On the optimal trimming of high-throughput mRNA sequence data

Matthew D. MacManes [1,2]*

[1] Department of Molecular, Cellular and Biomedical Sciences, University of New Hampshire, Durham, NH, USA
[2] Hubbard Center for Genome Studies, Durham, NH, USA

The widespread and rapid adoption of high-throughput sequencing technologies has afforded researchers the opportunity to gain a deep understanding of genome level processes that underlie evolutionary change, and perhaps more importantly, the links between genotype and phenotype. 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.

Keywords: quality trimming, quality control, illumina, RNAseq, assembly error

INTRODUCTION

The popularity of genome-enabled biology has increased dramatically over the last few years. While researchers involved in the study of model organisms have had the ability to leverage the power of genomics for nearly a decade, this power is only now available for the study of non-model organisms. For many, the primary goal of these newer 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 transcriptome assembly is less challenging (Earl et al., 2011; Bradnam et al., 2013), significant computational hurdles still exist. Amongst the most difficult of challenges in transcriptome assembly involves the reconstruction of isoforms (Pyrkosz et al., 2013), simultaneous assembly of transcripts where read coverage (=expression) varies by orders of magnitude, and overcoming biases related to random hexamer (Hansen et al., 2010) and GC content (Dohm et al., 2008).

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 result 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 improve 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), or bio-pieces (http://www.biopieces.org/)]. These tools typically use either a sliding window approach, discarding nucleotides falling below a given (user selected) average quality threshold, or trimming of low-quality nucleotides at one or both ends of the sequencing read. Though the absolute number will surely be decreased in the trimmed dataset, aggressive quality trimming may remove a substantial portion of the total read dataset, which in transcriptome studies may disproportionately effect lower expression transcripts.

Although the process of nucleotide quality trimming is commonplace, particularly in the assembly-based HTS analysis pipelines [e.g., SNP development (Milano et al., 2011; Helyar

et al., 2012), functional studies (Ansell et al., 2013; Bhardwaj et al., 2013), and more general studies of transcriptome characterization (MacManes and Lacey, 2012; Liu et al., 2013)], its 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 this 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 amount 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 provide recommendations regarding the efficient trimming of high-throughput sequence reads, specifically for mRNASeq reads from the Illumina platform. To do this, I used publicly available datasets containing Illumina reads derived from *Mus musculus*. Subsets of these data (10, 20, 50, 75, 100 million reads) were randomly chosen, trimmed to various levels of stringency, assembled then analyzed for assembly error and content. In addition to this, I develop a set of metrics that may be generally useful in evaluating the quality of transcriptome assemblies. 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, particularly in non-model organisms suggests its widespread utility.

MATERIALS AND METHODS

Because I was interested in understanding the effects of sequence read quality trimming on the quality of vertebrate transcriptome assembly, I elected to 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, 20, 50, 75, 100 million read datasets. To test the robustness of my findings, I evaluated a second dataset (SRR385624, Macfarlan et al., 2012) as well as a technical replicate of the primary dataset, both at the 10 M read dataset size.

Read datasets were trimmed at varying quality thresholds using the software package TRIMMOMATIC version 0.30 (Lohse et al., 2012), which was selected as it 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. Other parameters (MINLEN = 25, ILLUMINACLIP = barcodes.fa:2:40:15, SLIDINGWINDOW size = 4) were held constant. Transcriptome assemblies were generated for each dataset using the default settings (except group_pairs_distance flag set to 999) of the program

TRINITY R2013-02-25 (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 v34 (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. Next, because kmers represent the fundamental unit of assembly, kmers ($k = 25$) were counted for each dataset using the program Jellyfish v1.1.11 (Marçais and Kingsford, 2011). 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 the full dataset (not subsampled) of adapter trimmed sequencing reads to each assembly using Bowtie2 v2.1.0 (Trapnell et al., 2010) using default settings, except for maximum insert size (-X 999) and number of multiple mappings (-k 30).

