On the optimal trimming of high-throughput mRNA sequence data

Matthew D MacManes[1,2*]

1 *University of California, Berkeley. Berkeley, CA 94720*
2 *California Institute of Quantitative Biology*

∗ Corresponding author: macmanes@gmail.com, Twitter: @PeroMHC

Abstract

The widespread and rapid adoption of high-throughput sequencing technologies has changed the face of modern studies of evolutionary genetics. Indeed, newer sequencing technologies, like Illumina sequencing, have afforded researchers the opportunity to gain a deep understanding of genome level processes that underlie evolutionary change. In particular, researchers interested in functional biology and adaptation have used these technologies to sequence mRNA transcriptomes of specific tissues, which in turn are often compared to other tissues, or other individuals with different phenotypes. While these techniques are extremely powerful, careful attention to data quality is required. In particular, because high-throughput sequencing is more error-prone than traditional Sanger sequencing, quality trimming of sequence reads should be an important step in all data processing pipelines. While several software packages for quality trimming exist, no general guidelines for the specifics of trimming have been developed. Here, using both simulated and empirically derived sequence data, as well as several of the available read-trimmers, I provide general recommendations regarding the optimal strength of trimming, specifically in mRNA-Seq studies. Although very aggressive quality trimming is common, this study suggests that a more gentle trimming, specifically of those nucleotides whose PHRED score <5, is superior.

Introduction

The popularity of genome–enabled biology has increased dramatically, particularly for researchers studying non-model organisms, over the last few years. For many, the primary goal of these works is to better understand the genomic underpinnings of adaptive (Linnen et al., 2013; Narum et al., 2013) or functional (Muñoz Merida et al., 2013; Hsu et al., 2012) traits. While extremely promising, the study of functional genomics in non-model organisms typically requires the generation of a reference transcriptome to which comparisons are made. Although compared to genome assembly (Bradnam et al., 2013; Earl et al., 2011). transcriptome assembly is less challenging, significant computational hurdles still exist. Amongst the most difficult of challenges involves the reconstruction of isoforms (Pyrkosz et al., 2013) and simultaneous assembly of transcripts where read coverage (=expression) varies by orders of magnitude.

These processes are further complicated by the error-prone nature of high-throughput sequencing reads. With regards to Illumina sequencing, error is distributed non-randomly over the length of the read, with the rate of error increasing from 5' to 3' end (Liu et al., 2012). These errors are overwhelmingly substitution errors (Yang et al., 2013), with the global error rate being between 1% and 3%. Although *de Bruijn* graph assemblers do a remarkable job in distinguishing error from correct sequence, sequence error does results in assembly error. While this type of error is problematic for all studies, it may be particularly troublesome for SNP-based population genetic studies. In addition to the biological concerns, sequencing read error may results in problems of a more technical importance. Because most transcriptome assemblers use a *de Bruijn* graph representation of sequence connectedness, sequencing error can dramatically increase the size and complexity of the graph, and thus increase both RAM requirements and runtime.

In addition to sequence error correction, which has been shown to improved accuracy of the *de novo* assembly MacManes and Eisen (2013), low quality (=high probability of error) nucleotides are commonly removed from the sequencing reads prior to assembly, using one of several available tools (TRIMMOMATIC (Lohse et al., 2012), FASTX TOOLKIT (http://hannonlab.cshl.edu/fastx_toolkit/index.html), BIOPIECES (http://www.biopieces.org/), SOLEXAQA (Cox et al., 2010)). These tools typically use a sliding window approach, discarding nucleotides falling below a given (user selected) average quality threshold. The trimmed dataset that remains will undoubtedly contain error, though the absolute number will surely be decreased.

Although the process of nucleotide quality trimming is commonplace in HTS analysis pipelines, it's optimal implementation has not been well defined. Though the rigor with which trimming is performed may be guided by the design of the experiment, a deeper understanding of the effects of trimming is desirable. As transcriptome-based studies of functional genomics continue to become more popular, understanding how quality trimming of mRNA-seq reads used in these types of experiments is urgently needed. Researchers currently working in these field appear to favor aggressive trimming (e.g. (Riesgo et al., 2012; Looso et al., 2013)), but this may not be optimal. Indeed, one can easily image aggressive trimming resulting in the removal of a large amout of high quality data (Even nucleotides removed with the commonly used PHRED=20 threshold are accurate 99% of the time), just as lackadaisical trimming (or no trimming) may result in nucleotide errors being incorporated into the assembled transcriptome.

