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

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

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

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

* Corresponding author: adamfreedman@fas.harvard.edu

ABSTRACT

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

31 **1. INTRODUCTION**

32 The development of computational methods for *de novo* transcriptome assembly (Birol et al.,
33 2009; Grabherr et al., 2011a) has empowered a wide array of evolutionary analyses in the
34 absence of a reference genome. Their application has led to significant advances in fields as
35 diverse as phylogenetics (Schwentner, Combosch, Pakes Nelson, & Giribet, 2017), evolutionary
36 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. Comparisons of
44 different assembly algorithms indicate that there is a degree of complementarity in
45 missingness—one assembler will recover a gene that another cannot, motivating methods to
46 combine assemblies, e.g. the Oyster River Protocol (MacManes, 2018). Highly fragmented
47 assemblies comprised of hundreds of thousands of relatively short contigs are pervasive in the
48 literature, presumably representing transcript fragments, haplotypes erroneously designated as
49 separate transcripts, or both. Intuitively, the challenge of mapping reads to fragmented gene
50 models should diminish the robustness of expression estimates. Unknown is whether the
51 compositional aspects of transcriptome assemblies negatively impact SNP calls, allele frequency
52 estimates, and other population genetic parameters.

53 Below, we evaluate the assumptions inherent to *de novo* transcriptome assembly, discuss
54 how their violation might distort evolutionary inferences, identify particular features of
55 transcriptome assemblies that should be considered during the study design phase of a project,
56 and demonstrate a method for correcting bias in expression estimates. While phylogenomic
57 studies can discard the majority of contigs using aggressive filtering schemes, these are less
58 appropriate for studies that seek to understand patterns of genetic variation or gene expression
59 across populations or closely related species. Therefore, we focus on methodological
60 considerations for this class of investigations.
61

62 **2. MATERIALS AND METHODS**

63 **2.1 RNA-seq data**

64 Data sets include the *Mus musculus* (C57BL/6J) dendritic cell data used in the original TRINITY
65 paper (MDC), a pool of six whole brains from albino inbred *Mus* (BALB/c), and 8-sample pools
66 of whole brain from wild *M. musculus domesticus* from Massif Central, France (FRA), Iran
67 (IRN), Kazakhstan (KZK), and Germany (DEU). To assess the generality of particular results
68 with respect to assembly composition, we also generated TRINITY assemblies for pufferfish
69 (*Tetraodon nigroviridis*) whole brain, clawed frog (*Xenopus tropicalis*) kidney, and fly
70 (*Drosophila melanogaster*) heads. SRA accessions are summarized in Supplementary Table 1.
71 Illumina reads were pre-processed using a standard pipeline (Supplementary Methods 1).
72

73 **2.4 *De novo* transcriptome assembly**

74 We built *de novo* transcriptome assemblies with TRINITY v. 2.1.1 (Grabherr et al., 2011b) using
75 default settings. For the *Mus* data sets, to evaluate the robustness of some key findings to the

76 algorithm employed, we also built assemblies using two alternate assemblers: SHANNON v. 0.0.2
77 (Kannan, Hui, Mazooji, Pachter, & Tse, 2016) and BINPACKER v. 1.0 (Liu et al., 2016). Where
78 strand-specific reads were available, we ran TRINITY and BINPACKER with the appropriate
79 strand-specific options. Because the strand-specific mode of SHANNON is in experimental phase
80 (Kannan, personal communication), we treated reads as un-stranded in all SHANNON assemblies.
81 We assessed assembly quality using standard metrics for coverage of conserved orthologs, read
82 mapping statistics, and assembly contiguity (Supplementary Methods 2). Downstream analyses
83 compare outputs of *de novo* transcriptome assemblies to benchmarks consisting of reference
84 annotations and genomes for each species (Supplementary Methods 3).
85
86 **2.5 Assembly and read functional composition**
87 For each contig within a *de novo* assembly, we quantified the proportion of unique assembly
88 nucleotides that were CDS, UTR, intronic, start or stop codon, noncoding elements (including
89 pseudogenic transcripts of genes with "polymorphic_pseudogene" gene biotype), introns of non-
90 coding transcripts, and intergenic/unnanotated sequence. We took a similar approach with
91 sequence reads to determine if the functional composition of RNA-seq reads differed from that
92 of the transcriptome assemblies. For details, see Supplementary Methods 4.
93
94 **2.7 Expression estimation**
95 For nine combinations of assembly method (TRINITY, SHANNNON, BINPACKER) and *Mus* sample
96 (MDC, BALB/c, Massif Central), we generated expression estimates with RSEM v. 1.2.22 (B. Li &
97 Dewey, 2011b) and BOWTIE2, v. 2.2.4 (Langmead & Salzberg, 2012). We also estimated "true"
98 (reference-based) expression by mapping reads directly onto transcripts extracted from the

99 reference genome, using the same RSEM/BOWTIE2 pipeline but with the ENSEMBL GTF
100 annotation.

101 To be able to evaluate the robustness of expression patterns derived from *de novo*
102 transcriptome assemblies, one must connect assembly contigs to the underlying reference
103 isoforms from which they originated. Thus, after extracting reference transcripts from the
104 genome annotation with RSEM v. 1.2.22 (B. Li & Dewey, 2011a) , we mapped transcriptome
105 assembly contigs to these transcripts using BLAT v. 3.1.6 (Kent, 2002). We required at least
106 95% identity in the alignment to consider a hit valid, but did not set a minimum alignment length
107 in the query (contig). Given the minimal evolutionary distance between query and target
108 sequences, and because we required a minimum of 95% identity for alignments, we did not
109 penalize coverage for mismatches. For each *de novo* assembly contig with > 1 BLAT hit, we
110 identified the true transcript of origin as that for which the alignment produced the maximal
111 coverage. In this way, each contig was uniquely assigned to a single reference isoform.

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

122 assembly gene-level expression was then calculated by building an RSEM index with the –
123 *transcript-to-gene-map* flag, such that contig expression values were aggregated to gene labels
124 specified in a table of gene-contig relationships. Genes identified as expressed in the map-to-
125 reference approach but that didn't have any associated *de novo* transcriptome assembly contigs
126 were considered missing.

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

132 We also evaluated how a typical filtering approach influenced the robustness of
133 expression estimates at the gene level. We employed three filtering steps. First, we discarded all
134 lowly expressed transcriptome assembly contigs with TPM ≤ 1. Second, we used TRANSDECODER
135 v. 3.0.0 (https://github.com/TransDecoder/TransDecoder/wiki) to predict contig peptide
136 sequences, performed BLASTP hit against Uniref90 (downloaded 2017-04-05), and only retained
137 contigs with an alignment e-value $\leq 1 \times 10^{-5}$. Finally, we restricted comparisons between *de novo*
138 transcriptome assemblies and our map-to-reference benchmark to protein coding genes for which
139 at least one isoform was ≥ 200bp; because the default minimum contig size for TRINITY is 200bp,
140 the minimum length requirement prevents artefactual downward bias in expression estimates.
141 Because expression estimates derived from TRINITY assemblies were in nearly all cases more
142 consistent with our benchmark data than those derived from SHANNON or BINPACKER, we
143 restricted this analysis to TRINITY assemblies.

144 The fragmentation of transcript and gene models in assemblies will affect effective
145 length, where the effective length of a transcript equals the length of the transcript minus the
146 mean sequence read fragment length, plus one. In other words, effective length estimates the
147 total number of possible start sites for read alignments to a sequence given the fragment length
148 distribution of reads. For example, if a 1kb transcript is broken into two 500bp contigs in a
149 transcriptome assembly, and the mean read fragment length is 200bp, the effective length of the
150 original transcript equals 801, whereas the summed effective lengths of the two transcriptome
151 assembly contigs equals 602. The observed length also contributes to abundance estimates, e.g. if
152 half of an expressed transcript is not assembled, no counts can be tallied for the missing
153 transcript interval. Because both effective and observed length directly impacts abundance
154 measurements, we assessed whether correcting for differences in both length metrics between
155 reference transcripts and *de novo* assembled contigs would improve concordance in expression
156 estimates. We did this by rescaling CPM as follows:

157 $$CPM_{rescaled} = CPM * \frac{map\text{-}to\text{-}ref\ effective\ gene\ length}{denovo\ effective\ gene\ length} \times \frac{denovo\ gene\ length}{map\text{-}to\text{-}ref\ gene\ length}.$$

