Error, noise and bias in *de novo* transcriptome assemblies

Adam H. Freedman*, Faculty of Arts and Sciences Informatics Group, Harvard University, Cambridge, Massachusetts 02138

Michele Clamp, Faculty of Arts and Sciences Informatics Group, Harvard University, Cambridge, Massachusetts 02138

Timothy B. Sackton, Faculty of Arts and Sciences Informatics Group, Harvard University, Cambridge, Massachusetts 02138

* Corresponding author: adamfreedman@fas.harvard.edu
ABSTRACT

De novo transcriptome assembly is a powerful tool, widely used over the last decade for making evolutionary inferences. However, it relies on two implicit assumptions: that the assembled transcriptome is an unbiased representation of the underlying expressed transcriptome, and that expression estimates from the assembly are good, if noisy approximations of the relative abundance of expressed transcripts. Using publicly available data for model organisms, we demonstrate that, across assembly algorithms, species, and data sets, these assumptions are consistently violated. Bias exists at the nucleotide level, with genotyping error rates ranging from 30-83%. As a result, diversity is underestimated in transcriptome assemblies, with consistent under-estimation of heterozygosity in all but the most inbred samples. Even at the gene level, expression estimates show wide deviations from map-to-reference estimates, and positive bias at lower expression levels. Standard filtering of transcriptome assemblies improves the robustness of gene expression estimates but leads to the loss of a meaningful number of protein coding genes, including many that are highly expressed. We demonstrate a computational method to partly alleviate noise and bias in expression estimates. Researchers should consider ways to minimize the impact of bias in transcriptome assemblies.
1. INTRODUCTION

The development of computational methods for *de novo* transcriptome assembly (Birol et al., 2009; Grabherr et al., 2011a) has empowered a wide array of evolutionary analyses in the absence of a reference genome. Their application has led to significant advances in fields as diverse as phylogenetics (Schwentner, Combosch, Pakes Nelson, & Giribet, 2017), evolutionary responses to climate change (Veilleux et al., 2015), the molecular basis of convergent evolution (Berens, Hunt, & Toth, 2015; Pankey, Minin, Imholte, Suchard, & Oakley, 2014; Velotta et al., 2017), speciation (K. Li et al., 2016), and local adaptation (De Wit & Palumbi, 2013).

All studies using *de novo* transcriptome assemblies rely on two implicit assumptions: that the assembled transcriptome represents an unbiased, if incomplete, representation of the true underlying expressed transcriptome, and that the expression estimates from the assembly are good, if noisy, approximations of the relative abundance of expressed transcripts. Indirect evidence suggests that violations of these assumptions may be frequent. Comparisons of different assembly algorithms indicate that there is a degree of complementarity in missingness—one assembler will recover a gene that another cannot, motivating methods to combine assemblies, e.g. the Oyster River Protocol (MacManes, 2018). Highly fragmented assemblies comprised of hundreds of thousands of relatively short contigs are pervasive in the literature, presumably representing transcript fragments, haplotypes erroneously designated as separate transcripts, or both. Intuitively, the challenge of mapping reads to fragmented gene models should diminish the robustness of expression estimates. Unknown is whether the compositional aspects of transcriptome assemblies negatively impact SNP calls, allele frequency estimates, and other population genetic parameters.
Below, we evaluate the assumptions inherent to de novo transcriptome assembly, discuss how their violation might distort evolutionary inferences, identify particular features of transcriptome assemblies that should be considered during the study design phase of a project, and demonstrate a method for correcting bias in expression estimates. While phylogenomic studies can discard the majority of contigs using aggressive filtering schemes, these are less appropriate for studies that seek to understand patterns of genetic variation or gene expression across populations or closely related species. Therefore, we focus on methodological considerations for this class of investigations.

2. MATERIALS AND METHODS

2.1 RNA-seq data

Data sets include the Mus musculus (C57BL/6J) dendritic cell data used in the original TRINITY paper (MDC), a pool of six whole brains from albino inbred Mus (BALB/c), and 8-sample pools of whole brain from wild M. musculus domesticus from Massif Central, France (FRA), Iran (IRN), Kazakhstan (KZK), and Germany (DEU). To assess the generality of particular results with respect to assembly composition, we also generated TRINITY assemblies for pufferfish (Tetraodon nigroviridis) whole brain, clawed frog (Xenopus tropicalis) kidney, and fly (Drosophila melanogaster) heads. SRA accessions are summarized in Supplementary Table 1. Illumina reads were pre-processed using a standard pipeline (Supplementary Methods 1).

2.4 De novo transcriptome assembly

We built de novo transcriptome assemblies with TRINITY v. 2.1.1 (Grabherr et al., 2011b) using default settings. For the Mus data sets, to evaluate the robustness of some key findings to the
algorithm employed, we also built assemblies using two alternate assemblers: SHANNON v. 0.0.2 (Kannan, Hui, Mazooji, Pachter, & Tse, 2016) and BINPACKER v. 1.0 (Liu et al., 2016). Where strand-specific reads were available, we ran TRINITY and BINPACKER with the appropriate strand-specific options. Because the strand-specific mode of SHANNON is in experimental phase (Kannan, personal communication), we treated reads as un-stranded in all SHANNON assemblies. We assessed assembly quality using standard metrics for coverage of conserved orthologs, read mapping statistics, and assembly contiguity (Supplementary Methods 2). Downstream analyses compare outputs of de novo transcriptome assemblies to benchmarks consisting of reference annotations and genomes for each species (Supplementary Methods 3).

2.5 Assembly and read functional composition

For each contig within a de novo assembly, we quantified the proportion of unique assembly nucleotides that were CDS, UTR, intronic, start or stop codon, noncoding elements (including pseudogenic transcripts of genes with “polymorphic_pseudogene” gene biotype), introns of non-coding transcripts, and intergenic/unnanotated sequence. We took a similar approach with sequence reads to determine if the functional composition of RNA-seq reads differed from that of the transcriptome assemblies. For details, see Supplementary Methods 4.

2.7 Expression estimation

For nine combinations of assembly method (TRINITY, SHANNON, BINPACKER) and Mus sample (MDC, BALB/c, Massif Central), we generated expression estimates with RSEM v. 1.2.22 (B. Li & Dewey, 2011b) and BOWTIE2, v. 2.2.4 (Langmead & Salzberg, 2012). We also estimated “true” (reference-based) expression by mapping reads directly onto transcripts extracted from the
reference genome, using the same RSEM/BOWTIE2 pipeline but with the ENSEMBL GTF annotation.

To be able to evaluate the robustness of expression patterns derived from de novo transcriptome assemblies, one must connect assembly contigs to the underlying reference isoforms from which they originated. Thus, after extracting reference transcripts from the genome annotation with RSEM v. 1.2.22 (B. Li & Dewey, 2011a), we mapped transcriptome assembly contigs to these transcripts using BLAT v. 3.1.6 (Kent, 2002). We required at least 95% identity in the alignment to consider a hit valid, but did not set a minimum alignment length in the query (contig). Given the minimal evolutionary distance between query and target sequences, and because we required a minimum of 95% identity for alignments, we did not penalize coverage for mismatches. For each de novo assembly contig with > 1 BLAT hit, we identified the true transcript of origin as that for which the alignment produced the maximal coverage. In this way, each contig was uniquely assigned to a single reference isoform.

We assessed concordance between de novo assembly and referenced-derived approaches in multiple ways. First, we compared estimates for reference isoforms and their best hit BLAT contig in units of counts per million (CPM), transcripts per million (TPM), and a rescaled CPM that adjusts for differences in effective and observed length estimates between methods (see below, hereafter rescaled CPM). Second, we evaluated concordance between reference gene expression and our best estimate of gene-level expression derived from the transcriptome assembly, by treating all contigs with a best BLAT hit to transcripts of a particular ENSEMBL gene as transcripts of that gene. Contigs without BLAT hits were aggregated by TRINITY components, i.e. “genes”. For SHANNON assemblies, the collection of contigs assembled from an individual graph do not necessarily belong to the same gene (Kannan, personal communication). De novo
assembly gene-level expression was then calculated by building an RSEM index with the –

transcript-to-gene-map flag, such that contig expression values were aggregated to gene labels
specified in a table of gene-contig relationships. Genes identified as expressed in the map-to-
reference approach but that didn’t have any associated de novo transcriptome assembly contigs
were considered missing.

