RNA-Seq *De Novo* Assembly and Analysis Pipeline using Trinity and Reproducing study results conducted by Haas et al. (2013)[1]

1 Introduction

RNA-Seq is a promising technology for understanding gene expression and metabolic networks. *De novo* transcriptome assembly is the preferred strategy to study non-model organisms, which often lack a reference genome.

RNA-Seq The development of the de Bruijn graph-based Trinity software package[2,3] has improved the de novo reconstruction of transcriptomes from RNA-Seq reads. Trinity was developed at the Broad Institute and the Hebrew University of Jerusalem[2] and features low base error rates and the ability to capture multiple isoforms.[2] In a recently published paper, Haas et al. demonstrate an example of the effective use of a workflow using the Trinity package for *de novo* transcriptome assembly and analysis of RNA-Seq data from the widely studied organism *S. pombe*.[1]

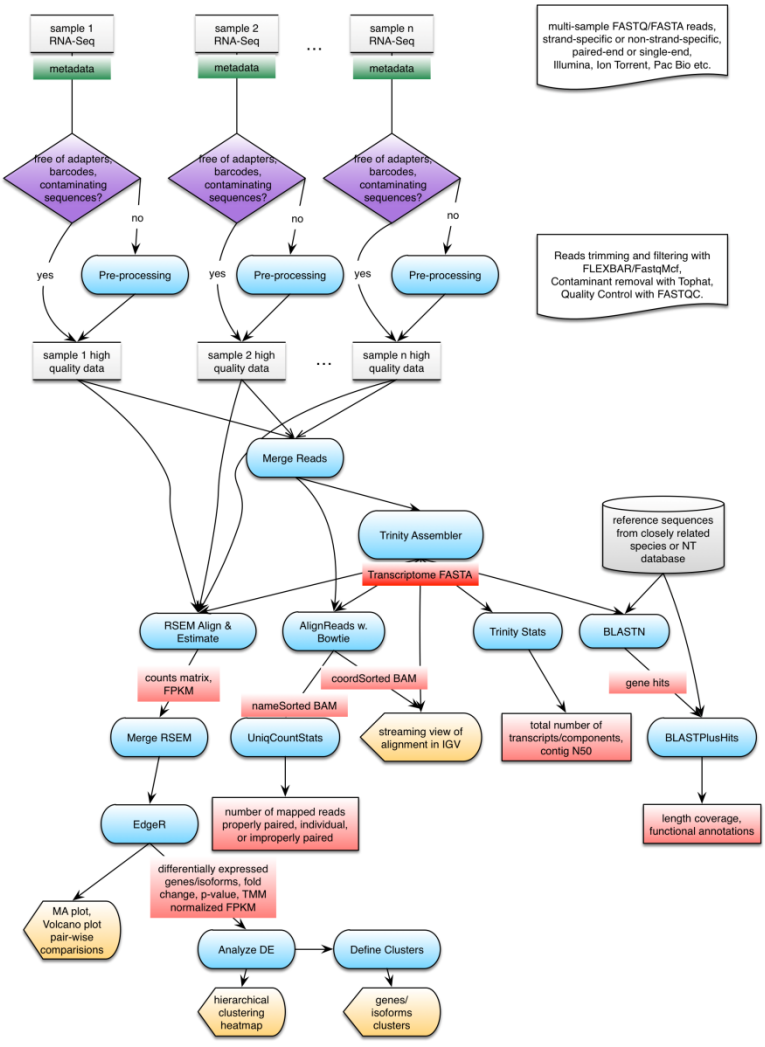
Despite the advantages that Trinity offers, setting up such a workflow for large-scale statistical analysis on local machines or clusters remains challenging. The demand for high capacity of hardware memory and CPUs inherent to big data, the significant knowledge required to effectively use algorithms to optimize results performance, and the substantial use of computational resources to optimize time performance add to the complexities of *de novo* assembly.

In this report, we present a cloud-based implementation of the Trinity pipeline on Seven Bridges Genomics's platform, which enables automatic, flexible, and scalable analysis of RNA-Seq data. Shown here also is testing analysis that demonstrates that the above-mentioned computational experiment[1] is reproducible, despite the fact that Trinity is a non-deterministic algorithm and that we applied different versions for some tools in this task.

