MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2016) **16**, 1315–1321 doi: 10.1111/1755-0998.12529

Evaluation of TagSeq, a reliable low-cost alternative for RNAseq

BRIAN K. LOHMAN, JESSE N. WEBER[a] and DANIEL I. BOLNICK
Department of Integrative Biology, University of Texas at Austin, One University Station C0990, Austin TX 78712, USA

Abstract

RNAseq is a relatively new tool for ecological genetics that offers researchers insight into changes in gene expression in response to a myriad of natural or experimental conditions. However, standard RNAseq methods (e.g., Illumina TruSeq® or NEBNext®) can be cost prohibitive, especially when study designs require large sample sizes. Consequently, RNAseq is often underused as a method, or is applied to small sample sizes that confer poor statistical power. Low cost RNAseq methods could therefore enable far greater and more powerful applications of transcriptomics in ecological genetics and beyond. Standard mRNAseq is costly partly because one sequences portions of the full length of all transcripts. Such whole-mRNA data are redundant for estimates of relative gene expression. TagSeq is an alternative method that focuses sequencing effort on mRNAs' 3′ end, reducing the necessary sequencing depth per sample, and thus cost. We present a revised TagSeq library construction procedure, and compare its performance against NEBNext®, the 'gold-standard' whole mRNAseq method. We built both TagSeq and NEBNext® libraries from the same biological samples, each spiked with control RNAs. We found that TagSeq measured the control RNA distribution more accurately than NEBNext®, for a fraction of the cost per sample (~10%). The higher accuracy of TagSeq was particularly apparent for transcripts of moderate to low abundance. Technical replicates of TagSeq libraries are highly correlated, and were correlated with NEBNext® results. Overall, we show that our modified TagSeq protocol is an efficient alternative to traditional whole mRNAseq, offering researchers comparable data at greatly reduced cost.

Keywords: 3′Tag-based sequencing, ecological genetics, RNAseq

Received 14 December 2015; revision received 18 March 2016; accepted 18 March 2016

Introduction

RNAseq has been widely used to describe differences in gene expression among wild populations, as well as changes in captive or wild individuals' expression following exposure to stimuli (mates, predators, parasites, abiotic stress, toxins). This work has helped uncover the genetic basis of complex traits, implicate genes underlying targets of natural selection, and measure the heritable and environmental components of variation in gene expression (Pickrell *et al.* 2010; Lenz *et al.* 2013; Barribeau *et al.* 2014; Foth *et al.* 2014; Lovell *et al.* 2015; Videvall *et al.* 2015). However, the most widely used RNAseq protocols are cost-prohibitive for many biologists, including but not limited to researchers in ecological genetics.

Correspondence: Brian K. Lohman, E-mail: lohman@utexas.edu
[a]Current address:Division of Biological Sciences, University of Montana, Missoula, MT 59812, USA.

Construction of any whole mRNAseq library for the Illumina platform (including Illumina TruSeq® and NEBNext® kits) involves isolating or enriching for mRNA, which is then fragmented and subject to massively parallel sequencing. The resulting data yields sequences for overlapping portions of the entire lengths of the original messenger RNAs (hence 'whole' mRNAseq). This requires high depth of coverage; although sequencing requirements vary depending on sample type, the ENCODE Consortium suggests ~30 million raw reads per sample as a 'best practice' for most RNAseq experiments (The ENCODE Consortium, 2011), limiting researchers to a maximum of eight samples per lane of Illumina HiSeq 2500. The high cost of sequencing, combined with high cost of library construction, has forced many studies to use small sample sizes, or pool samples within treatments. This is cause for concern, as meaningful differences in gene expression simply may not be detected with such low-powered sampling designs, and pooled RNAseq may fail to properly account for residual variation in expression (Todd *et al.* 2016).

