome. *A. thaliana* will be referred to as the host, and *S. octosporus* will be referred to as the pathogen.

2.2 | RNA-seq simulations

Flux Simulator was used to generate simulated RNA-seq data (Griebel et al., 2012). Flux Simulator produced sequencing reads from a reference genome according to annotated transcripts.

RNA-seq data were simulated separately for the host and pathogen. For each, four datasets were simulated for a factorial combination of two read lengths (76-bp or 150-bp) and configuration (single end or paired end). Each dataset included 10 million reads, similar to the simulations used in Baruzzo et al., 2016. All other simulation parameters were run as default. Simulated datasets were output as FASTQ files. Reads within each dataset were labelled with unique species-identifying tags (either 'HOST' or 'PATH') to facilitate downstream assessments of alignments.

To create dual RNA-seq datasets, we randomly selected reads from complementing datasets (same read length and configuration) of the host and pathogen and mixed them together. For each of the four sets of sequencing parameters, we created 12 10-million read datasets that ranged from 1 to 100% pathogen reads. As dual RNA-seq is sequenced from RNA extracted from pathogen-i nfected host tissue, typical datasets are comprised of a very low percentage of pathogen reads. The analysis of datasets with higher percentages of pathogen reads was conducted to show when and how patterns changed across the range of the percentages of pathogen reads. Some systems investigate simultaneous gene expression of host and pathogens by mechanically separating cells of each species prior to sequencing (Ellison, DiRenzo, McDonald, Lips, & Zamudio, 2017). Datasets with higher percentages of pathogen reads that still include host reads could represent sequencing from RNA extracted after imperfect cell sorting.

2.3 | Dual RNA-seq analysis approaches

For each reference-based approach, the reference genomes and annotations for *S. octosporus* (the target pathogen), *S. cryophilus*, *S. japonicus and S. pombe* were used, downloaded in April 2018 from Fungi Ensembl ((Rhind et al., 2011), Table 1). As *S. octosporus* was simulated as the pathogen within the generated dual RNA-seq datasets, the genomic resources for the other species within the *Schizosaccharomyces* genus facilitated the investigation of how different levels of evolutionary distance between the genome of the target species and reference genome affect mapping accuracy. To generate a reference transcriptome, the BEDTools 'getfasta' utility version 2.25.0 was used to extract the transcript sequences from each of these downloaded genomes as specified by coordinates in the complementing annotation files (Quinlan & Hall, 2010). These transcript sequences were used as reference transcriptomes for alignment of reads from each dual RNA-seq dataset.

TABLE 1 *Schizosaccharomyces* species information

Species	1:1 Ortholog amino acid identity to target[a]	Genome size	GC content (%)
*S. octosporus**	–	11.5 Mb	38
S. cryophilus	85%	12.5 Mb	38
S. pombe	66%	12.5 Mb	36
S. japonicus	56%	12.5 Mb	44

[a]Target Species: Species used for dual RNA-seq simulations (*S. octosporus*).

Read mapping was conducted with four different aligners: TopHat2, STAR, MapSplice2 and NextGenMap. TopHat2 (version 2.1.1) and STAR (version 2.5.1b) (Dobin et al., 2013; Trapnell, Pachter, & Salzberg, 2009) are both splice-aware aligners. TopHat2 relies on a Burrows–Wheeler transform and FM-index to search for matches between a reference genome and RNA-seq reads. STAR, which uses a seed and anchor approach based on a Maximal Mappable Prefix, is more robust to non-continuous reads and some mismatches. Default parameter settings were used for both methods.

In addition to these two splice-aware aligners, a de novo aligner, MapSplice2 (version 2.2.1) was used to map reads to a reference genome (Wang et al., 2010). MapSplice2 detects splice junctions without any dependence on splice site features (an annotation file). We also mapped reads with an unspliced aligner, NextGenMap (version 0.4.12) to map reads from each simulated dataset to each reference transcriptome (Li & Durbin, 2010; Sedlazeck, Rescheneder, & von Haeseler, 2013). The hash-based variable mismatch threshold algorithm of NextGenMap maximizes its ability to utilize divergent reads. Default parameter settings were used.

Using the genomic resources and aligners discussed above, we processed reads in four different approaches to investigate the effectiveness and accuracy of analytical methods for dual RNA-seq data, (Workflow in Figure 1):

2.3.1 | Raw read mapping

First, we investigated the accuracy of mapping the raw sequencing reads that were comprised of both host and pathogen reads. We conducted alignments with the reference genome of the target pathogen species, *S. octosporus*, and those of species closely related to the target species. Each of the four alignment algorithms discussed above was utilized.

2.3.2 | Sequential mapping approach: map to host genome first

While the decision of mapping first to the host or pathogen is somewhat arbitrary, we believed that mapping to the host first would provide insight into the potential biases of dual RNA-seq and the impact of unintentional sequencing of a pathogen along with the host (i.e. unwittingly sequencing an infected host). Thus, we first mapped simulated dual RNA-seq datasets to the host genome, then took the reads left unmapped and mapped them to the genome of the target species and those of closely related species. Reads from each simulated dataset were mapped to the host genome using TopHat2 under default parameters. Following alignments, the output BAM files containing unmapped reads were converted to FASTA files using the 'bam2fq' function within Samtools version 1.3.1 (Li et al., 2009). These reads that did not map to the host genome were then mapped to each of the *Schizosaccharomyces* genomes with NextGenMap and STAR.

2.3.3 | Place-to-go approach: mapping to concatenated genome

We further investigated potential alignment methods when a host genome is available by mapping reads to concatenated genomes of the host genome and either the target pathogen genome or the genome of a species closely related to the target pathogen species. First, we investigated mapping accuracy with the composite genome of the host *A. thaliana* and of the target pathogen species *S. octosporus*, the two species used to simulate the dual RNA-seq datasets. Second, to assess the effectiveness and accuracy of this method when only the genomes of species closely related to the target pathogen species are available, we also created composite genomes of *A. thaliana* and each of the three other *Schizosaccharomyces* genomes. Read mapping was conducted with the four different aligners as described above.

2.3.4 | Assembly approach: de novo assembly

We investigated whether de novo assembly of reads prior to mapping affected the effectiveness and accuracy of alignments. Prior to mapping reads, Trinity (Haas et al., 2013, version 2.2.0) was used for de novo assembly. Default parameters were used. To determine which reads comprised each contig, Bowtie2, version 2.3.4.1, was used to align reads back to the assembled contigs (Langmead & Salzberg, 2012). If only pathogen reads mapped to a contig, the contig was tagged 'pathogen'. If only host reads mapped to a contig, the contig was tagged 'host'. If both pathogen and host reads mapped to a contig, the contig was tagged 'undetermined'. After Trinity assembly and tagging, contigs were mapped to each of the *Schizosaccharomyces* genomes with NextGenMap and STARlong under default settings.

