Title	Optimization of Sequencing Parameters for de novo Transcriptome Assembly by Comprehensive Analysis of Public RNA-Seq Data
Sub Title	
Author	板谷, 英駿(Itaya, Hidetoshi)
Publisher	慶應義塾大学湘南藤沢学会
Publication year	2014
Jtitle	生命と情報 No.21 (2014.) ,p.89- 98
Abstract	The advent of the second generation sequencing technology has made RNA sequencing (RNA-Seq) a preferable choice over existing transcriptome profiling methods. As RNA-seq experiments can be designed to output data which do not require a reference genome to analyze, it is a cost efficient and high-throughput choice for transcriptome profiling of non-model organisms which only have partial, if not any reference genome. However, as many second generation sequencers including the popular Illumina system are not capable of reading the full length of many transcripts, the reads must first be assembled in effort to reconstruct the original transcriptome. There have been studies independently addressing the influences of some sequencing parameters in de novo transcriptome assembly; namely the read depth and read length. Both reports suggest the existence of a fundamental limit for performance enhancement, but to date there is no study which comprehensively analyzes the influences of such parameters in actual RNA-seq data which are publicly accessible. In this research, RNA-seq data for four model organisms were obtained from the sequencing read archive (SRA), categorized by read depth and read length, assembled using the SOAPdenovo-Trans and Trinity software, and evaluated with several assembly metrics and by searching against a well defined subset of Clusters of Orthologous Groups (COGs) included in the Core Eukaryotic Genes Mapping Approach (CEGMA) pipeline. By studying these results we aimed to generate guidelines to selecting sequencing parameters for RNA-seq experiments targeted to non-model organisms. However, assembly results were not consistent most likely due to lack of data.
Notes	慶應義塾大学湘南藤沢キャンパス先端生命科学研究会 2014年度学生論文集 卒業論文ダイジェスト
Genre	Technical Report
URL	http://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=KO92001004-00000021-0089

Graduation thesis digest:

Optimization of Sequencing Parameters for *de novo* Transcriptome Assembly by Comprehensive Analysis of Public RNA-Seq Data

71140802

Hidetoshi Itaya

Abstract

The advent of the second generation sequencing technology has made RNA sequencing (RNA-Seq) a preferable choice over existing transcriptome profiling methods. As RNA-seq experiments can be designed to output data which do not require a reference genome to analyze, it is a cost efficient and high-throughput choice for transcriptome profiling of non-model organisms which only have partial, if not any reference genome. However, as many second generation sequencers including the popular Illumina system are not capable of reading the full length of many transcripts, the reads must first be assembled in effort to reconstruct the original transcriptome. There have been studies independently addressing the influences of some sequencing parameters in *de novo* transcriptome assembly; namely the read depth and read length. Both reports suggest the existence of a fundamental limit for performance enhancement, but to date there is no study which comprehensively analyzes the influences of such parameters in actual RNA-seq data which are publicly accessible. In this research, RNA-seq data for four model organisms were obtained from the sequencing read archive (SRA), categorized by read depth and read length, assembled using the SOAPdenovo-Trans and Trinity software, and evaluated with several assembly metrics and by searching against a well defined subset of Clusters of Orthologous Groups (COGs) included in the Core Eukaryotic Genes Mapping Approach (CEGMA) pipeline. By studying these results we aimed to generate guidelines to selecting sequencing parameters for RNA-seq experiments targeted to non-model organisms. However, assembly results were not consistent most likely due to lack of data.

Keywords: Transcriptomics, *de novo* Transcriptome Assembly, Non-classical Model Organisms

1 Background

The advent of massively parallel sequencing technologies has opened a new possibility for profiling transcriptome abundance. The method designated RNA sequencing (widely known as RNA-seq) (1,2), is a promising technology which has many advantages over existing transcriptome profiling techniques (3-7). RNA-seq yields a high-throughput, is capable of detecting a wide dynamic range of gene expression as well as novel isoforms, and can be designed to be analyzed without a reference genome (1,8). These benefits make RNA-seq a high-throughput and cost efficient experiment for transcriptome profiling of non-classical model organisms. However, popular sequencing platforms including sequencers from Illumina, Roche 454, and Life Technologies all produce short reads from few tens to few hundreds of bases which is insufficient to cover the full-length of every transcript. On the contrary, the single molecule real time (SMRT) sequencing technology developed by Pacific Biosciences is capable of sequencing up to a few tens of kilobases but has moderate throughput and poor quality compared to methods listed above. According to the metadata of the Sequence Read Archive (SRA) (9-11) accessible at the DBCLS SRA (12), Illumina currently dominates the number of applications in RNA-seq, used in more than 4,000 studies out of 4,915 transcriptome analysis projects total, exemplifying the significance of high-throughput short read sequencing technology in the field today.

