Received: 28 July 2021 | Revised: 10 January 2022 | Accepted: 24 January 2022

DOI: 10.1111/1755-0998.13593

MOLECULAR ECOLOGY RESOURCES WILEY

RESOURCE ARTICLE

TransPi—a comprehensive TRanscriptome ANalysiS PIpeline for *de novo* transcriptome assembly

Ramón E. Rivera-Vicéns[1] | Catalina A. Garcia-Escudero[1,2] | Nicola Conci[1] | Michael Eitel[1] | Gert Wörheide[1,3,4]

[1]Department of Earth and Environmental Science, Paleontology & Geobiology, Ludwig-Maximilians-Universität München, München, Germany

[2]Graduate School for Evolution, Ecology and Systematics, Faculty of Biology, Ludwig-Maximilians-Universität München, Planegg-Martinsried, Germany

[3]GeoBio-Center, Ludwig-Maximilians-Universität München, München, Germany

[4]SNSB-Bayerische Staatssammlung für Paläontologie und Geologie, München, Germany

Correspondence
Gert Wörheide, Department of Earth and Environmental Sciences, Paleontology & Geobiology, Ludwig-Maximilians-Universität München, München, Germany.
Email: woerheide@lmu.de

Funding information
Horizon 2020 Framework Programme, Grant/Award Number: 764840

Handling Editor: Alana Alexander

Abstract

The use of RNA sequencing (RNA-Seq) data and the generation of *de novo* transcriptome assemblies have been pivotal for studies in ecology and evolution. This is especially true for nonmodel organisms, where no genome information is available. In such organisms, studies of differential gene expression, DNA enrichment bait design and phylogenetics can all be accomplished with *de novo* transcriptome assemblies. Multiple tools are available for transcriptome assembly, but no single tool can provide the best assembly for all data sets. Therefore, a multi-assembler approach, followed by a reduction step, is often sought to generate an improved representation of the assembly. To reduce errors in these complex analyses while at the same time attaining reproducibility and scalability, automated workflows have been essential in the analysis of RNA-Seq data. However, most of these tools are designed for species where genome data are used as reference for the assembly process, limiting their use in nonmodel organisms. We present TransPi, a comprehensive pipeline for *de novo* transcriptome assembly, with minimum user input but without losing the ability of a thorough analysis. A combination of different model organisms, k-mer sets, read lengths and read quantities was used for assessing the tool. Furthermore, a total of 49 nonmodel organisms, spanning different phyla, were also analysed. Compared to approaches using single assemblers only, TransPi produces higher BUSCO completeness percentages, and a concurrent significant reduction in duplication rates. TransPi is easy to configure and can be deployed seamlessly using Conda, Docker and Singularity.

KEYWORDS
annotation, assembly, *de novo*, Nextflow, nonmodel, pipeline, RNA-Seq, transcriptome

1 | INTRODUCTION

In recent decades, technology improvements have rendered next generation sequencing (NGS) a robust and cost-effective technique of wide applicability in research fields that require large-scale DNA sequencing. Among the different NGS-based approaches, RNA sequencing (RNA-Seq) allows the generation of the so-called transcriptomes *de novo* (i.e., without the need for a reference genome). Transcriptomes are applicable for several downstream applications, including the analysis of differential gene expression (Pita et al., 2018), gene model prediction (Chan et al., 2017), DNA enrichment bait design (Quek et al., 2020), genome annotation (Holt & Yandell,

2011; Testa et al., 2015), detection of whole-genome duplication (Yang et al., 2019) and phylogenetics (Cheon et al., 2020; Lozano-Fernandez et al., 2019).

Various software has been developed for the generation of *de novo* transcriptome assembly. Commonly used tools include TRINITY (Grabherr et al., 2011), RNASPADES (Bushmanova et al., 2019), TRANSABYSS (Robertson et al., 2010), and SOAPDENOVO-TRANS (Xie et al., 2014). However, a recent study compared 10 assemblers with nine data sets (i.e., different species and samples) and demonstrated that the performance of each tool varies by data set; no single tool was able to generate optimal assemblies for all data sets (Hölzer & Marz, 2019). The assembler's performance measurement was based on a combination of biological-based measures (e.g., number of Benchmarking Universal Single-Copy Orthologs—BUSCO), and reference-free measures (e.g., TRANSRATE's optimal assembly score; Smith-Unna et al., 2016). Therefore, combining multiple assemblers probably represents a valuable approach to increase the quality of reference assemblies (Lu et al., 2013). Additionally, factors such as read length and number also play important roles in the assembly process (Francis et al., 2013; Grabherr et al., 2011; Schulz et al., 2012).

Transcriptome *de novo* assemblies tend to produce thousands to hundreds of thousands of different transcripts of which a significant amount can be misassembled (Bushmanova et al., 2019; Schulz et al., 2012). To reduce the complexity within a transcriptome and to identify true transcripts and isoforms, one common approach is to remove duplicated and misassembled sequences. Clustering methods are often employed for this, where similar transcripts are combined into groups. One of the tools commonly used for clustering transcripts is CDHIT-EST (Fu et al., 2012), which tends to keep the longest transcripts only. However, clustering and selecting for the longest transcripts is not always the best strategy (Gilbert, 2013) since they often result from misassemblies (i.e., not real transcripts) and may include frameshift errors. On the other hand, tools such as EVIDENTIALGENE (Gilbert, 2013, 2019) use a combination of clustering and classification methods (i.e., sequence features such as coding sequence [CDS] content and length) to generate a nonredundant consensus assembly. The latter approach is more accurate for the cost of longer computing time and higher computation demands (e.g., higher memory usage). Combining multiple assemblers with a thorough reduction of each assembly individually thus increases the complexity of the analyses.

The ideal path to optimal reference transcriptomes should, therefore, include the use of multiple assemblers, followed by thorough filtering of each assembly individually. Generating, combining and filtering all resulting assemblies step by step individually (cf. Cerveau, & Jackson, 2016; MacManes, 2018) is impractical, of limited reproducibility and can be prone to human error. Consequently, the design of streamlined RNA-Seq analysis pipelines has gained popularity in recent years. However, most of these pipelines require a reference genome for the transcriptome assembly (i.e., reference-guided assembly) and are, consequently, not suitable for *de novo* approaches (Cornwell et al., 2018; D'Antonio et al., 2015; Kohen et al., 2019; Martin et al., 2010; Wang, 2018; Zhang & Jonassen, 2020). This represents a major limitation for transcriptomics in nonmodel organisms, where genome reference data are usually lacking.

To address these shortcomings, we developed TransPi, a comprehensive Transcriptome ANalysiS Pipeline, for *de novo* transcriptome assembly. TransPi is implemented using the scientific workflow manager NEXTFLOW (Di Tommaso et al., 2017), which provides a user-friendly environment, easy deployment, scalability and reproducibility. TransPi performs all steps of standard RNA-Seq analysis workflows, from raw read quality control up to annotation against multiple databases (e.g., SwissProt, PFAM). To reduce possible biases, duplication and misassemblies, TransPi utilizes various assemblers and k-mers (i.e., k length sequences used for the assembly) to generate an over-assembled transcriptome that is then reduced to a nonredundant consensus transcriptome with the software EVIDENTIALGENE (Gilbert, 2013, 2019). Here we show that, when compared to approaches using single assemblers only, TransPi produces higher BUSCO completeness percentages, and a concurrently significant reduction in duplication rates (i.e., higher single-copy genes). Higher BUSCO scores in the complete and single-copy categories indicate a less erroneous consensus assembly (Simão et al., 2015; Waterhouse et al., 2011).

In sum, TransPi provides a useful resource for the generation of *de novo* transcriptome assemblies, with minimum user input but without losing the ability of a thorough analysis. TransPi and all documentation is available at https://github.com/palmuc/TransPi.git.

2 | METHODS

2.1 | Pipeline implementation and configuration

TransPi is based on the scientific workflow manager NEXTFLOW (Di Tommaso et al., 2017). The pipeline is easy to configure and can be deployed using the package management system Conda, Docker, Singularity or cloud environments (e.g., AWS). Real-time monitoring of the pipeline can be performed by using Nextflow Tower with no modification needed to the TransPi script. Deployment of TransPi in computing clusters is accomplished by the native communication of Nextflow with scheduling managers such as SLURM, PBS and Torque. TransPi can deploy hundreds of jobs depending on user configurations and needs. Multiple data sets can be run in parallel given that enough computing resources are available. Running time of the pipeline is dependent on factors such as number of data sets, read quantity, k-mer selection, complexity of the transcriptome being assembled, and user-specified additional options selected (filtration, SIGNALP, etc.). TransPi consists of two major components: a precheck script to install dependencies, and the main script to run the assemblers, perform the reduction and transcriptome annotation.

2.2 | Precheck script

TransPi integrates several programs and external databases (e.g., SwissProt, Boeckmann et al., 2003; PFAM, El-Gebali et al., 2018) for the generation and annotation of the reference transcriptome. To facilitate the setup of all necessary dependencies, TransPi includes

an installation script. This will first install, if necessary, the Conda package management system, all dependencies, and download and configure the required databases. The script is designed to recognize when a previous run of the script was performed, thus preventing the repetition of previous steps. Another advantage of the precheck script is that it will automatically create the configuration file needed by Nextflow to execute the pipeline with all the necessary information. As a result, the user will only have to make some minor changes to the file (e.g., number of allocated threads, the amount of working memory, scheduling manager, node names and queue) before running the pipeline. Essentially, the precheck has to be run entirely only once for the dependencies and database installation. Subsequent pipeline runs can be done with the same configuration file. Auxiliary scripts for the automated update of the databases such as PFAM and SwissProt are also provided.

2.3 | Main script

A diagram of TransPi version 1.0.0 is shown in Figure 1. First, reads are checked for adapter presence, low-quality bases and over-represented sequences with FASTQC version 0.11.9 (Andrews, 2010). Filtration of the reads (by default reads with an average phred quality >5 are kept, see MacManes, 2014) and trimming of adapters (if present) is performed with FASTP version 0.20.1 (Chen et al., 2018). Optionally, removal of rRNA is performed with SORTMERNA version 4.2.0 (Kopylova et al., 2012). Filtered reads are subsequently normalized before being assembled (Grabherr et al., 2011). The assembly step combines five different assemblers and uses multiple k-mer lengths. The assemblers used by TransPi are RNASPADES version 3.14.0 (Bushmanova et al., 2019), TRANS-ABYSS version 2.0.1 (Robertson et al., 2010), SOAPDENOVOTRANS version 1.03 (Xie et al., 2014), TRINITY version 2.9.1 (Grabherr et al., 2011) and VELVET version 1.2.12/OASES version 0.2.09 (Schulz et al., 2012; Zerbino & Birney, 2008). All assemblers, but TRINITY, will use the k-mer list to produce individual assemblies per k-mer. After the assembly stage is performed, the combined transcriptomes (i.e., all assemblers and k-mers) are reduced with EVIDENTIALGENE v2019.05.14 (Gilbert, 2013, 2019). Briefly, EVIDENTIALGENE will merge perfect duplicates, cluster protein sequences and perform local similarities searches between the transcripts using BLAST version 2.2.31 (Altschul et al., 1997) (for more details see Gilbert, 2019).

Next, TransPi uses the nonredundant reference transcriptome to run several downstream analyses commonly applied to *de novo* transcriptomes projects: (i) RNAQUAST version 2.0.1 for quality assessment (Bushmanova et al., 2016), (ii) BOWTIE2 version 2.3.5.1 to map the reads against the transcriptome (Langmead & Salzberg, 2012), (iii) BUSCO (Simão et al., 2015; versions 3 and 4) to quantitatively assess the completeness in terms of expected universal single copy gene content, (iv) TRANSDECODER version 5.5.0 (https://transdecoder.github.io) to identify open reading frames (ORFs), with the option to perform homology searches of all ORFs to known proteins via BLAST, in order to retain ORFs that may have functional significance but do not pass the coding likelihood scores, and (v) TRINOTATE version 3.2.0 (Bryant et al., 2017) to provide automatic functional annotation.

