Assembly-free rapid differential gene expression analysis in non-model organisms using DNA-protein alignment

Anish M.S. Shrestha[1,2], Joyce Emlyn B. Guiao[1,3], Kyle Christian R. Santiago[1,2]

[1] Bioinformatics Lab, Advanced Research Institute for Informatics, Computing, and Networking (AdRIC), De La Salle University-Manila, Philippines
[2] Software Technology Department, College of Computer Studies, De La Salle University-Manila, Philippines
[3] Mathematics and Statistics Department, College of Science, De La Salle University-Manila, Philippines

Abstract

RNA-seq is being increasingly adopted for gene expression studies in a panoply of non-model organisms, with applications spanning the fields of agriculture, aquaculture, ecology, and environment. Conventional differential expression analysis for organisms without reference sequences requires performing computationally expensive and error-prone de-novo transcriptome assembly, followed by homology search against a high-confidence protein database for functional annotation. We propose a shortcut, where we obtain counts for differential expression analysis by directly aligning RNA-seq reads to the protein database. Through experiments on simulated and real data, we show drastic reductions in run-time and memory usage, with no loss in accuracy. A Snakemake implementation of our workflow is available at: `https://bitbucket.org/project_samar/samar`

Background

RNA-seq has become the principal technique for measuring variation of genome-wide gene expression levels across conditions [1]. Differential expression analysis usually begins by mapping RNA-seq reads to either a reference genome or transcriptome sequence. On one hand, accurate genome annotation has not kept up with the increase in sequence data [2]. Consequently, well-annotated and high-quality reference sequences are available for only a handful of model organisms. On the other hand, driven by declining costs, RNA-seq is becoming increasingly accessible to labs with modest resources; and as a result, it is being employed on an ever-expanding catalog of non-model organisms, pervading the fields of agriculture, aquaculture, ecology, and environment. A very short list of recent studies include: environmental stress response in sea-trout [3], coral [4], ryegrass [5], pigeonpea [6], tiger barb [7]; immune response to parasites and pathogens in guppy [8], eel [9], silkworm [10], peanut [11], sunflower [12]; mechanisms of phenotypic divergence in hares [13], bats [14], grass carps [15]; effect of diet in the growth and development in shrimp [16], yellow perch [17], mandarin fish [18], grenadier anchovy [19], catfish [20], tilapia [21], bass [22]. It is only likely that RNA-seq will continue to rapidly proliferate while high-quality reference databases grow at a slow pace.

The conventional strategy to adapt standard reference-based RNA-seq analysis workflows to the case of non-model organisms, has been to first compute a de-novo transcriptome assembly by pooling all reads and to annotate the assembly against a high-confidence protein database. Since assemblers typically do not provide a read-to-contig mapping, the subsequent step is

to map the reads to the assembly. This is followed by a quantification step in which reads mapping to each contig are counted. The count data is used for statistical testing of differential expression. Since the number of differentially expressed genes tends to be quite large, inference of biological function is done computationally by using the annotations to perform GO-term enrichment or pathway analysis.

A major drawback of de-novo assembly is that it requires massive computational resources. In most cases, the goal is to characterize the expression profile of the protein-coding fraction of the transcriptome, and not necessarily to obtain an assembly. Accordingly, samples are sequenced at a much lower depth than would be required for a reliable assembly [25]. Assembly errors such as over-extension, fragmentation, and incompleteness of contigs can adversely impact downstream expression analysis [26, 27]. Furthermore, assemblers tend to over-estimate the number of isoforms/contigs per gene, introducing complications for statistical test of differential expression as well as interpretation of results since many genes could appear multiple times in the final result [28]. In special cases such as comparison of gene expression across species, it might not even be reasonable to compute a single assembly.

We provide an alternative strategy that circumvents the need for assembly and annotation. The first step of our proposed pipeline uses LAST [29, 30] to directly align RNA-seq reads to the high-confidence protein set which would otherwise have been used for annotation. This is followed by a simple counting step that employs a traditional rescue strategy to resolve multimaps [34]. The counts can be fed into standard count-based differential gene expression analysis tools, e.g. DESeq2 [35]. Our main proposition here is that since functional analyses in non-model organisms rely on a database of homologous proteins in order to draw conclusions, it might be more reasonable to directly allocate reads to those homologous proteins using DNA-protein alignment, instead of introducing an error-prone yet computationally heavy intermediary step of assembly.

The main and obvious advantage of our method is that it drastically brings down computational costs. For example, for a typical RNA-seq containing 2 groups of 3 replicates each and 20 million paired-end reads per replicate, our approach takes under half an hour, whereas computing an assembly would take several tens of hours. Additionally we show, through experiments on simulated and real RNA-seq datasets, that our method is more accurate in identifying differentially expressed genes than an assembly-mapping-quantification pipeline. Another advantage is that it is easier to interpret results, as each homologous gene is reported as differentially expressed or not, along with associated statistical measures. In contrast, with assembly-based pipelines, there might be a need to consolidate results across several fragmented contigs. Furthermore, reference proteomes, for example in UniProt, come with GO annotations, allowing for a straightforward transition to downstream functional analysis; whereas with assembly-based pipelines, there is a need to post-process the multiple local alignments that might be reported by the annotation software for each contig.

Our pipeline is publicly available at `https://bitbucket.org/project_samar/samar`.

