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 [1, 2] or functional [3, 4] 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 [5, 6]. transcriptome assembly is less challenging, significant computational hurdles

still exist. Amongst the most difficult of challenges involves the reconstruction of isoforms [7] 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 [8]. These errors are overwhelmingly substitution errors [9], 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 [**?**], low quality (=high probability of error) nucleotides are commonly removed from the sequencing reads prior to assembly, using one of several available tools (TRIMMOMATIC [10], FASTX TOOLKIT (http://hannonlab.cshl.edu/fastx_toolkit/index.html), BIOPIECES (http://www.biopieces.org/), SOLEXAQA [11]). 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. [12, 13]), 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 [14], 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 [10], 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 [15**?**]. Code for running TRINITY is available at https://gist.github.com/macmanes/5859956. 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 [16] 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 untrimmed sequencing reads to each assembly using Bowtie2 [17].

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 BLAST+ [18]. As the number of transcripts matching a given reference increases, so may assembly quality.

Results

Quality trimming of sequence reads had a relatively small 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 more modest improvements may be garnered with more aggressive trimming.

Figure 1

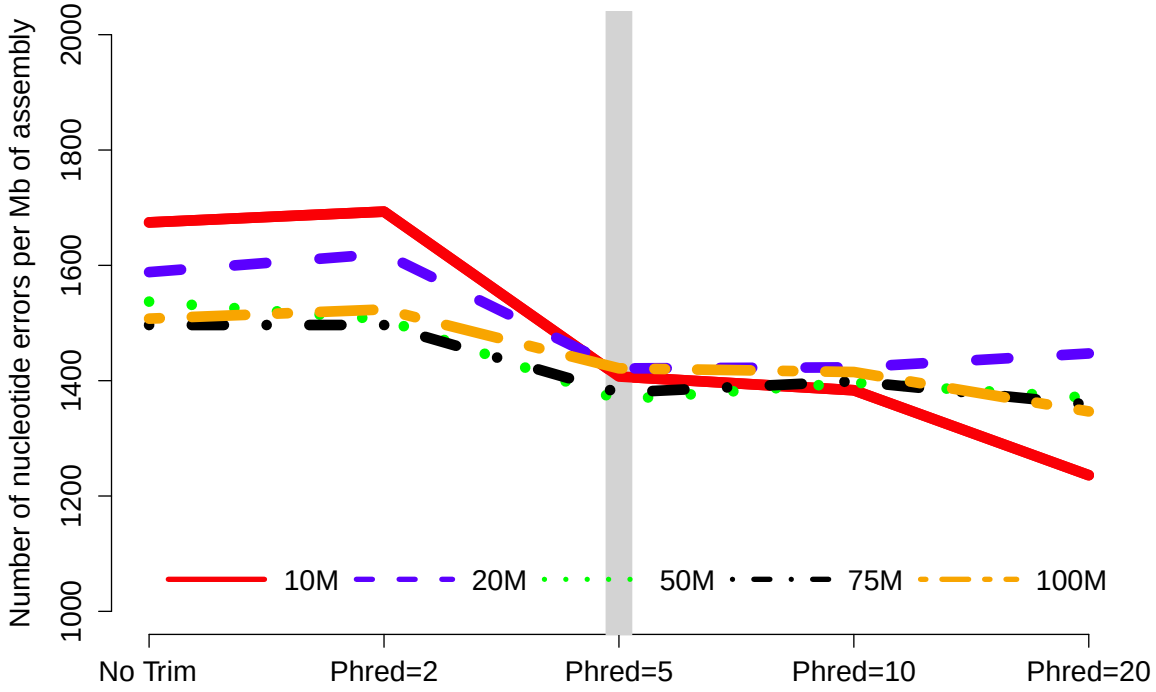

Figure 1. The number of nucleotide errors contained in the final transcriptome assembly, normalized to assembly size, is related to the strength of quality trimming (Trimming of nucleotides whose error scores are: PHRED >20, 10, 5, 2, or no trimming, though most benefits are observed at a modest level 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 [19]. 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 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 assembly of untrimmed reads.