We present key statistics and plots from this analysis, which are compared and assessed for consistency against previously published results.[1] Finally, we describe one of our extended features: streaming visualization of alignments in IGV, which enables exploring transcript expression profiles in greater detail. This step is useful for verification of strand-specific library type, and ensures the quality of assembly.

Our pre-built ready-to-run complete pipeline is available at:
https://igor.sbgenomics.com/lab/pipeline/view/52cf2564d79f0008ddb87e3f/
An alternative fast and basic version for single sample assembly is available at:
https://igor.sbgenomics.com/lab/pipeline/view/5346ba3dd79f0049c0c944a6/

Figure 1. The overall Trinity workflow for *de novo* transcriptome assembly and gene expression study using RNA-Seq data from non-model organisms that lack a reference genome. To facilitate downstream analysis (i.e. comparison between multiple samples, tissues, environmental conditions, and so on), reads from all samples will be merged to generate a single assembly.



2 Trinity Pipeline Overview

Figure 1 describes the Trinity computational workflow. The major components of this workflow are the Trinity assembler[1], Bowtie aligner[6], RSEM[4] for transcripts abundance estimation, EdgeR[5] for differential expression analysis, and BLASTN[7] for functional annotation and assembly quality control (see Appendix for tool versions). Default parameter values are set according to best practice recommendations.[1] Individual modules can be upgraded, replaced by other tools, or customized.

Computational tasks using this pipeline are executed in Seven Bridges Genomics's cloud framework. Sub-jobs are automatically distributed over different machines and are run in parallel mode when possible. The machine types are chosen based on the memory and CPU requirements from different algorithms. For example, Trinity assembly tasks are run on cluster instances with 244GB RAM and 32 CPUs. Large-scale, independent experiments can be conducted simultaneously using the "batch" functionality.

3 Results of a case study (SBG task ID: 17146)

3.1 Source data

- RNA-Seq reads of *Schizosaccharomyces pombe* grown in four conditions, from Haas et al. (2013)[1] (see Appendix)
- A total of 4 million Illumina paired-end reads
- Strand-specific library type: RF

3.2 Pre-processing

The pipeline requires input data to be of high quality: free from adapters, barcodes, and other contaminating sub-sequences. If raw data needs pre-processing, Flexbar/FastqMcf can be used for read trimming and filtering by quality scores, and Tophat can be used for examining and removing contaminant sequences. Last, FASTQC can be used to assess the quality of processed reads, to make sure that they are acceptable for downstream analysis.

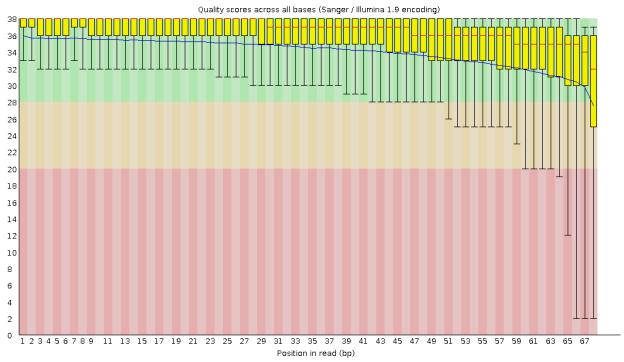


Figure 2. FASTQC representation of diauxic shift data for quality assessment. X-axis: position in read; y-axis: base quality score.

3.3 Transcriptome assembly

Trinity

Trinity software assembles full-length or nearly full-length transcripts using RNA-Seq data in three steps: 1. Greedily and efficiently assembling reads in unique sequences of transcript contigs (Inchworm); 2. Clustering related contigs that correspond to portions of alternatively spliced transcripts or otherwise unique portions of paralogous genes, forming a de Bruijn graph for each cluster of related contigs (Chrysalis); and 3. Analyzing the paths taken by reads and read pairings in the context of the corresponding de Bruijn graph and reporting all plausible transcript sequences (Butterfly)[2].

Statistical summary

Total trinity transcripts (isoforms): 9349
Total trinity components (genes): 8698
Contig N50: 1584

Note: "gene" used loosely here.

