CRSP: Comparative RNA-seq pipeline for species lacking both of sequenced genomes and reference transcripts

Atefeh Bagheri[1], Colin Dewey[2,3], Ron Stewart[4], Peng Jiang[1,5,6*]

[1]Department of Biological, Geological and Environmental Sciences, Cleveland State University, Cleveland, OH 44115, USA

[2]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI 53706, USA

[3]Department of Computer Science, University of Wisconsin-Madison, Madison, WI 53706, USA

[4]Morgridge Institute For Research, Madison, WI 53706, USA

[5]Center for Gene Regulation in Health and Disease, Cleveland State University, Cleveland, OH 44115, USA

[6]Center for Applied Data Analysis and Modeling, Cleveland State University, Cleveland, OH 44115, USA

*To whom correspondence should be addressed: p.jiang@csuohio.edu

ABSTRACT

Summary: RNA sequencing (RNA-seq) has been a widely used high-throughput method to characterize transcriptomic dynamics spatiotemporally. However, typical RNA-seq data analysis pipelines depend on either a sequenced genome or reference transcripts. This constriction makes the use of RNA-seq for species lacking both of sequenced genomes and reference transcripts challenging. To solve this problem, we developed CRSP, an RNA-seq pipeline integrating multiple comparative species strategy but not depending on a specific sequenced genome or reference transcripts. Benchmarking suggests the CRSP tool can achieve high accuracy to quantify gene expression levels.

Availability: CRSP and documents are available at: https://github.com/pjiang1105/CRSP

1 INTRODUCTION

RNA sequencing (RNA-seq) has been widely used for characterization of transcriptomic dynamics with different biological contexts, such as development and diseases. Current RNA-seq tools either rely on an annotated reference genome or reference transcripts (e.g., RSEM). For species lacking both of sequenced genomes and reference transcripts, performing RNA-seq data analysis remains challenging. Our prior studies along with others suggest combing *de novo* transcripts assembly (Grabherr, et al., 2011) with several cross-species comparative computational strategies, such as mapping contigs to a protein database, can partially solve this problem (Jiang, et al., 2017). However, no such standard freely available pipelines are available to implement these approaches. Furthermore, there is no benchmarking to evaluate these approaches, and thus it is largely unknown that how good of these comparative approaches. There are 8.7 million eukaryotic species on our planet while only 100 have been genomic sequenced with high quality. This limits most of the RNA-seq studies to these high quality genomic sequenced model species.

To facilitate a general use of RNA-seq to species without any prior references (either genome or transcripts), we developed CRSP (Comparative RNA-seq pipeline). CRSP integrates a set of computational strategies, such as mapping *de novo* transcriptomic assembly contigs to a protein database, using Expectation-Maximization (EM) algorithm to assign reads mapping uncertainty, and integrative statistics to quantify gene expression values. Benchmarking showed the CRSP estimated gene expression values are highly correlated with gene expression values estimated by directly mapping to a

