Bioinformatics, 35(13), 2019, 2199–2207
doi: 10.1093/bioinformatics/bty903
Advance Access Publication Date: 19 November 2018
Original Paper

OXFORD

Sequence analysis

CAARS: comparative assembly and annotation of RNA-Seq data

Carine Rey[1,*], Philippe Veber[2], Bastien Boussau[2,†] and Marie Sémon[1,†]

[1]UnivLyon, Université Claude Bernard Lyon 1, ENS de Lyon, CNRS UMR 5239, INSERM U1210, LBMC, F-69007, Lyon, France and [2]UnivLyon, Université Claude Bernard Lyon 1, CNRS, UMR 5588, LBBE, F-69100, Villeurbanne, France

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the last two authors should be regarded as Joint Last Authors.
Associate Editor: Janet Kelso

Received on June 28, 2017; revised on September 13, 2018; editorial decision on October 10, 2018; accepted on November 16, 2018

Abstract

Motivation: RNA sequencing (RNA-Seq) is a widely used approach to obtain transcript sequences in non-model organisms, notably for performing comparative analyses. However, current bioinformatic pipelines do not take full advantage of pre-existing reference data in related species for improving RNA-Seq assembly, annotation and gene family reconstruction.
Results: We built an automated pipeline named CAARS to combine novel data from RNA-Seq experiments with existing multi-species gene family alignments. RNA-Seq reads are assembled into transcripts by both *de novo* and assisted assemblies. Then, CAARS incorporates transcripts into gene families, builds gene alignments and trees and uses phylogenetic information to classify the genes as orthologs and paralogs of existing genes. We used CAARS to assemble and annotate RNA-Seq data in rodents and fishes using distantly related genomes as reference, a difficult case for this kind of analysis. We showed CAARS assemblies are more complete and accurate than those assembled by a standard pipeline consisting of *de novo* assembly coupled with annotation by sequence similarity on a guide species. In addition to annotated transcripts, CAARS provides gene family alignments and trees, annotated with orthology relationships, directly usable for downstream comparative analyses.
Availability and implementation: CAARS is implemented in Python and Ocaml and is freely available at https://github.com/carinerey/caars.
Contact: carine.rey@ens-lyon.org
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Large scale RNA sequencing (RNA-Seq) is often used in non-model species as a pragmatic alternative to genome sequencing, in particular for comparative analyses (Ozsolak and Milos, 2011; Todd *et al.*, 2016; Wang *et al.*, 2009). However, the assembly of short reads from transcriptome assays into full length transcript sequences poses difficult issues related to repeated regions, variable expression levels, alternative splicing, sequencing errors and composition biases (Garber *et al.*, 2011). Further, the clustering of those sequences into gene families, their alignment and the step of gene tree reconstruction all represent challenges that studies of comparative genomics face without agreed-upon standards.

Different strategies can be used for transcript assembly, depending on the existence of genomic data for closely related species (Conesa *et al.*, 2016; Ockendon *et al.*, 2016). If no sister species with a sequenced genome is available, reads are assembled *de novo* based on overlapping sequences [e.g. Trinity, Grabherr *et al.* (2011)]. Otherwise, genome-guided assembly may be used [e.g. Tophat, Trapnell *et al.* (2009) and Cufflinks, Trapnell *et al.* (2010)]. In that case, reads are aligned to this guide genome, creating clusters

© The Author(s) 2018. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

of reads that are used for local transcript assembly. This strategy is obviously restricted to very closely related species, for which trans-species read mapping is feasible. On more distantly related species, no approach has been proposed for RNA-Seq assembly, but developments have been made for genome assembly. In particular, the Target Restricted Assembly Method (TRAM) by Johnson *et al.* (2013), automated in aTRAM (Allen *et al.*, 2015), reconstructs a gene sequence by an iterative process where reads are collected by sequence similarity to a reference genome using BLAST (Camacho *et al.*, 2009) and then assembled. A different implementation was proposed in Kollector (Kucuk *et al.*, 2017) based on a *k*-mers approach. These methods show encouraging results, but have not been designed to be used on RNA-Seq data and for thousands of genes at a time.

After assembly, transcripts should ideally be annotated with a gene name. Commonly, transcriptome annotation is based on sequence similarity between the transcripts and the transcriptome of already annotated species. This step is most often treated by Reciprocal Best Hits (RBHs) (Rivera *et al.*, 1998), typically using BLAST (Camacho *et al.*, 2009), which cannot handle species–specific duplications (Altenhoff and Dessimoz, 2009; Tekaia, 2016). This is an issue because many genes are duplicated. For instance, in the Ensembl database (Herrero *et al.*, 2016; Yates *et al.*, 2016) 10% of all Human genes have no one-to-one orthology relationships with mouse genes.

In principle, relying on gene phylogenies instead of RBH for annotation allows handling complex homology relationships (Chen *et al.*, 2007; Kristensen *et al.*, 2011; Kuzniar *et al.*, 2008; Tekaia, 2016). We suggest to take such an approach: genes from annotated transcriptomes are clustered into homologous gene families either *de novo* (Kristensen *et al.*, 2011) or using existing families [EnsemblCompara (Herrero *et al.*, 2016), TreeFam (Finn *et al.*, 2014), Hogenom (Penel *et al.*, 2009), PhylomeDB (Huerta-Cepas *et al.*, 2014)]. Then, reconstructed transcripts are integrated into these gene families based on sequence similarity. Alignments and trees are reconstructed for these enlarged gene families. Quality of the trees can be improved by using reconstruction methods that use the information provided by the species tree (Boussau *et al.*, 2013; Ullah *et al.*, 2015). Finally, gene trees are reconciled with a species tree to annotate speciations, duplications and losses (Kristensen *et al.*, 2011). Based on this scenario of gene family evolution, orthology and paralogy relationships are derived, and gene names are propagated from annotated sequences to novel transcripts (Kristensen *et al.*, 2011). In this approach, accurate annotations are an outcome of accurate gene trees.

Here, we present an automated tool, named CAARS, to assemble and annotate the whole transcriptome of non-model organisms from RNA-Seq data, using sequences from one or several species that can be closely or distantly related to guide transcript assembly and annotation. CAARS relies on reference gene alignments and outputs homologous gene sets with high quality phylogenetic trees and orthology relationships that can directly be used for downstream comparative analyses. CAARS improves upon a well-established pipeline in terms of transcriptome completeness, transcript accuracy and annotation accuracy. Thanks to its high quality output gene trees, CAARS also improves upon Ensembl Compara in terms of the number of orthologs it can recover.