Aside from these metrics, measures of assembly content were also assayed. Here, open reading frames (ORFs) were identified using the default settings of the program TRANSDECODER R20131110 (http://transdecoder.sourceforge.net/), and were subsequently translated into amino acid sequences, both using default settings. The larger the number of complete ORFs (containing both start and stop codons) the better the assembly. Next, unique transcripts were identified using the blastP program within the BLAST+ package version 2.2.28 (Camacho et al., 2009). Blastp hits were retained only if the sequence similarity was >80% over at least 100 amino acids, and *e*-value < 10^{-10}. As the number of transcripts matching a given reference increases, so may assembly quality. Lastly, because the effects of trimming may vary with expression, I estimated expression (e.g., FPKM) for each assembled contig using default settings of the the program EXPRESS v1.5.0 (Roberts and Pachter, 2013) and the BAM file produced by Bowtie2 as described above. 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/recreate_ms_analyses/scripts

RESULTS

Quality trimming of sequence reads had a relatively large effect 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**).
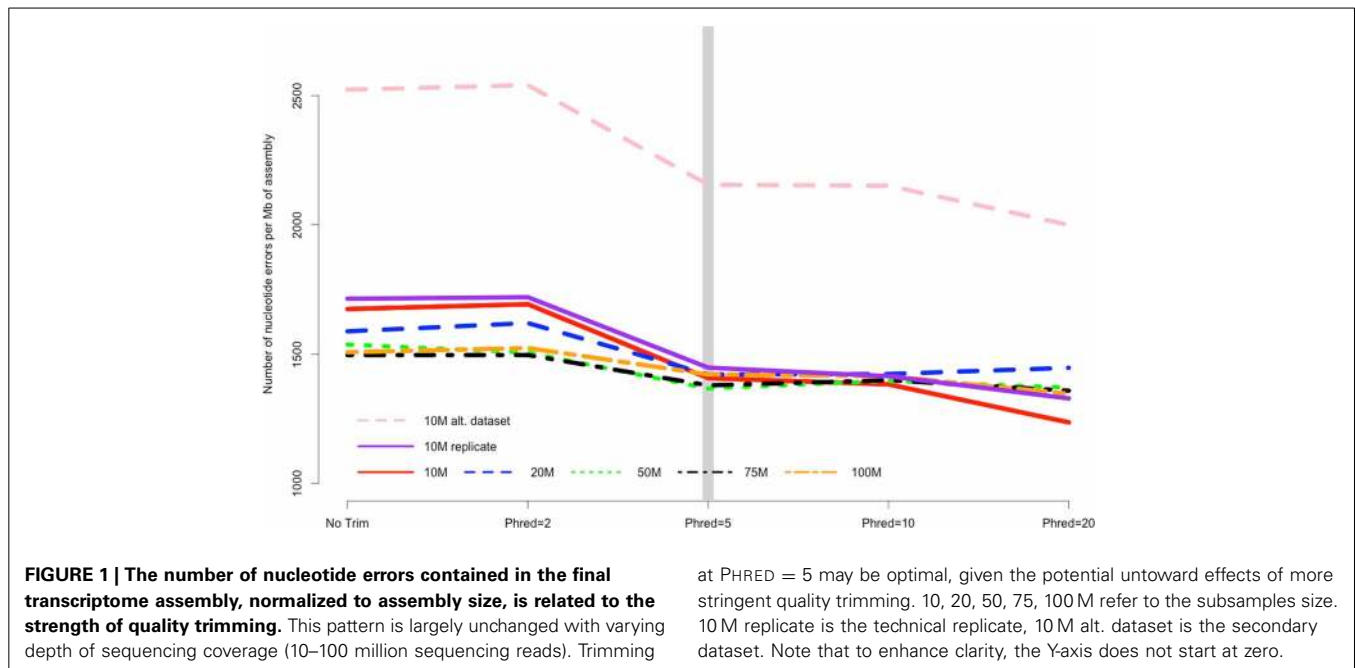
FIGURE 1 | The number of nucleotide errors contained in the final transcriptome assembly, normalized to assembly size, is related to the strength of quality trimming. This pattern is largely unchanged with varying depth of sequencing coverage (10–100 million sequencing reads). Trimming at PHRED = 5 may be optimal, given the potential untoward effects of more stringent quality trimming. 10, 20, 50, 75, 100 M refer to the subsamples size. 10 M replicate is the technical replicate, 10 M alt. dataset is the secondary dataset. Note that to enhance clarity, the Y-axis does not start at zero.