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

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

165 $$log_2\left(\frac{sample\ 1\ CPM_{denovo}}{sample\ 1\ CPM_{map\text{-}to\text{-}ref}}\right) - log_2\left(\frac{sample\ 2\ CPM_{denovo}}{sample\ 2\ CPM_{map\text{-}to\text{-}ref}}\right),$$

166 which can be re-arranged as:

167 $$log_2\left(\frac{sample\ 1\ CPM_{denovo}}{sample\ 2\ CPM_{denovo}}\right) - log_2\left(\frac{sample\ 1\ CPM_{map-to-ref}}{sample\ 2\ CPM_{map-to-ref}}\right),$$

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

170 $$2^{EXP}\left[abs\left(log_2\left(\frac{sample\ 1\ CPM_{denovo}}{sample\ 2\ CPM_{denovo}}\right) - log_2\left(\frac{sample\ 1\ CPM_{map-to-ref}}{sample\ 2\ CPM_{map-to-ref}}\right)\right)\right]$$

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

174 Two standard ways of filtering a *de novo* transcriptome assembly are to discard contigs
175 that are lowly expressed or that cannot be annotated. For the latter, the usual practice is to find
176 open reading frames, and use BLASTX or BLASTP against a peptide database. We investigated
177 how such filtering impacts the robustness of gene-level expression estimates by recalculating
178 expression estimates after removing contigs with TPM < 1, contigs for which no ORF was
179 detected, and contigs for which a translated ORF did not return a BLASTP hit to UNIREF90. We
180 retained the original transcript and gene assignments for assembly contigs derived from BLAT,
181 such that protein hits are simply treated as filter. This approach is optimistic given that it
182 circumvents possible annotation errors, and also due to the fact that *Mus* proteins are prevalent in
183 UNIPROT. Because filtering will shorten the total length of sequence that is used to infer
184 expression estimates, this may lead to noisier expression estimates. Thus, we quantified the
185 reduction in coverage of reference annotations induced by our filtering scheme and related it to
186 expression estimates. For details on coverage calculations and filtering, see Supplementary
187 Methods 5.
188

189 **2.9 Evaluation of variant calls from transcriptome assemblies**

190 For these analyses, we focused on the *Mus* samples, and compared the map-to-reference
191 genotypes to those derived from *de novo* TRINITY assemblies. We called variants off of the
192 reference genome using the GATK v. 4 (McKenna et al., 2010), following the Broad Institute's
193 best practices for RNA-seq data
194 (https://software.broadinstitute.org/gatk/documentation/article.php?id=3891; last updated 2017-
195 01-21). For *de novo* TRINITY assemblies, we first constructed SuperTranscripts, (Davidson,
196 Hawkins, & Oshlack, 2017), which attempt to collapse overlapping contigs within TRINITY
197 "genes", and then called variants with the GATK pipeline, based upon python scripts provided
198 with TRINITY v. 2.5.1. SuperTranscript alignments to the genome were used to define unique
199 polymorphic site ids that included cases where (a) both methods called genotypes, (b) only map-
200 to-reference called a genotype (including sites with and without valid SuperTranscript
201 alignments), or (c) a genotype was only called off a SuperTranscript. We recorded cases where
202 multiple SuperTranscripts were aligned to the same genomic position, taking the union of all
203 reported alleles as the ST genotype for that position. We restricted our comparison of the two
204 methods by filtering out REF sites that were indels or contained > 2 alleles. Using genotype calls
205 based upon the reference annotation as the gold standard, we calculated 11 different metrics,
206 defined as follows:

207 - **Error rate:** # positions with a genotype call in either map-to-reference or map-to-
208 SuperTranscripts, where SuperTranscript alleles don't match map-to-reference alleles ÷
209 total # positions with a genotype call in either map-to-reference or map-to-
210 SuperTrasncripts; includes positions with no overlapping ST.

211 • **Error-Snv2Indel**: # biallelic map-to-reference SNVs called as an indel in map-to-
212 SuperTranscripts ÷ # bi-allelic map-to-reference SNVs.
213 • **FN**: # map-to-reference polymorphic sites with an invariant or missing genotype in map-
214 to-SuperTranscripts ÷ # map-to-reference polymorphic sites; includes sites with no
215 overlapping SuperTranscripts.
216 • **FN-FractionNoAlignment**: # FN with no aligned SuperTranscript ÷ # FN
217 • **FN-withAlignment**: Same as **FN**, but restricted to sites with ≥1 overlapping
218 SuperTranscript alignments
219 • **FP**: # map-to-reference invariant sites with a polymorphic map-to-SuperTranscript
220 genotype ÷ # map-to-reference invariant sites.
221 • **FP-indel**: # map-to-SuperTranscript indel polymorphic sites at map-to-reference
222 invariant sites ÷ # false positive sites.
223 • **FP-hetSNV**: # false positive sites with map-to-reference biallelic SNV genotypes ÷ #
224 false positive sites.
225 • **Recall**: # polymorphic map-to-reference genotypes with matching map-to-
226 SuperTranscripts genotypes ÷ # polymorphic map-to-reference genotypes
227 • **Recall-withAlignment**: same as **Recall** restricted to sites with ≥1 overlapping
228 SuperTranscript alignments
229 • **Precision:** # polymorphic map-to-SuperTranscript genotypes that match map-to-
230 reference genotypes ÷ # polymorphic map-to-SuperTranscript genotypes.
231 See Supplementary Methods 6 for more details.
232
233 **2.10 Analysis reproducibility**

234 Software command lines, scripts and accessory files used to implement analyses described above
235 are available at https://github.com/harvardinformatics/TranscriptomeAssemblyEvaluation.
236
237
238 **3. RESULTS**
239 **3.1 "High quality" assemblies are highly fragmented**
240 Brain tissue is both transcriptionally complex and expected to express a large proportion of an
241 organism's overall transcriptional profile. It thus represents a challenging test case for *de novo*
242 transcriptome assembly. For this reason, the bulk of our analyses focus on data generated for
243 experiments involving mouse (*Mus* spp.) brain.
244 Our assemblies were highly fragmented, consisting of anywhere from 70,375 – 582,133
245 contigs, with median contig sizes ranging from just a few hundred bases to slightly over 1kb, and
246 N_{50} ranging from 721 to 3021 bp (Supplementary Table 2). This is typical of many published
247 transcriptome assemblies, e.g. Bentley et al., 2017; Faherty et al., 2018; Nespolo et al., 2018.
248 Yet, typically employed measures of accuracy and completeness indicate that all assemblies
249 were of reasonably high quality. 86.4 – 98.9% of read pairs aligned to their respective
250 assemblies, and upwards of 95% of these alignments occurred in a concordant fashion, with both
251 reads aligning to the same contig (Supplementary Table 3). For the *Mus* assemblies, 77.1 –
252 90.4% of BUSCOs were recovered in either complete or fragmented form; BUSCO recovery was
253 considerably lower for the *Xenopus*, *Drosophila*, and *Tetraodon* assemblies (Supplementary
254 Table 4). The large number of short contigs strongly suggest that *de novo* transcriptome
255 assemblies do not typically assembly full length transcripts.
256

257 **3.2 Assembly functional composition is biased**

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

268

269 **3.3 Frequent genotyping errors**

270 SNP calls and allele frequencies inferred from them and that are derived from transcriptome
271 assemblies have been used in a variety of contexts, including estimation genetic diversity in
272 evolutionary analyses (Romiguier et al., 2014), population structure (Tepolt & Palumbi, 2015),
273 reconstruction of demographic history and gene flow (De Wit & Palumbi, 2013; Le Gac et al.,
274 2016; Roda, Mendes, Hahn, & Hopkins, 2017), estimation of dN/dS (Cokus, Gugger, & Sork,
275 2015), F_{ST} outlier analysis (Bankers et al., 2017) , and tests for allele-specific expression (Singh
276 et al., 2018). In principle, redundancy through assembly of haplotypes as distinct contigs should
277 lead to erroneous genotype calls and biased estimates of allele frequency spectra. For example,
278 two contigs can actually be haplotypes of a single locus, such that a truly heterozygous position
279 is instead represented as two separate contigs, one homozygous for the reference allele and one

280 homozygous for the alternative allele. Two lines of evidence indicate such redundancy is present
281 in transcriptome assemblies. First, BLAT all-by-all alignments indicate that many contigs are
282 nearly identical (Supplementary Figures 4-6), in some cases differing only by single nucleotide
283 variants (e.g. Supplementary Figure 4b). Surprisingly, TRINITY assemblies will produce such
284 highly similar contigs from different assembly graphs which are purported to assembly isoforms
285 of a gene or sets of closely related paralogs (Supplementary Figure 6). Second, there is a high
286 frequency of cases where all nucleotides of a transcript from a single-isoform *Mus* gene are
287 covered by ≥ 1 *de novo* assembly contig (Supplementary Figure 7). The latter are primarily due
288 to slight variations in the assembly of intronic bases, and secondarily due to differences in CDS
289 content (Supplementary Figure 8).