Results

DNA-protein alignment attains similar performance to using a transcriptome reference

We first demonstrate, under ideal conditions, the soundness of our idea of aligning RNA-seq reads to a proteome reference for differential gene expression analysis, by comparing our performance to that of the traditional case of using an established transcriptome reference.

For this we simulated an RNA-seq dataset from the fruit fly *D. melanogaster* protein-coding transcriptome. We chose this extensively studied transcriptome since the transcripts and protein products associated with each gene is known for a huge number of protein-coding *D.*

melanogaster genes; and such a mapping between Flybase ID and UniProt ID can be obtained from Ensembl, UniProt, Flybase [37], etc. The data generation process is detailed in the Methods section.

We aligned the simulated reads to the UniProt *D. melanogaster* proteome UP000000803, which contains 1 representative protein sequence per gene, and fed the counts obtained by our method to DESeq2 [35] for differential analysis. To serve as baseline for comparison, we additionally ran a typical pipeline consisting of Bowtie2 [40] for read alignment to the *D. melanogaster* transcriptome, followed by RSEM for transcript quantification, tximport [41] for gene-level aggregation of counts, and finally DESeq2 [35] for differential analysis at the gene level. Details of the two pipelines are provided in the Supplementary Material.

We evaluated the two approaches based on their recall and precision in predicting differentially expressed (DE) genes. Recall is the proportion of actual DE genes that were correctly predicted to be DE, and precision is the proportion of predicted DE genes that were actually DE. We require that the direction of fold-change (up/down) match between the ground truth and prediction to be classified as a correct prediction. To compute recall and precision of our method, we mapped our predictions from the set of proteins to the set of genes using the mapping between UniProt protein ID and FlyBase gene ID obtained from Ensembl.

Figure 1(left) shows the Precision-Recall curves obtained by varying the false discovery rate (FDR) threshold of DESeq2, and Figure 1(right) shows the distribution of the estimated log fold-change at an FDR threshold of 0.1. There is almost no difference between the two approaches in the ability to detect DE genes and in the trend of under- or over-estimation of true fold change. This result demonstrates that for differential gene expression analysis, there is no performance degradation when aligning RNA-seq reads to a proteome reference, even though we are effectively only using reads from the coding region of transcripts.

In the absence of a close reference, our method outperforms assembly-based approach

Next we simulated the scenario faced in the case of non-model organisms by pretending that the *D. melanogaster* reference sequences are not available, and that the closest species with a well-annotated reference proteome is a distant relative, the mosquito *Anopheles gambiae*. The two species are of the same order Diptera with their lineages thought to have separated roughly 250 million years ago [43]. The evaluation process described below uses the mosquito proteome as reference; but to calibrate the effect of the degree of evolutionary and sequence divergence, we repeated the process described below with reference proteomes of closer relatives of *D. melanogaster*: *D. ananassae* and *D. grimshawi*.

We applied our method to the same simulated RNA-seq data as before, this time using the *A. gambiae* proteome as reference. We compared our performance to that of a typical assembly-based pipeline consisting of: Trinity [44] for de-novo transcriptome assembly, followed by Bowtie2 for mapping the reads to the assembly, RSEM for counting, tximport for gene-level aggregation using the gene-to-transcript mapping provided by Trinity, and finally DESeq2 for differential analysis. We used the Dammit pipeline [45] to annotate the assembly against the mosquito proteome. Details of the two pipelines are provided in the Supplementary Material.

To facilitate the comparison, we obtained a pre-computed orthology map between *A. gambiae* and *D. melanogaster*, from the website of InParanoid [46]. Consider a *D. melanogaster* protein-coding gene g, and let F_g be the set of protein products of g. For a *D. melanogaster* protein f, let O_f be the set of mosquito proteins in the same ortholog group as f. We associate with g the set M_g of mosquito proteins m such that $m \in O_f$ for some $f \in F_g$.

We computed recall and precision of our method as follows. An actual up-regulated (down-regulated) *D. melanogaster* DE gene g was defined as correctly predicted if there was at least one protein in M_g that was predicted to be up-regulated (down-regulated). Recall was defined as the number of correctly predicted DE genes divided by the number of actual DE genes.

Precision was defined as the number of correctly predicted DE genes divided by the number of genes g for which at least one protein in M_g was predicted to be DE.

We computed recall and precision of the assembly-based approach as follows. For a Trinity gene t, let D_t be the set of mosquito proteins that Dammit assigned to the isoforms of t (if there were multiple alignments for an isoform, we kept only one with the lowest E-value). With an actual *D. melanogaster* gene g, we associated a set T_g of Trinity genes, where $t \in T_g$ if $D_t \cap M_g \neq \emptyset$. An actual up-regulated (down-regulated) DE *D. melanogaster* gene g was defined to be correctly predicted if there was at least one up-regulated (down-regulated) Trinity gene in T_g. Recall was defined as the number of correctly predicted DE genes divided by the number of actual DE genes. Precision was defined as the number of correctly predicted genes divided by the number of genes g for which at least one gene in T_g was predicted to be DE.

The definitions of recall and precision are necessarily slightly different for the two approaches. Our hope is that they convey a similar meaning – that an actual *D. melanogaster* DE gene is represented by a set of orthologous mosquito proteins (in our method) or by a set of Trinity genes for which there was an annotation to an orthologous mosquito protein (in the assembly-based method), and that the gene is considered to be correctly predicted if at least one of the representatives are predicted to be DE.

