On the optimal trimming of high-throughput mRNA sequence data

Matthew D MacManes[1,2,3*]

1 *University of New Hampshire. Durham, NH 03824*
2 *Department of Molecular, Cellular & Biomedical Sciences*
3 *Hubbard Center for Genome Studies*

* 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 empirically derived sequence data, 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 <2 or <5, is optimal for most studies across a wide variety of metrics.

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 (Hsu et al., 2012; Muñoz-Mérida et al., 2013) 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 (MacManes and Eisen, 2013). 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 sequencing read 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. (Looso et al., 2013; Riesgo et al., 2012)), 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 for mRNASeq reads from the Illumina platform. To do this, I used a publicly available dataset containing Illumina reads derived from *Mus musculus*. Subsets of these data (10 million, 20 million, 50 million, 75 million, 100 million reads) were randomly chosen, trimmed to various levels of stringency, assembled then analyzed for assembly error and content 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, I elected analyze a publicly available (SRR797058) paired-end Illumina read dataset. This dataset is fully described in a previous publication (Han et al., 2013), and contains 232 million paired-end 100nt Illumina reads. To investigate how sequencing depth influences the choice of trimming level, reads data were randomly subsetted into 10 million, 20 million, 50 million, 75 million, 100 million read datasets.

Read datasets were trimmed at varying quality thresholds using the software package TRIMMOMATIC (Lohse et al., 2012), which was selected as is appears to be amongst the most popular of read trimming tools. Specifically, sequences were trimmed at both 5' and 3' ends using PHRED $=0$ (adapter trimming only), ≤ 2, ≤ 5, ≤ 10, and ≤ 20. Transcriptome assemblies were generated for each dataset using the default settings of the program TRINITY (Grabherr et al., 2011; Haas et al.,

2013). Assemblies were evaluated using a variety of different metrics, many of them comparing assemblies to the complete collection of *Mus* cDNA's, available at http://useast.ensembl.org/info/data/ftp/index.html.

Quality trimming may have substantial effect on assembly quality, and as such, I sought to identify high quality transcriptome assemblies. Assemblies with few nucleotide errors relative to a known reference may indicate high quality. The program BLAT (Kent, 2002) was used to identify and count nucleotide mismatches between reconstructed transcripts and their corresponding reference. To eliminate spurious short matches between query and template inflating estimates of error, only unique transcripts that covered more than 90% of their reference sequence were used. Another potential assessment of assembly quality may be related to the number of paired-end sequencing reads that concordantly map to the assembly. As the number of reads concordantly mapping increased, so does assembly quality. To characterize this, I mapped raw (adapter trimmed) sequencing reads to each assembly using Bowtie2 (Trapnell et al., 2010).

Aside from these metrics, measures of assembly content were also assayed. Here, open reading frames (ORFs) were identified using the program TRANSDECODER (http://transdecoder.sourceforge.net/), and were subsequently translated into amino acid sequences. The larger the number of complete open reading frames (containing both start and stop codons) the better the assembly. Lastly, unique transcripts were identified using the blastP program within the BLAST+ package (Camacho et al., 2009). Blastp hits were retained only if the sequence similarity was >80% over at least 100 amino acids. As the number of transcripts matching a given reference increases, so may assembly quality. Code for performing the subsetting, trimming, assembly, peptide and ORF prediction and blast analyses can be found in the following Github folder https://github.com/macmanes/trimming_paper/tree/master/scripts.

Results

Quality trimming of sequence reads had a relatively large on the total number of errors contained in the final assembly (Figure 1), which was reduced by between 9 and 26% when comparing the assemblies of untrimmed versus PHRED=20 trimmed sequence reads. Most of the improvement in

accuracy is gained when trimming at the level of PHRED=5 or greater, with modest improvements potentially garnered with more aggressive trimming at certain coverage levels (Table 1).