Table 1 | The PHRED trimming levels that resulted in optimal assemblies across the 4 metrics tested in the different size datasets.

Dataset size	Error	Map	ORF	BLAST
10 M	20	0	2	2
10 M REP.	20	2	2	2
10 M ALT	20	2	0	0
20 M	5	5	2	2
50 M	5	10	5	2
75 M	20	10	5	0
100 M	20	0	2	2

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. 10 M rep. is the technical replicate, 10 M alt. is the secondary dataset.

In *de Bruijn* graph-based assemblers, the kmer is the fundamental unit of assembly. Even in transcriptome datasets, unique kmers are likely to be formed as a result of sequencing error, and therefore may be removed during the trimming process. **Figure 2A** shows the pattern of unique kmer loss across the various trimming levels and read datasets. What is apparent, is that trimming at PHRED = 5 removes a large fraction of unique kmers, with either less- or more-aggressive trimming resulting in smaller effects. In contrast to the removal of unique kmers, those kmers whose frequency is >1 are more likely to be real, and therefore should be retained. **Figure 2B** shows that while PHRED = 5 removes unique kmers, it may also reduce the number of non-unique kmers, which may hamper the assembly process.

In addition to looking at nucleotide error and kmer distributions, 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, I found small but important effects of trimming. Specifically, assembling with aggressively quality trimmed reads decreased the proportion of reads that map concordantly. For instance, the percent of reads successfully mapped to the assembly of 10 million Q20 trimmed reads was decreased by 0.6% or approximately 1.4 million reads (compared to mapping of untrimmed reads) while the effects on the assembly of 100 million Q20 trimmed reads was more blunted, with only 381,000 fewer reads mapping. Though the differences in mapping rates are exceptionally small, when working with extremely large datasets, the absolute difference in reads utilization may be substantial.

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. The maximum number of BLAST hits for each dataset were 10 M = 27452 hits, 20 M = 29563 hits, 50 M = 31848 hits, 75 M = 32786 hits, and 100 M = 33338 hits.

When counting complete ORFs recovered in the different assemblies, all datasets were all worsened by aggressive trimming, as evidenced by negative values in **Figure 4**. Trimming at PHRED = 20 was the most poorly performing level at all read depths. The maximum number of complete ORFs for each dataset were 10 M = 11429 ORFs, 20 M = 19463 ORFs, 50 M = 35632 ORFs, 75 M = 42205 ORFs, and 100 M = 48434 ORFs.

Of note, all assembly files are available for download on dataDryad (http://dx.doi.org/10.5061/dryad.7rm34).