Figure 2 shows the precision and recall of our method and the assembly-based approach when using the mosquito reference proteome. It also contains the PR-curves when using the *D. ananassae* and *D. grimshawi* reference proteomes. The curves were obtained by varying the FDR threshold of DESeq2. When using the two *Drosophila* reference proteomes, the performance of our method varied slightly, but in both cases, outperformed the assembly-based approach. When using the *A. agambiae* reference, recall was lower for both methods, mainly because the orthology map contains only 60% of the fruit fly proteins – there were 7341 ortholog clusters involving 7863 fruit fly proteins and 8090 mosquito proteins. Overall, across any setting of FDR threshold or any choice of a reference proteome, our approach outperformed the assembly-based approach.

So far, to compute recall and precision of the assembly-based approach, we used all the alignments reported by the Dammit pipeline, even including many short local alignments. It is not uncommon in practice to filter short alignments. We repeated the analysis by keeping only those alignments predicted by the Dammit pipeline that covered at least 50% of a contig. The precision-recall curves for this cases is shown in Figure 3, which shows a significant drop in recall of the assembly-based approach.

With real data too, our method outperforms the assembly-based approach

We applied our pipeline and the assembly-based approach to a recently published real RNA-seq dataset ArrayExpress E-MTAB-8090 ERR3393437–42. The dataset contains RNA-seq reads of the hemocyte tissue of *D. melanogaster* samples with and without injury, with 3 replicates for each condition. After cleaning and trimming low-quality reads using fastp [47], there were roughly 110 million pairs of reads.

Continuing with the assumption that no reference sequences are available for *D. melanogaster*, we applied our pipeline and the assembly-based pipeline as in the previous section using the mosquito and two *Drosophila* reference proteomes. Since we do not know the ground truth for this dataset, to serve as baseline, we additionally ran the Bowtie2-RSEM-DESeq2 pipeline using *D. melanogaster* reference transcriptome. To be able to compare the DE call sets, we mapped our predicted DE genes to the *D. melanogaster* gene names using the same Inparanoid orthology maps as before.

At FDR threshold of 0.01, there were 104 genes identified as DE by the baseline method. Based on the observation from Figure 1 that the Bowtie2-RSEM-DESeq2 pipeline has high precision at FDR 0.01, let us assume that all of these baseline calls are correct and that they constitute the empirical ground truth. At the same FDR threshold, when using the *D. ananassae*

reference proteome, our method was slightly more sensitive than the assembly-based approach, being able to predict 68 out of the 104 baseline DE genes, compared to 58 by the assembly-based approach. Our method was also slightly more precise, with the 68 calls corresponding to roughly 78% of the calls, compared to 74% for the assembly-based method. This is in line with observation from Figure 2 that our method has slightly better sensitivity and precision than the assembly-based approach. Similar results were obtained when using the *D. grimshawi* reference proteome.

When using the *A. gambiae* proteome, there is a significant decrease in the size of the overlaps between the baseline and the two approaches, consistent with the drop in sensitivity observed in Figure 2. The two approaches are similarly sensitive (25 calls by our approach vs. 27 by assembly-based) while our method is more precise (25 out of 34 calls by our approach vs. 27 out of 46 calls by assembly-based).

Avoiding assembly dramatically reduces running time and memory usage

All the experiments above were carried out on a system with Intel Xeon Silver 4114 Processor with 10 cores and 20 threads. For the real dataset E-MTAB-8090 which contains roughly 110 million pairs of cleaned reads, de-novo assembly alone took more than 24 hours. In contrast, DNA-protein alignment, which is the most compute-intensive part of our pipeline, takes less than 20 minutes per sample containing roughly 20 million pairs of reads, using 20 threads. While the de-novo assembly had a massive peak memory usage of $\sim$65 Gbytes, the memory requirement of our method is dominated by the size of the proteome index, which was just $\sim$33 Mbytes for the mosquito proteome. In general, the index size is roughly $5 \times n$ bytes, where n is the length of the proteome.

Discussion

Summary of results

We have shown that aligning RNA-seq reads to a proteome reference followed by a simple counting procedure provides an extremely fast and light-weight alternative to the current resource-intensive assembly-and-annotation based approach for differential gene expression analysis. We have shown through experiments on simulated and real datasets that our approach is more sensitive and precise than the assembly-based approach.

Isoform-level quantification

In this paper, we focused on differential expression analysis at the gene level, as it has been shown that it is advantageous to perform statistical inference of differential expression at the gene level even if the quantification is done at the transcript level [41]. Since we used reference sets with only 1 protein entry per gene, our counts were automatically at the gene level. However, it is also possible to get isoform-level counts and aggregate the counts at the gene level for differential analysis. We saw no loss of performance with this approach (Supplementary Material). An advantage of isoform-level counts is that it can be used for other kinds of statistical tests such as differential usage of isoforms across condition. This is akin to the differential transcript expression/usage studies.

Choice of reference

Our results suggest – not surprisingly – that the choice of reference has a huge influence on the outcome of differential expression analysis, since a distant reference means fewer reads are aligned (correctly). One source to find a closest possible proteome is the UniProt Reference

Proteome database. This database currently contains almost 20,000 proteomes of organisms which are relatively well-studied and "provide broad coverage of the three of life".

Apart from single-species reference proteomes, it is also common to use cross-species proteins sets such as Swiss-Prot for annotating transcriptome assemblies. In theory, our method can also use Swiss-Prot as reference. However, Swiss-Prot is extremely redundant due to presence of orthologous proteins, which can needlessly aggravate the issue of multi-mapping. To use Swiss-Prot as reference, it is advisable to remove sequence redundancies by using tools such as CD-HIT [50] or MMSeq2 [51] and by selecting a subset of Swiss-Prot based on taxa.