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

296 The frequency of errors, in which the allele set recovered by the two methods differed or
297 in which a genomic position with a called genotype contained no overlapping SuperTranscript
298 nucleotides, was high, ranging from 30-37% for the wild samples, and 49-83% for the inbred
299 samples (Table 1). Error rates varied among SuperTranscripts within a *de novo* transcriptome
300 assembly, with some SuperTranscripts demonstrating extremely high error rates. While the
301 median error rate per SuperTranscript was 0.14% per base or less, maximum per SuperTranscript
302 rates typically exceeded 10% per base (Supplementary Table 5). For five of six samples, the

303 median number of errors per SuperTranscript was 2, but, depending upon the sample, as many as
304 56-187 errors could occur within a single SuperTranscript (Supplementary Table 5). Thus, within
305 any *de novo* transcriptome assembly, a small number of contigs may contribute
306 disproportionately to noise that may obscure the true patterns of polymorphism. On a per-sample
307 basis, errors show a modest yet statistically significant bias towards the distal end of a
308 SuperTranscript for all samples except MDC (sample-level Wilcoxon 1-sided tests, $H_0 = 0.5$, $p <$
309 2×10^{-16}; Supplementary Figure 9).

310 The false negative rate (FN; 23-71%) was greater than the false positive rate (FP; 15-
311 87%) in four of six samples. The latter includes all classes of errors including indel
312 polymorphism. Anywhere from 42-53% of FN were due to there being no SuperTranscript
313 aligned at a genomic position where a map-to-reference heterozygote genotype call was made
314 (Table 1 FN-FractionNoAlignment). Given that rates of mapping SuperTranscripts to the
315 genome were extremely high (>99%), these FN instances are primarily due to a failure to
316 assemble transcripts overlapping polymorphic genomic positions. Even at sites called as
317 heterozygotes in map-to-reference analyses that were assembled into SuperTranscripts, false
318 negative rates were still relatively high, from 13-15% and 37-59% in the wild and inbred
319 samples, respectively (Table 1).

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

326 reference at sites where SuperTranscripts mapped to the genome (Supplementary Figure 12).
327 Given the magnitude of coverage reduction is small, we suspect it may not be the only factor
328 contributing to false negatives at sites where a SuperTranscript covers a polymorphic genomic
329 position.

330 The false positive (FP) rate ranged from 15-87%. Incorrect heterozygous and indel
331 genotype calls at truly invariant sites are expected when paralogs are collapsed, and both occur
332 with considerable frequency in our samples. From 6-22% of FPs are erroneous indel genotypes,
333 8-62% of FPs are erroneous heterozygous genotype calls, and the remainder are other classes of
334 SuperTranscript polymorphisms, including both indel and single nucleotide variants, with > 2
335 alleles. The inference of frequent paralog collapses is also supported by the high rate at which
336 true SNV polymorphisms are called as indels in SuperTranscripts (Error-Snv2Indel, 16-31%).
337 When genotypes were called in SuperTranscripts, the frequency of calls that were correct (i.e.
338 precision) ranged from 66-83% for the wild samples, but precision for the inbred samples,
339 particularly MDC, was substantially lower. Recall ranged from 76-81% in the wild samples; as
340 with precision, recall for both inbred samples was substantially lower. The frequency of
341 genotyping errors seems likely to distort evolutionary inferences and reduce power across a wide
342 range of statistical methods, particularly those aimed at detecting signals of selection on
343 particular polymorphisms, allele-phenotype and allele frequency-environment associations.
344

345 **3.4 Negatively biased heterozygosity estimates**

346 In all but the inbred strain samples (MDC and BALB/c), we observed reduced heterozygosity in
347 the exonic regions of SuperTranscripts relative to mapping to those regions in the reference
348 genome (Figure 1). A substantial contributor to this bias is the fact that, as noted above,

349 approximately half of map-to-reference heterozygous positions are not overlapped by any
350 SuperTranscript. In addition, this bias does not result from SuperTranscripts being preferentially
351 assembled in regions of lower heterozygosity, given that, for the wild samples, map-to-reference
352 heterozygosity is actually higher in SuperTranscript-overlapping regions (Supplementary Figure
353 13). Furthermore, the bias cannot be due to some unique feature of genotype calls in exons, as
354 the same pattern holds when considering all genotype calls (Supplementary Figure 14). Another
355 possible explanation may be lower power to detect low frequency alternative alleles in map-to-
356 SuperTranscript analyses. However, in our five samples consisting of pooled individuals, minor
357 allele frequency at map-to-reference heterozygous sites did not vary between correct
358 SuperTranscript heterozygous calls, false negatives, and other genotyping errors (Supplementary
359 Figure 15). Downward bias in heterozygosity estimates occurs despite a relatively high
360 prevalence of false positive heterozygotes. Because this bias shifts all estimates towards zero, it
361 truncates the observed range of variation, making it more difficult to make meaningful inferences
362 from comparisons of diversity, whether those comparisons take place among assemblies derived
363 from different populations or species, or at the level of individual genes within or between
364 assemblies.
365
366 **3.5 Single contigs are poor expression estimators**
367 Expression estimates derived from the contigs that are the best BLAT hit relative to the reference
368 transcripts are only weakly correlated (spearman's ρ: CPM: 0.070-0.387; TPM: 0.045-0.392;
369 rescaled CPM: 0.083-0.415), with expression for those transcripts, and show positive bias for
370 lowly expressed reference transcripts (Supplementary Figure 16ab, Supplementary Table 6).
371 Correlations between best-hit contig expression and expression of the associated reference gene,

372 after accounting for effective length differences, are slightly better but still relatively weak
373 (spearman's ρ: rescaled CPM: 0.103-0.657 Supplementary Table 6). Overall, the use of single
374 contigs as proxies for gene-level expression appears problematic.
375
376 **3.6 Correctable gene-level expression bias**
377 Expression estimates derived from *de novo* transcriptome assemblies are far more robust at the
378 gene level, when one is able to group contigs into genes via annotation (Figure 2, Table 2). It is
379 reassuring that the correlation between the methods is quite strong, although there is sizeable
380 positive bias at lower expression levels for TPM, which is substantially reduced for CPM.
381 Furthermore, for many genes there are substantial deviations from estimates obtained with
382 mapping to the reference, even at high expression levels (Figure 2, Supplementary Figure 17).
383 For estimates generated with RSEM, bias and deviations from the map-to-reference estimates are
384 reduced when one rescales CPMs to account for differences in effective length estimates (Figure
385 2c). This indicates that effective length differences—that is, the difference in the number of start
386 sites for read alignments, given a particular sequenced fragment size, between a full-length
387 transcript, and a set of transcript sub-sequences represented by contigs—are part of the problem,
388 and in particular, that effective length is not properly estimated on the fragmented gene models
389 that are typical in *de novo* transcriptome assemblies. Biases in gene expression are reduced as the
390 difference in effective length between map-to-reference and map-to-transcriptome-assembly
391 decreases (Figure 3). Even so, disparities in effective length estimates between map-to-reference
392 and map-to-*de novo*-transcriptome-assembly approaches can persist even at high reference gene
393 expression, with many *de novo* transcriptome-based estimates of gene expression showing a

394 fourfold or greater deviation relative to the map-to-reference benchmark (Figure 2,
395 Supplementary Figure 18).
396 Correlations between reference gene expression and gene-level summing across assembly
397 contigs obtained with KALLISTO were very similar to those obtained with RSEM (Table 3,
398 Supplementary Figure 17). Adjusting for effective length had little effect on KALLISTO CPM,
399 and in some cases reduced the correlation between methods. KALLISTO may be doing a better job
400 at estimating effective length on transcriptome assemblies, such that it's inferior performance
401 relative to RSEM must be due to other reasons.
402
403 **3.7 Expression bias and DE testing**
404 Biases in expression estimates derived from transcriptome assemblies might not have large
405 effects on downstream differential expression analyses if biases are consistent across samples.
406 While the pooled *Mus* data sets we used did not permit us to look at inter-individual variation
407 amongst individual libraries used to build a single assembly, we looked across the wild *Mus*
408 datasets to see if biases in gene-level expression were concordant, i.e. were biases for a particular
409 gene always of similar magnitude and in the same direction. We focus on results for TRINITY
410 assemblies of wild *Mus* samples, as TRINITY consistently showed stronger correlations with map-
411 to-reference based estimates, such that our findings are conservative with respect to assemblers
412 showing weaker correlations. Reassuringly, log-fold change of transcriptome assembly vs. map-
413 to-reference expression estimators (LFC_{METHOD}) showed broad patterns of moderate concordance
414 at the gene level indicating that gene-level features influence the magnitude of bias
415 (Supplementary Figure 19). Inter-sample correlations were, for CPM, 0.625-0.727; TPM, 0.790-
416 0.830; and rescaled CPM, 0.605-0.671. These correlations are weaker for CPM and rescaled

