Error, noise and bias in de novo transcriptome assemblies

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

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

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

* Corresponding author: adamfreedman@fas.harvard.edu

ABSTRACT

De novo transcriptome assembly is a powerful tool, widely used over the last decade for making evolutionary inferences. However, it relies on two implicit, untested assumptions: that the assembled transcriptome represents an unbiased, if incomplete, representation of the 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. Using standard filtering approaches, coverage of annotated genes by transcriptome assemblies falls far below complete coverage, even at the highest expression levels, and hundreds of expressed genes are not assembled by any of the algorithms we evaluated. Assemblies are compositionally biased towards non-coding elements relative to their frequency in reads. Bias also exists at the nucleotide level, with genotyping error rates ranging from 30-87%. As a result, diversity is underestimated in transcriptome assemblies, with consistent under-estimation of heterozygosity in all but the most inbred sample we assayed. Finally, even at the gene level and when one sums across contigs annotated to those genes, expression estimates show wide deviations from map-to-reference estimates, and positive bias at lower expression levels. We demonstrate a computational method to partly alleviate noise and bias in expression estimates. Genome assembly, where feasible, may alleviate some of these problems for non-model organisms, although genome assemblies are themselves imperfect in many ways. Researchers should consider ways to minimize the impact of bias in transcriptome assemblies.

36 **1. INTRODUCTION**

37 Although early advances fueled by RNA-seq were restricted to genome-enabled species, the
38 advent of computational methods for de novo transcriptome assembly (Birol et al., 2009;
39 Grabherr et al., 2011a) has empowered a wide array of evolutionary analyses in the absence of a
40 reference genome. Their application has led to significant advances in fields as diverse as
41 phylogenetics (Schwentner, Combosch, Pakes Nelson, & Giribet, 2017), evolutionary responses
42 to climate change (Veilleux et al., 2015), the molecular basis of convergent evolution (Berens,
43 Hunt, & Toth, 2015; Pankey, Minin, Imholte, Suchard, & Oakley, 2014; Velotta et al., 2017),
44 speciation (K. Li et al., 2016), and local adaptation (De Wit & Palumbi, 2013).

45 All studies utilizing de novo transcriptome assemblies rely on two implicit assumptions:
46 that the assembled transcriptome represents an unbiased, if incomplete, representation of the true
47 underlying expressed transcriptome, and that the expression estimates from the assembly are
48 good, if noisy approximations of the relative abundance of expressed transcripts. These
49 assumptions are rarely tested. When tests are carried out, they typically employ reads simulated
50 from curated transcript annotations that represent idealized data (e.g. Vijay, Poelstra, Künstner,
51 & Wolf, 2013) . Similarly, empirical investigations typically rely on metrics of contiguity such
52 as N_{50}, completeness with respect to the proportion of conserved orthologs recovered, i.e.
53 BUSCOs (Simão, Waterhouse, Ioannidis, Kriventseva, & Zdobnov, 2015), and dimensionless
54 quality metrics used to guide assembly parameter tuning using tools such as DETONATE (B. Li et
55 al., 2014) or TRANSRATE (Smith-Unna, Boursnell, Patro, Hibberd, & Kelly, 2016). While these
56 are standard metrics used to benchmark new assembly tools, they do not explicitly measure or
57 test the robustness of expression estimates derived from the assembled contigs, or whether

58 genotypes, allele frequencies and population genetic parameters that are important for
59 evolutionary inference can be accurately estimated from a transcriptome assembly.
60 Indirect evidence suggests that violations of the assumptions noted above may be
61 frequent. Comparisons of different assembly algorithms indicate that there is a degree of
62 complementarity in missingness—one assembler will recover a gene that another cannot,
63 motivating methods to combine assemblies, e.g. the Oyster River Protocol (MacManes, 2018).
64 While gene absence is often attributed to the difficulties of assembling lowly expressed
65 transcripts, little is known about that kinds of genes that are missing, whether expression level is
66 their only important unifying feature, and whether they are biased towards functional categories
67 of interest to evolutionary ecologists. Second, highly fragmented assemblies comprised of
68 hundreds of thousands of relatively short contigs are pervasive in the literature, presumably
69 representing transcript fragments, haplotypes erroneously designated as separate transcripts, or
70 both. Intuitively, the challenge of mapping reads to fragmented gene models should diminish the
71 robustness of expression estimates. Unknown is whether the compositional aspects of
72 transcriptome assemblies negatively impact SNP calls and the allele frequency estimates and
73 other population genetic parameters derived from them.
74 Below, we evaluate the robustness of the assumptions inherent to de novo transcriptome
75 assembly, how their violation might distort evolutionary inferences, and identify particular
76 features of transcriptome assemblies that should be considered during the study design phase of a
77 project. Where possible, we demonstrate methods that ameliorate assumption violations and
78 augment the robustness of downstream results. While phylogenomic studies can discard the
79 majority of contigs using aggressive filtering schemes and still obtain recover a large number of
80 phylogenetically informative sites in conserved orthologs, these aggressive filtering options are