By using DIAMOND version 0.9.30 (Buchfink et al., 2015), the similarity searches of the transcripts used for the annotation step against the SwissProt and UniProt databases (chosen by the user) are accelerated. RNAMMER version 1.2 (Lagesen et al., 2007), TMHMM version 2.0 (Krogh et al., 2001) and SIGNALP 4.1 (Petersen et al., 2011) are used to search for rRNA, signal peptide proteins and transmembrane domain prediction, respectively. Protein domain searches are performed with HMMER version 3.3 (Finn et al., 2011) against the latest version of the PFAM database. All this information is combined into an annotation report which includes: (i) information on Gene Ontology (GO); (ii) evolutionary genealogy of genes: Non-supervised Orthologous Groups (eggNOG), and (iii) Kyoto Encyclopedia of Genes and Genomes (KEGG). It also contains the similarity search against SwissProt and the user-specified UniProt database. TransPi will also produce a custom Hypertext Markup Language (HTML) report that summarizes the steps and provides interactive plots for straightforward exploration of the data. Plots from the interactive report can also be saved in SVG format. Other plots are also saved automatically (PDF and SVG) in the results directory generated by the pipeline. Altogether, TransPi provides the user with the ability to assess and evaluate the final assembly and to compare it to other commonly used methods for reference transcriptome generation (e.g., a TRINITY-only assembly).

2.4 | K-mer selection, read length effect and chimera detection

To test the performance of the pipeline and the effect of k-mer selection (Prjibelski et al., 2020), read quantities and read lengths, data sets from the model organisms *Caenorhabditis elegans*, *Drosophila melanogaster* and *Mus musculus* were used. These species were selected given the vast amount of transcriptomic data available with various read length and quantities (Table S1). Three k-mer sets (A, B, C) depending on read length were designed, since the selection of this parameter will modify how the assembly graph is constructed (Table S2). For the read length test, data consisting of paired-end reads of 50, 75, 100 and 150 bp (Table S1) were analysed. All statistical analyses, such as ANOVA and Kruskal–Wallis test, were performed in R (version 3.6.2).

To measure the percentage of chimeric transcripts and transcript accuracy, a similar approach to Kerkvliet et al. (2019) was used. First, gene sets for the model organisms *C. elegans* (i.e., c_elegans.PRJNA13758.WS279.mRNA_transcripts.fa from Wormbase), *D. melanogaster* (i.e., Dmel-all-transcript-r6.39.fasta from Flybase) and *M. musculus* (i.e., GCF_000001635.27_GRCm39_rna_from_genomic.fna from NCBI) were downloaded. Then a BLASTN search was performed using the transcriptomes from TransPi and TRINITY against each corresponding gene set. Parameters used were as specified by Kerkvliet et al. (2019) (perc_identity.90 -evalue.001). BLASTN output was filtered using a minimum length of 300 bp for each match.

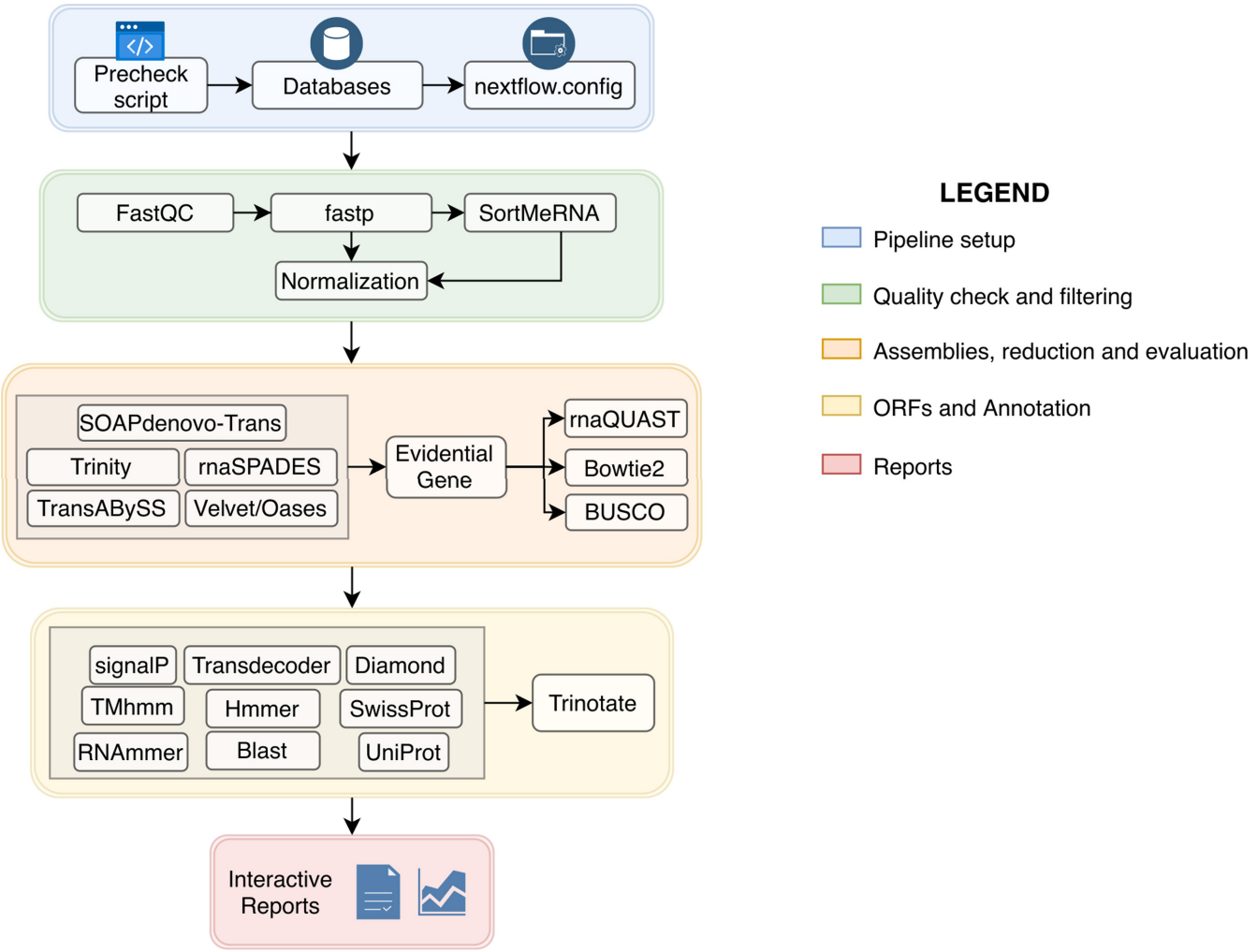

FIGURE 1 TransPi version 1.0.0 flowchart showing the various steps and analyses it can perform. For simplicity, this diagram does not show all the connections between the processes. Also, it omits other additional options such as the BUSCO distribution and transcriptome filtering with PSYTRANS (see Section 2.6). ORFs, open reading frames

Nonchimeric transcripts were identified as transcripts with one match per gene. Transcripts with two or more matches were classified as chimeras.

2.5 | TransPi on nonmodel organisms

Given TransPi is focused on species with no (or scarce) genome information, the pipeline was tested with multiple species from different phyla (Table 1). K-mer set C was applied for all assemblies of nonmodel organisms as this gave the best BUSCO percentages (see Section 3). The nonmodel organism data sets consisted of various read lengths ranging from 50 to 150 bp. Given TRINITY is by far the most commonly used *de novo* transcriptome assembly tool (by number of citations and excluding usage of some tools for genome assemblies), the performance of TransPi was evaluated by comparing the output transcriptome to the TRINITY assembly for each data set.

2.6 | Additional options

Various additional options were implemented in TransPi to obtain more insight into the transcriptomes being assembled. One of these options is filtering symbionts and/or contaminants from the assembly using the Parasite & Symbiont Transcriptome Separation software (PSYTRANS) (https://github.com/sylvainforet/psytrans). The filtration step was tested with the data set of the coral *Porites pukoensis* (accession SRR8491966) using sequences of its symbiont *Symbiodinium microadriaticum* (Uniprot Taxon Identifier: 2951) and sequences of the order Scleractinia (Uniprot Taxon Identifier: 6125) as host. Another option of TransPi examines the presence and absence of BUSCO genes in all the generated assemblies and creates a heatmap of gene distribution. This option was tested with the epadomorph barnacle *Octolasmis warwickii* data set (SRR10527303) given the difference between TransPi and TRINITY BUSCO scores for the missing category (see Section 3).

TABLE 1 Nonmodel organism data sets used in this study

Phylum	Class	Order	Species	SRA	No. of reads	Length (bp)
Cnidaria	Anthozoa	Alcyonacea	*Pinnigorgia flava*	ERR3026433	30,545,400	50
Cnidaria	Anthozoa	Alcyonacea	*Sinularia cruciata*	ERR3026434	22,160,908	50
Cnidaria	Anthozoa	Alcyonacea	*Tubipora musica*	ERR3026435	23,006,724	50
Cnidaria	Anthozoa	Helioporacea	*Heliopora coerulea*	ERR3040053	29,000,821	50
Cnidaria	Anthozoa	Scleractinia	*Acropora palmata*	SRR5569439	10,476,071	75
Cnidaria	Anthozoa	Scleractinia	*Acropora pulchra*	SRR8601367	14,037,157	75
Cnidaria	Anthozoa	Scleractinia	*Porites pukoensis*	SRR8491966	16,448,725	150
Cnidaria	Hydrozoa	Anthoathecata	*Millepora alcicornis*	SRR4294206	24,645,545	150
Porifera	Homoscleromorpha	Homosclerophorida	*Oscarella pearsei*	SRR1042012	11,306,242	100
Porifera	Homoscleromorpha	Homosclerophorida	*Corticium candelabrum*	SRR504694	18,897,095	150
Porifera	Demospongiae	Spongillida	*Ephydatia muelleri*	SRR1041944	11,425,188	100
Porifera	Demospongiae	Spongillida	*Spongilla lacustris*	SRR1168575	5,136,881	100
Porifera	Demospongiae	Poecilosclerida	*Mycale phylophylla*	SRR1711043	11,408,543	100
Porifera	Demospongiae	Haplosclerida	*Haliclona tubifera*	SRR1793376	16,356,602	100
Porifera	Demospongiae	Dictyoceratida	*Ircinia fasciculate*	SRR7655554	13,420,109	100
Porifera	Calcarea	Leucosolenida	*Sycon coactum*	SRR504690	9,098,097	100
Mollusca	Gastropoda	Trochida	*Monodonta labio*	SRR1505119	10,388,770	100
Mollusca	Bivalvia	Pholadomyoida	*Lyonsia floridana*	SRR1560310	9,919,645	100
Mollusca	Bivalvia	Veneroida	*Mercenaria campechiensis*	SRR1560359	11,935,267	100
Mollusca	Bivalvia	Trigoniida	*Neotrigonia margaritacea*	SRR1560432	11,215,767	100
Mollusca	Bivalvia	Veneroida	*Cardites antiquatus*	SRR1560458	11,916,756	100
Mollusca	Bivalvia	Veneroida	*Sphaerium nucleus*	SRR1561723	18,539,173	100
Mollusca	Bivalvia	Nuculoida	*Ennucula tenuis*	SRR331123	14,420,942	100
Mollusca	Bivalvia	Ostreoida	*Dimya lima*	SRR3350463	5,426,850	150
Rotifera	Monogononta	Ploima	*Brachionus plicatilis*	SRR3404576	7,403,847	150
Arthropoda	Branchiopoda	Diplostraca	*Eoleptestheria* cf *ticinensis*	SRR5140141	5,471,351	150
Arthropoda	Remipedia	Nectiopoda	*Godzilliognomus frondosus*	SRR8280777	14,086,834	75
Arthropoda	Arachnida	Solifugae	*Galeodes* sp.	SRR8745910	6,356,774	75
Arthropoda	Hexanauplia	Calanoida	*Neocalanus flemingeri*	SRR5873556	4,112,626	150
Arthropoda	Hexanauplia	Calanoida	*Calanus finmarchicus*	SRR4113507	10,633,606	150
Arthropoda	Hexanauplia	Pedunculata	*Octolasmis warwickii*	SRR10527303	15,813,391	150
Echinodermata	Holothuroidea	Aspidochirotida	*Apostichopus japonicus*	SRR8393254	8,289,770	150
Echinodermata	Crinoidea	Comatulida	*Florometra*	SRR3097584	32,710,859	100
Echinodermata	Echinoidea	Echinoida	*Paracentrotus lividus*	ERR1000783	6,803,316	75
Echinodermata	Echinoidea	Echinoida	*Paracentrotus lividus*	SRR10744002	13,583,857	75
Xenacoelomorpha	—	Acoela	*Childia submaculatum*	SRR3105702	6,089,955	100
Chaetognatha	Sagittoidea	Aphragmophora	*Krohnitta subtilis*	SRR7754744	15,954,007	100
Brachiopoda	Rhynchonellata	Rhynchonellida	*Hemithiris psittacea*	SRR1611556	9,221,875	100
Nemertea	Enopla	Bdellonemertea	*Malacobdella grossa*	SRR1611560	8,307,739	100
Nemertea	Palaeonemertea	—	*Cephalothrix linearis*	SRR1273789	4,869,244	75
Phoronida	—	—	*Phoronis psammophila*	SRR1611565	12,949,999	100
Platyhelminthes	Catenulida	—	*Catenula lemnae*	SRR1796434	3,028,636	100

