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

Adam H. Freedman*, Michele Clamp, Timothy B. Sackton

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

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

This manuscript includes:

-- Main Text

-- Figures 1 to 5

-- Table 1

-- Supplemental information including Supplementary Methods S1 – S5, Supplementary Results S1 – S3, Supplementary Tables S1 – S8, Supplementary Figures S1 – S19

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 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, length-rescaled CPM, 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

37 (Berens, Hunt, & Toth, 2015; Pankey, Minin, Imholte, Suchard, & Oakley, 2014; Velotta et al.,
38 2017), speciation (K. Li et al., 2016), and local adaptation (De Wit & Palumbi, 2013).

39 All studies using *de novo* transcriptome assemblies rely on two implicit assumptions: that
40 the assembled transcriptome represents an unbiased, if incomplete, representation of the true
41 underlying expressed transcriptome, and that the expression estimates from the assembly are
42 good, if noisy, approximations of the relative abundance of expressed transcripts. Indirect
43 evidence suggests that violations of these assumptions may be frequent. For example,
44 comparisons of different assembly algorithms indicate that there is a degree of complementarity
45 in missingness—one assembler will recover a gene that another cannot, motivating methods to
46 combine assemblies, e.g. the Oyster River Protocol (MacManes, 2018). In other words, particular
47 assemblers may not be able to assemble particular kinds of genes. A separate issue, related to the
48 robustness of expression estimates, is that highly fragmented assemblies comprised of hundreds
49 of thousands of relatively short contigs are pervasive in the literature, presumably representing
50 transcript fragments, haplotypes erroneously designated as separate transcripts, or both.
51 Intuitively, the challenge of mapping reads to fragmented gene models should diminish the
52 robustness of expression estimates. Unknown is whether the compositional aspects of
53 transcriptome assemblies negatively impact SNP calls, allele frequency estimates, and other
54 population genetic parameters.

55 Below, we evaluate the assumptions inherent to *de novo* transcriptome assembly, discuss
56 how their violation might distort evolutionary inferences, identify particular features of
57 transcriptome assemblies that should be considered during the study design phase of a project,
58 and demonstrate a method for correcting bias in expression estimates. While phylogenomic
59 studies can discard the majority of contigs using aggressive filtering schemes, these are less

60 appropriate for studies that seek to understand patterns of genetic variation or gene expression
61 across populations or closely related species. Therefore, we focus on methodological
62 considerations for this class of investigations.
63

64 **2. MATERIALS AND METHODS**

65 **2.1 RNA-seq data**

66 Data sets include the *Mus musculus* (C57BL/6J) dendritic cell data used in the original TRINITY
67 paper (MDC), a pool of six whole brains from albino inbred *Mus* (BALB/c), and 8-sample pools
68 of whole brain from wild *M. musculus domesticus* from Massif Central, France (FRA), Iran
69 (IRN), Kazakhstan (KZK), and Germany (DEU). Brain tissue is both transcriptionally complex
70 and expected to express a large proportion of an organism's overall transcriptional profile. It
71 thus represents a challenging test case for *de novo* transcriptome assembly. Illumina reads were
72 pre-processed using a standard pipeline (Methods S1).
73

74 **2.2 *De novo* transcriptome assembly**

75 We built *de novo* transcriptome assemblies with Trinity v. 2.1.1 (Grabherr et al., 2011b) using
76 default settings. For the *Mus* data sets, to evaluate the robustness of some key findings to the
77 algorithm employed, we also built assemblies using two alternate assemblers: Shannon v. 0.0.2
78 (Kannan, Hui, Mazooji, Pachter, & Tse, 2016) and BinPacker v. 1.0 (Liu et al., 2016). Where
79 strand-specific reads were available, we ran Trinity and BinPacker with the appropriate strand-
80 specific options. Because the strand-specific mode of Shannon is in experimental phase (Kannan,
81 personal communication), we treated reads as un-stranded in all Shannon assemblies. We
82 assessed assembly quality using standard, commonly used metrics for a) coverage of conserved

83 orthologs to estimate recovery of conserved genes that should be expressed in many tissues, b)
84 the proportion of reads that are aligned to the assembly in a concordant fashion, i.e. whether the
85 assembly is well-supported by the reads and c) assembly contiguity metrics, namely N_{50},which
86 captures the extent to which the assembly is fragmented (Methods S2). Downstream analyses
87 compare outputs of *de novo* transcriptome assemblies to results obtained from mapping RNA-seq
88 reads directly to the reference genome and its associated annotations of exons, transcripts and
89 genes. (Methods S3).
90
91 **2.3 Assembly and Read Functional Composition**
92 For each contig within a *de novo* assembly, we quantified the proportion of unique assembly
93 nucleotides that were CDS, UTR, intronic, start or stop codon, noncoding elements (including
94 pseudogenic transcripts of genes with "polymorphic_pseudogene" gene biotype), introns of non-
95 coding transcripts, and intergenic/unnanotated sequence. We took a similar approach with
96 sequence reads to determine if the functional composition of RNA-seq reads differed from that
97 of the transcriptome assemblies. For details, see Methods S4.
98
99 **2.4 Expression Estimation**
100 For nine combinations of assembly method (Trinity, Shannon, BinPacker) and *Mus* sample
101 (MDC, BALB/c, Massif Central), we generated expression estimates with RSEM v. 1.2.22 (B. Li
102 & Dewey, 2011a) and Bowtie2, v. 2.2.4 (Langmead & Salzberg, 2012), using RSEM default
103 parameters (which internally sets Bowtie2 arguments). For libraries that were strand-specific, we
104 set the appropriate flags so strand-specificity could be leveraged during the read-mapping stage.
105 We also estimated "true" (reference-based) expression by mapping reads directly onto transcripts

106 extracted from the reference genome, using the same RSEM/Bowtie2 pipeline but with the
107 ENSEMBL GTF annotation.

108 To be able to evaluate the robustness of expression patterns derived from *de novo*
109 transcriptome assemblies, one must connect assembly contigs to the underlying reference
110 isoforms from which they originated. Thus, after extracting reference transcripts from the
111 genome annotation with RSEM v. 1.2.22 (B. Li & Dewey, 2011b) , we mapped transcriptome
112 assembly contigs to these transcripts using BLAT v. 3.1.6 (Kent, 2002). We required at least
113 95% identity in the alignment to consider a hit valid, but did not set a minimum alignment length
114 in the query (contig). Given the minimal evolutionary distance between query and target
115 sequences, and because we required a minimum of 95% identity for alignments, we did not
116 penalize coverage for mismatches. For each *de novo* assembly contig with > 1 BLAT hit, we
117 identified the true transcript of origin as that for which the alignment produced the maximal
118 coverage. In this way, each contig was uniquely assigned to a single reference isoform.

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

129 assembly gene-level expression was then calculated by building an RSEM index with the –
130 *transcript-to-gene-map* flag, such that contig expression values were aggregated to gene labels
131 specified in a table of gene-contig relationships. Genes identified as expressed in the map-to-
132 reference approach but that did not have any associated *de novo* transcriptome assembly contigs
133 were considered missing.

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

139 We also evaluated how a typical filtering approach influenced the robustness of
140 expression estimates at the gene level. We employed three filtering steps. First, we discarded all
141 lowly expressed transcriptome assembly contigs with TPM ≤ 1. Second, we used Transdecoder
142 v. 3.0.0 (https://github.com/TransDecoder/TransDecoder/wiki) to predict contig peptide
143 sequences, performed BLASTP hit against Uniref90 (downloaded 2017-04-05), and only
144 retained contigs with an alignment e-value $\leq 1 \times 10^{-5}$. Finally, we restricted comparisons
145 between *de novo* transcriptome assemblies and our map-to-reference benchmark to protein
146 coding genes for which at least one isoform was ≥ 200bp; because the default minimum contig
147 size for Trinity is 200bp, the minimum length requirement prevents artefactual downward bias in
148 expression estimates.