We also compared gene-level expression analyses between map-to-reference and map-to-
de novo using expression values estimated with KALLISTO v. 0.43.1 (Bray, Pimentel, Melsted, &
Pachter, 2016). Because KALLISTO operates on the transcript/contig level, we generated gene-
level estimates by aggregating to the gene level using the R package TXIMPORT (Soneson, Love,
& Robinson, 2015).

We also evaluated how a typical filtering approach influenced the robustness of
expression estimates at the gene level. We employed three filtering steps. First, we discarded all
lowly expressed transcriptome assembly contigs with TPM ≤ 1. Second, we used TRANSDECODER
v. 3.0.0 (https://github.com/TransDecoder/TransDecoder/wiki) to predict contig peptide
sequences, performed BLASTP hit against Uniref90 (downloaded 2017-04-05), and only retained
contigs with an alignment e-value ≤ 1 × 10⁻⁵. Finally, we restricted comparisons between de novo
transcriptome assemblies and our map-to-reference benchmark to protein coding genes for which
at least one isoform was ≥ 200bp; because the default minimum contig size for TRINITY is 200bp,
the minimum length requirement prevents artefactual downward bias in expression estimates.
Because expression estimates derived from TRINITY assemblies were in nearly all cases more
consistent with our benchmark data than those derived from SHANNON or BINPACKER, we
restricted this analysis to TRINITY assemblies.
The fragmentation of transcript and gene models in assemblies will affect effective length, where the effective length of a transcript equals the length of the transcript minus the mean sequence read fragment length, plus one. In other words, effective length estimates the total number of possible start sites for read alignments to a sequence given the fragment length distribution of reads. For example, if a 1kb transcript is broken into two 500bp contigs in a transcriptome assembly, and the mean read fragment length is 200bp, the effective length of the original transcript equals 801, whereas the summed effective lengths of the two transcriptome assembly contigs equals 602. The observed length also contributes to abundance estimates, e.g. if half of an expressed transcript is not assembled, no counts can be tallied for the missing transcript interval. Because both effective and observed length directly impacts abundance measurements, we assessed whether correcting for differences in both length metrics between reference transcripts and de novo assembled contigs would improve concordance in expression estimates. We did this by rescaling CPM as follows:

\[ CPM_{rescaled} = CPM \times \frac{\text{map-to-ref effective gene length}}{\text{denovo effective gene length}} \times \frac{\text{denovo gene length}}{\text{map-to-ref gene length}}. \]

The first scalar adjusts for the difference in estimated effective length, while the second scalar corrects for the difference in observed length. We only evaluated reference features to which at least one transcriptome assembly contig could be aligned.

If biases in expression estimates derived from de novo transcriptome assemblies are correlated across samples, this would mitigate downstream effects on differential expression testing. We examined the distribution of differences in such biases among samples, measured as effect sizes. Bias can be quantified as:

\[ \log_2 \left( \frac{\text{sample 1 CPM}_{\text{denovo}}}{\text{sample 1 CPM}_{\text{map-to-ref}}} \right) - \log_2 \left( \frac{\text{sample 2 CPM}_{\text{denovo}}}{\text{sample 2 CPM}_{\text{map-to-ref}}} \right), \]

which can be re-arranged as:
which is the bias observed when comparing log-fold change derived from different methods. We can estimate an effect size as:

\[
2^{\text{EXP}} \left[ \text{abs} \left( \log_2 \left( \frac{\text{sample 1 CPM}_{\text{denovo}}}{\text{sample 2 CPM}_{\text{denovo}}} \right) - \log_2 \left( \frac{\text{sample 1 CPM}_{\text{map-to-ref}}}{\text{sample 2 CPM}_{\text{map-to-ref}}} \right) \right) \right]
\]

For example, given \textit{de novo} CPM values for samples 1 and 2 of 15 and 20, and for map-to-reference, 5 and 10, respectively, we would estimate that the magnitude of difference between samples would be 1.5 times as great as if one had used the map-to-reference approach.

Two standard ways of filtering a \textit{de novo} transcriptome assembly are to discard contigs that are lowly expressed or that cannot be annotated. For the latter, the usual practice is to find open reading frames, and use BLASTX or BLASTP against a peptide database. We investigated how such filtering impacts the robustness of gene-level expression estimates by recalculating expression estimates after removing contigs with TPM < 1, contigs for which no ORF was detected, and contigs for which a translated ORF did not return a BLASTP hit to UNIREF90. We retained the original transcript and gene assignments for assembly contigs derived from BLAT, such that protein hits are simply treated as filter. This approach is optimistic given that it circumvents possible annotation errors, and also due to the fact that \textit{Mus} proteins are prevalent in UNIPROT. Because filtering will shorten the total length of sequence that is used to infer expression estimates, this may lead to noisier expression estimates. Thus, we quantified the reduction in coverage of reference annotations induced by our filtering scheme and related it to expression estimates. For details on coverage calculations and filtering, see Supplementary Methods 5.
2.9 Evaluation of variant calls from transcriptome assemblies

For these analyses, we focused on the Mus samples, and compared the map-to-reference genotypes to those derived from de novo TRINITY assemblies. We called variants off of the reference genome using the GATK v. 4 (McKenna et al., 2010), following the Broad Institute’s best practices for RNA-seq data (https://software.broadinstitute.org/gatk/documentation/article.php?id=3891; last updated 2017-01-21). For de novo TRINITY assemblies, we first constructed SuperTranscripts, (Davidson, Hawkins, & Oshlack, 2017), which attempt to collapse overlapping contigs within TRINITY “genes”, and then called variants with the GATK pipeline, based upon python scripts provided with TRINITY v. 2.5.1. SuperTranscript alignments to the genome were used to define unique polymorphic site ids that included cases where (a) both methods called genotypes, (b) only map-to-reference called a genotype (including sites with and without valid SuperTranscript alignments), or (c) a genotype was only called off a SuperTranscript. We recorded cases where multiple SuperTranscripts were aligned to the same genomic position, taking the union of all reported alleles as the ST genotype for that position. We restricted our comparison of the two methods by filtering out REF sites that were indels or contained > 2 alleles. Using genotype calls based upon the reference annotation as the gold standard, we calculated 11 different metrics, defined as follows:

- **Error rate**: # positions with a genotype call in either map-to-reference or map-to-SuperTranscripts, where SuperTranscript alleles don’t match map-to-reference alleles ÷ total # positions with a genotype call in either map-to-reference or map-to-SuperTranscripts; includes positions with no overlapping ST.
• **Error-Snv2Indel**: # bi-alleleic map-to-reference SNVs called as an indel in map-to-SuperTranscripts ÷ # bi-alleleic map-to-reference SNVs.

• **FN**: # map-to-reference polymorphic sites with an invariant or missing genotype in map-to-SuperTranscripts ÷ # map-to-reference polymorphic sites; includes sites with no overlapping SuperTranscripts.

• **FN-FractionNoAlignment**: # FN with no aligned SuperTranscript ÷ # FN

• **FN-withAlignment**: Same as FN, but restricted to sites with ≥1 overlapping SuperTranscript alignments

• **FP**: # map-to-reference invariant sites with a polymorphic map-to-SuperTranscript genotype ÷ # map-to-reference invariant sites.

• **FP-indel**: # map-to-SuperTranscript indel polymorphic sites at map-to-reference invariant sites ÷ # false positive sites.

• **FP-hetSNV**: # false positive sites with map-to-reference biallelic SNV genotypes ÷ # false positive sites.

• **Recall**: # polymorphic map-to-reference genotypes with matching map-to-SuperTranscripts genotypes ÷ # polymorphic map-to-reference genotypes

• **Recall-withAlignment**: same as Recall restricted to sites with ≥1 overlapping SuperTranscript alignments

• **Precision**: # polymorphic map-to-SuperTranscript genotypes that match map-to-reference genotypes ÷ # polymorphic map-to-SuperTranscript genotypes.

