bioRxiv preprint doi: https://doi.org/10.1101/2022.01.02.473666; this version posted November 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

ClusTrast: a short read *de novo* transcript isoform assembler guided by clustered contigs

Karl Johan Westrin, westrin@kth.se[1]
Warren W. Kretzschmar, wk@warrenwk.com[1,2]
Olof Emanuelsson, olofem@kth.se[1]

[1] Science for Life Laboratory, Department of Gene Technology, KTH Royal Institute of Technology, SE-171 65, Solna, Sweden
[2] Center for Hematology and Regenerative Medicine (HERM), Department of Medicine Huddinge, Karolinska Institute, SE-141 52, Flemingsberg, Sweden

Abstract

Background: Transcriptome assembly from RNA-sequencing data in species without a reliable reference genome has to be performed *de novo*, but studies have shown that *de novo* methods often have inadequate ability to reconstruct transcript isoforms. We address this issue by constructing an assembly pipeline whose main purpose is to produce a comprehensive set of transcript isoforms.
Results: We present the *de novo* transcript isoform assembler ClusTrast, which clusters a set of guiding contigs by similarity, aligns short reads to the guiding contigs, and assembles each clustered set of short reads individually. We tested ClusTrast on datasets from six eukaryotic species, and showed that ClusTrast reconstructed more expressed known isoforms than any of the other tested *de novo* assemblers, at a moderate reduction in precision. For recall, ClusTrast was on top in the lower end of expression levels (<15% percentile) for all tested datasets, and over the entire range for almost all datasets. Reference transcripts were often (35–69% for the six datasets) reconstructed to at least 95% of their length by ClusTrast, and more than half of reference transcripts (58–81%) were reconstructed with contigs that exhibited polymorphism, measuring on a subset of reliably predicted contigs.
Conclusion: We suggest that ClusTrast can be a useful tool for studying isoforms in species without a reliable reference genome, in particular when the goal is to produce a comprehensive transcriptome set with polymorphic variants.

1 Background

In eukaryotes, many genes can produce RNA transcripts of differing base sequences called transcript isoforms. Transcript isoforms are created by alternative transcriptional start sites, splicing, or polyadenylation. Controlling transcript isoform expression is one way for a cell to regulate protein expression and thereby its behavior (Wang et al., 2008; Barbosa-Morais et al., 2012; Floor and Doudna, 2016). Changes in transcript isoform expression have been associated with developmental changes and tissue specificity in eukaryotes, and disease in humans (Fackenthal and Godley, 2008; Sterne-Weiler and Sanford, 2014; Xiong et al., 2015; Akhter et al., 2018).

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.02.473666; this version posted November 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Thus, it is often important to clarify not only what genes are expressed but also which transcript isoforms are expressed.

The expression of genes and transcripts is often studied by RNA-sequencing, where short reads (SRs) derived from massively parallel shotgun sequencing are aligned to an organism's reference genome. With this approach, reconstructing transcripts is possible by using the reference genome as a guide (Garber et al., 2011). However, many non-model organisms do not have a high-quality reference genome available. In such cases, a commonly used approach is *de novo* assembly in which transcripts are assembled from the reads only. The assembled transcripts are sometimes referred to as contigs or reconstructed transcripts. Popular tools to perform *de novo* transcriptome assembly include Trans-ABySS (Robertson et al., 2010), Trinity (Grabherr et al., 2011), Oases (Schulz et al., 2012), and SOAP-denovo-Trans (Xie et al., 2014). An overview of current transcriptome assemblers is available, e.g., in Hölzer and Marz (2019).

In principle, these tools can also reconstruct transcript isoforms of the expressed genes, but in practice their sensitivity is poor. In *Mus musculus*, Schulz et al. (2012) reported that Oases, Trans-ABySS, and Trinity assembled 1.21, 1.25, and 1.01 transcripts per gene, respectively, whereas a reference-based assembler reconstructed 1.56 transcripts per gene. Bushmanova et al. (2019) also observed poor transcript isoform reconstruction performance of transcriptome assembly methods: while their method, rnaSPAdes, outperformed the other compared assemblers in gene reconstruction in *Mus musculus*, it assembled only 1.02 transcripts per gene. In the same comparison, Trinity managed to assemble the most transcripts, with a ratio of 1.11 transcripts per gene. The insufficient ability of current *de novo* transcriptome assembly approaches to reconstruct all expressed transcript isoforms of a gene was evident to us in our work on the DAL19 gene in spruce, *Picea abies* (Akhter et al., 2018): Only one out of four confirmed DAL19 transcript isoforms was reconstructed to at least 90% using Oases and two using Trinity. We performed a directed assembly that managed to reconstruct three of the four transcript isoforms, but this method did not scale to whole transcriptome assembly. These examples, and others, e.g. Hayer et al. (2015) and Thind et al. (2021), demonstrate that there is still much room for improvement in *de novo* transcript isoform assembly.

Another observation concerns the imperfect overlap between the sets of reconstructed transcripts from different *de novo* assembly tools. Smith-Unna et al. (2016) noted that out of Oases, Trinity, and SOAP-denovo-Trans, each assembler reconstructed a large number of *bona fide* transcripts that neither of the other assemblers managed to reconstruct. They concluded that combining assembly methods may be an effective way to improve the detection rate of transcripts.

We report the *de novo* transcriptome assembler ClusTrast, which builds upon our previous experience of transcript isoform assembly (Akhter et al., 2018). The main purpose of ClusTrast is to provide a comprehensive set of transcript isoforms, using only sequence reads as input, and with the explicit intent to prioritize recall. The ClusTrast pipeline combines two assembly methods, Trans-ABySS and Shannon, incorporates a novel approach to clustering guiding contigs, assigns short reads to the clusters, and finally performs a cluster-wise assembly of the clustered short reads. We assessed transcript isoform reconstruction performance of ClusTrast and several *de novo* transcriptome assemblers in six eukaryotic organisms and found that ClusTrast reconstructed more known transcript isoforms than any other assembler and reconstructed unknown (including misassembled) transcripts at a rate comparable to other assemblers.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.02.473666; this version posted November 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

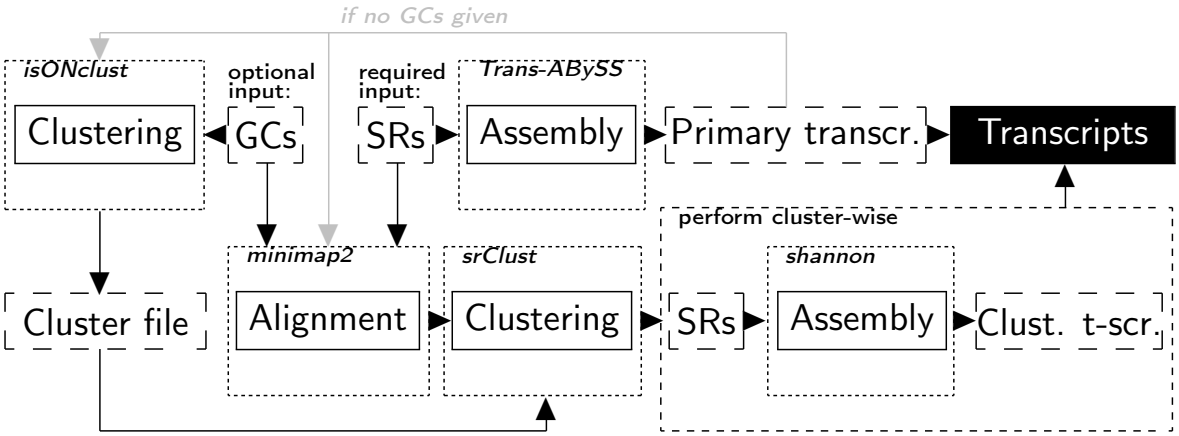

Figure 1: The ClusTrast pipeline. SRs = short reads, GCs = guiding contigs. The only required input data is the set of SRs from an RNA-seq experiment.

2 Implementation

2.1 ClusTrast method

We developed an approach for transcriptome assembly from short reads called ClusTrast.

2.1.1 Overview

Figure 1 shows a flowchart of the ClusTrast pipeline. The only required input to ClusTrast is a file with short RNA-seq reads, referred to as SRs (short reads). Supplementary Figure S.1 illustrates an example of how the method works.

