# 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 $<5$ or $<10$, is optimal for most studies across a wide variety of metrics.

## 1 Introduction

2 The popularity of genome-enabled biology has increased dramatically, particularly for researchers

3 studying non-model organisms, over the last few years. For many, the primary goal of these works is to

4 better understand the genomic underpinnings of adaptive [1, 2] or functional [3, 4] traits. While

5 extremely promising, the study of functional genomics in non-model organisms typically requires the

6 generation of a reference transcriptome to which comparisons are made. Although compared to

7 genome assembly [5, 6]. transcriptome assembly is less challenging, significant computational hurdles

8    still exist. Amongst the most difficult of challenges involves the reconstruction of isoforms [7] and

9    simultaneous assembly of transcripts where read coverage (=expression) varies by orders of magnitude.

10    These processes are further complicated by the error-prone nature of high-throughput sequencing

11    reads. With regards to Illumina sequencing, error is distributed non-randomly over the length of the

12    read, with the rate of error increasing from 5' to 3' end [8]. These errors are overwhelmingly

13    substitution errors [9], with the global error rate being between 1% and 3%. Although *de Bruijn* graph

14    assemblers do a remarkable job in distinguishing error from correct sequence, sequence error does

15    results in assembly error. While this type of error is problematic for all studies, it may be particularly

16    troublesome for SNP-based population genetic studies. In addition to the biological concerns,

17    sequencing read error may results in problems of a more technical importance. Because most

18    transcriptome assemblers use a *de Bruijn* graph representation of sequence connectedness, sequencing

19    error can dramatically increase the size and complexity of the graph, and thus increase both RAM

20    requirements and runtime.

21    In addition to sequence error correction, which has been shown to improved accuracy of the *de novo*

22    assembly [10], low quality (=high probability of error) nucleotides are commonly removed from the

23    sequencing reads prior to assembly, using one of several available tools (TRIMMOMATIC [11], FASTX

24    TOOLKIT (http://hannonlab.cshl.edu/fastx_toolkit/index.html), BIOPIECES

25    (http://www.biopieces.org/), SOLEXAQA [12]). These tools typically use a sliding window

26    approach, discarding nucleotides falling below a given (user selected) average quality threshold. The

27    trimmed sequencing read dataset that remains will undoubtedly contain error, though the absolute

28    number will surely be decreased.

29    Although the process of nucleotide quality trimming is commonplace in HTS analysis pipelines, it's

30    optimal implementation has not been well defined. Though the rigor with which trimming is

31    performed may be guided by the design of the experiment, a deeper understanding of the effects of

32    trimming is desirable. As transcriptome-based studies of functional genomics continue to become more

33    popular, understanding how quality trimming of mRNA-seq reads used in these types of experiments is

34    urgently needed. Researchers currently working in these field appear to favor aggressive trimming (e.g.

35    [13, 14]), 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 [15], 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 [11], 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 $\leq 2$, $\leq 5$, $\leq 10$, and $\leq 20$. Transcriptome assemblies were generated for each dataset using the default settings of the program TRINITY [16, 17]. 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

62  Quality trimming may have substantial effect on assembly quality, and as such, I sought to identify

63  high quality transcriptome assembles. Assemblies with few nucleotide errors relative to a known

64  reference may indicate high quality. The program BLAT [18] was used to identify and count nucleotide

65  mismatches between reconstructed transcripts and their corresponding reference. To eliminate spurious

66  short matches between query and template inflating estimates of error, only unique transcripts that

67  covered more than 90% of their reference sequence were used. Another potential assessment of

68  assembly quality may be related to the number of paired-end sequencing reads that concordantly map

69  to the assembly. As the number of reads concordantly mapping increased, so does assembly quality.