Figure 2

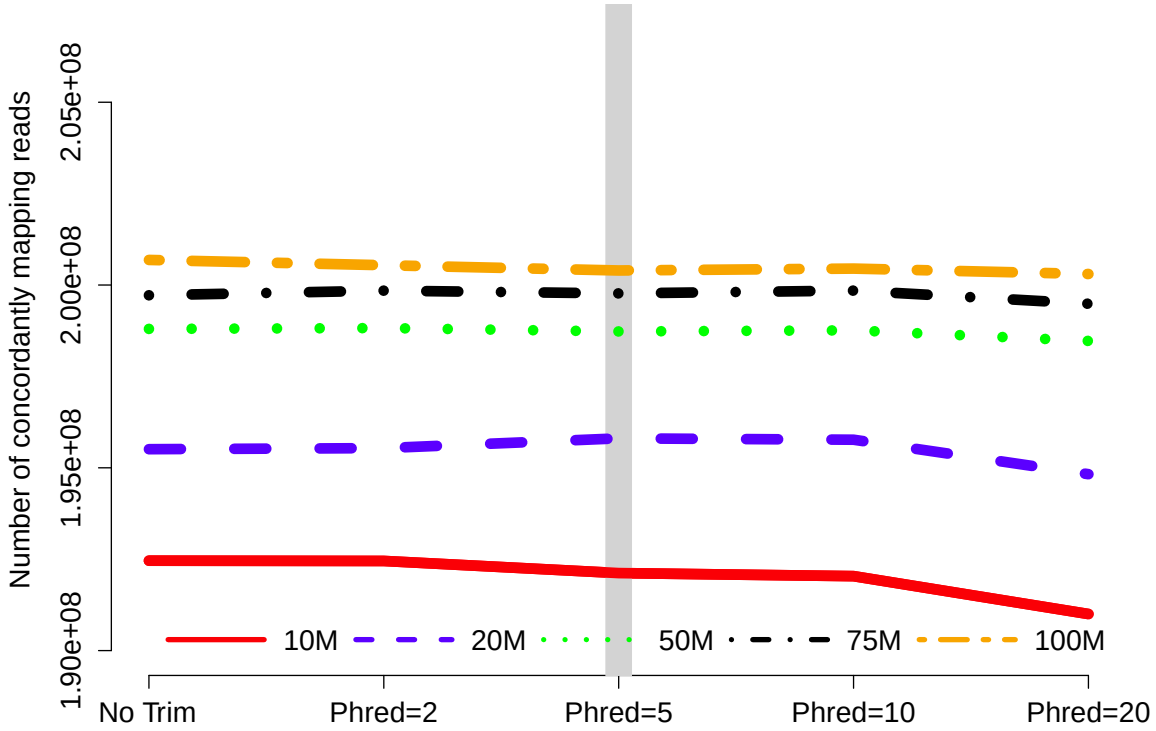

Figure 2. The number of concordantly mapping reads was reduced by trimming. The pattern is particularly salient with trimming at PHRED=20 which was always associated with the successful mapping of hundreds of thousands of fewer reads.

Analysis of assembly content painted a similar picture, with trimming having a relatively small, though tangible effect. The number of BLAST+ matches decreased as the stringency of trimming increased from PHRED=0 to PHRED=20 (Figure 3), though trimming at PHRED=20 was associated with particularly poor performance.