2.1.2 Primary assembly and guiding contigs (GCs)

In ClusTrast, Trans-ABySS (Robertson et al., 2010) is employed to create a "primary assembly" from the short reads. The primary assembly will by default be used as the set of guiding contigs (GCs) in ClusTrast. The guiding contigs are used in the next step to cluster the short reads (2.1.3). Guiding contigs may also be provided separately by the user, and could then serve as primary assembly if desired. The primary assembly is by default also merged into the final set of assembled transcripts (2.1.5), in order to capture isoforms that the cluster-wise assemblies (2.1.4) might fail to assemble due to low coverage.

The original version of Trans-ABySS used several different k-mers and merged the resulting assemblies, in order to get both recall from small values of k and precision from high values of k. Since a single-k run uses much less memory (or is substantially faster) than a multi-k run, we tried both strategies with ClusTrast. In this report, we have appended **-M** to the name of a method if it used a multi-k strategy.

2.1.3 Clustering of (a) Guiding Contigs and (b) Short Reads

(a) The clustering of the guiding contigs is performed with isONclust (Sahlin and Medvedev, 2020), a tool originally developed for clustering of PacBio CCS reads or ONT reads into gene families. It uses a greedy algorithm for the clustering and handles variable error rates by the means of the quality values in the FASTQ input files. When the set of guiding contigs is in FASTA-format, ClusTrast will convert it to FASTQ-format with a static quality.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.02.473666; this version posted November 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table 1: Short read RNA-seq datasets accessed from the NCBI SRA database. RL=read length in bases. Species column, indicated in bold is the name to which the data set is referred to throughout this article. RPs=million read pairs, before pre-processing (on the left) and after pre-processing (on the right).

SRA ID	RL	Species		RPs	
SRR5133163.1	2 × 150	**Human**	*Homo sapiens*	29.51	29.05
SRR8632985	2 × 76	**Mouse**	*Mus musculus*	31	30
SRR11341576	2 × 150	**Rice**	*Oryza sativa*	24	23
SRR11278019	2 × 126	**Arabidopsis**	*Arabidopsis thaliana*	11.2	11.1
SRR10728575	2 × 150	**Zebrafish**	*Danio rerio*	21	20
SRR5986240.1	2 × 150	**Poplar**	*Populus trichocarpa*	25.1	24.4

(b) The short reads are aligned to the guiding contigs with minimap2 (Li, 2018), using the preset option `-x sr`, intended for short read alignment, but included secondary alignments. Secondary alignments can optionally be excluded in ClusTrast. Next, the short reads are assigned to the guiding contig clusters based on the alignment results. If a short read x is aligned to guiding contigs X_1 and X_2, and X_1 belongs to cluster n_1 and X_2 belongs to cluster n_2, x will be included in n_1 and n_2. Thus, the short reads have now been clustered. A read can only occur once per cluster. See also Supplementary Section C.2.

2.1.4 Clusterwise assembly

The cluster-wise assembly in ClusTrast is performed by the transcriptome assembler Shannon (Kannan et al., 2016) (also used in refShannon, a genome-guided transcript assembler (Mao et al., 2020)), and aims to be information theoretically optimal. Kannan et al. claim that Shannon can finish in linear time given (*i*) sufficient diversity of transcript abundance and (*ii*) no loops in the graph, but do not address how it will deal with datasets not meeting these criteria. However, dividing the reads in the short read dataset into clusters before assembly will reduce the complexity of each individual assembly and lower the risk of violating these requirements. Because of this, and its aim to reconstruct as many transcripts as possible, Shannon is used for the cluster-wise assemblies in ClusTrast.

2.1.5 Merging the assemblies

The final step of ClusTrast is to merge the cluster-wise assemblies with the primary assembly by concatenation. Duplicate instances of the reconstructed transcripts are by default removed.

2.2 Datasets and annotations

2.2.1 Short read RNA-seq datasets for assembly generation

We evaluated the *de novo* transcriptome assemblies using the NCBI SRA datasets in Table 1. They were all non-stranded paired-end short read RNA-seq datasets. We pre-processed the datasets with fastp (Chen et al., 2018) with default parameters, which means removal of any remaining adapter sequences, quality pruning (max 40% of the bases were allowed to have base quality < 16, and at most five Ns per read), and exclusion of reads that ended up shorter than 15bp (see the supplementary material for details).

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.02.473666; this version posted November 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table 2: Reference transcriptome sequence, genome sequence and annotation versions accessed from Ensembl (`https://www.ensembl.org/index.html` and `https://plants.ensembl.org/index.html` for non-plant and plant species, respectively), where the Version id suffix shows the Ensembl version. The number of genes and isoforms are counted from the reference transcriptome. Genes and isoforms are considered expressed if TPM>0 as calculated by RSEM on the datasets in Table 1.

Reference		Total		Expressed	
Species	Version id	Genes	Isoforms	Genes	Isoforms
Human	GRCh38.99	40491	190432	22510	102552
Mouse	GRCm38.99	36711	119353	17400	52845
Arabidopsis	TAIR10.48	27655	48359	23085	38004
Rice	IRGSP-1.0.48	37967	44761	29703	34956
Zebrafish	GRCz11.99	30628	57775	23963	35189
Poplar	Pop_tri_v3.46	41335	73012	29400	44652

2.2.2 Reference datasets for assembly evaluation

We downloaded reference genome sequences as well as reference transcript annotations from Ensembl for each of the six species. We used the GTF file annotations of genes and transcripts, not including the abinitio annotations. We estimated the expression of all reference transcripts in each of the six datasets using RSEM (Li and Dewey, 2011) and defined a transcript isoform as expressed if the transcripts per million (TPM) reported by RSEM was greater than zero. Versions and commands for RSEM are listed in the supplementary material. We defined a gene as expressed if at least one of the transcript isoforms associated with that gene was expressed. The versions of the annotations used and the number of genes and isoforms we detected in each dataset are shown in Table 2.

2.3 Transcriptome assembly generation

We assembled the transcriptomes for all six datasets (Table 1) using Trans-ABySS (Robertson et al., 2010), Trinity (Grabherr et al., 2011), Oases (Schulz et al., 2012), SOAP-denovo-Trans (Xie et al., 2014), BinPacker (Liu et al., 2016), Shannon (Kannan et al., 2016), rnaSPAdes (Bushmanova et al., 2019), TransLiG (Liu et al., 2019), RNA-Bloom (Nip et al., 2020), and ClusTrast. We used each assembler's own default parameters. Trans-ABySS and Oases can be run in a "multi-k" mode where the assembler is first run with a single k-mer ("single-k" mode; where a k-mer is a substring, with fixed length k, of a read) for several different k-mers and the resulting assemblies are merged into a single assembly. We used both the single-k and multi-k strategies for these two assemblers. We append **-M** to the name of a method if it uses a multi-k strategy, and **-S** if it uses a single-k strategy. Oases-**M** uses by default all odd k-mers from 19 to 31, but it only finished within less than 58 hours on the mouse and arabidopsis datasets. On the rice dataset, it finished after ∼400 hours. Therefore, for human and zebrafish, we used only Oases-**S** with $k = 31$. The program versions and the executed commands are listed in the supplementary material.

We also generated a concatenated assembly from the Trans-ABySS and Shannon transcriptomes, referred to as TrAB+Sh, to examine if the clustering approach of ClusTrast improves the assembly quality.

2.4 Transcriptome assembly evaluation

We evaluated the transcriptome assemblies by estimating precision (positive predicted value,

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.02.473666; this version posted November 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

PPV) and recall (sensitivity or true positive rate, TPR). For this, we used the reference based transcriptome comparison tools Conditional Reciprocal Best BLAST (CRBB) (Aubry et al., 2014), as implemented in the TransRate package (Smith-Unna et al., 2016), and SQANTI (Tardaguila et al., 2018). Versions and commands for these tools can be found in the supplementary material. We only used reference transcripts that were considered expressed (2.2.2). All assembled transcripts were considered expressed, since they were reconstructed from actual RNA-seq data.

2.4.1 Using SQANTI in evaluation

We used SQANTI (Structural and Quality Annotation of Novel Transcript Isoforms) (Tardaguila et al., 2018) to classify assembled transcripts according to their splice junction matches with reference genes and transcript isoforms. When an assembled transcript is anti-sense to an annotated gene, SQANTI will classify that transcript as anti-sense. We extracted all transcripts classified as anti-sense, reverse-complemented them, and then reclassified them with SQANTI.