149 The fragmentation of transcript and gene models in assemblies will affect effective
150 length, where the effective length of a transcript equals the length of the transcript minus the
151 mean sequence read fragment length, plus one. In other words, effective length estimates the

152 total number of possible start sites for read alignments to a sequence given the fragment length
153 distribution of reads. For example, if a 1kb transcript is broken into two 500bp contigs in a
154 transcriptome assembly, and the mean read fragment length is 200bp, the effective length of the
155 original transcript equals 801, whereas the summed effective lengths of the two transcriptome
156 assembly contigs equals 602. The observed length also contributes to abundance estimates, e.g. if
157 half of an expressed transcript is not assembled, no counts can be tallied for the missing
158 transcript interval. Because both effective and observed length directly impacts abundance
159 measurements, we assessed whether correcting for differences in both length metrics between
160 reference transcripts and *de novo* assembled contigs would improve concordance in expression
161 estimates. We did this by rescaling CPM as follows:

162 $$CPM_{rescaled} = CPM * \frac{map\text{-}to\text{-}ref\ effective\ gene\ length}{denovo\ effective\ gene\ length} \times \frac{denovo\ gene\ length}{map\text{-}to\text{-}ref\ gene\ length}.$$

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

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

170 $$log_2\left(\frac{sample\ 1\ CPM_{denovo}}{sample\ 1\ CPM_{map\text{-}to\text{-}ref}}\right) - log_2\left(\frac{sample\ 2\ CPM_{denovo}}{sample\ 2\ CPM_{map\text{-}to\text{-}ref}}\right),$$

171 which can be re-arranged as:

172 $$log_2\left(\frac{sample\ 1\ CPM_{denovo}}{sample\ 2\ CPM_{denovo}}\right) - log_2\left(\frac{sample\ 1\ CPM_{map\text{-}to\text{-}ref}}{sample\ 2\ CPM_{map\text{-}to\text{-}ref}}\right),$$

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

175 $$2^{EXP}\left[abs\left(log_2\left(\frac{sample\ 1\ CPM_{denovo}}{sample\ 2\ CPM_{denovo}}\right) - log_2\left(\frac{sample\ 1\ CPM_{map-to-ref}}{sample\ 2\ CPM_{map-to-ref}}\right)\right)\right]$$

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

179 Two standard ways of filtering a *de novo* transcriptome assembly are to discard contigs
180 that are lowly expressed or that cannot be annotated. For the latter, the usual practice is to find
181 open reading frames, and use BLASTX or BLASTP against a peptide database. We investigated
182 how such filtering impacts the robustness of gene-level expression estimates by recalculating
183 expression estimates after removing contigs with TPM < 1, contigs for which no ORF was
184 detected, and contigs for which a translated ORF did not return a BLASTP hit to Uniref90. We
185 retained the original transcript and gene assignments for assembly contigs derived from BLAT,
186 such that protein hits are simply treated as filter. This approach is optimistic given that it
187 circumvents possible annotation errors, and also due to the fact that *Mus* proteins are prevalent in
188 Uniprot. Because filtering will shorten the total length of sequence that is used to infer
189 expression estimates, this may lead to noisier expression estimates. Thus, we quantified the
190 reduction in coverage of reference annotations induced by our filtering scheme and related it to
191 expression estimates. For details on coverage calculations and filtering, see Methods S5.
192

193 **2.5 Evaluation of variant calls**

194 We compared the map-to-reference genotypes—those derived from mapping reads directly to the
195 genome—to those derived from *de novo* Trinity assemblies. We called variants off of the

196 reference genome using the GATK v. 4 (McKenna et al., 2010), following the Broad Institute's
197 best practices for RNA-seq data
198 (https://software.broadinstitute.org/gatk/documentation/article.php?id=3891; last updated 2017-
199 01-21). For *de novo* Trinity assemblies, we first constructed SuperTranscripts, (Davidson,
200 Hawkins, & Oshlack, 2017), which attempt to collapse overlapping contigs within Trinity
201 "genes", and then called variants with the GATK pipeline, based upon python scripts provided
202 with Trinity v. 2.6.5 (see 16ploidy_run_variant_calling_gatk4.py, in the genotying directory of
203 the below-referenced Github repository, which as an example, has the –ploidy argument set to 16
204 as would be used for the wild *Mus* pools) . SuperTranscript alignments to the genome were used
205 to define unique polymorphic site ids that included cases where (a) both methods called
206 genotypes, (b) only map-to-reference called a genotype (including sites with and without valid
207 SuperTranscript alignments), or (c) a genotype was only called off a SuperTranscript. We
208 recorded cases where multiple SuperTranscripts were aligned to the same genomic position,
209 taking the union of all reported alleles as the ST genotype for that position. We restricted our
210 comparison of the two methods by filtering out map-to-reference sites that were indels or
211 contained > 2 alleles. Using genotype calls based upon the reference annotation as the gold
212 standard, we calculated 11 different metrics, defined as follows:

213 - **Error rate:** # positions with a genotype call in either map-to-reference or map-to-
214 SuperTranscripts, where SuperTranscript alleles do not match map-to-reference alleles ÷
215 total # positions with a genotype call in either map-to-reference or map-to-
216 SuperTrasncripts; includes positions with no overlapping ST.
217 - **Error-Snv2Indel**: # biallelic map-to-reference SNVs called as an indel in map-to-
218 SuperTranscripts ÷ # bi-allelic map-to-reference SNVs.

219 • **FN**: # map-to-reference polymorphic sites with an invariant or missing genotype in map-
220 to-SuperTranscripts ÷ # map-to-reference polymorphic sites; includes sites with no
221 overlapping SuperTranscripts.
222 • **FN-FractionNoAlignment**: # FN with no aligned SuperTranscript ÷ # FN
223 • **FN-withAlignment**: Same as **FN**, but restricted to sites with ≥1 overlapping
224 SuperTranscript alignments
225 • **FP**: # map-to-reference invariant sites with a polymorphic map-to-SuperTranscript
226 genotype ÷ # map-to-reference invariant sites.
227 • **FP-indel**: # map-to-SuperTranscript indel polymorphic sites at map-to-reference
228 invariant sites ÷ # false positive sites.
229 • **FP-hetSNV**: # false positive sites with map-to-reference biallelic SNV genotypes ÷ #
230 false positive sites.
231 • **Recall**: # polymorphic map-to-reference genotypes with matching map-to-
232 SuperTranscripts genotypes ÷ # polymorphic map-to-reference genotypes
233 • **Recall-withAlignment**: same as **Recall** restricted to sites with ≥1 overlapping
234 SuperTranscript alignments
235 • **Precision:** # polymorphic map-to-SuperTranscript genotypes that match map-to-
236 reference genotypes ÷ # polymorphic map-to-SuperTranscript genotypes.
237 In order to determine if SuperTranscripts with either high false negative, false positive or overall
238 genotyping error rates were enriched for particular functions, we performed statistical
239 enrichment tests for Biological Process, Molecular Function and Cellular Component Gene
240 Ontology categories using Panther (Mi, Muruganujan, Casagrande, & Thomas, 2013). For these
241 tests, we assigned SuperTranscripts to gene annotations based upon sequence alignment, and