2.4 | Evaluation of alignments

SAM/BAM conversions, sorting and indexing were performed with SAMtools version 1.3.1 and Picard version 2.2.4 (Li et al., 2009). For each alignment, the number of mapped and unmapped reads originating from *A. thaliana* and *S. octosporus* was counted by parsing BAM files for the previously added unique tags for each species.

To investigate how biological insight would be affected by alignment method, gene-wise counts were obtained with featureCounts (Liao, Smyth, & Shi, 2013) and differential gene expression analysis was performed following the instructions of the 'DESeq2' package (Love, Huber, & Anders, 2014) deposited in Bioconductor. Specifically, we quantified gene counts (reads overlapping exons as described in the annotation build) for alignments to the target pathogen genome. Gene expression was compared between alignments of the same sequencing dataset among the different aligner methods. These pairwise comparisons did not have replicates because our simulations were performed without stochastic sampling. To address that limitation, we used the rlog transformation function, which transforms the average of the genes across samples to a log2 scale, as well as accounts for genes for which the evidence for strong fold changes is weak due to low counts. This protocol does not produce *p*-values but provides a ranked list of genes by regularized fold changes.

2.5 | Applications to other host–pathogen systems

To investigate if patterns observed with the above approaches held across other host–pathogen systems, we also simulated dual RNA-seq involving another fungal pathogen, *Candida albicans*, and a different host species, *Homo sapiens*. Additionally, we simulated a bacterial pathogen system with *Homo sapiens* and *Escherichia coli*. Many species within the *Candida* and *Escherichia* genera have reference genomes available. We utilized two sister species of each *C. dubliniensis* and *C. parapsilosis*, and *E. fergusonii* and *E. albertii* as reference species to evaluate impact on dual RNA-seq analytical methods (Table S2).

3 | RESULTS

3.1 | Generation of simulated datasets

We simulated dual RNA-seq datasets to investigate if and how dual RNA-seq can be utilized in host–pathogen systems in which genomic resources are limited. Simulated datasets represented a factorial combination of read length (76 bp vs. 150 bp), sequencing configuration (paired or single ended), and 12 different ratios of host reads to pathogen reads. In total, 48 dual RNA-seq datasets, each with 10 million reads, were simulated. Similarly, we simulated dual RNA-seq datasets for *Homo sapiens* and *Candida albicans*, and *Homo sapiens* and *Escherichia coli*, also with varying ratios of pathogen reads to host reads. We will first discuss the main results from the 76-bp single-end *A. thaliana* and *S. octosporus* datasets relegating the extensions and ancillary results to the supplement.

3.2 | Comparison of raw read mapping

We first assessed the accuracy of alignments of dual RNA-seq raw reads with different aligners, representing a cross section of alignment algorithms, when using the correct target genome. When mapping raw reads to the target genome of the pathogen of interest, the four aligners had comparable mapping rates of reads originating from the pathogen (Figure 2a). TopHat2 and MapSplice2 aligned c. 88% of pathogen reads; STAR and NextGenMap each aligned over 99% of pathogen reads to the target reference. Mapping rate of pathogen reads was unaffected by the percentage of pathogen reads in the sequencing datasets. While STAR and NextGenMap achieved a high mapping rate of pathogen reads, both aligners also mismapped host reads to the genome of the target pathogen species (Figure 2b, S3). For STAR and NextGenMap alignments of datasets in which there were more host reads than pathogen reads, a common occurrence among real dual RNA-seq datasets, mismapped host reads comprised 25–98% of the total reads mapped. In sum, most pathogen reads from a dual RNA-seq experiment were aligned by common aligners, but the aligners that mapped the most pathogen

