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, untested 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. Using publicly available data for model organisms, we demonstrate that, across assembly algorithms, species, and data sets, these assumptions are consistently violated. Using standard filtering approaches, coverage of annotated genes by transcriptome assemblies falls far below complete coverage, even at the highest expression levels, and hundreds of expressed genes are not assembled by any of the algorithms we evaluated. Assembly contigs are compositionally biased towards non-coding bases relative to what is observed in the sequence reads, with a large fraction of contigs being entirely composed of intron or UTR sequence. Assemblies are compositionally biased at the nucleotide level as well, with genotyping error rates ranging from 30-87%. Most genotyping errors are due to false negatives, which themselves are dominated by cases in which contigs were not assembled for particular genes. Diversity is underestimated in transcriptome assemblies, with consistent under-estimation of heterozygosity in all but the most inbred sample we assayed. Finally, even at the gene level and when one sums across contigs annotated to those genes, expression estimates show wide deviations from map-to-reference estimates, and positive bias at lower expression levels. Much of the bias and noise is due to incorrect estimation of the effective length of transcripts and genes, which is fundamental to abundance calculations. Length-scaled abundance estimators partly alleviate this problem, and more pipelines should be developed to leverage them. Given the rapidly decreasing costs of genome assembly, researchers should

carefully consider genome assembly, where feasible, as a prelude to expression and population genetic analyses for non-model organisms, and consider ways to minimize the impact of bias in transcriptome assemblies when genome assembly is infeasible.

1. INTRODUCTION

Although early advances fueled by RNA-seq were restricted to genome-enabled species, the advent 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 utilizing 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. These assumptions are rarely tested. When tests are carried out, they typically employ reads simulated from curated transcript annotations that represent idealized data (e.g. Vijay, Poelstra, Künstner, & Wolf, 2013) . Similarly, empirical investigations typically rely on metrics of contiguity such as N_{50}, completeness with respect to the proportion of conserved orthologs recovered, i.e. BUSCOs (Simão, Waterhouse, Ioannidis, Kriventseva, & Zdobnov, 2015), and dimensionless quality metrics used to guide assembly parameter tuning using tools such as DETONATE (B. Li et

al., 2014) or TRANSRATE (Smith-Unna, Boursnell, Patro, Hibberd, & Kelly, 2016). While these are standard metrics used to benchmark new assembly tools, they do not explicitly measure or test the robustness of expression estimates derived from the assembled contigs, or whether genotypes, allele frequencies and population genetic parameters that are important for evolutionary inference can be accurately estimated from a transcriptome assembly.

Indirect evidence suggests that violations of the assumptions noted above 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). While gene absence is often attributed to the difficulties of assembling lowly expressed transcripts, little is known about that kinds of genes that are missing, whether expression level is their only important unifying feature, and whether they are biased towards functional categories of interest to evolutionary ecologists. Second, 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 and the allele frequency estimates and other population genetic parameters derived from them.

Below, we evaluate the robustness of the assumptions inherent to de novo transcriptome assembly, how their violation might distort evolutionary inferences, and identify particular features of transcriptome assemblies that should be considered during the study design phase of a project. Where possible, we demonstrate methods that ameliorate assumption violations and

augment the robustness of downstream results. While phylogenomic studies can discard the majority of contigs using aggressive filtering schemes and still obtain recover a large number of phylogenetically informative sites in conserved orthologs, these aggressive filtering options 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 Reference genomes

For the purpose of benchmarking de novo transcriptome assemblies, and comparing assembly and reference-based expression estimates, we downloaded from ENSEMBL genome and gtf annotation files for the following organisms: house mouse, *Mus musculus* C57BL/6J (GRCm38); clawed frog, *Xenopus tropicalis* (JGI_4.2); pufferfish, *Tetraodon nigroviridis* (TETRAODON8); and fly, *Drosophila melanogaster* (BDGP6).

2.2 RNA-seq data

Brain tissue is both transcriptionally complex and expected to express a large proportion of an organism's overall transcriptional profile. For this reason, the bulk of our analyses focus on de novo assemblies for data generated for experiments involving mouse (*Mus* spp.) brain. These 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 samples from wild *M. musculus domesticus* from Massif Central, France (FRA), Iran (IRN), Kazakhstan (KZK), and Germany (DEU). To assess the generality or particular

results with respect to assembly composition, we also generated assemblies for pufferfish (*Tetraodon nigroviridis*) whole brain, clawed frog (*Xenopus tropicalis*) kidney, and fly (*Drosophila melanogaster*) heads. Data set SRA accessions, sequencing strategy and depth are summarized in Supplementary Table 1. All libraries except for MDC were sequenced on an Illumina HiSeq 2000; MDC was sequenced on an Illumina GAII.

2.3 Short read processing

After an initial assessment of sequences reads with FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), we performed kmer-based error correction with RCORRECTOR (Song & Florea, 2015), followed by adapter and quality trimming with TRIMGALORE! (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), setting both the minimum phred quality score and adapter filtering stringency parameter to 5, and the minimum post-trimming read length to 36. We based our quality trimming threshold upon previous findings that performing stricter quality filtering can negatively impact assemblies (MacManes, 2014). A second run of FASTQC was used to confirm that there was no remaining adapter sequence in the filtered reads.

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, and did not perform *in silico* normalization for any of the data sets. For the *Mus* data sets, to evaluate the robustness of some key findings to the algorithm employed, we also built assemblies using two promising, recently develop 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, on the recommendation of its author (Kannan, personal communication), we treated reads as un-stranded in all SHANNON assemblies.

To perform an initial assessment of assembly quality, we calculated standard metrics typically reported in de novo transcriptome studies. First, we quantified the total number of contigs ("transcripts" in TRINITY's nomenclature), the number of TRINITY-defined "genes" comprising contigs deriving from the same graph, N_{50}, and the percent of contigs that were successfully mapped to the genome. We then assessed read support for the assembly, by mapping reads back to the assembly to assess the proportion that mapped concordantly, using methods specified by the developers of TRINITY (https://github.com/trinityrnaseq/trinityrnaseq/wiki/RNA-Seq-Read-Representation-by-Trinity-Assembly). Finally, we quantified the number of completely covered, fragmented, and missing single copy conserved orthologs using the program BUSCO v. 3.0.2 (Simão et al., 2015), using the vertebrate ortholog model set for all assemblies but the fly, for which we used the arthropod set.

2.5 Assembly and read functional composition

For each contig within an assembly and for the assembly as a whole, 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 first mapped contigs to their respective reference genomes with GMAP v. 2016-06-30 (Wu & Watanabe, 2005). We then filtered out unmapped contigs with

SAMTOOLS v. 1.5. (H. Li et al., 2009). For the analysis of whole assemblies, we then used BEDTOOLS v. 2.25 (Quinlan & Hall, 2010) to sort, and merge intervals such that each reported interval consistent of unique genomic nucleotides. Mapped intervals for GMAP alignments were converted to bed format with the bamToBed function, with the '-splitD' argument such that each mapped interval for a contig—excluding alignment gaps corresponding to putative splicing events—was converted into a separate row.

To calculate the proportions of transcriptome assemblies corresponding to particular functional elements, we employed a hierarchical classification of genomic nucleotides (Supplementary Text). In order to determine if the functional composition of RNA-seq reads differed from that of the transcriptome assemblies, we randomly sub-sampled one million paired-end reads, and aligned them to their respective genomes using HISAT2 v. 2.0.4 (Kim, Langmead, & Salzberg, 2015). We then performed the same hierarchical classification as was done for assembled contigs, calculating frequencies for individual sequence fragments, i.e. across both sequence reads when both were properly mapped. For multi-mapped reads, we calculated a weighted average across all mappings. Visualization of functional composition was obtained by randomly sub-sampling 100,000 reads that successfully aligned to the genome.

2.6 Coverage of reference transcripts and genes

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 (but no minimum alignment length in the query contig) to consider a hit valid. We quantified transcript coverage in two ways. First, for each reference transcript, we identified the contig with the best alignment to

it, and calculated coverage as the sum of the BLAT alignment intervals. 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. Second, we calculated the coverage of a reference transcript by summing coverage across all contigs with alignments that passed our 95% percent identity criterion. Because a contig map produce valid mappings to multiple transcripts of a gene, in calculating summed coverage we allowed a contig to be used in coverage calculations for >1 reference transcript. Finally, we estimated gene-level coverage as a TPM-weighted average of coverage for individual transcripts belonging to that ENSEMBL gene id. We calculated these three coverage metrics for 1) an unfiltered transcriptome assembly, and 2) for only protein coding transcripts with a minimum length of 200bp—the default minimum contig length for TRINITY—using only contigs with transcripts per million (TPM) ≥ 1, and with a predicted peptide sequence with a BLASTP hit against Uniref90 (downloaded 2017-04-05), We used TRANSDECODER v. 3.0.0 (https://github.com/TransDecoder/TransDecoder/wiki) to predict peptide sequences. For BLASTP alignment, we set the maximum number of target sequences to one, and with a maximum allowable e-value of 1×10^{-5}.

2.7 Expression estimation

For nine combinations of assembly method (TRINITY, SHANNNON, 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). For TRINITY-based analyses, isoform and gene level expression estimates were initially calculated using TRINITY-specified gene and isoform identities for each contig. We also estimated "true" (reference-based)

expression by mapping reads directly onto reference transcripts, using the same RSEM/BOWTIE2 pipeline but with the ENSEMBL GTF annotation.