242 calculated mean SuperTranscript error rate per gene as scores for statistical testing. See Methods
243 S5 for more details.
244
245 **2.6 Code Availability**
246 Software command lines, scripts and accessory files used to implement analyses described above
247 are available at https://github.com/harvardinformatics/TranscriptomeAssemblyEvaluation.
248
249 **3. RESULTS**
250 **3.1 "High quality" assemblies are highly fragmented**
251 Our assemblies were highly fragmented, consisting of anywhere from 70,375 – 582,133 contigs,
252 with median contig sizes ranging from just a few hundred bases to slightly over 1kb, and N_{50}
253 ranging from 870 to 3021 bp (Table S2). This is typical of many published transcriptome
254 assemblies (Bentley, Haas, Tedeschi, & Berry, 2017; Faherty, Villanueva-Cañas, Blanco, Albà,
255 & Yoder, 2018; Nespolo et al., 2018) .Yet, typically employed measures of accuracy and
256 completeness indicate that all assemblies were of reasonably high quality (Results S1, Tables S3
257 and S4). The large number of short contigs strongly suggest that *de novo* transcriptome
258 assemblies do not typically assembly full length transcripts.
259
260 **3.2 Assembly functional composition is biased**
261 Large fractions of *de novo* transcriptome contigs are dominated by intronic, UTR, and intergenic
262 sequence (Figs S1 and S2). The excess of intronic sequence in contigs is not reflected in the
263 composition of reads (Figs. S1 and S2). The prevalence of intronic and intergenic sequence in
264 assembly contigs (Fig. S3) indicates that a substantial proportion of assembled bases are derived

265 from some combination of transcriptional read-through across exon junctions, assembly of low-
266 abundance pre-mRNA, and low levels of spurious genome-wide transcription.
267
268 **3.3 Frequent genotyping errors**
269 In principle, redundancy through assembly of haplotypes as distinct contigs should lead to
270 erroneous genotype calls and biased estimates of allele frequency spectra. For example, two
271 contigs can actually be haplotypes of a single locus, such that a truly heterozygous position is
272 instead represented as two separate contigs, one homozygous for the reference allele and one
273 homozygous for the alternative allele. Two lines of evidence indicate such redundancy is present
274 in transcriptome assemblies. First, BLAT all-by-all alignments indicate that many contigs are
275 nearly identical (Figs. S4-S6), in some cases differing only by single nucleotide variants (e.g.
276 Figure S4 B). Surprisingly, Trinity assemblies will produce such highly similar contigs from
277 different assembly graphs which are purported to assemble isoforms of a gene or sets of closely
278 related paralogs (Fig. S6). Second, there is a high frequency of cases where all nucleotides of a
279 transcript from a single-isoform *Mus* gene are covered by ≥ 1 *de novo* assembly contig (Fig. S7).
280 The latter are primarily due to slight variations in the assembly of intronic bases, and secondarily
281 due to differences in CDS content (Results S2, Fig. S8).
282 The opposite problem can happen when paralogous loci are collapsed, leading to a
283 heterozygous genotype call where there would otherwise be two invariant sites, or perhaps a
284 mixture of alleles due to, for example, collapsing of an invariant and a heterozygous position. To
285 assess the degree to which these and other errors affect transcriptome assemblies, we compared
286 genotype calls and derived diversity metrics between the map-to-reference and map-to-
287 transcriptome approaches using 11 different metrics.

288 The frequency of errors, in which the allele set recovered by the two methods differed or
289 in which a genomic position with a called genotype contained no overlapping SuperTranscript
290 nucleotides, was high, ranging from 30-37% for the wild samples, and 49-83% for the inbred
291 samples (Table 1). Error rates varied among SuperTranscripts within a *de novo* transcriptome
292 assembly, with some SuperTranscripts demonstrating extremely high error rates. While the
293 median error rate per SuperTranscript was 0.14% per base or less, maximum per SuperTranscript
294 rates typically exceeded 10% per base (Table S5). For five of six samples, the median number of
295 errors per SuperTranscript was 2, but, depending upon the sample, as many as 56-187 errors
296 could occur within a single SuperTranscript (Table S5). Thus, within any *de novo* transcriptome
297 assembly, a small number of contigs may contribute disproportionately to noise that may obscure
298 the true patterns of polymorphism. On a per-sample basis, errors show a modest yet statistically
299 significant bias towards the distal end of a SuperTranscript for all samples except MDC (sample-
300 level Wilcoxon 1-sided tests, $H_0 = 0.5$, $p < 2 \times 10^{-16}$; Fig. S9).
301 The false negative rate (FN; 23-71%) was greater than the false positive rate (FP; 15-
302 87%) in four of six samples. The latter includes all classes of errors including indel
303 polymorphism. Anywhere from 42-53% of FN were due to there being no SuperTranscript
304 aligned at a genomic position where a map-to-reference heterozygote genotype call was made
305 (Table 1 FN-FractionNoAlignment). Given that rates of mapping SuperTranscripts to the
306 genome were extremely high (>99%), these FN instances are primarily due to a failure to
307 assemble transcripts overlapping polymorphic genomic positions. Even at sites called as
308 heterozygotes in map-to-reference analyses that were assembled into SuperTranscripts, false
309 negative rates were still relatively high, from 13-15% and 37-59% in the wild and inbred
310 samples, respectively (Table 1). False negative rates demonstrated significant negative

311 correlations with SuperTranscript length in five of six samples, but those correlations were weak
312 (Spearman's ρ, -0.24- -0.10), indicating a minimal increase in false negatives in shorter
313 SuperTranscripts.

314 Because read depth influences the ability of tools such as GATK to call genotypes, we
315 suspected false negatives might be due to reduced read depth at homologous genomic positions
316 called as heterozygotes with the map-to-reference approach. While map-to-reference and map-
317 to-SuperTranscript coverage were largely colinear for concordant and false positive map-to-
318 SuperTranscript genotypes (Figs. S10 and S11), false negative sites with overlapping
319 SuperTranscripts displayed small reductions in SuperTranscript read depth relative to map-to-
320 reference at sites where SuperTranscripts mapped to the genome (Fig. S12). Given the
321 magnitude of coverage reduction is small, we suspect it may not be the only factor contributing
322 to false negatives at sites where a SuperTranscript covers a polymorphic genomic position.

323 The false positive (FP) rate ranged from 15-87%. Incorrect heterozygous and indel
324 genotype calls at truly invariant sites are expected when paralogs are collapsed, and both occur
325 with considerable frequency in our samples. From 6-22% of FPs are erroneous indel genotypes,
326 8-62% of FPs are erroneous heterozygous genotype calls, and the remainder are other classes of
327 SuperTranscript polymorphisms, including both indel and single nucleotide variants, with > 2
328 alleles. The inference of frequent paralog collapses is also supported by the high rate at which
329 true SNV polymorphisms are called as indels in SuperTranscripts (Error-Snv2Indel, 16-31%).
330 False positive rate was significantly associated with SuperTranscript length in all samples, but
331 showed contrasting signals between inbred (MDC, Balb/c) and wild (FRA, DEU, IRN, KZK)
332 samples: correlations for inbred samples were moderate and negative (Spearman's ρ, -0.27 to -
333 0.49), but weak and positive for wild samples (Spearman's ρ, 0.01 to 0.05).