TABLE 1 (Continued)

Phylum	Class	Order	Species	SRA	No. of reads	Length (bp)
Onychophora	Udeonychophora	Euonychophora	*Peripatopsis capensis*	SRR1145776	11,638,180	100
Onychophora	Udeonychophora	Euonychophora	*Peripatoides novaezealandiae*	SRR8745911	5,768,550	75
Gastrotricha	–	Macrodasyida	*Macrodasys* sp	SRR1271706	3,204,609	75
Gastrotricha	–	Chaetonotida	*Lepidodermella squamata*	SRR1273732	4,370,938	75
Annelida	Polychaeta	Phyllodocida	*Nephtys caeca*	SRR1232685	1,576,665	75
Gnathostomulida	–	Bursovaginoidea	*Gnathostomula paradoxa*	SRR1271607	5,954,962	75

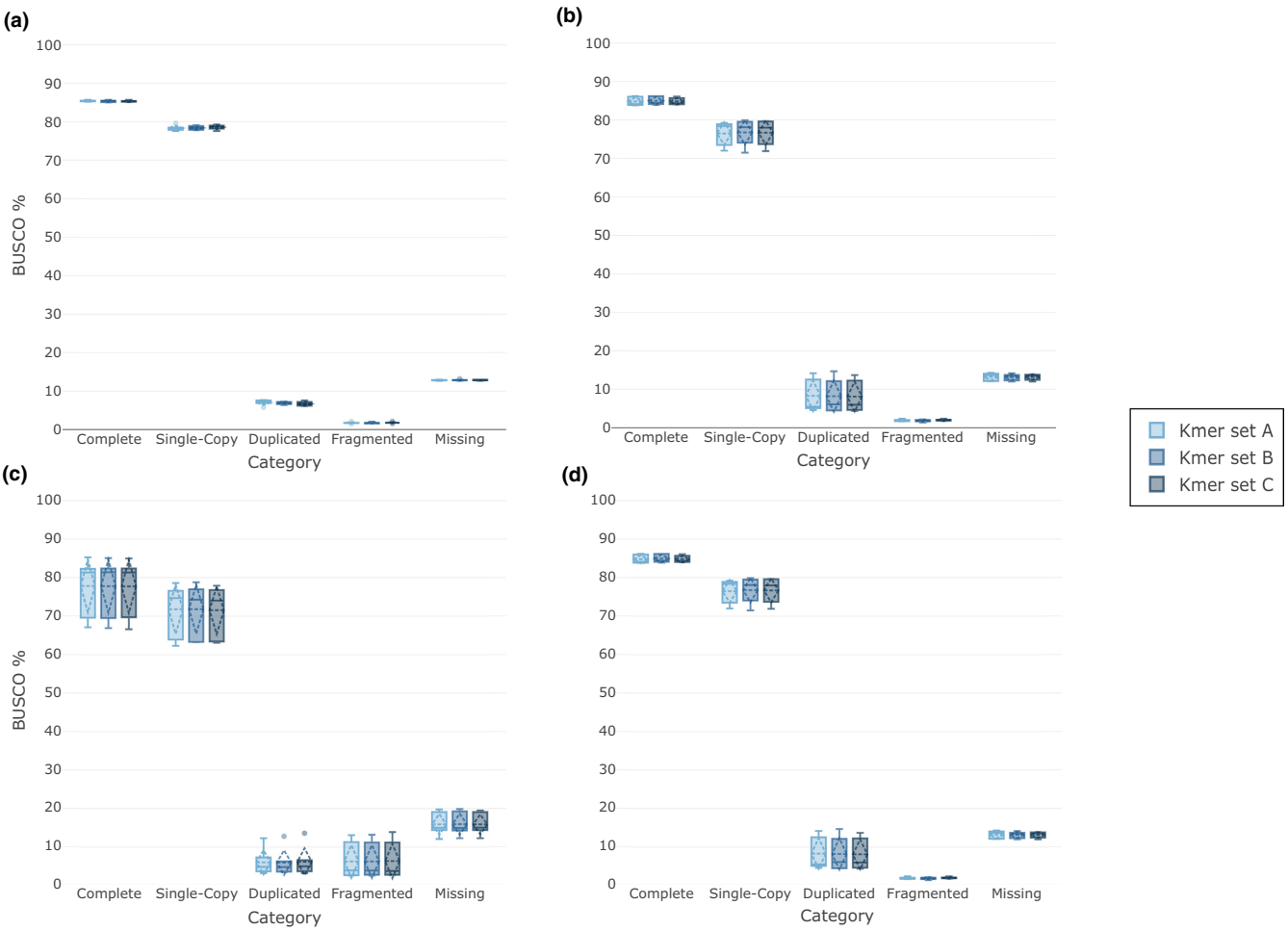

FIGURE 2 K-mer selection tests on the model organism *Caenorhabditis elegans*. Shown are the TransPi results for three different k-mer settings for read lengths of 50 bp (a), 75 bp (b), 100 bp (c), and 150 bp (d). For the k-mer test performed with *Mus musculus* and *Drosophila melanogaster* see Appendices S1–S4

3 | RESULTS

3.1 | K-mer selection, reads length effect and chimera detection

K-mer tests carried out on the model organisms used here (Table S1) showed that differences in BUSCO percentages between k-mer sets (i.e., A, B, C) were not significant (Table S2). However, slightly higher single-copy and lower duplication BUSCO percentages were observed with k-mer set C (Figure 2; Table S2; Appendices S1–S4). This pattern was observed in all three model organisms: *Caenorhabditis elegans* (worm), *Drosophila melanogaster* (fly) and *Mus musculus* (mouse) (Appendices S1 and S4). The read length test (i.e., 50, 75, 100 and 150 bp) also showed no significant difference in complete BUSCO percentages in favour of TransPi (Appendices S1 and S4). However, it should be noted that *D. melanogaster* paired-end reads of 50 bp

produced low complete BUSCO percentages for TransPi and TRINITY (complete BUSCO mean <45%). By contrast, *D. melanogaster* libraries with paired-end reads of 75, 100 and 150 bp length showed high BUSCO percentages for both, TransPi and TRINITY, where TRINITY surpasses TransPi by 1.0% (Appendices S1–S4). A similar pattern of a marginal difference between TransPi and TRINITY (with 1.0% higher complete BUSCO percentage in TRINITY) was also observed for the *C. elegans* and *M. musculus* data sets (Appendices S1–S4).

The major difference between TransPi and TRINITY in the model organisms was observed in the single-copy BUSCO category. This difference was more significant in the *D. melanogaster* and *M. musculus* data sets. For the *M. musculus* 150-bp reads, the difference between the single-copy BUSCO percentages was over 37% (Table S2). In terms of fragmented and missing BUSCO genes, TransPi scores were slightly higher (<1.0% of difference) than for TRINITY alone in most cases (Table S2; Appendices S1–S4). The increase of read length showed no clear effect on producing better BUSCO percentages on the majority of the model organism data sets (Appendices S1–S4). The same was observed for the increase of the read quantities in the data sets (Appendices S1–S4). Only for *D. melanogaster* 50-bp reads was an increase observed in complete BUSCO percentages when incrementing read quantity from 10 million to 26 million. The other model organism data sets did not show significant differences with respect to read quantities (Appendices S1–S4).

Results for the chimera detection test are presented in Table 2. A similar trend was observed in all model species (i.e., *C. elegans*, *D. melanogaster* and *M. musculus*). The number of non-chimeric transcripts (i.e., percentage of unique BLASTN matches) in TransPi (i.e., lowest: 3.07%; highest: 39.13%) was higher than in TRINITY alone (i.e., lowest: 3.66%; highest: 38.32%). Only in one sample (i.e., *M. musculus* SRR8329326) was the percentage of nonchimeric transcripts of TRINITY higher than TransPi. However, the TRINITY assembly had over 215,000 more transcripts than the TransPi transcriptome. Nevertheless, the percentage difference was only 0.59%. BUSCO scores followed the same pattern as explained above.

3.2 | TransPi on nonmodel organisms

A similar trend as seen in the model organisms was observed in the nonmodel organism data sets (Figure 3; Table S3). However, there were some key differences. First, results of complete BUSCO percentages were higher for TransPi in 41 of the 49 data sets tested in the study. The mean of complete BUSCO percentages was 79.57 ± 18.60% (median: 85%) for TransPi and 78.14 ± 19.30 (median: 84.2%) for the TRINITY assemblies. Of all data sets, 21 had complete BUSCO percentages higher than 90% with TransPi and 17 with TRINITY (Figure 4). Eleven and 13 data sets resulted in 80%–90% identified complete BUSCO genes with TransPi and TRINITY, respectively. However, Kruskal–Wallis tests showed no significant differences between TransPi and TRINITY (Table 3).

Second, there was a significant improvement of the percentage of identified complete single-copy BUSCO genes. Mean percentages with TransPi and TRINITY were 67.57 ± 16.75% (median: 72.9%) and 42.03 ± 15.37% (median: 40.8%), respectively. For the single-copy BUSCO genes, 16 data sets obtained scores higher than 80% with TransPi and none with TRINITY (Figure 4). For the range of 70%–80%, 11 data sets obtained scores in this range when using TransPi, whereas only one data set in this range was obtained when using TRINITY (Figure 4). Statistical testing (i.e., Kruskal–Wallis) demonstrated a significant difference for the single-copy BUSCO percentages between TransPi and TRINITY (p = 5.6e−10, Table 3). In the case of the nemertean worm *Malacobdella grossa* (accession SRR1611560), single-copy BUSCO genes had a substantial change from 20.4% for TRINITY to 83% with TransPi (Appendix S5). Other data sets with significant changes included the crinoid echinoderm *Florometra serratissima* (accession SRR3097584), where the scores for TRINITY and TransPi were 41.3% and 87.3%, respectively (Appendix S5).