81 less appropriate for studies that seek to understand patterns of genetic variation or gene
82 expression across populations or closely related species. Therefore, we focus on methodological
83 considerations for this class of investigations.
84
85 **2. MATERIALS AND METHODS**
86 **2.1 Reference genomes**
87 For the purpose of benchmarking de novo transcriptome assemblies, and comparing assembly
88 and reference-based expression estimates, we downloaded from ENSEMBL genome and gtf
89 annotation files for the following organisms: house mouse, *Mus musculus* C57BL/6J
90 (GRCm38); clawed frog, *Xenopus tropicalis* (JGI_4.2); pufferfish, *Tetraodon nigroviridis*
91 (TETRAODON8); and fly, *Drosophila melanogaster* (BDGP6).
92
93 **2.2 RNA-seq data**
94 Brain tissue is both transcriptionally complex and expected to express a large proportion of an
95 organism's overall transcriptional profile. For this reason, the bulk of our analyses focus on de
96 novo assemblies for data generated for experiments involving mouse (*Mus* spp.) brain. These
97 data sets include the *Mus musculus* (C57BL/6J) dendritic cell data used in the original Trinity
98 paper (MDC), a pool of six whole brains from albino inbred *Mus* (BALB/c), and 8-sample pools
99 of whole brain samples from wild *M. musculus domesticus* from Massif Central, France (FRA),
100 Iran (IRN), Kazakhstan (KZK), and Germany (DEU). To assess the generality or particular
101 results with respect to assembly composition, we also generated assemblies for pufferfish
102 (*Tetraodon nigroviridis*) whole brain, clawed frog (*Xenopus tropicalis*) kidney, and fly
103 (*Drosophila melanogaster*) heads. Data set SRA accessions, sequencing strategy and depth are

104 summarized in Supplementary Table 1. All libraries except for MDC were sequenced on an
105 Illumina HiSeq 2000; MDC was sequenced on an Illumina GAII.
106
107 **2.3 Short read processing**
108 After an initial assessment of sequences reads with FASTQC
109 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), we performed kmer-based error
110 correction with RCORRECTOR (Song & Florea, 2015), followed by adapter and quality trimming
111 with TRIMGALORE! (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), setting
112 both the minimum phred quality score and adapter filtering stringency parameter to 5, and the
113 minimum post-trimming read length to 36. We based our quality trimming threshold upon
114 previous findings that performing stricter quality filtering can negatively impact assemblies
115 (MacManes, 2014). A second run of FASTQC was used to confirm that there was no remaining
116 adapter sequence in the filtered reads.
117
118 **2.4 De novo transcriptome assembly**
119 We built de novo transcriptome assemblies with TRINITY v. 2.1.1 (Grabherr et al., 2011b) using
120 default settings, and did not perform *in silico* normalization for any of the data sets. For the *Mus*
121 data sets, to evaluate the robustness of some key findings to the algorithm employed, we also
122 built assemblies using two promising, recently develop assemblers: SHANNON v. 0.0.2 (Kannan,
123 Hui, Mazooji, Pachter, & Tse, 2016) and BINPACKER v. 1.0 (Liu et al., 2016). Where strand-
124 specific reads were available, we ran TRINITY and BINPACKER with the appropriate strand-
125 specific options. Because the strand-specific mode of SHANNON is in experimental phase, on the

126 recommendation of its author (Kannan, personal communication), we treated reads as un-
127 stranded in all SHANNON assemblies.
128 To perform an initial assessment of assembly quality, we calculated standard metrics
129 typically reported in de novo transcriptome studies. First, we quantified the total number of
130 contigs ("transcripts" in TRINITY's nomenclature), the number of TRINITY-defined "genes"
131 comprising contigs deriving from the same graph, N_{50}, and the percent of contigs that were
132 successfully mapped to the genome. We then assessed read support for the assembly, by
133 mapping reads back to the assembly to assess the proportion that mapped concordantly, using
134 methods specified by the developers of TRINITY
135 (https://github.com/trinityrnaseq/trinityrnaseq/wiki/RNA-Seq-Read-Representation-by-Trinity-
136 Assembly). Finally, we quantified the number of completely covered, fragmented, and missing
137 single copy conserved orthologs using the program BUSCO v. 3.0.2 (Simão et al., 2015), using the
138 vertebrate ortholog model set for all assemblies but the fly, for which we used the arthropod set.
139