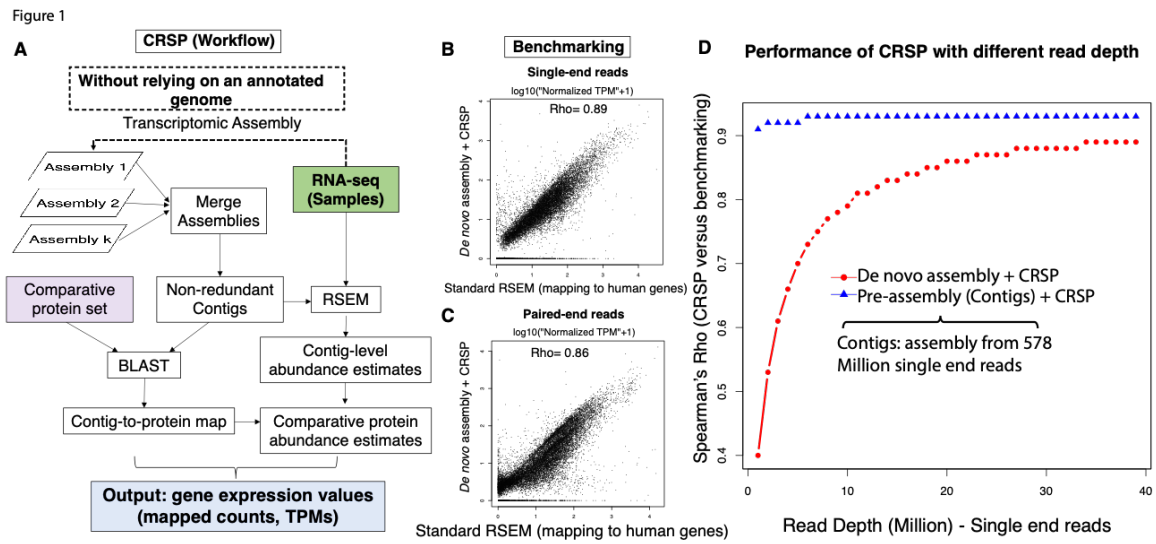

Figure 1: CRSP tool. (A) Workflow of CRSP; (B-C) CRSP estimated gene expression values are highly correlated with gene expression values estimated by directly mapping to a reference genome; (D) The performance of CRSP with different read depth.

reference genome. To facilitate a general use of CRSP, we further addressed the minimum RNA-seq reads requirement to run CRSP via simulation studies. Our simulation study suggests that 10 to 20 million single-end reads are sufficient to achieve reasonable gene expression quantification accuracy while a pre-compiled *de novo* transcripts assembly from deep sequencing can dramatically decrease the minimal reads requirement for the rest of RNA-seq experiments.

2 METHODS

CRSP: As shown in **Figure 1A**, CRSP takes RNA-seq raw reads and the transcriptomic assembly contigs (e.g., generated by Trinity (Grabherr, et al., 2011)) as input, and then leverage multiple computational strategies to estimate gene expression levels. The workflow of CRSP will contain the following steps: (a) Merging assembly (Contigs): The contigs from different tissues, conditions or batches will be merged, and then generated

a non-redundant contig set; (b) Mapping contigs to a protein database: the non-redundant transcriptomic contig set will be mapped to a protein database (e.g., human or mouse protein database) via NCBI BLASTX tool (Altschul, et al., 1990). CRSP excludes hits with E-value >= 10^{-5}, and then the best hit will be selected; (c) Mapping RNA-seq reads to contigs: The RNA-seq reads are mapped to contigs, and then a contig level abundance estimation is performed via RSEM tool (Li and Dewey, 2011); (d) For each RNA-seq sample, gene-level RNA-seq mapped read counts are obtained by summing the contig read mapping counts that are annotated with the same gene symbol (mapped to the same proteins); (f) Relative abundances, in terms of transcripts per million (TPM), for genes are computed by first normalizing each gene's fragment count by the sum of the "effective lengths" (length less the read length) of the contigs mapped to that gene and then scaling the resulting values such that they summed to one million over all genes. A more detailed workflow of CRSP can be found in **Supplementary Figure S1**.

RNA-seq datasets: The single-end reads (human, embryonic stem cell (ES), H1 cell line) was from our published work (Hou, et al., 2015). We also generated 80 million × 2 paired-end reads (human, H1 ES cells) as a second dataset to evaluate our CRSP tool. The detailed RNA-seq protocols for this newly generated RNA-seq data for this study can be found in **Supplementary Methods**. We have submitted the paired-end RNA-seq data to GEO (accession number: GSE194237).

3 RESULTS

To evaluate the performance of CRSP, we pooled 578 million human ES single-end RNA-seq reads (Hou, et al., 2015), and then generated a gold-standard gene expression value set by directly mapping to human reference genes via RSEM tool (Li and Dewey, 2011). As shown in **Figure 1B**, the gene expression values which were estimated by CRSP (*de novo* assembly, mapping to mouse protein database) are highly correlated with the gold-standard gene expression values (Spearman's rank correlation coefficient Rho= 0.89). We also generated we 80 million × 2 paired-end reads (human, H1 ES cells) as a second dataset testing. As shown in **Figure 1C**, the CRSP also achieved a high accuracy (Spearman's Rho = 0.86). To further estimate the minimal reads requirement to run CRSP, we randomly sampled RNA-seq reads from pooled 578 million human ES single-end RNA-seq reads, and then estimated the Spearman's Rho of gene expression values estimated from CRSP and from directly mapping to human reference genes, respectively. As shown in Figure 1D, although the performance of CRSP progressively increased with more sequencing reads, 10-20 million single-end reads can be sufficiently to achieve reasonable accuracy. Importantly, if we have a pre-compiled transcriptomic contigs set from a deep sequencing, CRSP does not require deep read depth to achieve high accuracy (**Figure 1D, blue line**). Together, our study suggest CRSP is a useful tool making RNA-seq as a general tool to interrogate transcriptomic dynamics but not only limited to species with sequenced genomes or known transcripts.