Through the reduction step performed by EVIDENTIALGENE in TransPi, an expected substantial decrease of the duplication rate was observed. The means for duplicated BUSCO genes with TransPi and TRINITY were 12.0 ± 9.96% (median: 9.7%) and 36.11 ± 20.52% (median: 31.1%), respectively (Figures 3 and 4; Table S3). Kruskal–Wallis tests demonstrated a significant difference for the duplicated BUSCO percentages (p = 9.60e−11, Table 3). Even though differences in fragmented BUSCO percentages were not statistically significant, these values were lower for data sets when using TransPi. In the case of missing BUSCO percentages, TransPi scores are higher than TRINITY (Appendix S5), although the differences were not significant. These genes were removed during the reduction step of EVIDENTIALGENE (see Section 4). Note that a few data sets were encountered where neither TransPi nor TRINITY obtained complete BUSCO percentages higher than 50%. These data sets are: the polychaete annelid *Nephtys caeca* (accession SRR1232685), and the bivalve molluscs *Mercenaria campechiensis* (accession SRR1560359), *Sphaerium nucleus* (accession SRR1561723) and *Cardites antiquatus* (accession SRR1560458) (Table S3). However, the majority of the identified complete BUSCO genes in these sets were single-copy in the TransPi assemblies (Appendix S5). On the other hand, data sets such as the scleractinian coral *Porites pukoensis* (accession SRR8491966) were observed with complete BUSCO percentages of 99.4% with both TransPi and TRINITY (with high duplication rates in both).

As expected due to the reduction in transcripts, the total number of transcripts in TransPi was lower than with TRINITY (Appendix S6). The mean for TransPi transcripts was 93,351 ± 89,863 (median: 73,435) and for TRINITY transcripts 157,130 ± 142,410 (median: 109,261). The reduction of transcripts was also observed for the numbers of transcripts larger than 500 and 1,000 bp (Figure 5). However, in terms of the longest transcript, the mean for TransPi was 23,684 ± 15,374 bp (median: 22,147 bp) and 20,668 ± 11,248 bp (median: 18,708 bp) for TRINITY (Figure 5). Mapping of sequencing reads to the assembled transcripts showed lower mapping rates obtained with TransPi than with TRINITY (Figure 5; Appendices S7 and S8). The mean of the predicted genes by TransPi and TRINITY was 34,659 ± 43,987 (median: 25,280) and 52,106 ± 47,273 (median: 43,783), respectively (Figure 5; Appendix S6). This reduction in TransPi vs. TRINITY mirrors the reduction of duplicated BUSCO results.

TABLE 2 Chimera test for model species *C. elegans*, *D. melanogaster*, and *M. musculus*

Sample	BLASTN hits	No. of transcripts	% unique	BUSCO version 4—Metazoa DB ($n = 954$)
C. elegans				
TRINITY				
SRR10407355	9,538	28,219	33.80	C:76.9% [S:65.2%, D:11.7%], F:2.2%, M:20.9%
SRR10407357	9,014	23,526	38.32	C:76.1% [S:65.7%, D:10.4%], F:2.5%, M:21.4%
SRR10407359	9,310	27,734	33.57	C:76.3% [S:65.8%, D:10.5%], F:2.6%, M:21.1%
TransPi				
SRR10407355	8,567	23,803	35.99	C:75.7% [S:68.9%, D:6.8%], F:2.3%, M:22.0%
SRR10407357	8,494	21,709	39.13	C:75.5% [S:68.7%, D:6.8%], F:2.5%, M:22.0%
SRR10407359	8,675	24,891	34.85	C:75.5% [S:69.6%, D:5.9%], F:2.8%, M:21.7%
D. melanogaster				
TRINITY				
SRR7716077	4,585	36,267	12.64	C:97.2% [S:84.6%, D:12.6%], F:1.5%, M:1.3%
SRR7716078	4,133	31,793	13.00	C:93.6% [S:63.6%, D:30.0%], F:1.0%, M:5.4%
SRR7716080	4,364	31,928	13.67	C:90.1% [S:63.4%, D:26.7%], F:3.1%, M:6.8%
TransPi				
SRR7716077	4,603	29,014	15.86	C:93.9% [S:81.0%, D:12.9%], F:1.3%, M:4.8%
SRR7716078	4,211	25,225	16.69	C:93.8% [S:84.1%, D:9.7%], F:1.5%, M:4.7%
SRR8491966	4,383	25,817	16.98	C:91.4% [S:82.9%, D:8.5%], F:2.2%, M:6.4%
M. musculus				
TRINITY				
SRR10560364	12,360	244,523	5.05	C:98.2% [S:53.1%, D:45.1%], F:0.9%, M:0.9%
SRR10560365	12,807	261,000	4.91	C:98.2% [S:48.7%, D:49.5%], F:0.7%, M:1.1%
SRR8329326	29,885	816,077	3.66	C:97.4% [S:36.8%, D:60.6%], F:1.4%, M:1.2%
TransPi				
SRR10560364	8,901	165,890	5.37	C:97.2% [S:84.6%, D:12.6%], F:1.5%, M:1.3%
SRR10560365	10,551	198,502	5.32	C:96.5% [S:81.0%, D:15.5%], F:1.3%, M:2.2%
SRR8329326	18,418	600,340	3.07	C:94.4% [S:81.7%, D:12.7%], F:2.4%, M:3.2%

3.3 | TransPi report

The report generated by TransPi is interactive (i.e., an HTML file is generated) and can be viewed with standard web browsers (Appendix S9). The report allows the user to comprehensively assess the data by zooming in on the figures, compare data sets and see detailed information by selecting specific data points. The report summarizes all major steps performed by the pipeline, including quality filtering, assembly metrics, ORF numbers, annotation and KEGG pathway assignment using IPATH3 (Darzi et al., 2018). TransPi provides the user

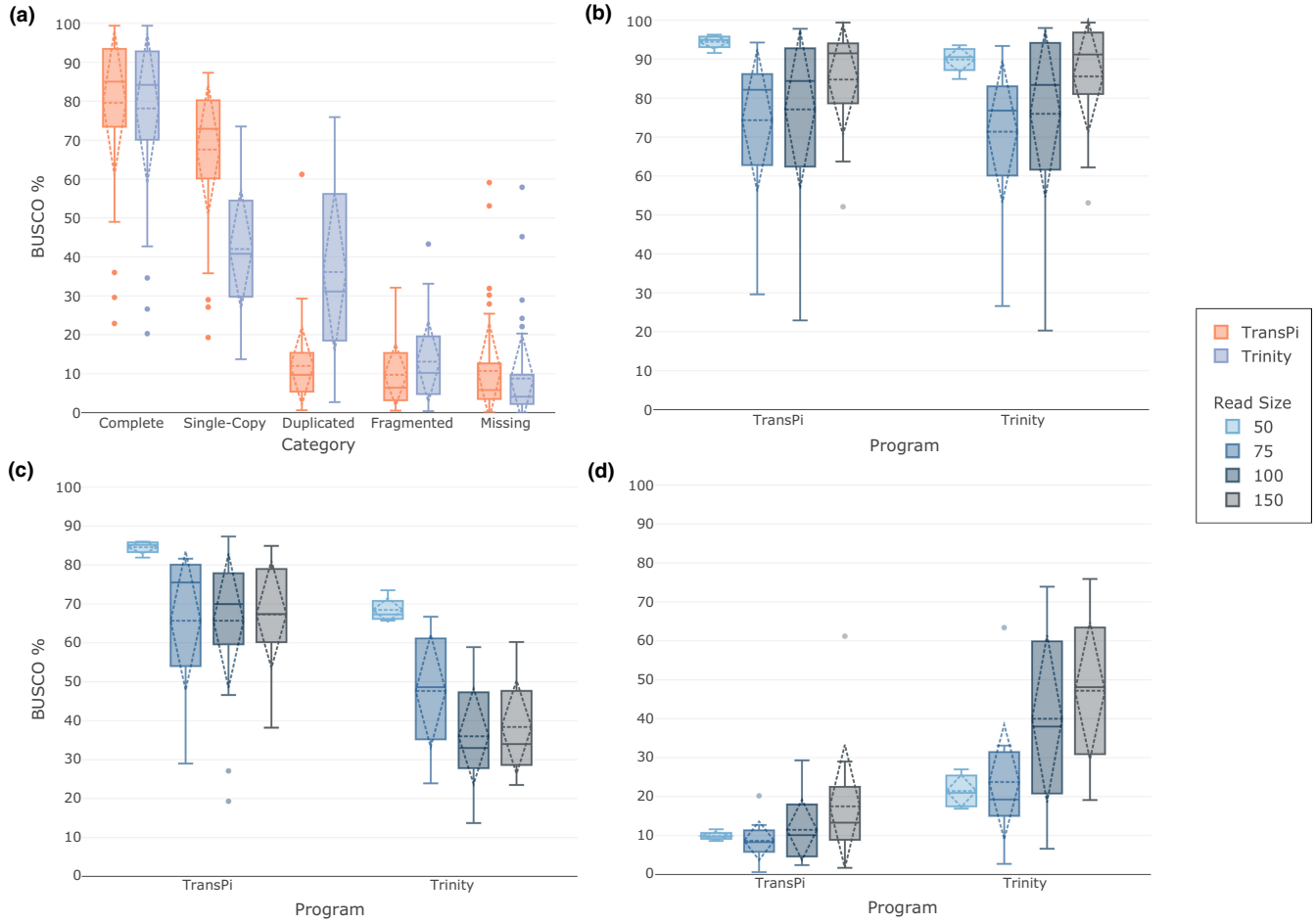

FIGURE 3 BUSCO results of nonmodel organisms (*n* 49). For a full list of analysed taxa see Table 1. (a) BUSCO percentages comparison for TransPi and TRINITY for all data sets. Comparisons of scores by read length for complete (b), single-copy (c) and duplicated (d) BUSCO genes. Significant differences (Kruskal–Wallis test $p < .05$) were obtained for (b) and (c)

with multiple files for further downstream analyses of the final reference transcriptome. For example, a file with all Gene Ontologies is created and can be directly used as input for TOPGO to perform enrichment analysis (Alexa & Rahnenfuhrer, 2016). All final and key intermediate files, including all plots, are stored in the user-selected output directory for manual inspection. Additionally, TransPi will save the execution report generated by Nextflow, in which the user can inspect how their system resources are being used in each process (example in Appendix S10).

3.4 | Additional TransPi options

The data set of *Porites pukoensis* (SRR8491966) produced a transcriptome with 567,526 sequences. Despite having a high BUSCO completeness (i.e., 99.4%), the majority of these were duplicates (i.e., 61.2%) (Appendix S5). Using the filtration step of TransPi, the number of transcripts was reduced by over 39% (from 567,526 to 343,832). The removed 223,694 transcripts had similarities with the *S. microadriaticum* sequences used for filtering (See Section 2.6). In the case of the "buscoDist" option, the *Octolasmis warwickii* data set (SRR10527303) was used and 30 genes that were missing from the TransPi assembly but were present in the other assemblies were found (Figure 6).

4 | DISCUSSION

De novo transcriptome assemblies are used in several applications such as: differential gene expression (Pita et al., 2018), gene model prediction (Chan et al., 2017), DNA target enrichment bait design (Quek et al., 2020), genome annotation (Holt & Yandell, 2011; Testa et al., 2015), detection of whole-genome duplication (Yang et al., 2019) and phylogenetics (Cheon et al., 2020; Lozano-Fernandez et al., 2019). Even though multiple softwares are currently available for transcriptome assembly, no single tool is able to generate optimal assemblies given various data sets (Hölzer & Marz, 2019). Thus, combining multiple assemblies, generated with various k-mers and software, represents a valuable approach to increase the quality of reference assemblies (Lu et al., 2013). Given the complexity of such analyses, automated workflows are desirable, including the need for standardization, reproducibility and scalability.