Figure 3

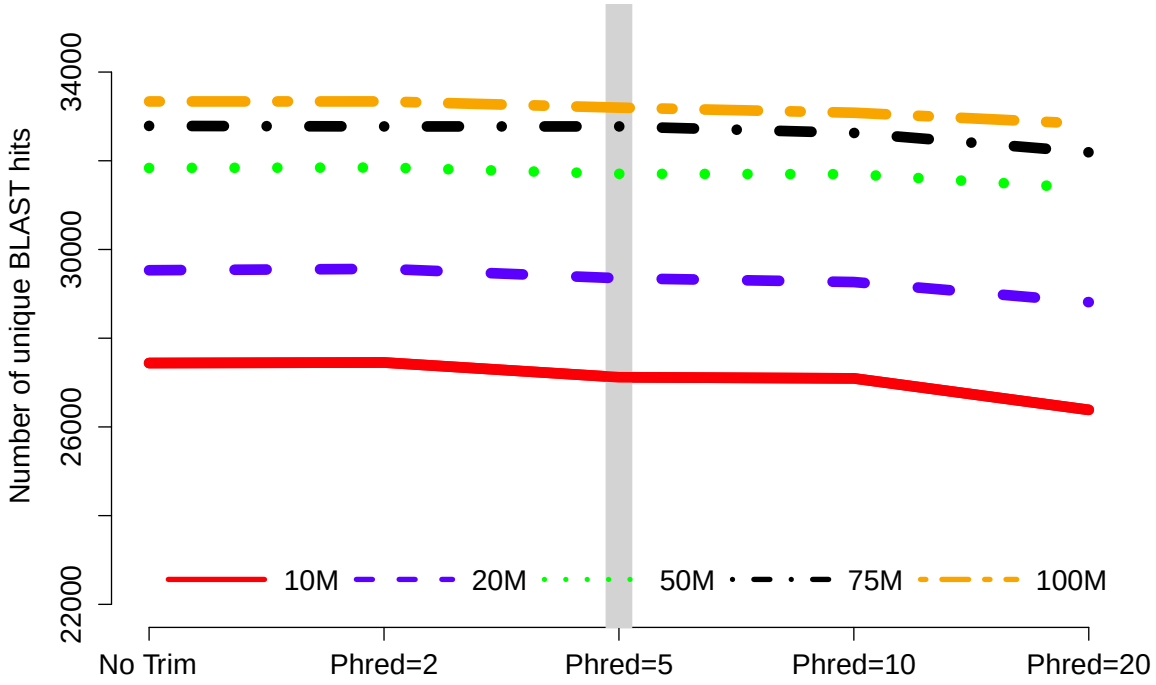

Figure 3. The number of unique BLAST matches contained in the final transcriptome assembly is related to the strength of quality trimming for any of the studied sequencing depths. The 'no trimming' strategy always yielded the most number of unique matches, while trimming at PHRED=20 was always associated with much poorer assembly content

When counting complete open reading frames, low and moderate coverage datasets (10M, 20M, 50M) were all worsened by all levels of trimming (Figure 4), the high overage datsets (75M and 100M) showed mild improvement at this metric at trimming at PHRED=5 OR 10 levels. Trimming at PHRED=20 was the most poorly performing level at all read depths.