See Supplementary Methods 6 for more details.

2.10 Analysis reproducibility
Software command lines, scripts and accessory files used to implement analyses described above are available at https://github.com/harvardinformatics/TranscriptomeAssemblyEvaluation.

3. RESULTS

3.1 “High quality” assemblies are highly fragmented

Brain tissue is both transcriptionally complex and expected to express a large proportion of an organism’s overall transcriptional profile. It thus represents a challenging test case for de novo transcriptome assembly. For this reason, the bulk of our analyses focus on data generated for experiments involving mouse (Mus spp.) brain.

Our assemblies were highly fragmented, consisting of anywhere from 70,375 – 582,133 contigs, with median contig sizes ranging from just a few hundred bases to slightly over 1kb, and \( N_{50} \) ranging from 721 to 3021 bp (Supplementary Table 2). This is typical of many published transcriptome assemblies, e.g. Bentley et al., 2017; Faherty et al., 2018; Nespolo et al., 2018. Yet, typically employed measures of accuracy and completeness indicate that all assemblies were of reasonably high quality. 86.4 – 98.9% of read pairs aligned to their respective assemblies, and upwards of 95% of these alignments occurred in a concordant fashion, with both reads aligning to the same contig (Supplementary Table 3). For the Mus assemblies, 77.1 – 90.4% of BUSCOs were recovered in either complete or fragmented form; BUSCO recovery was considerably lower for the *Xenopus, Drosophila*, and *Tetraodon* assemblies (Supplementary Table 4). The large number of short contigs strongly suggest that de novo transcriptome assemblies do not typically assemble full length transcripts.
3.2 Assembly functional composition is biased

Across all sample-assembler combinations, large fractions of de novo transcriptome contigs are dominated by intronic, UTR, and intergenic sequence (Supplementary Figures 1-2). The excess of intronic sequence in contigs is not reflected in the composition of reads (Supplementary Figures 1-2). The prevalence of intronic and intergenic sequence in assembly contigs (Supplementary figure 3) indicates that a substantial proportion of assembled bases are derived from some combination of transcriptional read-through across exon junctions, assembly of low-abundance pre-mRNA, and low levels of spurious genome-wide transcription. Even assuming intronic and intergenic sequences derive from few RNA-seq reads, contigs mostly deriving from UTR will be challenging to annotate. Excluding un-annotated UTR contigs will decrease post-filtering read counts, potentially reducing power for downstream expression analyses.

3.3 Frequent genotyping errors

SNP calls and allele frequencies inferred from them and that are derived from transcriptome assemblies have been used in a variety of contexts, including estimation genetic diversity in evolutionary analyses (Romiguier et al., 2014), population structure (Tepolt & Palumbi, 2015), reconstruction of demographic history and gene flow (De Wit & Palumbi, 2013; Le Gac et al., 2016; Roda, Mendes, Hahn, & Hopkins, 2017), estimation of dN/dS (Cokus, Gugger, & Sork, 2015), FST outlier analysis (Bankers et al., 2017), and tests for allele-specific expression (Singh et al., 2018). In principle, redundancy through assembly of haplotypes as distinct contigs should lead to erroneous genotype calls and biased estimates of allele frequency spectra. For example, two contigs can actually be haplotypes of a single locus, such that a truly heterozygous position is instead represented as two separate contigs, one homozygous for the reference allele and one
homozygous for the alternative allele. Two lines of evidence indicate such redundancy is present in transcriptome assemblies. First, BLAT all-by-all alignments indicate that many contigs are nearly identical (Supplementary Figures 4-6), in some cases differing only by single nucleotide variants (e.g. Supplementary Figure 4b). Surprisingly, TRINITY assemblies will produce such highly similar contigs from different assembly graphs which are purported to assembly isoforms of a gene or sets of closely related paralogs (Supplementary Figure 6). Second, there is a high frequency of cases where all nucleotides of a transcript from a single-isoform Mus gene are covered by ≥ 1 de novo assembly contig (Supplementary Figure 7). The latter are primarily due to slight variations in the assembly of intronic bases, and secondarily due to differences in CDS content (Supplementary Figure 8).

The opposite problem can happen when paralogous loci are collapsed, leading to a heterozygous genotype call where there would otherwise be two invariant sites, or perhaps a mixture of alleles due to, for example, collapsing of an invariant and a heterozygous position. To assess the degree to which these and other errors affect transcriptome assemblies, we compared genotype calls and derived diversity metrics between the map-to-reference and map-to-transcriptome approaches using 11 different metrics.

The frequency of errors, in which the allele set recovered by the two methods differed or in which a genomic position with a called genotype contained no overlapping SuperTranscript nucleotides, was high, ranging from 30-37% for the wild samples, and 49-83% for the inbred samples (Table 1). Error rates varied among SuperTranscripts within a de novo transcriptome assembly, with some SuperTranscripts demonstrating extremely high error rates. While the median error rate per SuperTranscript was 0.14% per base or less, maximum per SuperTranscript rates typically exceeded 10% per base (Supplementary Table 5). For five of six samples, the
median number of errors per SuperTranscript was 2, but, depending upon the sample, as many as 56-187 errors could occur within a single SuperTranscript (Supplementary Table 5). Thus, within any de novo transcriptome assembly, a small number of contigs may contribute disproportionately to noise that may obscure the true patterns of polymorphism. On a per-sample basis, errors show a modest yet statistically significant bias towards the distal end of a SuperTranscript for all samples except MDC (sample-level Wilcoxon 1-sided tests, $H_0 = 0.5, p < 2 \times 10^{-16}$; Supplementary Figure 9).

The false negative rate (FN; 23-71%) was greater than the false positive rate (FP; 15-87%) in four of six samples. The latter includes all classes of errors including indel polymorphism. Anywhere from 42-53% of FN were due to there being no SuperTranscript aligned at a genomic position where a map-to-reference heterozygote genotype call was made (Table 1 FN-FractionNoAlignment). Given that rates of mapping SuperTranscripts to the genome were extremely high (>99%), these FN instances are primarily due to a failure to assemble transcripts overlapping polymorphic genomic positions. Even at sites called as heterozygotes in map-to-reference analyses that were assembled into SuperTranscripts, false negative rates were still relatively high, from 13-15% and 37-59% in the wild and inbred samples, respectively (Table 1).

Because read depth influences the ability of tools such as GATK to call genotypes, we suspected false negatives might be due to reduced read depth at homologous genomic positions called as heterozygotes with the map-to-reference approach. While map-to-reference and map-to-SuperTranscript coverage were largely colinear for concordant and false positive map-to-SuperTranscript genotypes (Supplementary Figures 10-11), false negative sites with overlapping SuperTranscripts displayed small reductions in SuperTranscript read depth relative to map-to-
reference at sites where SuperTranscripts mapped to the genome (Supplementary Figure 1).

Given the magnitude of coverage reduction is small, we suspect it may not be the only factor contributing to false negatives at sites where a SuperTranscript covers a polymorphic genomic position.

The false positive (FP) rate ranged from 15-87%. Incorrect heterozygous and indel genotype calls at truly invariant sites are expected when paralogs are collapsed, and both occur with considerable frequency in our samples. From 6-22% of FPs are erroneous indel genotypes, 8-62% of FPs are erroneous heterozygous genotype calls, and the remainder are other classes of SuperTranscript polymorphisms, including both indel and single nucleotide variants, with > 2 alleles. The inference of frequent paralog collapses is also supported by the high rate at which true SNV polymorphisms are called as indels in SuperTranscripts (Error-Snv2Indel, 16-31%).

When genotypes were called in SuperTranscripts, the frequency of calls that were correct (i.e. precision) ranged from 66-83% for the wild samples, but precision for the inbred samples, particularly MDC, was substantially lower. Recall ranged from 76-81% in the wild samples; as with precision, recall for both inbred samples was substantially lower. The frequency of genotyping errors seems likely to distort evolutionary inferences and reduce power across a wide range of statistical methods, particularly those aimed at detecting signals of selection on particular polymorphisms, allele-phenotype and allele frequency-environment associations.