We assessed concordance between assembly and referenced-derived approaches in three ways. First, we compared transcripts per million (TPM) estimates for reference isoforms and their best hit BLAT contig. Second, we evaluated concordance between reference gene expression and our best estimate of gene-level expression derived from the transcriptome assembly. Specifically, we assigned assembly contigs to ENSEMBL genes based upon their BLAT alignments to ENSEMBL isoforms, complied these assignments into a contig-to-gene map file, and provided this file to RSEM when building the mapping index. RSEM *genes.results output files produced by this protocol contain expression estimates summed across contigs aggregated by reference label; for TRINITY, 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). For all analyses summing across contigs, in addition to TPM we also compared expected counts to determine if such an approach could effectively recover values comparable to gene-level estimates using the map-to-reference approach. Finally, we assessed whether correcting for effective length differences between reference transcripts and assembled contigs using length-scaled TPMs (Soneson, Love, & Robinson, 2015) would improve concordance between de novo assembly and map-to-reference based abundance estimates. We calculated length-scaled TPMs by multiplying TPM by the gene-level effective length (estimated across constituent isoforms or contigs) and divided by the sum of all expected counts; length-scaled TPMs are dimensionless quantities that lie between 0 and 1.

2.8 Analysis of missing genes

We investigated how often expressed genes are missing from transcriptome assemblies and whether those genes are enriched for particular ontologies. For three of the *Mus* brain samples (MDC, BALB/c, FRA), we first identified the set of expressed transcripts with TPM > 0 based upon mapping to their respective reference genomes. We considered a gene present if any of its transcripts were expressed. Conversely, we considered a gene absent from a de novo assembly if no contig with TPM > 0 had a BLAT alignment to mapped to any of a gene's constituent transcripts.

We performed two separate missingness analyses for each sample-assembler combination. These involve differing levels of a) filtering on the assembly, and b) which reference transcripts are considered as the "background" for enrichment analyses. For the first analysis, we included all assembled contigs and their respective BLAT alignments, and used all genes with at least one expressed transcript as background; this includes reference transcripts < 200 bp that Trinity, with default settings, will not assemble, as well as non-coding transcripts (hereafter termed "$\text{background}_{\text{unfiltered}}$"). Second, we only retained contigs with TPM ≥1, and restricted our background to protein-coding reference transcripts ≥ 200 bp ("$\text{background}_{\text{filtered}}$"). In both analyses, we only included in the background gene set those genes with a reasonable probability of detection, which we defined as TPM ≥1 for at least one of a gene's constituent isoforms.

To determine if particular functional categories were over-represented in genes that were missing, we conducted a statistical overrepresentation test using the PANTHER classification system (Mi, Muruganujan, Casagrande, & Thomas, 2013), via http://www.pantherdb.org/. We used the ENSEMBL gene ids associated with the expressed reference transcripts as our "reference

list" and the ids of the missing genes as the "analyzed list", calculating statistical significance of over/under-representation using a Fisher's Exact test.

To determine whether genes that are missing from assemblies are less conserved than those that are present, we downloaded from the UCSC Genome Browser (http://hgdownload.cse.ucsc.edu/goldenpath/mm10/phastCons60way/) the phastCons (Siepel, 2005) score bigWig track derived from multiple alignments of 59 vertebrate genomes to the mouse genome (mm10). We then converted from bigWig format to bedGraph with the KENTUTILS (https://genome-source.gi.ucsc.edu/gitlist/kent.git/blob/master/README) bigWigToBedGraph tool. We calculated mean phastCons scores per ENSEMBL gene (defined as the set of non-redundant CDS, start codon, 3` UTR, 5` UTR, stop codon, and exons associated with an Ensembl gene ID). We compared phastCons scores of genes missing from the assembly to actually expressed genes using a non-parametric Wilcoxon rank sum test.

2.9 Evaluating assembly redundancy

To evaluate contig redundancy, we first quantified the frequency and degree of overlaps between contigs whose best BLAT hits are known, single-isoform genes in *Mus*. For four of these data sets, there are independent estimates of genetic diversity obtained from genome-sequencing (Harr et al., 2016), so we predicted that, if false assembly of haplotypes increased with genetic variation, that the frequency of overlaps would correlate with diversity. We also calculated the redundancy from BLAT all-by-all alignments of individual assemblies. If haplotypes are assembled into different contigs, or contigs are assembled that only differ by a small number of bases, e.g. by differences in assembled intronic sequence, this should manifest in an excess of

cases where alignments cover nearly all of the constituent contigs in those alignments, with few or no mismatches.

2.10 SNP-based analysis of assembly composition

For these analyses, we focused on the *Mus* samples, and compared the map-to-reference genotypes (hereafter MR) to those derived from TRINITY assemblies. We called variants off of the reference genome using the GATK v. 4 (McKenna et al., 2010), and 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 the *Mus* samples representing pools, following the GATK developers' recommendations, we also used the --sample_poidy argument, with a value set to two times the number of individuals in the pool. For TRINITY assemblies, we first constructed SuperTranscripts, hereafter (ST), (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. Similar to our reference-based approach to obtain our truth set, we edited the variant calling script to include the --sample_ploidy argument. We did not use either base quality score or variant quality score recalibration, to reflect the fact that the truth sets needed for these methods are rarely, if ever, available for ecological genomics data. We applied the same strand bias, depth, and quality-by-depth filters to raw variant calls from both methods to produce a final filtered set. To define callable sites for the MR approach, we extracted read alignments to annotated exonic regions, and then calculated the number of exonic sites with a coverage depth ≥ 5. For ST, we performed similar operations, defining exonic coordinates from mappings of ST to the *Mus* genome using

GMAP, and callable sites within exons in the same way as with MR, i.e. coverage depth ≥ 5. We defined heterozygous calls in our filtered vcf files as all bi-allelic SNP sites that passed filtering criteria. In cases where sample ploidy was greater than diploid, we collapsed genotypes into the union of all unique recorded alleles.

We obtained exonic regions for ST by mapping them to the *Mus* genome using GMAP, then intersected these mappings with a bed file of exonic intervals with BEDTOOLS. We used a slightly newer version of GMAP (v. 2018-07-04) than we did for the contig composition, which fixed minor issues with CIGAR string handling that had the potential to generate spurious genotype errors. Python scripts (https://github.com/harvardinformatics/TranscriptomeAssemblyEvaluation) were then used make transformations between genomic and ST coordinates, for the purpose of extracting ST genotypes and writing genotype bed files back to genomic coordinates. ST alignments to the genome were used to define unique polymorphic site ids that included cases where (a) both methods called genotypes, (b) only MR called a genotype (including sites with and without valid ST alignments), or (c) genotypes were only called from ST. We recorded cases where multiple ST 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 MR sites that were indels or contained > 2 alleles. Using genotype calls based upon the reference annotation as the gold standard, we calculated 15 different metrics, defined as follows:

- Error rate: the proportion of ids where the set of alleles recorded by the two methods were not the same, including cases where one method did not call a genotype
- Error-MrefAlleleIncluded: the proportion of genotyping errors (in ST) where the reported alleles included at least one of those reported using MR

- Error-Snv2Indel: the proportion of errors, where a biallelic SNV in MR was called as an indel in ST
- Error-MultipleSuperTranscripts: the proportion of errors where multiple STs were aligned to the genomic position
- FN: the proportion of MR variable sites with no ST genotype
- FN-noalign: the proportion of false negatives that are due to the absence of ST alignments to the genome at the variable position.
- FN-het: the proportion of heterozygous MR genotype calls with either no ST genotype, a fixed alternative (relative to ST) genotype, an indel heterozygote call, or a multi-allelic ST genotype. ST SNV heterozygous genotypes that specify the wrong alleles are not considered FN, which is relevant for comparisons of transcriptome-level heterozygosity between methods.
- FP: the proportion of ST genotypes with either no associated MR genotype, or where ST is polymorphic and MR is fixed for an alternative allele
- FP-het: the proportion of heterozygous biallelic ST genotypes with either no associated MR genotype or a fixed alternative MR genotype
- FP-NAtoFixed: the proportion of FP cases with no MR genotype for which the ST genotype is fixed for an alternative allele (relative to the ST reference).
- FP-NAtoHet: the proportion of FP cases with no MR genotype for which ST is heterozygous
- FP-indel: the proportion of all FPs (both missing and MR-fixed) with an ST indel
- Recall: the proportion of MR genotypes recovered by ST

- Recall-het: the proportion of MR heterozygous genotypes thar are correctly recovered in ST; a genotype is not considered recovered if ST genotype is a heterozygous call comprised of the wrong nucleotides
- Specificity: the proportion of ST genotypes that match the MR genotype

2.11 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 Assemblies

We assembled de novo a total of 21 transcriptomes—six mouse brain data sets using TRINITY, SHANNON, and BINPACKER, as well as TRINITY assemblies for frog kidney, pufferfish brain and fly head. Consistent with previously published reports, assemblies were highly fragmented, with median contig sizes ranging from just a few hundred bases to slightly over 1kb (Supplementary Table 2). Yet, typically employed measures of accuracy and completeness indicate that all assemblies were of reasonably high quality (Supplementary Tables 2-4), particularly those assembled from the *Mus* data sets comprising the majority of our analyses. Important in the context of gene expression studies, large fractions of reads used to build assemblies mapped back to them in a concordant fashion, usually exceeding 90% of the input library (Supplementary Table 3)

3.2 Transcript and gene coverage

Without filtering on the reference transcriptome or the de novo assembly, transcript coverage is increased when one sums coverage across orthologous assembly contigs (Figure 1a-b, Supplementary Figure 1-3). In both methods, and at the level of genes (Figure 1c, Supplementary