70  To characterize this, I mapped raw untrimmed sequencing reads to each assembly using Bowtie2 [19].

71  Aside from these metrics, measures of assembly content were also assayed. Here, open reading frames

72  (ORFs) were identified using the program TRANSDECODER

73  (http://transdecoder.sourceforge.net/), and were subsequently translated into amino acid

74  sequences. The larger the number of complete open reading frames (containing both start and stop

75  codons) the better the assembly. Lastly, unique transcripts were identified using the blastP program

76  within BLAST+ [20]. As the number of transcripts matching a given reference increases, so may

77  assembly content.


# Results

79  Quality trimming of sequence reads had a relatively small effect on the total number of errors

80  contained in the final assembly (Figure 1), and this effect is largely a function of the size of the

81  assembly, which was reduced by between 5 and 13% when comparing the assemblies of untrimmed

82  versus PHRED=20 trimmed sequence reads. Indeed, after normalizing error by assembly size in

83  nucleotides (Figure Y), trimming increased the error rate for low coverage (10M and 20M reads), and

84  had negligible effect on higher coverage datasets. This result was extremely surprising.
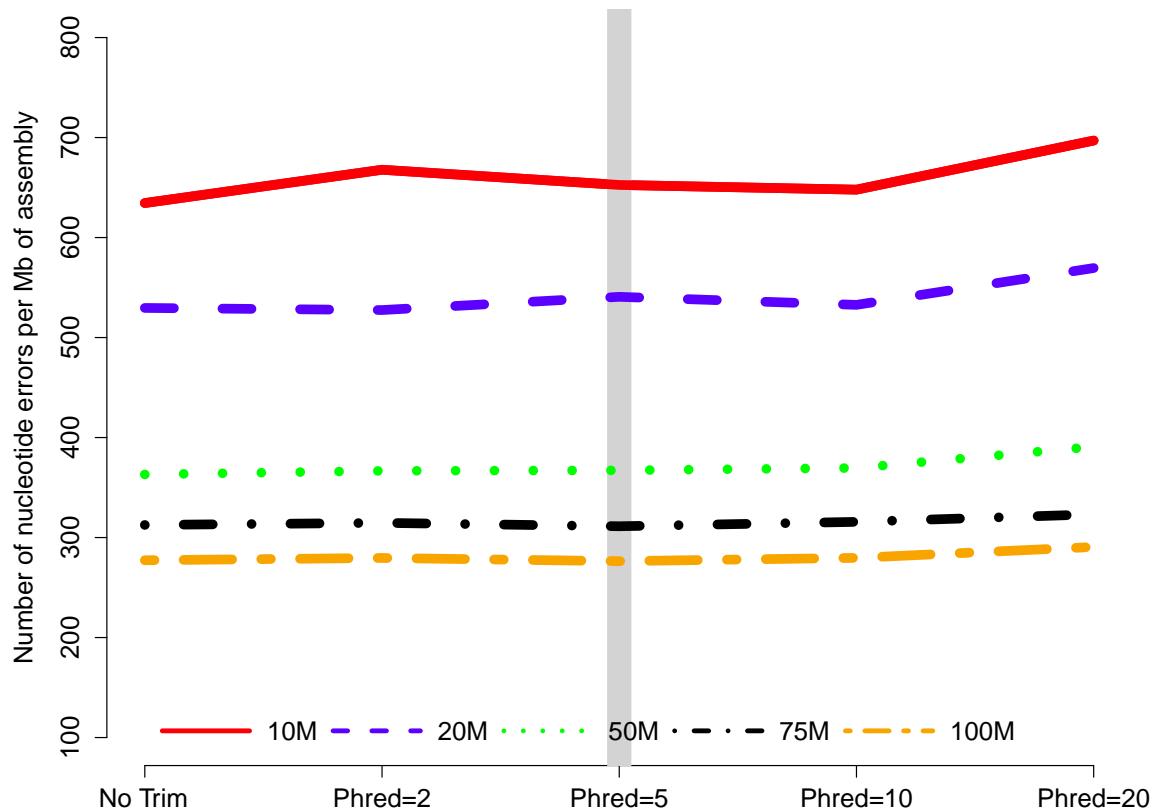
85  **Figure 1**

Figure 1. The number of nucleotide errors contained in the final transcriptome assembly, normalized to assembly size, is not strongly related to the strength of quality trimming (Trimming of nucleotides whose error scores are: PHRED >20, 10, 5, 2, or no trimming This patterns is largely unchanged with varying depth of sequencing coverage (10 million to 100 million sequencing reads). With higher coverage, trimming at PHRED = 5 may be optimal.