334 When genotypes were called in SuperTranscripts, the frequency of calls that were correct
335 (i.e. precision) ranged from 66-83% for the wild samples, but precision for the inbred samples,
336 particularly MDC, was substantially lower. Recall ranged from 76-81% in the wild samples; as
337 with precision, recall for both inbred samples was substantially lower.
338 We examined false negative, false positive, and overall error rates more closely,
339 evaluating whether particular GO terms were enriched in SuperTranscripts with higher error
340 rates. Panther statistical enrichment tests indicated that there were no categories enriched for
341 false negatives, and only a few for false positives in the FRA sample, indicating there is no
342 broader pattern of enrichment for these types of errors (Table S6). With respect to overall error
343 rates, FRA SuperTranscripts with high error rates were enriched for MHC class 1 protein
344 complex genes, which contains hundreds of paralogs, but these terms were not enriched in any
345 other sample (Table S6). Strong signals for gene families were not detected, with the only terms
346 that consistently showed up as being enriched related to ribosomal components (Table S6).
347 Overall, the frequency of genotyping errors seems likely to distort evolutionary inferences and
348 reduce power across a wide range of statistical methods, particularly those aimed at detecting
349 signals of selection on particular polymorphisms, allele-phenotype and allele frequency-
350 environment associations.
351
352 **3.4 Negatively biased heterozygosity estimates**
353 In all but the inbred strain samples (MDC and BALB/c), we observed reduced heterozygosity in
354 the exonic regions of SuperTranscripts relative to mapping to those regions in the reference
355 genome (Fig. 1). A substantial contributor to this bias is the fact that, as noted above,
356 approximately half of map-to-reference heterozygous positions are not overlapped by any

357 SuperTranscript. In addition, this bias does not result from SuperTranscripts being preferentially
358 assembled in regions of lower heterozygosity, given that, for the wild samples, map-to-reference
359 heterozygosity is actually higher in SuperTranscript-overlapping regions (Fig. S13).
360 Furthermore, the bias cannot be due to some unique feature of genotype calls in exons, as the
361 same pattern holds when considering all genotype calls (Fig. S14). Another possible explanation
362 may be lower power to detect low frequency alternative alleles in map-to-SuperTranscript
363 analyses. However, in our five samples consisting of pooled individuals, minor allele frequency
364 at map-to-reference heterozygous sites did not vary between correct SuperTranscript
365 heterozygous calls, false negatives, and other genotyping errors (Fig. S15). Downward bias in
366 heterozygosity estimates occurs despite a relatively high prevalence of false positive
367 heterozygotes. Because this bias shifts all estimates towards zero, it truncates the observed range
368 of variation, making it more difficult to make meaningful inferences from comparisons of
369 diversity across genes, populations, and species.
370

371 **3.5 Single contigs are poor expression estimators**

372 Expression estimates derived from the contigs that are the best BLAT hit relative to the reference
373 transcripts are only weakly correlated (Spearman's ρ: CPM: 0.070-0.387; TPM: 0.045-0.392;
374 rescaled CPM: 0.083-0.415), with expression for those transcripts, and show positive bias for
375 lowly expressed reference transcripts (Fig. S16 A and B, Table S7). Correlations between best-
376 hit contig expression and expression of the associated reference gene, after accounting for
377 effective length differences, are slightly better but still relatively weak (Spearman's ρ: rescaled
378 CPM: 0.103-0.657, Table S6). Overall, the use of single contigs as proxies for gene-level
379 expression appears problematic.

380

381 **3.6 Correctable gene-level expression bias**

382 Expression estimates derived from *de novo* transcriptome assemblies are far more robust at the
383 gene level, when one is able to group contigs into genes via annotation (Fig. 2, Table S8). It is
384 reassuring that the correlation between the methods is quite strong, although there is sizeable
385 positive bias at lower expression levels for TPM, which is substantially reduced for CPM.
386 Furthermore, for many genes there are substantial deviations from estimates obtained with
387 mapping to the reference, even at high expression levels (Fig. 2, Fig. S7). For estimates
388 generated with RSEM, bias and deviations from the map-to-reference estimates are reduced
389 when one rescales CPMs to account for differences in effective length estimates (Fig. 2C). This
390 indicates that effective length differences—that is, the difference in the number of start sites for
391 read alignments, given a particular sequenced fragment size, between a full-length transcript, and
392 a set of transcript sub-sequences represented by contigs—are part of the problem, and in
393 particular, that effective length is not properly estimated on the fragmented gene models that are
394 typical in *de novo* transcriptome assemblies. Biases in gene expression are reduced as the
395 difference in effective length between map-to-reference and map-to-transcriptome-assembly
396 decreases (Fig. 3). Even so, disparities in effective length estimates between map-to-reference
397 and map-to-*de novo*-transcriptome-assembly approaches can persist even at high reference gene
398 expression, with many *de novo* transcriptome-based estimates of gene expression showing a
399 fourfold or greater deviation relative to the map-to-reference benchmark (Fig. 2, Fig. S18).
400 Correlations between reference gene expression and gene-level summing across assembly
401 contigs obtained with Kallisto were very similar, if slightly inferior, to those obtained with
402 RSEM (Results S3, Table S8, Fig. S17).

403

404 **3.7 Expression bias and DE testing**

405 Biases in expression estimates derived from transcriptome assemblies might not have large
406 effects on downstream differential expression analyses if biases are consistent across samples.
407 While the pooled *Mus* data sets we used did not permit us to look at inter-individual variation
408 amongst individual libraries used to build a single assembly, we looked across the wild *Mus*
409 datasets to see if biases in gene-level expression were concordant, i.e. were biases for a particular
410 gene always of similar magnitude and in the same direction. We focus on results for Trinity
411 assemblies of wild *Mus* samples, as Trinity consistently showed stronger correlations with map-
412 to-reference based estimates, such that our findings are conservative with respect to assemblers
413 showing weaker correlations. Reassuringly, log-fold change of transcriptome assembly vs. map-
414 to-reference expression estimators (LFC_{METHOD}) showed broad patterns of moderate concordance
415 at the gene level indicating that gene-level features influence the magnitude of bias (Fig. S19).
416 Inter-sample correlations were, for CPM, 0.625-0.727; TPM, 0.790-0.830; and rescaled CPM,
417 0.605-0.671. These correlations are weaker for CPM and rescaled CPM, in part, because biases
418 are reduced overall, compressing LFC_{METHOD} toward zero. Despite these broad correlations, a
419 number of genes may show excessive bias. We measured this as a bias effect size that indicates
420 how much greater a difference in expression would be detected using a *de novo* transcriptome
421 approach, relative to mapping to reference annotation. Comparing the FRA and DEU samples,
422 for CPM, TPM, and rescaled CPM, 5.24%, 5.16% and 3.35% of genes showed bias effect sizes ≥
423 5, with more than 2% of genes showing an order of magnitude of difference between methods of
424 ≥ 10 for CPM and TPM (Fig. 4).

425

426 **3.8 Effects of filtering on expression estimates**

427 Filtering out lowly expressed *de novo* transcriptome assembly contigs as well as those contigs
428 without a BLASTP hit to Uniref90 led to a noticeable improvement in the correlation between
429 map-to-transriptome assembly and map-to-reference abundance estimates for protein coding
430 genes (Fig. 2 D, E, and F). Post-filtering, bias and noise in TPM was dramatically reduced.
431 However, despite a correlation nearly identical to that for CPM and rescaled CPM, there is
432 consistent positive bias across the range of map-to-reference expression (Fig. 2E). Nevertheless,
433 the improvement across all three metrics relative to unfiltered assemblies belies a tradeoff that is
434 hidden from view from researchers without access to an annotated genome: filtering leads to a
435 loss of many expressed protein coding genes that would be recovered without filtering. For the
436 FRA assembly, 21.0% of such genes are filtered out, although only 10.6% of genes are lost if
437 one excludes from consideration lowly expressed genes with map-to-reference TPM < 1 (Fig. 5).
438 Depending upon the sample-assembler combination, 9.1-32.5% of expressed protein coding
439 genes can be lost due to filtering, and for genes with map-to-reference TPM ≥ 1, 6.0-27.1%
440 (Table S7). In nearly all sample-assembler combinations, the proportion of genes lost due to
441 filtering far exceeded the proportion that were missing due to a failure to assemble those genes,
442 with most exceptions occurring for BinPacker assemblies, which performed poorly with respect
443 to recovery of expressed genes (Table S7).

444

445 **4. DISCUSSION**

446 *De novo* transcriptome assemblies have been an important entry point for functional and
447 population genomics in non-model organisms, but the two key assumptions underlying their
448 application—that the composition of *de novo* transcriptomes is unbiased if incomplete, and that

449 expression estimates derived from them are unbiased if noisy—are consistently violated, albeit in
450 unexpected ways. Below, we explain how particular compositional aspects of *de novo*
451 transcriptome assemblies may impact estimates of gene and isoform expression, genotype calls,
452 and genetic diversity—estimates which are frequently used for making evolutionary inferences
453 for non-model organisms.