140 **2.5 Assembly and read functional composition**

141 For each contig within an assembly and for the assembly as a whole, we quantified the
142 proportion of unique assembly nucleotides that were CDS, UTR, intronic, start or stop codon,
143 noncoding elements (including pseudogenic transcripts of genes with
144 "polymorphic_pseudogene" gene biotype), introns of non-coding transcripts, and
145 intergenic/unnanotated sequence. We first mapped contigs to their respective reference genomes
146 with GMAP v. 2016-06-30 (Wu & Watanabe, 2005). We then filtered out unmapped contigs with
147 SAMTOOLS v. 1.5. (H. Li et al., 2009). For the analysis of whole assemblies, we then used
148 BEDTOOLS v. 2.25 (Quinlan & Hall, 2010) to sort, and merge intervals such that each reported

149 interval consistent of unique genomic nucleotides. Mapped intervals for GMAP alignments were
150 converted to bed format with the bamToBed function, with the '-splitD' argument such that each
151 mapped interval for a contig—excluding alignment gaps corresponding to putative splicing
152 events—was converted into a separate row.

153 To calculate the proportions of transcriptome assemblies corresponding to particular
154 functional elements, we employed a hierarchical classification of genomic nucleotides
155 (Supplementary Text). In order to determine if the functional composition of RNA-seq reads
156 differed from that of the transcriptome assemblies, we randomly sub-sampled one million paired-
157 end reads, and aligned them to their respective genomes using HISAT2 v. 2.0.4 (Kim, Langmead,
158 & Salzberg, 2015). We then performed the same hierarchical classification as was done for
159 assembled contigs, calculating frequencies for individual sequence fragments, i.e. across both
160 sequence reads when both were properly mapped. For multi-mapped reads, we calculated a
161 weighted average across all mappings. Visualization of functional composition was obtained by
162 randomly sub-sampling 100,000 reads that successfully aligned to the genome.
163

164 **2.6 Coverage of reference transcripts and genes**

165 After extracting reference transcripts from the genome annotation with RSEM v. 1.2.22 (B. Li &
166 Dewey, 2011a) , we mapped transcriptome assembly contigs to these transcripts using BLAT v.
167 3.1.6 (Kent, 2002). We required at least 95% identity in the alignment (but no minimum
168 alignment length in the query contig) to consider a hit valid. We quantified transcript coverage in
169 two ways. First, for each reference transcript, we identified the contig with the best alignment to
170 it, and calculated coverage as the sum of the BLAT alignment intervals. Given the minimal
171 evolutionary distance between query and target sequences, and because we required a minimum

172 of 95% identity for alignments, we did not penalize coverage for mismatches. Second, we
173 calculated the coverage of a reference transcript by summing coverage across all contigs with
174 alignments that passed our 95% percent identity criterion. Because a contig map produce valid
175 mappings to multiple transcripts of a gene, in calculating summed coverage we allowed a contig
176 to be used in coverage calculations for >1 reference transcript. Finally, we estimated gene-level
177 coverage as a TPM-weighted average of coverage for individual transcripts belonging to that
178 ENSEMBL gene id. We calculated these three coverage metrics for 1) an unfiltered transcriptome
179 assembly, and 2) for only protein coding transcripts with a minimum length of 200bp—the
180 default minimum contig length for TRINITY—using only contigs with transcripts per million
181 (TPM) ≥ 1, and with a predicted peptide sequence with a BLASTP hit against Uniref90
182 (downloaded 2017-04-05), We used TRANSDECODER v. 3.0.0
183 (https://github.com/TransDecoder/TransDecoder/wiki) to predict peptide sequences. For BLASTP
184 alignment, we set the maximum number of target sequences to one, and with a maximum
185 allowable e-value of 1×10^{-5}.
186

187 **2.7 Expression estimation**

188 For nine combinations of assembly method (TRINITY, SHANNNON, BINPACKER) and *Mus* sample
189 (MDC, BALB/c, Massif Central), we generated expression estimates with RSEM v. 1.2.22 (B. Li &
190 Dewey, 2011b) and BOWTIE2, v. 2.2.4 (Langmead & Salzberg, 2012). For TRINITY-based
191 analyses, isoform and gene level expression estimates were initially calculated using TRINITY-
192 specified gene and isoform identities for each contig. We also estimated "true" (reference-based)
193 expression by mapping reads directly onto reference transcripts, using the same RSEM/BOWTIE2
194 pipeline but with the ENSEMBL GTF annotation.