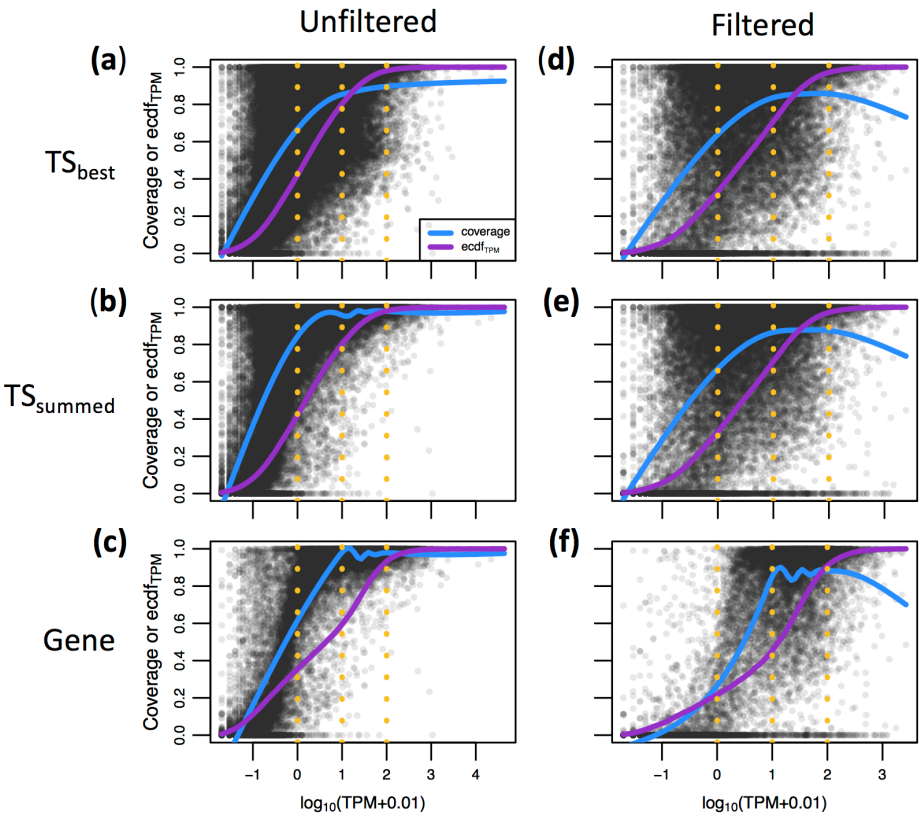

Figure 1. Coverage of individual *Mus* reference transcripts and genes by the TRINITY assembly. Coverage of all expressed reference sequences by the unfiltered assembly is defined by BLAT alignments with a minimum percent identity of 95% for **(a)** individual contigs with the highest scoring alignment, **(b)** summed coverage of all BLAT hits and **(c)** gene-level coverage as the expression-weighted average of summed coverage across isoforms. Comparable coverage estimates for only protein-coding transcripts/genes of minimum length 200bp, by contigs with TPM ≥ 1 and an open-reading frame alignment to UNIREF90 are represented by **(d-f)**. Blue lines indicate smoothed, fitted trends on the coverage-TPM relationship, while purple lines indicate the empirical cumulative frequency distribution on TPM.

Figures 1-3), coverage generally tracks reference expression level, although these correlations are weaker at the transcript level, and weakest for coverage of a single best-hit contig. For summed coverage of transcripts and gene coverage, while there is an expectation of complete

coverage above TPM=10, there are still many transcripts and, to a lesser extent, genes that fall far below complete coverage. This "unfiltered" approach represents an idealized scenario in which contigs are mapped directly to a reference annotation originating from the same species. If one considers only protein-coding genes and transcripts ≥ 200bp—which should favor assembly coverage—but takes a standard approach of filtering out lowly expressed contigs, and only considering contigs with hits to UNIREF90 (a standard approach for assigning orthology without an annotated genome)—coverage decreases substantially at both the transcript and gene level (Figure 1d-f, Supplementary Figures 1-3). The expectation of coverage never reaches 100%, regardless of expression level of the underlying reference structures, and there is far more variance around the coverage-TPM trend. The decreases observed in these trends at the highest expression levels result from poorly covered annotations making up an increasing fraction of the most highly expressed transcripts and genes.

3.3 Missing genes

Only considering genes that are expressed at a level to permit a reasonable chance of being detected (TPM ≥ 1), depending upon the sample, 17-78 (0.015-0.052%) of genes were missing from all methods' unfiltered assemblies, and 25-115 (0.23-0.77%) genes were unrecovered by 1-2 assembly methods (Table 1, Supplementary Figure 4a). Practically, unfiltered assemblies are rarely, if ever, useable, so to test a more realistic scenario we also estimated the number of genes missing from filtered assemblies. Missing from TRINITY assemblies were 3.5% (293), 2.6% (277), and 4.7% (529) protein coding genes for MDC, BALB/c and FRA, respectively. Numbers were substantially worse for SHANNON (MDC, 19.1% (1585); BALB/c, 9.9% (1039); FRA, 7.5% (836)), and better in two of three samples for BINPACKER (MDC, 3.0% (251); BALB/c 2.9%

Table 1. Counts of missing genes by assembler for three *Mus* samples.

	Sample					
	MDC		BALB/c		FRA	
Missingness pattern[a]	**Unfiltered**	**Filtered[b]**	**Unfiltered**	**Filtered[b]**	**Unfiltered**	**Filtered[b]**
PPP	11055 (99.6)	6646 (79.8)	14777 (98.7)	9167 (87.2)	16350 (99.3)	10170 (90.6)
MPP	1 (0.0)	46 (0.6)	47 (0.3)	197 (1.9)	7 (0.0)	106 (0.9)
PMP	2 (0.0)	1341 (16.1)	15 (0.1)	638 (6.1)	12 (0.1)	379 (3.4)
PPM	20 (0.2)	31 (0.4)	27 (0.2)	52 (0.5)	54 (0.3)	81 (0.7)
PMM	0 (0)	18 (0.2)	6 (0.0)	40 (0.4)	3 (0.0)	33 (0.3)
MPM	2 (0.0)	21 (0.3)	13 (0.1)	58 (0.6)	5 (0.0)	28 (0.2)
MMP	0 (0)	45 (0.5)	7 (0.0)	205 (1.9)	0 (0)	226 (2.0)
MMM	17 (0.2)	181 (2.2)	78 (0.5)	156 (1.5)	32 (0.2)	198 (1.8)

[a] Present (P) and missing (M) codes reflect the following assembler order: Trinity, Shannon, BinPacker.

[b] Restricted to protein-coding genes with minimum transcript length of 200 and TPM ≥ 1, based upon contigs with TPM ≥ 1 and a hit to Uniref90.

(306); FRA 3.0% (340)). Overall, any one assembler-sample combination led to hundreds to over 1,000 expressed genes being missed, at least after standard assembly filtering, which unavoidably removes both spurious contigs and contigs representing real transcripts. In general, genes not recovered by one method tended to be missing from all assemblers (Table 1, Supplementary Figure 4b); SHANNON tended to show greater missingness for protein coding genes than TRINITY and BINPACKER. With the exception of MDC, genes missed by all or some assemblers had significantly lower reference gene expression and conservation (Figure 2, Supplementary Figure 5, Table 2). Even so, there were plenty of cases where conserved genes expressed at a relatively high level (TPM ≥ 100) were not recovered by any assembler (Figure 2). Correlations between TPM and phastCons scores were, while statistically significant, relatively weak (MDC, Spearman's $\rho = 0.082$, $p = 2.2 \times 10^{-16}$; BALB/c, Spearman's $\rho = 0.18$, $p = 2.2 \times 10^{-16}$; FRA, Spearman's $\rho = 0.23$, $p = 2.2 \times 10^{-16}$). Missing genes were not enriched for any GO-Slim Biological Process terms.

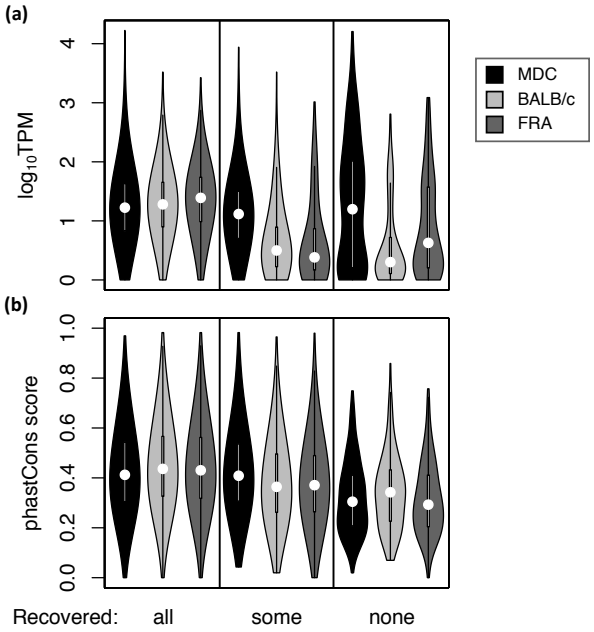

Figure 2. Violin plots of **(a)** TPM and **(b)** phastCons scores for protein coding genes with a minimum transcript length of 200bp, which are recovered in all three transcriptome assemblers (TRINITY, SHANNON, BINPACKER), 1-2 of three assemblers, or that are missing from all three assemblers. Missingness is determined using filtered transcriptome assemblies, that only include contigs with TPM ≥ 1 and an open-reading frame alignment to UNIREF90.

Table 2. For filtered assemblies, median genomic alignment-based transcript per million (TPM) and phastCons scores for ENSEMBL genes that are recovered by all de novo transcriptome assembly methods (All), recovered by 1-2 of the 3 methods (Some), and recovered by no assembler (None). Flagged tests are significant after Bonferroni correction; * P < 0.005, ** P <0.0005, and *** P < 1.0 x 10^{-10}.