Figure 1

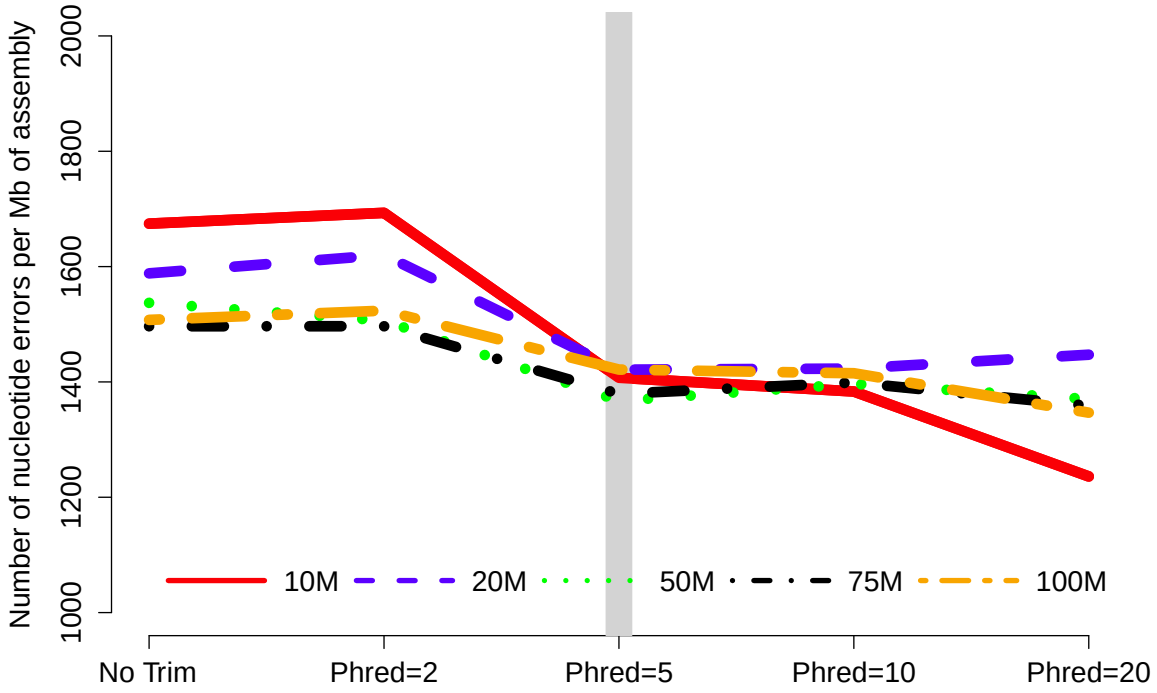

Figure 1. The number of nucleotide errors contained in the final transcriptome assembly, normalized to assembly size, is related to the strength of quality trimming (Trimming of nucleotides whose error scores are: PHRED >20, 10, 5, 2, or no trimming, though most benefits are observed at a modest level of trimming. This patterns is largely unchanged with varying depth of sequencing coverage (10 million to 100 million sequencing reads). Trimming at PHRED $= 5$ may be optimal, given the potential untoward effects of more stringent quality trimming.

In addition to looking at nucleotide errors, assembly quality may be measured by the the proportion of sequencing reads that map concordantly to a given transcriptome assembly (Hunt et al., 2013). As such, the analysis of assembly quality includes study of the mapping rates. Here, we found small but significant effects of trimming. Specifically, assembling with aggressively quality trimmed reads decreased the proportion of reads that map concordantly to a given contig (Figure 2). The pattern is

particularly salient with trimming at the PHRED = 20 level. Here, several hundred thousand fewer reads mapped compared to mapping against the assembly of untrimmed reads.