Room for improvement

Currently we use a simple technique of rescuing multi-mapping reads. It would be interesting to explore a more sophisticated way to handle multi-mapping issues similar to a statistical model in RSEM.

LAST currently does not handle quality data present in the fastq records during alignment, and as far as we know, nor do other DNA-protein aligners. It is an interesting open problem to investigate if incorporating the quality data improves alignment accuracy, not just in this application to RNA-seq data analysis but to other applications of DNA-protein alignment.

Long reads

This paper is focused on short-read datasets, since from our cursory literature search in the Introduction section, it appears that long-read technologies are currently not widespread as in the applied fields that deal with non-model organisms. Theoretically, the core idea of DNA-protein alignment carries over just as well to long reads. Longer sequences can potentially improve accuracy as it would be easier to disambiguate counts among paralogous genes. However, application to long reads warrants a separate benchmarking process as one needs to account for error profiles and error rates characteristic to long-read technologies.

Conclusions

The flip side of RNA-seq becoming accessible to even labs with limited resources, is that the time, labor, and infrastructure cost of bioinformatics analysis has grown. Transcriptome assembly is one such resource-hungry process, which take several tens of hours on typical datasets, even on high-performance computing systems. For many labs, such requirements can impose a serious bottleneck. By avoiding assembly, our pipeline allows for quick-and-easy RNA-seq-based gene expression studies in non-model organisms.

Methods

Our proposed method

The first step in our proposed method is to align RNA-seq reads to a reference set of proteins using the DNA-protein alignment feature of LAST [29, 30]. We chose LAST over numerous other aligners capable of DNA-protein alignment – BLASTX [31] being a prominent example – for its unique combination of features. It scales well to high-throughput sequencing data. The probabilistic framework for incorporating paired information from paired-end reads, which was originally designed for read-to-genome alignment [32], works out of the box for the case of read-to-proteome alignment. It allows training the substitution matrix and gap penalties to reflect the sequence divergence between the (translated) RNA-seq reads and the reference proteome [33].

In the second step, from the alignments produced by LAST, we compute counts of reads originating from each entry in the reference. This is not trivial due to multi-mapping, an issue that becomes more pronounced when the reference contains isoforms with high sequence similarity. We employ the simple strategy of rescuing multi-mapping reads proposed in [34]. Suppose the reference is a set of protein sequences indexed by $P = \{1, 2, \ldots n\}$. The counting proceeds in two passes. In the first pass, we obtain u_i, which is the number of reads aligning uniquely to sequence i. In the second pass, for each read multi-mapping to a subset $P' \subseteq P$, we update the count of sequence $i \in P'$ in proportion to u_i, i.e. to the current count of sequence i, we add c_i, where

$$c_i = \frac{u_i}{\sum_{j \in P'} u_j}.$$

If the denominator is zero, we distribute the count evenly among P'.

The counts obtained in the second step can be fed directly to count-based differential gene expression analysis tools such as DESeq2 [35].

We implemented our strategy as a Snakemake workflow [36], which is available at `https://bitbucket.org/project_samar/samar`.

Generation of benchmarking dataset

We downloaded the transcripts of protein-coding genes from the fruit fly assembly BDGP6.28 obtained from Ensembl Genes 101. After removing sequence duplicates and transcripts with no corresponding protein entries, there were 28,692 transcripts of 13,320 genes. From this transcriptome, we simulated 2 groups of RNA-seq reads with 3 replicates per group using Polyester [38]. In the first group, the mean expression levels of the transcripts were set to be proportional to the FPKM values computed from an arbitrarily chosen poly-A+ enriched real RNA-seq data (ArrayExpress E-MTAB-6584). The FPKM values were estimated using RSEM [39] on Bowtie2 [40] alignments of the reads to the transcriptome. In the second group, a subset of roughly 30% of the transcripts were set to be differentially expressed at varying levels of up- and down-regulation: 1.5, 2, and 4-fold. The transcripts were chosen by randomly selecting genes and setting only the highest expressing isoform to be differentially expressed. Since inference of differential expression is typically done at the gene level, having at most one isoform to differentially expressed simplifies the evaluation process [41] as we can define a gene to be differentially expressed if one of its transcripts was differentially expressed. In fact, it might not be too far from reality as it has been shown that most highly expressed protein-coding genes have a single dominant isoform [42]. Each read set had roughly 20 million pairs of 100 bp reads with mean fragment length of 250 bp.

Abbreviations

Not Applicable

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The software is available at `https://bitbucket.org/project_samar/samar`. Scripts used in this paper for generation of simulated data and benchmarking study is available at: `https://bitbucket.org/project_samar/benchmarking`.

Competing interests

The authors declare that they have no competing interests.

Funding

AMSS was partially funded by University Research Coordination Office, De La Salle University-Manila. KCRS was partially funded by the Department of Science and Technology (DOST) Engineering Research and Development for Technology (ERDT) scholarship program. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Authors' contributions

AMSS planned the study. AMSS and JEBG performed the benchmarking study. AMSS, JEBG, and KCRS wrote the software. AMSS wrote the manuscript with inputs from all authors. All authors read and approved the final manuscript.

Acknowledgements

We are grateful to Martin Frith for his advice on DNA-protein alignment and LAST usage, and for providing helpful comments on the manuscript. We thank Hugues Richard for suggesting many improvements to the manuscript.