417 CPM, in part, because biases are reduced overall, compressing LFC_{METHOD} toward zero. Despite
418 these broad correlations, a number of genes may show excessive bias. We measured this as a bias
419 effect size that indicates how much greater a difference in expression would be detected using a
420 *de novo* transcriptome approach, relative to mapping to reference annotation. Comparing the
421 FRA and DEU samples, for CPM, TPM, and rescaled CPM, 5.24%, 5.16% and 3.35% of genes
422 showed bias effect sizes ≥ 5, with more than 2% of genes showing an order of magnitude of
423 difference between methods of ≥ 10 for CPM and TPM (Figure 4).

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

436

437 **3.8 Effects of filtering on expression estimates**

438 Filtering out lowly expressed *de novo* transcriptome assembly contigs as well as those contigs
439 without a BLASTP hit to Uniref90 led to a noticeable improvement in the correlation between

440 map-to-transriptome assembly and map-to-reference abundance estimates for protein coding
441 genes (Figure 2d-f). Post-filtering, bias and noise in TPM was dramatically reduced. However,
442 despite a nearly a correlation nearly identical to that for CPM and rescaled CPM, there is
443 consistent positive bias across the range of map-to-reference expression (Figure 2e).
444 Nevertheless, the improvement across all three metrics relative to unfiltered assemblies belies a
445 tradeoff that is hidden from view from researchers without access to an annotated genome:
446 filtering leads to a loss of many expressed protein coding genes that would be recovered without
447 filtering. For the FRA assembly, 21.0% of such genes are filtered out, although only 10.6% of
448 genes are lost if one excludes from consideration lowly expressed genes with map-to-reference
449 TPM < 1 (Figure 5). Depending upon the sample-assembler combination, 9.1-32.5% of
450 expressed protein coding genes can be lost due to filtering, and for genes with map-to-reference
451 TPM ≥ 1, 6.0-27.1% (Table 2). In nearly all sample-assembler combinations, the proportion of
452 genes lost due to filtering far exceeded the proportion that were missing due to a failure to
453 assemble those genes, with most exceptions occurring for BINPACKER assemblies, which
454 performed poorly with respect to recovery of expressed genes (Table 2).
455

456 **DISCUSSION**

457 *De novo* transcriptome assemblies have been an important entry point for functional and
458 population genomics in non-model organisms, but the two key assumptions underlying their
459 application are consistently violated, albeit in unexpected ways. Below, we explain how
460 particular compositional aspects of *de novo* transcriptome assemblies may impact the power of
461 evolutionary inference.

462 By far the most unexpected findings from our study were the extremely high genotyping
463 error rates derived from transcriptome assemblies, specifically SuperTranscript representations
464 of TRNITY assemblies. False negatives were due, in part, to the absence of a SuperTranscript
465 aligned to a putatively true polymorphic site recovered with a map-to-reference approach. False
466 negatives are due to either the failure to assembly any contigs at all for a gene (i.e. missing
467 genes), or the incomplete coverage of genes not extending over some fraction of polymorphic
468 sites. That genotyping errors were biased towards the distal end of contigs suggest that assembly
469 graphs are getting broken at hard to assemble features containing low complexity and/or
470 increased polymorphism. When such graph-splitting occurs, the intervening polymorphic sites
471 will, by definition, not be covered by a SuperTranscript, leading to downwardly biased estimates
472 of heterozygosity in *de novo* assemblies. False negatives may have also been caused by the
473 slightly reduced read depth at homologous positions in read alignments to SuperTranscripts. That
474 the reduction in read depth in the latter, was, on average, small suggests that other factors may
475 play a role. It may be that genotype inference methods such at the GATK HAPLOTYPE CALLER
476 that leverage haplotypic information, have poorer performance when iterating over alignments to
477 relatively short assembly contigs. Another possibility is that assembly errors may generate
478 patterns that get flagged by standard genotyping filters. In contrast, false positives appear to be
479 driven by paralog collapses given the bias towards 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 weighted-adjusted CPM will apply to similarly
534 transformed counts. Adjusting counts across a set of species would require treating one species
535 as the "reference", and rescaling counts for other species relative to it. Outside of the realm of
536 formal tests of differential expression, we suggest that rescaled CPM could be useful in
537 phylogenetic comparative analyses of expression at the gene level. One of the big challenges of
538 these methods is accounting for length differences between transcript/contig sets belonging to
539 orthologous genes, and the adjusted CPM appears to substantially outperform other metrics in
540 this regard. We should note that our adjustment method is relatively simple, and that it may be
541 possible to adopt more sophisticated approaches that leverage such information as sequence
542 composition and complexity for additional improvements.

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

552 We do not test here datasets amenable to differential expression analyses, and a counter-
553 argument to our findings of bias and noise is that neither should have a substantial effect, so long

554 as bias and deviations from map-to-reference expectations are consistent across replicates and
555 conditions. While we did find broad congruence in bias across samples, gene-level differences in
556 bias among samples were frequent enough such that they will likely elevate false positive rates
557 for differential expression tests derived from *de novo* transcriptome assemblies. This effect can
558 be minimized by utilizing metrics that adjust for effective length differences.
559

560 **Conclusion**

561 Given the challenges associated with eco-evolutionary analyses that rely on *de novo*
562 transcriptome assemblies and the rapidly falling costs of long read sequencing, an unavoidable
563 recommendation for research teams launching new studies may be to generate a provisional
564 genome assembly, e.g. using a combination of Oxford Nanopore or PacBio long reads and
565 Illumina paired-end short reads (Austin et al., 2017; Marcionetti, Rossier, Bertrand, Litsios, &
566 Salamin, 2018; Michael et al., 2018; Nowoshilow et al., 2018), at least in some cases. Such an
567 assembly can be annotated with *ab initio* gene prediction tools and Illumina RNA-seq paired-end
568 reads. Given a reasonable quality annotated genome, fewer reads per sample would have to be
569 allocated for downstream expression analyses, and the cost savings from lighter RNA
570 sequencing would partially offset costs of genomic sequencing and *de novo* assembly.
571 Nevertheless, there are three reasons whole *de novo* transcriptome assembly might still be
572 attractive. First, genome assembly remains challenging and expensive species with large
573 genomes or with unusual genome structure precluding standard methods. Second, thousands of
574 RNA-seq data sets now exist for species without reference genomes, and it is highly unlikely that
575 genome assemblies will be carried out any time soon for anything but a small fraction of these
576 species. Third, analyses based upon genome assemblies will depend strongly on the quality of

577 the assembly. Even for genomes of reasonably high quality, protein coding genes may be
578 missing, either through problems with the annotation process or due to the fact that genes fall
579 into assembly gaps (Peona, Weissensteiner, & Suh, 2018). In other words, genome assembly in
580 not necessarily a panacea for all problems related to expression analyses. However analytically
581 challenging, there is great potential for employing *de novo* transcriptome assemblies in
582 interesting comparative analyses, so long as the conducted in an informed fashion—avoiding
583 analyses modes that suffer from fundamental biases, mitigating biases where possible for others,
584 and, always, placing appropriate caveats given the biases that will inevitably remain.
585

586 **ACKNOWLEDGMENTS**

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

594 **DATA AVAILABILITY**

595 Upon acceptance for publication, all transcriptome assemblies, SuperTranscript assemblies, and
596 vcf genotype files will be made available on Dryad.
597