	TPM						phastCons					
Sample	**All**	**Some**	**None**	**All vs. Some**	**All vs. None**	**Some vs. None**	**All**	**Some**	**None**	**All vs. Some**	**All vs. None**	**Some vs. None**
MDC	16.8	13.1	15.8	***	NS	NS	0.41	0.41	0.30	NS	***	***
BALB/c	19.0	3.2	2.0	***	***	**	0.44	0.36	0.34	***	***	*
FRA	24.5	2.4	4.2	***	***	**	0.43	0.37	0.29	***	***	***

3.4 Assembly vs. read functional composition

For a gene to be recovered in an assembly, it has to be annotated. And, because transcriptome assembly is motivated by the absence of a closely related reference genome, in nearly all cases transcriptome annotation involves predicting open reading frames and searching protein databases such as UNIREF90 and SWISSPROT. The greater the fraction of non-coding bases in a

contig, the harder it will be to align it to an annotated peptide. Thus, the functional composition of contigs sheds light on patterns of missingness at both the transcript and gene levels. Across all sample-assembler combinations, large fractions of contigs are dominated by intronic, UTR, and intergenic sequence (Figure 3, Supplementary Figure 9). The extent of intronic sequence is particularly surprising for *Mus* and *Drosophila* assemblies, as current annotations should be inclusive of the majority of alternative splice variants. That the excess of intronic sequence in contigs is not reflected in the composition of reads (Figure 3, Supplementary Figure 6) can be attributed to the sensitivity of the assembly algorithms, which require minimal kmer coverage from just a few intron-containing reads to extend contigs across exon boundaries. The prevalence of intronic and intergenic sequence in assembly contigs indicate 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 (Supplementary figure 7). The mean contig fraction comprised of intron sequence for TRINITY assemblies ranged from 7.5% (*Xenopus*) to 56.5% (FRA) (Supplementary table 5). Mean fractions for intergenic sequence ranged from 8.1% (MDC) to 39.0% (*Xenopus*). Results for SHANNON and BINPACKER assemblies of MDC, BALB/c and FRA were similar to those observed in TRINITY.

Redundant assembly of identical contigs from and assembly of different haplotypes as unique contigs can create complications for downstream analyses. Contig multiplicity can influence downward expression estimates of contig copies—particularly if redundant contigs are not properly assigned to the same gene-level annotation—and decrease the signal-to-noise ratio for differential expression analyses. Assembling haplotypes as contigs can lead to false positive differential expression results, e.g. if frequencies of haplotypes differ between populations. It

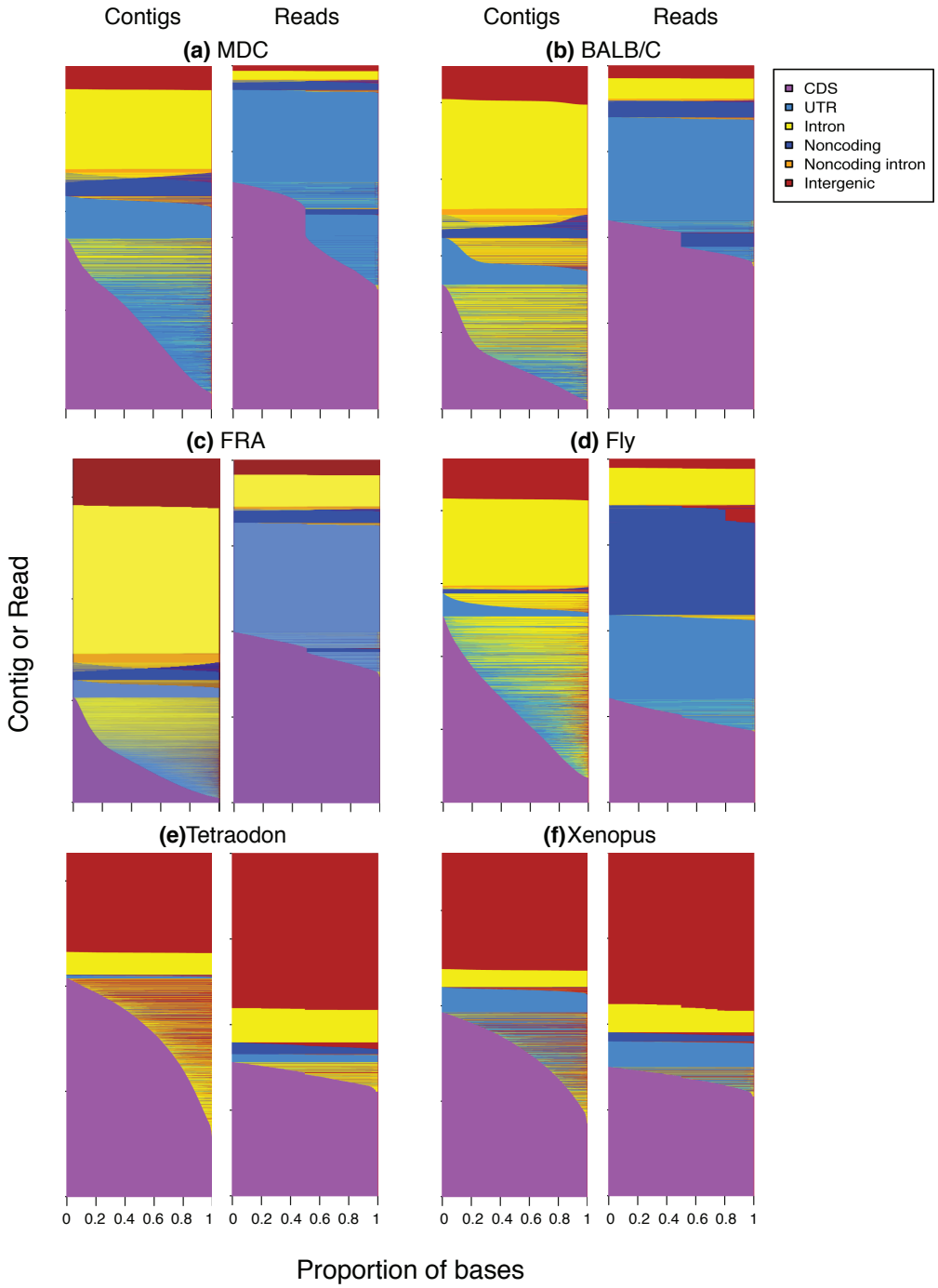

Figure 3. Base composition by TRINITY contig (left panels), and sequence reads (right panels) for **(a)** *M. musculus* dendritic cells, **(b)** *M. musculus* BALB/c strain brain, **(c)** wild outcrossed *M. musculus* brain, **(d)** *D. melanogaster* head, **(e)** *T. nigroviridis* brain, and **(f)** *X. tropicalis* kidney. For each panel, each y-axis row represents a single contig or read. Sequence panels for reads are based upon 100,000 randomly sampled mapped reads. Intersections of contigs and reads with annotations are based upon mapping with GMAP and HISAT2, respectively (see **Methods** for details). For visualization purposes, sequence proportions are arranged from the bottom to the top of a panel, sorted in descending order, consecutively by CDS, UTR, non-coding, non-coding intron, intron, and intergenic proportions.

can also distort allele frequency spectra that are critical for analyses of allele-specific expression. Across all three assemblers, the frequency of contig overlaps within known single-isoform genes was bimodal. Most genes were overlapped at few or no positions by ≥ 1 contig, but a large number of genes were also covered by multiple contigs at most or all of their positions (Figure 4). SHANNON, in particular, showed a troubling increasing trend in frequency with proportional reference gene coverage (Figure 4b), indicating it may be more prone to redundancy than either TRINITY or BINPACKER. Overall, 11.9%-12.2%, 33.2%-35.3%, and 16.9%-18.7% of single isoform genes were covered by more than one contig at all transcript positions for TRINITY, SHANNON, and BINPACKER, respectively. The frequency of contig overlaps has no relationship with estimates of genetic diversity obtained from whole genome sequencing (Figure 4), a pattern not to be expected to occur if redundancy is driven solely by the assembly of polymorphic haplotypes into separate contigs. Considering the variation in CDS and intron content among contigs that cover a single-isoform gene, while there are plenty of cases with variable assembly of CDS among contigs, the event of highest frequency across all assemblers appears to be one where there is no variation in CDS among contigs but variable intron retention, due to either transcription of intron sequence or the inclusion of un-spliced pre-mRNA in sequencing libraries (Supplemental Figure 8). Looking across all contigs, regardless of their genomic position of origin, all assemblies contain large numbers of contigs that are nearly identical (Figure 5a, Supplemental Figure 9). For TRINITY assemblies that putatively report "genes" as components, while the redundancy is greater for contigs assembled from the same graph, there are a substantial number of contigs with high similarity that originate from different TRINITY "genes" (Supplementary Figure 10). From different component graphs, Trinity will erroneously assemble haplotypes that differ only by a few SNVs (Figure 5b).