When an assembled transcript and a reference isoform have the same number of exons and same splice junctions, then SQANTI classifies it as a full splice match (FSM). When the assembled transcript has fewer exons than the reference but the splice junctions in the assembled transcript all exist in the reference, it is classified as an incomplete splice match (ISM). In order for SQANTI to classify an assembled transcript as an ISM, all junctions in the assembled transcript must match the reference, but the exact start and end can differ. In case there are several possible consistent reference isoforms, SQANTI assigns the assembled transcript to the shortest of the matching references. Assembled transcripts classified by SQANTI as novel in catalog (NIC, when the splice junctions are known but there is a novel combination) and novel not in catalog (NNC, with novel splice junctions) were not classified as true positives.

For recall, we counted all expressed reference isoforms with at least one assembled transcript that SQANTI classified as FSM or ISM (and with a certain fraction, 0.25-1.0, of the exons covered) as a true positive, and divided the total number of true positives by the total number of expressed reference isoforms. For precision, we counted each assembled isoform classified by SQANTI as FSM or ISM and covering at least a certain fraction (from 0.25 to 1.0) of the reference exons as a true positive, and divided the total number of true positives by the total number of assembled isoforms.

2.4.2 Using CRBB in evaluation

We used CRBB (Conditional Reciprocal Best BLAST) (Aubry et al., 2014) to classify assembled transcripts according to their similarity to reference transcripts. To this end, we used TransRate (Smith-Unna et al., 2016), which in turn used BLAST (Altschul et al., 1990) to align each assembled transcript to the set of reference transcripts, and each reference transcript to the set of assembled transcripts. By using all transcripts which are top hits in both BLAST alignments reciprocally, an appropriate E-value cutoff is calculated. Transcripts with lower E-values than this cutoff are then considered CRBB hits. We defined recall as the proportion of reference transcripts that have a CRBB hit covering the reference transcript to at least 25%–100%. We defined precision as the proportion of assembled transcripts that are a CRBB hit covering the reference isoform to at least 25%–100%.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.02.473666; this version posted November 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

3 Results

3.1 Transcriptome assembly evaluation

Transcriptome assemblies for all compared assemblers, including ClusTrast, were generated as described in 2.3. Basic statistics of all assembled transcriptomes are available in Supplementary Table S.2–S.7. We collectively refer to all tested approaches as "assemblers", although assembly pipeline (e.g., ClusTrast) or concatenation (TrAB+Sh) may be more accurate.

3.1.1 Evaluation with SQANTI

We investigated how recall and precision changed when we varied the proportion of exons that an assembled transcript needs to recover in order to be considered a true positive. As this proportion was relaxed for the ISM classifications from 1.0 to 0.25 (for the FSM category it is by definition 1.0), the recall (Figure 2) and precision (Figure 3) increased. ClusTrast-**M** had the highest SQANTI recall of any assembler for all of the six datasets over the entire range (except roughly tied with TrAB+Sh for arabidopsis). The assembler with the highest precision varied across datasets; it was RNA-Bloom in human, ClusTrast-**M** in rice and (with TransLiG) poplar, Oases-**M** in mouse, and TransLiG in arabidopsis and zebrafish.

Fixing the proportion at 0.5 (i.e., at least 50% of exons recovered for ISM), ClusTrast-**M** detected more transcript isoforms than the established assemblers Trinity (1.23–2.15 fold increase), Oases-**S** (1.33–2.59 fold increase), and Trans-ABySS-**M** (1.1–1.52 fold increase) (Supplementary Figure S.2 and Supplementary Table S.14). Precision was comparable to Trinity (0.9–1.76 fold change), Oases-**S** (0.78–1.5 fold change) and Trans-ABySS-**M** (0.73–1.63 fold change).

3.1.2 Evaluation with CRBB

We investigated CRBB recall and precision over the same proportion of required recovered exons as for SQANTI and observed an increase in recall and precision as this proportion was decreased from 1.0 to 0.25. We observed some changes in the relative ordering of assemblers as shown in Figure 4 (CRBB recall) and Figure 5 (CRBB precision). In particular, rnaSPAdes performance levelled off in the lower end.

Fixing the proportion at 0.5, CRBB recall was higher for ClusTrast-**M** than for Trinity (1.01–1.34 fold increase) across all datasets, but not compared to all assemblers (Supplementary Figure S.3 and Supplementary Table S.15). ClusTrast-**M** performed the best on human, mouse, rice, and zebrafish, it ranked second for arabidopsis, while its ranking varied from first to third as the required reference transcript coverage decreased. ClusTrast-**M** clearly underperformed compared with Trinity with regard to CRBB precision (0.33–0.68 fold change); the assembler with highest precision in a dataset was always TransLiG or RNA-Bloom.

3.1.3 SQANTI and CRBB evaluation metrics were correlated; true positive sets still differed

The number of transcripts that were considered as true positives by both SQANTI and CRBB or exclusively by only one of them varied between datasets and between assemblers (Supplementary Tables S.18 and S.19). ClusTrast, Trans-ABySS, Oases-**M**, and (with one exception) Shannon and SOAP-denovo-Trans consistently predicted more transcripts that were considered as true positives exclusively by SQANTI and not by CRBB, while TransLiG was the only assembler that consistently predicted more transcripts that were considered true positives exclusively by CRBB. We used SQANTI categories to classify the ClusTrast true positives that were exclusively detected by either SQANTI or CRBB (Supplementary Tables S.23 and S.24, respectively). We

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.02.473666; this version posted November 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

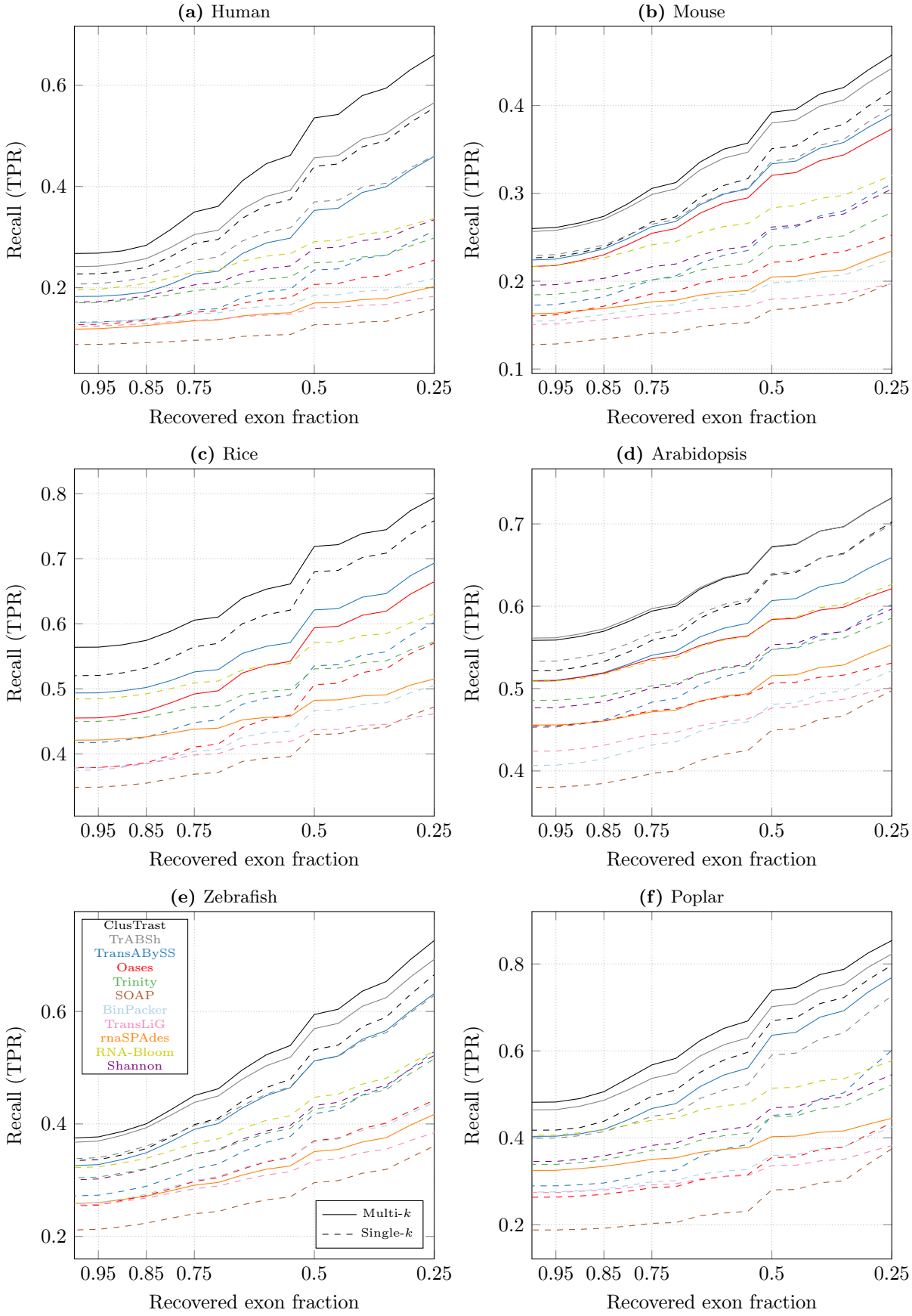

Figure 2: Proportion of reference isoforms with at least one SQANTI classification of FSM or ISM vs. the cumulative proportion of exons recovered by the assembly.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.02.473666; this version posted November 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