References

[1] Rory Stark, Marta Grzelak, and James Hadfield. Rna sequencing: the teenage years. *Nature reviews. Genetics*, 20:631–656, November 2019.

[2] Steven L. Salzberg. Next-generation genome annotation: we still struggle to get it right. *Genome Biology*, 20(1), may 2019.

[3] Jingwei Song and Jan R. McDowell. Comparative transcriptomics of spotted seatrout (cynoscion nebulosus) populations to cold and heat stress. *Ecology and Evolution*, 11(3):1352–1367, dec 2020.

[4] Jing Hou, Tao Xu, Dingjia Su, Ying Wu, Li Cheng, Jun Wang, Zhi Zhou, and Yan Wang. RNA-seq reveals extensive transcriptional response to heat stress in the stony coral galaxea fascicularis. *Frontiers in Genetics*, 9, feb 2018.

[5] Zhaoyang Hu, Yufei Zhang, Yue He, Qingqing Cao, Ting Zhang, Laiqing Lou, and Qingsheng Cai. Full-length transcriptome assembly of italian ryegrass root integrated with RNA-seq to identify genes in response to plant cadmium stress. *International Journal of Molecular Sciences*, 21(3):1067, feb 2020.

[6] Zhaoxu Gao, Biying Dong, Hongyan Cao, Hang He, Qing Yang, Dong Meng, and Yujie Fu. Time series RNA-seq in pigeonpea revealed the core genes in metabolic pathways under aluminum stress. *Genes*, 11(4):380, apr 2020.

[7] Lili Liu, Rong Zhang, Xiaowen Wang, Hua Zhu, and Zhaohui Tian. Transcriptome analysis reveals molecular mechanisms responsive to acute cold stress in the tropical stenothermal fish tiger barb (puntius tetrazona). *BMC Genomics*, 21(1), oct 2020.

[8] Mateusz Konczal, Amy R. Ellison, Karl P. Phillips, Jacek Radwan, Ryan S. Mohammed, Joanne Cable, and Magdalena Chadzinska. RNA-seq analysis of the guppy immune response against gyrodactylus bullatarudis infection. *Parasite Immunology*, 42(12), sep 2020.

[9] Seraina E. Bracamonte, Paul R. Johnston, Michael T. Monaghan, and Klaus Knopf. Gene expression response to a nematode parasite in novel and native eel hosts. *Ecology and Evolution*, 9(23):13069–13084, oct 2019.

[10] Qiang Sun, Huizhen Guo, Qingyou Xia, Liang Jiang, and Ping Zhao. Transcriptome analysis of the immune response of silkworm at the early stage of bombyx mori bidensovirus infection. *Developmental & Comparative Immunology*, 106:103601, may 2020.

[11] Tejas C. Bosamia, Sneha M. Dodia, Gyan P. Mishra, Suhail Ahmad, Binal Joshi, Polavakkalipalayam P. Thirumalaisamy, Narendra Kumar, Arulthambi L. Rathnakumar, Chandramohan Sangh, Abhay Kumar, and Radhakrishnan Thankappan. Unraveling the mechanisms of resistance to sclerotium rolfsii in peanut (arachis hypogaea l.) using comparative RNA-seq analysis of resistant and susceptible genotypes. *PLOS ONE*, 15(8):e0236823, aug 2020.

[12] Mónica I. Fass, Máximo Rivarola, Guillermo F. Ehrenbolger, Carla A. Maringolo, Juan F. Montecchia, Facundo Quiroz, Francisco García-García, Joaquín Dopazo Blázquez, H. Esteban Hopp, Ruth A. Heinz, Norma B. Paniego, and Verónica V. Lia. Exploring sunflower responses to sclerotinia head rot at early stages of infection using RNA-seq analysis. *Scientific Reports*, 10(1), aug 2020.

[13] Mafalda S. Ferreira, Paulo C. Alves, Colin M. Callahan, Iwona Giska, Liliana Farelo, Hannes Jenny, L. Scott Mills, Klaus Hackländer, Jeffrey M. Good, and José Melo-Ferreira. Transcriptomic regulation of seasonal coat color change in hares. *Ecology and Evolution*, 10(3):1180–1192, jan 2020.

[14] Hanbo Zhao, Hui Wang, Tong Liu, Sen Liu, Longru Jin, Xiaobin Huang, Wentao Dai, Keping Sun, and Jiang Feng. Gene expression vs. sequence divergence: comparative transcriptome sequencing among natural *rhinolophus ferrumequinum*, populations with different acoustic phenotypes. *Frontiers in zoology*, 16:37, 2019.

[15] Xue Lu, Hui-Min Chen, Xue-Qiao Qian, and Jian-Fang Gui. Transcriptome analysis of grass carp (ctenopharyngodon idella) between fast- and slow-growing fish. *Comparative Biochemistry and Physiology Part D: Genomics and Proteomics*, 35:100688, sep 2020.

[16] Yichao Wang, Baojie Wang, Mei Liu, Keyong Jiang, Mengqiang Wang, and Lei Wang. Comparative transcriptome analysis reveals the potential influencing mechanism of dietary astaxanthin on growth and metabolism in litopenaeus vannamei. *Aquaculture Reports*, 16:100259, mar 2020.

[17] Megan M. Kemski, Chad A. Rappleye, Konrad Dabrowski, Richard S. Bruno, and Macdonald Wick. Transcriptomic response to soybean meal-based diets as the first formulated feed in juvenile yellow perch (perca flavescens). *Scientific Reports*, 10(1), mar 2020.