454 SNP calls and allele frequencies inferred from them and that are derived from
455 transcriptome assemblies have been used in a variety of contexts, including estimation genetic
456 diversity in evolutionary analyses (Romiguier et al., 2014), population structure (Tepolt &
457 Palumbi, 2015), reconstruction of demographic history and gene flow (De Wit & Palumbi, 2013;
458 Le Gac et al., 2016; Roda, Mendes, Hahn, & Hopkins, 2017), estimation of dN/dS (Cokus,
459 Gugger, & Sork, 2015), F_{ST} outlier analysis (Bankers et al., 2017) , and tests for allele-specific
460 expression (Singh et al., 2018). Given this wide application of *de novo* transcriptome assembly
461 genotypes in studies of non-model organisms, by far the most unexpected findings from our
462 study were the extremely high genotyping error rates derived from transcriptome assemblies,
463 specifically SuperTranscript representations of TRNITY assemblies. False negatives were due, in
464 part, to the absence of a SuperTranscript aligned to a putatively true polymorphic site recovered
465 with a map-to-reference approach. False negatives are due to either the failure to assemble any
466 contigs at all for a gene (i.e. missing genes), or the incomplete coverage of genes not extending
467 over some fraction of polymorphic sites. That genotyping errors were biased towards the distal
468 end of contigs suggest that assembly graphs are getting broken at hard to assemble features
469 containing low complexity and/or increased polymorphism. When such graph-splitting occurs,
470 the intervening polymorphic sites will, by definition, not be covered by a SuperTranscript,
471 leading to downwardly biased estimates of heterozygosity in *de novo* assemblies. False negatives

472 may have also been caused by the slightly reduced read depth at homologous positions in read
473 alignments to SuperTranscripts. That the reduction in read depth, was, on average, small
474 suggests that other factors may play a role. It may be that genotype inference methods such at the
475 GATK Haplotype Caller that leverage haplotypic information, have poorer performance when
476 iterating over alignments to relatively short assembly contigs. Another possibility is that
477 assembly errors may generate patterns that get flagged by standard genotyping filters. In
478 contrast, false positives appear to be driven by paralog collapses given the bias towards
479 heterozygous false positive sites.

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

491 The negative bias we observed in heterozygosity estimates derived from *de novo*
492 transcriptome assemblies indicates that the underlying site-frequency spectrum derived from an
493 SuperTranscript assembly will be biased in ways that distort other population genetic summary
494 statistics typically reported in molecular ecological investigations that utilize genotypes called

495 from transcriptome assemblies. More broadly, the error rates we document suggest that several
496 lines of transcriptome-assembly based research should be approached with caution. First,
497 erroneous heterozygous calls, when added to the redundancy problems described above, will
498 substantially reduce power and elevate false positive rates in allele-specific expression analyses.
499 Second, levels of genome-wide patterns of genetic diversity derived from transcriptome
500 assemblies will likely be negatively biased, at least for diploid species. The distinction
501 concerning ploidy is an important one because, in the absence of a genome or karyotypic
502 information, there can be no a priori expectation of how events such as whole-genome or even
503 chromosome-level duplication will impact genotyping inference. Thus, both integration of
504 diversity metrics across transcriptome assemblies with differing, unknown levels of ploidy, and
505 comparing transcriptome and genome assembly-derived metrics will confound method with
506 biology and can lead to spurious inferences. Even when applied solely to diploid transcriptome
507 assemblies, as haplotypes will get assembled into different SuperTranscripts, the expectation is
508 that negative bias will be greater for more heterozygous genomes, such that estimated levels of
509 diversity across assemblies will artificially appear more similar than they are in actuality,
510 impeding researchers' ability to detect underlying patterns of variation. Overall, the distortion of
511 allele frequencies derived from transcriptome assemblies strongly suggests that they will inject
512 bias into analyses of population structure, demographic inference, and efforts to detect positive
513 selection. Whether or not improved clustering of contigs can mitigate this distortion remains an
514 open question.

515 Our results clearly indicate that isoform-level analyses and using individual contigs as
516 proxies for isoform or gene-level expression, should be avoided. Even when annotation-based
517 groupings of contigs are performed in an optimal setting, such as our analyses of *Mus*

518 transcriptomes, there is substantial noise in gene-level expression estimates derived from
519 transcriptome assemblies relative to a map-to-reference based approach, and there is positive
520 bias at lower expression levels. Incorrect estimation of effective length metrics used in RNA
521 abundance calculations play a substantial role in the observed patterns of bias. Nevertheless, it
522 may be surprising to researchers that biases persist for standard length-adjusted metrics such as
523 TPM. CPM has less bias compared to TPM. Bias and noise can be reduced by rescaling CPM to
524 account for differences in effective and observed length, such that the rescaled CPM more
525 faithfully recapitulates map-to-reference based expression estimates, even if substantial noise
526 persists. Adjusting expression estimates to account for effective and observed length differences
527 can be beneficially used in two different applications. First, one could conduct differential
528 expression estimates across different transcriptome assemblies when it is necessary to generate
529 different assemblies for different conditions, such as divergent species or sub-species. Because
530 one of the first steps differential expression tools execute is to normalize by library size, one can
531 simply apply the correction directly on the counts themselves. CPM and count scale are directly
532 proportional, with the former effectively being the latter multiplied by a sample-specific
533 constant, such that all benefits we report on rescaled CPM will apply to similarly transformed
534 counts. Rescaling counts across a set of species would require treating one species as the
535 "reference", and rescaling counts for other species relative to it. Outside of the realm of formal
536 tests of differential expression, we suggest that rescaled CPM could be useful in phylogenetic
537 comparative analyses of expression at the gene level. One of the big challenges of these methods
538 is accounting for length differences between transcript/contig sets belonging to orthologous
539 genes, and the adjusted CPM appears to substantially outperform other metrics in this regard.
540 While our method for correcting for effective length differences substantially reduced the

541 frequency of bias outliers, it does not eliminate them entirely. Further investigation of other
542 sources of bias is necessary in order to devise methods to mitigate their impacts on expression
543 analyses. Even so, we should note that our adjustment method is relatively simple, and that it
544 may be possible to adopt more sophisticated approaches that leverage such information as
545 sequence composition and complexity for additional improvements.

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

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

562

563

564 **5. CONCLUSIONS**

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

587 analyses modes that suffer from fundamental biases, mitigating biases where possible for others,
588 and, always, placing appropriate caveats given the biases that will inevitably remain.
589

590 **ACKNOWLEDGMENTS**

591 The computations in this paper were run on the Odyssey cluster supported by the FAS Division
592 of Science, Research Computing Group at Harvard University. We thank members of the FAS
593 Informatics Group for discussion. We also thank Nadia Davidson, Alicia Oshlack, Robin
594 Hopkins, and Mara Laslo for feedback on an earlier version of this manuscript. This work was
595 conducted on the traditional territory of the Wampanoag and Massachusett peoples.
596

597 **DATA AVAILABILITY**

598 Upon acceptance for publication, all transcriptome assemblies, SuperTranscript assemblies, and
599 vcf genotype files will be made available on Dryad.
600

601 **AUTHOR CONTRIBUTIONS**

602 The research project was designed by A. H. F., M. C., and T. B. S. Analyses were conducted by
603 A. H. F. with input from M.C. and T. B. S. A. H. F. wrote the paper with T. B. S. All authors
604 read and approved the final manuscript.
605

606 **REFERENCES**

607 Austin, C. M., Tan, M. H., Harrisson, K. A., Lee, Y. P., Croft, L. J., Sunnucks, P., … Gan, H. M.
608 (2017). De novo genome assembly and annotation of Australia's largest freshwater fish,