To resolve problems with whole mRNAseq, several low-cost alternatives have been developed. Most notably, Meyer *et al.* (2011) presented a 3′ Tag-based approach to RNAseq, called TagSeq, that requires little input RNA, involves low library construction costs, and requires many fewer raw reads per sample. By focusing on the 3′ end of mRNA fragments, TagSeq reduces the sequencing effort required to characterize a population of mRNAs in a biological sample. This cost-saving does come with some constraints: TagSeq cannot distinguish between alternatively spliced transcripts from a single locus, and will not identify polymorphism or allele-specific expression in much of a gene's coding sequence. Furthermore, while it may be possible to use TagSeq for genotyping we choose not to explore that issue here, focusing instead on gene expression. The benefits of precisely measuring locus-level transcriptional differences with high replication may outweigh the lack of splicing or SNP information for many experiments in ecological systems. However, as presented in Meyer *et al.* (2011), TagSeq uses a number of outdated methods and enzymes, which may skew the distribution of RNA fragments in the library, with respect to both fragment size and GC content (Aird *et al.* 2011). In addition, the accuracy of TagSeq has not yet been compared to the industry standard TruSeq®/NEBNext® which reliably measures moderate and high abundance mRNAs in a sample (New England BioLabs, NEB Next Ultra Directional RNA Library Prep Kit for Illumina (NEB #E7420S/L), Ipswich, MA, USA).

Here, we present a modified protocol intended to increase the accuracy and precision of TagSeq, by incorporating recent findings on polymerase performance (Aird *et al.* 2011), fragmentation methods and bead-based purification technology into the library construction process. We then tested the accuracy of TagSeq against the industry standard NEBNext® by sequencing technical replicates of a biological sample, each containing an artificial set of diverse RNAs of known concentration, designed by the External RNA Controls Consortium (hereafter simply 'ERCC').

Materials and methods

Improvements to TagSeq library construction

Briefly, our improved TagSeq library construction method involves 11 steps: (i) isolate total RNA, (ii) remove genomic DNA with DNase (if not included in total RNA isolation), (iii) fragment total RNA with Mg+ buffer via hydrolysis (NEB), (iv) synthesize cDNA with a poly-dT oligo, (v) PCR amplify cDNA, (vi) purify PCR products with DNA-binding magnetic beads [Agencourt, or made in-house (Rohland & Reich 2012)], (vii) fluormetrically quantify PCR products (PicoGreen, Life Technologies), (viii) normalize among-sample concentrations, (ix) add sample-specific barcodes via PCR, (x) pool samples and select a small range of fragment sizes (to maximize output on the Illumina platform) via automated gel extraction (400–500 bp, Sage Pippin Prep 2% agarose), (xi) quantify concentrations of postextraction products via Qubit (dsDNA high sensitivity following the manufacture's instruction), (xii) normalize among pools. This protocol can be completed by a single researcher in three days, and this approach is optimized for 96-well format plates. Improvements over the original protocol are described in Table 1.

Sample acquisition and library preparation

Total RNA was extracted (with Ambion AM1912) from six freshly isolated stickleback (*Gasterosteus aculeatus*) head kidneys stored in RNAlater (Ambion). All fish were lab-raised, nonsibling, nongravid females, bred via *in vitro* crosses of wild caught parents. Three fish originated from crosses between parents from Gosling Lake, British Columbia and three fish from crosses between parents from Roberts Lake, British Columbia (two populations of the same species). Total RNA from all six head kidneys was then split, and libraries were constructed with both whole mRNAseq and TagSeq. The Genomic Sequencing and Analysis Facility at the University of Texas at Austin prepared whole mRNAseq libraries (NEBNext® directional RNA libraries with poly-A enrichment), according to the manufacturer's instruction (New England BioLabs, NEB Next Ultra Directional RNA Library Prep Kit for Illumina (NEB #E7420S/L), Ipswich, MA, USA), for four of the six RNA samples (two samples dropped to ensure adequate sequencing depth). ERCC (2 μL of 1:100 dilution for every 1 μg of total RNA) was added to each sample before library construction began, according to the manufacturer's instructions. Whole mRNAsq samples were sequenced on a single lane of Illumina HiSeq 2500 2 × 100, producing an average of 40.5 million paired-end reads per sample (81 million reads total per sample). Following the addition of ERCC to one technical replicate per biological sample, TagSeq samples were prepared according to Meyer *et al.* (2011), but with changes detailed in Table 1. Four TagSeq samples had two technical replicates (totally independent library builds from total RNA), a fifth sample had three technical replicates, and a sixth sample had no technical replicates. 12 TagSeq libraries were sequenced using three partial lanes of HiSeq 2500 1 × 100 (average of 10.3 million raw reads per sample).