Figure 2

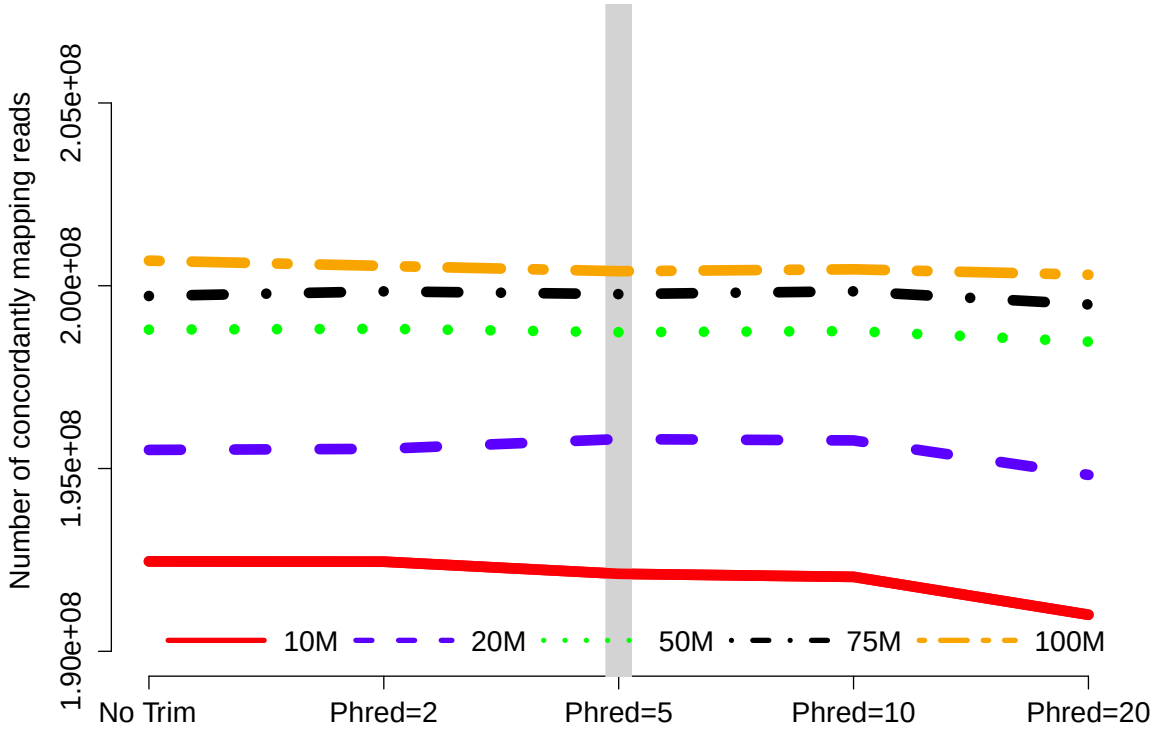

Figure 2. The number of concordantly mapping reads was reduced by trimming. The pattern is particularly salient with trimming at PHRED=20 which was always associated with the successful mapping of hundreds of thousands of fewer reads.

Analysis of assembly content painted a similar picture, with trimming having a relatively small, though tangible effect. The number of BLAST+ matches decreased with stringent trimming (Figure 3), with trimming at PHRED=20 associated with particularly poor performance.