This result is highly consistent with the stats from Haas et al. (2013),[1] which reported 9299 transcripts, 8694 components, and Contig N50 value 1585.

Note: Contig N50 value is the maximum length whereby at least 50% of the total assembled sequence resides in contigs of at least that length on the basis of assembled transcripts[1]. Here the metric is used to confirm the assembly success.

3.4 Functional annotation

BLASTN

BLASTN is a local alignment and search algorithm that can rapidly compare a query sequence with a database of nucleotide sequences[7]. To examine the breadth of genetic composition and transcript contiguity, assembled transcripts were aligned against known gene sequences with BLASTN. Gene hits and their functional annotations were reported in the *.pct_hit_length.txt file. BlastPlusHits examines the length coverage of top database hits, and provides the top hit's length, percent of the hit's length,

and distribution of percent length coverage for the top matching database entries.

Here, we use *S. pombe*'s known transcripts as the reference database. If well-annotated relative species are not available, we recommended changing the database option to NCBI's nt.

Statistical summary

Shown here, 4766 (4765[1]) of the reference *S. pombe* transcripts have a BLASTN hit with an E-value less than 1e-20. 3401 (3401[1]) are considered to be of approximately 'full length', with the Trinity contigs aligning by greater than 90% of the matching reference transcript's length.[1]

Table 1. Distribution of BLASTN hit coverage of reference transcripts.

#hit_pct_cov_bin	count_in_bin	>bin_below
100	3401	3401
90	194	3595
80	165	3760
70	197	3957
60	224	4181
50	202	4383
40	160	4543
30	139	4682
20	84	4766
10	0	4766
0	0	4766

3.5 Expression quantification

RSEM

RSEM is a utility that enables accurate gene and isoform quantification for species without sequenced genome[4]. Transcripts abundance estimation was performed using Trinity's RSEM Align & Estimate module, in which the number of RNA-Seq fragments mapped to each transcript were counted and expression level was represented by FPKM in the *.isoforms.results and *.genes.results files.

3.6 Differential expression analysis

EdgeR

EdgeR is a Bioconductor package for differential expression analysis of RNA-Seq data. In this section, differentially expressed transcripts for each pair of samples were identified using EdgeR. By default, we reported transcripts having a significance p-value of <0.001 and at least a four-fold difference in expression value between two samples.

Gene and isoform levels were compared across each pair of samples. This step generates an *.edgeR.DE_results file that includes differential transcripts, fold change, significance p-value, and FDR. TMM (trimmed mean of M-values) scaling normalization was performed in order to account for differences in total cellular RNA production across all samples. The result file TMM_info.txt contains the effective library size for each sample after TMM normalization, and *.TMM_normalized.FPKM contains normalized transcript expression values according to the transcript and sample, measured as FPKM.

Table 2. The top 5 transcripts that are differentially expressed between logarithmic versus plateau conditions. According to logFC, logCPM, p-value, and FDR, this result is highly consistent with Table 3 in Haas et al. (2013)[1].

Transcript	logFC	logCPM	p-value	FDR
comp5132_c0_seq1	10.3	11.1	2.13e-22	1.22e-18
comp5235_c0_seq1	10.1	10.9	1.09e-21	3.13e-18
comp5101_c0_seq1	8.7	11.3	5.73e-20	1.10e-16
comp2035_c0_seq1	9.2	10.4	1.01e-19	1.46e-16
comp1972_c0_seq1	8.3	11.5	2.81e-19	3.23e-16

Note: Transcript identifiers are randomly assigned.

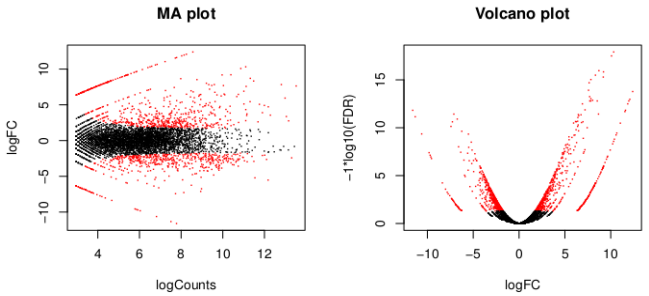
Statistical summary

The number of differentially expressed transcripts is 668, which is comparable to 659, the number reported in Haas et al. (2013)[1].