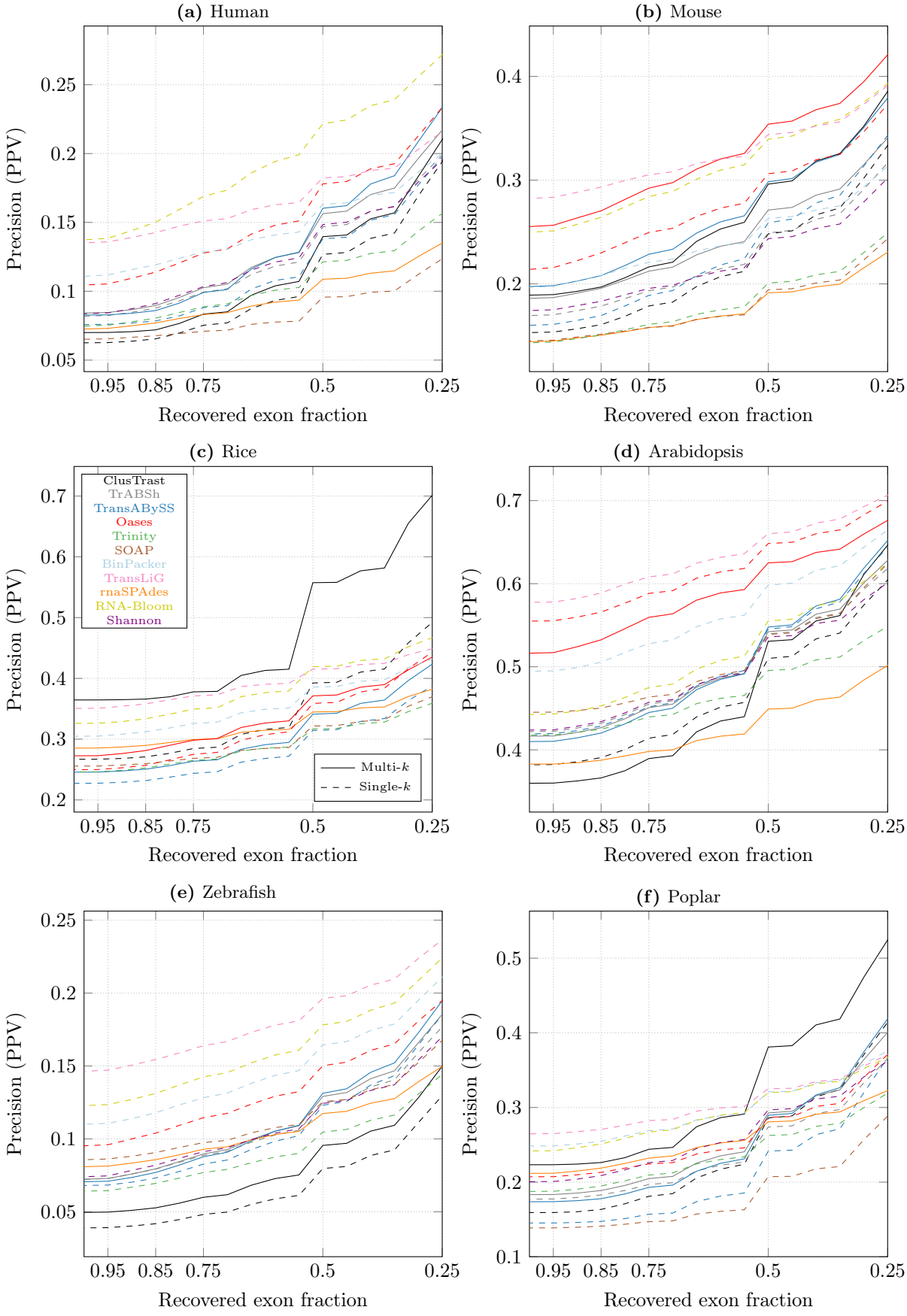

Figure 3: Proportion of reconstructed isoforms classified by SQANTI as FSM or ISM vs. the cumulative proportion of recovered exons from the reference.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.02.473666; this version posted November 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

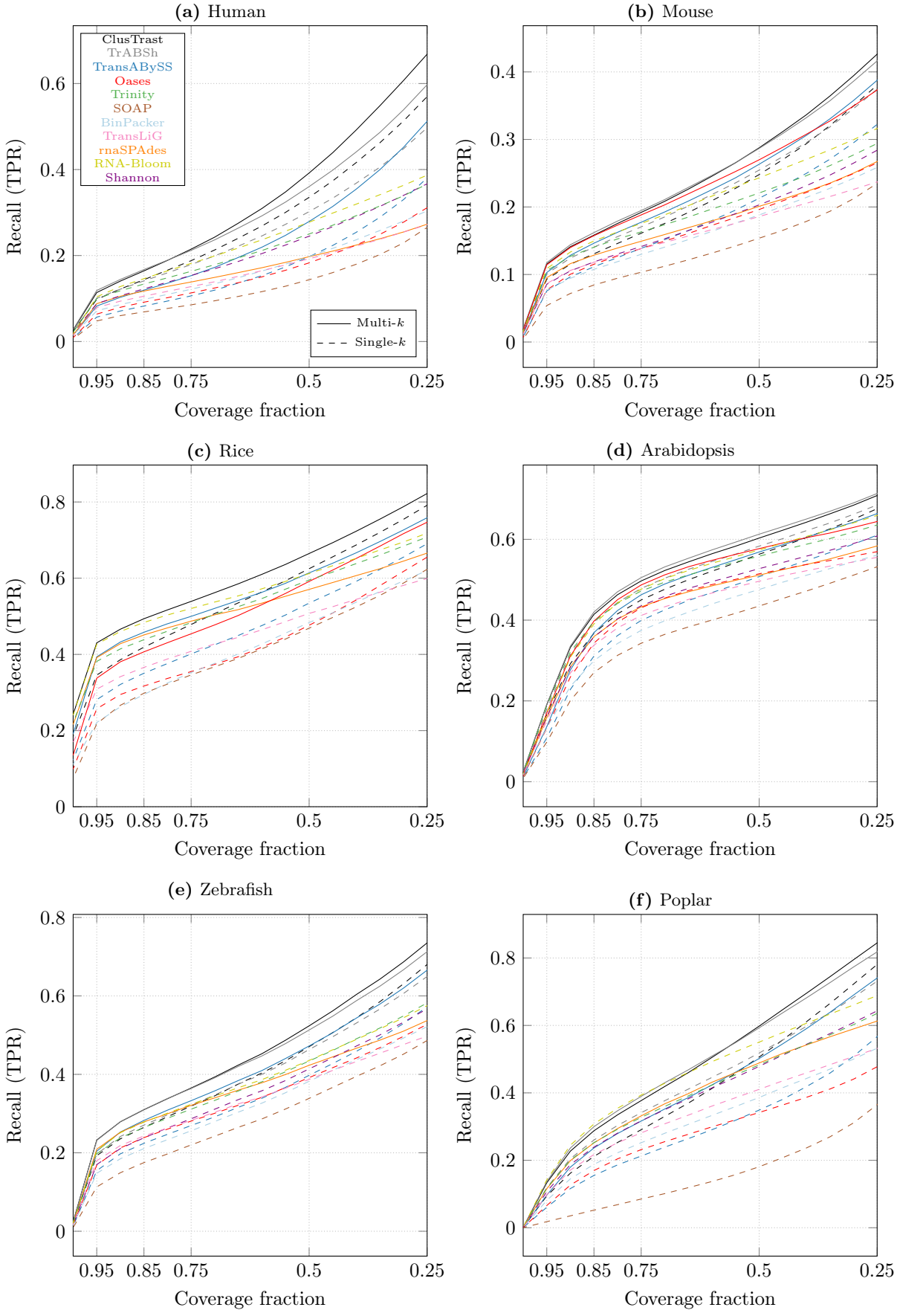

Figure 4: Proportion of references with a CRBB hit vs. the cumulative proportion of recovered reference length.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.02.473666; this version posted November 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