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 [21]. As such, the analysis of assembly quality included study of the mapping rates. Here, we found small but significant effects of trimming. Specifically, quality trimming at most levels decreased the proportion of reads that map concordantly to a given contig (Figure 2) in low coverage datasets, while when higher coverage,

97 mapping rates was equivalent. The exception to these patters occurred with trimming at $\textsc{Phred} =$

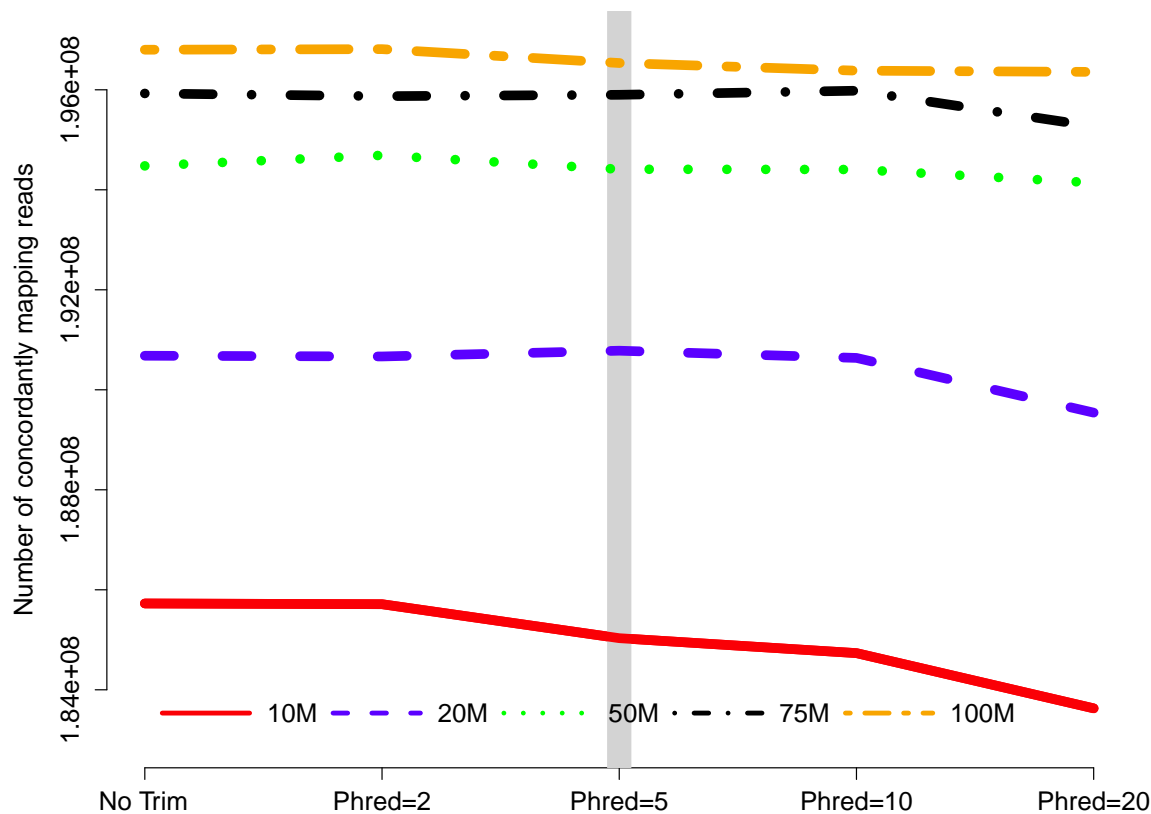98 20, which was universally linked to lower read mapping rates.