Figure 3

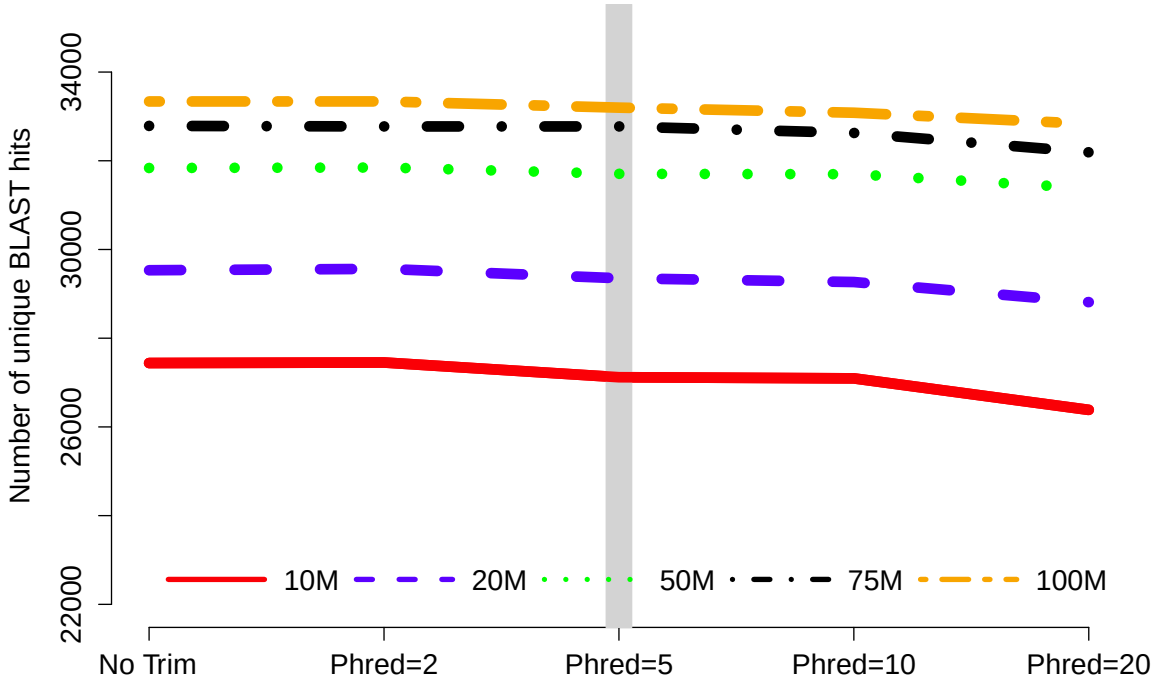

Figure 3. The number of unique BLAST matches contained in the final transcriptome assembly is related to the strength of quality trimming for any of the studied sequencing depths. A gentle trimming strategy typically yielded the most number of unique matches, while trimming at PHRED=20 was always associated with much poorer assembly content

When counting complete open reading frames, low and moderate coverage datasets (10M, 20M, 50M) were all worsened by aggressive trimming (Figure 4). Trimming at PHRED=20 was the most poorly performing level at all read depths.

Figure 4

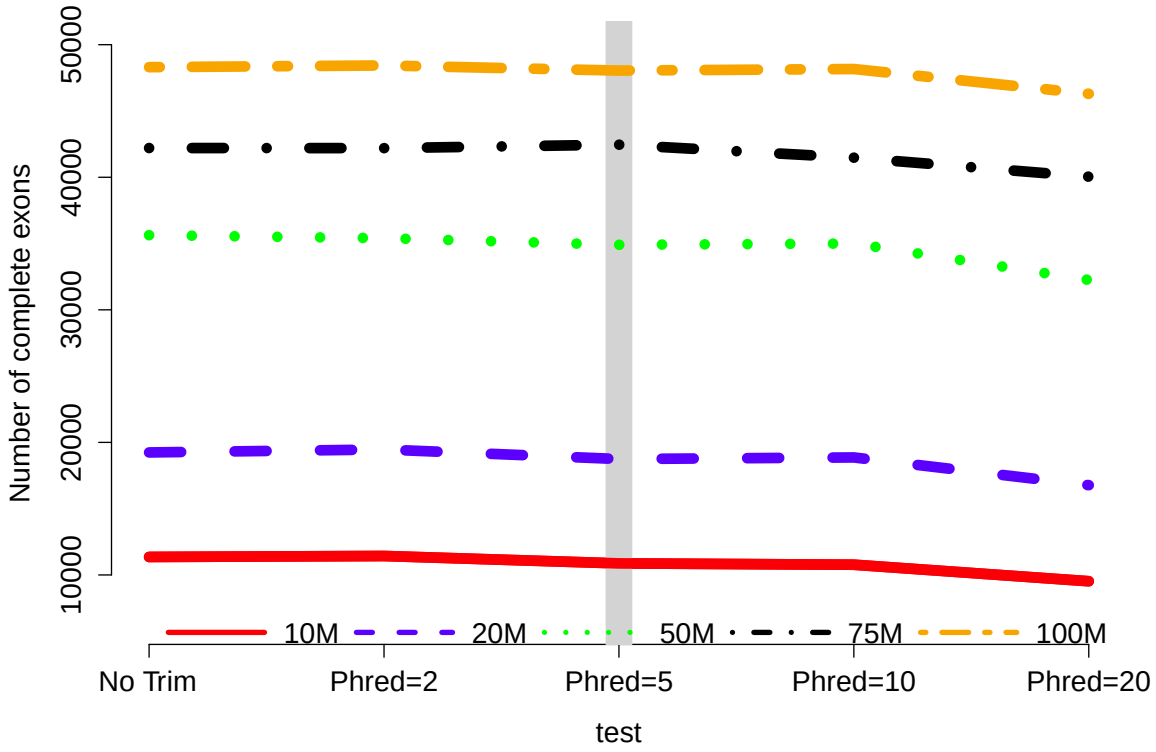

Figure 4. The number of complete exons contained in the final transcriptome assembly is not strongly related to the strength of quality trimming for any of the studies sequencing depths, though trimming at PHRED=20 was always associated with fewer identified exons.