[18] Wen-Zhi Guan, Gao-Feng Qiu, and Feng-Liu. Transcriptome analysis of the growth performance of hybrid mandarin fish after food conversion. *PLOS ONE*, 15(10):e0240308, oct 2020.

[19] Fengjiao Ma, Denghua Yin, Di-An Fang, Yanping Yang, Min Jiang, Lei You, Jia-Li Tian, Pao Xu, and Kai Liu. Insights into response to food intake in anadromous coilia nasus through stomach transcriptome analysis. *Aquaculture Research*, 51(7):2799–2812, apr 2020.

[20] Ming Xiao Li, Jun Qiang, Jing Wen Bao, Yi Fan Tao, Hao Jun Zhu, and Pao Xu. Growth performance, physiological parameters, and transcript levels of lipid metabolism-related genes in hybrid yellow catfish Tachysurus fulvidraco x Pseudobagrus vachelliis fed diets containing siberian ginseng. *PLOS ONE*, 16(2):e0246417, feb 2021.

[21] Yao Zheng, Wei Wu, Gengdong Hu, Liping Qiu, and Jiazhang Chen. Transcriptome analysis of juvenile tilapia (oreochromis niloticus) blood, fed with different concentrations of resveratrol. *Frontiers in Physiology*, 11, dec 2020.

[22] Shao-Kui Yi, Han-Ping Wang, Peng Xie, Xiao-Xia Li, and Hong Yao. Evaluation of growth and gene expression patterns of different strains related to SMD utilization in largemouth bass. *Aquaculture*, 523:735214, jun 2020.

[23] Sonia Rey, Xingkun Jin, Børge Damsgård, Marie-Laure Bégout, and Simon Mackenzie. Analysis across diverse fish species highlights no conserved transcriptome signature for proactive behaviour. *BMC genomics*, 22:33, January 2021.

[24] Wen Kin Lim and Ajay S Mathuru. Design, challenges, and the potential of transcriptomics to understand social behavior. *Current Zoology*, 66(3):321–330, feb 2020.

[25] Jordan Patterson, Eric J. Carpenter, Zhenzhen Zhu, Dan An, Xinming Liang, Chunyu Geng, Radoje Drmanac, and Gane Ka-Shu Wong. Impact of sequencing depth and technology on de novo RNA-seq assembly. *BMC Genomics*, 20(1), jul 2019.

[26] Nagarjun Vijay, Jelmer W. Poelstra, Axel Künstner, and Jochen B. W. Wolf. Challenges and strategies in transcriptome assembly and differential gene expression quantification. a comprehensive in-silico assessment of RNA-seq experiments. *Molecular Ecology*, 22(3):620–634, 2012.

[27] Ping-Han Hsieh, Yen-Jen Oyang, and Chien-Yu Chen. Effect of de novo transcriptome assembly on transcript quantification. *Scientific Reports*, 9(1), jun 2019.

[28] Nadia M Davidson and Alicia Oshlack. Corset: enabling differential gene expression analysis for de novoassembled transcriptomes. *Genome Biology*, 15(7), jul 2014.

[29] S. M. Kielbasa, R. Wan, K. Sato, P. Horton, and M. C. Frith. Adaptive seeds tame genomic sequence comparison. *Genome Research*, 21(3):487–493, jan 2011.

[30] Sergey L. Sheetlin, Yonil Park, Martin C. Frith, and John L. Spouge. Frameshift alignment: statistics and post-genomic applications. *Bioinformatics*, 30(24):3575–3582, 08 2014.

[31] Scott McGinnis and Thomas L. Madden. Blast: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic acids research*, 32:W20–W25, July 2004.

[32] Anish Man Singh Shrestha and Martin C. Frith. An approximate bayesian approach for mapping paired-end DNA reads to a reference genome. *Bioinformatics*, 29(8):965–972, feb 2013.

[33] Michiaki Hamada, Yukiteru Ono, Kiyoshi Asai, and Martin C Frith. Training alignment parameters for arbitrary sequencers with last-train. *Bioinformatics (Oxford, England)*, 33:926–928, March 2017.

[34] Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nature Methods*, 5(7):621–628, may 2008.

[35] Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15:550, 2014.

[36] J. Koster and S. Rahmann. Snakemake–a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, aug 2012.

[37] Jim Thurmond, Joshua L Goodman, Victor B Strelets, Helen Attrill, L Sian Gramates, Steven J Marygold, Beverley B Matthews, Gillian Millburn, Giulia Antonazzo, Vitor Trovisco, Thomas C Kaufman, Brian R Calvi, Norbert Perrimon, Susan Russo Gelbart, Julie Agapite, Kris Broll, Lynn Crosby, Gilberto dos Santos, David Emmert, L. Sian Gramates, Kathleen Falls, Victoria Jenkins, Beverley Matthews, Carol Sutherland, Christopher Tabone, Pinglei Zhou, Mark Zytkovicz, Nick Brown, Giulia Antonazzo, Helen Attrill, Phani Garapati, Alex Holmes, Aoife Larkin, Steven Marygold, Gillian Millburn, Clare Pilgrim, Vitor Trovisco, Pepe Urbano, Thomas Kaufman, Brian Calvi, Bryon Czoch, Josh Goodman, Victor Strelets, Jim Thurmond, Richard Cripps, and Phillip Baker and. FlyBase 2.0: the next generation. *Nucleic Acids Research*, 47(D1):D759–D765, oct 2018.