3.4 Negatively biased heterozygosity estimates

In all but the inbred strain samples (MDC and BALB/c), we observed reduced heterozygosity in the exonic regions of SuperTranscripts relative to mapping to those regions in the reference genome (Figure 1). A substantial contributor to this bias is the fact that, as noted above,
approximately half of map-to-reference heterozygous positions are not overlapped by any SuperTranscript. In addition, this bias does not result from SuperTranscripts being preferentially assembled in regions of lower heterozygosity, given that, for the wild samples, map-to-reference heterozygosity is actually higher in SuperTranscript-overlapping regions (Supplementary Figure 13). Furthermore, the bias cannot be due to some unique feature of genotype calls in exons, as the same pattern holds when considering all genotype calls (Supplementary Figure 14). Another possible explanation may be lower power to detect low frequency alternative alleles in map-to-SuperTranscript analyses. However, in our five samples consisting of pooled individuals, minor allele frequency at map-to-reference heterozygous sites did not vary between correct SuperTranscript heterozygous calls, false negatives, and other genotyping errors (Supplementary Figure 15). Downward bias in heterozygosity estimates occurs despite a relatively high prevalence of false positive heterozygotes. Because this bias shifts all estimates towards zero, it truncates the observed range of variation, making it more difficult to make meaningful inferences from comparisons of diversity, whether those comparisons take place among assemblies derived from different populations or species, or at the level of individual genes within or between assemblies.

3.5 Single contigs are poor expression estimators

Expression estimates derived from the contigs that are the best BLAT hit relative to the reference transcripts are only weakly correlated (spearman’s $p$: CPM: 0.070-0.387; TPM: 0.045-0.392; rescaled CPM: 0.083-0.415), with expression for those transcripts, and show positive bias for lowly expressed reference transcripts (Supplementary Figure 16ab, Supplementary Table 6). Correlations between best-hit contig expression and expression of the associated reference gene,
after accounting for effective length differences, are slightly better but still relatively weak 
(spearman’s $\rho$: rescaled CPM: 0.103-0.657 Supplementary Table 6). Overall, the use of single 
contigs as proxies for gene-level expression appears problematic. 

3.6 Correctable gene-level expression bias

Expression estimates derived from *de novo* transcriptome assemblies are far more robust at the 
gene level, when one is able to group contigs into genes via annotation (Figure 2, Table 2). It is 
reassuring that the correlation between the methods is quite strong, although there is sizeable 
positive bias at lower expression levels for TPM, which is substantially reduced for CPM. 
Furthermore, for many genes there are substantial deviations from estimates obtained with 
mapping to the reference, even at high expression levels (Figure 2, Supplementary Figure 17). 
For estimates generated with RSEM, bias and deviations from the map-to-reference estimates are 
reduced when one rescales CPMs to account for differences in effective length estimates (Figure 
2c). This indicates that effective length differences—that is, the difference in the number of start 
sites for read alignments, given a particular sequenced fragment size, between a full-length 
transcript, and a set of transcript sub-sequences represented by contigs—are part of the problem, 
and in particular, that effective length is not properly estimated on the fragmented gene models 
that are typical in *de novo* transcriptome assemblies. Biases in gene expression are reduced as the 
difference in effective length between map-to-reference and map-to-transcriptome-assembly 
decreases (Figure 3). Even so, disparities in effective length estimates between map-to-reference 
and map-to-*de novo*-transcriptome-assembly approaches can persist even at high reference gene 
expression, with many *de novo* transcriptome-based estimates of gene expression showing a
Correlations between reference gene expression and gene-level summing across assembly contigs obtained with KALLISTO were very similar to those obtained with RSEM (Table 3, Supplementary Figure 17). Adjusting for effective length had little effect on KALLISTO CPM, and in some cases reduced the correlation between methods. KALLISTO may be doing a better job at estimating effective length on transcriptome assemblies, such that it’s inferior performance relative to RSEM must be due to other reasons.

3.7 Expression bias and DE testing

Biases in expression estimates derived from transcriptome assemblies might not have large effects on downstream differential expression analyses if biases are consistent across samples. While the pooled Mus data sets we used did not permit us to look at inter-individual variation amongst individual libraries used to build a single assembly, we looked across the wild Mus datasets to see if biases in gene-level expression were concordant, i.e. were biases for a particular gene always of similar magnitude and in the same direction. We focus on results for TRINITY assemblies of wild Mus samples, as TRINITY consistently showed stronger correlations with map-to-reference based estimates, such that our findings are conservative with respect to assemblers showing weaker correlations. Reassuringly, log-fold change of transcriptome assembly vs. map-to-reference expression estimators (LFCMETHOD) showed broad patterns of moderate concordance at the gene level indicating that gene-level features influence the magnitude of bias (Supplementary Figure 19). Inter-sample correlations were, for CPM, 0.625-0.727; TPM, 0.790-0.830; and rescaled CPM, 0.605-0.671. These correlations are weaker for CPM and rescaled
CPM, in part, because biases are reduced overall, compressing LFC_{METHOD} toward zero. Despite these broad correlations, a number of genes may show excessive bias. We measured this as a bias effect size that indicates how much greater a difference in expression would be detected using a \textit{de novo} transcriptome approach, relative to mapping to reference annotation. Comparing the FRA and DEU samples, for CPM, TPM, and rescaled CPM, 5.24\%, 5.16\% and 3.35\% of genes showed bias effect sizes $\geq 5$, with more than 2\% of genes showing an order of magnitude of difference between methods of $\geq 10$ for CPM and TPM (Figure 4).

Overall, our results suggest three practical considerations for statistical tests on gene expression estimates derived from \textit{de novo} transcriptome assemblies. First, RSEM is preferable over KALLISTO for estimating expression from \textit{de novo} transcriptome assemblies. Second, although molecular ecologists do not typically have a gold-standard dataset derived from a well-annotated genome, the effective length adjustments we carry out can be employed across a set of transcriptome assemblies, simply by setting one assembly as the standard against which to conduct rescaling. This opens up the possibility of robust comparative analyses of gene expression across multiple species for which separate \textit{de novo} transcriptome assemblies must be built. Finally, while correcting for effective length differences substantially reduced the frequency of bias outliers, it does not eliminate them entirely. Further investigation of other sources of bias is necessary in order to devise methods to mitigate their impacts on expression analyses.

3.8 Effects of filtering on expression estimates

Filtering out lowly expressed \textit{de novo} transcriptome assembly contigs as well as those contigs without a BLASTP hit to Uniref90 led to a noticeable improvement in the correlation between
map-to-transcriptome assembly and map-to-reference abundance estimates for protein coding genes (Figure 2d-f). Post-filtering, bias and noise in TPM was dramatically reduced. However, despite a nearly a correlation nearly identical to that for CPM and rescaled CPM, there is consistent positive bias across the range of map-to-reference expression (Figure 2e).

Nevertheless, the improvement across all three metrics relative to unfiltered assemblies belies a tradeoff that is hidden from view from researchers without access to an annotated genome: filtering leads to a loss of many expressed protein coding genes that would be recovered without filtering. For the FRA assembly, 21.0% of such genes are filtered out, although only 10.6% of genes are lost if one excludes from consideration lowly expressed genes with map-to-reference TPM < 1 (Figure 5). Depending upon the sample-assembler combination, 9.1-32.5% of expressed protein coding genes can be lost due to filtering, and for genes with map-to-reference TPM ≥ 1, 6.0-27.1% (Table 2). In nearly all sample-assembler combinations, the proportion of genes lost due to filtering far exceeded the proportion that were missing due to a failure to assemble those genes, with most exceptions occurring for BINPACKER assemblies, which performed poorly with respect to recovery of expressed genes (Table 2).