**Figure 2**



Figure 2. The number of concordantly mapping reads was reduced by trimming in low coverage datasets, while in higher coverage datasets where was no strong effect. The exception to this came with trimming at $\textsc{Phred}$=20 was always associated with much poorer mapping rates.

104 Analysis of assembly content painted a similar picture, with trimming having a relatively small, though

105 tangible effect. The number of $\textsc{Blast}+$ matches was essentially unchanged when looking at

106 untrimmed reads or reads trimmed at $\textsc{Phred}$=2,5, $\textsc{or}$10 (Figure 3), though trimming at

107 $\textsc{Phred}$=20 was always the most poorly performing.
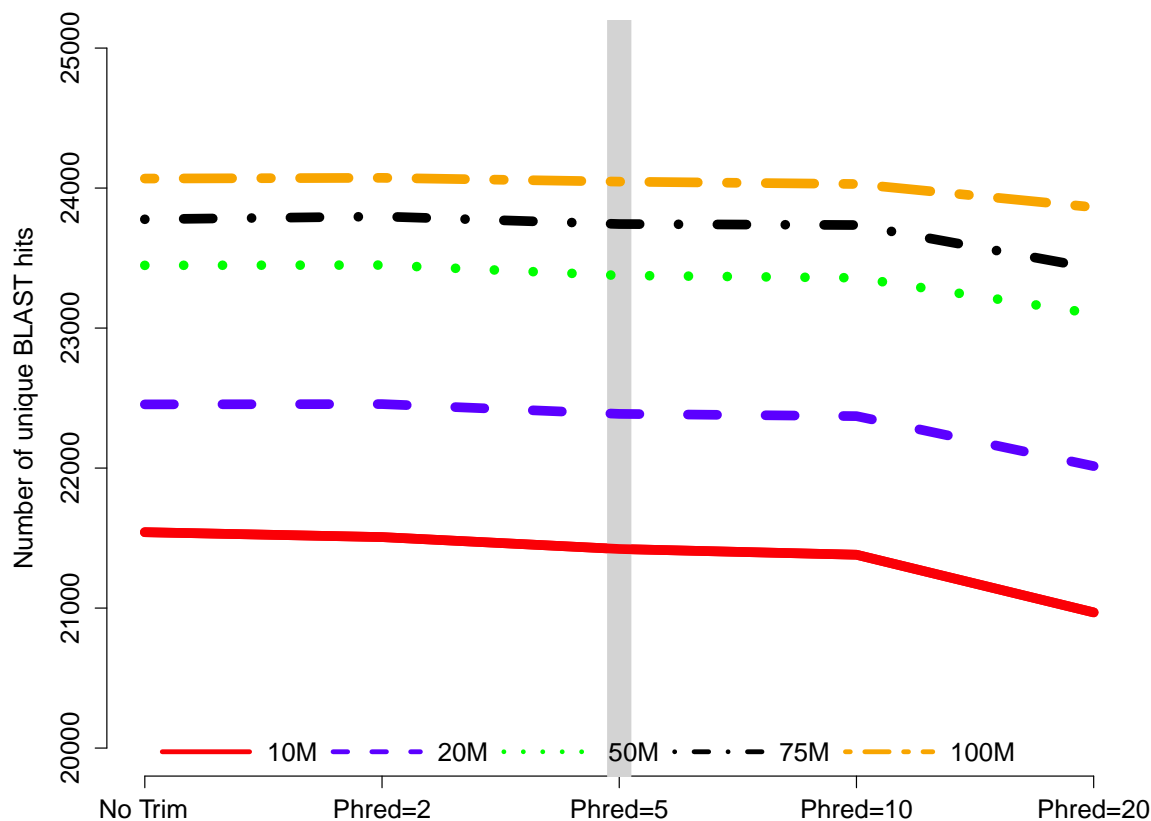
**Figure 3**