Bioinformatics: How many genes are identified by each method?

Raw whole mRNAseq reads were trimmed with CUTADAPT v 1.3 (Martin 2011) to remove any adapter

Table 1 Changes to Meyer *et al*. (2011). We identified a number of areas where the Meyer protocol could be improved and implemented changes to address these concerns

Meyer *et al*. (2011)	Problem	Our approach
Measured DNA/RNA via spectroscopy	Quantification of DNA/RNA by spectroscopy is inaccurate.	Fluorescent based quantification with Quant-iT assays.
Included DNase with total RNA extraction	Genomic DNA contamination leads to nonspecific amplification.	Increase DNase treatment to 1.5× concentration at 37 °C for 1 h.
Heat and Tris buffer to fragment total RNA	Fragmentation of total RNA with Tris buffer produces a wide distribution of fragment sizes.	Precisely fragment total RNA with a specialized Mg2+ buffer.
No normalizations	Yield of first strand synthesis is too variable.	Normalize RNA input to 1 μg.
Titanium Taq (Clontech)	Variable GC content among fragments can cause dropout of transcripts (Aird *et al*. 2011).	Use AccuPrime Taq polymerase and associated thermal profile for PCR steps (Aird *et al*. 2011).
Suggest users limit PCR to 15–17 cycles or less	Excessive PCR amplification increases the number of PCR duplicates.	Reduce number of PCR cycles to 12 or less.
PCR cleanups with spin columns	Purification using solid-phase methods (e.g. spin columns) is not high throughput compatible, inefficient and costly.	Clean with Agencourt AMPure beads, which can be made in-house (Rohland & Reich 2012)
No normalizations	Post-PCR cDNA amplification yield is highly variable.	Normalize input to 40 ng total.
Size selection on a per sample basis via gel extraction	Size selection by standard gel extraction is highly variable.	Precise size selection with Pippin Prep automated gel extraction.
Mix individuals samples after qPCR of every library	Mixing individual libraries based on qPCR is slow and expensive.	Normalize lanes of Pippin Prep with Qubit Fluorimeter.

contamination. We then mapped the trimmed reads to version 79 of the stickleback genome (with ERCC sequences appended) using BWA-MEM (Li & Durbin 2010), and counted genes using Bedtools (Quinlan & Hall 2010), producing 20 678 total genes. TagSeq reads were processed according to the iRNAseq pipeline (https://github.com/z0on/tag-based_RNAseq) (Dixon *et al*. 2015), producing 19 145 total genes.

Statistical analysis of control transcripts: How accurately does each method estimate a known distribution?

For each sample, we plotted observed counts of artificial ERCC transcripts against expected values, fitting a simple linear model (observed ~ expected). We tested for a difference in mean adjusted R^2 value between library construction methods with a paired *t*-test (paired by biological sample).

We calculated the Spearman correlation between observed logtransformed counts of ERCC transcripts and expected transcript quantity. We tested for a difference in mean Rho values between library construction methods using a paired *t*-test. We also considered Rho separately for abundance quartiles.

Statistical analysis of stickleback transcripts: How similar are estimates of biological RNAs within and between methods?

We calculated the mean Spearman correlation among TagSeq technical replicates (n = 5, calculate Rho for each biological sample and average). We calculated the Spearman correlation between stickleback head kidney samples which had been prepared using both library construction methods.

Statistical analysis of inline barcodes: Is the current system for detecting PCR duplicates working?