MA plot & Volcano plot

Figure 3. Comparison of transcript expression profiles between the logarithmic growth and plateau growth samples from *S. pombe.* In the MA plot (left), the $\log_2$(fold change) between the two samples is plotted (y-axis) against the gene's $\log_2$(average expression) in the two samples (x-axis). Volcano plot (right) reports $-\log_{10}$(false discovery rate) (y-axis) as a function of

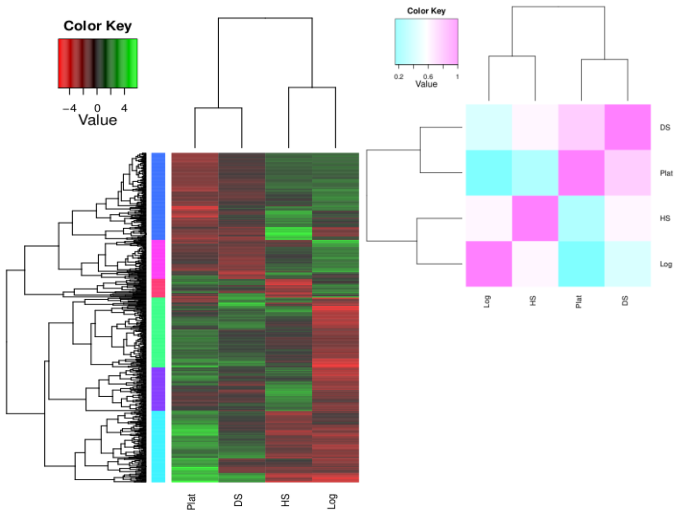
log_2(fold change) between the samples (x-axis). Transcripts that are identified as significantly, differentially expressed at most 0.1% FDR are colored in red.



Analyze DE

This step generates hierarchical clustering of transcripts and samples based on normalized transcriptional expression profiles.

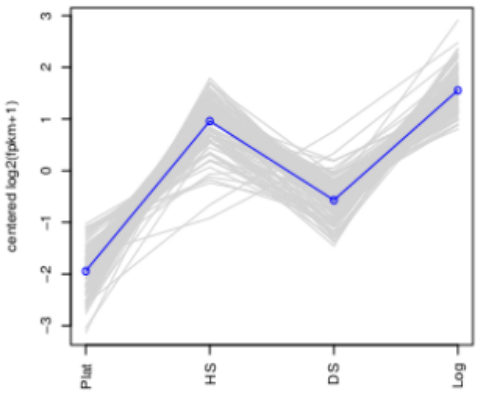
Figure 4. Hierarchical clustering of assembled transcripts and the four *S. pombe* samples. Color indicates the log2-transformed, median-centered expression value for each transcript. On the left is a heat map showing the relative expression levels of each transcript (rows) in each sample (column). On the right is a heat map showing the hierarchically clustered Spearman correlation matrix resulting from comparing the transcript expression values for each pair of samples[1].



Extract clusters

This step defines clusters of transcripts with common expression patterns and displays their expression changes across samples. The clusters are extracted from the hierarchical clustering tree (Figure 4). The function cuts the tree based on given criteria either at a certain height (20% of its height for this case study) or at a point that will generate a specific number of clusters.[1]

Figure 5. Viewing an example cluster of transcripts and their common expression profiles across the four conditions in the study. The grey lines indicate individual transcripts, and the blue lines indicate average expression level for each cluster.



3.7 Visualization of alignment

One advantage of using RNA-Seq over Microarray is that gene expression can be quantified at single nucleotide resolution with RNA-Seq. In this section, we use IGV to visualize reads alignment along transcripts and explore expression structure in greater detail.

Bowtie

Bowtie is an ultrafast, memory-efficient short read aligner[6]. Trinity's AlignReads module uses Bowtie to align left and right fragment reads separately to the assembled contigs and groups the reads together into pairs while retaining those single-read alignments that are not found to be properly paired with their mates[1]. For strand-specific data, as in this case study, AlignReads separates the aligned reads that map to the sense strand (' + ') from those that align to the anti-sense strand (' − ')[1].