Figure 4

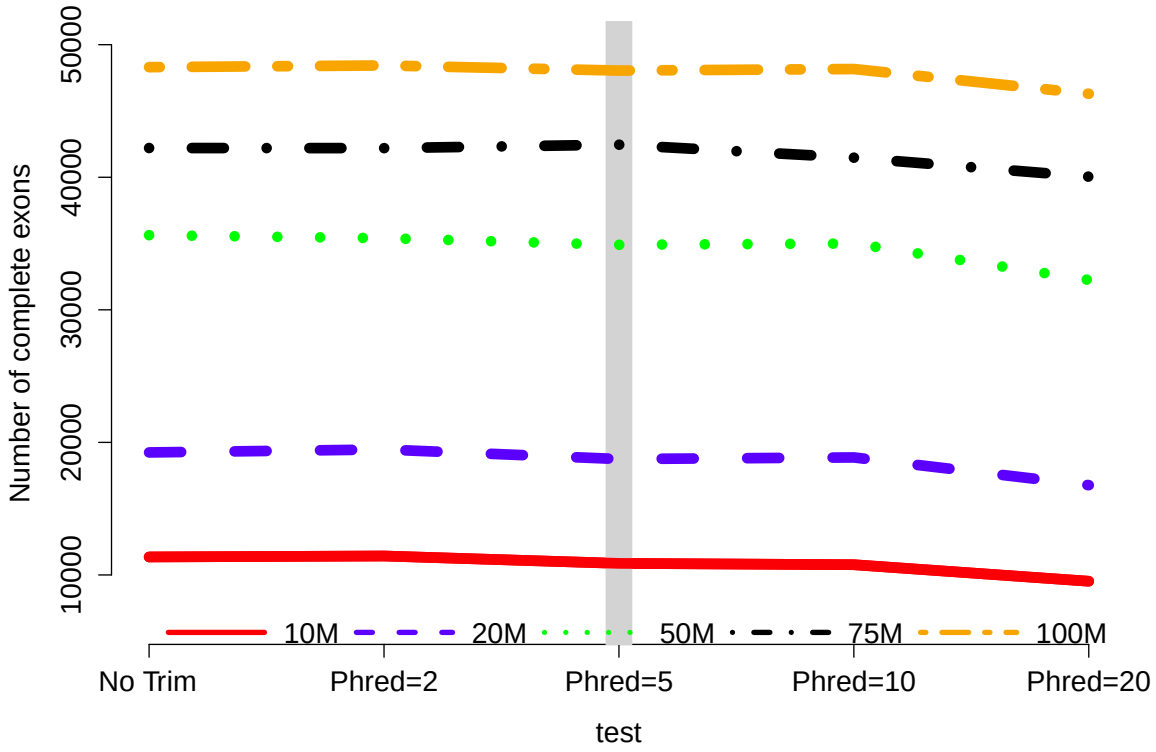

Figure 4. The number of complete exons contained in the final transcriptome assembly is not strongly related to the strength of quality trimming for any of the studies sequencing depths, though trimming at PHRED=20 was always associated with fewer identified exons.

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 [20–23]. This study provides strong evidence that stringent quality trimming of nucleotides whose quality scores are ≥ 20 results in a poorer transcriptome assembly across all metrics. Instead, researchers interested in assembling transcriptomes *de novo* should elect for a much more gentle quality trimming, or no trimming at all, particularly when fewer than 20 million reads are generated.

Activity	Pre	2014	2015	2016	Post
Recruit UG, GS & PDF	X	X			
Collect Animals for Genome	X				
Genome sequencing & assembly – Aim 1	X	X			
Collect and analyze physiology data – Aim 2		X	X		
Collect and analyze RNAseq data – Aim 3		X	X	X	
Write papers and submit			X	X	
Present results at international conference			X	X	X
Train UG, GS & PDF	X	X	X	X	X
Disseminate info	X	X	X	X	X

The results of this study were surprising. In fact, much of my own work assembling transcriptomes included a vigorous trimming step. That trimming had such 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 [11]. Indeed, the program modifies the distribution of Phred in the predicted fashion (Figure 5) yet downstream effects are minimal.

Why does quality trimming have such small effect, and even a negative effect at Phred=20?

Effects of read depth — 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 [24], 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.

Acknowledgments

References

[1] Catherine R Linnen, Yu-Ping Poh, Brant K Peterson, Rowan D H Barrett, Joanna G Larson, Jeffrey D Jensen, and Hopi E Hoekstra. Adaptive evolution of multiple traits through multiple mutations at a single gene. *Science (New York, NY)*, 339(6125):1312–1316, March 2013.

[2] Shawn R Narum, Shawn R Narum, Nathan R Campbell, Nathan R Campbell, Kevin A Meyer, Kevin A Meyer, Michael R Miller, Michael R Miller, Ronald W Hardy, and Ronald W Hardy. Thermal adaptation and acclimation of ectotherms from differing aquatic climates. *Molecular Ecology*, pages 1–8, March 2013.

[3] A Mu noz Merida, A Mu noz Merida, J J Gonzalez-Plaza, J J Gonzalez-Plaza, A Canada, A Canada, A M Blanco, A M Blanco, M d C Garcia-Lopez, M d C Garcia-Lopez, J M Rodriguez, J M Rodriguez, L Pedrola, L Pedrola, M D Sicardo, M D Sicardo, M L Hernandez, M L Hernandez, R De la Rosa, R De la Rosa, A Belaj, A Belaj, M Gil-Borja, M Gil-Borja, F Luque, F Luque, J M Martinez-Rivas, J M Martinez-Rivas, D G Pisano, D G Pisano, O Trelles, O Trelles, V Valpuesta, V Valpuesta, C R Beuzon, and C R Beuzon. De Novo Assembly and Functional Annotation of the Olive (*Olea europaea*) Transcriptome. *DNA Research*, 20(1):93–108, February 2013.