TagSeq, as presented by Meyer *et al*. (2011) and here, uses degenerate inline barcodes on the 5′ end of each fragment to identify PCR duplicates. We tested for the random incorporation of these barcodes with a chi-squared test. We also tested for the effect of increased GC content within each barcode on the number of times that barcode was observed with a Poisson GLM. All statistical analyses were carried out in base R (R Development Core Team, 2007).

Results

We found that, when fitting a linear model between the expected concentrations of ERCC to observed transcript counts, TagSeq had a significantly higher mean adjusted R^2 value (R^2 = 0.89) than NEBNext® (R^2 = 0.80, Fig. 1, observed ~ expected, paired *t*-test, t = 18.63, d.f. = 3, $P < 0.001$). Similarly, the rank correlation between observed and expected ERCC fragments was consistently higher for TagSeq (mean Rho = 0.94) than NEBNext® (mean Rho = 0.87, Fig. 2, paired *t*-test, t = 12.20, d.f. = 3, P = 0.001). TagSeq showed higher mean Rho values for all abundance classes except the

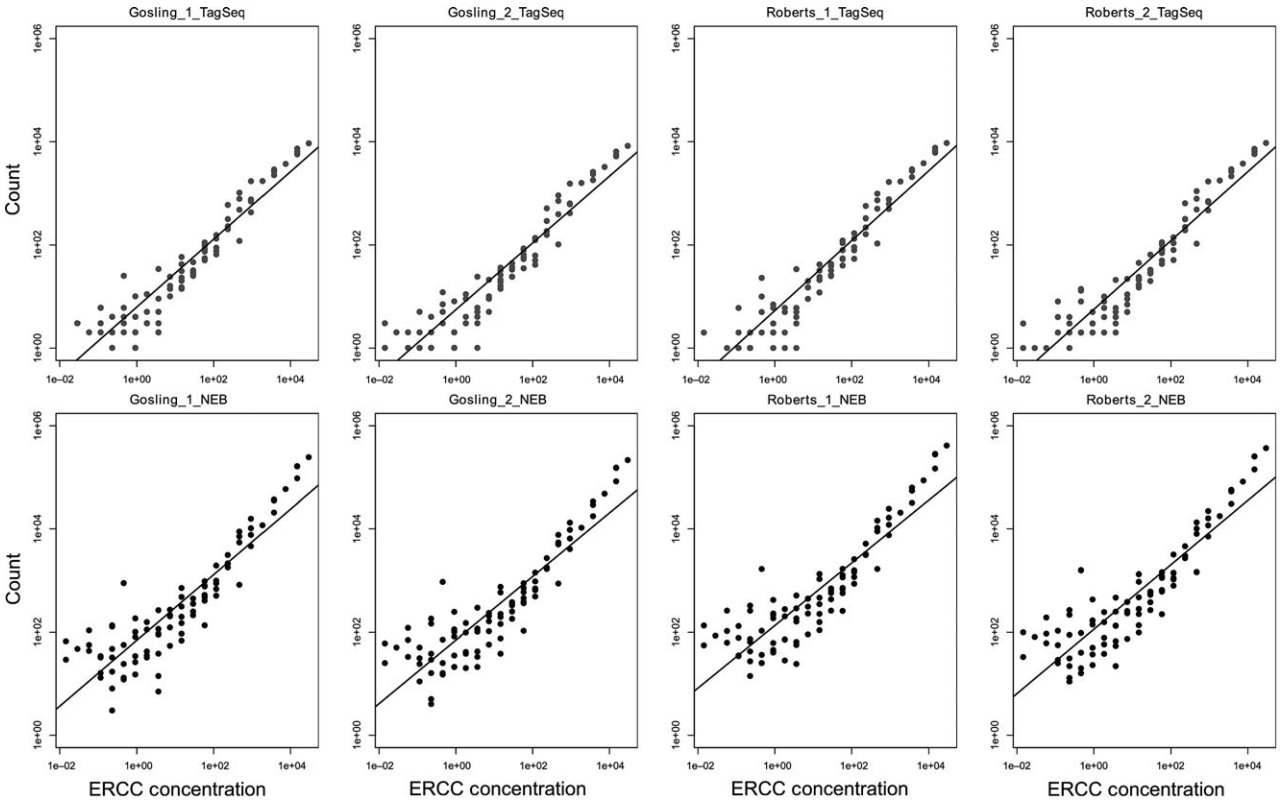

Fig. 1 Regression of observed vs. expected ERCC transcripts shows TagSeq has higher adjusted R^2 values for four different biological samples prepared with both methods (paired *t*-test, $t = 18.63$, d.f. = 3, $P < 0.001$).