**DISCUSSION**

*De novo* transcriptome assemblies have been an important entry point for functional and population genomics in non-model organisms, but the two key assumptions underlying their application are consistently violated, albeit in unexpected ways. Below, we explain how particular compositional aspects of *de novo* transcriptome assemblies may impact the power of evolutionary inference.
By far the most unexpected findings from our study were the extremely high genotyping error rates derived from transcriptome assemblies, specifically Super Transcript representations of TRNITY assemblies. False negatives were due, in part, to the absence of a Super Transcript aligned to a putatively true polymorphic site recovered with a map-to-reference approach. False negatives are due to either the failure to assemble any contigs at all for a gene (i.e. missing genes), or the incomplete coverage of genes not extending over some fraction of polymorphic sites. That genotyping errors were biased towards the distal end of contigs suggest that assembly graphs are getting broken at hard to assemble features containing low complexity and/or increased polymorphism. When such graph-splitting occurs, the intervening polymorphic sites will, by definition, not be covered by a Super Transcript, leading to downwardly biased estimates of heterozygosity in de novo assemblies. False negatives may have also been caused by the slightly reduced read depth at homologous positions in read alignments to Super Transcripts. That the reduction in read depth in the latter, was, on average, small suggests that other factors may play a role. It may be that genotype inference methods such at the GATK HAPLOTYPE CALLER that leverage haplotypic information, have poorer performance when iterating over alignments to relatively short assembly contigs. Another possibility is that assembly errors may generate patterns that get flagged by standard genotyping filters. In contrast, false positives appear to be driven by paralog collapses given the bias towards heterozygous false positive sites.

While we document overall high error rates, the range of precision we observed was broadly consistent with what was observed in the original paper that describes the LACE algorithm (~ 80%) for converting transcriptome assemblies into Super Transcripts (Davidson et al., 2017). It is important to note that, in that paper, recall and false negative rates are not reported, and genotyping performance is restricted to Super Transcript genotype calls relative to
known variants in Genome in a Bottle Consortium high confidence regions for humans (Zook et al., 2016). Such well-curated, filtered benchmark SNV sets are, by definition, not available for non-model organisms with neither genomes nor extensive resequencing data. Results reported on them are optimistic relative to what can be expected for organisms without such genomic resources. For this reason and to create more realistic expectations for molecular ecologists, we use genotypes obtained from mapping to reference genome as our truth set.

The negative bias we observed in heterozygosity estimates derived from de novo transcriptome assemblies indicates that the underlying site-frequency spectrum derived from an SuperTranscript assembly will be biased in ways that distort other population genetic summary statistics typically reported in molecular ecological investigations that utilize genotypes called from transcriptome assemblies. More broadly, the error rates we document suggest that several lines of transcriptome-assembly based research should be approached with caution. First, erroneous heterozygous calls, when added to the redundancy problems described above, will substantially reduce power and elevate false positive rates in allele-specific expression analyses. Second, levels of genome-wide patterns of genetic diversity derived from transcriptome assemblies will likely be negatively biased, at least for diploid species. The distinction concerning ploidy is an important one because, in the absence of a genome or karyotypic information, there can be no a priori expectation of how events such as whole-genome or even chromosome-level duplication will impact genotyping inference. Thus, both integration of diversity metrics across transcriptome assemblies with differing, unknown levels of ploidy, and comparing transcriptome and genome assembly-derived metrics will confound method with biology and can lead to spurious inferences. Even when applied solely to diploid transcriptome assemblies, as haplotypes will get assembled into different SuperTranscripts, the expectation is
that negative bias will be greater for more heterozygous genomes, such that estimated levels of diversity across assemblies will artificially appear more similar than they are in actuality, impeding researchers’ ability to detect underlying patterns of variation. Overall, the distortion of allele frequencies derived from transcriptome assemblies strongly suggests that they will inject bias into analyses of population structure, demographic inference, and efforts to detect positive selection. Whether or not improved clustering of contigs can mitigate this distortion remains an open question.

Our results clearly indicate that isoform-level analyses and using individual contigs as proxies for isoform or gene-level expression, should be avoided. Even when annotation-based groupings of contigs are performed in an optimal setting, such as our analyses of *Mus* transcriptomes, there is substantial noise in gene-level expression estimates derived from transcriptome assemblies relative to a map-to-reference based approach, and there is positive bias at lower expression levels. Incorrect estimation of effective length metrics used in RNA abundance calculations play a substantial role in the observed patterns of bias. Nevertheless, it may be surprising to researchers that biases persist for standard length-adjusted metrics such as TPM. CPM has less bias compared to TPM. Bias and noise can be reduced by rescaling CPM to account for differences in effective and observed length, such that the rescaled CPM more faithfully recapitulates map-to-reference based expression estimates, even if substantial noise persists. Adjusting expression estimates to account for effective and observed length differences can be beneficially used in two different applications. First, one could conduct differential expression estimates across different transcriptome assemblies when it is necessary to generate different assemblies for different conditions, such as divergent species or sub-species. Because one of the first steps differential expression tools execute is to normalize by library size, one can
simply apply the correction directly on the counts themselves. CPM and count scale are directly proportional, with the former effectively being the latter multiplied by a sample-specific constant, such that all benefits we report on weighted-adjusted CPM will apply to similarly transformed counts. Adjusting counts across a set of species would require treating one species as the “reference”, and rescaling counts for other species relative to it. Outside of the realm of formal tests of differential expression, we suggest that rescaled CPM could be useful in phylogenetic comparative analyses of expression at the gene level. One of the big challenges of these methods is accounting for length differences between transcript/contig sets belonging to orthologous genes, and the adjusted CPM appears to substantially outperform other metrics in this regard. We should note that our adjustment method is relatively simple, and that it may be possible to adopt more sophisticated approaches that leverage such information as sequence composition and complexity for additional improvements.

Standard filtering of de novo transcriptome assemblies, including removing lowly expressed and un-annotated contigs, can measurably improve the concordance between transcriptome assembly and map-to-reference-based expression estimates. Unfortunately, filtering leads to the loss of many expressed protein coding genes that are detected in a de novo transcriptome assembly prior to filtering. Researchers relying on de novo transcriptome assemblies for expression analyses thus face a difficult tradeoff—get less robust expression estimates for more genes, and risk a larger fraction of erroneous differential expression results, or get more robust estimates for far fewer genes, and risk missing those that are functionally relevant to the phenotypes being investigated.

We do not test here datasets amenable to differential expression analyses, and a counter-argument to our findings of bias and noise is that neither should have a substantial effect, so long
as bias and deviations from map-to-reference expectations are consistent across replicates and conditions. While we did find broad congruence in bias across samples, gene-level differences in bias among samples were frequent enough such that they will likely elevate false positive rates for differential expression tests derived from de novo transcriptome assemblies. This effect can be minimized by utilizing metrics that adjust for effective length differences.

**Conclusion**

Given the challenges associated with eco-evolutionary analyses that rely on de novo transcriptome assemblies and the rapidly falling costs of long read sequencing, an unavoidable recommendation for research teams launching new studies may be to generate a provisional genome assembly, e.g. using a combination of Oxford Nanopore or PacBio long reads and Illumina paired-end short reads (Austin et al., 2017; Marcionetti, Rossier, Bertrand, Litsios, & Salamin, 2018; Michael et al., 2018; Nowoshilow et al., 2018), at least in some cases. Such an assembly can be annotated with ab initio gene prediction tools and Illumina RNA-seq paired-end reads. Given a reasonable quality annotated genome, fewer reads per sample would have to be allocated for downstream expression analyses, and the cost savings from lighter RNA sequencing would partially offset costs of genomic sequencing and de novo assembly. Nevertheless, there are three reasons whole de novo transcriptome assembly might still be attractive. First, genome assembly remains challenging and expensive species with large genomes or with unusual genome structure precluding standard methods. Second, thousands of RNA-seq data sets now exist for species without reference genomes, and it is highly unlikely that genome assemblies will be carried out any time soon for anything but a small fraction of these species. Third, analyses based upon genome assemblies will depend strongly on the quality of
the assembly. Even for genomes of reasonably high quality, protein coding genes may be missing, either through problems with the annotation process or due to the fact that genes fall into assembly gaps (Peona, Weissensteiner, & Suh, 2018). In other words, genome assembly in not necessarily a panacea for all problems related to expression analyses. However analytically challenging, there is great potential for employing de novo transcriptome assemblies in interesting comparative analyses, so long as the conducted in an informed fashion—avoiding analyses modes that suffer from fundamental biases, mitigating biases where possible for others, and, always, placing appropriate caveats given the biases that will inevitably remain.