Data from short read sequencers can be mapped to a reference genome to calculate transcriptome abundance in the case where a reference is available, but because this is not the case for a great majority of non-classical model organisms, the reads must first be assembled *de novo* in effort to reconstruct a reference transcriptome for further downstream analyses. To date, various software have been implemented to perform this task (13-16), each possessing their own strengths and weaknesses, but there is no common consensus for a *de facto* standard program yet. Another open question is the impact of sequencing parameters such as read depth and read length on *de novo* assembly metrics, which is crucial not only for *de novo* transcriptome assembly but also for cost effective experiment design. Effects of read depth, or more intuitively the number of RNA-seq reads, on *de novo* transcriptome assembly is discussed in a preceding research (17). Reads from a mouse RNA-seq experiment (SRR453174) obtained from the ENCODE project (18) hosted at SRA, along with six other invertebrate species sequenced by the team were used. The results showed a capping of assembly performance at around 30 million reads with some cases of mis-assemblies reported for samples with too many reads. They conclude that 30 million reads is a good choice both from their results and experience. Effects of read length and transcriptome complexity have been discussed in another paper (19) which performed *de novo* assembly on data which were computationally simulated from reference genomes of *Saccharomyces cerevisiae*, *Mus musculus*, and *Homo sapiens*. Reads were generated with a read depth of 30 million reads as proposed in (17), and various read lengths and number of transcript isoforms. For the read length of the data, there was an observable limit to the number of reconstructed transcripts which were likely to be unique to each organism which had been known to exist for *de novo* genome assembly (20). Assembly performance of *S. cerevisiae* reads topped at a shorter length compared to *Mus musculus* and *H. sapiens* which showed similar trends. Assembly quality dropped drastically as more and more isoforms were introduced across different *de novo* transcriptome assemblers, confirming transcriptome complexity as a major factor to consider in *de novo* assembly. A factor to consider when using *de novo* assemblers with variable k option is the setting of the k parameter of the assembly software. The general algorithm used in transcriptome assembly is similar to that of genome assembly, breaking up reads into k-mers and constructing de Bruijn graphs to searching for eulerian paths representing transcripts. Smaller choices of k will make the graphs sensitive to transcripts with low expression and larger values of k resolve more repetitions and also may decrease graph complexity. Trinity uses a fixed k of 25 for the internal pipeline, which according to their report is chosen from experience (14).

Given such background our question is simple: what combination of software and parameters produce the best assembly metrics? This study aimed to elucidate the relationships of various parameters (read depth, read length, transcriptome complexity, assembly algorithm, and k) against multiple assembly metrics (number of reconstructed scaffolds, mean length and median length of the scaffolds, N50, N90, and reconstruction of well conserved genes) by comprehensively assessing RNA-seq data from SRA.

2 Materials and Methods

2.1 Data Sampling

From the massive amount of data available at SRA, samples were reduced according to the following procedures. As the first step, the metadata of read entries of transcriptome studies registered in SRA were collected from the metadata XML dump available at the FTP site of the National Center for Biotechnology Information (NCBI) (21) and analyzed. From the metadata the samples which met all of the following criteria were extracted:

a) Sequenced on an Illumina platform
b) Sequenced with a paired-end protocol
c) Sequenced with read length ≥75 bases per end

The Illumina platform is preferred over other short read sequencers for the total number of read data available at SRA and reports of superiority in *de novo* transcriptome assembly over the 454 system (22). Of the 82,251 RNA-seq read data registered in SRA 4,269 samples met all three criteria. The read data were then separated into bins based on the distribution of read depth and read length. The bins were partitioned to 5, 10, 15, 20, 30, 40, and 50 million reads for read depth, 75, 100, and 150 bases for read length, and by organism (*Homo sapiens*, *Drosophila melanogaster*, *Saccharomyces cerevisiae*, and *Arabidopsis thaliana*) for transcriptome complexity. From each bin, 5% of the read data with a minimum of 5 reads data per bin were chosen while maximizing the heterogeneity of SRA study much as possible. All data for organisms which had small amounts of data in total were used for analyses. Reads which had insufficient number of reads left after filtering were discarded during the assembly pipeline.

2.2 Preprocessing and *de novo* Assembly

Each of the reads were preprocessed by filtering out reads which had a mean Phred score <10 with a Perl script. Assembly was performed with two software: SOAPdenovo-Trans (13) which requires little computational resources as well as execution time and Trinity (14) which executes a series of sophisticated pipelines to resolve transcript isoforms but consumes significant amount of computational resources and time. Processed reads were directly fed to Trinity, and additional procedures for estimating the mean insert size was necessary for SOAPdenovo-Trans. The insert sizes were estimated using the Burrows-Wheeler Aligner (BWA) (23) to map 1 million subsampled paired reads to the reference genome of the organism. As the choice of k for the partitioning of the reads is variable in SOAPdenovo-Trans, assembly has been performed for a range of values for k starting at 21 and incremented by 6 up to 63. The assembly process has been conducted using the supercomputer system of the National Institute of Genetics (NIG).

2.3 Evaluation

Assembly scaffolds were assessed based on several metrics; namely the total number of assembled scaffolds, mean scaffold length, median scaffold length, N50, and N90. Reconstruction of conserved genes were evaluated by performing a similarity search using BLAST (24) with E value less than 1e^{-6} following the protocol described in (17) to identify the number of reconstructed Core Eukaryotic Genes (CEGs); a set of 248 EuKaryotic Orthologous Groups (KOGs) (25) defined in the Core Eukaryotic Genes Mapping Approach (CEGMA) pipeline (26). The number of CEGs reported were counted and were annotated as full length reconstructed if the assembled transcript length was within range of the longest and shortest sequences of the CEG. Both the count of BLAST hits and full length reconstructed transcripts were used for further evaluation of the data.