195 We assessed concordance between assembly and referenced-derived approaches in three
196 ways. First, we compared estimates for reference isoforms and their best hit BLAT contig in units
197 of counts per million (CPM), transcripts per million (TPM), and a corrected CPM that adjusts for
198 differences in effective length estimates between methods (see below, hereafter termed
199 weighted-adjusted CPM). We examine CPMs rather than raw counts because, despite using the
200 same RNA-seq data, differences in read mapping rates lead to differences in total mapped
201 counts, i.e. the estimates of library size by which differential expression algorithms normalize
202 individual samples. Second, we evaluated concordance between reference gene expression and
203 our best estimate of gene-level expression derived from the transcriptome assembly. Specifically,
204 we assigned assembly contigs to ENSEMBL genes based upon their BLAT alignments to ENSEMBL
205 isoforms, complied these assignments into a contig-to-gene map file, and provided this file to
206 RSEM when building the mapping index. RSEM *genes.results output files produced by this
207 protocol contain expression estimates summed across contigs aggregated by reference label; for
208 TRINITY, contigs without BLAT hits were aggregated by TRINITY components, i.e. "genes". For
209 SHANNON assemblies, the collection of contigs assembled from an individual graph do not
210 necessarily belong to the same gene (Kannan, personal communication). For all analyses
211 summing across contigs, Because preliminary analyses indicated that there were differences in
212 effective length estimates between the denovo assembly and map-to-reference approaches, and
213 because these lengths influence RNA abundance estimators, we assessed whether correcting for
214 effective length differences between reference transcripts and assembled contigs would improve
215 concordance between de novo assembly and map-to-reference based abundance estimates. We
216 did this by rescaling CPM as follows:

217 $$CPM_{wt-adj}\frac{map-to-ref\ effective\ gene\ length}{denovo\ effective\ gene\ length}\times\frac{denovo\ gene\ length}{map-to-ref\ gene\ length}\times 10^6$$

218

219 The first scalar adjusts for the difference in estimated effective length, while the second scalar
220 downweighs the correction as the denovo-based effective length approaches that obtained by
221 mapping to the reference annotation. In order to focus directly on structural aspects of
222 transcriptome assemblies (such as fragmentation and mis-assembly) that might negatively impact
223 expression estimate robustness, we only evaluated reference features to which at least one
224 transcriptome assembly contig could be aligned. In other words, we did not consider the trivial
225 case where expression is, practically speaking, zero when no contigs exist that can be used to
226 obtain counts.

227 Because it was suggested to us that RSEM might not perform as well when aligning to
228 transcriptome assemblies (N. Davidson, personal communication), we also compared gene-level
229 expression analyses between map-to-reference and map-to-denovo using expression values
230 estimated with KALLISTO v. 0.43.1 (Bray, Pimentel, Melsted, & Pachter, 2016). Because
231 KALLISTO operates on the transcript/contig level, we generated gene-level estimates by
232 aggregating to the gene level using the R package TXIMPORT (Soneson, Love, & Robinson,
233 2015).

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

238 $$log_2\left(\frac{sample\ 1\ CPM_{denovo}}{sample\ 1\ CPM_{map-to-ref}}\right) - log_2\left(\frac{sample\ 2\ CPM_{denovo}}{sample\ 2\ CPM_{map-to-ref}}\right),$$

239 which can be re-arranged as:

240 $$log_2\left(\frac{sample\ 1\ CPM_{denovo}}{sample\ 2\ CPM_{denovo}}\right) - log_2\left(\frac{sample\ 1\ CPM_{ma-to-ref}}{sample\ 2\ CPM_{map-to-ref}}\right),$$

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

243 $$2^{EXP}\left[abs\left(log_2\left(\frac{sample\ 1\ CPM_{denovo}}{sample\ 2\ CPM_{denovo}}\right) - log_2\left(\frac{sample\ 1\ CPM_{ma-to-ref}}{sample\ 2\ CPM_{map-to-ref}}\right)\right)\right]$$

244 For example, given de novo CPM values for samples 1 and 2 of 15 and 20, and for map-to-ref 5
245 and 10, respectively, we would estimate that the magnitude of difference between samples would
246 be 1.5 times as great as if one had used the map-to-reference approach.
247

248 **2.8 Analysis of missing genes**

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

256 We performed two separate missingness analyses for each sample-assembler
257 combination. These involve differing levels of a) filtering on the assembly, and b) which
258 reference transcripts are considered as the "background" for enrichment analyses. For the first
259 analysis, we included all assembled contigs and their respective BLAT alignments, and used all
260 genes with at least one expressed transcript as background; this includes reference transcripts <
261 200 bp that Trinity, with default settings, will not assemble, as well as non-coding transcripts
262 (hereafter termed "$\text{background}_{\text{unfiltered}}$"). Second, we only retained contigs with TPM ≥1, and
263 restricted our background to protein-coding reference transcripts ≥ 200 bp ("$\text{background}_{\text{filtered}}$").