ACKNOWLEDGMENTS

The computations in this paper were run on the Odyssey cluster supported by the FAS Division of Science, Research Computing Group at Harvard University. We thank members of the FAS Informatics Group for discussion. We also thank Nadia Davidson, Alicia Oshlack, Robin Hopkins, and Mara Laslo for feedback on an earlier version of this manuscript. This work was conducted on the traditional territory of the Wampanoag and Massachusett peoples.

DATA AVAILABILITY

Upon acceptance for publication, all transcriptome assemblies, SuperTranscript assemblies, and vcf genotype files will be made available on Dryad.

AUTHOR CONTRIBUTIONS
The research project was designed by A. H. F., M. C., and T. B. S. Analyses were conducted by A. H. F. with input from M.C. and T. B. S. A. H. F. wrote the paper with T. B. S. All authors read and approved the final manuscript.

REFERENCES

Austin, C. M., Tan, M. H., Harrisson, K. A., Lee, Y. P., Croft, L. J., Sunnucks, P., ... Gan, H. M. (2017). De novo genome assembly and annotation of Australia’s largest freshwater fish, the Murray cod (Maccullochella peelii), from Illumina and Nanopore sequencing read. *GigaScience*, 6(8), 1–6. doi: 10.1093/gigascience/gix063

Bankers, L., Fields, P., McElroy, K. E., Boore, J. L., Logsdon, J. M., & Neiman, M. (2017). Genomic evidence for population-specific responses to co-evolving parasites in a New Zealand freshwater snail. *Molecular Ecology*, 26(14), 3663–3675. doi: 10.1111/mec.14146

Bentley, B. P., Haas, B. J., Tedeschi, J. N., & Berry, O. (2017). Loggerhead sea turtle embryos (Caretta caretta) regulate expression of stress response and developmental genes when exposed to a biologically realistic heat stress. *Molecular Ecology*, 26(11), 2978–2992. doi: 10.1111/mec.14087

Berens, A. J., Hunt, J. H., & Toth, A. L. (2015). Comparative transcriptomics of convergent evolution: different genes but conserved pathways underlie caste phenotypes across lineages of eusocial insects. *Molecular Biology and Evolution*, 32(3), 690–703. doi: 10.1093/molbev/msu330
Birol, I., Jackman, S. D., Nielsen, C. B., Qian, J. Q., Varhol, R., Stazyk, G., … Jones, S. J. M. (2009). De novo transcriptome assembly with ABySS. *Bioinformatics*, 25(21), 2872–2877. doi: 10.1093/bioinformatics/btp367

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, 34(5), 525–527. doi: 10.1038/nbt.3519

Cokus, S. J., Gugger, P. F., & Sork, V. L. (2015). Evolutionary insights from de novo transcriptome assembly and SNP discovery in California white oaks. *BMC Genomics*, 16, 552. doi: 10.1186/s12864-015-1761-4

Cokus, S. J., Gugger, P. F., & Sork, V. L. (2015). Evolutionary insights from de novo transcriptome assembly and SNP discovery in California white oaks. *BMC Genomics*, 16, 552. doi: 10.1186/s12864-015-1761-4

Davidson, N. M., Hawkins, A. D. K., & Oshlack, A. (2017). SuperTranscripts: a data driven reference for analysis and visualisation of transcriptomes. *Genome Biology*, 18(1). doi: 10.1186/s13059-017-1284-1

Davidson, N. M., & Oshlack, A. (2014). Corset: enabling differential gene expression analysis for de novoassembled transcriptomes. *Genome Biology*, 15(7), 410. doi: 10.1186/s13059-014-0410-6

De Wit, P., & Palumbi, S. R. (2013). Transcriptome-wide polymorphisms of red abalone (Haliotis rufescens) reveal patterns of gene flow and local adaptation. *Molecular Ecology*, 22(11), 2884–2897. doi: 10.1111/mec.12081

Faherty, S. L., Villanueva-Cañas, J. L., Blanco, M. B., Albà, M. M., & Yoder, A. D. (2018). Transcriptomics in the wild: Hibernation physiology in free-ranging dwarf lemurs. *Molecular Ecology*, 27(3), 709–722. doi: 10.1111/mec.14483
Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., ... Regev, A. (2011a). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology, 29*(7), 644–652. doi: 10.1038/nbt.1883

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., ... Regev, A. (2011b). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology, 29*(7), 644–652. doi: 10.1038/nbt.1883

Hornett, E. A., & Wheat, C. W. (2012). Quantitative RNA-Seq analysis in non-model species: assessing transcriptome assemblies as a scaffold and the utility of evolutionary divergent genomic reference species. *BMC Genomics, 13*, 361. doi: 10.1186/1471-2164-13-361

Kannan, S., Hui, J., Mazooji, K., Pachter, L., & Tse, D. (2016). Shannon: An Information-Optimal de Novo RNA-Seq Assembler. *BioRxiv*, 039230. doi: 10.1101/039230

Kent, W. J. (2002). BLAT—the BLAST-like alignment tool. *Genome Research, 12*(4), 656–664. doi: 10.1101/gr.229202. Article published online before March 2002

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods, 9*(4), 357–359. doi: 10.1038/nmeth.1923

Le Gac, M., Metegnier, G., Chomérat, N., Malestroit, P., Quéré, J., Bouchez, O., ... Chapelle, A. (2016). Evolutionary processes and cellular functions underlying divergence in *Alexandrium minutum. Molecular Ecology, 25*(20), 5129–5143. doi: 10.1111/mec.13815

Li, B., & Dewey, C. N. (2011a). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics, 12*, 323. doi: 10.1186/1471-2105-12-323
Li, B., & Dewey, C. N. (2011b). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics, 12*, 323. doi: 10.1186/1471-2105-12-323

Li, K., Wang, H., Cai, Z., Wang, L., Xu, Q., Lövy, M., … Nevo, E. (2016). Sympatric speciation of spiny mice, *Acomys*, unfolded transcriptomically at Evolution Canyon, Israel. *Proceedings of the National Academy of Sciences, 113*(29), 8254–8259. doi: 10.1073/pnas.1608743113

Liu, J., Li, G., Chang, Z., Yu, T., Liu, B., McMullen, R., … Huang, X. (2016). BinPacker: Packing-Based De Novo Transcriptome Assembly from RNA-seq Data. *PLOS Computational Biology, 12*(2), e1004772. doi: 10.1371/journal.pcbi.1004772

MacManes, M. D. (2018). The Oyster River Protocol: a multi-assembler and kmer approach for de novo transcriptome assembly. *PeerJ, 6*. doi: 10.7717/peerj.5428

Malik, L., Almodaresi, F., & Patro, R. (2018). Grouper: graph-based clustering and annotation for improved de novo transcriptome analysis. *Bioinformatics (Oxford, England), 34*(19), 3265–3272. doi: 10.1093/bioinformatics/bty378

Marcionetti, A., Rossier, V., Bertrand, J. A. M., Litsios, G., & Salamin, N. (2018). First draft genome of an iconic clownfish species (Amphiprion frenatus). *Molecular Ecology Resources, 18*(5), 1092–1101. doi: 10.1111/1755-0998.12772

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res, 20*(9), 1297–1303. doi: 10.1101/gr.107524.110
Mi, H., Muruganujan, A., Casagrande, J. T., & Thomas, P. D. (2013). Large-scale gene function analysis with the PANTHER classification system. *Nature Protocols, 8*(8), 1551–1566. doi: 10.1038/nprot.2013.092