[38] Alyssa C. Frazee, Andrew E. Jaffe, Ben Langmead, and Jeffrey T. Leek. Polyester: simulating rna-seq datasets with differential transcript expression. *Bioinformatics (Oxford, England)*, 31:2778–2784, September 2015.

[39] Bo Li and Colin N Dewey. RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinformatics*, 12(1), aug 2011.

[40] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):R25, 2009.

[41] Charlotte Soneson, Michael I. Love, and Mark D. Robinson. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research*, 4:1521, dec 2015.

[42] Iakes Ezkurdia, Jose Manuel Rodriguez, Enrique Carrillo de Santa Pau, Jesús Vázquez, Alfonso Valencia, and Michael L. Tress. Most highly expressed protein-coding genes have a single dominant isoform. *Journal of Proteome Research*, 14(4):1880–1887, mar 2015.

[43] Viacheslav N. Bolshakov, Pantelis Topalis, Claudia Blass, Elena Kokoza, Alessandra della Torre, Fotis C. Kafatos, and Christos Louis. A comparative genomic analysis of two distant diptera, the fruit fly, drosophila melanogaster, and the malaria mosquito, anopheles gambiae. *Genome research*, 12:57–66, January 2002.

[44] Brian J Haas, Alexie Papanicolaou, Moran Yassour, Manfred Grabherr, Philip D Blood, Joshua Bowden, Matthew Brian Couger, David Eccles, Bo Li, Matthias Lieber, Matthew D MacManes, Michael Ott, Joshua Orvis, Nathalie Pochet, Francesco Strozzi, Nathan Weeks,

Rick Westerman, Thomas William, Colin N Dewey, Robert Henschel, Richard D LeDuc, Nir Friedman, and Aviv Regev. De novo transcript sequence reconstruction from RNA-seq using the trinity platform for reference generation and analysis. *Nature Protocols*, 8(8):1494–1512, jul 2013.

[45] Dammit pipeline. `https://github.com/dib-lab/dammit`.

[46] Erik L.L. Sonnhammer and Gabriel Östlund. InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Research*, 43(D1):D234–D239, nov 2014.

[47] Shifu Chen, Yanqing Zhou, Yaru Chen, and Jia Gu. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 34(17):i884–i890, sep 2018.

[48] Wei Zhao, Xiaping He, Katherine A Hoadley, Joel S Parker, David Hayes, and Charles M Perou. Comparison of RNA-seq by poly (a) capture, ribosomal RNA depletion, and DNA microarray for expression profiling. *BMC Genomics*, 15(1):419, 2014.

[49] Shanrong Zhao, Ying Zhang, Ramya Gamini, Baohong Zhang, and David von Schack. Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion. *Scientific Reports*, 8(1), mar 2018. share of ncRNA in poly-A.

[50] W. Li and A. Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, may 2006.

[51] Martin Steinegger and Johannes Söding. Clustering huge protein sequence sets in linear time. *Nature Communications*, 9(1), jun 2018.

Figures

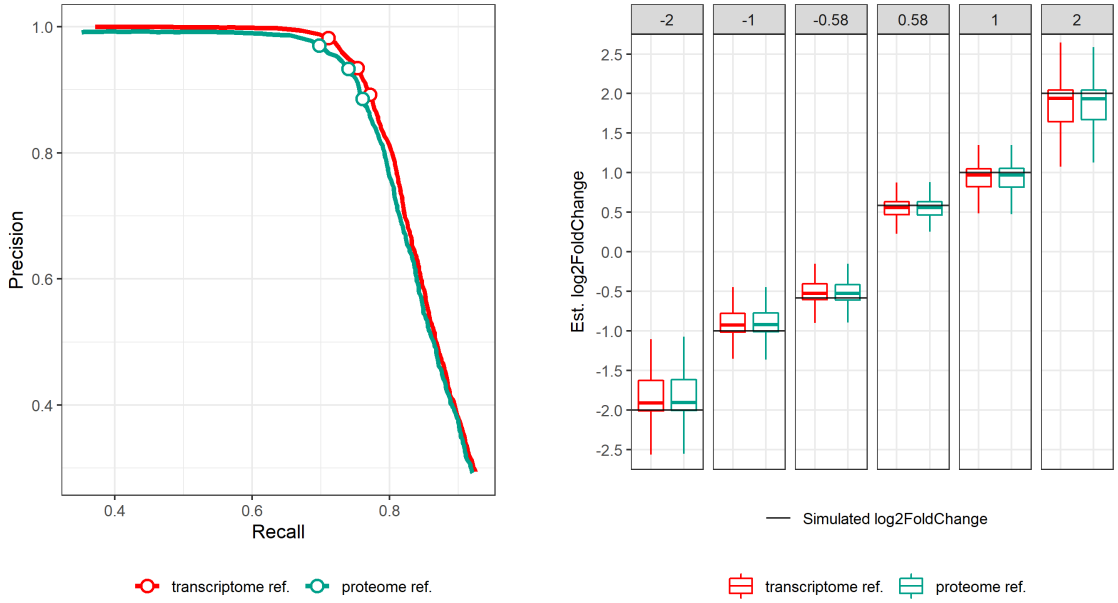

Figure 1: (Left) Precision-recall curves of our method using the *D. melanogaster* proteome reference and the Bowtie2-RSEM-DESeq2 pipeline using its transcriptome reference. The three open dots in each curve correspond to setting the FDR threshold of DESeq2 to 0.01, 0.05, and 0.1. (Right) Log fold change of true positive DE genes estimated by DESeq2 at FDR threshold of 0.1, compared against the 6 simulated log-fold change levels.