DISCUSSION
Although the process of nucleotide quality trimming is commonplace in HTS analysis pipelines, particularly those involving assembly, its optimal implementation has not been well defined.
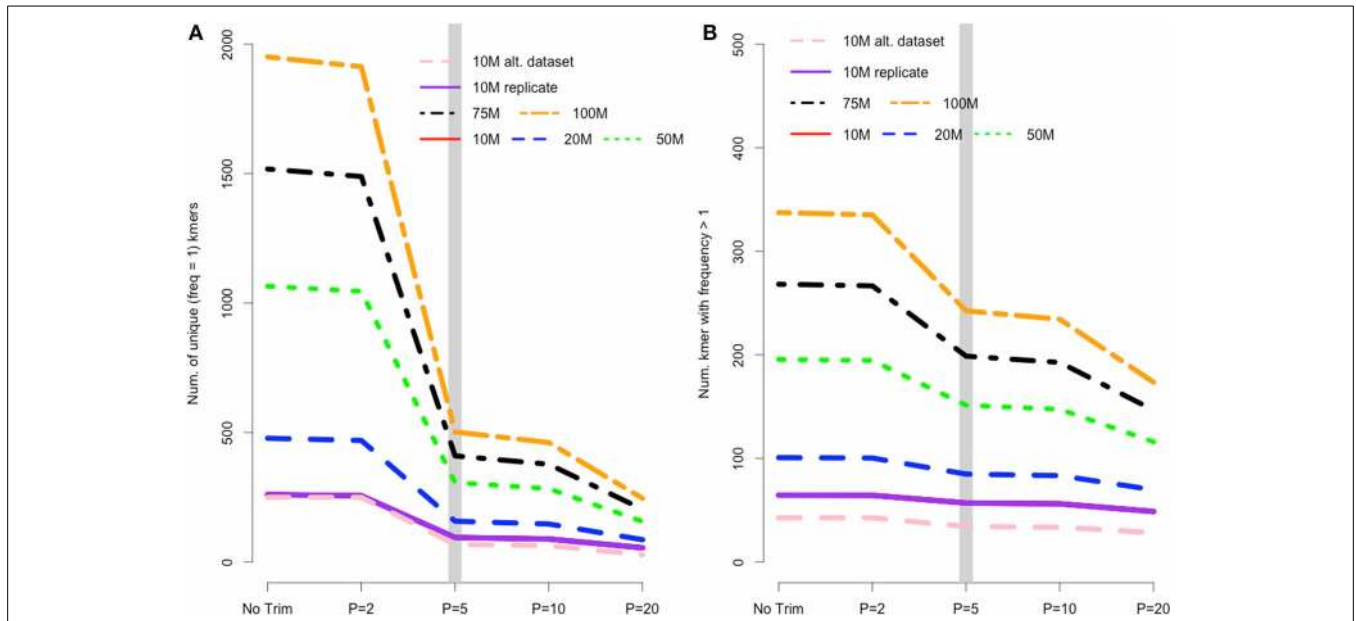
FIGURE 2 | (A) The number of unique kmers removed with various trimming levels across all datasets. Trimming at PHRED = 5 results in a substantial loss of likely erroneous kmers, while the effect of more and less aggressive trimming is more diminished. **(B)** Depicts the relationship between trimming and non-unique kmers, whose pattern is similar to that of unique kmers.
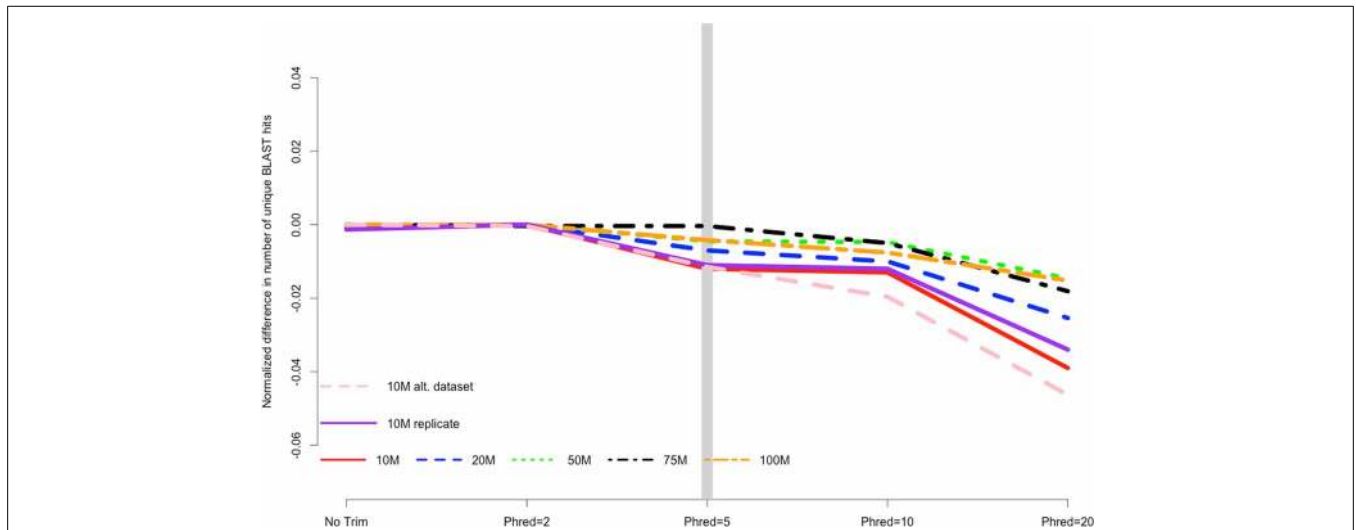