Figure 3. The number of unique SMALL CAPS BLAST matches contained in the final transcriptome assembly is not strongly related to the strength of quality trimming for any of the studies 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.
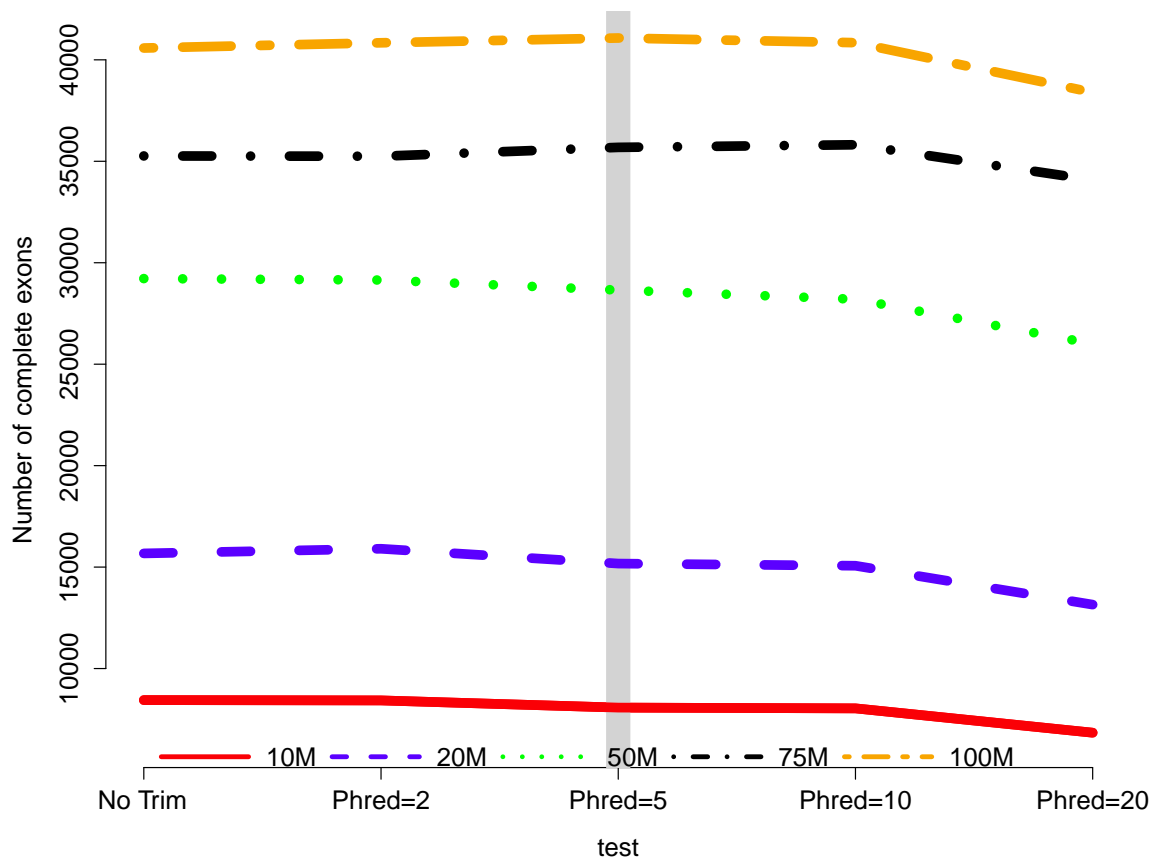
**Figure 4**

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 $\textsc{Phred}=20$ 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 [? ? ? ? ]. This study provides strong evidence that stringent quality trimming of nucleotides whose quality scores are $\geq 20$ 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.

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 $\textsc{Phred}=20$ was unexpected. To understand if trimming changes the distribution of quality scores along the read, we generated plots with the program SolexaQA [12]. Indeed, the program modifies the distribution of $\textsc{Phred}$ 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 $\textsc{Phred}=20$?