609 the Murray cod (Maccullochella peelii), from Illumina and Nanopore sequencing read.
610 *GigaScience*, *6*(8), 1–6. doi: 10.1093/gigascience/gix063
611 Bankers, L., Fields, P., McElroy, K. E., Boore, J. L., Logsdon, J. M., & Neiman, M. (2017). Genomic
612 evidence for population-specific responses to co-evolving parasites in a New Zealand
613 freshwater snail. *Molecular Ecology*, *26*(14), 3663–3675. doi: 10.1111/mec.14146
614 Bentley, B. P., Haas, B. J., Tedeschi, J. N., & Berry, O. (2017). Loggerhead sea turtle embryos
615 (Caretta caretta) regulate expression of stress response and developmental genes when
616 exposed to a biologically realistic heat stress. *Molecular Ecology*, *26*(11), 2978–2992.
617 doi: 10.1111/mec.14087
618 Berens, A. J., Hunt, J. H., & Toth, A. L. (2015). Comparative transcriptomics of convergent
619 evolution: different genes but conserved pathways underlie caste phenotypes across
620 lineages of eusocial insects. *Molecular Biology and Evolution*, *32*(3), 690–703. doi:
621 10.1093/molbev/msu330
622 Birol, I., Jackman, S. D., Nielsen, C. B., Qian, J. Q., Varhol, R., Stazyk, G., … Jones, S. J. M. (2009).
623 De novo transcriptome assembly with ABySS. *Bioinformatics*, *25*(21), 2872–2877. doi:
624 10.1093/bioinformatics/btp367
625 Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq
626 quantification. *Nature Biotechnology*, *34*(5), 525–527. doi: 10.1038/nbt.3519
627 Cokus, S. J., Gugger, P. F., & Sork, V. L. (2015). Evolutionary insights from de novo transcriptome
628 assembly and SNP discovery in California white oaks. *BMC Genomics*, *16*, 552. doi:
629 10.1186/s12864-015-1761-4

630 Davidson, N. M., Hawkins, A. D. K., & Oshlack, A. (2017). SuperTranscripts: a data driven
631 reference for analysis and visualisation of transcriptomes. *Genome Biology*, *18*(1). doi:
632 10.1186/s13059-017-1284-1
633 De Wit, P., & Palumbi, S. R. (2013). Transcriptome-wide polymorphisms of red abalone (Haliotis
634 rufescens) reveal patterns of gene flow and local adaptation. *Molecular Ecology*, *22*(11),
635 2884–2897. doi: 10.1111/mec.12081
636 Faherty, S. L., Villanueva-Cañas, J. L., Blanco, M. B., Albà, M. M., & Yoder, A. D. (2018).
637 Transcriptomics in the wild: Hibernation physiology in free-ranging dwarf lemurs.
638 *Molecular Ecology*, *27*(3), 709–722. doi: 10.1111/mec.14483
639 Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., … Regev, A.
640 (2011a). Full-length transcriptome assembly from RNA-Seq data without a reference
641 genome. *Nature Biotechnology*, *29*(7), 644–652. doi: 10.1038/nbt.1883
642 Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., … Regev, A.
643 (2011b). Full-length transcriptome assembly from RNA-Seq data without a reference
644 genome. *Nature Biotechnology*, *29*(7), 644–652. doi: 10.1038/nbt.1883
645 Kannan, S., Hui, J., Mazooji, K., Pachter, L., & Tse, D. (2016). Shannon: An Information-Optimal
646 de Novo RNA-Seq Assembler. *BioRxiv*, 039230. doi: 10.1101/039230
647 Kent, W. J. (2002). BLAT--the BLAST-like alignment tool. *Genome Research*, *12*(4), 656–664. doi:
648 10.1101/gr.229202. Article published online before March 2002
649 Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature*
650 *Methods*, *9*(4), 357–359. doi: 10.1038/nmeth.1923

651 Le Gac, M., Metegnier, G., Chomérat, N., Malestroit, P., Quéré, J., Bouchez, O., … Chapelle, A.
652 (2016). Evolutionary processes and cellular functions underlying divergence in
653 Alexandrium minutum. *Molecular Ecology*, *25*(20), 5129–5143. doi: 10.1111/mec.13815
654 Li, B., & Dewey, C. N. (2011a). RSEM: accurate transcript quantification from RNA-Seq data with
655 or without a reference genome. *BMC Bioinformatics*, *12*, 323. doi: 10.1186/1471-2105-
656 12-323
657 Li, B., & Dewey, C. N. (2011b). RSEM: accurate transcript quantification from RNA-Seq data with
658 or without a reference genome. *BMC Bioinformatics*, *12*, 323. doi: 10.1186/1471-2105-
659 12-323
660 Li, K., Wang, H., Cai, Z., Wang, L., Xu, Q., Lövy, M., … Nevo, E. (2016). Sympatric speciation of
661 spiny mice, *Acomys* , unfolded transcriptomically at Evolution Canyon, Israel.
662 *Proceedings of the National Academy of Sciences*, *113*(29), 8254–8259. doi:
663 10.1073/pnas.1608743113
664 Liu, J., Li, G., Chang, Z., Yu, T., Liu, B., McMullen, R., … Huang, X. (2016). BinPacker: Packing-
665 Based De Novo Transcriptome Assembly from RNA-seq Data. *PLOS Computational*
666 *Biology*, *12*(2), e1004772. doi: 10.1371/journal.pcbi.1004772
667 MacManes, M. D. (2018). The Oyster River Protocol: a multi-assembler and kmer approach for
668 de novo transcriptome assembly. *PeerJ*, *6*. doi: 10.7717/peerj.5428
669 Marcionetti, A., Rossier, V., Bertrand, J. A. M., Litsios, G., & Salamin, N. (2018). First draft
670 genome of an iconic clownfish species (Amphiprion frenatus). *Molecular Ecology*
671 *Resources*, *18*(5), 1092–1101. doi: 10.1111/1755-0998.12772

672 McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … DePristo, M. A.
673 (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-
674 generation DNA sequencing data. *Genome Res*, *20*(9), 1297–1303. doi:
675 10.1101/gr.107524.110
676 Mi, H., Muruganujan, A., Casagrande, J. T., & Thomas, P. D. (2013). Large-scale gene function
677 analysis with the PANTHER classification system. *Nature Protocols*, *8*(8), 1551–1566. doi:
678 10.1038/nprot.2013.092
679 Michael, T. P., Jupe, F., Bemm, F., Motley, S. T., Sandoval, J. P., Lanz, C., … Ecker, J. R. (2018).
680 High contiguity Arabidopsis thaliana genome assembly with a single nanopore flow cell.
681 *Nature Communications*, *9*(1), 541. doi: 10.1038/s41467-018-03016-2
682 Nespolo, R. F., Gaitan-Espitia, J. D., Quintero-Galvis, J. F., Fernandez, F. V., Silva, A. X., Molina,
683 C., … Bozinovic, F. (2018). A functional transcriptomic analysis in the relict marsupial
684 Dromiciops gliroides reveals adaptive regulation of protective functions during
685 hibernation. *Molecular Ecology*, *27*(22), 4489–4500. doi: 10.1111/mec.14876
686 Nowoshilow, S., Schloissnig, S., Fei, J.-F., Dahl, A., Pang, A. W. C., Pippel, M., … Myers, E. W.
687 (2018). The axolotl genome and the evolution of key tissue formation regulators.
688 *Nature*, *554*(7690), 50–55. doi: 10.1038/nature25458
689 Pankey, M. S., Minin, V. N., Imholte, G. C., Suchard, M. A., & Oakley, T. H. (2014). Predictable
690 transcriptome evolution in the convergent and complex bioluminescent organs of squid.
691 *Proceedings of the National Academy of Sciences*, *111*(44), E4736–E4742. doi:
692 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

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

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

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

726 **TABLES**

727 **Table 1**. Selected performance metrics comparing genotypes derived from SuperTranscript
728 collapse of TRINITY de novo transcriptome assemblies with those derived from a map-to-
729 reference genome approach. Samples are *Mus dendritic cell* (MDC), albino inbred mouse
730 (Balb/c), and wild mouse from France (FRA), Germany (DEU), Iran (IRN), and Kazakhstan
731 (KZK). See **MATERIALS AND METHODS** for information on metric calculations.