FIGURE 3 | The number of unique BLAST matches contained in the final transcriptome assembly is related to the strength of quality trimming, with more aggressive trimming resulting in worse performance. Data are normalized to the number of BLAST hits obtained in the most favorable trimming level for each dataset. Negative numbers indicate the detrimental affect of trimming. 10, 20, 50, 75, 100 M refer to the subsamples size. 10 M replicate is the technical replicate, 10 M alt. dataset is the secondary dataset.

Though the rigor with which trimming is performed seems to vary, there is a bias toward stringent trimming (Barrett and Davis, 2012; Ansell et al., 2013; 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.
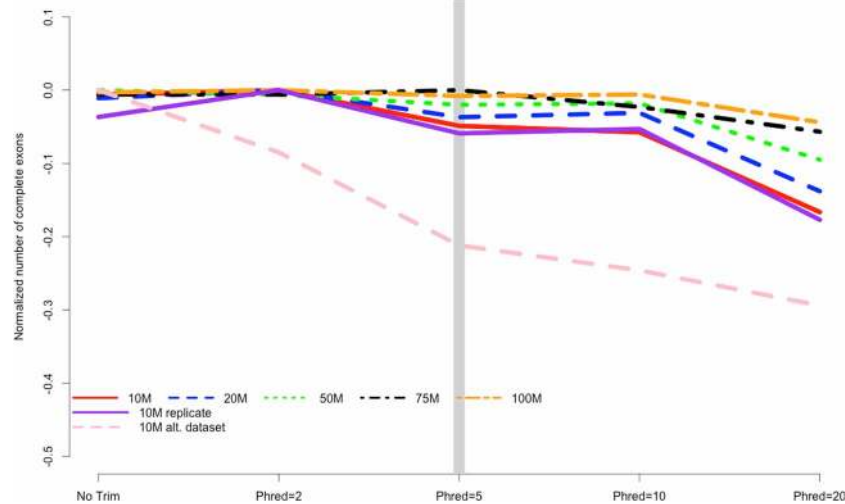
FIGURE 4 | The number of complete exons contained in the final transcriptome assembly is related to the strength of quality trimming for any of the studied sequencing depths, Trimming at PHRED = 20 was always associated with poor performance. Data are normalized to the number of complete exons obtained in the most favorable trimming level for each dataset. Negative numbers indicate the detrimental affect of trimming. 10, 20, 50, 75, 100 M refer to the subsamples size. 10 M replicate is the technical replicate, 10 M alt. dataset is the secondary dataset.

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 the bubble popping algorithms included in TRINITY and other *de Bruijn* graph assemblers.

The majority of the results presented here stem from the analysis of a single Illumina dataset and specific properties of that dataset may have biased the results. Though the dataset was selected for its "typical" Illumina error profile, other datasets may produce different results. To evaluate this possibility, a second dataset was evaluated at the 10 M subsampling level. Interestingly, although the assemblies based on this dataset contained more error (e.g., **Figure 1**), aggressive trimming did not improve quality for any of the assessed metrics, though like other datasets, the absolute number of errors were reduced.