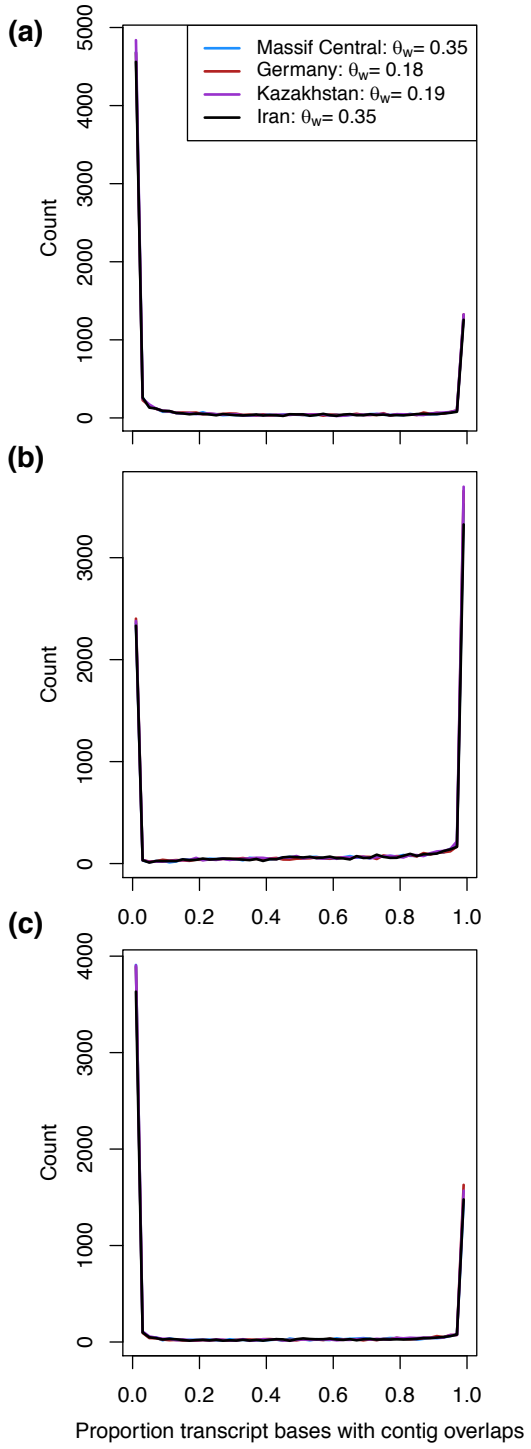

Figure 4. Frequency of proportional overlaps between contigs, based upon BLAT all-by-all hits, for known *Mus* single isoform genes, for three 8-individual pools of wild *M. m. domesticus* (Iran; Massif Central, France and Germany) and one 8-individual pool of *M. m. musculus* (Kazakhstan), for **(a)** TRINITY, **(b)** SHANNON and **(c)** BINPACKER. Estimates of Watterson's θ are those reported from whole genome sequence data in Harr et al. (2016).

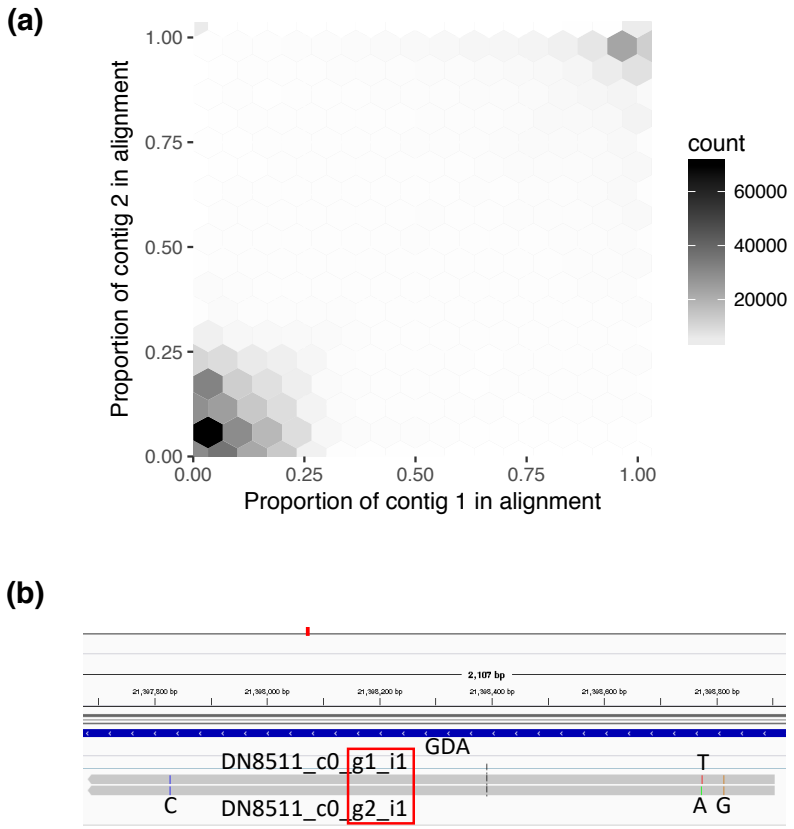

Figure 5. (a) For FRA TRINITY assembly, hexbin 2D histogram of the proportion of contig lengths comprising BLAT pairwise alignments resulting from all-by-all searches. **(b)** Example alignment to the *Mus* genome, to a protein-coding transcript of the GDA gene, of nearly identically sized (1220 bp) FRA TRINITY contigs. These contigs have a BLAT-based 99.8 % identity in an alignment equal to the entirety of the contig length, and which are assembled from different TRINITY components (i.e. "genes"), as indicated by the red box around their labels. The contigs differ by only two adjacent SNPs (A and T), and also share fixed C (far left) and G (far right) differences relative to the reference genome. Thus, different Trinity components have assembled different haplotypes.

3.5 SNP calling and heterozygosity

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), F_{ST} outlier analysis (Bankers et al., 2017) , and tests for allele-specific expression (Singh

et al., 2018). Our ST mapped at extremely high rates to the *Mus* genome (MDC, 99.9%; BALB/c, 99.8%; FRA, 99.8%; DEU, 99.8%; IRN, 99.8%; KZK, 99.7%), such that any discrepancies between genotyping methods cannot be attributed to low alignment efficiency. 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, when two contigs represent haplotypes: at a variant position, the true heterozygous position is instead represented as two separate genotypes, one homozygous for the reference allele and one homozygous for the alternative allele. 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. We compared genotype calls and derived diversity metrics between the map-to-reference and map-to-transcriptome approaches using 15 different metrics, in order to determine whether such genotypic distortions occur in de novo transcriptome assemblies. The map-to-reference (MR) approach involves mapping sequence reads to the genome, then calling genotypes with GATK based upon those genomic alignments. In contrast, the transcriptome assembly-based approach involves aligning reads to the ST collapse of a TRINITY assembly, then calling variants with GATK relative to the ST. Comparisons between MR and ST genotypes are based upon orthology determined by aligning the ST to the reference genome. To see example cases of how particular errors are flagged, see Figure 6.

The frequency of errors, in which the allele set recovered by the two methods differed or in which only one of the two methods called a genotype at all, was high, ranging from 30-38% for the wild samples, and surprisingly, ≥ 50% for the inbred samples (Table 3). While the median per-base error rate per ST was 0.2% or less, maximum per ST rates typically exceeded 10%

Genome sequence	**G**	**G**	**G**	**G**	**G**	**G**	**G**	**G**
MR genotype	**G/G**	**G/G**	**G/G**	**A/G**	**A/A**	**A/G**	**A/G**	**A/A**
ST sequence	**G**	**G**	**G**	**AT**	**A**	**G**	**_**	**G**
ST genotype	**A/G**	**G/GT**	**T/T**	**A/AT**	**A/T**	**G/G**	**_**	**G/T**
STs aligned to genome	**1**	**1**	**1**	**1**	**1**	**1**	**0**	**2**
Error classification								
Error	●	●	●	●		●		●
Error-MrefAlleleIncluded	●	●		●	●			
Error-MultipleSuperTranscripts								●
Error-Snv2Indel		●		●				
FN						●	●	
FN-noalign							●	
FN-het						●	●	
FP	●	●	●					
FP-het	●	●			●			
FP-NAtoFixed			●					
FP-NAtoHet	●	●						
FP-indel		●						

Figure 6. Examples of map-to-reference and and SuperTranscript derived genotype combinations that produce particular error classes.

Table 3. 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. Values in parentheses are after converting SuperTranscripts with no protein hit to UNIREF90 to NA.

	Sample					
Metric	**MDC**	**BALB/c**	**FRA**	**DEU**	**IRN**	**KZK**
Error rate	0.870 (0.892)	0.500 (0.477)	0.359 (0.453)	0.376	0.336	0.301
Error-MrefAlleleIncluded	1.000 (1.000)	0.898 (0.898)	0.969 (0.971)	0.977	0.969	0.959
Error-Snv2Indel	0.083 (0.111)	0.250 (0.250)	0.446 (0.560)	0.457	0.484	0.449
Error-MultipleSuperTranscripts	0.004 (0.004)	0.005 (0.002)	0.003 (0.001)	0.004	0.003	0.002
FN	0.662 (0.754)	0.389 (0.412)	0.260 (0.396)	0.265	0.246	0.239
FN-noalign	0.728 (0.868)	0.745 (0.881)	0.585 (0.860)	0.585	0.572	0.668
FN-het	0.623 (0.752)	0.742 (0.885)	0.259 (0.451)	0.259	0.242	0.268
FP	0.609 (0.558)	0.179 (0.110)	0.129 (0.091)	0.147	0.115	0.078
FP-het	0.677 (0.660)	0.188 (0.118)	0.295 (0.224)	0.352	0.303	0.113
FP-NAtoFixed	0.118 (0.094)	0.053 (0.051)	0.027 (0.019)	0.028	0.026	0.024
FP-NAtoHet	0.882 (0.906)	0.919 (0.928)	0.916 (0.923)	0.909	0.917	0.915
FP-indel	0.188 (0.196)	0.063 (0.048)	0.184 (0.146)	0.225	0.194	0.070
Recall	0.334 (0.243)	0.609 (0.587)	0.735 (0.602)	0.731	0.750	0.757
Recall-het	0.199 (0.131)	0.038 (0.017)	0.510 (0.377)	0.536	0.558	0.288
Specificity	0.176 (0.161)	0.735 (0.826)	0.829 (0.856)	0.805	0.849	0.897

(Supplementary Table 6). For five of six samples, the median number of errors per ST was 2, but, depending upon the sample, as many as 56-187 errors could occur within a single ST (Supplementary Table 6). On a per-sample basis, errors show a modest yet statistically significant bias towards the distal end of a ST (sample-level Wilcoxon 1-sided tests, $H_0 = 0.5$, all tests, $p \leq 0.002$; Supplementary Figure 11). Considering only contigs with an ORF hit to UNIREF90 changed error rates in a predictable direction, e.g. reduced false positives and recall, and increased false negatives (Table 3).