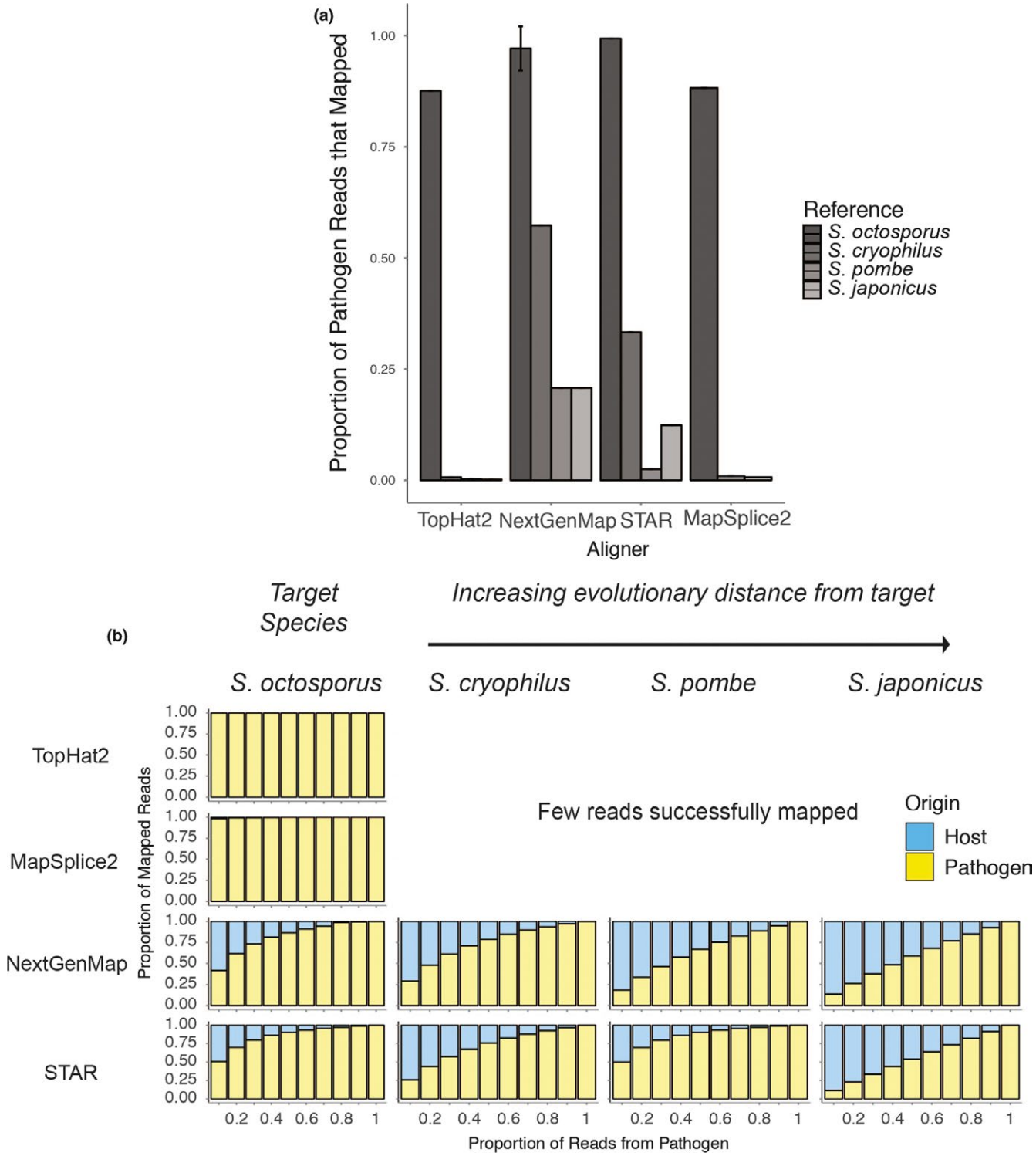

FIGURE 2 Comparison of utility and accuracy of several aligners for dual RNA-seq analysis. Factorial combination of four aligners and the four *Schizosaccharomyces* reference genomes. (a) Bars indicate the proportion of pathogen reads that mapped to each genome/transcriptome. Alignments to target pathogen species are shown in black, and greyscale gradient indicates evolutionary distance from the target. TopHat2 and MapSplice2 can only map pathogen reads when mapping to target pathogen genome and are unable to effectively map reads to genomes of related species. STAR and NextGenMap are able to map pathogen reads to reference genomes of related species. (b) Origin of Mapped Reads. For each bar plot, blue reads are those that originated from host (*Arabidopsis thaliana*) and yellow reads are those that originated from pathogen (*Schizosaccharomyces octosporus*). TopHat2 and MapSplice2 can only effectively map pathogen reads when mapping to target pathogen genome, and are unable to effectively map reads to the reference genomes of related species. These plots are therefore excluded. STAR and NextGenMap are able to map pathogen reads to reference genomes of related species, but both aligners result in host reads mismapping

TABLE 2 Results of sequential approach

		Target species		Increasing evolutionary distance from target					
		S. octosporus		*S. cryophilus*		*S. pombe*		*S. japonicus*	
	Proportion of pathogen reads in dataset	Raw reads	After host mapping	Raw reads	After host mapping	Raw reads	After host mapping	Raw reads	After host mapping
NextGenMap	0.1	0.70	0.68	0.80	0.78	0.86	0.83	0.89	0.88
	0.2	0.51	0.49	0.64	0.61	0.72	0.68	0.78	0.77
	0.3	0.38	0.35	0.50	0.48	0.61	0.56	0.68	0.66
	0.4	0.28	0.25	0.39	0.37	0.50	0.45	0.57	0.55
	0.5	0.21	0.19	0.30	0.28	0.40	0.35	0.47	0.45
	0.6	0.15	0.12	0.22	0.21	0.31	0.26	0.38	0.36
	0.7	0.10	0.08	0.16	0.14	0.22	0.19	0.28	0.26
	0.8	0.06	0.04	0.10	0.09	0.14	0.12	0.18	0.17
	0.9	0.01	0.00	0.05	0.04	0.07	0.06	0.09	0.08
	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
STAR	0.1	0.67	0.67	0.84	0.84	0.9	0.9	0.83	0.83
	0.2	0.47	0.47	0.7	0.7	0.8	0.81	0.69	0.69
	0.3	0.34	0.34	0.57	0.57	0.71	0.71	0.57	0.57
	0.4	0.25	0.25	0.46	0.46	0.61	0.61	0.46	0.46
	0.5	0.18	0.18	0.37	0.37	0.51	0.51	0.36	0.36
	0.6	0.13	0.13	0.28	0.28	0.41	0.41	0.27	0.27
	0.7	0.09	0.09	0.2	0.2	0.31	0.31	0.19	0.19
	0.8	0.05	0.05	0.13	0.13	0.21	0.21	0.12	0.12
	0.9	0.02	0.02	0.06	0.06	0.1	0.1	0.06	0.06
	1	0	0	0	0	0	0	0	0

Host mismapping rates (proportion of reads that mapped to pathogen genome or that of a closely related species that originated from host) are shown for raw read approach in which sequencing reads were mapped without any prior steps and sequential approach, in which reads that did not map to an initial alignment to host genome were mapped. Results are shown for each factorial combination of reference species and the proportion of pathogen reads in the simulated dataset. Cells' colour indicates the magnitude of host read mismapping rate: dark orange (>50% of total mapped reads are from host), intermediate orange (10–50%), light orange (<10%). When using NextGenMap, there are slight decreases when reads are first filtered by mapping to host genome first, but there is still a substantial amount of host mismapping with both methods. When using STAR, there is no effect of first filtering datasets by mapping to host genome first on host read mismapping.

reads to the source pathogen genome also incorrectly aligned the most host reads to the pathogen genome.

As many pathogen species do not have genomic resources, we investigated how each of the four aligners performed when mapping reads to the genome of a species closely related to the pathogen. TopHat2 and MapSplice2 performed poorly, mapping <0.5% of pathogen reads when mapping to any genome of a related species. STAR and NextGenMap were able to map pathogen reads when using the genomic information of a related species as a reference (Figure 2a). Mapping rates of pathogen reads remained unaffected by the percentage of pathogen reads in the sequencing datasets. When using STAR, c. 36% of pathogen reads mapped to the genome of *S. cryophilus*, the species most closely related to the target species, *S. octosporus*. 19% of pathogen reads mapped to the *S. japonicus* genome, and 3% mapped to the *S. pombe* genome (Figure 2a). When using NextGenMap, c. 60% of pathogen reads mapped to the *S. cryophilus* transcriptome, 22% mapped to the *S. pombe* transcriptome and 28% mapped to the *S. japonicus* transcriptome (Figure 2a). For both STAR and NextGenMap, the percentage of pathogen reads mapped generally decreased as evolutionary distance between the target and reference genomes increased. Thus, there is a distinct bifurcation between aligners that can and cannot effectively map pathogen reads when only a related reference genome is available.