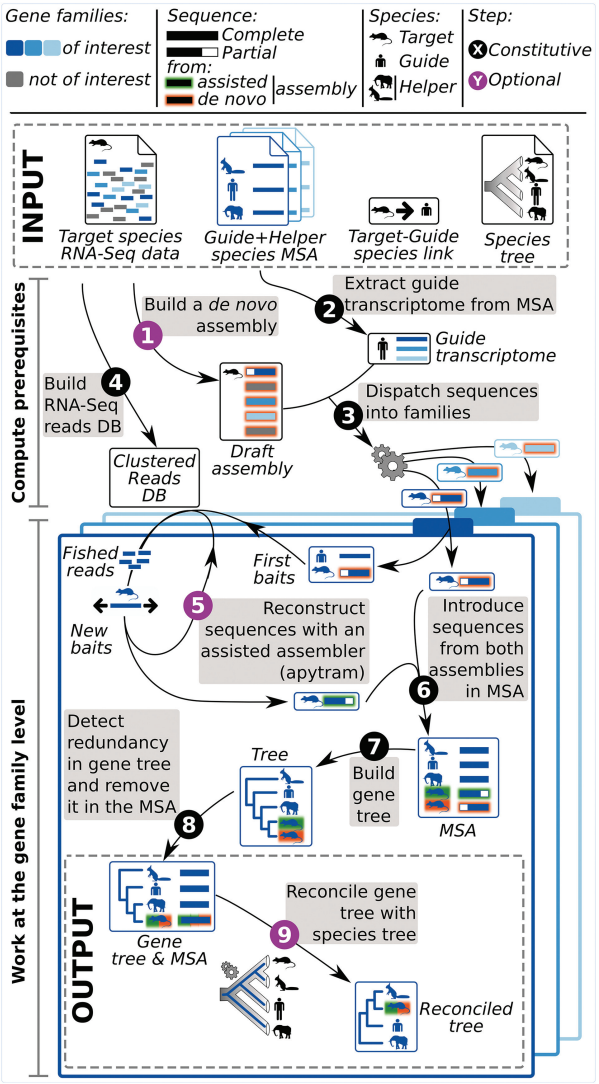

Fig. 1. CAARS overview. Representation of the major steps of CAARS. Steps 1–4 group pre-requisite computations. (**1**) If no draft transcriptome is given in input, RNA-Seq data are *de novo* assembled into a draft transcriptome and coding sequences are parsed to remove 5′ and 3′ UTR. (**2**) Transcriptomes from guide species are extracted from input MSAs to form guide transcriptomes. (**3**) Transcripts from the draft transcriptome are associated to the corresponding gene families by BH against guide transcriptomes. Steps 5–10 group computations made for each family. (**4**) RNA-Seq reads are clustered and formated into a database. (**5**) Transcripts are assembled again with an assisted and iterative method (Apytram). At the first iteration, genes from the guide species and target transcripts from the draft transcriptome corresponding to this family are used as bait sequences to fish reads in the reads database. Mate reads are used to enlarge this batch of reads. Reads are then *de novo* assembled, and a new iteration can begin with the reconstructed sequences as baits. (**6**) Coding sequences from both assemblies are added to the existing gene family alignments. (**7**) A primary gene tree is obtained for the family. (**8**) Redundancy is removed by selecting the longest sequence or by merging sequences from the same species when appropriate. Then, sequences (from target species) with a low-scoring alignment to their sister sequences (from guide or helper species) can be discarded (not shown). A gene tree is re-computed to take into account potential changes. (**9**) The species tree and the gene family tree are used jointly to infer a reconciled gene tree placing gene losses and duplications along the gene family tree

2 Materials and methods

2.1 Outline of CAARS and implementation

The general structure of CAARS is illustrated in Figure 1. As input CAARS requires data from three types of species: the species whose transcriptomes need to be assembled, which we call target species, the species with transcriptomes serving as guide for assembly and annotation, which we call guide species and the species with already assembled transcriptomes because they improve the resolution of

the gene trees in the pipeline, which we call helper species. More particularly, CAARS requires Multiple Sequence Alignments (MSAs) corresponding to gene families containing sequences from guide and helper species, a rooted species tree with all the species and RNA-Seq data for the target species. Finally, we require to specify a set of guide species for each target species (Fig. 1 top): since target species may belong to various taxonomic groups, it may be useful to adapt guide species to each target species.

CAARS is organized in two major parts. The first one sets up several pre-requisites for the second, which is the execution in parallel of a series of steps for each gene family.

First, CAARS performs a *de novo* assembly using the commonly used program Trinity (Grabherr *et al.*, 2011). The *de novo* assembled transcripts are dispatched to gene families using BLAST. In addition, RNA-Seq reads are formatted as BLAST databases, one per target species.

Second, independently for each gene family, CAARS performs another assembly assisted by sequences from the guide species and by *de novo* assembled transcripts. This latter assembly is performed by our in-house software Apytram (Supplementary Fig. S1) (Rey *et al.*, 2017), a multi-species implementation of the TRAM algorithm (Johnson *et al.*, 2013). Importantly Apytram is able to deal with several RNA-Seq samples simultaneously, which improves on the initial implementation (Allen *et al.*, 2015).

Coding regions of transcripts from both *de novo* and assisted assemblies are extracted using Transdecoder (v3.0.1) (http://transdecoder.github.io) and then integrated into the MSAs. At this step, gene families typically include redundant transcripts, which can be alternative transcripts of the same gene at the same locus, or identical transcripts that have been assembled independently by the two methods.

To remove this redundancy, the default option is to select the longest sequence (raw or aligned length), following Yang and Smith (2014). Alternatively, it is possible to merge transcripts from the same species that branch at the same position in the tree, by maximizing the information content in the alignments (Supplementary Fig. S3). Partial sequences may be filtered out based on their alignment to their sister sequence in the tree (see Section 3 or the detailed implementation on the CAARS website).

Then, accurate gene trees are inferred using a phylogenetic pipeline that uses the information coming from both the alignments and the species tree (Boussau *et al.*, 2013) and identifies events of gene duplication and loss in a reconciliation step. Orthology and paralogy relationships are naturally deduced from the reconciled gene trees. Because CAARS grounds assembly and annotation on several guide species at the same time, and not on a single one, it is robust to species-specific gene duplications or losses in the guide transcriptomes.