ACKNOWLEDGEMENTS

This study was supported by Cleveland State University startup to PJ.

REFERENCES

Altschul, S.F., *et al.* Basic local alignment search tool. *J Mol Biol* 1990;215(3):403-410.

Grabherr, M.G., *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 2011;29(7):644-652.

Hou, Z., *et al.* A cost-effective RNA sequencing protocol for large-scale gene expression studies. *Sci Rep* 2015;5:9570.

Jiang, P., *et al.* Analysis of embryonic development in the unsequenced axolotl: Waves of transcriptomic upheaval and stability. *Dev Biol* 2017;426(2):143-154.

Li, B. and Dewey, C.N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 2011;12:323.

Supplementary Data

Supplementary Methods

Paired-end RNA-seq: To further evaluate CRSP on paired-end RNA-seq data, we generated 80 million × 2 paired-end RNA-seq reads on human H1 ES cells. Briefly, total RNA was purified from RLT-Plus Buffer using RNeasy Plus Mini Kits (Qiagen, Netherlands) and quantified with both the Quant-It RNA Assay Kit (Thermofisher, Massachusetts) and the Bioanalyzer RNA 6000 Pico Kit (Agilent, California). The TruSeq RNA Library Prep Kit v2 (Illumina, San Diego), as per the manufacturer's directions, was used to prepare and index all cDNA libraries. The RNA-seq library was sequenced on Illumina HiSeq 2500. We have submitted the paired-end RNA-seq data to GEO (accession number: GSE194237).

Supplementary Figures

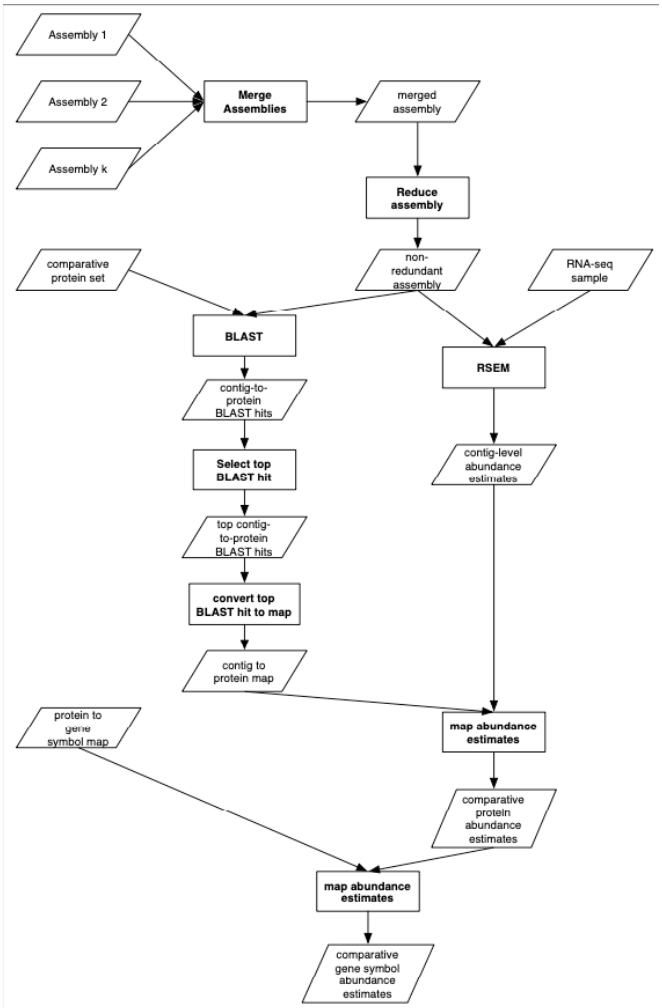

Figure S1: Detailed workflow of CRSP: CRSP takes RNA-seq raw reads and the transcriptomic assembly contigs (e.g., generated by Trinity) as input, and then leverage multiple computational strategies to estimate gene expression levels.