For STAR and NextGenMap, host reads mismapping to the incorrect genome increased as the evolutionary distance between the target pathogen species and the reference species increased, while TopHat2 and MapSplice2 only mapped a few host reads (maximum of 19 reads) to any of the *Schizosaccharomyces* reference genomes under any sequencing parameters (Figure 2b, S3). When aligning with STAR, c. 21% of host reads mapped to the *S. cryophilus* and *S. pombe* genomes. Although only c. 1% of host reads mapped to the *S. pombe* genome, only 3% of pathogen reads mapped to the same genome, so overall mapping rate was very low. When aligning with NextGenMap, c. 25% of host reads mapped to the *S. cryophilus* and *S. japonicus* transcriptomes, and c. 15% of host reads mapped to *S. pombe*. The effects of evolutionary distance-related mismapping is greatest in datasets in which the proportion of pathogen reads was low, as host reads comprise the majority of total reads mapped.

3.3 | Gene counts and differential expression analysis

To investigate how biological insight would be affected by host read mismapping, we quantified gene-wise counts of the host reads that STAR mismapped to the target pathogen genome. The sequencing dataset that resulted in the highest number of host reads mismapped to the target pathogen genome in which the highest number of host reads mismapped was the datatset with highest proportion of host reads relative to pathogen reads (specifically, 99% host reads/1% pathogen reads). The vast majority of the mismapped host reads derived from repetitive parts of the genome. As determined by featureCounts, only 1.6% of *S. octosporus* genes had more than 50 host reads mismap to them (Table S5), and these genes varied in biological function (Table S4).

We compared two alignments of the same simulated dataset to the target genome of *S. octosporus*; one alignment was performed by mapping raw reads to the genome with TopHat2 (which did not mismap host reads) and the other was performed by mapping raw reads to the genome with STAR (which included host reads mismapping). A ranked list of regularized fold changes on the log2 scale is in the supplementary material (Table S3). 97.1% of genes that were expressed differently (above 0.1-fold change on the log2 scale) between alignments were overexpressed in the alignment in which host read mismapping occurred. This suggests that mismapping of reads to the wrong reference in dual RNA-seq can result in upward biases in estimates of gene expression. Thus, compared to an uninfected control, these upwardly biased genes would appear to be 'induced' by infection.

3.4 | Alternative mapping strategies may reduce mapping problems

We considered three alternative approaches that could reduce poor mapping and mismapping of dual RNA-seq data. We first investigated approaches that would be possible if a host genome was available. We tried to filter out host reads by first aligning dual RNA-seq datasets to the host genome, and then mapping the reads left unmapped to the target pathogen genome (or those of related species). This 'sequential' approach decreased the amount of host read mismapping only slightly (Table 2). For most alignments, the percentage of total mapped reads that originated from the host only decreased by 1–3%.

We investigated a second approach in which reads were mapped to concatenated genomes of the host and either the target pathogen or that of a closely related species to the pathogen. (a 'place-to-go' design). The place-to-go method resulted in alignments using STAR and NextGenMap having substantially fewer host reads mismapping to the genome of the pathogen or that of a species closely related to the pathogen compared to the alignments to genomes excluding the host genome. Furthermore, both aligners retained their ability to map pathogen reads to the genomes of closely related species to the target pathogen (Figure 3, Figure S4). Therefore, the place-to-go method may overcome some of the limitations of mapping to a related reference in dual RNA-seq.

Finally, we investigated a third approach in which reads were first assembled de novo into longer fragments, then those fragments were mapped to the genomes of the pathogen and species closely related to the pathogen ('assembly' approach). The majority of assembled contigs were comprised entirely of host reads or entirely of pathogen reads (Assembly metrics in Figure S1). A small fraction of contigs were chimeras—that is, a mix of host and pathogen reads (labelled as 'undetermined'). Alignments of these assembled transcripts to each of the reference genomes resulted in a substantial decrease in host read mismapping while preserving the ability to map pathogen contigs (Figure 4, Figure S5). Across all reference species and proportions of