264 In both analyses, we only included in the background gene set those genes with a reasonable
265 probability of detection, which we defined as TPM ≥1 for at least one of a gene's constituent
266 isoforms.

267 To determine if particular functional categories were over-represented in genes that were
268 missing, we conducted a statistical overrepresentation test using the PANTHER classification
269 system (Mi, Muruganujan, Casagrande, & Thomas, 2013), via http://www.pantherdb.org/. We
270 used the ENSEMBL gene ids associated with the expressed reference transcripts as our "reference
271 list" and the ids of the missing genes as the "analyzed list", calculating statistical significance of
272 over/under-representation using a Fisher's Exact test.

273 To determine whether genes that are missing from assemblies are less conserved than
274 those that are present, we downloaded from the UCSC Genome Browser
275 (http://hgdownload.cse.ucsc.edu/goldenpath/mm10/phastCons60way/) the phastCons (Siepel,
276 2005) score bigWig track derived from multiple alignments of 59 vertebrate genomes to the
277 mouse genome (mm10). We then converted from bigWig format to bedGraph with the
278 KENTUTILS (https://genome-source.gi.ucsc.edu/gitlist/kent.git/blob/master/README)
279 bigWigToBedGraph tool. We calculated mean phastCons scores per ENSEMBL gene (defined as
280 the set of non-redundant CDS, start codon, 3` UTR, 5` UTR, stop codon, and exons associated
281 with an Ensembl gene ID). We compared phastCons scores of genes missing from the assembly
282 to actually expressed genes using a non-parametric Wilcoxon rank sum test.
283

284 **2.9 Evaluating assembly redundancy**

285 To evaluate contig redundancy, we first quantified the frequency and degree of overlaps between
286 contigs whose best BLAT hits are known, single-isoform genes in *Mus*. For four of these data

287 sets, there are independent estimates of genetic diversity obtained from genome-sequencing
288 (Harr et al., 2016), so we predicted that, if false assembly of haplotypes increased with genetic
289 variation, that the frequency of overlaps would correlate with diversity. We also calculated the
290 redundancy from BLAT all-by-all alignments of individual assemblies. If haplotypes are
291 assembled into different contigs, or contigs are assembled that only differ by a small number of
292 bases, e.g. by differences in assembled intronic sequence, this should manifest in an excess of
293 cases where alignments cover nearly all of the constituent contigs in those alignments, with few
294 or no mismatches.
295
296 **2.10 Evaluation of variant calls from transcriptome assemblies**
297 For these analyses, we focused on the *Mus* samples, and compared the map-to-reference
298 genotypes (hereafter MR) to those derived from TRINITY assemblies. We called variants off of
299 the reference genome using the GATK v. 4 (McKenna et al., 2010), and the Broad Institute's best
300 practices for RNA-seq data
301 (https://software.broadinstitute.org/gatk/documentation/article.php?id=3891; last updated 2017-
302 01-21). For the *Mus* samples representing pools, following the GATK developers'
303 recommendations, we also used the --sample_poidy argument, with a value set to two times the
304 number of individuals in the pool. For TRINITY assemblies, we first constructed
305 SuperTranscripts, hereafter (ST), (Davidson, Hawkins, & Oshlack, 2017), which attempt to
306 collapse overlapping contigs within Trinity "genes", and then called variants with the GATK
307 pipeline, based upon python scripts provided with TRINITY v. 2.5.1. Similar to our reference-based
308 approach to obtain our truth set, we edited the variant calling script to include the --
309 sample_ploidy argument. We did not use either base quality score or variant quality score

310 recalibration, to reflect the fact that the truth sets needed for these methods are rarely, if ever,
311 available for ecological genomics data. We applied the same strand bias, depth, and quality-by-
312 depth filters to raw variant calls from both methods to produce a final filtered set. To define
313 callable sites for the MR approach, we extracted read alignments to annotated exonic regions,
314 and then calculated the number of exonic sites with a coverage depth ≥ 5. For ST, we performed
315 similar operations, defining exonic coordinates from mappings of ST to the *Mus* genome using
316 GMAP, and callable sites within exons in the same way as with MR, i.e. coverage depth ≥ 5. We
317 defined heterozygous calls in our filtered vcf files as all bi-allelic SNP sites that passed filtering
318 criteria. In cases where sample ploidy was greater than diploid, we collapsed genotypes into the
319 union of all unique recorded alleles.