3 Results

3.1 Overall Statistics

Overall, there were only a small amount of data with a read depth of 150 bases available across every organism. For organisms besides *H.sapiens*, reads in one or two bins contained more than half of the number of data available for the organism in total. These data were mostly from a single large scale study following a single experimental protocol. In order to reduce the number of data used in the analysis, 5% of the *H. sapiens* data were subsampled with a minimum of 5 data per bin.

3.2 Evaluation of Assembly Results

Resulting assembly metrics were plotted against read depth, read length, and value of k (for SOAPdenovo-Trans only), and Pearson correlation coefficients for each combination of parameters and metrics were calculated. Number of scaffolds, mean scaffold length, median scaffold length, N50, and N90 were all converted to common logarithm. From the overall results no immediately obvious patterns besides a gradual improvement in some assembly metrics with the increase of read depth were observable for output from both of the assemblers (Figure 1), only SOAPdenoo-Trans (Figure 2) and only Trinity (Figure 3). The plots split by organism which are omitted for brevity, revealed a drop in the number of scaffolds and CEGMA coverage with the increase of *k* for SOAPdenovo-Trans results. An increase in the number of scaffolds following the increase of read depth and decrease in variance of CEGMA BLAST hits can be observed in *H. sapiens* plots and *D. melanogaster* plots.