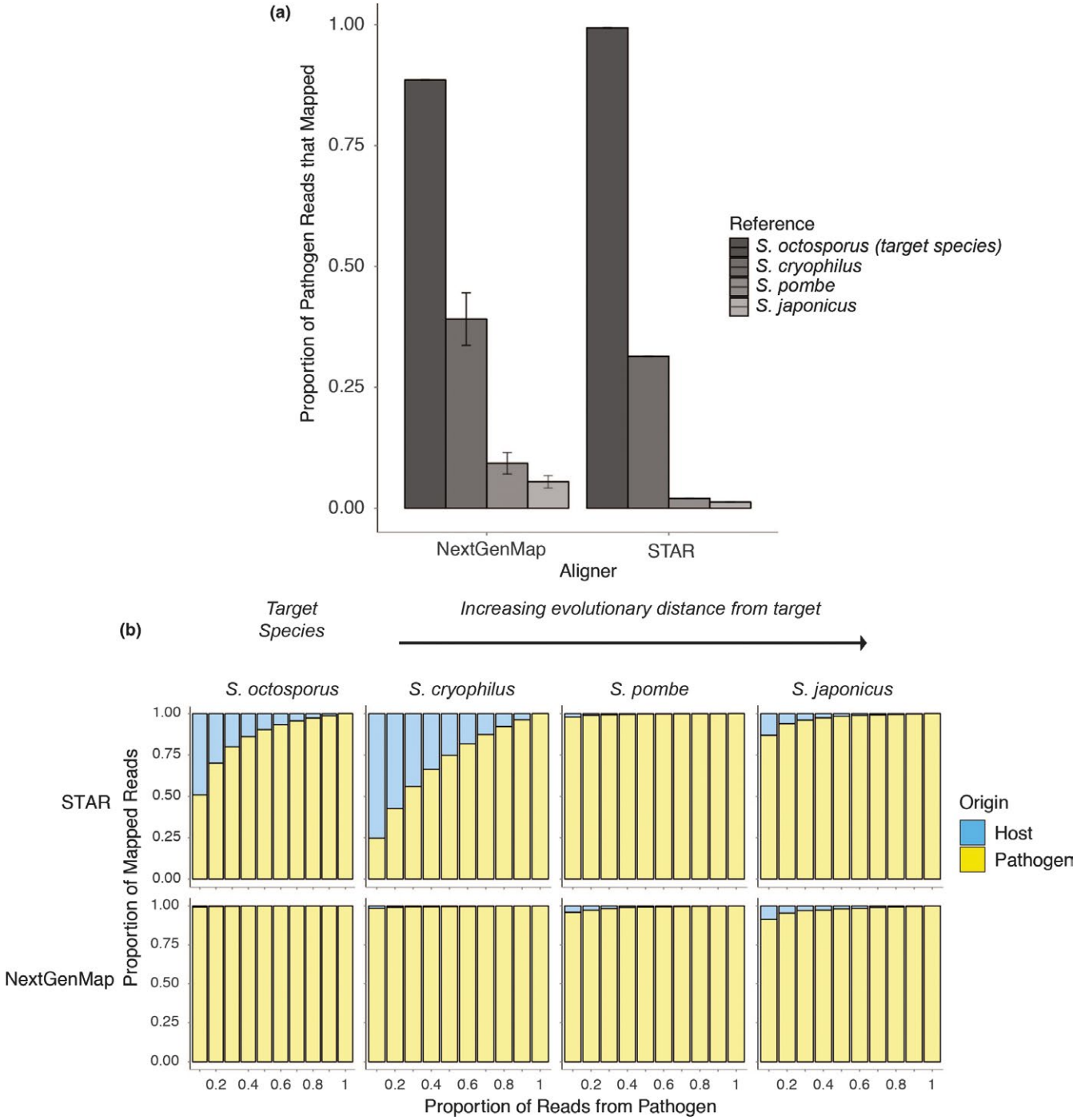

FIGURE 3 Concatenated genome mapping (Place-to-go approach) improve dual RNA-seq analysis. Results are shown for aligners STAR and NextGenMap and four composite genomes comprised of the *Arabidopsis thaliana* (host) genome and each of four *Schizosaccharomyces* reference genomes. (a) Bars indicate the proportion of pathogen reads that mapped to the part of the composite genome originating from the genome of the pathogen or closely related species to pathogen. Alignments to target genome of the pathogen species of interest are shown in black, and greyscale gradient indicates evolutionary distance from target. STAR is able to map almost all pathogen reads to target genome, and 27.7% of pathogen reads to the composite genome with genome of the most closely related genome to the target. NextGenMap is able to map c. 89% reads to target, and 39% reads to the composite genome with genome of most closely related genome to target. (b) Origin of reads that mapped to genome of pathogen or closely related species. For each bar plot, blue reads are those that originated from host (*Arabidopsis thaliana*) and yellow reads are those that originated from target pathogen (*Schizosaccharomyces octosporus*). Bar plots represent the composition of reads that mapped to the component of each composite genome corresponding to either the target pathogen genome or the genome of a closely related species

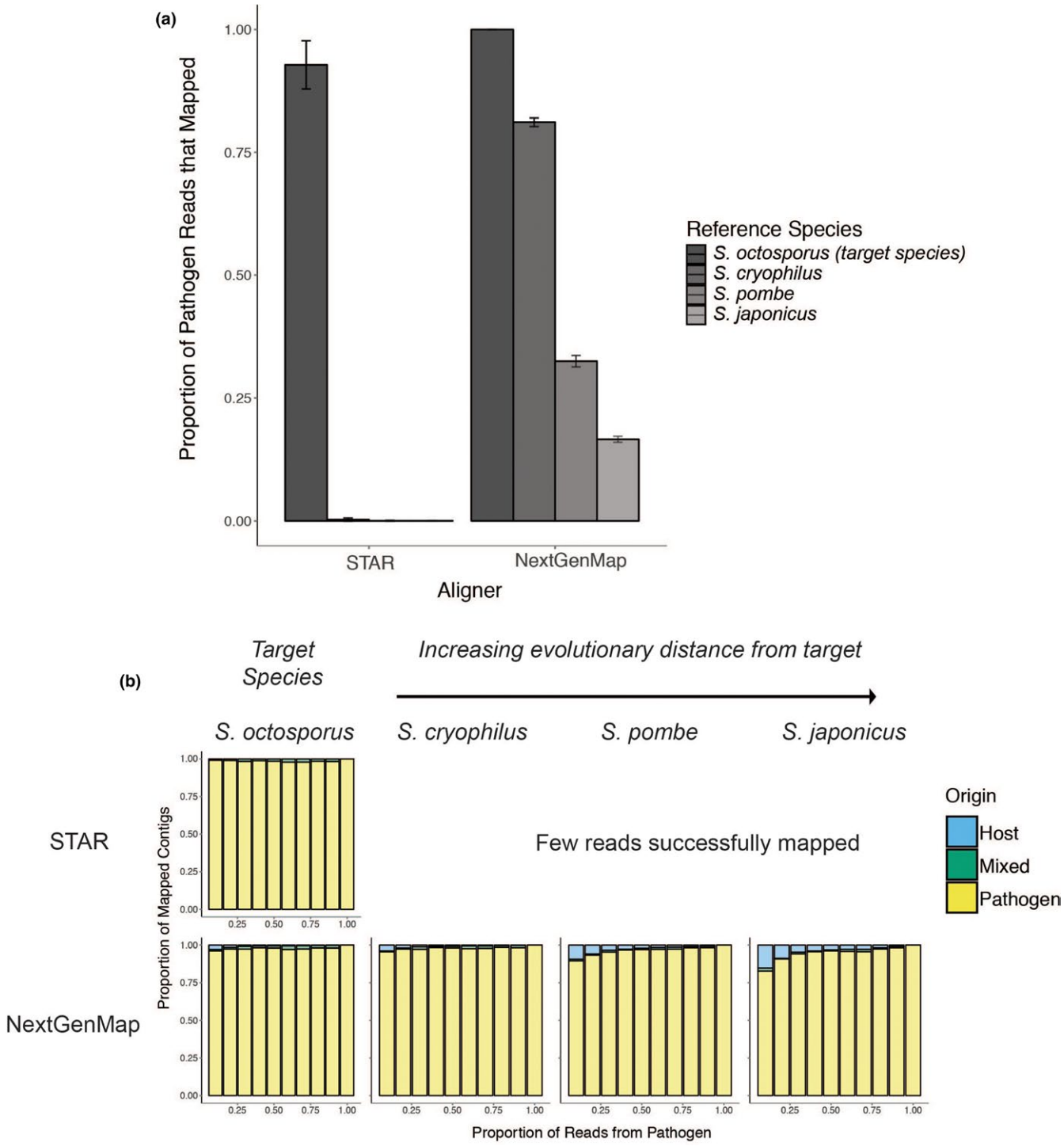

FIGURE 4 The Assembly Approach improves alignment of dual RNA-seq reads. Results are shown for aligners STAR and NextGenMap and four *Schizosaccharomyces* reference genomes. (a) Bars indicate the proportion of pathogen reads that mapped to each genome/transcriptome. Alignments to target pathogen species are shown in black, and greyscale gradient indicates evolutionary distance from target. STAR can only map pathogen contigs when mapping to reference genome of the pathogen and is unable to effectively map contigs to reference genomes of related species. NextGenMap retains its ability to map pathogen contigs to transcriptomes of related species. (b) Origin of assembled reads that mapped to genome of pathogen or closely related species. For each bar plot, blue contigs are those that originated from host (*Arabidopsis thaliana*) and yellow contigs are those that originated from pathogen (*Schizosaccharomyces octosporus*). Contigs that were unable to be determined as comprised of host or pathogen reads are coloured green. The bar plots represent composition of contigs that mapped to each reference genome. STAR was unable to align more than a few contigs for *S. cryophilus*, *S.pombe* and *S. japonicus*, so those plots have been excluded. Host mismapping was strongly reduced