Michael, T. P., Jupe, F., Bemm, F., Motley, S. T., Sandoval, J. P., Lanz, C., ... Ecker, J. R. (2018). High contiguity Arabidopsis thaliana genome assembly with a single nanopore flow cell. *Nature Communications, 9*(1), 541. doi: 10.1038/s41467-018-03016-2

Nespolo, R. F., Gaitan-Espitia, J. D., Quintero-Galvis, J. F., Fernandez, F. V., Silva, A. X., Molina, C., ... Bozinovic, F. (2018). A functional transcriptomic analysis in the relict marsupial Dromiciops gliroides reveals adaptive regulation of protective functions during hibernation. *Molecular Ecology, 27*(22), 4489–4500. doi: 10.1111/mec.14876

Nowoshilow, S., Schloissnig, S., Fei, J.-F., Dahl, A., Pang, A. W. C., Pippel, M., ... Myers, E. W. (2018). The axolotl genome and the evolution of key tissue formation regulators. *Nature, 554*(7690), 50–55. doi: 10.1038/nature25458

Ockendon, N. F., O’Connell, L. A., Bush, S. J., Monzón-Sandoval, J., Barnes, H., Székely, T., ... Urrutia, A. O. (2016). Optimization of next-generation sequencing transcriptome annotation for species lacking sequenced genomes. *Molecular Ecology Resources, 16*(2), 446–458. doi: 10.1111/1755-0998.12465

Pankey, M. S., Minin, V. N., Imholte, G. C., Suchard, M. A., & Oakley, T. H. (2014). Predictable transcriptome evolution in the convergent and complex bioluminescent organs of squid. *Proceedings of the National Academy of Sciences, 111*(44), E4736–E4742. doi: 10.1073/pnas.1416574111
Peona, V., Weissensteiner, M. H., & Suh, A. (2018). How complete are “complete” genome assemblies?—An avian perspective. *Molecular Ecology Resources, 18*(6), 1188–1195. doi: 10.1111/1755-0998.12933

Roda, F., Mendes, F. K., Hahn, M. W., & Hopkins, R. (2017). Genomic evidence of gene flow during reinforcement in Texas Phlox. *Molecular Ecology, 26*(8), 2317–2330. doi: 10.1111/mec.14041

Romiguier, J., Gayral, P., Ballenghien, M., Bernard, A., Cahais, V., Chenuil, A., ... Galtier, N. (2014). Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature, 515*(7526), 261–263. doi: 10.1038/nature13685

Schwentner, M., Combosch, D. J., Pakes Nelson, J., & Giribet, G. (2017). A Phylogenomic Solution to the Origin of Insects by Resolving Crustacean-Hexapod Relationships. *Current Biology, 27*(12), 1818-1824.e5. doi: 10.1016/j.cub.2017.05.040

Siepel, A. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research, 15*(8), 1034–1050. doi: 10.1101/gr.3715005

Singh, R., Jones, T., Wai, C. M., Jifon, J., Nagai, C., Ming, R., & Yu, Q. (2018). Transcriptomic analysis of transgressive segregants revealed the central role of photosynthetic capacity and efficiency in biomass accumulation in sugarcane. *Scientific Reports, 8*(1), 4415. doi: 10.1038/s41598-018-22798-5

Soneson, C., Love, M. I., & Robinson, M. D. (2015). Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research, 4*, 1521. doi: 10.12688/f1000research.7563.1
Tepolt, C. K., & Palumbi, S. R. (2015). Transcriptome sequencing reveals both neutral and adaptive genome dynamics in a marine invader. *Molecular Ecology, 24*(16), 4145–4158. doi: 10.1111/mec.13294

Todd, E. V., Black, M. A., & Gemmell, N. J. (2016). The power and promise of RNA-seq in ecology and evolution. *Molecular Ecology, 25*(6), 1224–1241. doi: 10.1111/mec.13526

Veilleux, H. D., Ryu, T., Donelson, J. M., van Herwerden, L., Seridi, L., Ghosheh, Y., ... Munday, P. L. (2015). Molecular processes of transgenerational acclimation to a warming ocean. *Nature Climate Change, 5*(12), 1074–1078. doi: 10.1038/nclimate2724

Velotta, J. P., Wegryn, J. L., Ginzburg, S., Kang, L., Czesny, S., O’Neill, R. J., ... Schultz, E. T. (2017). Transcriptomic imprints of adaptation to fresh water: parallel evolution of osmoregulatory gene expression in the Alewife. *Molecular Ecology, 26*(3), 831–848. doi: 10.1111/mec.13983

Vijay, N., Poelstra, J. W., Künstner, A., & Wolf, J. B. W. (2013). Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive *in silico* assessment of RNA-seq experiments. *Molecular Ecology, 22*(3), 620–634. doi: 10.1111/mec.12014

Zook, J. M., Catoe, D., McDaniel, J., Vang, L., Spies, N., Sidow, A., ... Salit, M. (2016). Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Scientific Data, 3*. doi: 10.1038/sdata.2016.25
Table 1. Selected performance metrics comparing genotypes derived from SuperTranscript collapse of TRINITY de novo transcriptome assemblies with those derived from a map-to-reference genome approach. See METHODS for information on metric calculations.

| Metric                          | Sample       |
|--------------------------------|--------------|
|                                | MDC  | BALB/c | FRA  | DEU  | IRN  | KZK  |
| Error rate                     | 0.829| 0.494  | 0.350| 0.367| 0.327| 0.296|
| Error-Snv2Indel                | 0.194| 0.160  | 0.282| 0.307| 0.281| 0.248|
| False negative (FN)            | 0.558| 0.712  | 0.244| 0.245| 0.228| 0.255|
| FN-withAlignment               | 0.371| 0.587  | 0.145| 0.144| 0.134| 0.150|
| FN-FractionNoAlignment         | 0.532| 0.424  | 0.474| 0.482| 0.475| 0.484|
| False positive (FP)            | 0.871| 0.267  | 0.388| 0.462| 0.396| 0.151|
| FP-indel                       | 0.187| 0.060  | 0.178| 0.219| 0.188| 0.067|
| FP-hetSNV                      | 0.621| 0.191  | 0.195| 0.225| 0.193| 0.077|
| Recall                         | 0.442| 0.287  | 0.754| 0.752| 0.769| 0.742|
| Recall-withAlignment           | 0.628| 0.410  | 0.852| 0.853| 0.862| 0.846|
| Precision                      | 0.145| 0.169  | 0.780| 0.762| 0.814| 0.774|
**Table 2.** Correlation between map-to-reference and map-to-de novo-transcriptome expression at the gene level, for both RSEM and KALLISTO, based upon CPM, TPM, and weighted-adjusted CPM. See equation in methods for details regarding the last of these. Correlations exclude expressed reference genes where there are not associated de novo assembly contigs, i.e. genes that are missing from the assembly. Correlations are calculated for entries with TPM > 0 for either map-to-reference or map-to-de novo assembly. RSEM values in parentheses are for filtered assemblies from which contigs with TPM ≤ 1 and without a BLASTP hit to Uniref90 are removed, and which are only compared for protein coding genes with a minimum transcript length ≥ 200bp. The % missing column indicates expressed protein coding genes that were not recovered by the unfiltered transcriptome assembly. The % lost-to-filtering column indicates what fraction of expressed protein coding genes that are detected in the unfiltered *de novo* transcriptome assemblies are lost due to filtering. Bold text indicates, for each assembly, the method that produces the strongest correlation with map-to-reference expression estimates.