The method and implementation are detailed in the Supplementary Material provided on the CAARS website. CAARS is written in the Python programing language for all intermediate steps and in the OCaml language for the main program orchestrating all computational steps. This program relies on bistro (Veber, 2017), an OCaml library that manages dependencies between steps, distributed computation and recovery upon error [also known as resume-on-failure ability (Leipzig, 2016)]. For ease of deployment, users do not need to install all external dependencies and may instead use the dedicated docker image available on DockerHub called *carinerey/caars*.

2.2 Assessing CAARS performance

We selected the Human as a guide species for assembling mouse and stickleback transcriptomes. We used their annotated transcriptomes as reference against which to compare the performance of CAARS and of a standard pipeline commonly used for *de novo* transcriptome assembly and annotation. In the following, to avoid ambiguities, we will use 'guide' to name the transcriptomes used to help the assemblies and 'reference' to name the transcriptomes used to benchmark the assemblies. In the standard pipeline, used e.g. in Marra *et al.* (2014); Konczal *et al.* (2014); Pereira *et al.* (2016); Thompson and Ortí (2016) and Ishikawa *et al.* (2016), the assembly is performed *de novo* by Trinity (Grabherr *et al.*, 2011) and the annotation by RBH using BLAST (Camacho *et al.*, 2009).

2.2.1 Dataset, common inputs

We used paired-end RNA-Seq libraries from adult mouse kidneys (2×51 bp, about 12.5–15.3 million reads per library, SRR636916, SRR636917, SRR636918) and from adult stickleback kidneys (2×100 bp, about 16.5 million reads per library, SRR528539, SRR528540).

2.2.2 CAARS additional inputs and assembly

In addition to the reads for the target species and an annotated transcriptome for the guide species, CAARS also requires MSAs corresponding to gene families containing sequences from guide and helper species. We downloaded the Ensembl Compara dataset from the Ensembl database [release 91, Herrero *et al.* (2016); Yates *et al.* (2016)]. This dataset contains MSAs for 22 340 gene families with sequences from 97 chordates, including 71 mammals. From this set of species, we extracted a subset of 17 species of which 13 representative mammals, 1 bird, 1 reptile and 2 fishes. We obtained their phylogeny from the Ensembl Github repository (Herrero *et al.*, 2016; Yates *et al.*, 2016) (Supplementary Fig. S2). We did not want to favor CAARS during the tests, and we voluntarily removed the rodents belonging to the mouse sub-order and the fishes belonging to the stickleback order.

To be usable in CAARS, MSAs must contain at least one sequence from the guide species (Human), and at least 3 species in total (for the reconciliation step). A total of 8622 MSAs satisfy both criteria.

We launched CAARS on a Linux server (16 threads, 64 G RAM) with a running installation of Docker and an imported CAARS image from DockerHub (*carinerey/caars*). CAARS tutorial contains the material (dataset and scripts) to replicate the analysis presented here as a demo and can be found on the wiki page of the Github repository. A smaller data set is also provided for a quick test.

2.2.3 Additional inputs for the standard pipeline and assembly

On the same hardware, for each target species, we first assembled RNA-Seq reads into transcripts with Trinity (Grabherr *et al.*, 2011) using default parameters. Then, we used Cap3 (Huang and Madan, 1999) (default parameters) to assemble overlapping Trinity contigs. We removed redundancy using CD-HIT-EST (Fu *et al.*, 2012), which clusters nucleotide sequences that meet a sequence identity threshold (99%) and finds one representative sequence per cluster (-c 0.99 -n 11 -d 0). We then used Transdecoder (v3.0.1, –retain_long_orfs 150) to extract the coding region of each transcript. Finally, we retained only transcripts associated by RBH with a guide species transcript using BLAST (Camacho *et al.*, 2009) with an *e*-value of $1e-6$ and using `blastn` as task option. In order to assess the impact of the evolutionary distance between the guide and the target species, we run this final step using two different guide species for each target species.

2.2.4 Reference transcriptome

To assess the accuracy of the assemblies and annotations made by CAARS or the standard pipeline, we compared them for each target species with their corresponding known transcriptome. We extracted mouse and stickleback sequences from the Ensembl Compara dataset (v91), as reference transcriptomes (Herrero *et al.*, 2016; Yates *et al.*, 2016). They are composed of 22 388 coding DNA sequences (CDSs) for the mouse and 20 072 for the stickleback distributed in 10 350 gene families. We removed strictly identical sequences using CD-HIT-EST (Fu *et al.*, 2012), keeping respectively 22 060 and 20 020 sequences.

Of note, the mouse and the stickleback sequences are distributed in more families (10 350) than were used as input for CAARS (8622) because CAARS needs Human homologous sequences as bait sequences. So, because they have no homolog in Human, in this intentionally difficult test, CAARS cannot find 1822 mouse and 1766 stickleback sequences. In a real-life situation, users can use more closely related genomes when available, or can use multiple genomes as bait sequences. This would drastically reduce the number of genes without homologs.

2.2.5 Sensitivity measure

We compared the completeness of each CAARS and standard assemblies with respect to the corresponding reference transcriptome. For each gene of the reference transcriptome and each assembly, we retrieved the RBH sequence when available, using BLAST with a stringent *e*-value threshold ($1e-10$) and otherwise default parameters (Camacho *et al.*, 2009). In the case where no RBH was found, we considered that this gene was missing from the assembly. Completeness statistics are provided in the Table 1.

2.2.6 Identification of partial and alternative transcripts

Assembled transcripts may be incomplete compared to the reference transcriptome, because they represent shorter alternative transcripts, or because the coverage for this transcript in the kidney expression data is low. To identify these partial and alternative transcripts, we aligned each transcript of each assembly to the reference transcriptome [BLAST (Camacho *et al.*, 2009) with `evalue=1e-10`, and otherwise default parameters], and retrieved the Best Hit sequence (BH). By using BH instead of RBH, we allow that several sequences in a given assembly match a single transcript of the reference transcriptome. This ensures all sequences of an assembly with a possible hit have an associated reference sequence.