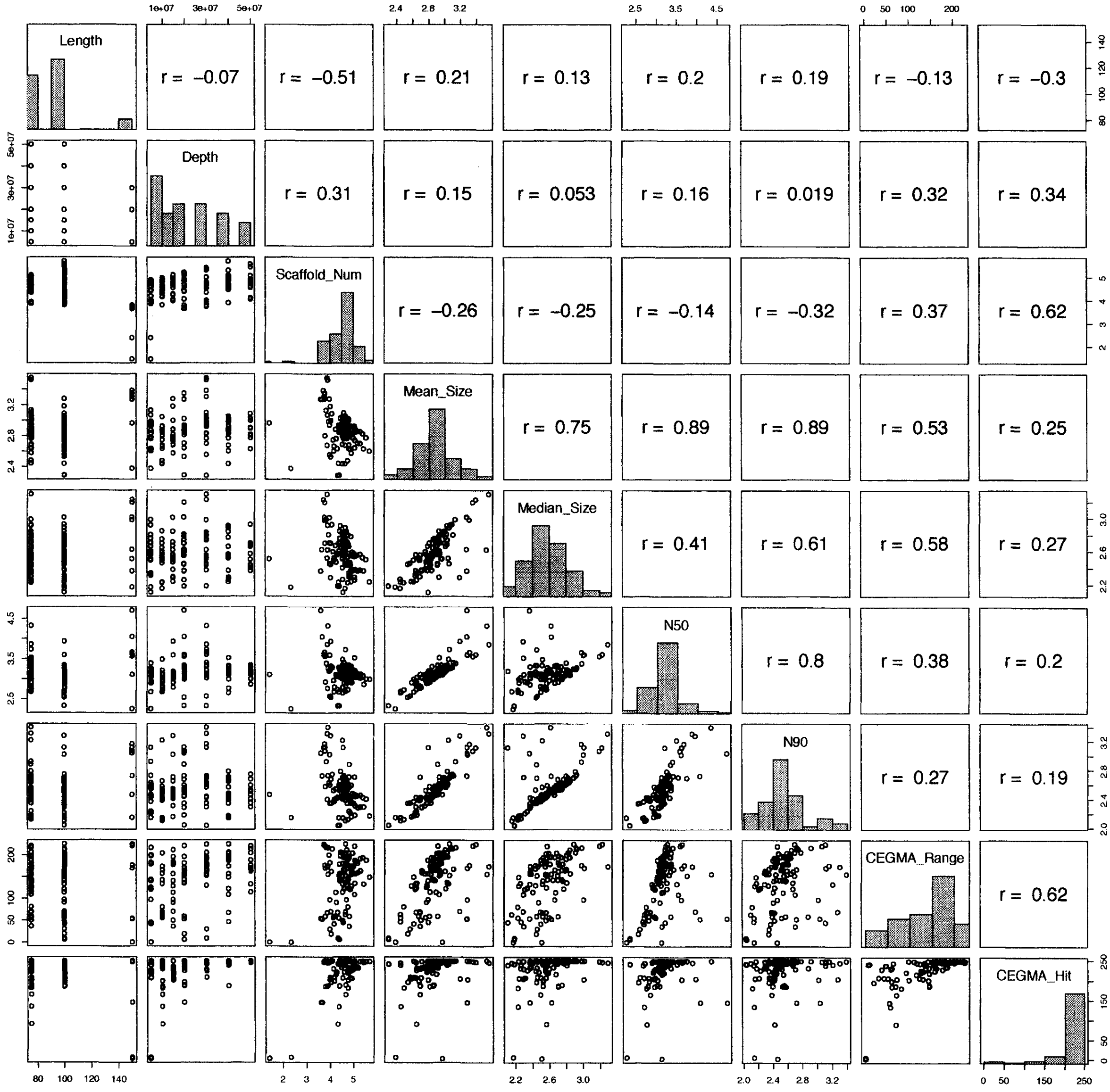

Figure 1. Plot of assembly metrics against parameters

The bottom left partition contains plots of SOAPdenovo-Trans and Trinity assembly metrics: number of scaffolds (Scaffold_Num), mean scaffold length (Mean_Size), median scaffold length (Median_Size), N50, N90, number of full length reconstructed CEGMA CEGs (CEGMA_Range), and number of CEGMA CEGs with a BLAST hit report (CEGMA_Hit), against read length and read depth. Scaffold_Num, Mean_Size, Median_Size, N50, and N90 were converted to common logarithm. The top right partition contains the Pearson correlation coefficients between variables. The diagonals represent the histograms of each variable. The value of *k* with the highest product of assembly metrics were selected as the best *k* and used in the plots.

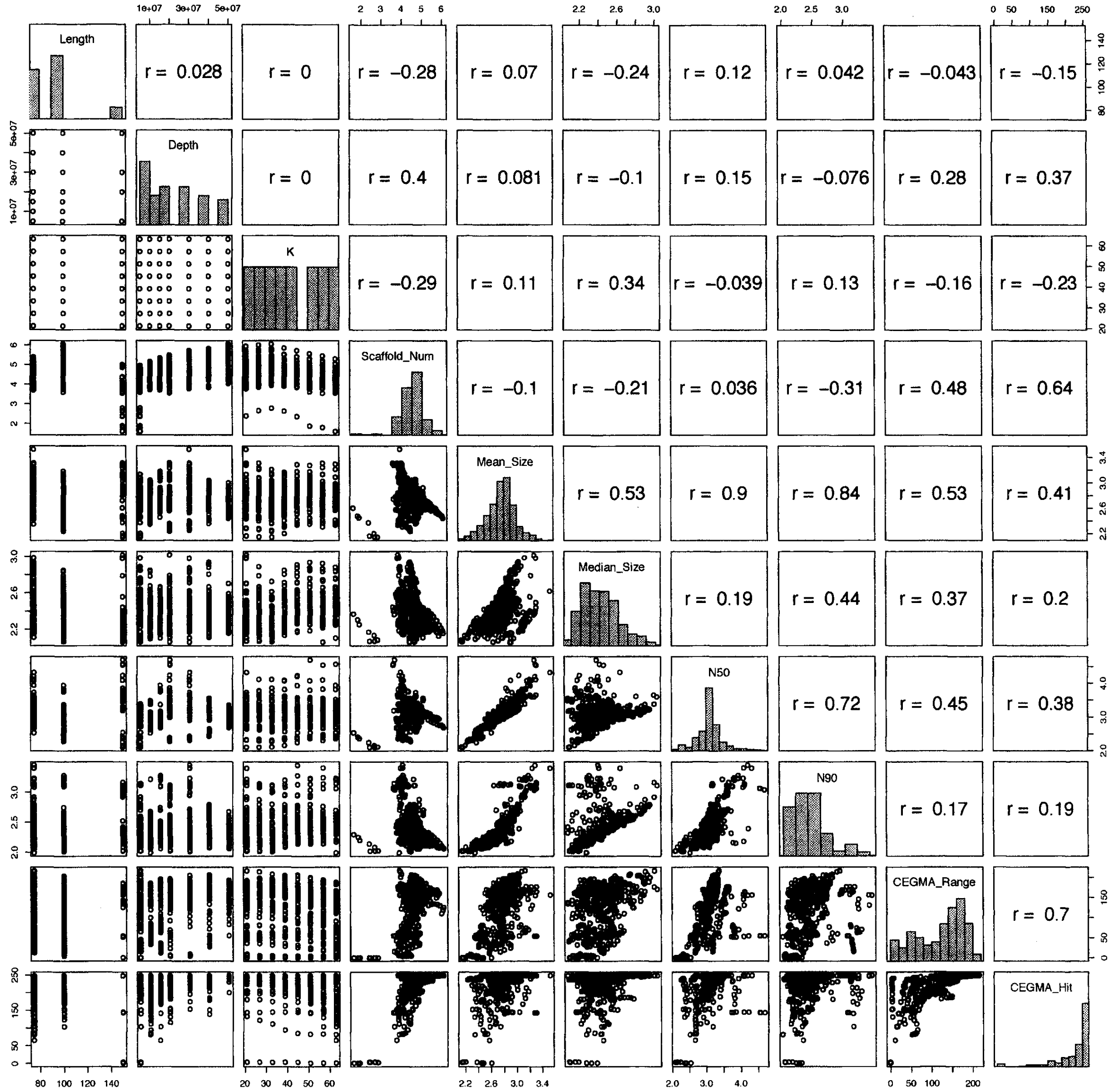

Figure 2. Plot of SOAPdenovo-Trans assembly metrics against parameters

The bottom left partition contains plots of SOAPdenovo-Trans assembly metrics: number of scaffolds (Scaffold_Num), mean scaffold length (Mean_Size), median scaffold length (Median_Size), N50, N90, number of full length reconstructed CEGMA CEGs (CEGMA_Range), and number of CEGMA CEGs with a BLAST hit report (CEGMA_Hit), against read length, read depth, and *k*. Scaffold_Num, Mean_Size, Median_Size, N50, and N90 were converted to common logarithm. The top right partition contains the Pearson correlation coefficients between variables. The diagonals represent the histograms of each variable.