	Sample					
Metric	**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

732

733

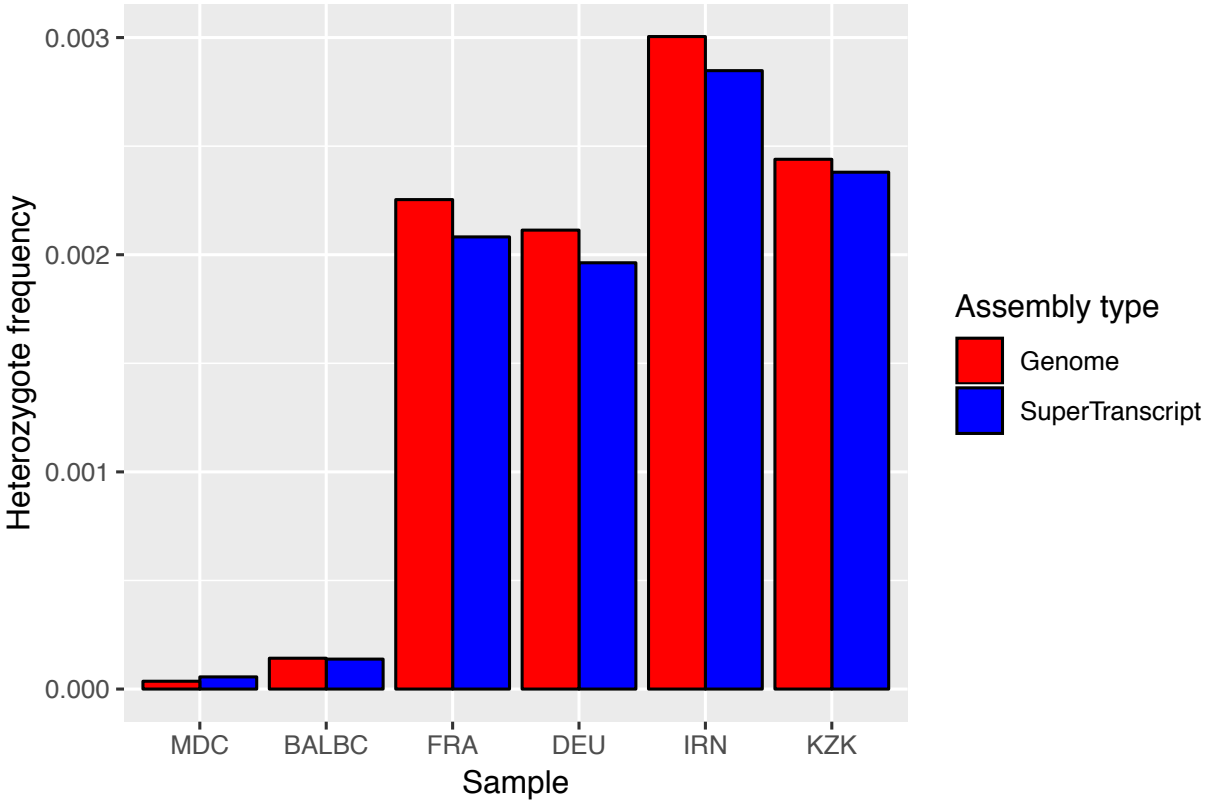

Figure 1. Heterozygosity per base for exonic biallelic SNV variants called relative to the *Mus* reference genome using the GATK best practices pipeline versus those obtained from supertranscripts constructed from TRINITY assemblies using scripts provided with the Trinity v. 2.6.5 package. See **Materials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

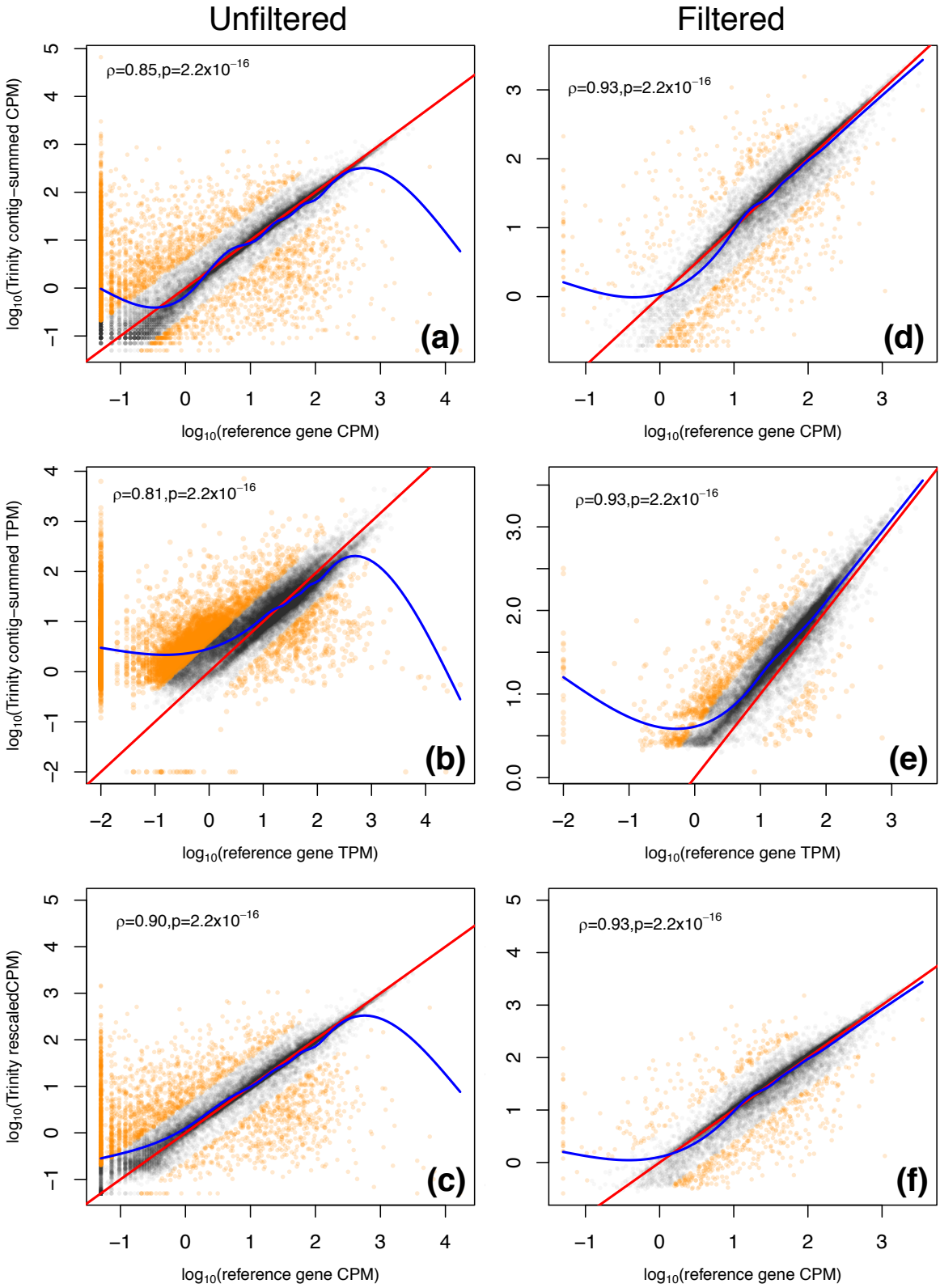

Figure 2. For the FRA data set, comparison of gene-level expression summed across BLAT alignments of TRINITY contigs to reference transcripts (grouped by gene), to reference gene expression in terms of **(a, b)** CPM, **(b, e)** TPM, and **(c, f)** rescaled CPM that accounts for differences in effective length estimates between methods. Left and right columns are for unfiltered and filtered transcriptome assemblies, respectively. 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.