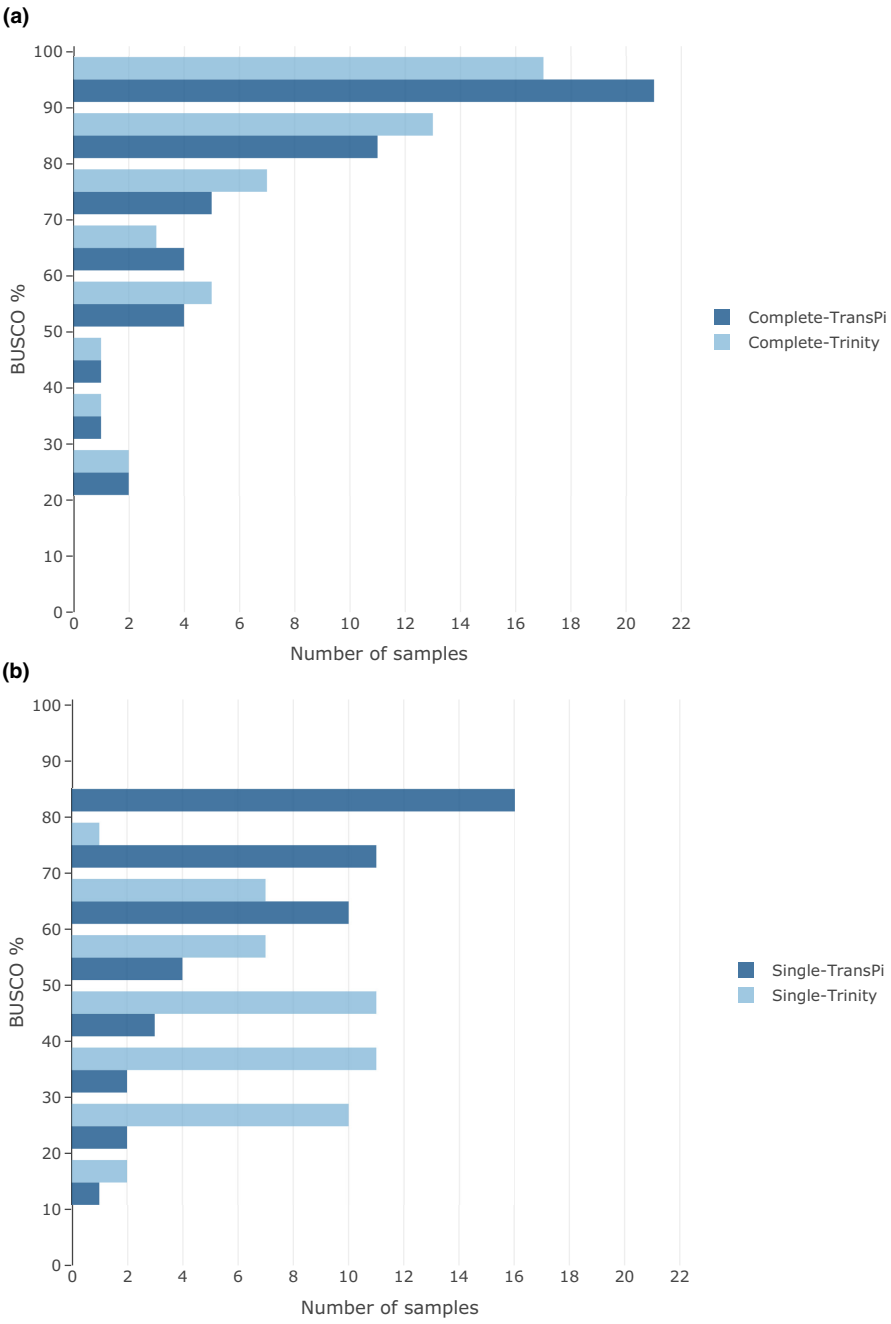

FIGURE 4 Histogram of number of data sets and BUSCO percentages in 10% bins. Comparison of identified complete (including duplicates) (a) and single-copy (b) BUSCO genes between TransPi and TRINITY

TABLE 3 Statistical tests on nonmodel organisms

	($p > .05$)					
Shapiro-Wilk	**TransPi**	**TRINITY**	**Normally distributed**	**ANOVA**	**Kruskal–Wallis ($p < .05$)**	**Significant**
Complete	5.09E−06	1.76E−05	No	—	0.6492791	No
Single-copy	0.0001667	0.1081	No	—	5.67E−10	Yes
Duplicated	2.51E−07	0.008433	No	—	9.60E−11	Yes
Fragmented	0.0005825	0.002121	No	—	0.1375242	No
Missing	1.81E−08	3.61E−09	No	—	0.1413003	No

The selection of assembler and k-mer list is the first step before performing an assembly. For the assemblers we chose the programs that produced better overall scores (i.e., TRINITY, RNASPADES, TRANS-ABYSS and SOAPDENOVO-TRANS) when compared using different data sets (Hölzer & Marz, 2019). VELVET/OASES was also included in the list of assemblers since this assembler performed better than the others when assembling long transcripts, while at the same time producing high BUSCO scores (Hölzer & Marz, 2019). Since the selection of

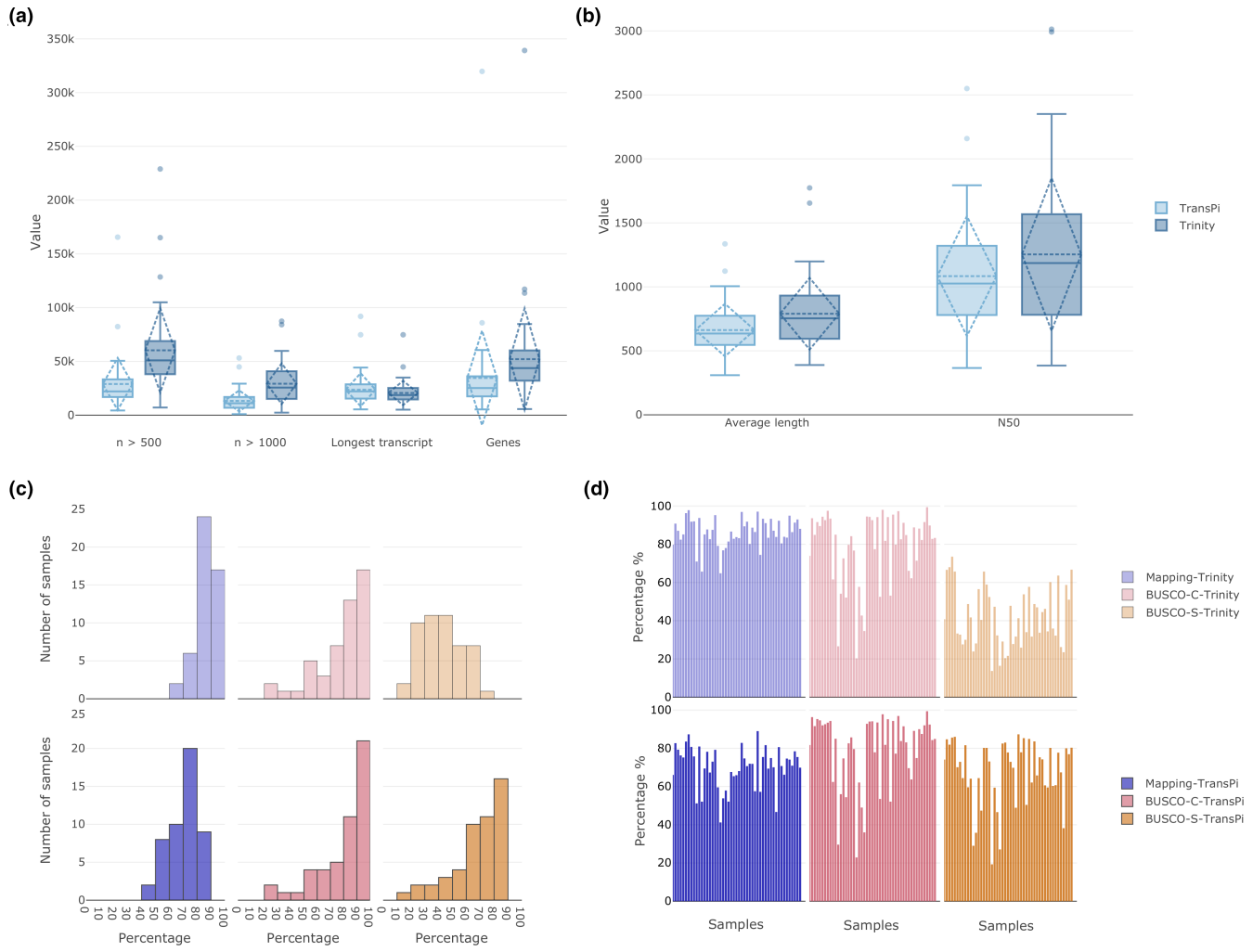

FIGURE 5 RNAQUAST results comparing TransPi and TRINITY. (a) Number of transcripts >500 bp, >1,000 bp, longest transcripts and number of predicted genes in the transcriptome. (b) Transcript average length and N50. (c) Histograms (10% bins) of all samples and percentage of mapping (reads to transcriptome), BUSCO-complete and BUSCO-single (light colour TRINITY, dark colour TransPi). (d) Percentage of mapping (reads to transcriptome), BUSCO-complete and BUSCO-single by individual samples (light colour TRINITY, dark colour TransPi)

the k-mers will modify the assembly graph creation, the effectiveness of TransPi was tested using multiple k-mer lists. These tests included different k-mer sizes, combinations of k-mers and different organisms (Table 1). Since TransPi relies on multiple assemblers and various k-mers, the effect on k-mer selection and their impact on the outcome of the pipeline is minimized. However, k-mer set C consistently resulted in moderately higher BUSCO percentages for single-copy genes and lower duplication levels, respectively. This k-mer set had a wider range of k-mer sizes (from small to long) than the other sets. Small k-mers tend to generate more transcripts but are more prone to misassemblies (Gibbons et al., 2009; Zerbino & Birney, 2008). By contrast, longer k-mers produce more contiguous assembly while decreasing transcript numbers (Robertson et al., 2010). Thus, by combining various k-mer sizes (i.e., short and long k-mers), a more comprehensive representation of the transcriptome can be achieved (Peng et al., 2013).