In addition to the specific dataset, the subsampling procedure may have resulted in undetected biases. To address these concerns, a technical replicate of the original dataset was produced at the 10 M subsampling level. This level was selected as a smaller sample of the total dataset is more likely to contain an unrepresentative sample than larger samples. The results, depicted in all figures as the solid purple line, are concordant. Therefore, I believe that sampling bias is unlikely to drive the patterns reported on here.

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 read 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; $W = 18591566$, p-value = 7.184e-13, non-parametric Wilcoxon test).

I believe that the untoward effects of trimming are linked to a reduction in coverage. For the datasets tested here, trimming at PHRED = 20 resulted in the loss of nearly 25% of the dataset, regardless of the size of the initial dataset. This relationship does suggest, however, that the magnitude of the negative effects of trimming should be reduced in larger datasets, and in fact may be completely erased with ultra-deep sequencing. Indeed, when looking at the differences in the magnitude of negative effects in the datasets presented here, it is apparent that trimming at PHRED = 20 is "less bad" in the 100 M read dataset than it is in the 10 M read datasets. For instance, **Figure 2** demonstrates that one of the untoward effects of trimming, the reduction of non-unique kmers, is reduced as the depth of sequencing is increased. **Figures 3** and **4** demonstrate a similar pattern, where the negative effects of aggressive trimming of higher coverage datasets are blunted relative to lower coverage datasets.

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; $W = 26790212$, p-value $< 2.2e$-16) that are differentially recovered in the untrimmed dataset relative to the PHRED = 20 trimmed dataset.

E!!ects of coverage on transcriptome assembly—
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 assem-
bly 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, as well as the number
of exons, and BLAST hits were improved 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.

Should quality trimming be replaced by unique
kmer filtering?—For transcriptome studies that revolve
around assembly, quality control of sequence data has been
thought to be a crucial step. Though the removal of erroneous
nucleotides is the goal, how best to accomplish this is less clear. As
described above, quality trimming has been a common method,
but in its commonplace usage, may be detrimental to assembly.
What if, instead of relying on quality scores, we instead rely on
the distribution of kmers to guide our quality control endeav-
ors? In transcriptomes of typical complexity, sequenced to even
moderate coverage, it is reasonable to expect that all but the most
exceptionally rare mRNA molecules are sequenced at a depth >1.
Following this, all kmer whose frequency is <2 are putative errors,
and should be removed before assembly, though this process may
result in the loss of kmers from extremely low abundance tran-
scripts or isoforms. This idea and its implementation are fodder
for future research.

In summary, the process of nucleotide quality trimming is
commonplace in many HTS analysis pipelines, but its optimal
implementation has not been well defined. A very aggressive strat-
egy, 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 vul-
nerable 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 lever-
aged 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., et al. (2013). Insights into the immuno-molecular biology of
Angiostrongylus vasorum through transcriptomics-prospects for new interven-
tions. Biotechnol. Adv. 31, 1486–1500. doi: 10.1016/j.biotechadv.2013.07.006

Barrett, C. F., and Davis, J. I. (2012). The plastid genome of the mycoheterotrophic
Corallorhiza striata (Orchidaceae) is in the relatively early stages of degradation.
Am. J. Bot. 99, 1513–1523. doi: 10.3732/ajb.1200256

Bhardwaj, J., Chauhan, R., Swarnkar, M. K., Chahota, R. K., Singh, A. K.,
Shankar, R., et al. (2013). Comprehensive transcriptomic study on horse gram
(Macrotyloma uniflorum): de novo assembly, functional characterization and
comparative analysis in relation to drought stress. BMC Genomics 14:647. doi:
10.1186/1471-2164-14-647

Bradnam, K. R., Fass, J. N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., et al.
(2013). Assemblathon 2: evaluating de novo methods of genome assembly in
three vertebrate species. Gigascience 2, 10. doi: 10.1186/2047-217X-2-10