Here, I attempt to provide recommendations regarding the efficient trimming of high-throughput sequence reads, specifically or mRNASeq reads from the Illumina platform. To do this, I used both simulated reads from the *Mus musculus* transcriptome, as well as an empirically derived dataset. These datasets were trimmed at various levels of stringency, assembled, then analyzed for assembly error. These results aim to guide researchers through this critical aspect of the analysis of high-throughput sequence data. While the results of this paper may not be applicable to all studies, that so many researchers are interested in the genomics of adaptation and phenotypic diversity suggests its widespread utility.

Materials and Methods

Because I was interested in understanding the effects of sequence read quality trimming on the assembly of vertebrate transcriptome assembly, we elected to simulate thirty million 100nt paired-end Illumina reads with the program Flux Simulator (Griebel et al., 2012). This dataset is fully described in a previous publication (MacManes and Eisen, 2013). In addition to this simulated dataset, an empirically derived dataset consisting of 30 million 76nt paired-end mRNA Illumina reads, which is a subset of the well-characterized 50M read set continued within the Trinity software package. Quality metrics for the raw and quality trimmed reads were generated using the program SolexaQA (Cox et al., 2010), and visualized using R (R Core Development Team, 2011).

Both simulated and empirical datasets were trimmed at varying quality thresholds using the software packages Trimmomatic (Lohse et al., 2012). Specifically, sequences were trimmed at both 5' and 3' ends using Phred ≤ 0, ≤ 1, ≤ 5, ≤ 10, and ≤ 20. Though adapter trimming is included in the functionality the Trimmomatic packages, because the simulated dataset was produced sans adapter contamination– for the purposes of comparison, the empirical dataset was cleansed of adapter before quality trimming.

Transcriptome assemblies were generated using the default settings of the program Trinity (Haas et al., 2013; Grabherr et al., 2011). Code for running Trinity is available at https://gist.github.com/macmanes/5859956. Assemblies were evaluated using a variety of different metrics. First, Blast+ (Camacho et al., 2009) was used to match assembled transcripts to their reference. TransDecoder (http://transdecoder.sourceforge.net/) was used to identify full-length transcripts. The program Blat (Kent, 2002) was used to identify and count nucleotide mismatches between reconstructed transcripts and their corresponding ref-

erence. Differences were visualized using the program R.

Results

Discussion

Acknowledgments

References

Bradnam, K.R., Fass, J.N., Alexandrov, A., Baranay, P., Bechner, M., Boisvert10, S., Chapman, J.A., Chapuis, G., Chikhi, R., Chitsaz, H., Chou, W.C., Corbeil, J., Del Fabbro, C., Docking, T.R., Durbin, R., Earl, D., Emrich, S., Fedotov, P., Fonseca, N.A., Ganapathy, G., Gibbs, R.A., Gnerre, S., Godzaridis, É., Goldstein, S., Haimel, M., Hall, G., Haussler, D., Hiatt, J.B., Ho, I.Y., Howard, J., Hunt, M., Jackman, S.D., Jaffe, D.B., Jarvis, E., Jiang, H., Kazakov, S., Kersey, P.J., Kitzman, J.O., Knight, J.R., Koren, S., Lam, T.W., Lavenier, D., Laviolette, F., Li, Y., Li, Z., Liu, B., Liu, Y., Luo, R., MacCallum, I., MacManes, M.D., Maillet, N., Melnikov, S., Vieira, B.M., Naquin, D., Ning, Z., Otto, T.D., Paten, B., Paulo, O.S., Phillippy, A.M., Pina-Martins, F., Place, M., Przybylski, D., Qin, X., Qu, C., Ribeiro, F.J., Richards, S., Rokhsar, D.S., Ruby, J.G., Scalabrin, S., Schatz, M.C., Schwartz, D.C., Sergushichev, A., Sharpe, T., Shaw, T.I., Shendure, J., Shi, Y., Simpson, J.T., Song, H., Tsarev, F., Vezzi, F., Vicedomini, R., Wang, J., Worley, K.C., Yin, S., Yiu, S.M., Yuan, J., Zhang, G., Zhang, H., Zhou, S., Korf, I.F., 2013. Assemblathon 2: evaluating *de novo* methods of genome assembly in three vertebrate species. arXiv.org arXiv:1301.5406v1.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L., 2009. BLAST+: architecture and applications. BMC Bioinformatics 10, 421.