In previous studies, it has been shown that using more than 30 million read pairs does not significantly improve the quality of the transcriptome assembly (Francis et al., 2013; MacManes, 2018). However, in our tests mixed results were observed when comparing read quantities and BUSCO scores in each organism respectively (Appendix S5). As previously demonstrated, assembly quality and characteristics are data-dependent (Hölzer & Marz, 2019). Consequently, to provide a profound conclusion on the effects of read quantities in *de novo* transcriptome assemblies, a larger number of data sets from a broad range of taxa, in addition to biological replicates for each taxon, are needed. Also, organisms with sources of contamination (e.g., of symbiotic origin, prey, parasites or eukaryotic overgrowth in the target tissue) may need higher quantities of reads. In cases of sizeable data sets or where multiple libraries are combined, TransPi by default performs a read normalization step. This option can be skipped, although we recommend always performing the normalization step. A dramatic reduction in the computing time and resources was achieved when using the normalization step with many data sets in our laboratory (e.g., coral data sets). Although it has been shown that 30 is more than enough for

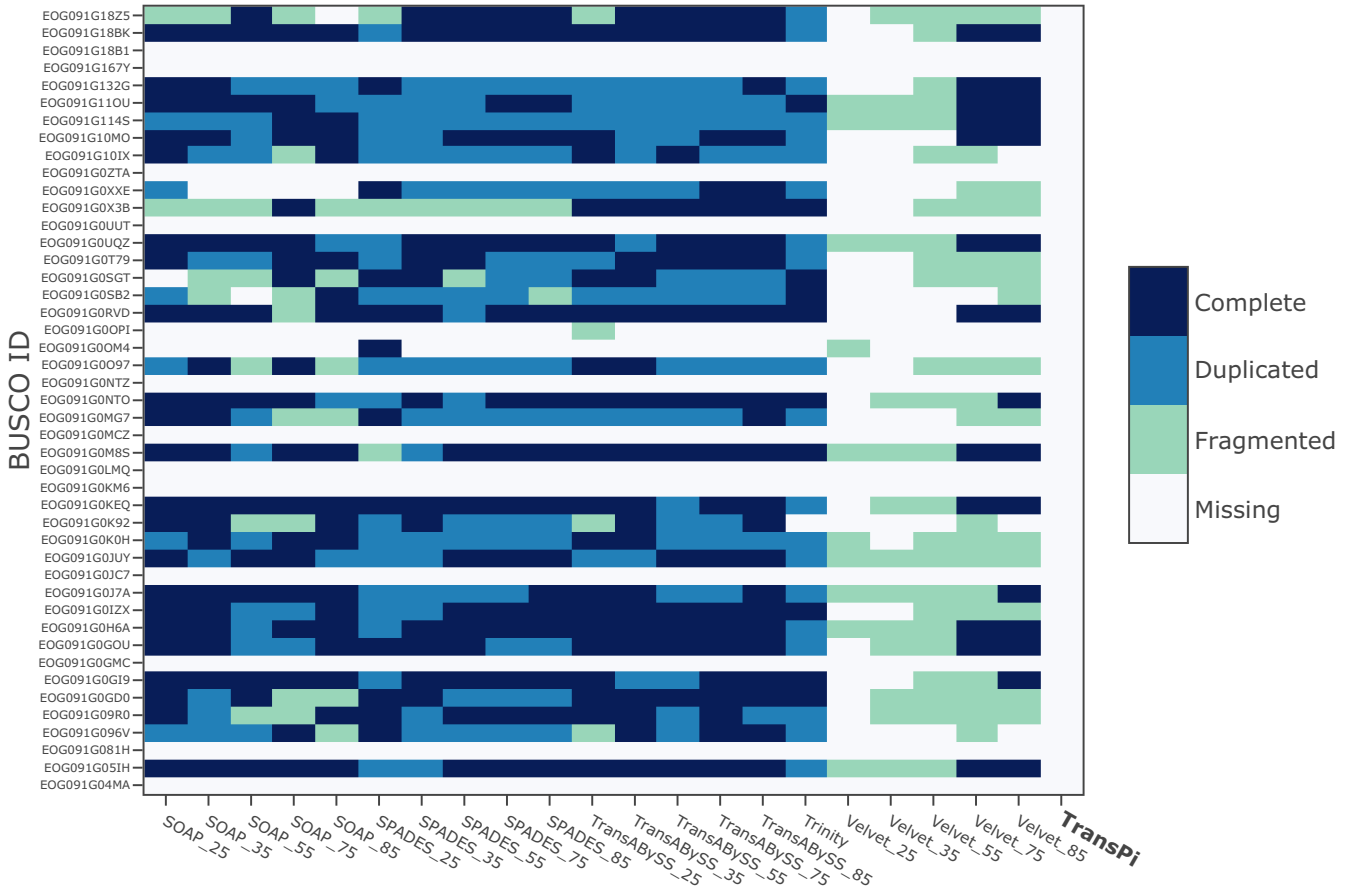

FIGURE 6 Heatmap of BUSCO gene presence in all assemblers with multiple k-mers that are found missing in TransPi for the data set of *Octolasmis warwickii* (SRR10527303)

the maximum coverage of reads during the normalization step (Haas et al., 2013), we have set the values for minimum and maximum read coverage to one and 100, respectively. However, since an assembly can vary widely depending on the organism and data sets we give the user the option to modify these values accordingly. The use of different read lengths did not yield significant differences between TransPi and TRINITY in all three model organisms (i.e., worm, fly and mouse) included in this study. However, the tests conducted on the three model organisms suggest strongly that use of longer reads (150 bp) should be preferred, because those generally yielded higher quality transcriptomes with respect to the BUSCO results.

The newly established TransPi pipeline performed well in model organisms (Tables S1 and S2, Appendices S1–S4). However, TRINITY performed slightly better than TransPi with respect to the "complete" and "fragmented" categories of the metazoa BUSCO genes set. The major advantage of TransPi in the model organisms, however, was the reduction of duplicated BUSCO genes (Figure 2). TransPi performed significantly better than the TRINITY assembler alone on nonmodel organisms (Figure 3). A high BUSCO completeness with a high number of single-copy BUSCO genes was obtained for the majority of the nonmodel data sets used here (Figures 3 and 4). In the case of the "fragmented" BUSCO genes category, TransPi produced lower scores than TRINITY due to the reduction step by EVIDENTIALGENE. Since the tool relies on sequence features such as CDS content and length (Gilbert, 2013, 2019), fragmented CDS will be less likely to pass the filtration step. The results for high number of single-copy BUSCO genes were statistically significant and are a major difference when comparing with the TransPi results of model organisms. The reduction of transcript duplication is obviously beneficial for studies where the presence of duplicates would bias the interpretation of the results. Another major disadvantage of keeping false isoforms is in phylogenomic analyses. Due to the relative ease of generation and affordability, many phylogenomic studies analyse multigene alignments based on transcriptome data instead of full genome data to estimate phylogenies (Cheon et al., 2020; Lozano-Fernandez et al., 2019). By using TransPi, the automation of large-scale phylogenomic approaches, focusing on thousands of proteins from many taxa, can be attained with ease in a scalable and reproducible way.

As expected, the final number of transcripts was consistently lower for TransPi given the reduction performed by EVIDENTIALGENE (Figure 5). In some cases (*Malacobdella grossa*) the reduction of transcripts was over 50% (Figure 5; Appendix S5). This explains why the mapping percentages for TransPi were also lower than for TRINITY. However, having reduced mapping rates (i.e., TransPi) did not affect the content of BUSCO genes in the transcriptomes. For example, in the *Malacobdella grossa* assembly, TransPi mapping was 65.41%

vs. 82.86% for TRINITY (Figure 5; Appendix S5), but the difference of complete BUSCO genes was only 0.30% (TransPi = 94.0%, TRINITY = 94.30%) and 62% (TransPi = 83.0%, TRINITY = 20.4%) for single-copy BUSCOs. Thus, the reduction in mapping percentages is due to the reduction of redundant transcripts (including allelic variants) rather than missing information from the assemblies. However, this reduction could potentially be an issue for differential gene expression studies where gene variants (i.e., isoforms) are removed from the samples via the reduction of transcripts. Nevertheless, performing a reduction step before a gene expression analysis is a common practice (DeLeo & Bracken-Grissom, 2020; Devens et al., 2020; Guo et al., 2017; Kashyap et al., 2020; Perez et al., 2021). Therefore, the nonredundant transcriptome generated by TransPi could be utilized for gene expression studies (Deshpande et al., 2021). However, further investigations are needed to support this. Currently, tests are being performed by our group to shed light on the effect of reduction algorithms (i.e., EVIDENTIALGENE, CD-HIT and others), in differential gene expression studies by employing data sets from various organisms.

The reduced mapping rates were observed throughout the nonmodel organisms analysed here (Figure 5; Appendix S5). In general, when the mapping percentages of TransPi were over 65%, satisfactory BUSCO content in the transcriptomes (i.e., high BUSCO presence and in single copies) were also observed. However, there were some cases where both TransPi and TRINITY produced equally low BUSCO scores, even though a relatively high mapping percentage was obtained (Figure 5; Appendix S7). This was the case for a *Catenula lemnae* data set (SRR1796434), where read mapping percentage was relatively high (74.69% and 89.37% for TransPi and TRINITY, respectively), while the BUSCO gene content (complete and single) was <53% (Figure 5; Appendix S7). In such cases, the assemblies may not be optimal and probably do not represent the complete transcriptome of the organism. (Figure 5; Appendix S7).

For the missing BUSCO category, TransPi produced assemblies with slightly higher values in comparison to the other assemblers. When a BUSCO gene is missing in TransPi (i.e., removed by the EVIDENTIALGENE step), in some cases these genes are found in the other individual assemblies (Figure 6). EVIDENTIALGENE aims to keep the most valid biological transcript, discards the probably not valid transcripts (based on specific measures), and decreases the redundancy of the multiple assemblers to obtain a nonredundant consensus transcriptome assembly (Gilbert, 2019). However, by doing so, some genes can be categorized as redundants, presumably because better candidates were selected. To gain more insight into cases like the one above, the TransPi option "buscoDist" was used with the *Octolasmis warwickii* data set. Comparing the missing genes of all generated assemblies and plotting the distribution of the BUSCO genes showed that TransPi had more missing genes that were categorized in other assemblers as being present (Figure 6). However, a considerable number of these genes were classified as duplicates by BUSCO. Since the BUSCO scores are indicators of the transcriptome completeness, correcting them will provide a more realistic estimation on the transcriptome quality of a given taxon. This TransPi option offers the user insight into the BUSCO gene content and transcript reduction by EVIDENTIALGENE to help better assess the quality of the assemblies.

In certain cases, significant numbers of BUSCO genes were not retrieved by TransPi, TRINITY or any of the assemblers. Although this could be related to assembler performance, other factors have been shown to alter transcriptome quality (RNA degradation, library preparation, sequencing depth, etc.) (Romero et al., 2014; Sultan et al., 2014). In the nonmodel organisms, four of the data sets yielded BUSCO complete percentages <50% in TransPi and TRINITY (Appendix S5). Three of these data sets (i.e., *Mercenaria campechiensis* [SRR1560359], *Sphaerium nucleus* [SRR1561723], *Carditesantiquatus* [SRR1560458]) stem from the same project and the same taxonomic group, the molluscan class Bivalvia. Extraction of nucleic acids in molluscs is known to be hampered by the presence of mucopolysaccharides and polyphenolic proteins, which can inhibit PCR and lead to biases in RNA preservation and/or the extraction quantities and/or qualities (Gayral et al., 2011; Knutson et al., 2020; Rzepecki et al., 1991). Moreover, since they are filter-feeders, possible high rates of contamination could arise depending on the tissue extracted and procedure. Nevertheless, BUSCO genes that were retrieved exhibited low rates of duplication, highlighting how the incorporation of EVIDENTIALGENE into TransPi can also decrease redundancy in cases of low transcriptome completeness. For the fourth data set (*Nephtys caeca* [SRR1232685], a polychaete annelid), only a small number (1.5 million) of read pairs were deposited in INSDC databases (i.e., NCBI's GenBank), which helps to explain the poor results (Appendix S5). Deeper sequencing of these particular specimens may well lead to an improved transcriptome. This also might indicate that the quantity of reads rather than the quality of input material was the limiting factor for the generation of a complete transcriptome.

BUSCO gene presence is one of the main metrics, together with mapping and number of transcripts, to assess transcriptome completeness and the quality of nonmodel organisms. Thus, the analyses for evaluating TransPi's performance were mostly based on this metric. However, since model organisms have established gene models, a procedure similar to the Bellerophon pipeline (Kerkvliet et al., 2019) was used as an additional metric for evaluation of TransPi. Overall, TransPi had a higher number of nonchimeric transcripts when compared to TRINITY alone (Table 2). These suggest that the filtration step performed by EVIDENTIALGENE is able to reduce the number of chimeric transcripts (i.e., erroneous assembled sequences) while maintaining the information of the different transcripts. These analyses were done in model organisms only and TransPi does not apply this procedure as part of its execution. The implementation of a procedure like the Bellerophon pipeline (Kerkvliet et al., 2019) in our tool is hindered by multiple factors. First, a reference transcriptome or gene set is needed to calculate the chimeras. TransPi is intended to be used mainly on nonmodel species where the majority do not have a reference transcriptome. Second, the Bellerophon pipeline makes use of a software (i.e., TRANSRATE) which has not been recently updated. This creates reproducibility problems since the tool relies on old versions for some dependencies. Also, the tool does not offer a conda

installation or container images. However, note that one of the critical steps in the Bellerophon pipeline is the use of CD-HIT-EST for decreasing redundancy in the assemblies. This step is already incorporated in the EVIDENTIALGENE software for the same purpose.