| Sample | CPM  | TPM  | Rescaled CPM |
|--------|------|------|--------------|
|        | RSEM | kallisto | RSEM | kallisto | RSEM | kallisto | % missing | % lost-to-filtering |
| MDC    |      |        |      |        |      |        |            |                  |
| Trinity| 0.853 | 0.858 | 0.756 | 0.675 | 0.901 | 0.823 | 20.0 (2.0) | 10.6 (6.0) |
|        | (0.951) | | (0.933) | | | | | |
| Shannon| 0.629 | 0.833 | 0.567 | 0.662 | 0.680 | 0.784 | 21.3 (2.5) | 31.5 (27.1) |
|        | (0.897) | | (0.901) | | | | | |
| BinPacker| 0.824 | 0.825 | 0.796 | 0.754 | 0.877 | 0.783 | 25.0 (3.3) | 9.1 (6.8) |
|        | (0.940) | | (0.937) | | | | | |
| BALB/c |      |        |      |        |      |        |            |                  |
| Trinity| 0.869 | 0.881 | 0.818 | 0.819 | 0.897 | 0.863 | 8.6 (0.6%) | 24.5 (11.5) |
|        | (0.934) | | (0.895) | | | | | |
| Shannon| 0.819 | 0.859 | 0.790 | 0.815 | 0.854 | 0.826 | 8.8 (0.7) | 32.5 (19.1) |
|        | (0.846) | | (0.848) | | | | | |
| BinPacker| 0.835 | 0.846 | 0.801 | 0.803 | 0.870 | 0.824 | 12.4 (1.7) | 20.6 (10.5) |
|        | (0.877) | | (0.864) | | | | | |
| FRA    |      |        |      |        |      |        |            |                  |
| Trinity| 0.850 | 0.861 | 0.809 | 0.790 | 0.903 | 0.841 | 11.8 (2.0) | 20.6 (10.6) |
|        | (0.927) | | (0.928) | | | | | |
| Shannon| 0.849 | 0.857 | 0.812 | 0.785 | 0.906 | 0.819 | 11.5 (1.6) | 24.7 (14.1) |
|       | BinPacker | Trinity   | Shannon  | BinPacker | Trinity   | Shannon  | BinPacker | Trinity   | Shannon  |
|-------|-----------|-----------|----------|-----------|-----------|----------|-----------|-----------|----------|
| DEU   | 0.815     | 0.821     | 0.808    | 0.786     | 0.872     | 0.793    | 18.3 (3.2) | 14.1 (10.5) |
|       | (0.888)   | (0.902)   | (0.896)  | (0.900)   | (0.902)   | (0.900)  | (0.902)   | (0.902)   |
| Trinity | 0.848     | 0.859     | 0.804    | 0.788     | 0.903     | 0.838    | 11.7 (2.0) | 21.9 (11.5) |
|       | (0.925)   | (0.926)   | (0.894)  | (0.902)   | (0.926)   | (0.888)  | (0.926)   | (0.926)   |
| Shannon | 0.847     | 0.855     | 0.812    | 0.784     | 0.905     | 0.813    | 11.3 (1.6) | 25.6 (14.9) |
|       | (0.879)   | (0.894)   | (0.894)  | (0.902)   | (0.888)   | (0.902)  | (0.902)   | (0.902)   |
| BinPacker | 0.811     | 0.818     | 0.808    | 0.784     | 0.873     | 0.789    | 17.6 (3.1) | 14.7 (10.5) |
|       | (0.880)   | (0.900)   | (0.890)  | (0.902)   | (0.890)   | (0.902)  | (0.902)   | (0.902)   |
| IRN   | 0.853     | 0.862     | 0.806    | 0.785     | 0.906     | 0.842    | 12.0 (2.0) | 18.9 (10.0) |
|       | (0.929)   | (0.928)   | (0.928)  | (0.928)   | (0.928)   | (0.928)  | (0.928)   | (0.928)   |
| Shinan | 0.847     | 0.854     | 0.804    | 0.777     | 0.904     | 0.817    | 11.5 (1.7) | 22.1 (12.6) |
|       | (0.893)   | (0.902)   | (0.902)  | (0.902)   | (0.902)   | (0.902)  | (0.902)   | (0.902)   |
| BinPacker | 0.809     | 0.816     | 0.799    | 0.773     | 0.869     | 0.787    | 19.0 (3.6) | 13.1 (10.2) |
|       | (0.888)   | (0.901)   | (0.901)  | (0.901)   | (0.898)   | (0.902)  | (0.902)   | (0.902)   |
| KZK   | 0.852     | 0.862     | 0.810    | 0.791     | 0.903     | 0.841    | 10.9 (2.0) | 22.8 (12.0) |
|       | (0.926)   | (0.925)   | (0.925)  | (0.925)   | (0.925)   | (0.925)  | (0.925)   | (0.925)   |
| Shannon | 0.852     | 0.858     | 0.814    | 0.793     | 0.907     | 0.824    | 10.7 (1.8) | 25.7 (14.8) |
|       | (0.885)   | (0.891)   | (0.891)  | (0.891)   | (0.893)   | (0.893)  | (0.893)   | (0.893)   |
| BinPacker | 0.816     | 0.821     | 0.809    | 0.788     | 0.872     | 0.795    | 16.8 (3.4) | 14.9 (10.7) |
|       | (0.894)   | (0.908)   | (0.908)  | (0.908)   | (0.902)   | (0.902)  | (0.902)   | (0.902)   |
FIGURE LEGENDS

Figure 1. Heterozygosity per base for exonic biallelic SNV variants called relative to the Mus reference genome using the GATK best practices pipeline versus those obtained from supertranscripts constructed from TRINITY assemblies using scripts provided with the TRINITY v. 2.6.5 package. See Methods for details concerning how exonic coordinates were obtained for SuperTranscripts.

Figure 2. For the FRA data set, comparison of gene-level expression summed across BLAT alignments of TRINITY contigs to reference transcripts (grouped by gene), to reference gene expression in terms of (a) CPM, (b) TPM, and (c) weighted-adjusted CPM that accounts for differences in effective length estimates between methods. For CPM, TPM, and rescaled CPM, 0.05, 0.01 and 0.05 are added to values prior to log-transformation, in order to visualize instances where expression = 0 for one of the methods being compared. The red line represents unity, the blue line is the smoothed fit to data. Gray and orange points indicate log-fold change of Trinity/map-to-reference genes less than and ≥ 2, respectively. We exclude all cases where an expressed reference gene has no TRINITY contigs aligned to it, such that it is missing from the de novo transcriptome assembly.

Figure 3. For the FRA data set, for genes with non-zero expression in both transcriptome assembly and map-to-reference, log-fold change between TRINITY and map-to-reference expression estimates as a function of difference in effective length (EL) for (a) CPM, (b) TPM, and (c) weighted-adjusted CPM.
Figure 4. Frequency of bias effect sizes for gene-level expression estimated by CPM, TPM, and weighted-adjusted CPM for a TRINITY-based comparison of the FRA and DEU samples.

Figure 5. Histogram of map-to-reference TPM for expressed protein-coding genes that were detected in the FRA TRINITY assembly prior to filtering out un-annotated and lowly expressed assembly contigs, but were lost due to filtering.
Unfiltered

\(\log_{10}(\text{Trinity contig-summed CPM})\) vs \(\log_{10}(\text{reference gene CPM})\)

\(\rho = 0.85, p = 2.2 \times 10^{-16}\)

\(\log_{10}(\text{Trinity contig-summed TPM})\) vs \(\log_{10}(\text{reference gene TPM})\)

\(\rho = 0.81, p = 2.2 \times 10^{-16}\)

\(\log_{10}(\text{Trinity rescaled CPM})\) vs \(\log_{10}(\text{reference gene CPM})\)

\(\rho = 0.90, p = 2.2 \times 10^{-16}\)

Filtered

\(\log_{10}(\text{Trinity contig-summed CPM})\) vs \(\log_{10}(\text{reference gene CPM})\)

\(\rho = 0.93, p = 2.2 \times 10^{-16}\)

\(\log_{10}(\text{Trinity contig-summed TPM})\) vs \(\log_{10}(\text{reference gene TPM})\)

\(\rho = 0.93, p = 2.2 \times 10^{-16}\)

\(\log_{10}(\text{Trinity rescaled CPM})\) vs \(\log_{10}(\text{reference gene CPM})\)

\(\rho = 0.93, p = 2.2 \times 10^{-16}\)
\(\rho = 0.50\)

\(\rho = -0.71\)

\(\rho = 0.049\)
all missing: n=3382 (20.1%)
TPM ≥ 1: n=1468 (10.6%)