pathogen reads in the original datasets, <1% of contigs comprised of host reads mapped with NextGenMap. Furthermore, c. 99% of contigs comprised of pathogen reads mapped to the target transcriptome of *S. octosporus*. Approximately 79% of contigs comprised of pathogen reads mapped to *S. cryophilus*, 29% of contigs comprised of pathogen reads mapped to *S. pombe* and 15% of contigs comprised of pathogen reads mapped to *S. japonicus*. Although some of the contigs that were unable to be identified as comprised of pathogen or host reads mapped, they comprised a minority of the total number of mapped contigs, with a maximum of 2.5%. With such a reduction in host read mismapping, the majority of all mapped contigs originated from the pathogen. Given that a good de novo assembly is possible, the assembly approach also clearly reduced mismapping.

3.5 | Effect of sequencing parameters

To investigate the effect of sequencing parameters—the size of sequencing read, paired-end vs. single-end reads—on the above approaches, we simulated dual RNA-seq datasets of the same system with a longer read length (150 bp) and with paired-end sequencing (Figure S11–S16). While the same patterns largely held—that raw read mapping resulted in host reads mismapping when aligning with STAR and NextGenMap and mapping reads to a concatenated genome or assembling reads de novo prior to mapping substantially reduced host read mismapping—there were some differences among the layouts. Specifically, longer read lengths not only resulted in overall lower host read mismapping rates (which is consistent with the results of mapping assembled reads) but also resulted in lower pathogen read mapping rate, especially when mapping to the genomes of species closely related to the pathogen. Additionally, assembling 76 paired-end reads de novo prior to mapping to the *Schizosaccharomyces* genomes resulted in more reads that could not be identified as from the host or pathogen (potentially chimeras) comprising the group of mapped reads. As expected, longer paired-end reads generally performed better than other configurations.

3.6 | Similarities and contrasts in other taxa

The simulations above focused on a fungal 'parasite' infecting a plant host as we believe that this could be a particularly problematic scenario, as plants and fungi are both eukaryotes, while other host–pathogen systems involve more diverged species. To investigate if and how the patterns observed above extend to other systems, we simulated two more sets of dual RNA-seq datasets. We simulated datasets across the same range of proportion of pathogen reads for another host–fungal pathogen system, *Homo sapiens* and *Candida albicans*, as well as a bacterial pathogen system, *Homo sapiens* and *Escherichia coli*. While alignments of the Human-*Candida* raw reads to the genomes of the target species, *C. albicans*, and two closely related species, *C. dublinensis* and *C. parapsilosis*, did result in a comparable level of host read mismapping to the above analyses (Figure S6), alignments of Human-*E. coli* raw reads had minimal if any host read mismapping (Figure S10). We conducted the same three approaches described above to minimize host read mismapping with the Human-*Candida* alignments, and we observed the same results as described above—that mapping to concatenated genomes of the host and closely related species, as well as de novo assembly prior to aligning, substantially reduce host read mismapping while retaining the ability to map pathogen reads (Figure S7–9). Therefore, the mismapping problems in dual RNA-seq held, to varying degrees, across the systems we investigated.

4 | DISCUSSION

Understanding the genetic mechanisms of host–pathogen interactions may be improved using dual RNA-seq (Westermann et al. 2012), but several limitations have contributed to the underutilization of this approach. Dual RNA-seq is inherently a mixture of host and pathogen reads that need to be parsed prior to analyses. This parsing relies on mapping reads to the genomes of each organism. Consequently, it was previously unknown whether and how dual RNA-seq could be applied to non-model host–pathogen systems, in which there are limited or no genomic resources. Our analyses of simulated sequencing identified as problematic several approaches that might be encountered by researchers applying dual RNA-seq to non-model host–pathogen systems. However, our systematic comparison of analytical approaches also revealed a workflow that can be used to identify the genetic mechanisms of host–parasite interactions for non-model organisms.

For non-model organisms, traditional approaches to analysing RNA-seq data include mapping reads to reference genomes of related species (Benjamin, Nichols, Burke, Ginsburg, & Lucas, 2014). Depending on the software for aligning sequencing reads, we found that using genomic resources of a closely related pathogen can result in one of two error modes. We found that aligners like TopHat2 and MapSplice2 are too restrictive with allowed mismatches by default, resulting in pathogen reads failing to map to genomes of the closely related species. In contrast, aligners such as NextGenMap and STAR are too lenient, allowing for too many mismatches by default and resulting in the mismapping of host reads to the genome of the species closely related to the pathogen. This was consistent when investigating simulated dual RNA-seq datasets of a plant (*Arabidopsis* thaliana) and fungus (*Schizosaccharomyces octosporus*) as well as simulated dual RNA-seq datasets of human and fungal pathogen *Candida albicans*. In contrast, host read mismapping was substantially reduced for a simulated dataset of human and bacterial pathogen, *Escherichia coli*, suggesting that this mapping inaccuracy may be a particular concern for studies of fungal pathogens.

The difference in the performance of the alignment tools likely reflects the classic trade-off between precision vs. sensitivity in the underlying algorithms. TopHat2 underutilizes the dual RNA-seq data as it relies on a Burrows–Wheeler transform and FM-index to quickly search for matches between the reference genome and the RNA-seq reads. This emphasis on fast and exact (or nearly so) match alignment struggles to map reads that have divergent bases. MapSplice2, a splice-aware aligner, applies a metric based on Shannon maximum entropy as applied to a weighted de Bruijn graph. This approach can detect splice junctions