In 90-100% of all erroneous ST genotype calls, at least one MR allele was included in the ST allele set (Table 3, Error-MrefAlleleIncluded). False negatives (FN; 24-66%) occurred far more frequently than false positives (FP; 8-61%). Anywhere from 57-75% of FN were due to there being no ST aligned at a genomic position where a MR genotype call was made (Table 3, FN-noalign). Given that the ST mapping rates to the genome were extremely high, these FP instances are primarily due to a failure to assembly transcripts overlapping polymorphic genomic positions. The FP rate includes both sites where there is no genotype call in MR, or where MR is fixed for an alternative allele and ST is polymorphic. Looking more closely, paralog collapsing should lead to FP heterozygous ST calls, and FP ST fixed-alternative genotypes will be due local assembly errors; in the latter case, mis-assembly is the most parsimonious explanation, given that the ST reference sequences were derived from the data used to call genotypes. FP-NAtoHet was 5-38 times higher than FP-NAtoFixed, indicating that paralog collapsing plays a larger role in FP than local assembly errors within ST. FP-NAtoHet includes ST indel heterozygotes, which are not unexpected in cases of paralog collapse; The frequencies of cases where MR SNVs are called as indels in ST (Error-Snv2Indel) approach 50% in the wild samples. This class of error as well as the somewhat less frequent occurrence of false positive indel calls (FP-indel) collectively

suggest that paralog collapses are a general problem, if not as large a contributor to overall error rates as FN. When genotypes were called in STs, the frequency of calls that were correct (i.e. specificity) ranged from 81-90% for the wild samples, but specificity for the inbred samples, particularly MDC, was substantially lower. Recall ranged from 73-76% in the wild samples; as with specificity, recall for the inbred MDC sample was extremely low.

Because read depth influences the ability of tools such as GATK to call genotypes, we compared read depth at homologous positions, comparing MR depth and median ST depth (to account for the possibility of > 1 ST mapping to a genomic position). Concordant genotypes displayed a high degree of concordance in depth-of-coverage (Supplementary Figure 12). In contrast, on average FN sites had very small reductions in ST read depth relative to MR at sites where STs mapped to the genome (Supplementary Figure 13, blue lines). FN at sites with modest loss of read depth is likely caused by multiple interacting factors that are beyond the scope of this study to investigate in detail. In contrast, FN due to failure to assemble transcripts leads to an effective ST read depth of zero, and including these sites leads to a much more dramatic reduction in ST depth relative to MR (Supplementary Figure 13, yellow lines). With the exception of a relatively small number of cases of zero MR coverage, MR and ST read depth were largely co-linear (Supplementary Figure 14); the causes of noise about those correlations, as with FNs, remain unclear.

One prediction resulting from such genotyping bias is that heterozygosity rates in ST will be negatively biased relative to MR, because alterative alleles will be partitioned amongst different assembly components (i.e. “genes”) and reads overlapping these alleles will preferentially map to their corresponding contigs, leading to separate homozygous reference genotype calls rather than one heterozygous call. Consistent with this prediction, in all but the

inbred (MDC, BALB/c) strain samples, we observe reduced heterozygosity in the exonic regions of ST relative to MR (Figure 7). This bias does not result from ST being preferentially assembled in regions of lower heterozygosity, given that, for the wild samples, MR heterozygosity is

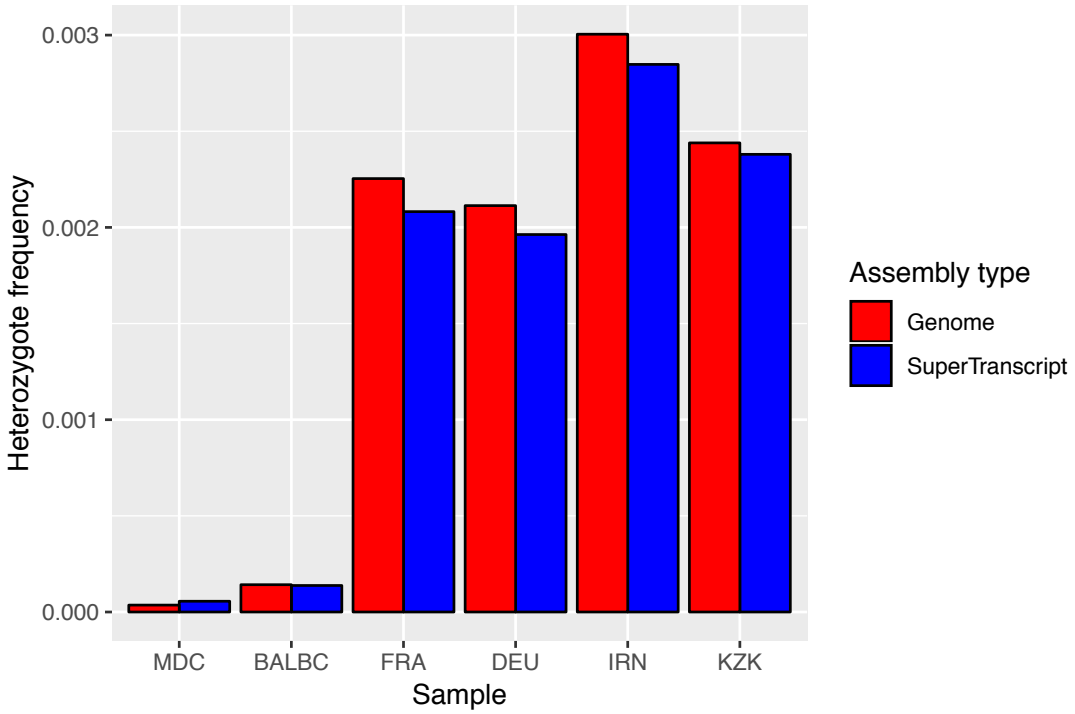

Figure 7. Heterozygosity per base for exonic biallelic SNV variants called relative to the *Mus* reference genome using the GATK best practices pipeline versus those obtained from supertranscripts constructed from TRINITY assemblies using scripts provided with the TRINITY v. 2.6.5 package. See **Methods** for details concerning how exonic coordinates were obtained for SuperTranscripts.

actually higher in ST-overlapping regions (Supplementary Figure 15). 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 16). The driving factor appears to be FN at heterozygous sites; as already noted, FN are dominated by cases where ST are not assembled (Table 3). An additional contributing factor may be erroneous indel polymorphism genotypes in ST. Given that indel genotypes are harder to call correctly than SNVs, a standard approach is to filter out indel sites. Because we perform the same filtering when estimating heterozygosity, we

effectively filter out many true heterozygous SNV positions, further reducing heterozygosity estimates in ST. Paradoxically, FP genotype calls may actually mask the magnitude of this bias.

Gene and isoform expression

The compositional deviations of de novo transcriptome assemblies relative to the true set of expressed transcripts identified by mapping reads to a reference annotation results have a net effect on expression estimates. Whether in terms of expected counts or TPM, expression estimates derived from the contigs that are the best BLAT hit relative to the reference transcripts are only weakly correlated with map-to-reference estimates for those transcripts, and show positive bias for lowly expressed reference transcripts (Supplementary Figure 17ab, Table 4). The practice of using single contigs as proxies for gene-level expression is problematic, as contig-based estimates are also poor estimators of reference gene expression. Since individual contigs typically represent a fraction of the mappable sequence of a set of transcripts belonging to a gene, one expects smaller counts on contigs relative to the full set of mappable positions encompassed by transcripts produced by a gene locus. Given this fact, it is surprising that, for lowly expressed genes, their associated best-hit contigs on average have larger counts (Supplementary Figure 17c). It is also surprising that TPM, a metric that should account for effective length differences, shows a similar pattern of bias (Supplementary Figure 17d). 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 8, Table 4). There are negative and positive biases in de novo-based estimates for expected counts and TPM, respectively, particularly at lower expression levels (Figure 8ab). That bias and deviations from the map-to-reference estimates are partially reduced for length-scaled TPM (Figure 8c). This

Table 4. Correlation between expression of the best BLAT hit assembly contig and its associated reference transcript, and between gene-level expression estimates derived from mapping to a reference transcriptome and summing across de novo transcriptome assembly contigs linked to reference genes via annotation. Correlations are calculated for entries where TPM > 0 for either the map-to-reference, map-to-denovo assembly, or both approaches.

Sample	Expected count		TPM		Length-scaled TPM	
	$\rho_{besthit}$	ρ_{summed}	$\rho_{besthit}$	ρ_{summed}	$\rho_{besthit}$	ρ_{summed}
MDC						
Trinity	**0.129**	**0.838**	0.155	0.795	**0.130**	**0.853**
Shannon	-0.143	0.704	-0.138	0.674	-0.144	0.628
BinPacker	0.112	0.822	**0.158**	**0.810**	0.111	0.824
BALB/c						
Trinity	**0.270**	**0.849**	**0.253**	**0.787**	**0.270**	**0.869**
Shannon	-0.102	0.813	-0.135	0.766	-0.105	0.819
BinPacker	0.221	0.819	0.207	0.772	0.221	0.835
Massif Central						
Trinity	**0.241**	0.829	**0.229**	0.797	**0.241**	**0.850**
Shannon	0.125	**0.833**	0.052	**0.807**	0.124	**0.850**
BinPacker	0.178	0.808	0.181	0.799	0.178	0.815
Germany						
Trinity	**0.240**	0.828	**0.227**	0.793	**0.240**	**0.848**
Shannon	0.129	**0.831**	0.057	**0.807**	0.128	0.847
BinPacker	0.193	0.806	0.195	0.799	0.192	0.811
Iran						
Trinity	**0.225**	**0.831**	**0.209**	0.796	**0.225**	**0.853**
Shannon	0.129	0.830	0.056	**0.800**	0.129	0.846
BinPacker	0.171	0.803	0.172	0.793	0.171	0.808
Kazakhstan						
Trinity	**0.256**	0.830	**0.245**	0.794	**0.256**	**0.852**
Shannon	0.133	**0.834**	0.060	**0.804**	0.132	0.851
BinPacker	0.190	0.804	0.187	0.792	0.190	0.816

indicates that effective length (EL) 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. As demonstrated earlier, high gene coverage is most frequently obtained for highly expressed genes, but transcriptome assembly-derived expression estimates can still deviate substantially from the map-to-reference gold standard at high coverage (Figure 9a). Such deviation is well explained by the disparity in EL, with estimates showing the least deviation where EL values are identical or nearly so (Figure 9b), EL disparities persist at high expression levels, even if those disparities are not as variable as at lower expression levels (Figure 9c). Length-scaled TPMs clearly correct