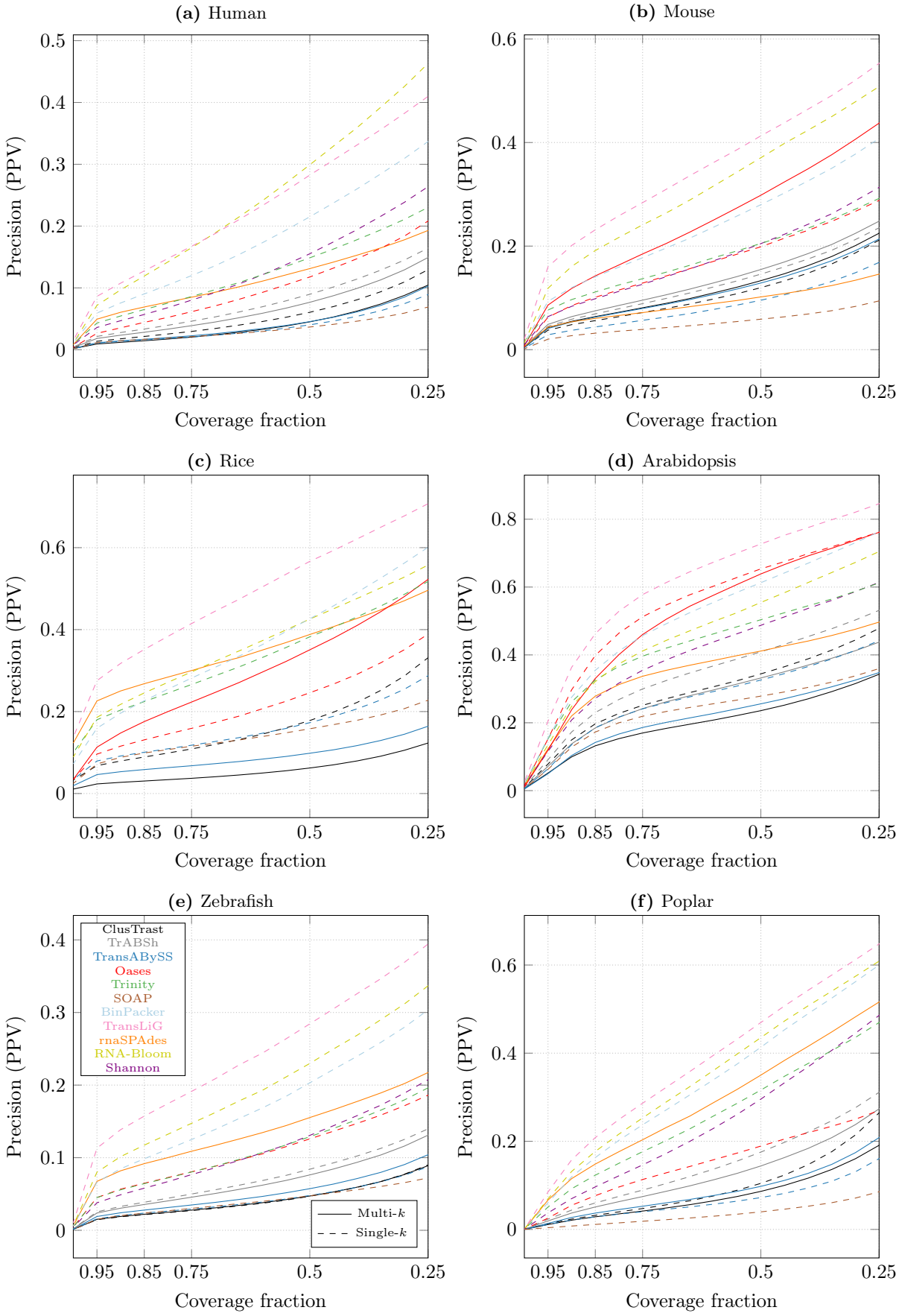

Figure 5: Proportion of reconstructed isoforms with a CRBB hit vs. the cumulative proportion of recovered reference length.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.02.473666; this version posted November 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

observed that the largest category of true positives according to SQANTI but not CRBB was the ISM mono-exon class. The largest category of true positives according to CRBB but not SQANTI was novel not in catalog (NNC) with novel splice sites.

SQANTI and CRBB recall measurements were highly correlated across all assemblies and datasets ($\rho = 0.93$; Supplementary Figure S.4) while SQANTI and CRBB precisions were less correlated ($\rho = 0.75$; Supplementary Figure S.5). We calculated the correlation of precision measurements for each assembler individually: ClusTrast-**M** obtained $\rho = 0.54$ while for all other assemblers $\rho \geq 0.82$. Next, we excluded the ISM mono-exon class from the set of true positives and recalculated the precision correlation for ClusTrast-**M**: it increased to $\rho = 0.94$.

3.1.4 Reference transcripts were often covered to at least 95% by FSMs

We investigated the number of expressed reference transcript isoforms that were reconstructed to at least 50% and 95% of their length by a single FSM according to SQANTI, Supplementary Table S.16. For all assemblies and both length requirements, either TrAB+Sh or ClusTrast reconstructed the most reference transcript isoforms, with small differences (<5%) except for rice where TrAB+Sh did not produce a result. Between 35.1% (arabidopsis) and 68.8% (rice) of the reference transcript isoforms that had an FSM match were reconstructed by the FSM-classified contig from ClusTrast to at least 95% of their length. The corresponding range for reconstruction to at least 50% of the reference transcript length was between 76.7% (human) and 91.0% (arabidopsis).

3.1.5 An appreciable fraction of reference transcripts were reconstructed with polymorphisms by ClusTrast

We used the subset of reference transcripts with FSM or CRBB hits to estimate how often these reference transcripts were reconstructed as polymorphic variants (SNPs, indels) or as alternatively spliced contigs. In Supplementary Tables S.20 and S.21, the sets labeled A contain the FSMs, while the sets labeled B contain the CRBB hits. By definition, FSM contigs corresponding to a specific reference transcript are not alternatively spliced, since they contain all splice junctions of their reference transcript. Two (or more) FSM contigs matching one and the same reference transcript are thus polymorphic variants of each other. This is the $A \setminus B$ and $A \cap B$ sets in Supplementary Tables S.20 and S.21. On the other hand, two (or more) contigs that are not FSMs but considered as CRBB hits to one and the same reference transcript, are potentially splice variants of that reference transcript. This is the $B \setminus A$ sets. We estimated that 58–81% of the reference transcripts reconstructed by ClusTrast were reconstructed with polymorphic variants, Supplementary Table S.22. Conversely, we estimated that 47–78% of ClusTrast assembled contigs contained polymorphic variants, Supplementary Table S.22.

3.1.6 Recall varied over expression levels and number of exons in isoforms

To determine if the assemblers differed in how well they recovered isoforms of genes with more than one annotated isoform, we calculated SQANTI recall of isoforms binned by genes according to the number of isoforms these genes expressed (Figure 6). In most cases the ranking of assemblers by recall did not change with increasing number of expressed isoforms per gene. ClusTrast-**M** came out on top over almost the entire range for 5 out of 6 datasets, although for mouse it was tied with TrAB+Sh and for arabidopsis it was tied with Oases-**M** and TrAB+Sh.

Next, we binned reference transcripts by expression quantiles as measured by RSEM. SQANTI recall increased with increased expression level, for all assemblers and for all data sets except that some assemblers levelled off in the range 80-100%. We observed that recall was higher for

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.02.473666; this version posted November 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

ClusTrast-**M** than all other assemblers in the lower end of expression levels (expression quantile <15%) and across the entire range of expression levels for all datasets except arabidopsis and zebrafish where ClusTrast-**M** was tied with TrAB+Sh (Figure 7).

We observed cases where ClusTrast detected highly expressed isoforms missed by other methods. We illustrate this with examples of genes where the highest expressed isoform (according to RSEM) was reconstructed only by ClusTrast and not by any other method. Supplementary Figures S.23 to S.27 contain Sahimi plots for these example genes, one for each of the six datasets (see Supplementary Section C.4 for more details).