IGV

IGV (Integrative Genomics Viewer) is a lightweight tool for visualizing diverse, large-scale genomic data on standard desktop computers[8]. The above-generated coordSorted BAM file can be imported into the IGV via its associated URL, which provides streaming visualization of reads alignment, information about read depth, and the difference in expression profiles between sense and anti-sense strands.

Figure 6. Steps for alignment visualization in IGV. Copy BAM file's associated URL and download assembled FASTA file. In IGV, load BAM from URL and load assembled FASTA as Genome.

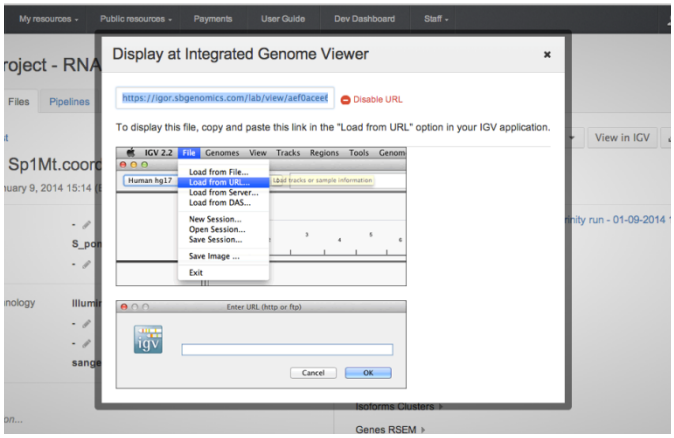
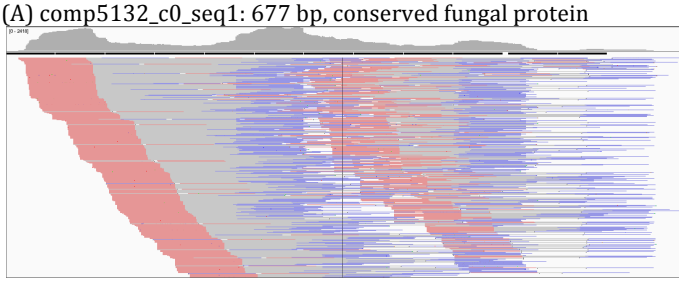
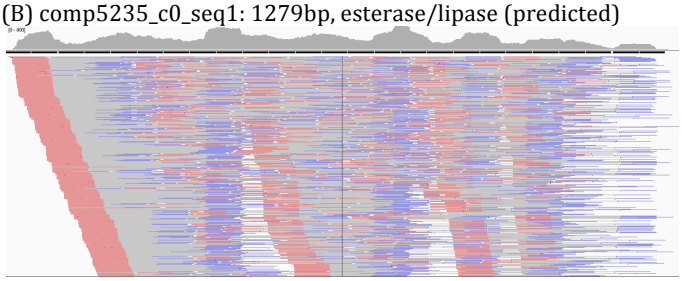


Figure 7. Exploration of three example transcripts (A-C) from Table 2, using IGV. For each plot, on the top is the track that showing pileup signals spanning each full transcript. The bottom track displays aligned reads. Color indicates first (purple) and second (red) in pair, and unpaired second read (grey).

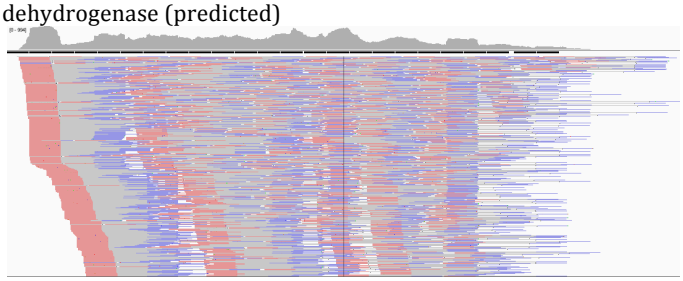
(A) comp5132_c0_seq1: 677 bp, conserved fungal protein



(B) comp5235_c0_seq1: 1279bp, esterase/lipase (predicted)



(C) comp2035_c0_seq1: 1474bp, 3-hydroxyacyl-CoA dehydrogenase (predicted)



According to the paired-end reads mappings in Figure 7, we observed that the first reads in pair (purple) are mapped to the reverse of the sense strand of transcripts. This pattern matches the description of strand-specific library type, i.e. reverse-forward (RF). We also observed that each transcript has sufficiently read coverage across the full sequence. We didn't observe obvious breakpoints or vertical discontinuities, indicating that the assembled contigs are not chimeric artifacts. This confirms the accuracy of reconstruction of transcripts.