Cox, M.P., Peterson, D.A., Biggs, P.J., 2010. SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. BMC Bioinformatics 11, 485.

Earl, D., Bradnam, K., St John, J., Darling, A., Lin, D., Fass, J., Yu, H.O.K., Buffalo, V., Zerbino, D.R., Diekhans, M., Nguyen, N., Ariyaratne, P.N., Sung, W.K., Ning, Z., Haimel, M., Simpson, J.T., Fonseca, N.A., Birol, I., Docking, T.R., Ho, I.Y., Rokhsar, D.S., Chikhi, R., Lavenier, D., Chapuis, G., Naquin, D., Maillet, N., Schatz, M.C., Kelley, D.R., Phillippy, A.M., Koren, S., Yang, S.P., Wu, W., Chou, W.C., Srivastava, A., Shaw, T.I., Ruby, J.G., Skewes-Cox, P., Betegon, M., Dimon, M.T., Solovyev, V., Seledtsov, I., Kosarev, P., Vorobyev, D., Ramirez-Gonzalez, R., Leggett, R., MacLean, D., Xia, F., Luo, R., Li, Z., Xie, Y., Liu, B., Gnerre, S., Maccallum, I., Przybylski, D., Ribeiro, F.J., Yin, S., Sharpe, T., Hall, G., Kersey, P.J., Durbin, R., Jackman, S.D., Chapman, J.A., Huang, X., DeRisi, J.L., Caccamo, M., Li, Y., Jaffe, D.B., Green, R.E., Haussler, D., Korf, I., Paten, B., 2011. Assemblathon 1: A competitive assessment of *de novo* short read assembly methods. Genome Research 21, 2224–2241.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, a., Rhind, N., di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., Regev, A., 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology 29, 644–652.

Griebel, T., Zacher, B., Ribeca, P., Raineri, E., Lacroix, V., Guigó, R., Sammeth, M., 2012. Modelling and simulating generic RNA-Seq experiments with the flux simulator. Nucleic Acids Research 40, 10073–10083.

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P., Bowden, J., Couger, M., Eccles, D., Li, B., Lieber, M., MacManes, M.D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C.N., Henschel, R., LeDuc, R.G., Friedman, N., Regev, A., 2013. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature protocols , 1–21.

Hsu, J.C., Chien, T.Y., Hu, C.C., Chen, M.J.M., Wu, W.J., Feng, H.T., Haymer, D.S., Chen, C.Y., 2012. Discovery of Genes Related to Insecticide Resistance in *Bactrocera dorsalis* by Functional Genomic Analysis of a *De Novo* Assembled Transcriptome. PLOS ONE 7, e40950.

Kent, W.J., 2002. BLAT—The BLAST-Like Alignment Tool. Genome Research 12, 656–664.

Linnen, C.R., Poh, Y.P., Peterson, B.K., Barrett, R.D.H., Larson, J.G., Jensen, J.D., Hoekstra, H.E., 2013. Adaptive Evolution of Multiple Traits Through Multiple Mutations at a Single Gene. Science (New York, NY) 339, 1312–1316.

Liu, B., Yuan, J., Yiu, S.M., Li, Z., Xie, Y., Chen, Y., Shi, Y., Zhang, H., Li, Y., Lam, T.W., Luo, R., 2012. COPE: an accurate k-mer-based pair-end reads connection tool to facilitate genome assembly. Bioinformatics (Oxford, England) 28, 2870–2874.

Lohse, M., Bolger, A.M., Nagel, A., Fernie, A.R., Lunn, J.E., Stitt, M., Usadel, B., 2012. RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. Nucleic Acids Research 40, W622–W627.

Looso, M., Preussner, J., Sousounis, K., Bruckskotten, M., Michel, C.S., Lignelli, E., Reinhardt, R., Hoeffner, S., Krueger, M., Tsonis, P.A., Borchardt, T., Braun, T., 2013. A *de novo* assembly of the newt transcriptome combined with proteomic validation identifies new protein families expressed during tissue regeneration. Genome Biology 14, R16.