2.3 Overlap between reconstructed and reference transcript sequences

We analyzed the sequences of transcripts present in both the reference transcriptome and a reconstructed assembly to estimate whether reconstructed transcripts are longer, shorter or generally different from the reference transcripts. To this end we computed two indices, for a given reference sequence R of length len_R and a given query sequence Q of length len_Q. We used Mafft (Katoh *et al.*, 2002) with default parameters to align the R and Q sequences and we computed the number of aligned positions, len_{ali}, between R and Q (without gaps) in the alignment.

The two indices are then calculated using these formula: $P_{\text{reference}} = \frac{len_{\text{ali}}}{len_R} \times 100$ and $P_{\text{query}} = \frac{len_{\text{ali}}}{len_Q} \times 100$.

2.4 Quantification of expression levels

For each target species, we quantified the levels of gene expression for the three transcriptomes (the reference transcriptome and the CAARS and standard assemblies) using Kallisto (Bray *et al.*, 2016) and the RNA-Seq libraries mentioned earlier.

2.5 Evaluation of sets of orthologs

Orthologs predicted by CAARS were extracted for different sets of species, including or not including target species. We compared the set of orthologs obtained without target species to the 'high confidence' orthologs available on the Ensembl Compara database.

3 Results

We used CAARS to assemble and annotate transcriptomes of target species using a guide species too divergent for a genome-guided assembly, and we compared it with a standard pipeline combining *de novo* assembly and annotation by RBH. We selected two target species, the mouse and the stickleback, for which gene sequences and annotations are well-established. Kidney RNA-Seq libraries of these two species were used for assemblies, and their genomes were used later to evaluate the accuracy of CAARS. We used the Human as

Table 1. Statistics of the CAARS assembly compared to a standard assembly

Target species	Assembly method	Options	Guide or reference species	Divergence (in Mya)	# of seqs in the assembly	Precision: # of seqs associated with a seq in the target species	Sensitivity: % of seqs of the target species associated with a seq in the assembly[a] (%)
Mouse	CAARS	by/By default	Human	90[b]	12 421	11 500 (92.6%)	88.1
		Filter at 25%	Human	90[b]	11 093	10 779 (97.2%)	82.6
	Standard pipeline	Filter based on RBH	Human	90[b]	10 808	10 749 (99.5%)	82.4
			Guinea pig	70[c]	10 572	10 496 (99.3%)	80.5
			Squirrel	70[c]	10 594	10 511 (99.2%)	80.6
Stickleback	CAARS	by/By default	Human	400[b]	10 878	9570 (88.0%)	66.7
		Filter at 25%	Human	400[b]	9789	8931 (91.2%)	62.2
	Standard pipeline	Filter based on RBH	Human	400[b]	8758	8189 (93.5%)	57.1
			Zebrafish	225[d]	11 273	10 483 (93.0%)	73.1

[a]This is calculated as the ratio of the number of target reference sequences with an associated sequence in the assembly over the number of target reference sequences expressed more than 1 count per base in the library (for the Mouse, 13 046 seqs, and for the Stickleback, 14 349 seqs).

[b]Hedges *et al.* (2015).

[c]Fabre *et al.* (2012).

[d]Betancur-R *et al.* (2015).

guide species because it is well-annotated and also because it is quite divergent from the mouse and the stickleback [divergence around 90 million years ago—Mya—and 400Mya (Hedges *et al.*, 2015)], far too distant for trans-species mapping (Conesa *et al.*, 2016; Ockendon *et al.*, 2016; Torres-Oliva *et al.*, 2016). Overall we conservatively chose unfavorable test settings to assess the performance of CAARS. To this end, first we chose a single distant guide species to assemble target transcriptomes. Second, we also chose a Teleost fish as one of our target species because it contains many duplicate genes due to whole genome duplications, which makes the reconstruction difficult.

3.1 CAARS has a better sensitivity than a standard assembly pipeline

CAARS took 4 days to reconstruct 12 421 mouse CDSs and 10 878 stickleback CDSs included in 7049 families (Table 1). These figures are in accordance with the fact that about 10 000 ∼ 13 000 genes are expressed in Human kidney [10 000 using a threshold > 5 FPKM (Fagerberg *et al.*, 2014; Uhlen *et al.*, 2015).)]. This demonstrates that CAARS may be used with a distant guide species in a reasonable amount of time.

Independently, we ran a standard *de novo* assembly on the same data and the same hardware, which took about 4 h. After filtering transcripts using RBH, there is a large influence of the divergence between the guide and the target species on the accuracy of the standard pipeline (Table 1). For example, for the stickleback assembly, we obtained 8758 sequences using the Human as guide species, against 11 273 if we used zebrafish, a less divergent guide species. The genes only recovered with the zebrafish as a guide species probably mostly correspond to fish-specific gene families. To interpret the differences between the pipelines in terms of their power to detect transcripts rather than in terms of whether they can detect fish-specific gene families or not, we discuss in the following the assembly with the Human as guide for both the mouse and the stickleback and for both methods.

To examine the completeness of both assemblies, we associated by RBH the target species reference transcripts to the reconstructed sequences. Out of all 22 060 mouse reference sequences, 10 672 sequences were found by both assemblies, 828 sequences only by CAARS, 77 only by the standard assembly (Fig. 2A). Therefore, in this case, CAARS retrieves 7% more transcripts than a pipeline classically used for *de novo* assemblies.

For the stickleback, on a total of 20 020 sequences, 7687 sequences were found by both assemblies, 1883 sequences only by CAARS, 502 only by the standard assembly (Fig. 2B), meaning CAARS allows a gain of 17% of transcripts.

The sequence of genes not expressed in the kidney cannot be reconstructed from our RNA-Seq data. Conversely, the sequence of highly expressed genes should be easier to obtain. We wished to establish more systematically the link between the level of gene expression and the accuracy of sequence assembly. For the mouse, we measured gene expression levels in kidney using the reference transcriptome. As expected, very weakly expressed genes ($\leq$1 count per base) are rarely retrieved by CAARS, or by the standard assembly.

If we focus only on genes expressed more than two counts per base pair (meaning with an average sequencing coverage $\geq 2\times$), CAARS detects more genes than the standard assembly. In the mouse, 93% of the reference sequences have an associated sequence in the CAARS assembly and 89% in the standard assembly (Fig. 2A). In the stickleback, the improvement of CAARS over the standard assembly is more pronounced: 66% compared to 59% (Fig. 2B).