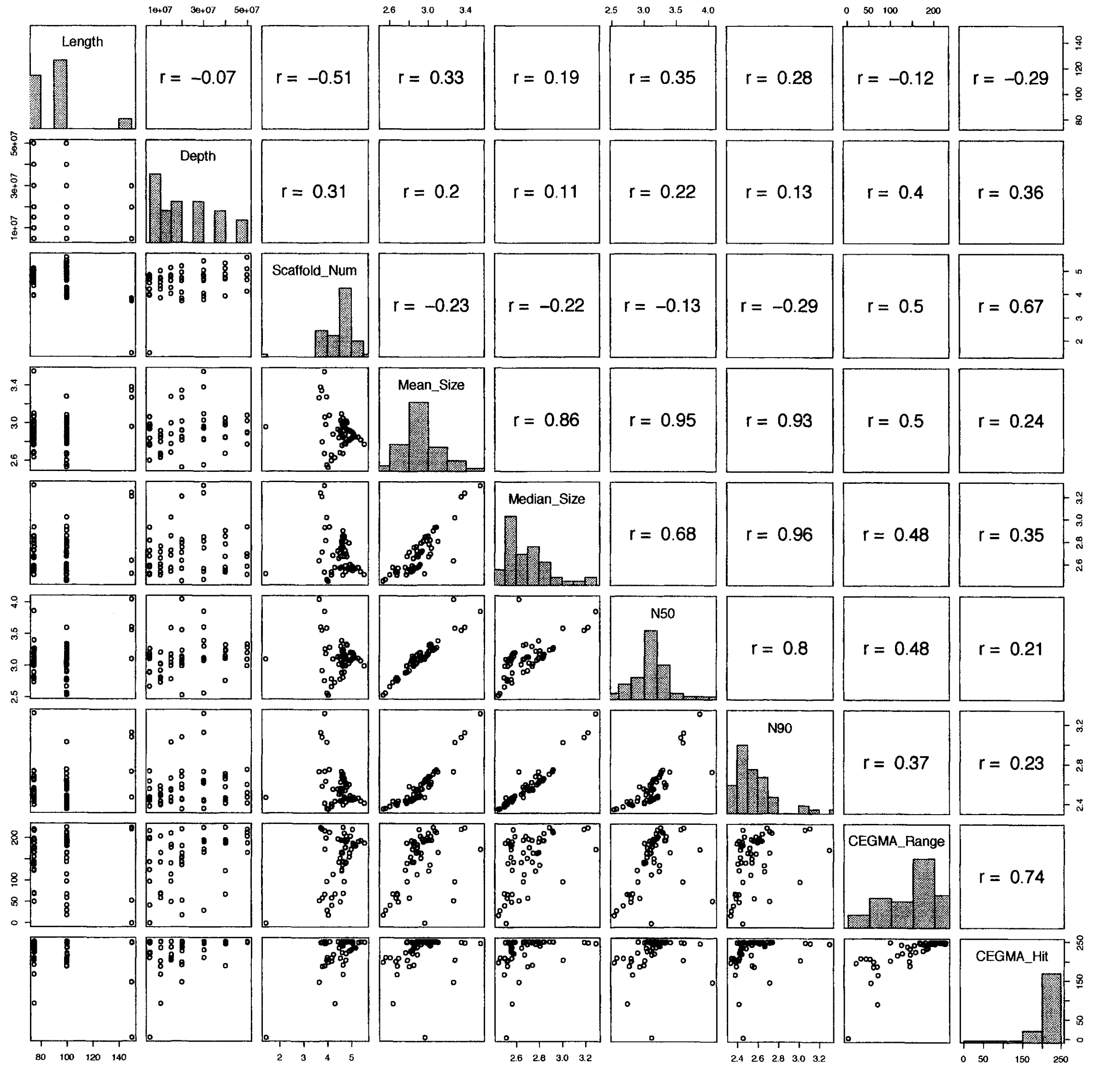

Figure 3. Plot of Trinity assembly metrics against parameters

The bottom left partition contains plots of SOAPdenovo-Trans assembly metrics: number of scaffolds (Scaffold_Num), mean scaffold length (Mean_Size), median scaffold length (Median_Size), N50, N90, number of full length reconstructed CEGMA CEGs (CEGMA_Range), and number of CEGMA CEGs with a BLAST hit report (CEGMA_Hit), against read length and read depth. Scaffold_Num, Mean_Size, Median_Size, N50, and N90 were converted to common logarithm. The top right partition contains the Pearson correlation coefficients between variables. The diagonals represent the histograms of each variable.