MacManes, M.D., Eisen, M.B., 2013. Improving transcriptome assembly through error correction of high-throughput sequence reads. arXiv.org arXiv:1304.0817v2.

Muñoz Merida, A., Gonzalez-Plaza, J.J., Canada, A., Blanco, A.M., Garcia-Lopez, M.d.C., Rodriguez, J.M., Pedrola, L., Sicardo, M.D., Hernandez, M.L., De la Rosa, R., Belaj, A., Gil-Borja, M., Luque, F., Martinez-Rivas, J.M., Pisano, D.G., Trelles, O., Valpuesta, V., Beuzon, C.R., 2013. De Novo Assembly and Functional Annotation of the Olive (*Olea europaea*) Transcriptome. DNA Research 20, 93–108.

Narum, S.R., Campbell, N.R., Meyer, K.A., Miller, M.R., Hardy, R.W., 2013. Thermal adaptation and acclimation of ectotherms from differing aquatic climates. Molecular Ecology , 1–8.

Pyrkosz, A.B., Cheng, H., Brown, C.T., 2013. RNA-Seq Mapping Errors When Using Incomplete Reference Transcriptomes of Vertebrates. http://arxiv.org/abs/1303.2411v1 arXiv:1303.2411v1.

R Core Development Team, F., 2011. R: A Language and Environment for Statistical Computing .

Riesgo, A., Perez-Porro, A.R., Carmona, S., Leys, S.P., Giribet, G., 2012. Optimization of preservation and storage time of sponge tissues to obtain quality mRNA for next-generation sequencing. Molecular Ecology Resources 12, 312–322.

Yang, X., Chockalingam, S.P., Aluru, S., 2013. A survey of error-correction methods for next-generation sequencing. Briefings In Bioinformatics 14, 56–66.