As CAARS annotates transcripts using a phylogeny, it is expected to be more effective than the standard pipeline, which used a RBH annotation, to retrieve genes in the target species that have been duplicated since the divergence with the guide species (one-to-many or many-to-many orthologs in Ensembl Compara). Such orthologs correspond to 8% of the expressed genes (>1 count/pb) in the mouse and 30% in the stickleback. Expectedly, they are more often retrieved by CAARS than the standard pipeline, whether for the mouse, where 63.0% of them are retrieved compared to 57.5% in the standard pipeline, or for the stickleback (58.9 versus 46.6%).

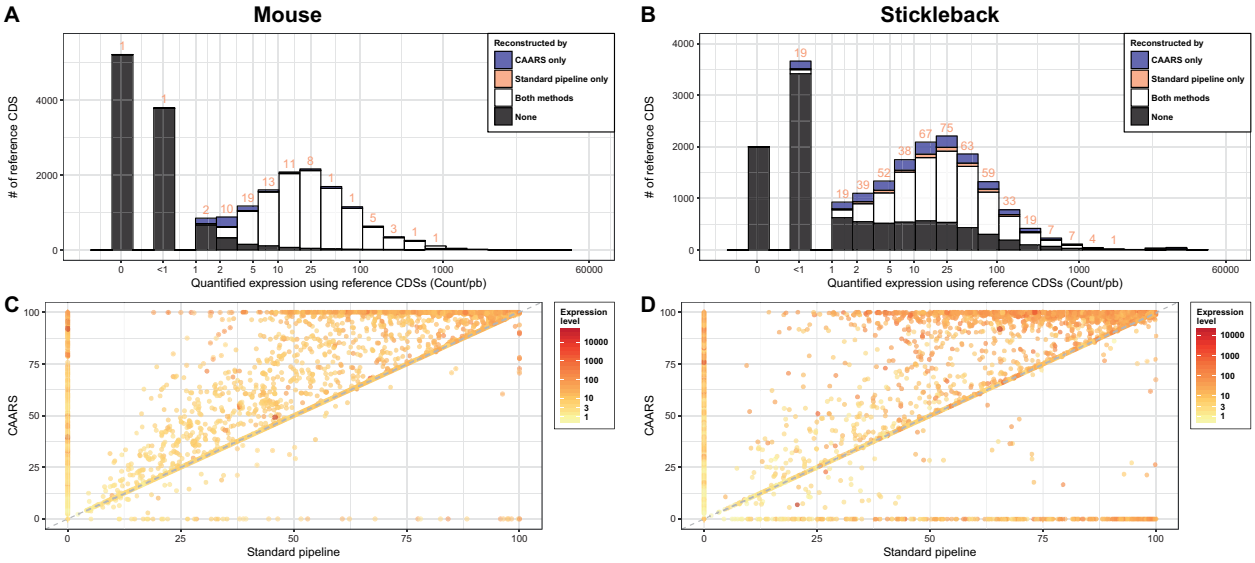

Fig. 2. Comparison between the CAARS and standard assemblies for the mouse and the sickleback. (**A, B**) Number of reference CDSs associated by a RBH with a transcript from CAARS assembly only, the standard assembly only, both assemblies or none. Expression was quantified on the reference transcriptome in Count per base; 0 means no detected expression. (**C, D**) Proportion of the reference transcript ($P_{reference}$) aligned with its CAARS assembly counterpart (*y* axis) or its standard assembly counterpart (*x* axis). Marginal distributions can be seen in Supplementary Figures S5A and S6A. Dots with a null value on one of the axes represent genes not recovered by the corresponding method. The results for CAARS are similar whether we use the filtering step or not

3.2 CAARS assembly provides more complete transcripts

The completeness of an assembly cannot be assessed solely on the basis of the number of recovered transcripts. We measured the coverage of each reference transcript, and compared the results between the two assemblies for each target species. Gene coverage was estimated on pairs of sequences with a transcript from the reference transcriptome and the corresponding reconstructed transcript. We computed the percentage of the reference transcript that aligns with the reconstructed transcript ($P_{\text{reference}}$) for 11 577 Ensembl CDS with a matched sequence in at least one of the assemblies (Fig. 2C). Both for the mouse and the stickleback, CAARS better recovers the reference transcripts. In the experiment on mouse libraries, the percentage of reconstruction of reference transcripts is identical for 8937 sequences, better in the CAARS assembly for 2509 sequences (with an average increase of 24.8%) and better in the standard assembly for 131 sequences (smaller improvement, 10.4%) (Fig. 2C). For the stickleback, 3813 sequences (out of 10 072) are better in the CAARS assembly (26.5% average increase) against 566 for the standard pipeline (12.7% average increase) (Fig. 2D).

CAARS transcripts are longer than those from the standard assembly but they are also more often complete. Total of 8134 CAARS mouse transcripts are complete or sub-complete ($P_{\text{reference}} > 95$), which is better than the 7246 sub-complete transcripts obtained with the standard pipeline and respectively 6465 compared to 4853 for the stickleback transcripts (Table 2).

A potential issue for assembled transcripts is the merging of two transcripts into a chimeric sequence. Such chimeric sequences will be longer than the reference CDSs and characterized by a low P_{query} value. For the mouse and the stickleback, distributions are similar for CAARS and the standard pipeline, with no excess of low values (Supplementary Figs S5 and S6), meaning that the potential numbers of chimeric sequences found in the CAARS assembly and in the standard pipeline are small.

We estimated the quality of the assembly using another criterion, gene expression levels. The levels of expression obtained from both RNA-Seq assemblies are well correlated to reference expression levels ($R^2 = 0.98$) (Supplementary Fig. S4A and B).

3.3 An optional filter can discard redundant and low quality sequences

CAARS assemblies have a better sensitivity than the standard pipeline but, with default parameters, the precision, the number of sequences associated by RBH with a reference transcript, is lower. Among sequences without RBH, most (914, ie. 92%) have a unidirectional blast hit in the reference transcriptome. Besides, they have a low $P_{\text{reference}}$ (Supplementary Fig. S5A left) and a high P_{query} (Supplementary Fig. S5B left), meaning they are partial assemblies or assemblies of small redundant transcripts.

Table 2. Comparison of the alignment statistics on reference genes of the CAARS assembly and a standard assembly using the Human as guide species

Assembly method	# of sub-complete CDSs[a]	
	Mouse	Stickleback
CAARS (by default)	8134 (65.5%)	6465 (59.4%)
CAARS (with filter at 25%)	8131 (73.3%)	6457 (66.0%)
Standard pipeline	7246 (67.0%)	4853 (55.4%)

[a]A CDS is counted as sub-complete if its $P_{reference}$ is superior to 95, ie. it covers at least 95% of its reference CDS. The proportion of sub-complete transcripts is obtained by dividing the number of sub-complete transcripts by the total number of transcripts predicted by the method.