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al.
(2009). BLAST+: architecture and applications. BMC Bioinformatics 10:421.
doi: 10.1186/1471-2105-10-421

Cox, M. P., Peterson, D. A., and Biggs, P. J. (2010). SolexaQA: At-a-glance quality
assessment of Illumina second-generation sequencing data. BMC Bioinformatics
11:485. doi: 10.1186/1471-2105-11-485

Dohm, J. C., Lottaz, C., Borodina, T., and Himmelbauer, H. (2008). Substantial
biases in ultra-short read data sets from high-throughput DNA sequencing.
Nucleic Acids Res. 36, e105–e105. doi: 10.1093/nar/gkn425

Earl, D., Bradnam, K., St. John, J., Darling, A., Lin, D., Fass, J., et al. (2011).
Assemblathon 1: a competitive assessment of de novo short read assembly
methods. Genome Res. 21, 2224–2241. doi: 10.1101/gr.1265599.111

Francis, W. R., Christianson, L. M., Kiko, R., Powers, M. L., Shaner, N. C., and
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. doi: 10.1186/1471-2164-14-167

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I.,
et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a
reference genome. Nat. Biotechnol. 29, 644–652. doi: 10.1038/nbt.1883

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden,
J., et al. (2013). De novo transcript sequence reconstruction from RNA-seq
using the Trinity platform for reference generation and analysis. Nat. Protoc.
8, 1494–1512. doi: 10.1038/nprot.2013.084

Han, H., Irimia, M., Ross, P. J., Sung, H.-K., Alipanahi, B., David, L., et al. (2013).
MBNL proteins repress ES-cell-specific alternative splicing and reprogramming.
Nature 498, 241–245. doi: 10.1038/nature12270

Hansen, K. D., Brenner, S. E., and Dudoit, S. (2010). Biases in Illumina transcrip-
tome sequencing caused by random hexamer priming. Nucleic Acids Res. 38,
e131. doi: 10.1093/nar/gkq224

Helyar, S. J., Limborg, M. T., Bekkevold, D., Babbucci, M., van Houdt, J., Maes,
G. E., et al. (2012). SNP discovery using Next Generation Transcriptomic
Sequencing in Atlantic herring (Clupea harengus). PLoS ONE 7:e42089. doi:
10.1371/journal.pone.0042089

Hsu, J.-C., Chien, T.-Y., Hu, C.-C., Chen, M.-J. M., Wu, W.-J., Feng, H.-T., et al.
(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. doi: 10.1371/journal.pone.0040950

Hunt, M., Kikuchi, T., Sanders, M., Newbold, C., Berriman, M., and Otto, T. D.
(2013). REAPR: a universal tool for genome assembly evaluation. Genome Biol.
14:R47. doi:10.1186/gb-2013-14-5-r47

Kent, W. J. (2002). BLAT—the BLAST-like alignment tool. Genome Res. 12,
656–664. doi: 10.1101/gr.229202

Linnen, C. R., Poh, Y.-P., Peterson, B. K., Barrett, R. D. H., Larson, J. G., Jensen, J. D.,
et al. (2013). Adaptive evolution of multiple traits through multiple mutations
at a single gene. Science 339, 1312–1316. doi: 10.1126/science.1233213

Liu, B., Yuan, J., Yiu, S.-M., Li, Z., Xie, Y., Chen, Y., et al. (2012). COPE: an accu-
rate k-mer-based pair-end reads connection tool to facilitate genome assembly.
Bioinformatics 28, 2870–2874. doi: 10.1093/bioinformatics/bts563

Liu, T., Zhu, S., Tang, Q., Chen, P., Yu, Y., and Tang, S. (2013). De novo assembly and
characterization of transcriptome using Illumina paired-end sequencing and
identification of CesA gene in ramie (Boehmeria nivea L. Gaud). BMC Genomics
14:125. doi: 10.1186/1471-2164-14-125