598 **AUTHOR CONTRIBUTIONS**

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

605 Austin, C. M., Tan, M. H., Harrisson, K. A., Lee, Y. P., Croft, L. J., Sunnucks, P., … Gan, H. M.
606 (2017). De novo genome assembly and annotation of Australia's largest freshwater fish,
607 the Murray cod (Maccullochella peelii), from Illumina and Nanopore sequencing read.
608 *GigaScience*, *6*(8), 1–6. doi: 10.1093/gigascience/gix063
609 Bankers, L., Fields, P., McElroy, K. E., Boore, J. L., Logsdon, J. M., & Neiman, M. (2017). Genomic
610 evidence for population-specific responses to co-evolving parasites in a New Zealand
611 freshwater snail. *Molecular Ecology*, *26*(14), 3663–3675. doi: 10.1111/mec.14146
612 Bentley, B. P., Haas, B. J., Tedeschi, J. N., & Berry, O. (2017). Loggerhead sea turtle embryos
613 (Caretta caretta) regulate expression of stress response and developmental genes when
614 exposed to a biologically realistic heat stress. *Molecular Ecology*, *26*(11), 2978–2992.
615 doi: 10.1111/mec.14087
616 Berens, A. J., Hunt, J. H., & Toth, A. L. (2015). Comparative transcriptomics of convergent
617 evolution: different genes but conserved pathways underlie caste phenotypes across
618 lineages of eusocial insects. *Molecular Biology and Evolution*, *32*(3), 690–703. doi:
619 10.1093/molbev/msu330

620 Birol, I., Jackman, S. D., Nielsen, C. B., Qian, J. Q., Varhol, R., Stazyk, G., … Jones, S. J. M. (2009).
621 De novo transcriptome assembly with ABySS. *Bioinformatics*, *25*(21), 2872–2877. doi:
622 10.1093/bioinformatics/btp367
623 Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq
624 quantification. *Nature Biotechnology*, *34*(5), 525–527. doi: 10.1038/nbt.3519
625 Cokus, S. J., Gugger, P. F., & Sork, V. L. (2015). Evolutionary insights from de novo transcriptome
626 assembly and SNP discovery in California white oaks. *BMC Genomics*, *16*, 552. doi:
627 10.1186/s12864-015-1761-4
628 Davidson, N. M., Hawkins, A. D. K., & Oshlack, A. (2017). SuperTranscripts: a data driven
629 reference for analysis and visualisation of transcriptomes. *Genome Biology*, *18*(1). doi:
630 10.1186/s13059-017-1284-1
631 Davidson, N. M., & Oshlack, A. (2014). Corset: enabling differential gene expression analysis for
632 de novoassembled transcriptomes. *Genome Biology*, *15*(7), 410. doi: 10.1186/s13059-
633 014-0410-6
634 De Wit, P., & Palumbi, S. R. (2013). Transcriptome-wide polymorphisms of red abalone (Haliotis
635 rufescens) reveal patterns of gene flow and local adaptation. *Molecular Ecology*, *22*(11),
636 2884–2897. doi: 10.1111/mec.12081
637 Faherty, S. L., Villanueva-Cañas, J. L., Blanco, M. B., Albà, M. M., & Yoder, A. D. (2018).
638 Transcriptomics in the wild: Hibernation physiology in free-ranging dwarf lemurs.
639 *Molecular Ecology*, *27*(3), 709–722. doi: 10.1111/mec.14483

640 Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., … Regev, A.
641 (2011a). Full-length transcriptome assembly from RNA-Seq data without a reference
642 genome. *Nature Biotechnology*, *29*(7), 644–652. doi: 10.1038/nbt.1883
643 Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., … Regev, A.
644 (2011b). Full-length transcriptome assembly from RNA-Seq data without a reference
645 genome. *Nature Biotechnology*, *29*(7), 644–652. doi: 10.1038/nbt.1883
646 Hornett, E. A., & Wheat, C. W. (2012). Quantitative RNA-Seq analysis in non-model species:
647 assessing transcriptome assemblies as a scaffold and the utility of evolutionary
648 divergent genomic reference species. *BMC Genomics*, *13*, 361. doi: 10.1186/1471-2164-
649 13-361
650 Kannan, S., Hui, J., Mazooji, K., Pachter, L., & Tse, D. (2016). Shannon: An Information-Optimal
651 de Novo RNA-Seq Assembler. *BioRxiv*, 039230. doi: 10.1101/039230
652 Kent, W. J. (2002). BLAT--the BLAST-like alignment tool. *Genome Research*, *12*(4), 656–664. doi:
653 10.1101/gr.229202. Article published online before March 2002
654 Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature*
655 *Methods*, *9*(4), 357–359. doi: 10.1038/nmeth.1923
656 Le Gac, M., Metegnier, G., Chomérat, N., Malestroit, P., Quéré, J., Bouchez, O., … Chapelle, A.
657 (2016). Evolutionary processes and cellular functions underlying divergence in
658 Alexandrium minutum. *Molecular Ecology*, *25*(20), 5129–5143. doi: 10.1111/mec.13815
659 Li, B., & Dewey, C. N. (2011a). RSEM: accurate transcript quantification from RNA-Seq data with
660 or without a reference genome. *BMC Bioinformatics*, *12*, 323. doi: 10.1186/1471-2105-
661 12-323

662 Li, B., & Dewey, C. N. (2011b). RSEM: accurate transcript quantification from RNA-Seq data with
663 or without a reference genome. *BMC Bioinformatics*, *12*, 323. doi: 10.1186/1471-2105-
664 12-323
665 Li, K., Wang, H., Cai, Z., Wang, L., Xu, Q., Lövy, M., … Nevo, E. (2016). Sympatric speciation of
666 spiny mice, *Acomys* , unfolded transcriptomically at Evolution Canyon, Israel.
667 *Proceedings of the National Academy of Sciences*, *113*(29), 8254–8259. doi:
668 10.1073/pnas.1608743113
669 Liu, J., Li, G., Chang, Z., Yu, T., Liu, B., McMullen, R., … Huang, X. (2016). BinPacker: Packing-
670 Based De Novo Transcriptome Assembly from RNA-seq Data. *PLOS Computational*
671 *Biology*, *12*(2), e1004772. doi: 10.1371/journal.pcbi.1004772
672 MacManes, M. D. (2018). The Oyster River Protocol: a multi-assembler and kmer approach for
673 de novo transcriptome assembly. *PeerJ*, *6*. doi: 10.7717/peerj.5428
674 Malik, L., Almodaresi, F., & Patro, R. (2018). Grouper: graph-based clustering and annotation for
675 improved de novo transcriptome analysis. *Bioinformatics (Oxford, England)*, *34*(19),
676 3265–3272. doi: 10.1093/bioinformatics/bty378
677 Marcionetti, A., Rossier, V., Bertrand, J. A. M., Litsios, G., & Salamin, N. (2018). First draft
678 genome of an iconic clownfish species (Amphiprion frenatus). *Molecular Ecology*
679 *Resources*, *18*(5), 1092–1101. doi: 10.1111/1755-0998.12772
680 McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … DePristo, M. A.
681 (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-
682 generation DNA sequencing data. *Genome Res*, *20*(9), 1297–1303. doi:
683 10.1101/gr.107524.110

Mi, H., Muruganujan, A., Casagrande, J. T., & Thomas, P. D. (2013). Large-scale gene function analysis with the PANTHER classification system. *Nature Protocols*, *8*(8), 1551–1566. doi: 10.1038/nprot.2013.092

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

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

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

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

Pankey, M. S., Minin, V. N., Imholte, G. C., Suchard, M. A., & Oakley, T. H. (2014). Predictable transcriptome evolution in the convergent and complex bioluminescent organs of squid. *Proceedings of the National Academy of Sciences*, *111*(44), E4736–E4742. doi: 10.1073/pnas.1416574111

705 Peona, V., Weissensteiner, M. H., & Suh, A. (2018). How complete are “complete” genome
706 assemblies?—An avian perspective. *Molecular Ecology Resources*, *18*(6), 1188–1195.
707 doi: 10.1111/1755-0998.12933
708 Roda, F., Mendes, F. K., Hahn, M. W., & Hopkins, R. (2017). Genomic evidence of gene flow
709 during reinforcement in Texas Phlox. *Molecular Ecology*, *26*(8), 2317–2330. doi:
710 10.1111/mec.14041
711 Romiguier, J., Gayral, P., Ballenghien, M., Bernard, A., Cahais, V., Chenuil, A., … Galtier, N.
712 (2014). Comparative population genomics in animals uncovers the determinants of
713 genetic diversity. *Nature*, *515*(7526), 261–263. doi: 10.1038/nature13685
714 Schwentner, M., Combosch, D. J., Pakes Nelson, J., & Giribet, G. (2017). A Phylogenomic
715 Solution to the Origin of Insects by Resolving Crustacean-Hexapod Relationships. *Current*
716 *Biology*, *27*(12), 1818-1824.e5. doi: 10.1016/j.cub.2017.05.040
717 Siepel, A. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast
718 genomes. *Genome Research*, *15*(8), 1034–1050. doi: 10.1101/gr.3715005
719 Singh, R., Jones, T., Wai, C. M., Jifon, J., Nagai, C., Ming, R., & Yu, Q. (2018). Transcriptomic
720 analysis of transgressive segregants revealed the central role of photosynthetic capacity
721 and efficiency in biomass accumulation in sugarcane. *Scientific Reports*, *8*(1), 4415. doi:
722 10.1038/s41598-018-22798-5
723 Soneson, C., Love, M. I., & Robinson, M. D. (2015). Differential analyses for RNA-seq: transcript-
724 level estimates improve gene-level inferences. *F1000Research*, *4*, 1521. doi:
725 10.12688/f1000research.7563.1