320 We obtained exonic regions for ST by mapping them to the *Mus* genome using GMAP, then
321 intersected these mappings with a bed file of exonic intervals with BEDTOOLS. We used a
322 slightly newer version of GMAP (v. 2018-07-04) than we did for the contig composition, which
323 fixed minor issues with CIGAR string handling that had the potential to generate spurious
324 genotype errors. Python scripts
325 (https://github.com/harvardinformatics/TranscriptomeAssemblyEvaluation) were then used make
326 transformations between genomic and ST coordinates, for the purpose of extracting ST
327 genotypes and writing genotype bed files back to genomic coordinates. ST alignments to the
328 genome were used to define unique polymorphic site ids that included cases where (a) both
329 methods called genotypes, (b) only MR called a genotype (including sites with and without valid
330 ST alignments), or (c) genotypes were only called from ST. We recorded cases where multiple
331 ST were aligned to the same genomic position, taking the union of all reported alleles as the ST
332 genotype for that position. We restricted our comparison of the two methods by filtering out MR

333 sites that were indels or contained > 2 alleles. Using genotype calls based upon the reference
334 annotation as the gold standard, we calculated 15 different metrics, defined as follows:

335 - **Error rate**: the proportion of ids where the set of alleles recorded by the two methods
336 were not the same, including cases where one method did not call a genotype; *total # of*
337 *incorrect ST calls aligned to genomic positions ÷ total # of genomic positions to which*
338 *STs were aligned*
339 - **Error-MrefAlleleIncluded**: the proportion of genotyping errors (in ST) where the
340 reported alleles included at least one of those reported using MR; *total # of incorrect ST*
341 *calls relative to MR, but that include on MR allele ÷ total # MR calls to which STs were*
342 *aligned*
343 - **Error-Snv2Indel**: the proportion of errors, where a biallelic SNV in MR was called as an
344 indel in ST; *# of ST indel calls at biallelic MR SNV sites ÷ total # bi-allelic MR SNV*
345 *calls*
346 - **Error-MultipleSuperTranscripts**: the proportion of errors where multiple STs were
347 aligned to the genomic position; *total # of ST genotyping errors at MR sites with multiple*
348 *ST alignments ÷ total # MR genotype calls with ST alignments*
349 - **FN**: false negative rate, the proportion of MR variable sites with no ST genotype; *# MR*
350 *variable sites with either no ST alignment or with an alignment but no genotype call ÷ #*
351 *MR variable sites*
352 - **False negatives-noalign**: the proportion of false negatives that are due to the absence of
353 ST alignments to the genome at the variable position; *# MR variable sites with no ST*
354 *alignment ÷ total # MR bi-allelic SNV sites*

355 • **FN-het**: the proportion of heterozygous MR genotype calls with either no ST genotype, a
356 fixed alternative (relative to ST) genotype, an indel heterozygote call, or a multi-allelic
357 ST genotype. ST SNV heterozygous genotypes that specify the wrong alleles are not
358 considered FN, which is relevant for comparisons of transcriptome-level heterozygosity
359 between methods*; # of bi-allelic SNV MR genotype calls with no corresponding ST SNV*
360 *biallelic genotype ÷ total # bi-allelic SNV MR genotype calls*
361 • **FP**: false positive rate, the proportion of ST genotypes with either no associated MR
362 genotype, or where ST is polymorphic and MR is fixed for an alternative allele; *# MR*
363 *invariant sites called as polymorphic with ST ÷ total # invariable MR sites*
364 • **FP-het**: the proportion of invariant MR sites that are represented by heterozygous bi-
365 allelic ST genotypes, i.e. with either no associated MR genotype or a fixed alternative
366 MR genotype; *# invariant MR sites with ST heterozygote genotypes ÷ total # invariant*
367 *MR sites*
368 • **FP-NAtoFixed**: the proportion of MR sites with no genotype call that are called as a
369 fixed variant in ST (relative to the ST reference); *# fixed-alternative ST genotypes at MR*
370 *sites with no genotype call ÷ total # sites with no MR genotype call*
371 • **FP-NAtoHet**: the proportion of MR sites with no genotype call that are called as
372 heterozygotes in ST ; *# heterozygote ST genotypes at MR sites with no genotype call ÷*
373 *total # of sites with no MR genotype call*
374 • **FP-indel**: the proportion of ST genotype calls at MR invariant sites (both missing and
375 MR-fixed) that are indel; *# invariant MR sites called as indels with ST ÷ # MR invariant*
376 *sites with ST genotype calls*