3.3 Cross evaluation of assembly metrics across software

Assembly metrics of scaffolds reconstructed by SOAPdenovo-Trans and Trinity were plotted against each other (Figure 4) and distributions were compared to assess performance differences (Figure 5). Of the multiple values of *k* available for SOAPdenovo-Trans, value of *k* with the highest product of assembly metric values were calculated and used. All assembly metrics showed strong correlation ($r > 0.7$) and had consistent distributions with the exception of median scaffold length, whose distribution differed significantly between the assemblers (p-value 8.965e^{-7} tested with ANOVA).

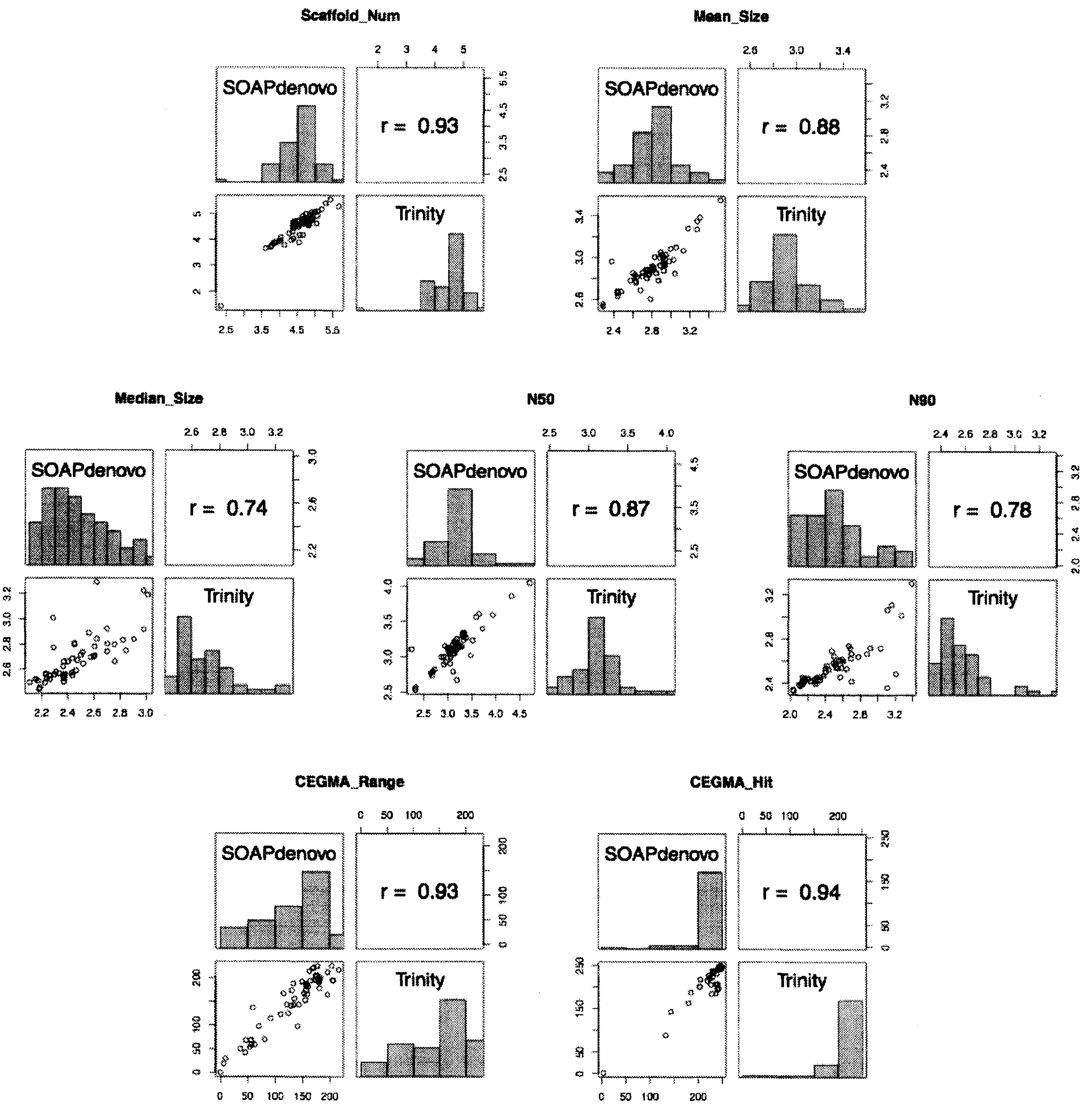

Figure 4. Plot of assembly metrics across SOAPdenovo-Trans and Trinity results

This figure shows the plot of seven assembly metics between assembly results from SOAPdenovo-Trans and Trinity. Scaffold_Num, Mean_Size, Median_Size, N50, and N90 were converted to common logarithm. All results show strong correlation.