726 Tepolt, C. K., & Palumbi, S. R. (2015). Transcriptome sequencing reveals both neutral and
727 adaptive genome dynamics in a marine invader. *Molecular Ecology*, *24*(16), 4145–4158.
728 doi: 10.1111/mec.13294
729 Todd, E. V., Black, M. A., & Gemmell, N. J. (2016). The power and promise of RNA-seq in ecology
730 and evolution. *Molecular Ecology*, *25*(6), 1224–1241. doi: 10.1111/mec.13526
731 Veilleux, H. D., Ryu, T., Donelson, J. M., van Herwerden, L., Seridi, L., Ghosheh, Y., … Munday, P.
732 L. (2015). Molecular processes of transgenerational acclimation to a warming ocean.
733 *Nature Climate Change*, *5*(12), 1074–1078. doi: 10.1038/nclimate2724
734 Velotta, J. P., Wegrzyn, J. L., Ginzburg, S., Kang, L., Czesny, S., O'Neill, R. J., … Schultz, E. T.
735 (2017). Transcriptomic imprints of adaptation to fresh water: parallel evolution of
736 osmoregulatory gene expression in the Alewife. *Molecular Ecology*, *26*(3), 831–848. doi:
737 10.1111/mec.13983
738 Vijay, N., Poelstra, J. W., Künstner, A., & Wolf, J. B. W. (2013). Challenges and strategies in
739 transcriptome assembly and differential gene expression quantification. A
740 comprehensive *in silico* assessment of RNA-seq experiments. *Molecular Ecology*, *22*(3),
741 620–634. doi: 10.1111/mec.12014
742 Zook, J. M., Catoe, D., McDaniel, J., Vang, L., Spies, N., Sidow, A., … Salit, M. (2016). Extensive
743 sequencing of seven human genomes to characterize benchmark reference materials.
744 *Scientific Data*, *3*. doi: 10.1038/sdata.2016.25
745

746 **TABLES**

747

748 **Table 1**. Selected performance metrics comparing genotypes derived from SuperTranscript

749 collapse of TRINITY de novo transcriptome assemblies with those derived from a map-to-

750 reference genome approach. See **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

751

752 **Table 2.** Correlation between map-to-reference and map-to-de novo-transcriptome expression at
753 the gene level, for both RSEM and KALLISTO, based upon CPM, TPM, and weighted-adjusted
754 CPM. See equation in methods for details regarding the last of these. Correlations exclude
755 expressed reference genes where there are not associated de novo assembly contigs, i.e. genes
756 that are missing from the assembly. Correlations are calculated for entries with TPM > 0 for
757 either map-to-reference or map-to-de novo assembly. RSEM values in parentheses are for
758 filtered assemblies from which contigs with TPM ≤ 1 and without a BLASTP hit to Uniref90 are
759 removed, and which are only compared for protein coding genes with a minimum transcript
760 length ≥ 200bp. The % missing column indicates expressed protein coding genes that were not
761 recovered by the unfiltered transcriptome assembly. The % lost-to-filtering column indicates
762 what fraction of expressed protein coding genes that are detected in the unfiltered *de novo*
763 transcriptome assemblies are lost due to filtering. Bold text indicates, for each assembly, the
764 method that produces the strongest correlation with map-to-reference expression estimates.
765

Sample	CPM		TPM		Rescaled CPM			
	RSEM	kallisto	RSEM	kallisto	RSEM	kallisto	% missing (TPM≥1)	% lost-to-filtering (TPM≥1)
MDC								
Trinity	0.853 (0.951)	0.858	0.756 (0.933)	0.675	0.901 **(0.957)**	0.823	20.0 (2.0)	10.6 (6.0)
Shannon	0.629 (0.897)	0.833	0.567 (0.901)	0.662	0.680 (0.906)	0.784	21.3 (2.5)	31.5 (27.1)
BinPacker	0.824 (0.940)	0.825	0.796 (0.937)	0.754	0.877 (0.947)	0.783	25.0 (3.3)	9.1 (6.8)
BALB/c								
Trinity	0.869 (0.934)	0.881	0.818 (0.895)	0.819	0.897 **(0.938)**	0.863	8.6 (0.6%)	24.5 (11.5)
Shannon	0.819 (0.846)	0.859	0.790 (0.848)	0.815	0.854 (0.853)	0.826	8.8 (0.7)	32.5 (19.1)
BinPacker	0.835 (0.877)	0.846	0.801 (0.864)	0.803	0.870 (0.883)	0.824	12.4 (1.7)	20.6 (10.5)
FRA								
Trinity	0.850 (0.927)	0.861	0.809 (0.928)	0.790	0.903 **(0.933)**	0.841	11.8 (2.0)	20.6 (10.6)
Shannon	0.849	0.857	0.812	0.785	0.906	0.819	11.5 (1.6)	24.7 (14.1)

BinPacker	0.815 (0.888)	0.821	0.808 (0.902)	0.786	0.872 (0.896)	0.793	18.3 (3.2)	14.1 (10.5)
DEU								
Trinity	0.848 (0.925)	0.859	0.804 (0.926)	0.788	0.903 **(0.933)**	0.838	11.7 (2.0)	21.9 (11.5)
Shannon	0.847 (0.879)	0.855	0.812 (0.894)	0.784	0.905 (0.888)	0.813	11.3 (1.6)	25.6 (14.9)
BinPacker	0.811 (0.880)	0.818	0.808 (0.900)	0.784	0.873 (0.890)	0.789	17.6 (3.1)	14.7 (10.5)
IRN								
Trinity	0.853 (0.929)	0.862	0.806 (0.928)	0.785	0.906 **(0.937)**	0.842	12.0 (2.0)	18.9 (10.0)
Shannon	0.847 (0.893)	0.854	0.804 (0.902)	0.777	0.904 (0.902)	0.817	11.5 (1.7)	22.1 (12.6)
BinPacker	0.809 (0.888)	0.816	0.799 (0.901)	0.773	0.869 (0.898)	0.787	19.0 (3.6)	13.1 (10.2)
KZK								
Trinity	0.852 (0.926)	0.862	0.810 (0.925)	0.791	0.903 **(0.933)**	0.841	10.9 (2.0)	22.8 (12.0)
Shannon	0.852 (0.885)	0.858	0.814 (0.891)	0.793	0.907 (0.893)	0.824	10.7 (1.8)	25.7 (14.8)
BinPacker	0.816 (0.894)	0.821	0.809 (0.908)	0.788	0.872 (0.902)	0.795	16.8 (3.4)	14.9 (10.7)

767 **FIGURE LEGENDS**

768 **Figure 1.** Heterozygosity per base for exonic biallelic SNV variants called relative to the *Mus*
769 reference genome using the GATK best practices pipeline versus those obtained from
770 supertranscripts constructed from TRINITY assemblies using scripts provided with the TRINITY v.
771 2.6.5 package. See **Methods** for details concerning how exonic coordinates were obtained for
772 SuperTranscripts.