Of note, all assembly files will be deposited in Dryad upon acceptance for publication. Until then, they can be accessed via https://www.dropbox.com/sh/oiem0v5jgr5c5ir/TYQdGcpYwP

Discussion

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 seems to vary, there seems to be a bias towards stringent trimming (Ansell et al., 2013; Barrett and Davis, 2012; Straub et al., 2013; Tao et al., 2013). This study provides strong evidence that stringent quality trimming of nucleotides whose quality scores are ≤ 20 results in a poorer transcriptome assembly across the majority metrics. Instead, researchers interested in assembling transcriptomes *de novo* should elect for a much more gentle quality trimming, or no trimming at all. Table 1 summarizes

my finding across all experiments, where the numbers represent the trimming level that resulted in the most favorable result. What is apparent, is that for typically-sized datasets, trimming at PHRED=2 or PHRED=5 optimizes assembly quality. The exception to this rule appears to be in studies where the identification of SNP markers from high (or very low) coverage datasets is the primary goal.

Table 1

DATASET SIZE	ERROR	MAP	ORF	BLAST
10M	20	0	2	2
20M	5	5	2	2
50M	5	5	5	2
75M	20	10	5	0
100M	20	0	2	2

Table 1. The PHRED trimming levels that resulted in optimal assemblies across the 4 metrics tested in the different size datasets. Error= the number of nucleotide errors in the assembly. Map= the number of concordantly mapped reads. ORF= the number of ORFs identified. BLAST= the number of unique BLAST hits.

The results of this study were surprising. In fact, much of my own work assembling transcriptomes included a vigorous trimming step. That trimming had generally small effects, and even negative effects when trimming at PHRED=20 was unexpected. To understand if trimming changes the distribution of quality scores along the read, we generated plots with the program SolexaQA (Cox et al., 2010). Indeed, the program modifies the distribution of PHRED scores in the predicted fashion yet downstream effects are minimal. This should be interpreted as speaking to the performance of the the bubble popping algorithms included in TRINITY and other *de Bruijn* assemblers.

The results presented here stem from the analysis of a single Illumina dataset and specific properties of that dataset may have biased the results. This dataset was selected from several evaluated SRA datasets for it's 'typical' error profile. The preliminary analysis of a 10 million read subset of another typical dataset were concordant with those presented here. Taken together, this suggests that the results presented here do not appear to be dependent on the particulars of this dataset, but instead are

typical of Illumina mRNAseq datasets.

WHAT IS MISSING IN TRIMMED DATASETS? — The question of differences in recovery of specific contigs is a difficult question to answer. Indeed, these relationships are complex, and could involve a stochastic process, or be related to differences in expression (low expression transcripts lost in trimmed datasets) or length (longer contigs lost in trimmed datasets). To investigate this, I attempted to understand how contigs recovered in the 10 million reads untrimmed dataset but not in the PHRED=20 trimmed dataset were different. Using the information on FPKM and length generated by the program EXPRESS, it was clear that the transcripts unique to the untrimmed dataset were more lowly expressed (mean FPKM=3.2) when compared to the entire untrimmed dataset (mean FPKM=11.1; t = -2.2255, df = 70773, p-value = 0.02605). Of note, a similar result was found when using the non-parametric Wilcoxan test (W = 18591566, p-value = 7.184e-13).

Turning my attention to length, when comparing uniquely recovered transcripts to the entire untrimmed dataset of 10 million reads, it appears to be the shorter contigs (mean length 857nt versus 954nt; t = -2.1285, df = 650.05, p-value = 0.03367, W = 26790212, p-value <2.2e-16) that are differentially recovered in the untrimmed dataset relative to the PHRED=20 trimmed dataset.

EFFECTS OF COVERAGE — Though the experiment was not designed to evaluate the effects of sequencing depth on assembly, the data speak well to this issue. Contrary to other studies, suggesting that 30 million paired end reads were sufficient to cover eukaryote transcriptomes (Francis et al., 2013), the results of the current study suggest that assembly content was more complete as sequencing depth increased; a pattern that holds at all trimming levels. Though the suggested 30 million read depth was not included in this study, all metrics, including the number of assembly errors was dramatically reduced, and the number of exons, and BLAST hits were increased as read depth increased. While generating more sequence data is expensive, given the assembled transcriptome reference often forms the core of future studies, this investment may be warranted.