[4] Ju-Chun Hsu, Ju-Chun Hsu, Ting-Ying Chien, Ting-Ying Chien, Chia-Cheng Hu, Chia-Cheng Hu, Mei-Ju May Chen, Mei-Ju May Chen, Wen-Jer Wu, Wen-Jer Wu, Hai-Tung Feng, Hai-Tung Feng, David S Haymer, David S Haymer, Chien-Yu Chen, and Chien-Yu Chen. Discovery of Genes Related to Insecticide Resistance in *Bactrocera dorsalis* by Functional Genomic Analysis of a *De Novo* Assembled Transcriptome. *PloS one*, 7(8):e40950, August 2012.

[5] Keith R Bradnam, Joseph N Fass, Anton Alexandrov, Paul Baranay, Michael Bechner, Inanç Birol, Sébastien Boisvert, Jarrod A Chapman, Guillaume Chapuis, Rayan Chikhi, Hamidreza Chitsaz, Wen-Chi Chou, Jacques Corbeil, Cristian Del Fabbro, T Roderick Docking, Richard Durbin, Dent Earl, Scott Emrich, Pavel Fedotov, Nuno A Fonseca, Ganeshkumar Ganapathy, Richard A Gibbs, Sante Gnerre, Elénie Godzaridis, Steve Goldstein, Matthias Haimel, Giles Hall, David Haussler, Joseph B Hiatt, Isaac Y Ho, Jason Howard, Martin Hunt, Shaun D Jackman, David B Jaffe, Erich Jarvis, Huaiyang Jiang, Sergey Kazakov, Paul J Kersey, Jacob O Kitzman, James R Knight, Sergey Koren, Tak-Wah Lam, Dominique Lavenier, François Laviolette, Yingrui Li, Zhenyu Li, Binghang Liu, Yue Liu, Ruibang Luo, Iain Maccallum, Matthew D Macmanes, Nicolas Maillet, Sergey Melnikov, Delphine Naquin, Zemin Ning, Thomas D Otto, Benedict Paten, Octávio S Paulo, Adam M Phillippy, Francisco Pina-Martins, Michael Place, Dariusz Przybylski, Xiang Qin, Carson Qu, Filipe J Ribeiro, Stephen Richards, Daniel S Rokhsar, J Graham Ruby, Simone Scalabrin, Michael C Schatz, David C Schwartz, Alexey Sergushichev, Ted Sharpe, Timothy I Shaw, Jay Shendure, Yujian Shi, Jared T Simpson, Henry Song, Fedor Tsarev, Francesco Vezzi, Riccardo Vicedomini, Bruno M Vieira, Jun Wang, Kim C Worley, Shuangye Yin, Siu-Ming Yiu, Jianying Yuan, Guojie Zhang, Hao Zhang, Shiguo Zhou, and Ian F Korf. Assemblathon 2: evaluating *de novo* methods of genome assembly in three vertebrate species. *GigaScience*, 2(1):10, July 2013.

[6] D Earl, D Earl, K Bradnam, K Bradnam, J St John, J St John, A Darling, A Darling, D Lin, D Lin, J Fass, J Fass, H O K Yu, H O K Yu, V Buffalo, V Buffalo, D R Zerbino, D R Zerbino,