Lohse, M., Bolger, A. M., Nagel, A., Fernie, A. R., Lunn, J. E., Stitt, M., et al.
(2012). RobiNA: a user-friendly, integrated software solution for RNA-Seq-
based transcriptomics. Nucleic Acids Res. 40, W622–W627. doi: 10.1093/nar/
gks540

Looso, M., Preussner, J., Sousounis, K., Bruckskotten, M., Michel, C. S., Lignelli, E., et al. (2013). A *de novo* assembly of the newt transcriptome combined with proteomic validation identifies new protein families expressed during tissue regeneration. *Genome Biol.* 14:R16. doi:10.1186/gb-2013-14-2-r16

Macfarlan, T. S., Gifford, W. D., Driscoll, S., Lettieri, K., Rowe, H. M., Bonanomi, D., et al. (2012). Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature* 487, 57–63. doi: 10.1038/nature11244

MacManes, M. D., and Eisen, M. B. (2013). Improving transcriptome assembly through error correction of high-throughput sequence reads. *PeerJ* 1:e113. doi: 10.7717/peerj.113

MacManes, M. D., and Lacey, E. A. (2012). The social brain: transcriptome assembly and characterization of the hippocampus from a social subterranean rodent, the colonial Tuco-Tuco (*Ctenomys sociabilis*). *PLoS ONE* 7:e45524. doi: 10.1371/journal.pone.0045524

Marçais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27, 764–770. doi: 10.1093/bioinformatics/btr011

Milano, I., Babbucci, M., Panitz, F., Ogden, R., Nielsen, R. O., Taylor, M. I., et al. (2011). Novel tools for conservation genomics: comparing two high-throughput approaches for SNP discovery in the transcriptome of the European hake. *PLoS ONE* 6:e28008. doi: 10.1371/journal.pone.0028008

Muñoz-Mérida, A., González-Plaza, J. J., Cañada, a., Blanco, A. M., García-López, M. D. C., Rodríguez, J. M., et al. (2013). *De novo* assembly and functional annotation of the olive (*Olea europaea*) transcriptome. *DNA Res.* 20, 93–108. doi: 10.1093/dnares/dss036

Narum, S. R., Campbell, N. R., Meyer, K. A., Miller, M. R., and Hardy, R. W. (2013). Thermal adaptation and acclimation of ectotherms from differing aquatic climates. *Mol. Ecol.* 22, 3090–3097. doi: 10.1111/mec.12240

Pyrkosz, A. B., Cheng, H., and 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., and Giribet, G. (2012). Optimization of preservation and storage time of sponge tissues to obtain quality mRNA for next-generation sequencing. *Mol. Ecol. Resour.* 12, 312–322. doi: 10.1111/j.1755-0998.2011.03097.x

Roberts, A., and Pachter, L. (2013). Streaming fragment assignment for real-time analysis of sequencing experiments. *Nat. Methods* 10, 71–73. doi: 10.1038/nmeth.2251

Straub, S. C. K., Cronn, R. C., Edwards, C., Fishbein, M., and Liston, A. (2013). Horizontal transfer of DNA from the mitochondrial to the plastid genome and its subsequent evolution in milkweeds (Apocynaceae). *Genome Biol. Evol.* 5, 1872–1885. doi: 10.1093/gbe/evt140

Tao, T., Zhao, L., Lv, Y., Chen, J., Hu, Y., Zhang, T., et al. (2013). Transcriptome sequencing and differential gene expression analysis of delayed gland morphogenesis in *Gossypium australe* during seed germination. *PLoS ONE* 8:e75323. doi: 10.1371/journal.pone.0075323

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., et al. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28, 511–515. doi: 10.1038/nbt.1621

Yang, X., Chockalingam, S. P., and Aluru, S. (2013). A survey of error-correction methods for next-generation sequencing. *Brief. Bioinform.* 14, 56–66. doi: 10.1093/bib/bbs015