On the optimal trimming of high-throughput mRNA sequence data

Matthew D MacManes\textsuperscript{1,2,3}\textsuperscript{*}

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

\textsuperscript{*} Corresponding author: macmanes@gmail.com, Twitter: @PeroMHC

\section*{Abstract}

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 \texttt{Phred} score $< 2$ or $< 5$, is optimal for most studies across a wide variety of metrics.

\section*{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 (Bradnam et al., 2013; Earl et al., 2011), significant computational hurdles still exist. Amongst the most difficult of challenges in transcriptome assembly involves the reconstruction of isoforms (Pykosz 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 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 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), BIOPICLES (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 (Helyar et al., 2012; Milano et al., 2011), functional studies (Ansell et al., 2013; Bhardwaj et al., 2013), and more general studies of transcriptome characterization (MacManes and Lacey, 2012; Touming et al., 2013)), 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 amount of high quality data (even nucleotides removed with the commonly used \texttt{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 \textit{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. 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 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 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 million, 20 million, 50 million, 75 million, 100 million read datasets. To test the robustness of my findings, I evaluated both a second dataset (SRR385624, (Macfarlan et al., 2012)) as well as a technical replicate of the primary dataset, both at the 10M 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 \texttt{Phred=0} (adapter trimming only), \(\leq 2\), \(\leq 5\), \(\leq 10\), and \(\leq 20\). Other parameters (\texttt{MINLEN=25}, \texttt{ILLUMINACLIP=bases:2:40:15}, \texttt{SLIDINGWINDOW size=4}) were held constant. Transcriptome assemblies were generated for each dataset using the default settings (except \texttt{group_pairs_distance flag set to 999}) of the program Trinity \texttt{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 \textit{Mus} cDNA’s, available at \texttt{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 \texttt{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 \texttt{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 \texttt{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 open reading frames (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 evalue \(< 10^{-10} \). 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 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 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 (Trimming of nucleotides whose error scores are: \( \text{Phred} > 20 \), 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 \( \text{Phred} = 5 \) may be optimal, given the potential untoward effects of more stringent quality trimming. 10M, 20M, 50M, 75M, 100M refer to the subsamples size. 10M replicate is the technical replicate, 10M alt. dataset is the secondary dataset. Note that to enhance clarity, the Y-axis does not start at zero.

In \textit{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 results of sequencing error, and therefore may be removed during the trimming process. Figure 1A shows the pattern of unique kmer loss across the various trimming levels and read datasets. What is apparent, is that trimming at \( \text{Phred}=5 \) 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 1B shows that while
**Figure 2**

**Figure 2A.** 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. 2B depicts the relationship between trimming and non-unique kmers, whose pattern is similar to that of unique kmers.

In addition to looking at nucleotide error and kmer distributions, assembly quality may be measured by 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 to a given contig (Figure 3). Though the patterns are not visually striking, mapping to an assembly of aggressively trimmed reads results in several hundred thousand fewer reads mapped compared to mapping against
the assembly of less aggressively trimmed reads.

Figure 3

![Graph showing proportion of reads mapped](image)

Figure 3. The proportion 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. 10M, 20M, 50M, 75M, 100M refer to the subsamples size. 10M replicate is the technical replicate, 10M alt. dataset is the secondary dataset. Note that to enhance clarity, the Y-axis does not start at zero.

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 4), with trimming at Phred=20 associated with particularly poor performance.
Figure 4. 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. 10M, 20M, 50M, 75M, 100M refer to the subsamples size. 10M replicate is the technical replicate, 10M alt. dataset is the secondary dataset. Note that to enhance clarity, the Y-axis does not start at zero.

When counting complete open reading frames, low and moderate coverage datasets (10M, 20M, 50M) were all worsened by aggressive trimming (Figure 5). Trimming at Phred=20 was the most poorly performing level at all read depths.

Figure 5
Figure 5. 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. 10M, 20M, 50M, 75M, 100M refer to the subsamples size. 10M replicate is the technical replicate, 10M alt. dataset is the secondary dataset.

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, particularly those involving assembly, its 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 $\text{Phred}=2$ or $\text{Phred}=5$ 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

<table>
<thead>
<tr>
<th>Dataset Size</th>
<th>Error</th>
<th>Map</th>
<th>Orf</th>
<th>Blast</th>
</tr>
</thead>
<tbody>
<tr>
<td>10M</td>
<td>20</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>10M rep.</td>
<td>20</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>10M alt</td>
<td>20</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>20M</td>
<td>5</td>
<td>5</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>50M</td>
<td>5</td>
<td>10</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>75M</td>
<td>20</td>
<td>10</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>100M</td>
<td>20</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 1. The $\text{Phred}$ 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. 10M rep. is the technical replicate, 10M alt. 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 $\text{Phred}=20$ 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 $\text{Phred}$ scores in the predicted fashion yet downstream effects are minimal. This should be interpreted as speaking to the performance of 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 10M 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 10M 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, sampling bias is unlikely do 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 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; W = 18591566, p-value = 7.184e-13, non-parametric Wilcoxon test).

Indeed, 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 the 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 100M read dataset than it is in the 10M read dataset (Figure 6).

Figure 6
Figure 6. The normalized difference in the number of complete exons contained in the
transcriptome assembly between trimming levels for the five different read datasets. Positive
numbers indicate increased exons recovered, while negative values indicate decreased recovery. At
the most aggressive trimming levels, the negative effects are greatest for the smaller datasets, while
more mitigated for larger datasets. This supports the hypothesis that the relationship between
assembly quality and trimming may be driven by differences in coverage. P0 - P2 indicates
differences between Phred=0 and Phred 2 trimming, P2 - P5 are differences between Phred=2 and
Phred=5 trimming, etc.

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.

Effects 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 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, 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 an 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 endeavors? 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 <1 are putative errors, and should be removed before assembly. 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 strategy, where sequence reads are trimmed when \( \text{PHRED} \) 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. \( \text{PHRED}=2 \) or \( \text{PHRED}=5 \)) 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

This paper was greatly improved by suggestions of C. Titus Brown and a second anonymous reviewer. In addition, the paper was first released as a bioRxiv preprint, and I received several comments based on that work both on that website as well as via Twitter. Let it be said here, that early use of a preprint archive, open access publication, and Twitter based discussion is a powerful way to rapidly disseminate (and get feedback on) work. I highly encourage its use!

References