Partial CDSs can introduce noise in subsequent analyses and users may wish to flag them. We reasoned that transcript length should not vary too much among related species. Hence, transcripts assembled by CAARS with a length similar to the length of their sister species in the tree are expected to be complete. We verified that, indeed, the percentage of reconstruction of the neighborhood sequence in the gene tree is a good proxy for the percentage of reconstruction of the reference sequence (Supplementary Figs S5B and S6B). An alternative would have been to filter them out based on their expression level but we have found that it is not correlated to the assembly quality (Supplementary Fig. S7A). A threshold can be applied on this criterion to discard partial CDSs or select high quality CDSs (Supplementary Fig. S7C).

For instance, with a threshold at 25%, CAARS puts aside 1328 (10.7%) sequences for the mouse and 1089 (10.0%) for the stickleback. This filter allows increasing the precision from 92.6 to 97.2% for the mouse and from 88.0 to 91.2% for the stickleback (Table 1). The filter increases the proportion of complete sequences in the assembly from 65.5 to 73.3% for the mouse and from 59.4 to 66.0% for the stickleback, which is better than the standard pipeline (Table 2). The sensitivity decreases a little but stays above the standard pipeline, so some sequences with a RBH have been discarded but these sequences are partial (low $P_{\text{reference}}$ and high P_{query}) (Supplementary Fig. S5C). The stringency of this filter can be set by the user when using CAARS.

3.4 CAARS produces sets of orthologs defined by phylogeny

CAARS returns MSA and reconciled gene trees ready to use for comparative analyses. From these reconciled gene trees, the user may use CAARS to infer orthology relationships between all sequences. This information is stored in a table which can be easily mined to retrieve all one-to-one orthologs for a given subset of species (Fig. 3). These subsets can include, or not, the target species. For example, we find 4850 sets of one-to-one orthologs with one gene per mammalian species of our dataset (Fig. 3). This is substantially more than the number found by the equivalent request in Ensembl Compara (4505 sets of genes with high confidence one-to-one orthology relationship), a gain attributable to our reconciliation step, which improves gene trees (Boussau *et al.*, 2013). We also extracted sets of orthologs for subset of species that include a target species, and obtained reasonable numbers (8705 for Human/mouse comparisons, 6435 for comparisons across rodents, 5 666 for comparisons across fishes Fig. 3). We cannot compare these numbers to numbers from the Compara database, because the mouse or stickleback gene complements are partial, being reconstructed from libraries of specific organs.

4 Discussion

CAARS is a pipeline that can be used to assemble transcriptomic data sets for comparative analyses. To assess its first steps, we compared CAARS with a standard pipeline for comparative assembly and annotation of RNA-Seq data. We found that CAARS is more sensitive since it finds more transcripts that are more complete. This better sensitivity is accompanied by a high precision (in particular with the optional filter), as a large majority of the sequences can be associated by RBH with a sequence of its reference transcriptome.

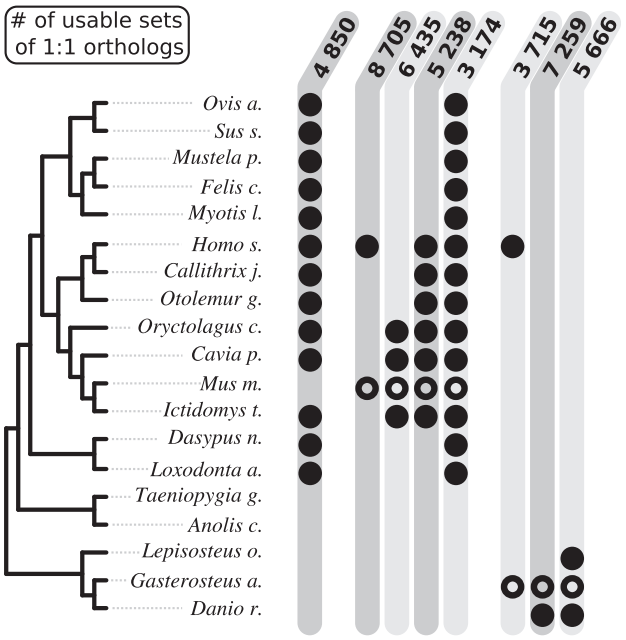

Fig. 3. CAARS outputs ready to use sets of one-to-one orthologs, inferred from reconciled phylogenies. Each column corresponds to the number of sets of one-to-one orthologs containing at least the species indicated by circles. Empty circles correspond to target species (reconstructed with RNA-Seq data), filled circles to guide or helper species (genome available)

Expression levels are at least as well estimated by CAARS as with the standard pipeline.

The improvement over the standard pipeline can be attributed to two novel features: (i) CAARS implements a trans-species assembly, based on one or several guide sequences, which can be distantly related. This is demonstrated here with mouse and stickleback RNA-Seq data assembled using a very distant Human guide species; (ii) CAARS annotates the transcripts by integrating them in phylogenies built with a set of helper sequences. This is demonstrated here with families from the Ensembl Compara database.

In addition to the steps of transcript detection and assembly, CAARS generates gene alignments, trees and sets of orthologs that can be directly used in subsequent comparative analyses. Notably, we found that it could recover more sets of orthologs than the Ensembl Compara pipeline and was better at recovering one-to-many or many-to-many orthologs than the standard pipeline, probably because it relies on gene trees reconstructed with a reconciliation approach. Besides, CAARS is easy to use, robust and modular.

4.1 CAARS uses one or several possibly divergent species to generate assemblies

In our test we used Human transcripts as a guide to assemble mouse and stickleback RNA-Seq data. Guide transcriptomes with good annotations are important as they reduce the likelihood that a gene has been mis-annotated or missed altogether.

In the case of the stickleback, the sensitivity of the CAARS assembly is better than that of the standard pipeline using the Human as guide species. It is not as good as the sensitivity of the standard pipeline using the zebrafish, but that is expected since the zebrafish is much more closely related to the stickleback than the Human is.

In the case of the mouse, very few transcripts are recovered by the standard pipeline only. For the stickleback, the number of transcripts found only by the standard pipeline is larger (see figures in red, Fig. 2B). This difference is due to a more stringent threshold inside CAARS (not shown). CAARS nonetheless clearly outperforms the standard pipeline in sensitivity in both cases (Fig. 2B).