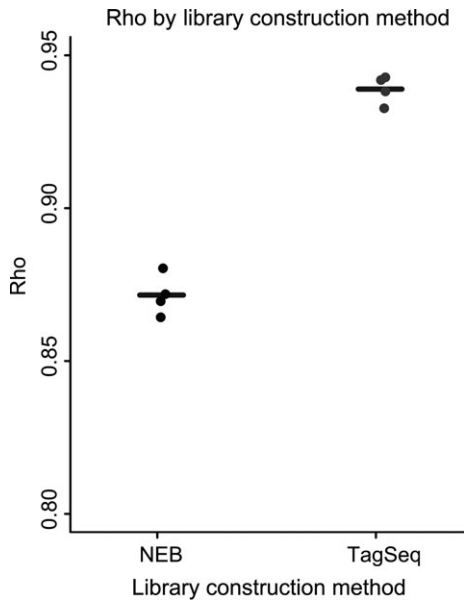

Fig. 2 TagSeq more accurately recovers a known distribution of control mRNA fragments (ERCC) than whole mRNAseq (mean Rho for TagSeq is higher than mean Rho for whole mRNAseq, paired *t*-test, $t = 12.20$, d.f. = 3, $P = 0.001$).

third quartile. Most notably, whole mRNAseq performed very poorly in the lowest abundance class (relative concentration of 0.014–0.45 attamols/μL), and TagSeq substantially outperformed whole mRNAseq in the second abundance class (relative concentration of 0.92–7.3 attamols/μL, Fig. 3).

With respect to stickleback (noncontrol) transcript counts, the mean Rho among technical replicates of TagSeq samples was 0.96. Due to the high cost of NEBNext® library generation and sequencing (~$340 per sample), we did not perform technical replicates using this method. We found a strong significant positive correlation between stickleback gene counts generated with TagSeq and whole mRNAseq (Rho = 0.74, $P < 0.001$). This is likely an underestimate of the actual correlation between the two library construction methods because whole mRNAseq performs very poorly when RNAs are in moderate to low abundance (first and second abundance classes, Fig. 3). Given that 9572 loci are in the bottom half of gene counts, even small differences in absolute counts between the methods will strongly influence the rank-based statistic.

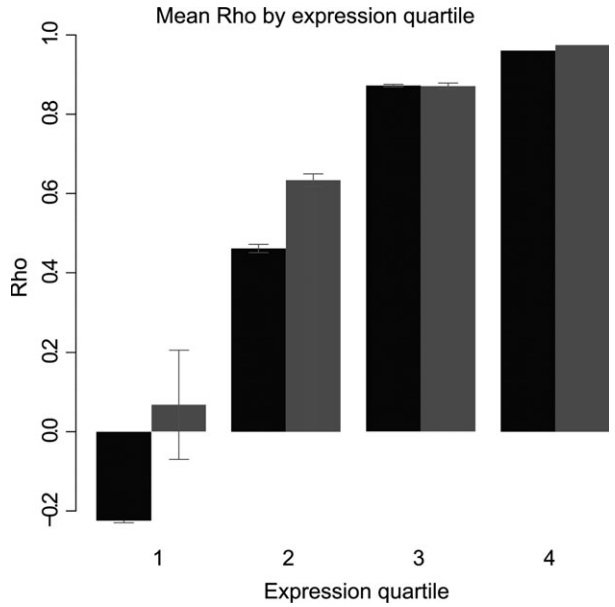

Fig. 3 Breakdown of control mRNAs by abundance class shows that TagSeq recovers mRNAs better than TruSeq, especially at lower abundances. Light grey bars are TagSeq, dark grey bars are whole mRNAseq. Fences indicate standard error.