TransPi also addresses putative contamination issues that might affect a transcriptome by providing an additional option that performs filtering of "contaminants." Data sets from organisms such as corals can represent a challenge during transcriptome assembly and downstream analyses due to their endosymbiotic zooxanthellae (Shinzato et al., 2014). Thus, a filtration step is usually performed to remove sequences that do not belong to the target (host) transcriptome (Veglia et al., 2018). The filtration step of TransPi is a useful step in cases of known contamination sources. For example, in the data set of the coral *Porites pukoensis*, both TransPi and TRINITY obtained high BUSCO completeness percentages. However, despite the reduction with EVIDENTIALGENE, single-copy BUSCO percentages were low and the percentage of duplicated BUSCO genes was high in both TransPi and TRINITY. Given the strong efficiency of TransPi in removing redundancy, the presence of many duplicates in this data set may indicate the presence of algae (symbionts) transcripts and/or contamination. Also, it has been previously reported that other eukaryotes, particularly fungi, are commonly found in *Porites pukoensis* (Li et al., 2014). This could potentially bias the outcomes and can strongly affect downstream analyses. Thus, using a contaminant filtration step, as performed by TransPi, is beneficial to generate a cleaner and accurate transcriptome assembly and provide the user with a host-only transcriptome to be further analysed. Note that PSYTRANS can work with any set of proteins. Thus, it is not only for the separation of host and symbiont sequences and can be used with any data set chosen by the user.

In summary, TransPi offers researchers working with nonmodel organisms the opportunity of a comprehensive *de novo* transcriptome analysis, requiring minimum user input but without losing the ability of a thorough analysis. TransPi is not intended as a one-stop solution for transcriptome assemblies, but rather as a broad start for gaining insight into the transcriptome of their nonmodel organisms of interest. New users can obtain a vast amount of information for exploring their transcriptome, while more experienced users also have the ability to modify the various pipeline processing options (if necessary). All files generated by TransPi (individual assemblies, nonredundant assembly, BUSCO files, BAM files, etc.) are stored and are available to the user for further exploration, use in other tools (e.g., Corset - Davidson & Oshlack, 2014) or to keep for future reference. The interactive report created by TransPi is key for data exploration and to help users decide if further processing is needed before using the generated nonredundant assembly directly in several downstream analyses. These analyses include but are not limited to gene modelling for genome annotations, bait design and phylogenetic studies. Another key advantage of using TransPi is that it offers reproducibility of the results with ease, where entire experiments can be repeated with defined versions of all programs included in the workflow. It also provides a user-friendly environment, easy deployment, and scalability by employing Nextflow. TransPi also has other additional features to help gain extra insight into the assemblies. Thus, we anticipate that TransPi will be a valuable tool for the generation of comprehensive *de novo* nonredundant transcriptome assemblies for nonmodel organisms.

ACKNOWLEDGEMENTS

Version 3 of this paper has been peer-reviewed and recommended by Peer Community In Genomics (https://doi.org/10.24072/pci.genomics.100009). R.E.R.V., M.E. and G.W. acknowledge funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 764840 (ITN IGNITE). C.G.E. acknowledges the Advanced Human Capital Program of the National Commission for Scientific and Technological Research (CONICYT) for the Becas-Chile Scholarship awarded to study at LMU. C.G.E. and M.E. acknowledges funding by Lehre@LMU (project no.: W19_F1; Studi_forscht@GEO). G.W. acknowledges funding through the LMU Munich's Institutional Strategy LMUexcellent within the framework of the German Excellence Initiative. The authors gratefully acknowledge the Leibniz Supercomputing Centre (LRZ) as a partner of ITN IGNITE for providing computing time and support on its Linux-Cluster and Compute Cloud system. Open access funding enabled and organized by ProjektDEAL.

CONFLICT OF INTEREST

The authors declare that they have no financial conflict of interest with the content of this article.

AUTHOR CONTRIBUTIONS

R.E.R.V. designed the pipeline and scripts, analysed data, and wrote the initial draft of the manuscript. C.A.G.E. analysed data and prepared tables. N.C. wrote pipeline processes and helped draft the manuscript. M.E. analysed data and helped draft the manuscript. G.W. acquired the funding, supervised the project, provided the infrastructure for data analysis, and helped draft the manuscript. All authors approved the final version of the manuscript.

DATA AVAILABILITY STATEMENT

Source code and scripts are available online at https://github.com/PalMuc/TransPi. Supplementary scripts and files are available at https://doi.org/10.5281/zenodo.5060054.

ORCID

Ramón E. Rivera-Vicéns https://orcid.org/0000-0002-6229-3537
Catalina A. Garcia-Escudero https://orcid.org/0000-0001-9704-7865
Nicola Conci https://orcid.org/0000-0001-5549-3197
Michael Eitel https://orcid.org/0000-0002-0531-0732
Gert Wörheide https://orcid.org/0000-0002-6380-7421

REFERENCES

Alexa, A., & Rahnenfuhrer, J. (2016). *topGO: Enrichment Analysis for Gene Ontology*. R package version 2.32.0.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–3402. https://doi.org/10.1093/nar/25.17.3389

Andrews, S. (2010). *FastQC: A quality control tool for high throughput sequence data*.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., & Pilbout, S. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research*, *31*(1), 365–370. https://doi.org/10.1093/nar/gkg095

Bryant, D. M., Johnson, K., DiTommaso, T., Tickle, T., Couger, M. B., Payzin-Dogru, D., Lee, T. J., Leigh, N. D., Kuo, T.-H., Davis, F. G., Bateman, J., Bryant, S., Guzikowski, A. R., Tsai, S. L., Coyne, S., Ye, W. W., Freeman, R. M., Peshkin, L., Tabin, C. J., … Whited, J. L. (2017). A tissue-mapped axolotl de novo transcriptome enables identification of limb regeneration factors. *Cell Reports*, *18*(3), 762–776. https://doi.org/10.1016/j.celrep.2016.12.063

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, *12*(1), 59–60. https://doi.org/10.1038/nmeth.3176

Bushmanova, E., Antipov, D., Lapidus, A., & Prjibelski, A. D. (2019). rnaSPAdes: A *de novo* transcriptome assembler and its application to RNA-Seq data. *GigaScience*, *8*(9), https://doi.org/10.1093/gigascience/giz100

Bushmanova, E., Antipov, D., Lapidus, A., Suvorov, V., & Prjibelski, A. D. (2016). rnaQUAST: A quality assessment tool for *de novo* transcriptome assemblies. *Bioinformatics*, *32*(14), 2210–2212. https://doi.org/10.1093/bioinformatics/btw218

Cerveau, N., & Jackson, D. J. (2016). Combining independent *de novo* assemblies optimizes the coding transcriptome for nonconventional model eukaryotic organisms. *BMC Bioinformatics*, *17*(1), 1–13. https://doi.org/10.1186/s12859-016-1406-x

Chan, K.-L., Rosli, R., Tatarinova, T. V., Hogan, M., Firdaus-Raih, M., & Low, E.-T.-L. (2017). Seqping: Gene prediction pipeline for plant genomes using self-training gene models and transcriptomic data. *BMC Bioinformatics*, *18*(S1), 1–7. https://doi.org/10.1186/s12859-016-1426-6

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*(17), i884–i890. https://doi.org/10.1093/bioinformatics/bty560

Cheon, S., Zhang, J., & Park, C. (2020). Is phylotranscriptomics as reliable as phylogenomics? *Molecular Biology and Evolution*, *37*(12), 3672–3683. https://doi.org/10.1093/molbev/msaa181

Cornwell, M. I., Vangala, M., Taing, L., Herbert, Z., Köster, J., Li, B., Sun, H., Li, T., Zhang, J., Qiu, X., Pun, M., Jeselsohn, R., Brown, M., Liu, X. S., & Long, H. W. (2018). VIPER: Visualization pipeline for RNA-seq, a Snakemake workflow for efficient and complete RNA-seq analysis. *BMC Bioinformatics*, *19*(1). 1–14. https://doi.org/10.1186/s12859-018-2139-9

D'Antonio, M., D'Onorio De Meo, P., Pallocca, M., Picardi, E., D'Erchia, A. M., Calogero, R. A., Castrignanò, T., & Pesole, G. (2015). RAP: RNA-Seq analysis pipeline, a new cloud-based NGS web application. *BMC Genomics*, *16*(Suppl 6), S3. https://doi.org/10.1186/1471-2164-16-s6-s3

Darzi, Y., Letunic, I., Bork, P., & Yamada, T. (2018). iPath3.0: Interactive pathways explorer v3. *Nucleic Acids Research*, *46*(W1), W510–W513. https://doi.org/10.1093/nar/gky299

Davidson, N. M., & Oshlack, A. (2014). Corset: Enabling differential gene expression analysis for *de novo* assembled transcriptomes. *Genome Biology*, *15*(7), 1–14. https://doi.org/10.1186/s13059-014-0410-6

DeLeo, D. M., & Bracken-Grissom, H. D. (2020). Illuminating the impact of diel vertical migration on visual gene expression in deep-sea shrimp. *Molecular Ecology*, *29*(18), 3494–3510. https://doi.org/10.1111/mec.15570

Deshpande, A., Rivera-Vicéns, R. E., Thakur, N. L., & Wörheide, G. (2021). Transcriptomic response of *Cinachyrella* cf. *cavernosa* sponges to spatial competition. *bioRxiv*, *451097*. https://doi.org/10.1101/2021.07.05.451097

Devens, H. R., Davidson, P. L., Deaker, D. J., Smith, K. E., Wray, G. A., & Byrne, M. (2020). Ocean acidification induces distinct transcriptomic responses across life history stages of the sea urchin *Heliocidaris erythrogramma*. *Molecular Ecology*, *29*(23), 4618–4636. https://doi.org/10.1111/mec.15664

Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nature Biotechnology*, *35*(4), 316–319. https://doi.org/10.1038/nbt.3820

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C. E., & Finn, R. D. (2018). The Pfam protein families database in 2019. *Nucleic Acids Research*, *47*(D1), D427–D432. https://doi.org/10.1093/nar/gky995

European Commission (2020). Horizon 2020 in brief: The EU framework programme for research & innovation.

Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Research*, *39*(suppl), W29–W37. https://doi.org/10.1093/nar/gkr367

Francis, W. R., Christianson, L. M., Kiko, R., Powers, M. L., Shaner, N. C., & D Haddock, S. H. (2013). A comparison across non-model animals suggests an optimal sequencing depth for *de novo* transcriptome assembly. *BMC Genomics*, *14*(1), 167. https://doi.org/10.1186/1471-2164-14-167

Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics*, *28*(23), 3150–3152. https://doi.org/10.1093/bioinformatics/bts565

Gayral, P., Weinert, L., Chiari, Y., Tsagkogeorga, G., Ballenghien, M., & Galtier, N. (2011). Next-generation sequencing of transcriptomes: A guide to RNA isolation in nonmodel animals. *Molecular Ecology Resources*, *11*(4), 650–661. https://doi.org/10.1111/j.1755-0998.2011.03010.x

Gibbons, J. G., Janson, E. M., Hittinger, C. T., Johnston, M., Abbot, P., & Rokas, A. (2009). Benchmarking next-generation transcriptome sequencing for functional and evolutionary genomics. *Molecular Biology and Evolution*, *26*(12), 2731–2744. https://doi.org/10.1093/molbev/msp188

Gilbert, D. (2013). *Gene-omes built from mRNA-seq not genome DNA*.