3.2 Run time and memory usage

Across all datasets, all assemblers except Oases-**M** completed within 48 hours and required less than 300 GB of memory (Supplementary Table S.25). ClusTrast-**M** took between 660 and 2145 minutes to complete, the longest time of all assemblers for mouse and arabidopsis. Oases-**M** took the longest time to complete for one dataset, and did not complete for the remaining three. SOAP-denovo-Trans was the fastest assembler for all datasets. ClusTrast-**M** peak memory use was between 57.15 and 267.2 GB for the six datasets, highest of all assemblers for one dataset, while Trinity and Oases-**M** had the highest peak memory use for two datasets each. RNA-Bloom had the lowest peak memory usage. Our computational setup is described in Supplementary Table S.26.

4 Discussion

We described and assessed the *de novo* transcriptome assembler ClusTrast. Across all six tested datasets, ClusTrast-**M** and ClusTrast-**S** created assemblies that had the highest or among the highest recall as measured by SQANTI and CRBB.

In our evaluation of ClusTrast and the other transcriptome assemblers, we have emphasized metrics that do not penalize the reconstruction of different isoforms of a gene. A compact transcriptome assembly is, for instance, preferable when the reconstructed transcripts are intended to be used for aligning reads to a "reference" for, e.g., differential gene expression analysis. In such a situation, our approach would not be helpful. However, when the goal is to find as many supported transcript isoforms as possible, compactness is not in itself desirable and could in fact be counter productive. In a recent review by Thind et al. (2021), the authors point out that there is a need for metrics that better capture the performance with regards to transcript isoforms. Our use of SQANTI's approach to evaluation is an attempt to address this. SQANTI (Tardaguila et al., 2018) was originally designed to evaluate long reads (e.g., PacBio CCS) but the categories it defined and its aim to classify each non-redundant transcript individually are useful also in evaluation of assembled transcripts. We have been conservative in that we included only the FSM and ISM categories as true positives. Also the NIC category, which encompasses reconstructed transcripts that contain annotated reference exons only, but in novel combinations, could have been included. We compared the results from SQANTI and CRBB (which has been used for transcriptome assembly benchmarking before), and detected a correlation between SQANTI and CRBB scores, particularly strong for recall. We noted that most of the contigs where SQANTI and CRBB classifications of true positives disagreed, belong to the SQANTI class ISM mono-exon. Excluding these from the set of true positives increased the correlation between SQANTI and CRBB precision measurements for ClusTrast (Section 3.1.3). We believe this supports the notion that SQANTI is possible to use for transcriptome assembly evaluation.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.02.473666; this version posted November 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

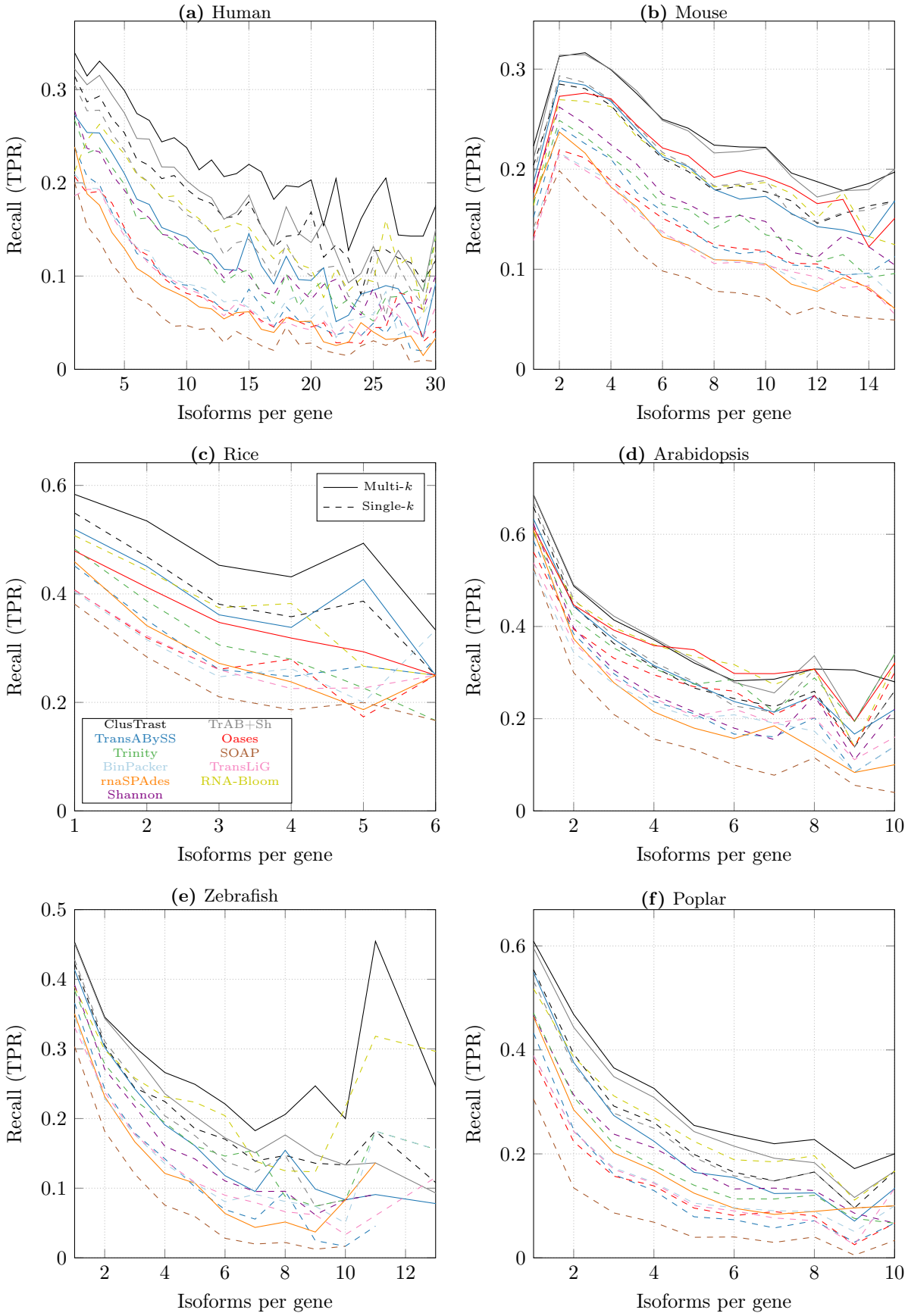

Figure 6: SQANTI recall of reference isoforms (FSM) binned by number of expressed isoforms per gene.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.02.473666; this version posted November 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