377 • **Recall:** the proportion of MR genotypes recovered by ST, including matching invariant
378 sites; (*# genomic positions with MR genotype calls with matching ST calls + matching*
379 *invariant MR and ST calls)÷(total # MR called genotypes +sites with matching invariant*
380 *calls),* i.e. *#TP/(#TP+#FN), where TP = true positives.*
381 • Recall-het: the proportion of MR heterozygous genotypes that are correctly recovered in
382 ST; a genotype is not considered recovered if ST genotype is a heterozygous call
383 comprised of the wrong nucleotides; *# MR bi-allelic heterozygote calls with matching ST*
384 *heterozygote calls ÷total # MR bi-allelic heterozygous calls*
385 • Precision: the proportion of ST genotypes that match the MR genotype; *# ST genotype*
386 *calls with a matching MR genotype call ÷total # ST genotype calls, i.e. TP/TP+FP*
387

388 **2.11 Analysis reproducibility**

389 Software command lines, scripts and accessory files used to implement analyses described above
390 are available at https://github.com/harvardinformatics/TranscriptomeAssemblyEvaluation.
391

392 **3. RESULTS**

393 **3.1 Assemblies**

394 We assembled de novo a total of 21 transcriptomes—six mouse brain data sets using TRINITY,
395 SHANNON, and BINPACKER, as well as TRINITY assemblies for frog kidney, pufferfish brain and
396 fly head. Consistent with previously published reports, assemblies were highly fragmented, with
397 median contig sizes ranging from just a few hundred bases to slightly over 1kb (Supplementary
398 Table 2). Yet, typically employed measures of accuracy and completeness indicate that all
399 assemblies were of reasonably high quality (Supplementary Tables 2-4), particularly those

400 assembled from the *Mus* data sets comprising the majority of our analyses. Important in the
401 context of gene expression studies, large fractions of reads used to build assemblies mapped back
402 to them in a concordant fashion, usually exceeding 90% of the input library (Supplementary
403 Table 3)
404
405 **3.2 Transcript and gene coverage**
406 Without filtering on the reference transcriptome or the de novo assembly, transcript coverage is
407 increased when one sums coverage across orthologous assembly contigs (Figure 1a-b,
408 Supplementary Figure 1-3). In both methods, and at the level of genes (Figure 1c, Supplementary
409 Figures 1-3), coverage generally tracks reference expression level, although these correlations
410 are weaker at the transcript level, and weakest for coverage of a single best-hit contig. For
411 summed coverage of transcripts and gene coverage, while there is an expectation of complete
412 coverage above TPM=10, there are still many transcripts and, to a lesser extent, genes that fall
413 far below complete coverage. This "unfiltered" approach represents an idealized scenario in
414 which contigs are mapped directly to a reference annotation originating from the same species. If
415 one considers only protein-coding genes and transcripts ≥ 200bp—which should favor assembly
416 coverage—but takes a standard approach of filtering out lowly expressed contigs, and only
417 considering contigs with hits to UNIREF90 (a standard approach for assigning orthology without
418 an annotated genome)—coverage decreases substantially at both the transcript and gene level
419 (Figure 1d-f, Supplementary Figures 1-3). The expectation of coverage never reaches 100%,
420 regardless of expression level of the underlying reference structures, and there is far more
421 variance around the coverage-TPM trend. The decreases observed in these trends at the highest

422 expression levels result from poorly covered annotations making up an increasing fraction of the
423 most highly expressed transcripts and genes.
424
425 **3.3 Missing genes**
426 Only considering genes that are expressed at a level to permit a reasonable chance of being
427 detected (TPM ≥ 1), depending upon the sample, 17-78 (0.015-0.052%) of genes were missing
428 from all methods' unfiltered assemblies, and 25-115 (0.23-0.77%) genes were unrecovered by 1-
429 2 assembly methods (Table 1, Supplementary Figure 4a). Practically, unfiltered assemblies are
430 rarely, if ever, useable, so to test a more realistic scenario we also estimated the number of genes
431 missing from filtered assemblies. Missing from TRINITY assemblies were 3.5% (293), 2.6%
432 (277), and 4.7% (529) protein coding genes for MDC, BALB/c and FRA, respectively. Numbers
433 were substantially worse for SHANNON (MDC, 19.1% (1585); BALB/c, 9.9% (1039); FRA, 7.5%
434 (836)), and better in two of three samples for BINPACKER (MDC, 3.0% (251); BALB/c 2.9%
435 (306); FRA 3.0% (340)). Overall, any one assembler-sample combination led to hundreds to over
436 1,000 expressed genes being missed, at least after standard assembly filtering, which
437 unavoidably removes both spurious contigs and contigs representing real transcripts. In general,
438 genes not recovered by one method tended to be missing from all assemblers (Table 1,
439 Supplementary Figure 4b); SHANNON tended to show greater missingness for protein coding
440 genes than TRINITY and BINPACKER. With the exception of MDC, genes missed by all or some
441 assemblers had significantly lower reference gene expression and conservation (Figure 2,
442 Supplementary Figure 5, Table 2). Even so, there were plenty of cases where conserved genes
443 expressed at a relatively high level (TPM ≥ 100) were not recovered by any assembler (Figure
444 2). Correlations between TPM and phastCons scores were, while statistically significant,