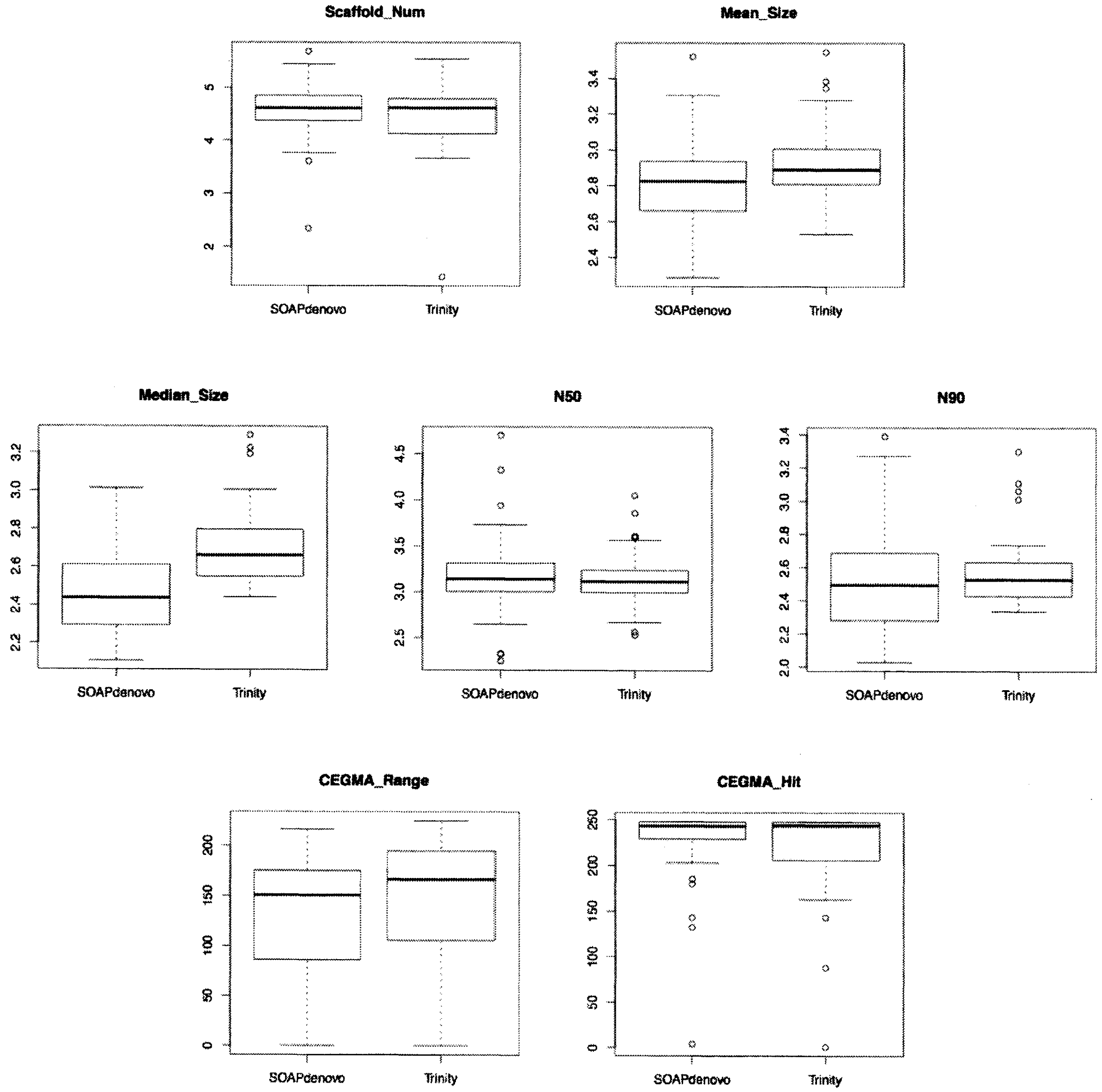

Figure 5. Distribution of assembly metrics from SOAPdenovo-Trans and Trinity results

This figure shows the distributions of the assembly metrics for the scaffolds output by SOAPdenovo-Trans and Trinity. Scaffold_Num, Mean_Size, Median_Size, N50, and N90 were converted to common logarithm. The mean of median scaffold size differed significantly (p-value 8.965e^{-7} tested with ANOVA).

3.4 Discussion

Regardless of the target organism, the number of CEGMA BLAST hits seemed to drop as k value was increased for SOAPdenovo-Trans. As the number of scaffolds and CEGMA coverage values show a moderate correlation (r = 0.62), it is intuitive to interpret that the decrease in the absolute number of scaffolds is lowering the chance of finding BLAST hits. It is logical for the number of scaffolds to have a negative relation with the choice of k, as larger k values reduce the complexity and heterogeneity of de Bruijn graphs generated in the assembly process. Such relationship between k and number of scaffolds holds true for *H. sapiens*, *D. melanogaster*, and *A. thaliana* assemblies. *S. cerevisiae* data shows the most inconsistent results, most likely due to lack of data for each bin with the exception of the 5 million read depth, 100 base read length bin. The assembly results from Trinity for *A. thaliana* reads are also inconsistent, which too is probably a result of insufficient data. For other plots, increase in the number of scaffolds assembled and CEGMA coverage can be observed as read depth accumulates, up to a point where not much improvement is apparent at around 30 to 40 million reads, which is consistent with the results of Francis, *et al.* (17). Effects of read length were not observable in all of the cases which is a surprising result as consistent trends were not observed even in between 75 base and 100 base read length data which in many cases have plenty amount of data.