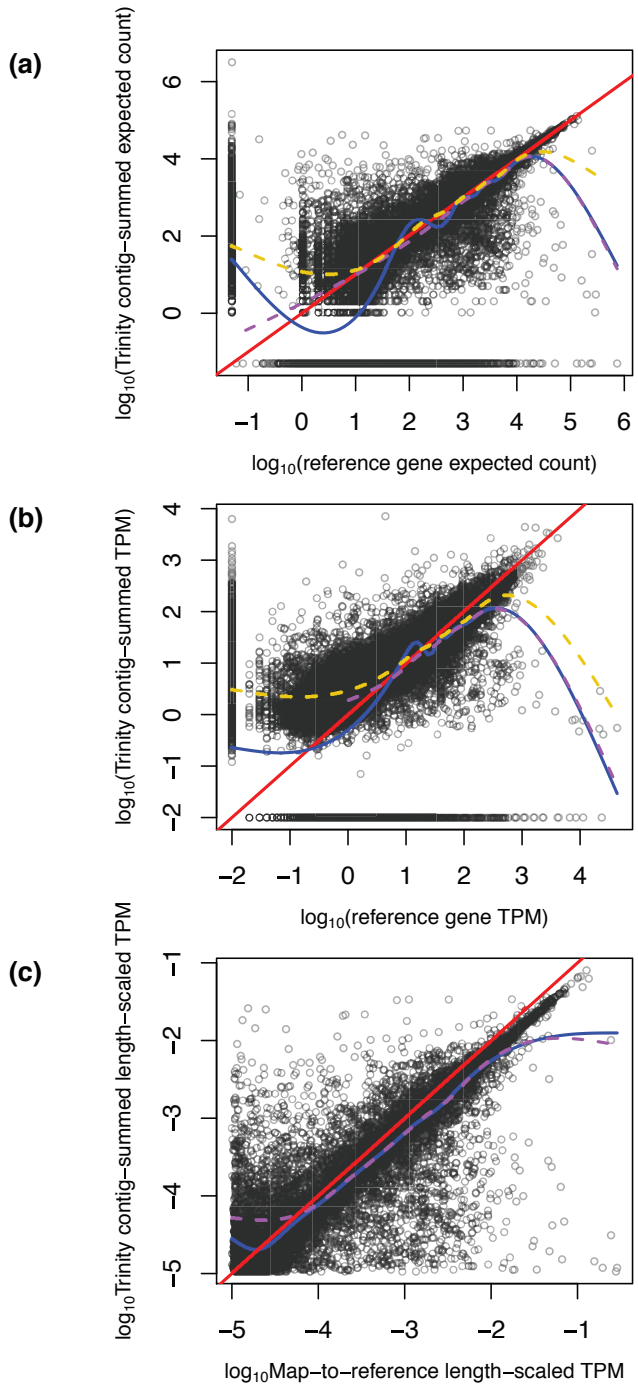

Figure 8. For the FRA data set, comparison of gene-level expression summed across BLAT alignments of TRINITY contigs to reference transcripts (grouped by gene), to reference gene expression in terms of **(a)** expected counts, **(b)** TPM, and **(c)** length-scaled TPM. The red line represents unity, the blue line is the smoothed fit to all of the data, and the pink dashed line is the smoothed line to all reference transcripts with TPM ≥ 1. The gold line is the smoothed line for all TRINITY expression estimates with TPM > 0, i.e. excluding false negatives. There is no comparable line in panel **c**, because it is impossible to rescale the effective length for TRINITY entries with no estimated length.

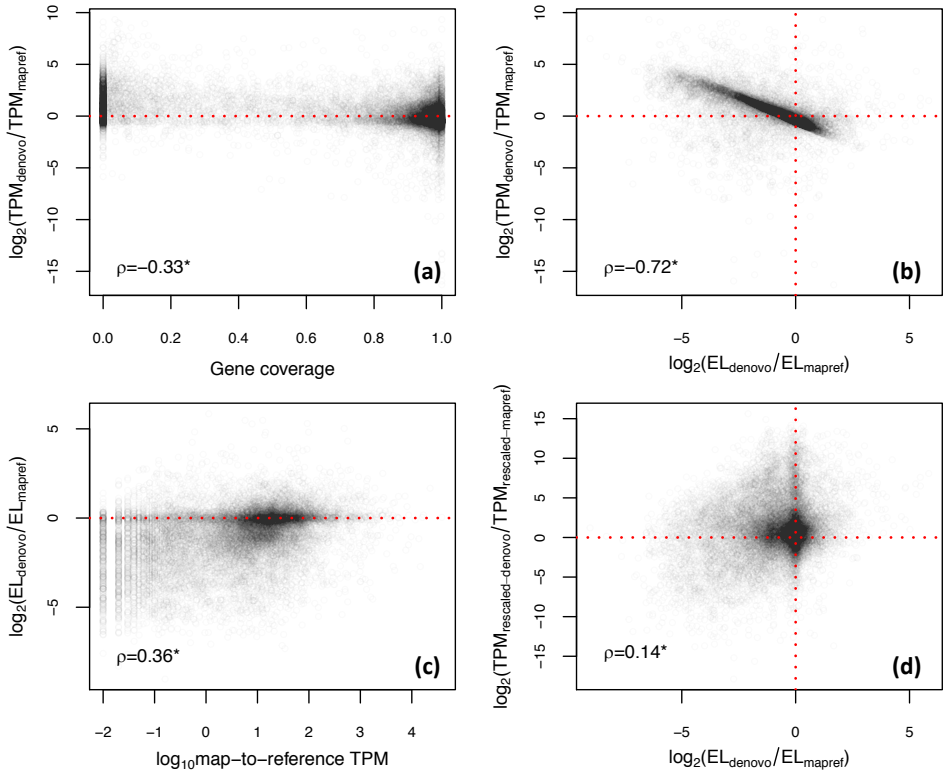

Figure 9. For the FRA data set, for genes with non-zero expression in both transcriptome assembly and map-to-reference, **(a)** difference between TRINITY and map-to-reference TPM visualized as log-fold change, as a function of gene coverage by the assembly, **(b)** TPM log-fold change in TPM as a function of log-fold change in the estimates of effective length (EL), **(c)** log-fold change in EL as a function of map-to-reference TPM, and **(d)** log-fold change in length-scaled TPM values as a function of log-fold change in EL.

for bias that would otherwise be produced by differences in EL (Figure 9d). Nevertheless, negative bias at lower expression levels and large deviations from map-to-reference expression estimates are ubiquitous.

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. Log-fold change of transcriptome assembly TPM vs. map-to-reference TPM at the gene level (LFC_{METHOD}) showed only moderate pairwise correlations between any two of the four wild *Mus* pools (Spearman's ρ: TRINITY, 0.49 – 0.56; SHANNON, 0.50-0.56; BINPACKER, 0.47-0.57), indicating that gene-level features, influence the magnitude of bias (Supplementary Figure 18a). Despite these broad correlations, there was substantial variation in LFC_{METHOD} between samples. The maximum difference in LFC_{METHOD} among the six possible pairwise comparisons between wild samples was ≥ 2 in 46.6 – 50.2 of expressed genes, depending upon the assembler. This magnitude difference corresponds to the threshold researchers often filter on in determining whether a difference is biologically meaningful, but in our case the difference is a statistical artefact of the assembly process. This difference was often far larger (Supplementary Figure 18b-d), indicating that biases may vary enough among samples to distort differential expression inferences. The gene-wise heterogeneity of bias manifests itself clearly when comparing expression differences between our samples: correlations are relatively weak between $\log_2$(transcriptome assembly $TPM_{sample1}$/ transcriptome assemblyTPM$_{sample2}$) and $\log_2$(map-to-reference $TPM_{sample1}$/ map-to-reference $TPM_{sample2}$): Spearman's ρ: TRINITY, 0.21-0.31; SHANNON, 0.22-0.33; BINPACKER, 0.18-0.27 (Supplementary Figure 19).

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.

While some genes not assembled by one algorithm are by another, dozens of genes will consistently fail to get assembled. While this seems like a modest fraction of genes, this occurs only in the idealized case where one can perform nucleotide-level searches against a transcript database from a closely related organism, or in our case, the same organism. Using a common approach that filters out lowly expressed contigs, and only retains those with a hit to a protein database, we found that hundreds of protein coding genes (from 1-2%) would not be recovered by any assembly method, and some assemblers missed several hundred to over 1,000 genes. Thus, while almost all expressed genes may be in the assembly somewhere, annotation challenges amplify missingness. Clearly, combining outputs from multiple assembly algorithms—a non-trivial challenge in its own right—will not prevent missing a sizeable fraction of genes. The efficiency of annotation will be further exacerbated by mis-assembly errors, such as those revealed by indel errors within SuperTranscripts. As *Mus* proteins are well represented in UNIREF90, our finding is overly optimistic relative to the typical predicament molecular ecologists find themselves in. Annotation problems should be greatest for more rapidly evolving genes as they will be more divergent from currently archived protein and nucleotide sequences. Previous studies have confirmed that increasing divergence to annotation databases increases annotation errors (Hornett & Wheat, 2012; Ockendon et al., 2016; Vijay et al., 2013), and these issues should be more precisely considered with respect to both taxon divergence and the evolutionary rate of genes within pathways of interest.