Gilbert, D. (2019). Longest protein, longest transcript or most expression, for accurate gene reconstruction of transcriptomes? *bioRxiv*, *829184*. https://doi.org/10.1101/829184

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., … Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, *29*(7), 644–652. https://doi.org/10.1038/nbt.1883

Guo, W., Wu, H., Zhang, Z., Yang, C., Hu, L., Shi, X., Jian, S., Shi, S., & Huang, Y. (2017). Comparative analysis of transcriptomes in rhizophoraceae provides insights into the origin and adaptive evolution of mangrove plants in intertidal environments. *Frontiers in Plant Science*, *8*, 795. https://doi.org/10.3389/fpls.2017.00795

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., Eccles, D., Li, B. O., Lieber, M., MacManes, M. D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman,

R., William, T., Dewey, C. N., … Regev, A. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, *8*(8), 1494–1512. https://doi.org/10.1038/nprot.2013.084

Holt, C., & Yandell, M. (2011). MAKER2: An annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*, *12*(1), 1–14. https://doi.org/10.1186/1471-2105-12-491

Hölzer, M., & Marz, M. (2019). De novo transcriptome assembly: A comprehensive cross-species comparison of short-read RNA-Seq assemblers. *Gigascience*, *8*(5), 1–16. https://doi.org/10.1093/gigascience/giz039

Kashyap, S. P., Prasanna, H. C., Kumari, N., Mishra, P., & Singh, B. (2020). Understanding salt tolerance mechanism using transcriptome profiling and *de novo* assembly of wild tomato *Solanum chilense*. *Scientific Reports*, *10*(1), 1–20. https://doi.org/10.1038/s41598-020-72474-w

Kerkvliet, J., de Fouchier, A., van Wijk, M., & Groot, A. T. (2019). The Bellerophon pipeline, improving *de novo* transcriptomes and removing chimeras. *Ecology and Evolution*, *9*(18), 10513–10521. https://doi.org/10.1002/ece3.5571

Knutson, V. L., Brenzinger, B., Schrödl, M., Wilson, N. G., & Giribet, G. (2020). Most Cephalaspidea have a shell, but transcriptomes can provide them with a backbone (Gastropoda: Heterobranchia). *Molecular Phylogenetics and Evolution*, *153*, 106943. https://doi.org/10.1016/j.ympev.2020.106943

Kohen, R., Barlev, J., Hornung, G., Stelzer, G., Feldmesser, E., Kogan, K., Safran, M., & Leshkowitz, D. (2019). UTAP: User-friendly transcriptome analysis pipeline. *BMC Bioinformatics*, *20*(1), 1–7. https://doi.org/10.1186/s12859-019-2728-2

Kopylova, E., Noé, L., & Touzet, H. (2012). SortMeRNA: Fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics*, *28*(24), 3211–3217. https://doi.org/10.1093/bioinformatics/bts611

Krogh, A., Larsson, B., von Heijne, G., & Sonnhammer, E. L. L. (2001). Predicting transmembrane protein topology with a hidden markov model: Application to complete genomes. *Journal of Molecular Biology*, *305*(3), 567–580. https://doi.org/10.1006/jmbi.2000.4315

Lagesen, K., Hallin, P., Rødland, E. A., Stærfeldt, H.-H., Rognes, T., & Ussery, D. W. (2007). RNAmmer: Consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research*, *35*(9), 3100–3108. https://doi.org/10.1093/nar/gkm160

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

Li, J., Zhong, M., Lei, X., Xiao, S., & Li, Z. (2014). Diversity and antibacterial activities of culturable fungi associated with coral Porites pukoensis. *World Journal of Microbiology and Biotechnology*, *30*(10), 2551–2558. https://doi.org/10.1007/s11274-014-1701-5

Lozano-Fernandez, J., Tanner, A. R., Giacomelli, M., Carton, R., Vinther, J., Edgecombe, G. D., & Pisani, D. (2019). Increasing species sampling in chelicerate genomic-scale datasets provides support for monophyly of Acari and Arachnida. *Nature Communications*, *10*(1), 1–8. https://doi.org/10.1038/s41467-019-10244-7

Lu, B. X., Zeng, Z. B., & Shi, T. L. (2013). Comparative study of *de novo* assembly and genome-guided assembly strategies for transcriptome reconstruction based on RNA-Seq. *Science China Life Sciences*, *56*(2), 143–155. https://doi.org/10.1007/s11427-013-4442-z

MacManes, M. D. (2014). On the optimal trimming of high-throughput mRNA sequence data. *Frontiers in Genetics*, *5*, 13. https://doi.org/10.3389/fgene.2014.00013

MacManes, M. D. (2018). The Oyster River Protocol: A multi-assembler and k-mer approach for *de novo* transcriptome assembly. *PeerJ*, *6*, e5428. https://doi.org/10.7717/peerj.5428

Martin, J., Bruno, V. M., Fang, Z., Meng, X., Blow, M., Zhang, T., Sherlock, G., Snyder, M., & Wang, Z. (2010). Rnnotator: An automated *de novo* transcriptome assembly pipeline from stranded RNA-Seq reads. *BMC Genomics*, *11*(1), 663. https://doi.org/10.1186/1471-2164-11-663

Peng, Y., Leung, H. C., Yiu, S. M., Lv, M. J., Zhu, X. G., & Chin, F. Y. (2013). IDBA-tran: A more robust *de novo* de Bruijn graph assembler for transcriptomes with uneven expression levels. *Bioinformatics*, *29*(13), i326–i334. https://doi.org/10.1093/bioinformatics/btt219

Perez, R., de Souza Araujo, N., Defrance, M., & Aron, S. (2021). Molecular adaptations to heat stress in the thermophilic ant genus *Cataglyphis*. *Molecular Ecology*, *30*(21), 5503–5516. https://doi.org/10.1111/mec.16134

Petersen, T. N., Brunak, S., von Heijne, G., & Nielsen, H. (2011). SignalP 4.0: Discriminating signal peptides from transmembrane regions. *Nature Methods*, *8*(10), 785–786. https://doi.org/10.1038/nmeth.1701

Pita, L., Hoeppner, M. P., Ribes, M., & Hentschel, U. (2018). Differential expression of immune receptors in two marine sponges upon exposure to microbial-associated molecular patterns. *Scientific Reports*, *8*(1), 1–15. https://doi.org/10.1038/s41598-018-34330-w

Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A., & Korobeynikov, A. (2020). Using SPAdes de novo assembler. *Current Protocols in Bioinformatics*, *70*(1), e102. https://doi.org/10.1002/cpbi.102

Quek, R. Z. B., Jain, S. S., Neo, M. L., Rouse, G. W., & Huang, D. (2020). Transcriptome-based target-enrichment baits for stony corals (Cnidaria: Anthozoa: Scleractinia). *Molecular Ecology Resources*, *20*(3), 807–818. https://doi.org/10.1111/1755-0998.13150

Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S. D., Mungall, K., Lee, S., Okada, H. M., Qian, J. Q., Griffith, M., Raymond, A., Thiessen, N., Cezard, T., Butterfield, Y. S., Newsome, R., Chan, S. K., She, R., Varhol, R., … Birol, I. (2010). De novo assembly and analysis of RNA-seq data. *Nature Methods*, *7*(11), 909–912. https://doi.org/10.1038/nmeth.1517

Romero, I. G., Pai, A. A., Tung, J., & Gilad, Y. (2014). RNA-seq: Impact of RNA degradation on transcript quantification. *BMC Biology*, *12*(1), 1–13. https://doi.org/10.1186/1741-7007-12-42

Rzepecki, L. M., Chin, S. S., Waite, J. H., & Lavin, M. F. (1991). Molecular diversity of marine glues: Polyphenolic proteins from five mussel species. *Molecular Marine Biology and Biotechnology*, *1*(1), 78–88.

Schulz, M. H., Zerbino, D. R., Vingron, M., & Birney, E. (2012). Oases: Robust *de novo* RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*, *28*(8), 1086–1092. https://doi.org/10.1093/bioinformatics/bts094

Shinzato, C., Inoue, M., & Kusakabe, M. (2014). A Snapshot of a Coral "Holobiont": A transcriptome assembly of the Scleractinian coral, Porites, captures a wide variety of genes from both the host and symbiotic zooxanthellae. *PLoS One*, *9*(1), e85182. https://doi.org/10.1371/journal.pone.0085182

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Smith-Unna, R., Boursnell, C., Patro, R., Hibberd, J. M., & Kelly, S. (2016). TransRate: Reference-free quality assessment of *de novo* transcriptome assemblies. *Genome Research*, *26*(8), 1134–1144. https://doi.org/10.1101/gr.196469.115

Sultan, M., Amstislavskiy, V., Risch, T., Schuette, M., Dökel, S., Ralser, M., Balzereit, D., Lehrach, H., & Yaspo, M.-L. (2014). Influence of RNA extraction methods and library selection schemes on RNA-seq data. *BMC Genomics*, *15*(1), 675. https://doi.org/10.1186/1471-2164-15-675

Testa, A. C., Hane, J. K., Ellwood, S. R., & Oliver, R. P. (2015). CodingQuarry: Highly accurate hidden Markov model gene prediction in fungal genomes using RNA-seq transcripts. *BMC Genomics*, *16*(1), 170. https://doi.org/10.1186/s12864-015-1344-4

Veglia, A. J., Hammerman, N. M., Rivera-Vicéns, R. E., & Schizas, N. V. (2018). De novo transcriptome assembly of the coral *Agaricia lamarcki* (Lamarck's sheet coral) from mesophotic depth in southwest Puerto Rico. *Marine Genomics*, *41*, 6–11. https://doi.org/10.1016/j.margen.2018.08.003

Wang, D. (2018). hppRNA—a Snakemake-based handy parameter-free pipeline for RNA-Seq analysis of numerous samples. *Briefings in Bioinformatics*, *19*(4), 622–626. https://doi.org/10.1093/bib/bbw143

Waterhouse, R. M., Zdobnov, E. M., & Kriventseva, E. V. (2011). Correlating traits of gene retention, sequence divergence, duplicability and essentiality in vertebrates, arthropods, and fungi. *Genome Biology and Evolution*, *3*, 75–86. https://doi.org/10.1093/gbe/evq083

Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., Huang, W., He, G., Gu, S., Li, S., Zhou, X., Lam, T.-W., Li, Y., Xu, X., Wong, G.-K.-S., & Wang, J. (2014). SOAPdenovo-Trans: *De novo* transcriptome assembly with short RNA-Seq reads. *Bioinformatics*, *30*(12), 1660–1666. https://doi.org/10.1093/bioinformatics/btu077

Yang, Y., Li, Y., Chen, Q., Sun, Y., & Lu, Z. (2019). WGDdetector: A pipeline for detecting whole genome duplication events using the genome or transcriptome annotations. *BMC Bioinformatics*, *20*(1), 1–6. https://doi.org/10.1186/s12859-019-2670-3

Zerbino, D. R., & Birney, E. (2008). Velvet: Algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Research*, *18*(5), 821–829. https://doi.org/10.1101/gr.074492.107

Zhang, X., & Jonassen, I. (2020). RASflow: An RNA-Seq analysis workflow with Snakemake. *BMC Bioinformatics*, *21*(1), 1–9. https://doi.org/10.1186/s12859-020-3433-x

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Rivera-Vicéns, R. E., Garcia-Escudero, C. A., Conci, N., Eitel, M., & Wörheide, G. (2022). TransPi—a comprehensive TRanscriptome ANalysiS PIpeline for *de novo* transcriptome assembly. *Molecular Ecology Resources*, 22, 2070–2086. https://doi.org/10.1111/1755-0998.13593