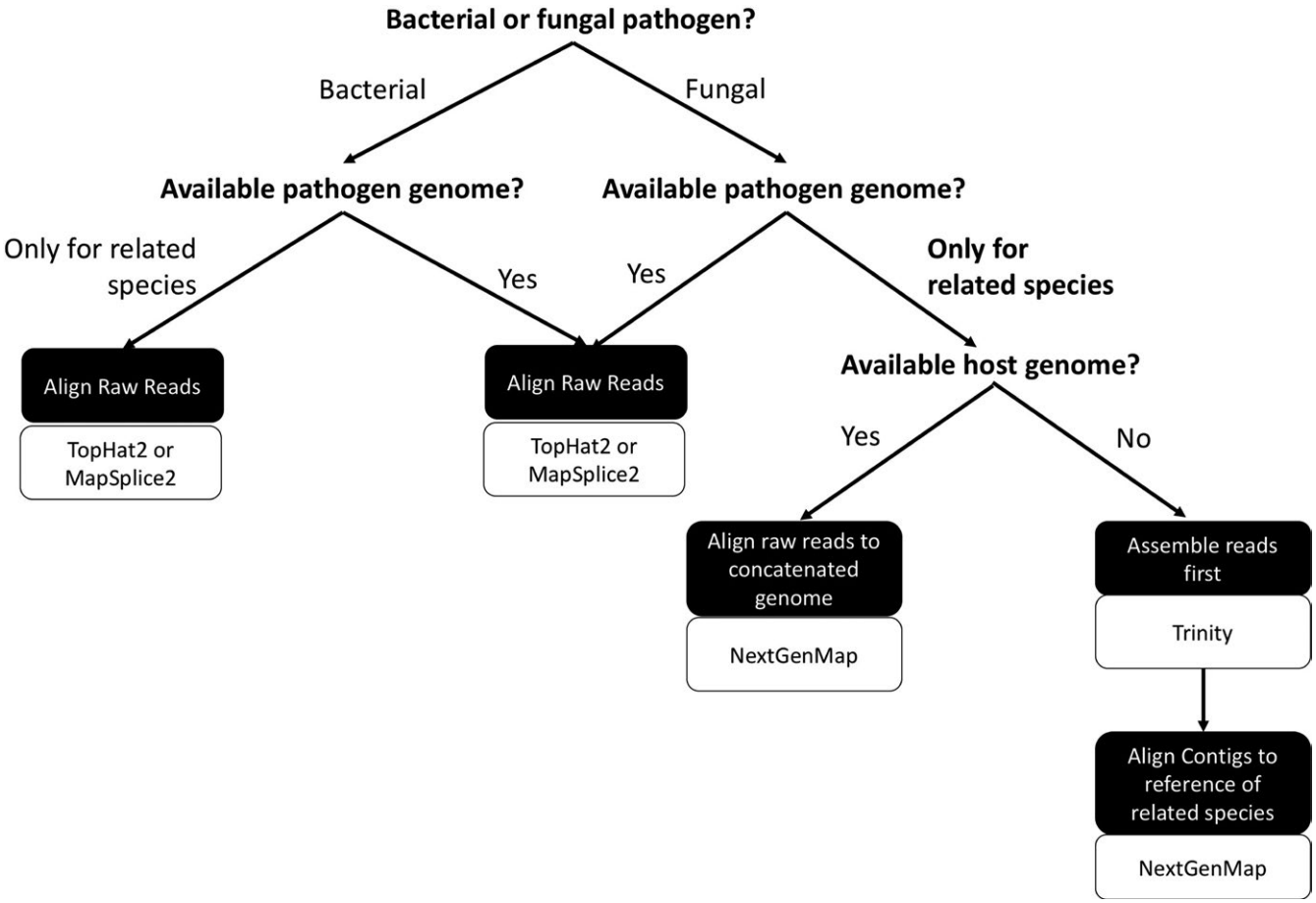

FIGURE 5 Suggested Workflow for dual RNA-seq Experiments of non-model host–pathogen systems

without any dependence on splice site features—potentially a critical feature when applying dual RNA-seq to poorly annotated genomes or closely related species. However, the base alignment approach relies on Bowtie algorithm and as a result suffers from the same limitations as TopHat2. In contrast, the hash-based variable mismatch threshold algorithm of NextGenMap maximizes its ability to utilize divergent reads but makes more erroneous assignments. STAR, which uses a seed and anchor approach based on a Maximal Mappable Prefix, is robust to non-continuous reads and some mismatches. It performed almost as well as NextGenMap in terms of data utilization, but again suffered from imprecision in the form of host reads mismapping to the wrong genome.

Host reads mismapping to the genome of the pathogen or the genome of a species closely related to the pathogen can have severe implications for the characterization of the gene expression profile of the pathogen during the infection process. Differential gene expression analyses between alignments of the same dual RNA-seq dataset that were produced by mapping raw reads to the target genome with TopHat2 (in which host reads did not mismap) and STAR (in which host reads did mismap) indicated that the alignments produced with STAR had overall higher levels of overexpression than that produced by TopHat2. This highlights that biological insight gained from dual RNA-seq data can be inaccurate if certain steps are not taken. As real sequencing is unable to definitively identify the species origin of transcripts, neglecting to take measures to avoid host reads mismapping would result in inaccurate genetic mechanisms implicated in the infection process.

From our results, we propose a workflow that should be followed to determine the best approaches to extending the use of dual RNA-seq to a wider array of systems, including non-model systems (specifically, eukaryotic host–fungal pathogen systems) in which genomic resources are available for the species closely related to the pathogen of interest (Figure 5).

- If a host genome is available, concatenating the genome of the host with the genome of species closely related to the pathogen of interest (place-to-go approach) results in more accurate alignments, in which host read mismapping is substantially reduced, with the aligners, STAR and NextGenMap.
- If a host genome is not available, assembling reads de novo, prior to aligning with NextGenMap and STAR decreased host read mismapping while retaining the ability to map pathogen reads. As expected, de novo assemblies do exclude some reads, which would consequently not be quantified. The reads excluded, however, were rare and a subset of them was low-complexity sequences (Figure S2).

As NGS technologies and their analytical tools continue to become more affordable and accessible, it is important to critically

assess how accurate genomic analyses with these tools are. While dual RNA-seq has been applied to many model disease systems, we remain woefully unaware of how the accuracy of mapping methods utilized to separate host and pathogen reads affects dual RNA-seq studies. The methods we used here allowed us to assess the accuracy of alignment approaches of dual RNA-seq in non-model systems through simulated sequencing, but the biological truth of the origin of transcripts in real dual RNA-seq data would remain unknown and the issues we identified would lead to misinterpretations of the data. As infectious diseases are expected to increase in the coming years, it is critical that we investigate proper methods of analyses to ensure accurate insights are gained as systems are explored.

ACKNOWLEDGEMENTS

We thank Charles E. Mitchell and Fletcher Halliday for helpful comments. We acknowledge the thorough and insightful suggestions of two reviewers on earlier drafts that led to a number of analyses in the work presented above. This work was supported by the NSF-USDA joint program in Ecology and Evolution of Infectious Diseases (USDA-NIFA AFRI grant 2016-67013-25762) and the University Cancer Research Fund. KRO was supported by graduate research fellowships from the Triangle Center for Evolutionary Medicine and the National Science Foundation.

AUTHORS' CONTRIBUTIONS

K.R.O. and C.D.J. conceived ideas and designed methodology; K.R.O. performed analyses; K.R.O. led writing of the manuscript. All authors contributed critically to drafts and gave final approval for publication.