773

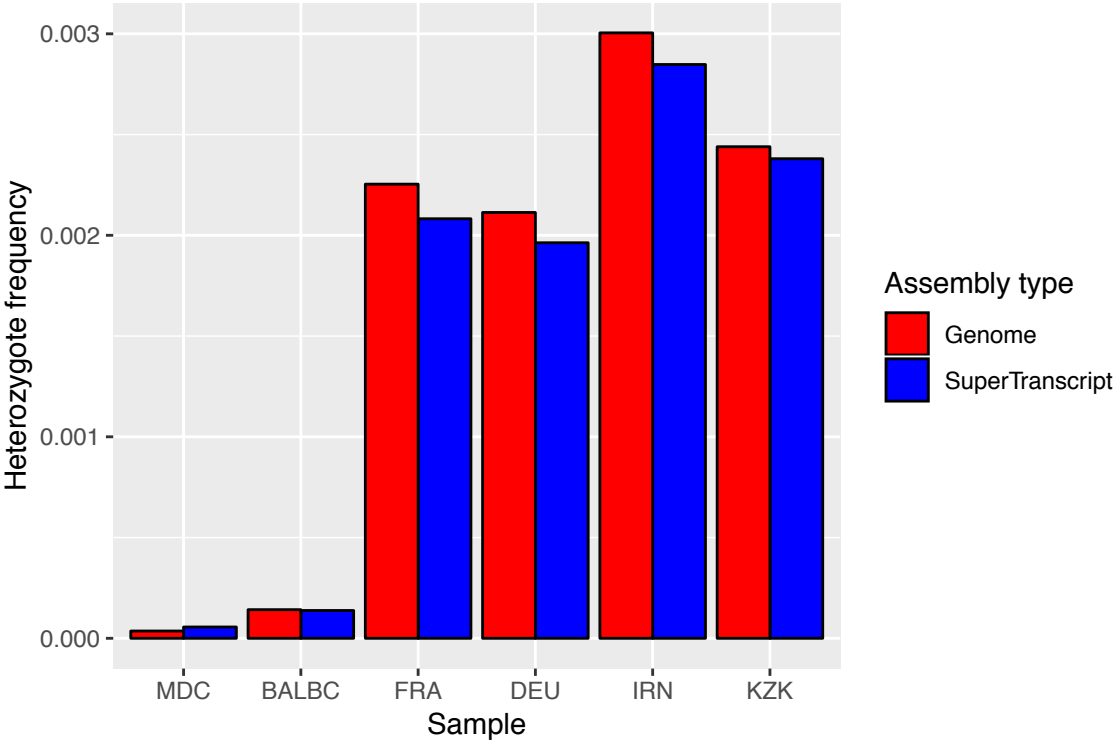
774 **Figure 2.** For the FRA data set, comparison of gene-level expression summed across BLAT
775 alignments of TRINITY contigs to reference transcripts (grouped by gene), to reference gene
776 expression in terms of **(a)** CPM, **(b)** TPM, and **(c)** weighted-adjusted CPM that accounts for
777 differences in effective length estimates between methods. For CPM, TPM, and rescaled CPM,
778 0.05, 0.01 and 0.05 are added to values prior to log-transformation, in order to visualize instances
779 where expression = 0 for one of the methods being compared. The red line represents unity, the
780 blue line is the smoothed fit to data. Gray and orange points indicate log-fold change of
781 Trinity/map-to-reference genes less than and ≥ 2, respectively. We exclude all cases where an
782 expressed reference gene has no TRINITY contigs aligned to it, such that it is missing from the *de*
783 *novo* transcriptome assembly.

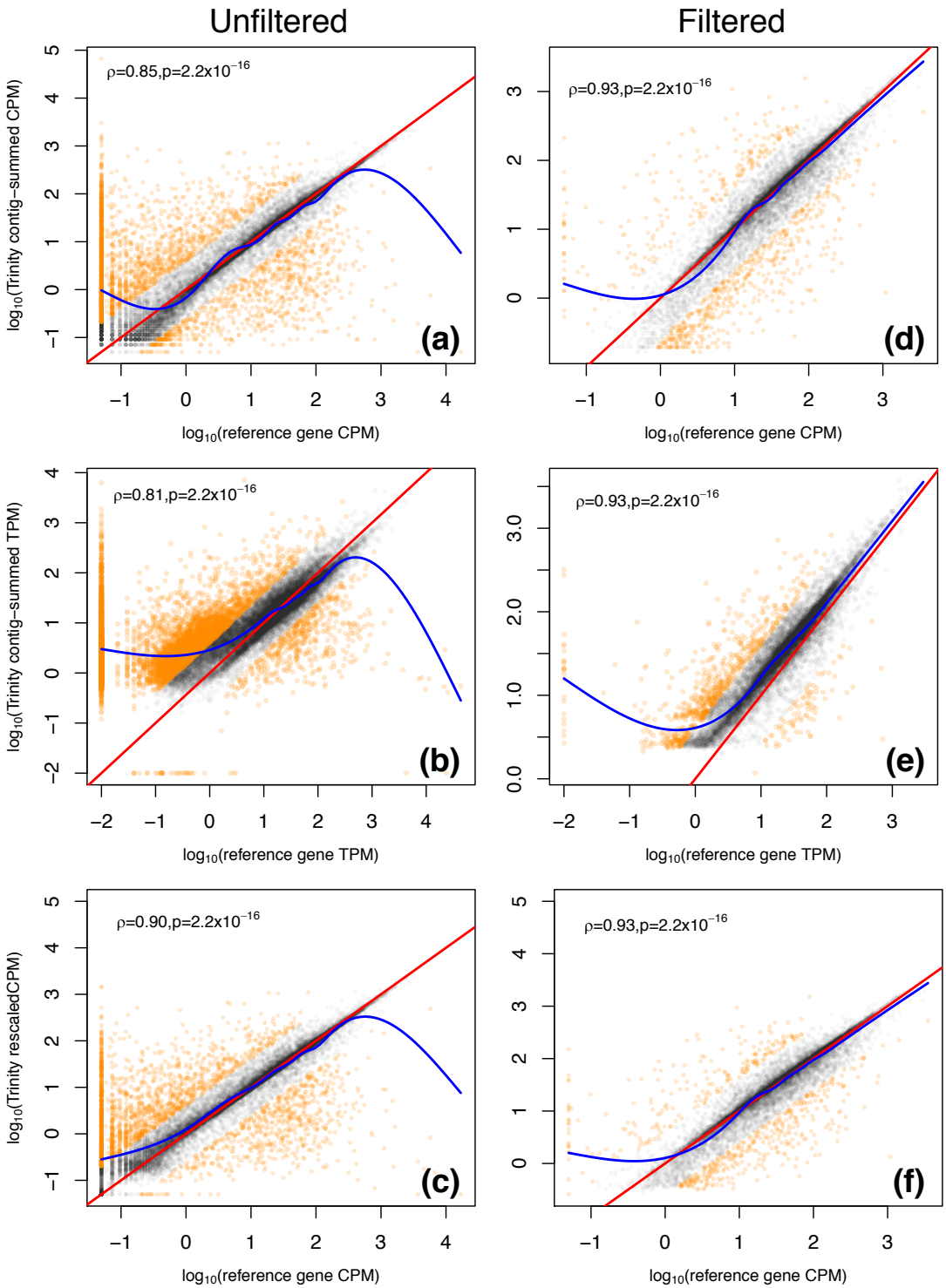
784

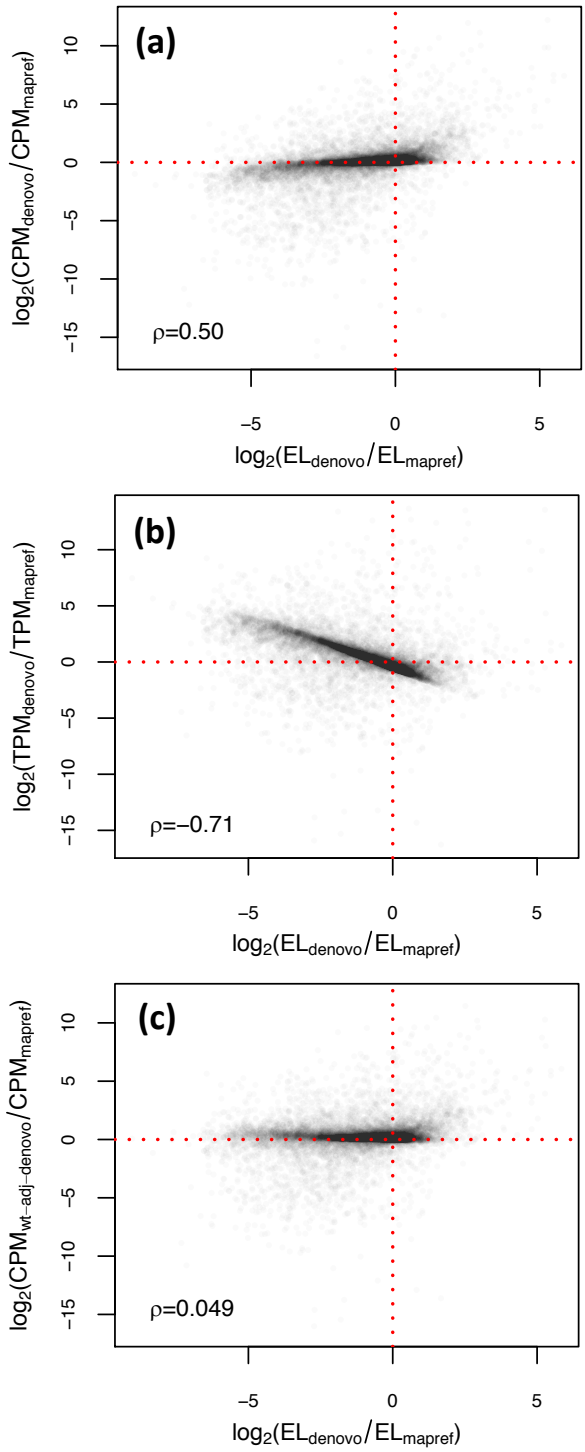
785 **Figure 3.** For the FRA data set, for genes with non-zero expression in both transcriptome
786 assembly and map-to-reference, log-fold change between TRINITY and map-to-reference
787 expression estimates as a function of difference in effective length (EL) for **(a)** CPM, **(b)** TPM,
788 and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