The difficulty in assembling genes that show less conservation is due, in part, to their being lowly expressed. Whether due to the difficulty assembling contigs from few reads, or

polymorphism in more rapidly evolving genes, the fact remains that genes tied to phenotypes of interest may be missing from transcriptome assemblies. To the extent that expression level drives missingness, conserved, lowly expressed genes such as transcription factors that produce large phenotypic effects may also be missing from transcriptome assemblies more often than expected by chance. Molecular ecologists must not confuse expression level with the capacity to drive phenotypic evolution and should find means to estimate the probability that genes of interest are likely to be recovered in a de novo transcriptome assembly.

While it is unsurprising that lowly expressed genes might be harder to assemble, the correlation between expression level and conservation is relatively weak, and in many cases highly expressed genes were missing from annotation and expression-filtered assemblies. Thus, factors correlated with conservation, such as polymorphism, may influence the probability an expressed gene is recovered. Future investigation aimed at quantifying their relative effects could shed important insights relevant to researchers interested in particular genes. Despite our incomplete understanding of all the factors that may lead to genes being missing from an assembly, the superficially good news is that missing genes do not represent a biased subset enriched for particular GO categories. The bad news is that, beyond the reduced probability of assembly for lowly expressed genes, it is not possible to predict in advance and with reasonable certainty that all genes relevant to a particular study being proposed will be recovered.

The nucleotide composition of contigs sheds light on how genes may wind up missing from a de novo transcriptome assembly. We found a relatively high frequency of non-coding bases in assembly contigs, which by definition, places constraints on the utility of standard protein database searches. To the extent that an underlying CDS is divided into a series of contigs, power to discover a protein homolog will be reduced; for contigs dominated by non-

coding bases containing only short strings of CDS, successful ORF prediction and annotation is unlikely.

We expect that missingness both actual and annotation-based, combined with the high frequency of sequence reads dominated by UTR bases, will also reduce power for differential expression studies in a previously unappreciated way. Practically speaking, one throws away all reads that map to unannotated contigs in addition to unmapped reads. Rules of thumb for sequencing depth sufficient for differential expression analyses are largely derived from map-to-reference studies, and do not consider this additional source of annotation-conditional filtering. Loss of coverage depth is most problematic for evolutionary studies of non-model organisms, given that they frequently employ small numbers of biological replicates (Todd, Black, & Gemmell, 2016) which places an additional constraint upon statistical power. We suggest that researchers employing transcriptome assemblies determine the number of reads per sample that remain after those mapping to unannotated contigs have been discarded and consider additional sequencing if read counts far well below desired targets. A second possibility is to "rescue" unannotated contigs, by finding ways to cluster them with annotated contigs. Standard clustering approaches such as CD-HIT will not perform well in this context, as they operate on sequence similarity. By definition, a contig derived mostly from UTR will share little sequence with an annotated contig comprised mostly of coding sequence. Underlying this type of phenomenon is the fact that assembly graphs will often get broken into pieces deriving from different functional compartments. For example, a short repeat in a UTR can split a transcript into an exclusively UTR contig, and a contig comprised of a few UTR bases plus the down or upstream CDS. Recently developed methods that leverage information from pseudo-alignments of paired-end reads that map to different contigs may help mitigate the loss of unannotated contigs, e.g.

GROUPER (Malik, Almodaresi, & Patro, 2018). Integration of this type of approach with annotation-based clustering is a promising avenue that requires further investigation.

In contrast to missingness, two kinds of artefactual redundancy create complications for downstream analyses. First, for any given genomic region from which an actual transcript is generated, transcriptome assembly algorithms will call multiple isoforms that differ only by the proportion of retained intron sequence. This adds an additional obstacle to performing isoform-level expression analyses beyond those imposed by the incomplete and fragmented recovery of transcripts, as counts will be allocated incorrectly across multiple spurious splice variants when they should be aggregated to one transcript, potentially obscuring true underlying inter-isoform variations in expression level. A second form of redundancy occurs in that haplotypes regularly get called as isoforms, and, at least for the commonly used Trinity SuperTranscript-based pipeline, with those haplotypes often being produced by different DeBruijn graphs that are supposed to represent different genes. This added source of false multiplicity makes robust inference of isoform-level expression even more challenging, although gene-level analyses might not be as seriously impacted, assuming that clustering approaches such as GROUPER yielded robust groupings of contigs.

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-based expression estimates derived from transcriptome assemblies relative to a map-to-reference based approach, and there is positive bias at lower expression levels. Furthermore, the practice of selecting a single best hit contig to represent a gene seems ill-advised given that correlations between contig and map-to-reference gene-level expression are poor, and subject not only to positive bias at lower expression levels, but negative bias at higher 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. At the gene-level, length-scaled TPM largely eliminates bias, and more faithfully recapitulates map-to-reference based expression estimates, even if substantial noise persists. The dimensionless scaling of this metric precludes its use in standard count-based differential expression analysis pipelines. Thus, we suggest a valuable avenue of research would be the development of pipelines that can take length scaled TPMs as input. Outside of the realm of formal tests of differential expression, we suggest that length-scaled TPMs 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 length-scaled TPM appears to substantially outperform TPM in this regard.

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. We find at least two reasons why this counter-argument is problematic. First, the observed noise and bias are the outcome of multiple processes, including incomplete coverage, missingness and the stochastic nature of read sampling (particularly at lower expression levels), such that the gene-wise mappability of any one sample to an assembled transcriptome may vary substantially. Consistent with this logic, we found that, on a per-gene basis, there could be large differences between our wild *Mus* samples with respect to the bias manifest in expression estimates derived from de novo transcriptome assemblies. In fact, the proportion of genes where the maximum difference between any pair of samples in log-fold change of transcriptome

assembly TPM vs. map-to-reference TPM was ≥ 2 approached 50%. In other words, half of genes showed a difference in bias that was at least as large as a value of log-fold change that biologists would consider meaningful when comparing conditions in an analysis of differential expression. Second, and more generally, most methods for inferring differential expression rely on estimating a mean-variance relationship. The elevated noise across the range of detectable expression will make fitting this relationship harder, with potentially complex effects on the inference of differential expression. In summary, pipelines that are best able to recapitulate map-to-reference estimates will be those most likely to produce robust gene-level differential expression calls, and isoform-level analyses should be avoided.

By far the most unexpected findings from our study were the extremely high genotyping error rates derived from transcriptome assemblies, specifically ST representations of TRNITY assemblies. False negatives were due, in part, to the absence of an ST aligned to a putatively true MR polymorphic site, which is not unexpected given the known fragmented nature of many *de novo* transcriptome assemblies. As our mapping rates of ST to genomes was extremely high, such FN are due to either the failure to assembly 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 ST, leading to downwardly biased estimates of heterozygosity in *de novo* assemblies. FN may have also been caused by the slightly reduced read depth at homologous positions in read alignments to ST. 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, may 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 FP sites. In the absence of a genome sequence, it is hard to correct FN biases for any particular transcriptome assembly. In contrast, sophisticated filtering methods might permit the learning of sequence features that produce FPs, and to derive metrics that could mitigate FP.

Even in the face of high FP heterozygous genotype calls, FNs led to an overall negative bias in estimates of heterozygosity estimates derived from ST relative to the gold standard MR approach, most prominently in the wild *Mus* samples. This finding strongly suggests that the underlying site-frequency spectrum derived from an ST 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 are problematic. 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 statistical method with biology and will lead to spurious inferences. Even when applied solely to diploid transcriptome assemblies, as haplotypes will get assembled into different ST, 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.

While it is possible that a combination of algorithmic and annotation-based clustering techniques can be employed to better collapse assembly redundancy, further research on this topic is necessary. Whether or not more robust clustering techniques ameliorate bias, naïve application of genotyping to ST derived from Trinity component "genes" will lead to high FP and FN rates. Finally, 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.

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 is 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). 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 two 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. However analytically challenging, there is great potential for 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. This work was conducted on the traditional territory of the Wampanoag and Massachusett peoples.

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

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

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

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

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

Harr, B., Karakoc, E., Neme, R., Teschke, M., Pfeifle, C., Pezer, Ž., … Tautz, D. (2016). Genomic resources for wild populations of the house mouse, Mus musculus and its close relative Mus spretus. *Scientific Data*, *3*, 160075. doi:10.1038/sdata.2016.75

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

Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*, *12*(4), 357–360. doi:10.1038/nmeth.3317

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, B., Fillmore, N., Bai, Y., Collins, M., Thomson, J. A., Stewart, R., & Dewey, C. N. (2014). Evaluation of de novo transcriptome assemblies from RNA-Seq data. *Genome Biology*, *15*(12). doi:10.1186/s13059-014-0553-5

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, *25*(16), 2078–2079. doi:10.1093/bioinformatics/btp352;

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. (2014). On the optimal trimming of high-throughput mRNA sequence data. *Bioinformatics and Computational Biology*, *5*, 13. doi:10.3389/fgene.2014.00013

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), 1297–303. 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

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

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. doi:10.1093/bioinformatics/btq033

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

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210–3212. doi:10.1093/bioinformatics/btv351

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

Smith-Unna, R., Boursnell, C., Patro, R., Hibberd, J. M., & Kelly, S. (2016). TransRate: reference-free quality assessment of de novo transcriptome assemblies. *Genome Research*, *26*(8), 1134–1144. doi:10.1101/gr.196469.115

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

Song, L., & Florea, L. (2015). Rcorrector: efficient and accurate error correction for Illumina RNA-seq reads. *GigaScience*, *4*, 48. doi:10.1186/s13742-015-0089-y

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., Wegrzyn, 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

Wu, T. D., & Watanabe, C. K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, *21*(9), 1859–1875. doi:10.1093/bioinformatics/bti310