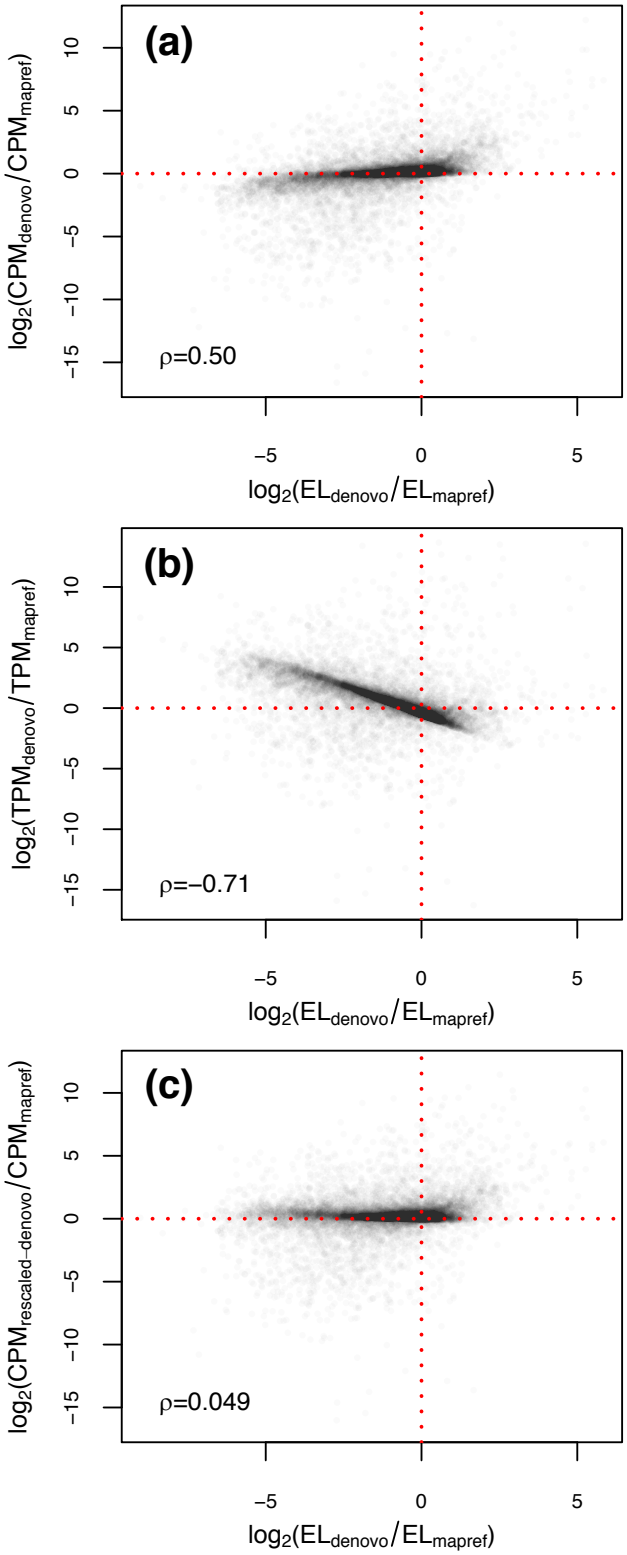

Figure 3. For the FRA data set, for genes with non-zero expression in both transcriptome assembly and map-to-reference, log-fold change between TRINITY and map-to-reference expression estimates as a function of difference in effective length (EL) for **(a)** CPM, **(b)** TPM, and **(c)** rescaled CPM.

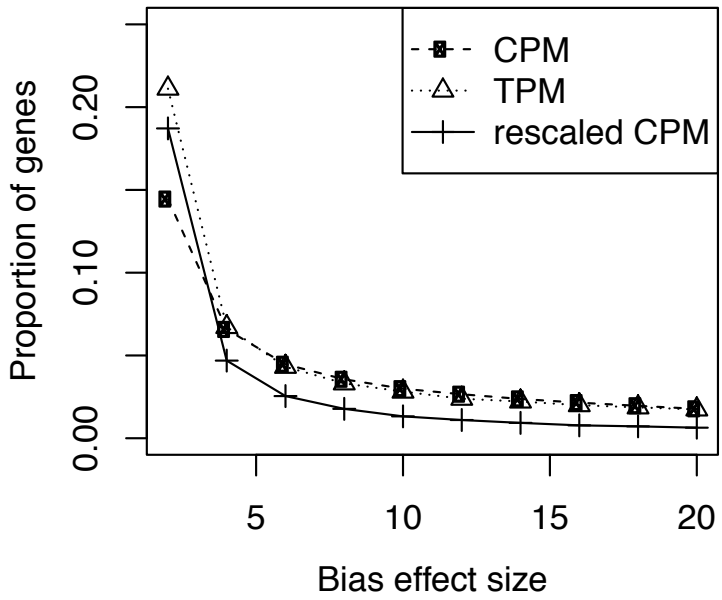

Figure 4. Frequency of bias effect sizes for gene-level expression estimated by CPM, TPM, and rescaled CPM for a TRINITY-based comparison of the FRA and DEU samples.

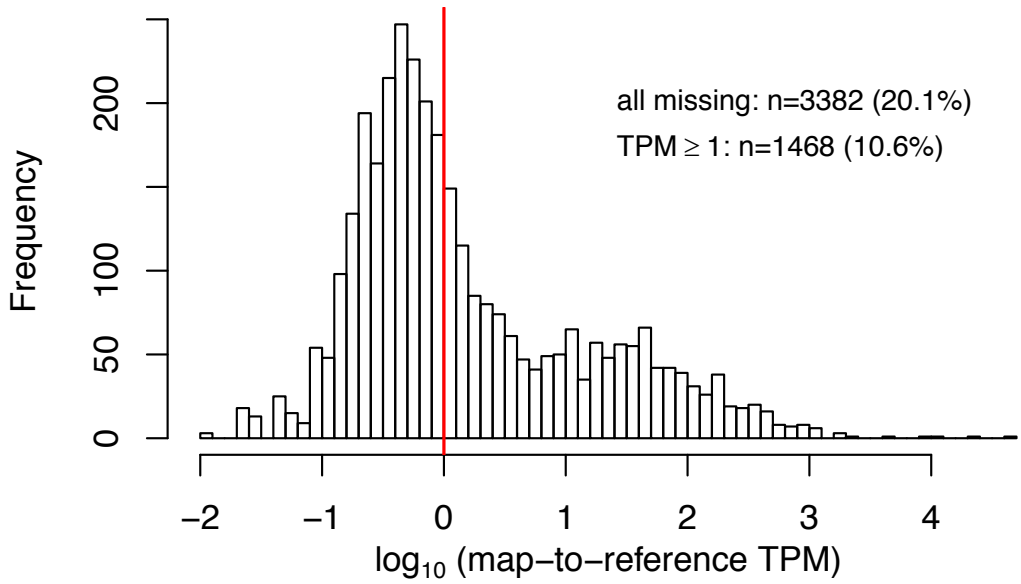

Figure 5. Histogram of map-to-reference TPM for expressed protein-coding genes that were detected in the FRA TRINITY assembly prior to filtering out un-annotated and lowly expressed assembly contigs, but were lost due to filtering.