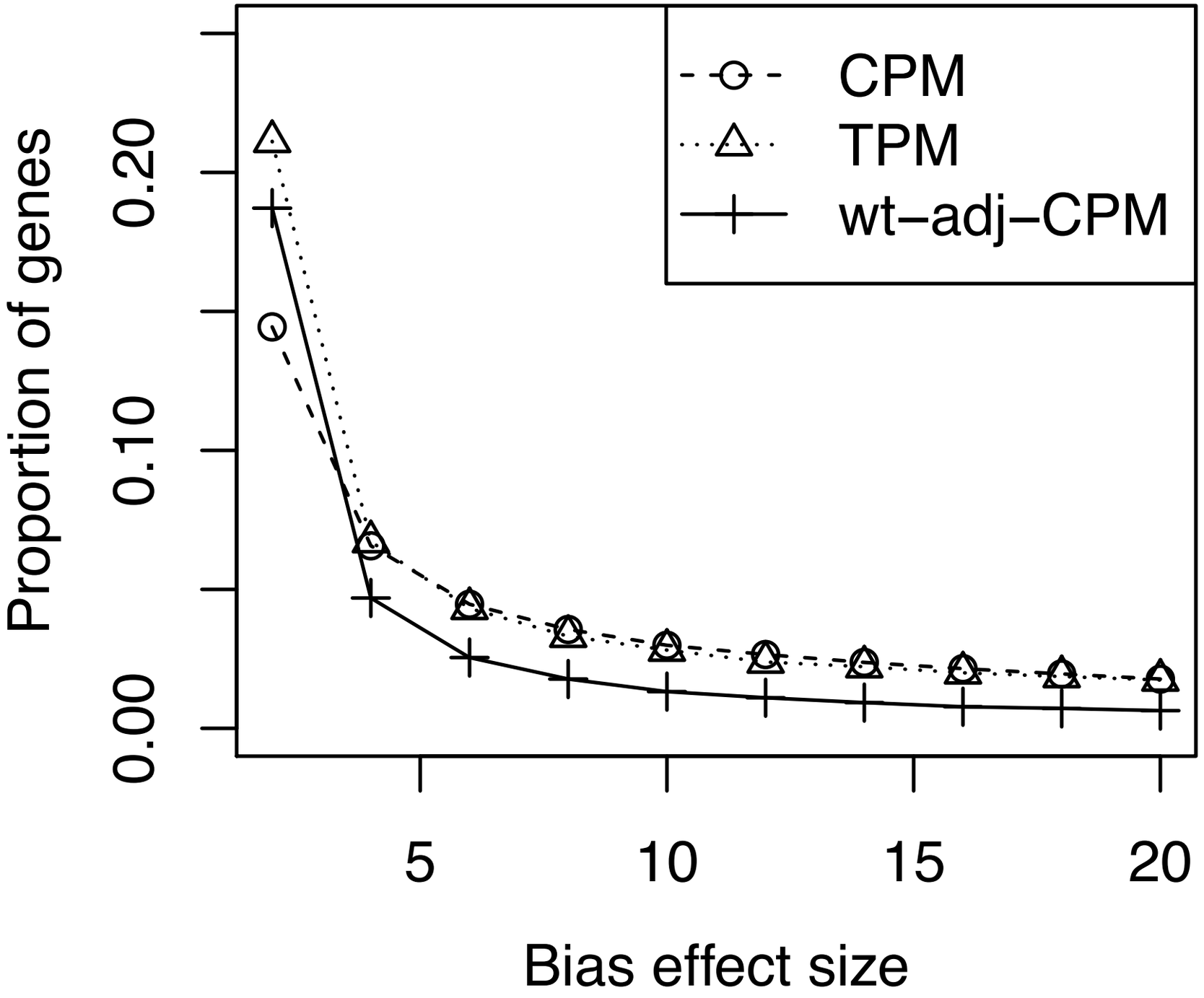
789

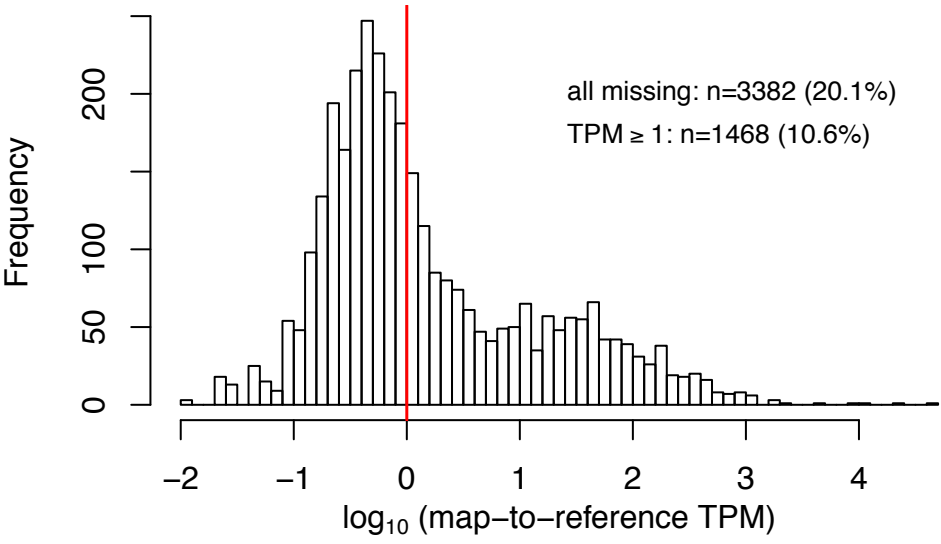
790 **Figure 4.** Frequency of bias effect sizes for gene-level expression estimated by CPM, TPM, and
791 weighted-adjusted CPM for a TRINITY-based comparison of the FRA and DEU samples.
792
793 **Figure 5.** Histogram of map-to-reference TPM for expressed protein-coding genes that were
794 detected in the FRA TRINITY assembly prior to filtering out un-annotated and lowly expressed
795 assembly contigs, but were lost due to filtering.
796


0.003
0.002
0.001
0.000
Heterozygote frequency
MDC
BALBC
FRA
DEU
IRN
KZK
Sample
Assembly type
Genome
SuperTranscript


Unfiltered
Filtered
ρ=0.85,p=2.2x10^−16
(a)
ρ=0.93,p=2.2x10^−16
(d)
ρ=0.81,p=2.2x10^−16
(b)
ρ=0.93,p=2.2x10^−16
(e)
ρ=0.90,p=2.2x10^−16
(c)
ρ=0.93,p=2.2x10^−16
(f)
log10(Trinity contig−summed CPM)
log10(Trinity contig−summed TPM)
log10(Trinity rescaledCPM)
log10(reference gene CPM)
log10(reference gene TPM)


(a)
ρ=0.50
log2(CPMdenovo/CPMmapref)
log2(ELdenovo/ELmapref)
(b)
ρ=−0.71
log2(TPMdenovo/TPMmapref)
log2(ELdenovo/ELmapref)
(c)
ρ=0.049
log2(CPMwt−adj−denovo/CPMmapref)
log2(ELdenovo/ELmapref)


CPM
TPM
wt−adj−CPM
Proportion of genes
0.00
0.10
0.20
5
10
15
20
Bias effect size


all missing: n=3382 (20.1%)
TPM ≥ 1: n=1468 (10.6%)
Frequency
0
50
100
200
−2
−1
0
1
2
3
4
log10 (map−to−reference TPM)