It is not always easy to find well-annotated and closely related species that can serve as guide species. In many cases, well annotated genomes will be distant and closely related genomes of weaker quality. To improve the performance of guide-based assembly in those situations, CAARS can use several guide species at the same time. This reduces the likelihood that a gene is missed because of a missing bait, since it would have to be absent from all guide species. In a real study, where we want to optimize the result and not challenge CAARS we would add the zebrafish and the spotted gar as guide species to assemble the stickleback transcriptome and the squirrel and the guinea pig to assemble the mouse transcriptome.

In addition, several target species can be assembled at the same time. This can benefit the assembly of target species because during the step of assisted assembly (by Apytram), all the target species sharing the same guide species will participate and help each other in fishing the reads. This also allows breaking the distance between guide and target species.

4.2 CAARS integrates assembled transcripts into families and builds gene phylogenies

CAARS belongs to a small group of pipelines [e.g. Agalma, Dunn *et al.* (2013)] that explicitly aim at assembling data sets for phylogenomic analyses providing homologous and orthologous sequences, MSAs and gene trees from RNA-Seq data. However making use of one or several distant guide species at the same time, using phylogeny for annotation and providing sets of one-to-one orthologs are to our knowledge new features.

Other automatized methods that can assemble RNA-Seq data using closely related helper species [Agalma, Dunn *et al.* (2013), BRANCH, Bao *et al.* (2013) or FRAMA, Bens *et al.* (2016)] are based on direct sequence similarities (mapping or alignment on guide genome). However studies showed the negative correlation between annotation quality and divergence with guide species used for trans-species assembly/annotation (Ockendon *et al.*, 2016; Torres-Oliva *et al.*, 2016; Ungaro *et al.*, 2017; Vijay *et al.*, 2013). For instance, Vijay *et al.* (2013) recommends not to map directly on the guide genome when there is more than 15% of sequence divergence between the target and the guide species, which corresponds to the median nucleotide divergence in one-to-one orthologs between mouse and Human (Church *et al.*, 2009).

The phylogenetic framework used by CAARS allows identifying redundant transcripts that have been assembled more than once, and collapsing them or selecting the longest (the default). Gene trees are also used by CAARS to identify incomplete transcripts by comparison with neighboring sequences and filter them out (Supplementary Fig. S5). It remains to be seen how CAARS would behave on datasets containing lots of recent duplicates; in particular, the option to merge monophyletic transcripts from the same species may create chimeric transcripts and should be used knowingly.

A limitation of CAARS remains the usage in input of MSAs. However, nowadays, there are several public database containing such MSAs [Orthomam (Ranwez *et al.*, 2007), Hogenom (Penel *et al.*, 2009), PhylomeDB (Huerta-Cepas *et al.*, 2014), TreeFam (Finn *et al.*, 2014), EnsemblCompara (Herrero *et al.*, 2016)].

4.3 CAARS is robust and easy to install and use

CAARS is a complex pipeline combining existing software and newly developed programs, and is built to be able to analyze

thousands of gene families at once. In particular, CAARS includes Apytram (Rey *et al.*, 2017), a multi-species and more accurate re-implementation of TRAM (Johnson *et al.*, 2013), which initially introduced the idea of a trans-species assembler at the level of a single gene. For robustness and traceability, and to enable recovery upon error or iterative use, it uses the *bistro* library (Veber, 2017).

For an easy installation, we packaged CAARS into a Docker image, so that there is no need to install any dependency once Docker has been installed on the system, which can be a Mac, Windows or Linux system. Further, by using Docker we ensure that the intended versions of all tools of the pipeline will be used, irrespective of what is installed on the host machine system. The results produced by CAARS are thus fully reproducible. Finally, the use of Docker ensures only minimal penalties on computational efficiency. However, users can also opt to install the full pipeline without using Docker.

Here we demonstrated the use of CAARS at the whole transcriptome level but the program can be used at a much smaller scale, for a single gene family, or even for a single gene. CAARS may also be used for integrating transcripts obtained from a pre-existing assembly into gene families. This is easily feasible by switching off the step of assisted reconstruction in the input option file (explained in the tutorial on CAARS's website).

Although CAARS is slower than the standard pipeline, it provides not only the assembly and annotation, but also gene family alignments, reconciled genes trees and sets of orthologous genes. In many cases, such data may be used directly for subsequent analyses.

5 Conclusion

We have introduced CAARS, a new pipeline for the comparative assembly and annotation of transcripts in non-model species. Because it operates within a phylogenetic framework, it can use both closely related and distantly related species. In addition to annotated transcripts, it also provides gene family alignments and trees built using state-of-the-art methods, which can be directly used for downstream analyses. On data coming from the Ensembl database, it compared favorably to a pipeline combining Trinity and BLAST, and provided more complete sets of orthologs than Ensembl. CAARS could therefore be used in a variety of situations where transcript assembly needs to be of high quality, for instance for comparing gene expression or gene sequences across species.

Acknowledgements

We would like to thank T. Lorin, L. Taulelle, R. Allio for their contributions to testing and releasing of CAARS. We also thank the French Institute of Bioinformatics (IFB, ANR-11-INBS-0013) and the PSMN computing center of ENS de Lyon for providing storage and computing resources.

Funding

The research presented here was supported by the Convergenomix project [ANR-15-CE32-0005]. C.R. was supported by a PhD fellowship (CDSN) from the Ecole Normale Supérieure of Lyon.

Conflict of Interest: none declared.

References

Allen,J.M. *et al.* (2015) aTRAM - automated target restricted assembly method: a fast method for assembling loci across divergent taxa from next-generation sequencing data. *BMC Bioinformatics*, **16**, 98.

Altenhoff,A.M. and Dessimoz,C. (2009) Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Comput. Biol.*, **5**, e1000262.

Bao,E. *et al.* (2013) BRANCH: boosting RNA-Seq assemblies with partial or related genomic sequences. *Bioinformatics*, **29**, 1250–1259.

Bens,M. *et al.* (2016) FRAMA: from RNA-seq data to annotated mRNA assemblies. *BMC Genomics*, **17**, 54.

Betancur-R,R. *et al.* (2015) Fossil-based comparative analyses reveal ancient marine ancestry erased by extinction in ray-finned fishes. *Ecol. Lett.*, **18**, 441–450.