It is clear that viewing alignment is useful for verification of strand-specific library type, and ensures the quality of assembly, as well as the applicability for interpretation of transcript abundance.

4 Runtime:

	Case 1	Case 2	Case 3
Species	*S. pombe*	*S. pombe*	Mouse
Number of samples	four	single	single
Total reads	4 million	1 million	100 million
Assembly	Yes	Yes	Yes
Alignment	Yes	Yes	Yes
Expression quantification	Yes	Yes	Yes
Differential expression	Yes	No	No
Clustering	Yes	No	No
Functional annotation	Yes	No	Yes
SBG Task ID	17146	27832	25264
Pipeline version	Complete	Fast	Customized
Runtime	1 hr 40 min	42 min	1 day 6 hr

5 Discussions:

In this report, we described the *de novo* transcriptome pipeline Trinity and data mining procedures. In the future, we will add more analytical functions, such as Trinotate, BLASTX, and integrate our interactive visualization tool Spectacle. We will also continue to explore the patterns observed in Figure 7, which holds promise for refining transcript structures.

6 References:

[1] Brian J. Haas et al. De novo transcript sequence reconstruction from RNA-Seq using the Trinity platform for reference generation and analysis. *Nature Protocols* 8, 1494–1512 (2013)

[2] Grabherr, M.G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol*. 29, 644–652 (2011).
[3] Martin, J.A. & Wang, Z. Next-generation transcriptome assembly. Nat. Rev. Genet. 12, 671–682 (2011).
[4] Li, B. & Dewey, C.N. RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinformatics* 12, 323 (2011).
[5] Robinson, M.D., McCarthy, D.J. & Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140 (2010).
[6] Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10, R25 (2009).
[7] Altschul, S.F. et al. Basic local alignment search tool. *J Mol Biol.* 251(3), 403-10 (1990).
[8] Robinson, J.T. et al. Integrative genomics viewer. *Nat. Biotechnol.* 29, 24–26 (2011).

7 Appendix

Input files specification

- Schizosaccharomyces pombe; logarithmic growth, 1 million Illumina paired-end reads:
  - Sp.log.1M.left.fq
  - Sp.log.1M.right.fq
- Schizosaccharomyces pombe; plateau phase, 1 million Illumina paired-end reads:
  - Sp.plat.1M.left.fq
  - Sp.plat.1M.right.fq
- Schizosaccharomyces pombe; diauxic shift, 1 million Illumina paired-end reads:
  - Sp.ds.1M.left.fq
  - Sp.ds.1M.right.fq
- Schizosaccharomyces pombe; heat shock, 1 million Illumina paired-end reads:
  - Sp.hs.1M.left.fq
  - Sp.hs.1M.right.fq
- Schizosaccharomyces pombe reference transcriptome:
  - S_pombe_refTrans.fasta
- Strand-specific description file:
  - samples_n_reads_described.txt

These files can be downloaded from Seven Bridges Public Data library: (https://igor.sbgenomics.com/lab/public/files/#q?page=1)

SBG software versions

Trinity version:
trinityrnaseq_r20130225
trinityrnaseq_r20121005
Bowtie version 0.12.9
Samtools version 0.1.18
R version 3.0.1
Blast + version 2.2.27

Package tools

Trinity.pl
TrinityStats.pl
analyze_blastPlus_topHit_coverage.pl
alignReads.pl
SAM_nameSorted_to_uniq_count_stats.pl
run_RSEM_align_n_estimate.pl
merge_RSEM_frag_counts_single_table.pl
run_DE_analysis.pl
analyze_diff_expr.pl
define_clusters_by_cutting_tree.pl