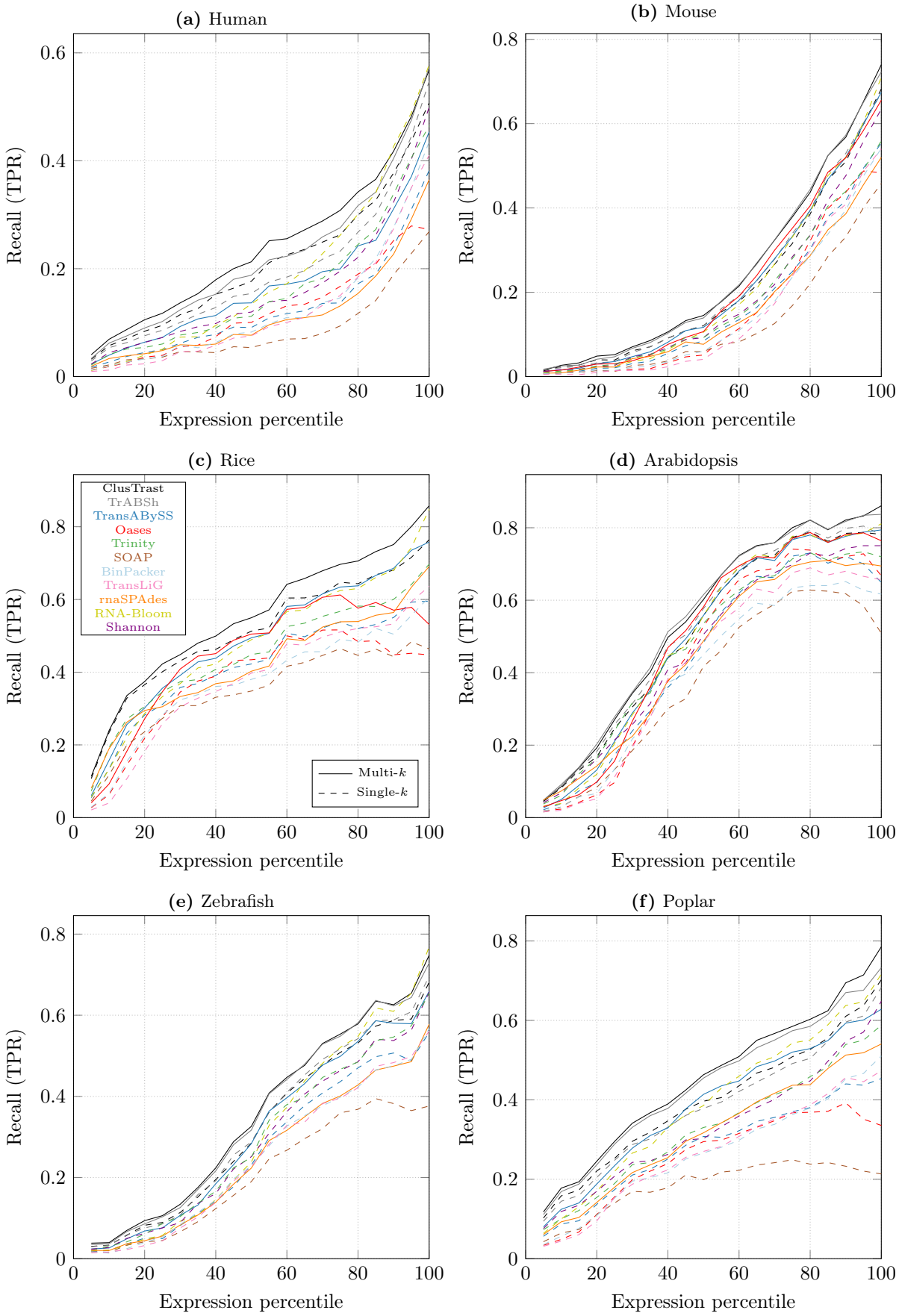

Figure 7: SQANTI recall of expressed transcript isoforms stratified according to RSEM expression. Recall within each bin (5 percentiles) is defined as the proportion of transcript isoforms that have an FSM match to an assembled contig that is $\geq$ 200bp.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.02.473666; this version posted November 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Comparing ClusTrast-**M** to one of the most popular transcriptome assemblers, Trinity, revealed that ClusTrast-**M** detected more transcript isoforms than Trinity, and also had a higher precision for isoforms as measured by SQANTI, but clearly underperformed according to CRBB precision. The difference in relative performance of ClusTrast and Trinity according to CRBB and SQANTI precision may be explained by how CRBB handles assembled transcripts with high similarity: If two or more highly similar transcripts exist in the assembly, and some of them have a lower E-value than others, then only the transcripts with E-values below the limit will be considered CRBB hits and thus true positives for precision. SQANTI, in contrast, annotates each transcript independently and therefore calls all assembled transcripts that are similar to a reference transcript as a true positive. We assessed this by recalculating SQANTI precision while only counting one transcript match for every reference transcript (Supplementary Figure S.6), and we observed a marked reduction in precision of Trans-ABySS-**M** and ClusTrast across all assemblies (compare Supplementary Figures S.2 and S.6). We also tested ClusTrast with secondary alignments switched off, and observed a slight improvement for CRBB precision, but at the cost of a reduction in both CRBB and SQANTI recall for most datasets (Supplementary Table S.27).

We observed that ClusTrast generally recovered as many or more known isoforms as TrAB+Sh as measured by SQANTI (Supplementary Figure S.2) while suffering only a small reduction in CRBB precision (Supplementary Figure S.3) and that ClusTrast finished successfully on the rice dataset, where Shannon (and thus TrAB+Sh) failed to create an assembly. A possible explanation for both observations is that the clustering performed in ClusTrast may simplify sub-graphs enough to allow better handling by the Shannon heuristic (Section 2.1.4) and thereby increase sensitivity. ClusTrast-**M** and TrAB-**M**+Sh were also the best in reconstructing isoforms to their full length according to SQANTI (Section 3.1.4).

In general the performance of the assemblers was rather consistent over species, regardless of evaluation approach (SQANTI or CRBB). We tested three additional human datasets (one of which simulated; Supplementary Table S.1), for a total of four, and observed that ClusTrast showed the highest SQANTI and CRBB recall over the range of transcript coverage as well as expression levels for all four human datasets (Supplementary Figures S.7–S.9,S.17). However, the precision performance of ClusTrast was mixed. ClusTrast performance was not correlated to the size of the datasets (Supplementary Figure S.19). For all results on all additional datasets, see Supplementary Section C.3.

We used RSEM to estimate the number of expressed isoforms, and in our recall calculations we used only transcripts with TPM>0. Using all reference transcripts in our evaluation, instead of only those that have a TPM>0, would mean a larger denominator when calculating recall, which would lower the recall of all compared assemblers alike. If RSEM makes a mistake and assigns TPM=0 to a reference transcript that in fact is transcribed, then the recall will be underestimated. If RSEM assigns a TPM>0 to a reference transcript which in fact is not transcribed, recall will be overestimated. Similarly, there might be assembled contigs that correspond to real isoforms that are not present in the reference. These contigs are counted as false negatives while they should be true positives, thus precision performance is likely underestimated. For the SQANTI evaluation, it is possible that many of these would be considered as true positives if we had included the NIC category among the true positives.

5 Conclusion

In our tests of model organisms, ClusTrast consistently detected the most transcript isoforms, not the least for the isoforms in the lower end of the expression range (Section 3.1.6, but at a cost of lower precision. This agrees with our intention of ClusTrast – to provide a comprehensive but

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.02.473666; this version posted November 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

non-redundant list of contigs. Therefore, we believe researchers interested in a more complete representation of transcript isoforms from eukaryotic organisms may wish to use ClusTrast. The resulting list of contigs is amenable for further processing and analysis tailored according to the research question at hand.

Availability and requirements

Project name: ClusTrast

Project home page: `https://github.com/karljohanw/clustrast`

Operating systems: Linux and MacOS

Programming language: Bashscript

Requirements: `transabyss`, `shannon_cpp`, `isONclust`, `minimap2`, `awk`.

License: GPLv3

Restrictions to use by non-academics: None

Funding

This work was supported by FORMAS [2013-650] and the Swedish Research Council [2018-05973]. Computations were enabled by resources, ParallellDatorCentrum (`PDC`) at KTH Royal Institute of Technology, provided by the Swedish National Infrastructure for Computing (SNIC), partially funded by the Swedish Research Council [2018-05973].

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

KJW implemented the method, ran all the experiments and wrote the original version of the manuscript. KJW and OE analyzed the results. OE and WWK supervised the project, and contributed to writing the manuscript. All authors read and approved the final version of the manuscript.

Acknowledgements

We also wish to thank Pelin Akan Sahlén at KTH for sharing access to the server `SAGA`.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.02.473666; this version posted November 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

References

S. Akhter, W. W. Kretzschmar, V. Nordal, N. Delhomme, N. Street, O. Nilsson, O. Emanuelsson, and J. F. Sundström. Integrative analysis of three RNA sequencing methods identifies mutually exclusive exons of MADS-box isoforms during early bud development in Picea abies. *Frontiers in Plant Science*, 9:1625, 2018. ISSN 1664-462X. doi: 10.3389/fpls.2018.01625.

S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic Local Alignment Search Tool. *Journal of Molecular Biology*, 215(3):403–410, 1990. ISSN 0022-2836. doi: https://doi.org/10.1016/S0022-2836(05)80360-2. URL `https://www.sciencedirect.com/science/article/pii/S0022283605803602`.

S. Aubry, S. Kelly, B. M. C. Kümpers, R. D. Smith-Unna, and J. M. Hibberd. Deep Evolutionary Comparison of Gene Expression Identifies Parallel Recruitment of Trans-Factors in Two Independent Origins of C4 Photosynthesis. *PLOS Genetics*, 10(6):1–16, 06 2014. doi: 10.1371/journal.pgen.1004365. URL `https://doi.org/10.1371/journal.pgen.1004365`.