Cross evaluation of the output from both assemblers yielded interesting results and provided some insights to the characteristics of each assembler. All of the assembly metrics show strong correlation of $r >$ 0.7, indicating that assembly metric trends are consistent between the two assemblers and suggest that the inconsistency in the pattern of the data is indeed not an assembler specific problem. The distributions of the assembly metrics from both software indicate performance differences between the two software. Trinity tends to have lower variance in many assembly metrics, more transcripts that is close to median length, and is slightly more suited to generating fully reconstructed KOGs compared to SOAPdenovo-Trans. On the contrary, SOAPdenovo-Trans seems to generate many scaffolds that have a BLAST hit report against the CEG dataset, and is producing scaffolds with length distributed in a wide range which explains the significantly low median length, These results suggest that Trinity output scaffolds have high quality as a trade-off with quantity, while SOAPdenovo-Trans generates large quantities of scaffolds with lower quality.

Acknowledgements

I hereby thank Prof. Kazuharu Arakawa for advisory and assistance throughout this research. Special thanks go to my advisee Konosuke Mark Ii and the members of the G-group. Finally, I would like to thank Prof. Masaru Tomita for providing me with a wonderful chance and environment to conduct this research.

References

1. McGettigan, P.A. (2013) Transcriptomics in the RNA-seq era. *Curr Opin Chem Biol*, **17**, 4-11.
2. Ozsolak, F. and Milos, P.M. (2011) RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet*, **12**, 87-98.
3. Velculescu, V.E., Zhang, L., Vogelstein, B. and Kinzler, K.W. (1995) Serial analysis of gene expression. *Science*, **270**, 484-487.
4. Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., Kodzius, R., Watahiki, A., Nakamura, M., Arakawa, T. *et al.* (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc Natl Acad Sci U S A*, **100**, 15776-15781.
5. Nakamura, M. and Carninci, P. (2004) Cap analysis gene expression: CAGE. *Tanpakushitsu Kakusan Koso*, **49**, 2688-2693.
6. Kodzius, R., Kojima, M., Nishiyori, H., Nakamura, M., Fukuda, S., Tagami, M., Sasaki, D., Imamura, K., Kai, C., Harbers, M. *et al.* (2006) CAGE: cap analysis of gene expression. *Nat Methods*, **3**, 211-222.
7. Gerhard, D.S., Wagner, L., Feingold, E.A., Shenmen, C.M., Grouse, L.H., Schuler, G., Klein, S.L., Old, S., Rasooly, R., Good, P. *et al.* (2004) The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC). *Genome Res*, **14**, 2121-2127.
8. Martin, J.A. and Wang, Z. (2011) Next-generation transcriptome assembly. *Nat Rev Genet*, **12**, 671-682.
9. Shumway, M., Cochrane, G. and Sugawara, H. (2010) Archiving next generation sequencing data. *Nucleic Acids Res.*, **38**, D870-871.
10. Leinonen, R., Sugawara, H., Shumway, M. and International Nucleotide Sequence Database, C. (2011) The sequence read archive. *Nucleic Acids Res.*, **39**, D19-21.
11. Kodama, Y., Shumway, M., Leinonen, R. and International Nucleotide Sequence Database, C. (2012) The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res.*, **40**, D54-56.
12. *DBCLS SRA*. http://sra.dbcls.jp/.
13. Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., Huang, W., He, G., Gu, S., Li, S. *et al.* (2014) SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics*, **30**, 1660-1666.
14. Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q. *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol*, **29**, 644-652.

15. Schulz, M.H., Zerbino, D.R., Vingron, M. and Birney, E. (2012) Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*, **28**, 1086-1092.
16. Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S.D., Mungall, K., Lee, S., Okada, H.M., Qian, J.Q. *et al.* (2010) De novo assembly and analysis of RNA-seq data. *Nat Methods*, **7**, 909-912.
17. Francis, W.R., Christianson, L.M., Kiko, R., Powers, M.L., Shaner, N.C. and Haddock, S.H.D. (2013) A comparison across non-model animals suggests an optimal sequencing depth for de novo transcriptome assembly. *BMC genomics*, **14**, 167.
18. Consortium, E.P., Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigo, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T. *et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799-816.
19. Chang, Z., Wang, Z. and Li, G. (2014) The impacts of read length and transcriptome complexity for de novo assembly: a simulation study. *PLoS One*, **9**, e94825.
20. Chaisson, M.J., Brinza, D. and Pevzner, P.A. (2009) De novo fragment assembly with short mate-paired reads: Does the read length matter? *Genome Res.*, **19**, 336-346.
21. *NCBI FTP*. ftp://ftp.ncbi.nlm.nih.gov.
22. Finseth, F.R. and Harrison, R.G. (2014) A comparison of next generation sequencing technologies for transcriptome assembly and utility for RNA-Seq in a non-model bird. *PloS one*, **9**, e108550.
23. Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754-1760.
24. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403-410.
25. Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.
26. Parra, G., Bradnam, K. and Korf, I. (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, **23**, 1061-1067.