Boussau,B. *et al.* (2013) Genome-scale coestimation of species and gene trees. *Genome Res.*, **23**, 323–330.

Bray,N.L. *et al.* (2016) Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.*, **34**, 525.

Camacho,C. *et al.* (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

Chen,F. *et al.* (2007) Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS One*, **2**, e383.

Church,D.M. *et al.* (2009) Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS Biol.*, **7**, e1000112.

Conesa,A. *et al.* (2016) A survey of best practices for RNA-seq data analysis. *Genome Biol.*, **17**, 13.

Dunn,C.W. *et al.* (2013) Agalma: an automated phylogenomics workflow. *BMC Bioinformatics*, **14**, 1–17.

Fabre,P.-H. *et al.* (2012) A glimpse on the pattern of rodent diversification: a phylogenetic approach. *BMC Evol. Biol.*, **12**, 88.

Fagerberg,L. *et al.* (2014) Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol. Cell. Proteomics*, **13**, 397–406.

Finn,R.D. *et al.* (2014) Pfam: the protein families database. *Nucleic Acids Res.*, **42**, D222–D230.

Fu,L. *et al.* (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–3152.

Garber,M. *et al.* (2011) Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat. Methods*, **8**, 469–477.

Grabherr,M.G. *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.*, **29**, 644–652.

Hedges,S.B. *et al.* (2015) Tree of life reveals clock-like speciation and diversification. *Mol. Biol. Evol.*, **32**, 835–845.

Herrero,J. *et al.* (2016) Ensembl comparative genomics resources. *Database (Oxford)*, **2016**, bav096.

Huang,X. and Madan,A. (1999) CAP3: a DNA sequence assembly program. *Genome Res.*, **9**, 868–877.

Huerta-Cepas,J. *et al.* (2014) PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome. *Nucleic Acids Res.*, **42**, D897–D902.

Ishikawa,M. *et al.* (2016) Different endosymbiotic interactions in two hydra species reflect the evolutionary history of endosymbiosis. *Genome Biol. Evol.*, **8**, evw142.

Johnson,K.P. *et al.* (2013) Next-generation phylogenomics using a target restricted assembly method. *Mol. Phylogenetics Evol.*, **66**, 417–422.

Katoh,K. *et al.* (2002) Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Res.*, **30**, 3059–3066.

Konczal,M. *et al.* (2014) Accuracy of allele frequency estimation using pooled RNA-Seq. *Mol. Ecol. Resour.*, **14**, 381–392.

Kristensen,D.M. *et al.* (2011) Computational methods for gene orthology inference. *Brief. Bioinform.*, **12**, 379–391.

Kucuk,E. *et al.* (2017) Kollector: transcript-informed, targeted de novo assembly of gene loci. *Bioinformatics*, **18**, 821–829.

Kuzniar,A. *et al.* (2008) The quest for orthologs: finding the corresponding gene across genomes. *Trends Genet.*, **24**, 539–551.

Leipzig,J. (2016) A review of bioinformatic pipeline frameworks. *Brief. Bioinform.*, **18**, 530–536.

Marra,N.J. *et al.* (2014) Natural selection and the genetic basis of osmoregulation in heteromyid rodents as revealed by RNA-seq. *Mol. Ecol.*, **23**, 2699–2711.

Ockendon,N.F. *et al.* (2016) Optimization of next-generation sequencing transcriptome annotation for species lacking sequenced genomes. *Mol. Ecol. Resour.*, **16**, 446–458.

Ozsolak,F. and Milos,P. (2011) RNA sequencing: advances, challenges and opportunities. *Nat. Rev. Genet.*, **12**, 87–98. Epub 2010 Dec 30. ST – RNA sequencing: adv.

Penel,S. *et al*. (2009) Databases of homologous gene families for comparative genomics. *BMC Bioinformatics*, **10** (Suppl. 6), S3.

Pereira,R.J. *et al*. (2016) Transcriptome-wide patterns of divergence during allopatric evolution. *Mol. Ecol.*, **25**, 1478–1493.

Ranwez,V. *et al*. (2007) Orthomam: a database of orthologous genomic markers for placental mammal phylogenetics. *BMC Evol. Biol.*, **7**, 241.

Rey,C. *et al*. (2017) apytram v1.1. *Zenodo*. (doi: 10.5281/zenodo.804416).

Rivera,M.C. *et al*. (1998) Genomic evidence for two functionally distinct gene classes. *Proc. Natl. Acad. Sci. USA*, **95**, 6239–6244.

Tekaia,F. (2016) Inferring orthologs: open questions and perspectives. *Genomics Insights*, **9**, 17–28.

Thompson,A.W. and Ortí,G. (2016) Annual Killifish transcriptomics and candidate genes for metazoan diapause. *Mol. Biol. Evol.*, **33**, 2391–2395.

Todd,E.V. *et al*. (2016) The power and promise of RNA-seq in ecology and evolution. *Mol. Ecol.*, **25**, 1224–1241.

Torres-Oliva,M. *et al*. (2016) A robust (re-)annotation approach to generate unbiased mapping references for RNA-seq-based analyses of differential expression across closely related species. *BMC Genomics*, **17**, 392.

Trapnell,C. *et al*. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.

Trapnell,C. *et al*. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28**, 511–515.

Uhlen,M. *et al*. (2015) Tissue-based map of the human proteome. *Science*, **347**, 1260419–1260419.

Ullah,I. *et al*. (2015) Integrating sequence evolution into probabilistic orthology analysis. *Syst. Biol.*, **64**, 969–982.

Ungaro,A. *et al*. (2017) Challenges and advances for transcriptome assembly in non-model species. *PLoS One*, **12**, e0185020.

Veber,P. (2017) bistro v0.3.0. *Zenodo*. (doi: 10.5281/zenodo.815611).

Vijay,N. *et al*. (2013) Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive in silico assessment of RNA-seq experiments. *Mol. Ecol.*, **22**, 620–634.

Wang,Z. *et al*. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, **10**, 57–63.

Yang,Y. and Smith,S.A. (2014) Orthology inference in nonmodel organisms using transcriptomes and low-coverage genomes: improving accuracy and matrix occupancy for phylogenomics. *Mol. Biol. Evol.*, **31**, 3081–3092.

Yates,A. *et al*. (2016) Ensembl 2016. *Nucleic Acids Res.*, **44**, D710–D716.