$\textsc{Effects of read depth}$ — 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 [**?** ], 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 $\textsc{BLAST}$ 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] C R Linnen, Y P Poh, B K Peterson, R D H Barrett, J G Larson, J D Jensen, and H 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, Nathan R Campbell, Kevin A Meyer, Michael R Miller, 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, J J Gonzalez-Plaza, A Canada, A M Blanco, M d C Garcia-Lopez, J M Rodriguez, L Pedrola, M D Sicardo, M L Hernandez, R De la Rosa, A Belaj, M Gil-Borja, F Luque, J M Martinez-Rivas, D G Pisano, O Trelles, V Valpuesta, 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, Ting-Ying Chien, Chia-Cheng Hu, Mei-Ju May Chen, Wen-Jer Wu, Hai-Tung Feng, David S Haymer, 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, Inan Birol, S bastien 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, l nie 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 D Jarvis, Huaiyang Jiang, Sergey Kazakov, Paul J Kersey, Jacob O Kitzman, James R Knight, Sergey Koren, Tak-Wah Lam, Dominique Lavenier, Fran ois 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, Oct vio 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. 2(1):1–1, July 2013.

[6] D Earl, K Bradnam, J St John, A Darling, D Lin, J Fass, H O K Yu, V Buffalo, D R Zerbino, M Diekhans, N Nguyen, P N Ariyaratne, W K Sung, Z Ning, M Haimel, J T Simpson, N A Fonseca, I Birol, T R Docking, I Y Ho, D S Rokhsar, R Chikhi, D Lavenier, G Chapuis, D Naquin, N Maillet, M C Schatz, D R Kelley, A M Phillippy, S Koren, S P Yang, W Wu, W C Chou, A Srivastava, T I Shaw, J G Ruby, P Skewes-Cox, M Betegon, M T Dimon, V Solovyev, I Seledtsov, P Kosarev, D Vorobyev, R Ramirez-Gonzalez, R Leggett, D MacLean, F Xia, R Luo, Z Li, Y Xie, B Liu, S Gnerre, I Maccallum, D Przybylski, F J Ribeiro, S Yin, T Sharpe, G Hall, P J Kersey, R Durbin, S D Jackman, J A Chapman, X Huang, J L DeRisi, M Caccamo, Y Li, D B Jaffe, R E Green, D Haussler, I Korf, 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, Hans Cheng, 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, J Yuan, S M Yiu, Z Li, Y Xie, Y Chen, Y Shi, H Zhang, Y Li, T W Lam, 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] Matthew D MacManes and Michael B Eisen. Improving transcriptome assembly through error correction of high-throughput sequence reads. *arXiv.org*, April 2013.

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

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

[13] Ana Riesgo, Alicia R Perez-Porro, Susana Carmona, Sally P Leys, 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.

[14] Mario Looso, Jens Preussner, Konstantinos Sousounis, Marc Bruckskotten, Christian S Michel, Ettore Lignelli, Richard Reinhardt, Sabrina Hoeffner, Marcus Krueger, Panagiotis A Tsonis, Thilo Borchardt, 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.

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

[16] Brian J Haas, Alexie Papanicolaou, Moran Yassour, Manfred Grabherr, Phillip Blood, Joshua Bowden, Matthew Couger, David Eccles, Bo Li, Matthias Lieber, Matthew D MacManes, Michael Ott, Joshua Orvis, Nathalie Pochet, Francisco Strozzi, Nathan Weeks, Rick Westerman, Thomas William, Colin N Dewey, Robert Henschel, Richard G LeDuc, Nir Friedman, and Aviv Regev. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature protocols*, pages 1494–1512, 2013.

[17] Manfred G Grabherr, Brian J Haas, Moran Yassour, Joshua Z Levin, Dawn A Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, Z 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.

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

[19] 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–U174, 2010.

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

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

# Figures & Tables

Fig. 1

Fig. 2