DATA ACCESSIBILITY

Simulated sequencing data available from the Dryad Digital Repository https://doi.org/10.5061/dryad.t40nj78 (O'Keeffe & Jones, 2018).

REFERENCES

Aprianto, R., Slager, J., Holsappel, S., & Veening, J.-W. (2016). Time-resolved dual RNA-seq reveals extensive rewiring of lung epithelial and pneumococcal transcriptomes during early infection. *Genome Biology*, *17*(1), 198. https://doi.org/10.1186/s13059-016-1054-5

Baddal, B., Muzzi, A., Censini, S., Calogero, R., Torricelli, G., Guidotti, S., … Paxxicoli, A. (2015). Dual RNA-seq of nontypeable haemophilus influenzae and host cell transcriptomes reveals novel insights into host-pathogen cross talk. *mBio*, *6*(e01765–15). https://doi.org/10.1128/mbio.01765-15

Baruzzo, G., Hayer, K. E., Kim, E. J., Di Camillo, B., FitzGerald, G. A., & Grant, G. R. (2016). Simulation-based comprehensive benchmarking of RNA-seq aligners. *Nature Methods*, *14*, 135. https://doi.org/10.1038/nmeth.4106

Benjamin, A. M., Nichols, M., Burke, T. W., Ginsburg, G. S., & Lucas, J. E. (2014). Comparing reference-based RNA-Seq mapping methods for non-human primate data. *BMC Genomics*, *15*(1), 570. https://doi.org/10.1186/1471-2164-15-570

Choi, Y.-J., Aliota, M. T., Mayhew, G. F., Erickson, S. M., & Christensen, B. M. (2014). Dual RNA-seq of parasite and host reveals gene expression dynamics during filarial worm–mosquito interactions. *PLOS Neglected Tropical Diseases*, *8*(5), e2905. https://doi.org/10.1371/journal.pntd.0002905

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., … Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics. 29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Ekblom, R., & Galindo, J. (2010). Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, *107*(1), 1–15. https://doi.org/10.1038/hdy.2010.152

Ellison, A. R., DiRenzo, G. V., McDonald, C. A., Lips, K. R., & Zamudio, K. R. (2017). First in vivo batrachochytrium dendrobatidis; transcriptomes reveal mechanisms of host exploitation, host-specific gene expression, and expressed genotype shifts. *G3: Genes|Genomes|Genetics*, *7*(1), 269, LP-278. http://www.g3journal.org/content/7/1/269.abstract

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., … Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotech*, *29*(7), 644–652. https://doi.org/10.1038/nbt.1883

Griebel, T., Zacher, B., Ribeca, P., Raineri, E., Lacroix, V., Guigó, R., & Sammeth, M. (2012). Modelling and simulating generic RNA-Seq experiments with the flux simulator. *Nucleic Acids Research*, *40*(20), 10073–10083. https://doi.org/10.1093/nar/gks666

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., … Regev, A. (2013). De novo transcript sequence reconstruction from RNA-Seq: Reference generation and analysis with Trinity. *Nature Protocols*, *8*, 1494–1512. https://doi.org/10.1038/nprot.2013.084

Hayden, K. J., Garbelotto, M., Knaus, B. J., Cronn, R. C., Rai, H., & Wright, J. W. (2014). Dual RNA-seq of the plant pathogen Phytophthora ramorum and its tanoak host. *Tree Genetics & Genomes*, *10*(3), 489–502. https://doi.org/10.1007/s11295-014-0698-0

Kawahara, Y., Oono, Y., Kanamori, H., Matsumoto, T., Itoh, T., & Minami, E. (2012). Simultaneous RNA-seq analysis of a mixed transcriptome of rice and blast fungus interaction. *PLoS ONE*, *7*(11), e49423. https://doi.org/10.1371/journal.pone.0049423

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, *26*(5), 589–595. https://doi.org/10.1093/bioinformatics/btp698

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., & Homer, N., … Subgroup, 1000 Genome Project Data Processing. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Liao, Y., Smyth, G. K., & Shi, W. (2013). The Subread aligner: Fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Research*, *41*(10), e108–e108. https://doi.org/10.1093/nar/gkt214

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Meth*, *5*(7), 621–628. https://doi.org/10.1038/nmeth.1226

O'Keeffe, K. R., & Jones, C. D. (2018). Data from: Challenges and solutions for analyzing dual RNA-seq data for non-model host/pathogen

systems. *Dryad Digital Repository*. https://doi.org/10.5061/dryad.t40nj78

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Rhind, N., Chen, Z., Yassour, M., Thompson, D. A., Haas, B. J., Habib, N., … Nusbaum, C. (2011). Comparative functional genomics of the fission yeasts. *Science*, *332*(6032), 930. LP-936. http://science.sciencemag.org/content/332/6032/930.abstract

Sedlazeck, F. J., Rescheneder, P., & von Haeseler, A. (2013). NextGenMap: Fast and accurate read mapping in highly polymorphic genomes. *Bioinformatics*, *29*(21), 2790–2791. https://doi.org/10.1093/bioinformatics/btt468

Teixeira, P. J. P. L., Thomazella, D. P. D. T., Reis, O., do Prado, P. F. V., do Rio, M. C. S., Fiorin, G. L., … Pereira, G. A. G. (2014). High-resolution transcript profiling of the atypical biotrophic interaction between theobroma cacao and the fungal pathogen moniliophthora perniciosa. *The Plant Cell Online*, *26*(11), 4245–4269. https://doi.org/10.1105/tpc.114.130807

Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics*, *25*(9), 1105–1111. https://doi.org/10.1093/bioinformatics/btp120

Wang, K., Singh, D., Zeng, Z., Coleman, S. J., Huang, Y., Savich, G. L., … Liu, J. (2010). MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Research*, *38*(18), e178–e178. https://doi.org/10.1093/nar/gkq622

Westermann, A. J., Förstner, K. U., Amman, F., Barquist, L., Chao, Y., Schulte, L. N., … Vogel, J. (2016). Dual RNA-seq unveils noncoding RNA functions in host–pathogen interactions. *Nature*, *529*, 496. https://doi.org/10.1038/nature16547

Westermann, A. J., Gorski, S. A., & Vogel, J. (2012). Dual RNA-seq of pathogen and host. *Nature Reviews Microbiology*, *10*, 618–630. https://doi.org/10.1038/nrmicro2852

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: O'Keeffe KR, Jones CD. Challenges and solutions for analysing dual RNA-seq data for non-model host–pathogen systems. *Methods Ecol Evol*. 2019;10:401–414. https://doi.org/10.1111/2041-210X.13135