In summary, the process of nucleotide quality trimming is commonplace in HTS analysis pipelines, but it's optimal implementation has not been well defined. A very aggressive strategy, where sequence reads are trimmed when PHRED scores fall below 20 is common. My analyses suggest that for studies

whose primary goal is transcript discovery, that a more gentle trimming strategy (*e.g.* PHRED=2 or PHRED=5) that removes only the lowest quality bases is optimal. In particular, it appears as if the shorter and more lowly expressed transcripts are particularly vulnerable to loss in studies involving more harsh trimming. The one potential exception to this general recommendation may be in studies of population genomics, where deep sequencing is leveraged to identify SNPs. Here, a more stringent trimming strategy may be warranted.

Acknowledgments

References

Ansell, B.R.E., Schnyder, M., Deplazes, P., Korhonen, P.K., Young, N.D., Hall, R.S., Mangiola, S., Boag, P.R., Hofmann, a., Sternberg, P.W., Jex, A.R., Gasser, R.B., 2013. Insights into the immuno-molecular biology of *Angiostrongylus vasorum* through transcriptomics-Prospects for new interventions. Biotechnology Advances .

Barrett, C.F., Davis, J.I., 2012. The plastid genome of the mycoheterotrophic *Corallorhiza striata* (Orchidaceae) is in the relatively early stages of degradation. American Journal of Botany 99, 1513–1523.

Bradnam, K.R., Fass, J.N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., Boisvert, 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, E., 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., 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., Vieira, B.M., 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. GigaScience 2, 10.

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.

Francis, W.R., Christianson, L.M., Kiko, R., Powers, M.L., Shaner, N.C., D Haddock, S.H., 2013. A comparison across non-model animals suggests an optimal sequencing depth for *de novo* transcriptome assembly. BMC Genomics 14, 167.

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.

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., 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.D., Friedman, N., Regev, A., 2013. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature protocols 8, 1494–1512.

Han, H., Irimia, M., Ross, P.J., Sung, H.K., Alipanahi, B., David, L., Golipour, A., Gabut, M., Michael, I.P., Nachman, E.N., Wang, E., Trcka, D., Thompson, T., O'Hanlon, D., Slobodeniuc, V., Barbosa-Morais, N.L., Burge, C.B., Moffat, J., Frey, B.J., Nagy, a., Ellis, J., Wrana, J.L., Blencowe, B.J., 2013. MBNL proteins repress ES-cell-specific alternative splicing and reprogramming. Nature 498, 241–245.

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.

Hunt, M., Kikuchi, T., Sanders, M., Newbold, C., Berriman, M., Otto, T.D., 2013. REAPR: a universal tool for genome assembly evaluation. Genome Biology 14, R47.

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, W622–7.

Looso, M., Preussner, J., Sousounis, K., Bruckskotten, M., Michel, C.S., Lignelli, E., Reinhardt, R., Höffner, S., Krüger, 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. PeerJ 1, e113.

Muñoz-Mérida, A., González-Plaza, J.J., Cañada, a., Blanco, A.M., García-López, M.d.C., Rodríguez, J.M., Pedrola, L., Sicardo, M.D., Hernández, M.L., De la Rosa, R., Belaj, A., Gil-Borja, M., Luque, F., Martínez-Rivas, J.M., Pisano, D.G., Trelles, O., Valpuesta, V., Beuzón, 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 22, 3090–3097.

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

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.

Straub, S.C.K., Cronn, R.C., Edwards, C., Fishbein, M., Liston, A., 2013. Horizontal transfer of DNA from the mitochondrial to the plastid genome and its subsequent evolution in milkweeds (Apocynaceae). Genome Biology and Evolution 5, 1872–1885.

Tao, T., Zhao, L., Lv, Y., Chen, J., Hu, Y., Zhang, T., Zhou, B., 2013. Transcriptome Sequencing and Differential Gene Expression Analysis of Delayed Gland Morphogenesis in *Gossypium australe* during Seed Germination. PLOS one .

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., Pachter, L., 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature Biotechnology 28, 511–515.

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