We also wished to compare our new method with that of the original TagSeq protocol, but cannot make a direct comparison with the available samples. Meyer *et al.* (2011) evaluated their accuracy by comparing fold-differences in differentially expressed genes (between experimental treatments), whereas we measured accuracy using estimates of relative ERCC abundance. Keeping in mind these different benchmarking methods, we can draw a rough comparison. The original TagSeq method yielded a correlation of $r = 0.86$ between TagSeq estimates of fold-change expression, and qPCR measures of the same fold change (a 'known' benchmark). In contrast, our protocol yields a correlation of Rho = 0.94 between our relative abundance estimates, and the known ERCC relative abundances. We infer that the new protocol performs at least as well, and probably better, than the previous protocol, at generating expression level estimates that resemble known values.

The iRNAseq pipeline includes the removal of PCR duplicates, which are a common problem in many library construction methods (Aird *et al.* 2011). Any reads which meet two criteria are called PCR duplicates and removed: (i) identical in-line barcodes (the four degenerate bases at the start of each read), and (ii) the first 30 bases of sequence after the in-line barcode are identical. The removal of PCR duplicates substantially reduces the number of TagSeq reads in each library (mean reduction of 70.3%, $n = 12$). However, this avoids potential bias introduced by PCR, namely over-representation of smaller fragments. We found that inline barcodes were incorporated nonrandomly (Chi Square = 10 500 000, d.f. = 63, $P \ll 0.001$). We found that increased GC content in the inline barcode significantly reduced the number of times a barcode was observed. For every G or C added to the inline barcode, the expected value of the number of observed barcodes is reduced by ~2.9% (count ~ gcContent, family = poisson, $\beta_{\text{gcContent}} = -0.133$, $P < 0.001$).

Discussion

We present a number of methodological improvements to the TagSeq method of Meyer *et al.* (2011), and take the important next step of comparing the new protocol to the NEBNext® kit, the industry-standard for whole mRNAseq. Overall, our results illustrate that the updated TagSeq method offers researchers the ability to dramatically increase sample sizes in gene expression analyses, which will greatly increase statistical power to detect subtle differences in transcript abundance than traditional whole mRNAseq methods. Furthermore, TagSeq promises improved accuracy when measuring medium and low abundance RNAs. We speculate that this increase in sensitivity derives from a more efficient distribution of reads among loci, probably due to a reduced connection between gene length and transcripts counted per locus. In total RNAseq, even if two transcripts are expressed at identical levels, random fragmentation and priming leads to greater representation of longer fragments in sequencing libraries (Trapnell *et al.* 2010; Roberts *et al.* 2011). In contrast, TagSeq only primes the 3′ poly-A tail, generating an essentially uniform distribution of fragments with respect to original RNA length.

While TagSeq has been used predominantly in corals (Des Marais *et al.* 2015; Dixon *et al.* 2015), it should be applicable to nearly all metazoans. However, we caution researchers to perform several basic checks during TagSeq library construction, most especially ensuring the size distribution of RNA fragments is as narrow as possible during total RNA fragmentation. We recommend evaluating the results of various total RNA fragmentation times via BioAnalyzer. Fragments should be larger than 100 bp and smaller than 500 bp (see Appendix S1, Supporting information). Here, we were interested in evaluating the robustness of the TagSeq method for threespine stickleback, and therefore sequenced stickleback transcripts more deeply than required for an accurate estimate of gene expression across the majority of expressed loci (we generated an average of 10.3 million raw reads per sample). We recommend that researchers dedicate roughly ~5–6 M raw reads per sample if the goal is to measure the top 75% of all expressed mRNAs in a sample, as this has produced sufficient gene counts for

robust statistical power in an invertebrate, a plant and stickleback (M. Matz and T. Jeunger, personal communications).