N. L. Barbosa-Morais, M. Irimia, Q. Pan, H. Y. Xiong, S. Gueroussov, L. J. Lee, V. Slobodeniuc, C. Kutter, S. Watt, R. Colak, T. Kim, C. M. Misquitta-Ali, M. D. Wilson, P. M. Kim, D. T. Odom, B. J. Frey, and B. J. Blencowe. The Evolutionary Landscape of Alternative Splicing in Vertebrate Species. *Science*, 338(6114):1587–1593, Dec. 2012. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1230612.

E. Bushmanova, D. Antipov, A. Lapidus, and A. D. Prjibelski. rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. *GigaScience*, 8(9), 09 2019. ISSN 2047-217X. doi: 10.1093/gigascience/giz100. URL `https://doi.org/10.1093/gigascience/giz100`. giz100.

S. Chen, Y. Zhou, Y. Chen, and J. Gu. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 34(17):i884–i890, 09 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty560. URL `https://doi.org/10.1093/bioinformatics/bty560`.

J. D. Fackenthal and L. A. Godley. Aberrant RNA splicing and its functional consequences in cancer cells. *Disease Models & Mechanisms*, 1(1):37–42, 2008. ISSN 1754-8403. doi: 10.1242/dmm.000331. URL `https://dmm.biologists.org/content/1/1/37`.

S. N. Floor and J. A. Doudna. Tunable protein synthesis by transcript isoforms in human cells. *eLife*, 5:e10921, Jan. 2016. ISSN 2050-084X. doi: 10.7554/eLife.10921.

M. Garber, M. G. Grabherr, M. Guttman, and C. Trapnell. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature Methods*, 8(6):469–477, Jun 2011. ISSN 1548-7105. doi: 10.1038/nmeth.1613. URL `https://doi.org/10.1038/nmeth.1613`.

M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, and A. Regev. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.*, 29(7):644–52, May 2011. ISSN 1546-1696. doi: 10.1038/nbt.1883. URL `http://www.ncbi.nlm.nih.gov/pubmed/21572440` `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3571712`.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.02.473666; this version posted November 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

K. E. Hayer, A. Pizarro, N. F. Lahens, J. B. Hogenesch, and G. R. Grant. Benchmark analysis of algorithms for determining and quantifying full-length mRNA splice forms from RNA-seq data. *Bioinformatics*, page btv488, Sept. 2015. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btv488.

M. Hölzer and M. Marz. De novo transcriptome assembly: A comprehensive cross-species comparison of short-read RNA-Seq assemblers. *GigaScience*, 8(5), 05 2019. ISSN 2047-217X. doi: 10.1093/gigascience/giz039. URL `https://doi.org/10.1093/gigascience/giz039`. giz039.

S. Kannan, J. Hui, K. Mazooji, L. Pachter, and D. Tse. Shannon: An Information-Optimal de Novo RNA-Seq Assembler. Preprint at *bioRxiv*, 2016. URL `https://www.biorxiv.org/content/early/2016/02/09/039230`.

B. Li and C. N. Dewey. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12(1):323, Aug 2011. ISSN 1471-2105. doi: 10.1186/1471-2105-12-323. URL `https://doi.org/10.1186/1471-2105-12-323`.

H. Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18): 3094–3100, 05 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty191. URL `https://doi.org/10.1093/bioinformatics/bty191`.

J. Liu, G. Li, Z. Chang, T. Yu, B. Liu, R. McMullen, P. Chen, and X. Huang. BinPacker: Packing-based de novo transcriptome assembly from RNA-seq data. *PLoS computational biology*, 12(2):e1004772–e1004772, Feb 2016. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1004772. URL `https://www.ncbi.nlm.nih.gov/pubmed/26894997`.

J. Liu, T. Yu, Z. Mu, and G. Li. TransLiG: a de novo transcriptome assembler that uses line graph iteration. *Genome Biology*, 20(1):81, 2019. ISSN 1474-760X. doi: 10.1186/s13059-019-1690-7. URL `https://doi.org/10.1186/s13059-019-1690-7`.

S. Mao, L. Pachter, D. Tse, and S. Kannan. RefShannon: A genome-guided transcriptome assembler using sparse flow decomposition. *PLOS ONE*, 15(6):1–14, 06 2020. doi: 10.1371/journal.pone.0232946. URL `https://doi.org/10.1371/journal.pone.0232946`.

K. M. Nip, R. Chiu, C. Yang, J. Chu, H. Mohamadi, R. L. Warren, and I. Birol. RNA-Bloom enables reference-free and reference-guided sequence assembly for single-cell transcriptomes. *Genome Research*, 30(8):1191–1200, 2020. doi: 10.1101/gr.260174.119. URL `http://genome.cshlp.org/content/30/8/1191.abstract`.

G. Robertson, J. Schein, R. Chiu, R. Corbett, M. Field, S. D. Jackman, K. Mungall, S. Lee, H. M. Okada, J. Q. Qian, M. Griffith, A. Raymond, N. Thiessen, T. Cezard, Y. S. Butterfield, R. Newsome, S. K. Chan, R. She, R. Varhol, B. Kamoh, A.-L. Prabhu, A. Tam, Y. Zhao, R. A. Moore, M. Hirst, M. A. Marra, S. J. M. Jones, P. A. Hoodless, and I. Birol. De novo assembly and analysis of RNA-seq data. *Nature Methods*, 7:909–912, Oct 2010. URL `https://doi.org/10.1038/nmeth.1517`.

K. Sahlin and P. Medvedev. De novo clustering of long-read transcriptome data using a greedy, quality value-based algorithm. *Journal of Computational Biology*, 27(4):472–484, 2020. doi: 10.1089/cmb.2019.0299. URL `https://doi.org/10.1089/cmb.2019.0299`. PMID: 32181688.

M. H. Schulz, D. R. Zerbino, M. Vingron, and E. Birney. Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*, 28(8):1086–1092, 2012. ISSN 13674803. doi: 10.1093/bioinformatics/bts094.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.02.473666; this version posted November 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

R. Smith-Unna, C. Boursnell, R. Patro, J. M. Hibberd, and S. Kelly. TransRate: reference-free quality assessment of de novo transcriptome assemblies. *Genome Research*, 26(8):1134–1144, 2016. doi: 10.1101/gr.196469.115. URL http://genome.cshlp.org/content/26/8/1134.abstract.

T. Sterne-Weiler and J. R. Sanford. Exon identity crisis: Disease-causing mutations that disrupt the splicing code. *Genome Biology*, 15(1):201, 2014. ISSN 1465-6906. doi: 10.1186/gb4150.

M. Tardaguila, L. de la Fuente, C. Marti, C. Pereira, F. J. Pardo-Palacios, H. del Risco, M. Ferrell, M. Mellado, M. Macchietto, K. Verheggen, M. Edelmann, I. Ezkurdia, J. Vazquez, M. Tress, A. Mortazavi, L. Martens, S. Rodriguez-Navarro, V. Moreno, and A. Conesa. SQANTI: extensive characterization of long read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Research*, 28(1):1–16, 2 2018. doi: 10.1101/gr.222976.117. URL http://genome.cshlp.org/content/early/2018/02/09/gr.222976.117.abstract.

A. S. Thind, I. Monga, P. K. Thakur, P. Kumari, K. Dindhoria, M. Krzak, M. Ranson, and B. Ashford. Demystifying emerging bulk RNA-Seq applications: the application and utility of bioinformatic methodology. *Briefings in Bioinformatics*, 22(6), 07 2021. ISSN 1477-4054. doi: 10.1093/bib/bbab259. URL https://doi.org/10.1093/bib/bbab259. bbab259.

E. T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth, and C. B. Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476, Nov. 2008. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature07509.

Y. Xie, G. Wu, J. Tang, R. Luo, J. Patterson, S. Liu, W. Huang, G. He, S. Gu, S. Li, X. Zhou, T.-W. Lam, Y. Li, X. Xu, G. K.-S. Wong, and J. Wang. SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics*, 30 (12):1660–1666, 02 2014. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu077. URL https://doi.org/10.1093/bioinformatics/btu077.

H. Y. Xiong, B. Alipanahi, L. J. Lee, H. Bretschneider, D. Merico, R. K. C. Yuen, Y. Hua, S. Gueroussov, H. S. Najafabadi, T. R. Hughes, Q. Morris, Y. Barash, A. R. Krainer, N. Jojic, S. W. Scherer, B. J. Blencowe, and B. J. Frey. The human splicing code reveals new insights into the genetic determinants of disease. *Science*, 347(6218): 1254806, Jan. 2015. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1254806. URL https://www.science.org/doi/abs/10.1126/science.1254806.