M Diekhans, M Diekhans, N Nguyen, N Nguyen, P N Ariyaratne, P N Ariyaratne, W K Sung, W K Sung, Z Ning, Z Ning, M Haimel, M Haimel, J T Simpson, J T Simpson, N A Fonseca, N A Fonseca, I Birol, I Birol, T R Docking, T R Docking, I Y Ho, I Y Ho, D S Rokhsar, D S Rokhsar, R Chikhi, R Chikhi, D Lavenier, D Lavenier, G Chapuis, G Chapuis, D Naquin, D Naquin, N Maillet, N Maillet, M C Schatz, M C Schatz, D R Kelley, D R Kelley, A M Phillippy, A M Phillippy, S Koren, S Koren, S P Yang, S P Yang, W Wu, W Wu, W C Chou, W C Chou, A Srivastava, A Srivastava, T I Shaw, T I Shaw, J G Ruby, J G Ruby, P Skewes-Cox, P Skewes-Cox, M Betegon, M Betegon, M T Dimon, M T Dimon, V Solovyev, V Solovyev, I Seledtsov, I Seledtsov, P Kosarev, P Kosarev, D Vorobyev, D Vorobyev, R Ramirez-Gonzalez, R Ramirez-Gonzalez, R Leggett, R Leggett, D MacLean, D MacLean, F Xia, F Xia, R Luo, R Luo, Z Li, Z Li, Y Xie, Y Xie, B Liu, B Liu, S Gnerre, S Gnerre, I Maccallum, I Maccallum, D Przybylski, D Przybylski, F J Ribeiro, F J Ribeiro, S Yin, S Yin, T Sharpe, T Sharpe, G Hall, G Hall, P J Kersey, P J Kersey, R Durbin, R Durbin, S D Jackman, S D Jackman, J A Chapman, J A Chapman, X Huang, X Huang, J L DeRisi, J L DeRisi, M Caccamo, M Caccamo, Y Li, Y Li, D B Jaffe, D B Jaffe, R E Green, R E Green, D Haussler, D Haussler, I Korf, I Korf, B Paten, and B Paten. Assemblathon 1: A competitive assessment of *de novo* short read assembly methods. *Genome Research*, 21(12):2224–2241, December 2011.

[7] Alexis Black Pyrkosz, Alexis Black Pyrkosz, Hans Cheng, Hans Cheng, C Titus Brown, and C Titus Brown. RNA-Seq Mapping Errors When Using Incomplete Reference Transcriptomes of Vertebrates. *http://arxiv.org/abs/1303.2411v1*, March 2013.

[8] B Liu, B Liu, J Yuan, J Yuan, S M Yiu, S M Yiu, Z Li, Z Li, Y Xie, Y Xie, Y Chen, Y Chen, Y Shi, Y Shi, H Zhang, H Zhang, Y Li, Y Li, T W Lam, T W Lam, R Luo, and R Luo. COPE: an accurate k-mer-based pair-end reads connection tool to facilitate genome assembly. *Bioinformatics (Oxford, England)*, 28(22):2870–2874, November 2012.

[9] X Yang, S P Chockalingam, and S Aluru. A survey of error-correction methods for next-generation sequencing. *Briefings In Bioinformatics*, 14(1):56–66, January 2013.

[10] M Lohse, M Lohse, A M Bolger, A M Bolger, A Nagel, A Nagel, A R Fernie, A R Fernie, J E Lunn, J E Lunn, M Stitt, M Stitt, B Usadel, and B Usadel. RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. *Nucleic Acids Research*, 40(W1):W622–W627, June 2012.

[11] Murray P Cox, Daniel A Peterson, and Patrick J Biggs. SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics*, 11(1):485, 2010.

[12] Ana Riesgo, Ana Riesgo, Alicia R Perez-Porro, Alicia R Perez-Porro, Susana Carmona, Susana Carmona, Sally P Leys, Sally P Leys, Gonzalo Giribet, and Gonzalo Giribet. Optimization of preservation and storage time of sponge tissues to obtain quality mRNA for next-generation sequencing. *Molecular ecology resources*, 12(2):312–322, March 2012.

[13] Mario Looso, Mario Looso, Jens Preussner, Jens Preussner, Konstantinos Sousounis, Konstantinos Sousounis, Marc Bruckskotten, Marc Bruckskotten, Christian S Michel, Christian S Michel, Ettore Lignelli, Ettore Lignelli, Richard Reinhardt, Richard Reinhardt, Sabrina Hoeffner, Sabrina Hoeffner, Marcus Krueger, Marcus Krueger, Panagiotis A Tsonis, Panagiotis A Tsonis, Thilo Borchardt, Thilo Borchardt, Thomas Braun, and Thomas Braun. A *de novo* assembly of the newt transcriptome combined with proteomic validation identifies new protein families expressed during tissue regeneration. *Genome Biology*, 14(2):R16, 2013.