In this project, we intentionally under-loaded our TagSeq libraries on the HiSeq lane by 15% (0.0017 pmols loaded), anticipating that low base diversity in the 5′ end of the fragments (caused by the inline barcode used to detect and remove PCR duplicates) would lead to poor clustering. However, quality metrics from the HiSeq run indicate that this is not a problem. We observed ~500–600 clusters per mm^2 on each tile, and the majority of these clusters passed filtering (low base diversity or overclustering would generate large numbers of clusters with few passing filtering). We therefore recommend that users load standard amounts of library (or even 10–20% extra material) on each lane of HiSeq (see Appendix S1, Supporting information). Specifically, overloading TagSeq libraries may help to increase raw read yield, relative to NEBNext®, which we found to optimally cluster when 0.002 pmols were loaded (~1000 clusters per mm^2). We also emphasize that small fragments need to be removed from TagSeq libraries, as they will more easily cluster on the HiSeq, reducing read output. These may be identified by BioAnalyzer and removed with additional bead clean-ups.

Several of our methodological changes aimed to mitigate the number of PCR duplicates, which are artefacts of all PCR-related methods. First, we predicted that increasing the degenerate inline barcodes from four to six bases would not only increase our ability to detect independent transcripts from PCR duplicates, but also increase base diversity on the 5′ end of fragments, thereby increasing the number of clusters passing Illumina's quality filters. However, this alteration did not completely remove the problem of PCR duplicates or increase the number of raw reads generated in each lane (data not shown). In the future we recommend that protocol users consider adding 3-nitropyrrole to the inline barcode region, as this should better randomize which bases are incorporated during initial round of PCR (Schweyen *et al.* 2014). Second, we limited our number of PCR cycles to 12. Empirically testing the effects of PCR cycle number on TagSeq accuracy was outside the scope of the present study. However, it is widely accepted that the best way to limit bias is to reduce the number of PCR cycles during cDNA amplification as much as possible (Aird *et al.* 2011).

Because TagSeq reads map to relatively small, 3′ regions of genes, when paralogs harbour few sequence differences, this method may only be able to distinguish patterns of expression at the gene family level. Although it is worth being aware of this limitation, paralogy problems are not unique to TagSeq. Quantifying the expression levels of paralogs (or copy number variants) can be very difficult; it often necessitates high-quality reference genome sequences, and factors such as the frequency, evolutionary age and expression differences between duplicated loci can greatly complicate analyses even when whole mRNA sequences are available.

In summary, we show that the improved TagSeq method has both benefits and drawbacks compared to traditional whole mRNA sequencing. While our TagSeq libraries did not generate optimal numbers of clusters on the HiSeq platforms, we identify several potential solutions to the problem. Regardless of the slightly lower number of raw reads, our improved TagSeq method is overwhelmingly more cost effective than whole mRNA-seq. At maximal efficiency (32 individuals per sequencing lane), our method was able to produce highly accurate, transcriptome-wide gene counts for only ~$33 per sample (down from Meyer *et al.* (2011)'s ~$50 per sample), including sequencing costs (HiSeq 2500 1 × 100 V3 chemistry with ~5.6 M raw reads per sample). This low cost and high reliability offers molecular ecologists the opportunity to vastly increase sample sizes and increase replication to uncover new patterns in gene expression.

Acknowledgements

All live animal research was approved by the UT Austin IACUC [protocol AUP-2013-00012]; and collections were approved by the British Columbia Ministry of Environment [Scientific Fish Collection permit NA09-52421]. We wish to thank Mikhail Matz, Marie Strader, and the Juenger lab for fruitful discussion on improvements to the TagSeq method both during library construction and analysis. Figures 2 and 3 were generated using plotting functions written by Luke Reding (https://github.com/lukereding/redingPlot). This work was supported by the Howard Hughes Medical Institute (DIB).

References

Aird D, Ross MG, Chen WS *et al.* (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology*, **12**, R18.