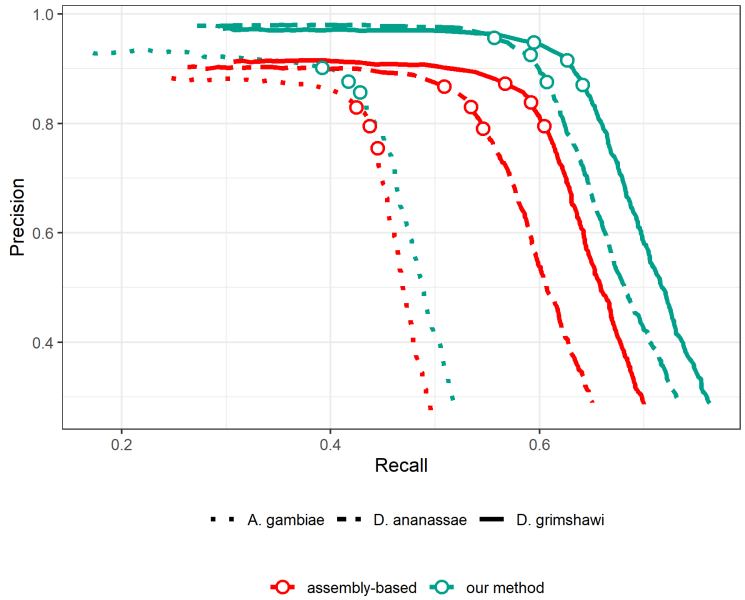

Figure 2: Precision-Recall curves of our method and the assembly-based pipeline, when using reference proteomes of close relatives (*D. ananassae* and *D. grimshawi*) and a distant relative (*A. gambiae*). The three open dots in each curve correspond to setting the FDR values of 0.01, 0.05, and 0.1.

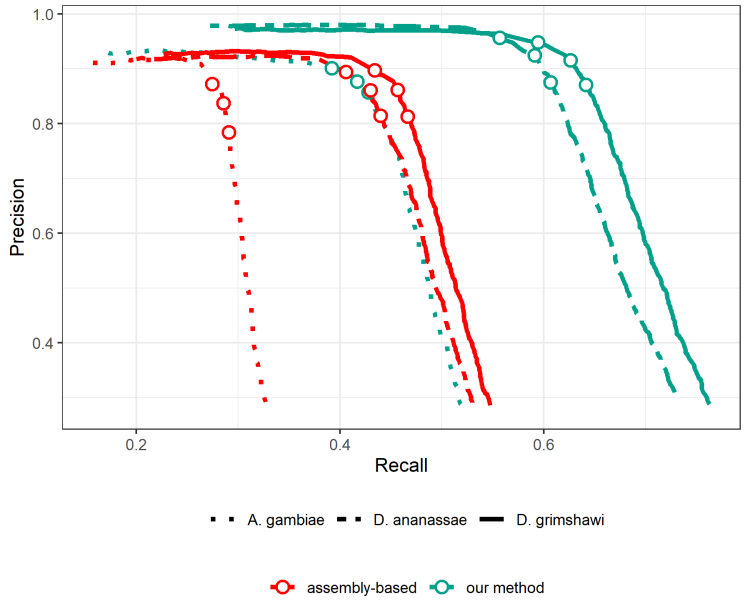

Figure 3: Precision-Recall curves of our method and the assembly-based pipeline, when using reference proteomes of close relatives (*D. ananassae* and *D. grimshawi*) and a distant relative (*A. gambiae*), and with the alignments produced by Dammit which covered less than 50% of the length of the contig removed. The three open dots in each curve correspond to setting the FDR values of 0.01, 0.05, and 0.1.

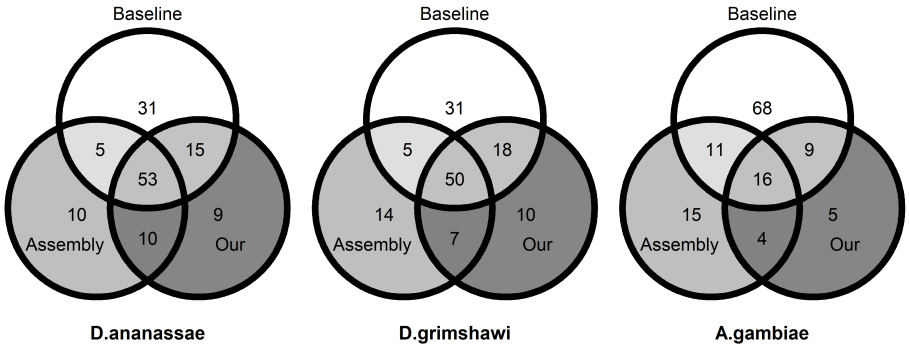

Figure 4: For the three reference proteomes, Venn diagrams showing the intersections among the **Baseline** set consisting of DE genes called by the baseline approach of Bowtie2-RSEM-DESeq2 using the *D. melanogaster* reference transcriptome, **Our** set consisting of *D. melanogaster* genes to which the DE genes called by our approach mapped to, and (3) **Assembly-based** set consisting of *D. melanogaster* genes to which Trinity DE genes mapped to. FDR threshold of 0.01 was used for all three approaches.