[14] Hong Han, Hong Han, Manuel Irimia, Manuel Irimia, P Joel Ross, P Joel Ross, Hoon-Ki Sung, Hoon-Ki Sung, Babak Alipanahi, Babak Alipanahi, Laurent David, Laurent David, Azadeh Golipour, Azadeh Golipour, Mathieu Gabut, Mathieu Gabut, Iacovos P Michael, Iacovos P Michael, Emil N Nachman, Emil N Nachman, Eric Wang, Eric Wang, Dan Trcka, Dan Trcka, Tadeo Thompson, Tadeo Thompson, Dave O'Hanlon, Dave O'Hanlon, Valentina Slobodeniuc, Valentina Slobodeniuc, Nuno L Barbosa-Morais, Nuno L Barbosa-Morais, Christopher B Burge, Christopher B Burge, Jason Moffat, Jason Moffat, Brendan J Frey, Brendan J Frey, andras Nagy, andras Nagy, James Ellis, James Ellis, Jeffrey L Wrana, Jeffrey L Wrana, Benjamin J Blencowe, and Benjamin J Blencowe. MBNL proteins repress ES-cell-specific alternative splicing and reprogramming. *Nature*, 498(7453):241–245, April 2014.

[15] Manfred G Grabherr, Brian J Haas, Moran Yassour, Joshua Z Levin, Dawn A Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, Zehua Chen, Evan Mauceli, Nir Hacohen, andreas Gnirke, Nicholas Rhind, Federica di Palma, Bruce W Birren, Chad Nusbaum, Kerstin Lindblad-Toh, Nir Friedman, and Aviv Regev. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, 29(7):644–652, July 2011.

[16] W J Kent and W J Kent. BLAT—The BLAST-Like Alignment Tool. *Genome Research*, 12(4):656–664, March 2002.

[17] Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28(5):511–515, May 2010.

[18] Christiam Camacho, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L Madden. BLAST+: architecture and applications. *BMC Bioinformatics*, 10(1):421, 2009.

[19] Martin Hunt, Martin Hunt, Taisei Kikuchi, Taisei Kikuchi, Mandy Sanders, Mandy Sanders, Chris Newbold, Chris Newbold, Matthew Berriman, Matthew Berriman, Thomas D Otto, and Thomas D Otto. REAPR: a universal tool for genome assembly evaluation. *Genome Biology*, 14(5):R47, May 2013.

[20] C F Barrett and J I Davis. The plastid genome of the mycoheterotrophic Corallorhiza striata (Orchidaceae) is in the relatively early stages of degradation. *American Journal of Botany*, 99(9):1513–1523, September 2012.

[21] Tao Tao, Tao Tao, Liang Zhao, Liang Zhao, Yuanda Lv, Yuanda Lv, Jiedan Chen, Jiedan Chen, Yan Hu, Yan Hu, Tianzhen Zhang, Tianzhen Zhang, Baoliang Zhou, and Baoliang Zhou. Transcriptome Sequencing and Differential Gene Expression Analysis of Delayed Gland Morphogenesis in Gossypium australe during Seed Germination. *PloS one*, 8(9):e75323, September 2013.

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

[23] Brendan R E Ansell, Brendan R E Ansell, Manuela Schnyder, Manuela Schnyder, Peter Deplazes, Peter Deplazes, Pasi K Korhonen, Pasi K Korhonen, Neil D Young, Neil D Young, Ross S Hall,

Ross S Hall, Stefano Mangiola, Stefano Mangiola, Peter R Boag, Peter R Boag, andreas Hofmann, andreas Hofmann, Paul W Sternberg, Paul W Sternberg, Aaron R Jex, Aaron R Jex, Robin B Gasser, and Robin B Gasser. Biotechnology Advances. *Biotechnology Advances*, pages 1–15, September 2013.

[24] Warren R Francis, Lynne M Christianson, Rainer Kiko, Meghan L Powers, Nathan C Shaner, and Steven H D Haddock. A comparison across non-model animals suggests an optimal sequencing depth for *de novo* transcriptome assembly. *BMC Genomics*, 14(1):167, 2013.

Figures & Tables

Fig. 1
Fig. 2