Barribeau SM, Sadd BM, du Plessis L, Schmid-Hempel P (2014) Gene expression differences underlying genotype-by-genotype specificty in a host-parasite system. *Proceeding of the National Academy of Sciences*, **111**, 3496–3501.

Des Marais DL, Skillern WD, Juenger TE (2015) Deeply diverged alleles in the Arabidopsis AREB1 transcription factor drive genome-wide differences in transcriptional response to the environment. *Molecular Biology and Evolution*, **32**, 956–969.

Dixon GB, Davies SW, Aglyamova GA, Meyer E, Bay LK, Matz MV (2015) Genomic determinants of coral heat tolerance across latitudes. *Science*, **348**, 1460–1462.

Foth BJ, Tsai IJ, Reid AJ *et al.* (2014) Whipworm genome and dual-species transcriptome analyses provide molecular insights into an intimate host-parasite interaction. *Nature Genetics*, **46**, 693–700.

Lenz TL, Eizaguirre C, Rotter B, Kalbe M, Milinski M (2013) Exploring local immunological adaptation of two stickleback ecotypes by

experimental infection and transcriptome-wide digital gene expression analysis. *Molecular Ecology*, **22**, 774–786.

Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, **26**, 589–595.

Lovell JT, Mullen JL, Lowry DB *et al.* (2015) Exploiting differential gene expression and epistasis to discover candidate genes for drought-associated QTLs in *Arabidopsis thaliana*. *The Plant Cell*, **27**, 969–983.

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet Journal*, **17**, 10–12.

Meyer E, Aglyamova GV, Matz MV (2011) Profiling gene expression responses of coral larvae (*Acropora millepora*) to elevated temperature and settlement inducers using a novel RNA-Seq procedure. *Molecular Ecology*, **20**, 3599–3616.

Pickrell JK, Marioni JC, Pai AA *et al.* (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, **464**, 768–772.

Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

R Development Core Team (2007) *R: a language and environment for statistical computing*. R Foundation for Statistical Computing: Vienna, Austria.

Roberts A, Trapnell C, Donaghey J, Rinn JL, Pachter L (2011) Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biology*, **12**, R22.

Rohland N, Reich D (2012) Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Research*, **22**, 939–946.

Schweyen H, Rozenberg A, Leese F (2014) Detection and removal of PCR duplicates in population genomic ddRAD studies by addition of a degenerate base region (DBR) in sequencing adapters. *The Biological Bulletin*, **227**, 146–160.

The ENCODE Consortium (2011) *Standards, Guidelines, and Best Practices for RNA-seq*. Available from: https://genome.ucsc.edu/ENCODE/protocols/dataStandards/ENCODE_RNAseq_Standards_V1.0.pdf.

Todd EV, Black MA, Gemmell NJ (2016) The power and promise of RNA-seq in ecology and evolution. *Molecular Ecology*, **25**, 1224–1241.

Trapnell C, Williams BA, Pertea G *et al.* (2010) Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms. *Nature Biotechnology*, **28**, 511.

Videvall E, Cornwallis CK, Palinauskas V, Valkiūnas G, Hellgren O (2015) The avian transcriptome response to malaria infection. *Molecular Biology and Evolution*, **32**, 1255–1267.

B.K.L., J.N.W. and D.I.B. jointly designed the research. B.K.L. carried out all library construction improvements. B.K.L., and J.N.W. analysed data. B.K.L. wrote the manuscript, with comments from J.N.W. and D.I.B. All authors approved the final version.

Data accessibility

Meta data, code for raw read processing, gene counts, code for statistical analysis and plotting of data, BioAnalyzer .XAD files, and HiSeq quality metrics, and detailed protocol are located in DRYAD entry: http://dx.doi.org/10.5061/dryad.vq275. Raw sequence reads are available on Corral, a permanent data repository with multiple, independent backups, located and owned by the University of Texas at Austin Texas Advanced Computing Center. Users can download data via 'wget http://web.corral.tacc.utexas.edu/Lohman_et_al_2016/*' from the terminal.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Appendix S1 Detailed TagSeq protocol as described in Lohman *et al.* 2016