Figures & Tables

Figure 1

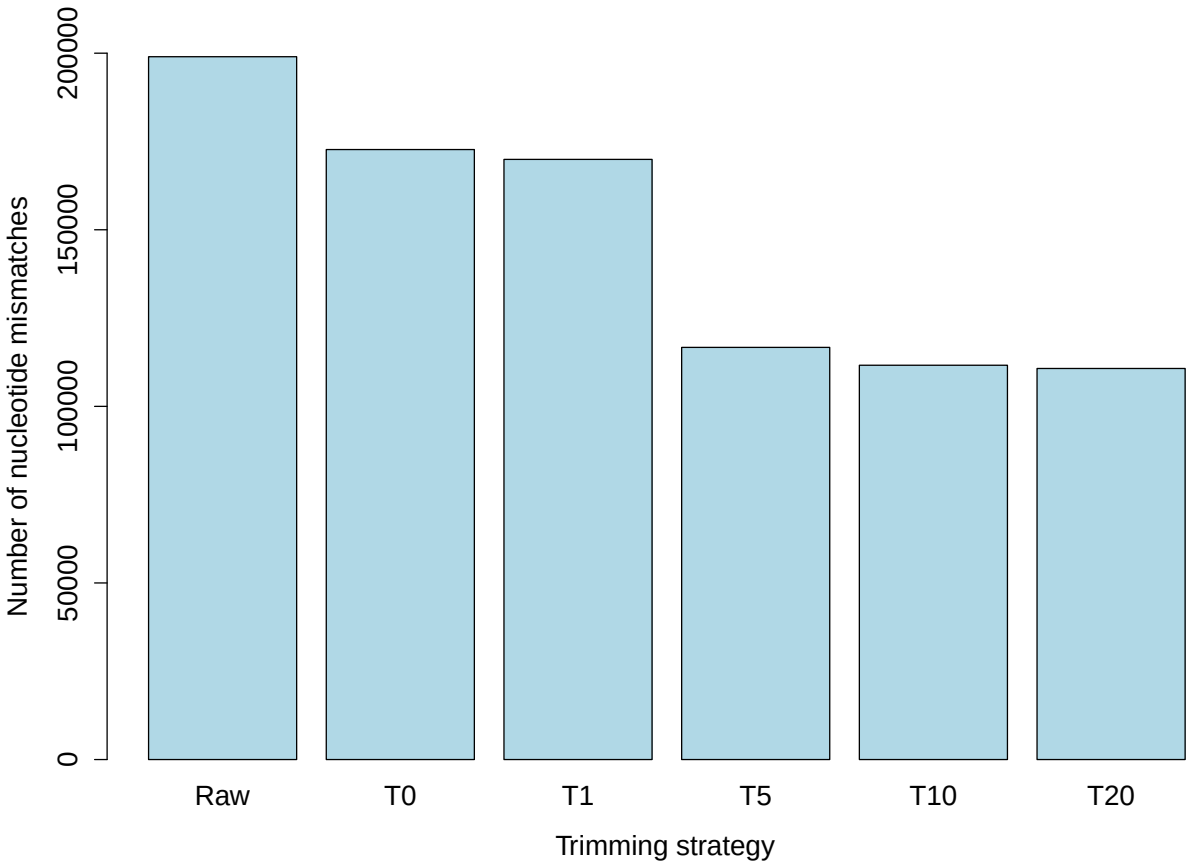

Fig. 1

Figure 2

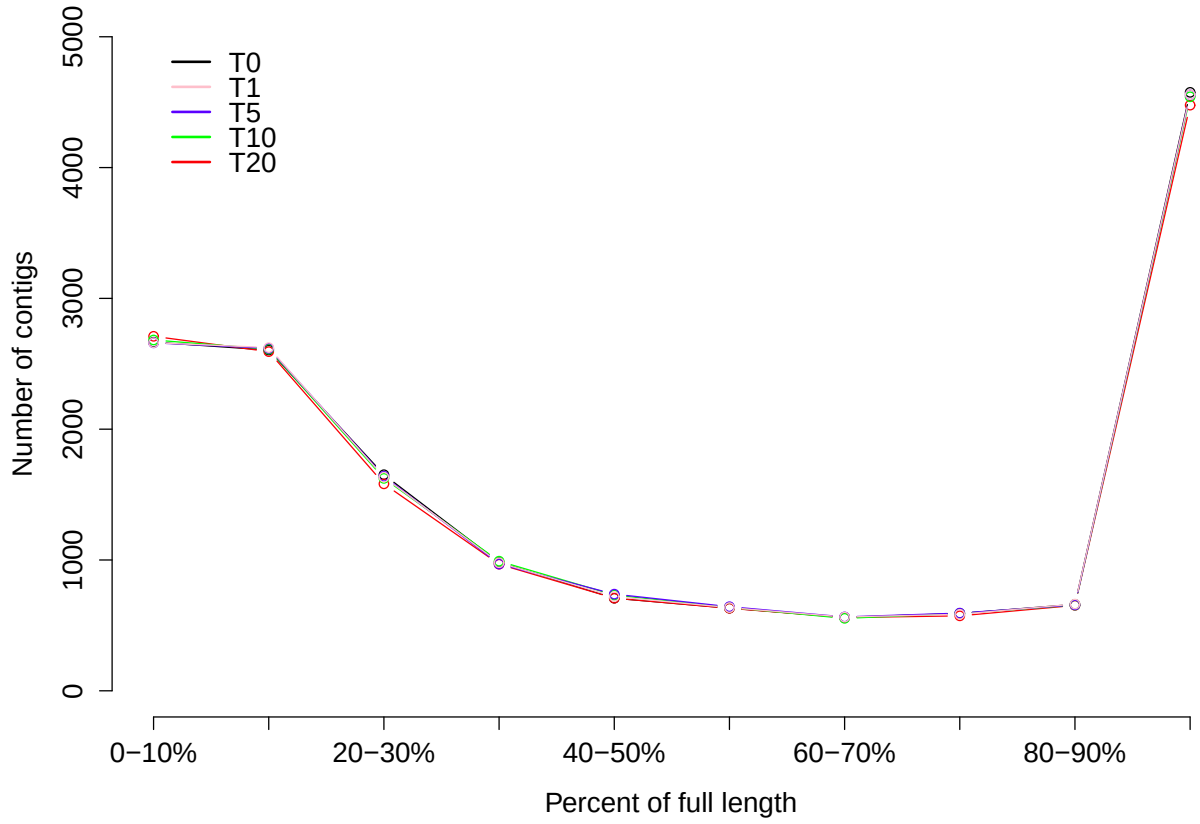
T0
T1
T5
T10
T20
Number of contigs
0
1000
2000
3000
4000
5000
0−10%
20−30%
40−50%
60−70%
80−90%
Percent of full length

Fig. 2