445 relatively weak (MDC, Spearman's $\rho = 0.082$, $p = 2.2 \times 10^{-16}$; BALB/c, Spearman's $\rho = 0.18$, p
446 $= 2.2 \times 10^{-16}$; FRA, Spearman's $\rho = 0.23$, $p = 2.2 \times 10^{-16}$). Missing genes were not enriched for
447 any GO-Slim Biological Process terms.
448
449 **3.4 Assembly vs. read functional composition**
450 For a gene to be recovered in an assembly, it has to be annotated. And, because transcriptome
451 assembly is motivated by the absence of a closely related reference genome, in nearly all cases
452 transcriptome annotation involves predicting open reading frames and searching protein
453 databases such as UNIREF90 and SWISSPROT. The greater the fraction of non-coding bases in a
454 contig, the harder it will be to align it to an annotated peptide. Thus, the functional composition
455 of contigs sheds light on patterns of missingness at both the transcript and gene levels.
456 Across all sample-assembler combinations, large fractions of contigs are dominated by intronic,
457 UTR, and intergenic sequence (Figure 3, Supplementary Figure 9). The extent of intronic
458 sequence is particularly surprising for *Mus* and *Drosophila* assemblies, as current annotations
459 should be inclusive of the majority of alternative splice variants. That the excess of intronic
460 sequence in contigs is not reflected in the composition of reads (Figure 3, Supplementary Figure
461 6) can be attributed to the sensitivity of the assembly algorithms, which require minimal kmer
462 coverage from just a few intron-containing reads to extend contigs across exon boundaries. The
463 prevalence of intronic and intergenic sequence in assembly contigs indicate that a substantial
464 proportion of assembled bases are derived from some combination of transcriptional read-
465 through across exon junctions, assembly of low-abundance pre-mRNA, and low levels of
466 spurious genome-wide transcription (Supplementary figure 7). The mean contig fraction
467 comprised of intron sequence for TRINITY assemblies ranged from 7.5% (*Xenopus*) to 56.5%
468 (FRA) (Supplementary table 5). Mean fractions for intergenic sequence ranged from 8.1%

469 (MDC) to 39.0% (*Xenopus*). Results for SHANNON and BINPACKER assemblies of MDC,
470 BALB/c and FRA were similar to those observed in TRINITY.

471 Redundant assembly of identical contigs from and assembly of different haplotypes as
472 unique contigs can create complications for downstream analyses. Contig multiplicity can
473 influence downward expression estimates of contig copies—particularly if redundant contigs are
474 not properly assigned to the same gene-level annotation—and decrease the signal-to-noise ratio
475 for differential expression analyses. Assembling haplotypes as contigs can lead to false positive
476 differential expression results, e.g. if frequencies of haplotypes differ between populations. It can
477 also distort allele frequency spectra that are critical for analyses of allele-specific expression.
478 Across all three assemblers, the frequency of contig overlaps within known single-isoform genes
479 was bimodal. Most genes were overlapped at few or no positions by ≥ 1 contig, but a large
480 number of genes were also covered by multiple contigs at most or all of their positions (Figure
481 4). SHANNON, in particular, showed a troubling increasing trend in frequency with proportional
482 reference gene coverage (Figure 4b), indicating it may be more prone to redundancy than either
483 TRINITY or BINPACKER. Overall, 11.9%-12.2%, 33.2%-35.3%, and 16.9%-18.7% of single
484 isoform genes were covered by more than one contig at all transcript positions for TRINITY,
485 SHANNON, and BINPACKER, respectively. The frequency of contig overlaps has no relationship
486 with estimates of genetic diversity obtained from whole genome sequencing (Figure 4), a pattern
487 not to be expected to occur if redundancy is driven solely by the assembly of polymorphic
488 haplotypes into separate contigs. Considering the variation in CDS and intron content among
489 contigs that cover a single-isoform gene, while there are plenty of cases with variable assembly
490 of CDS among contigs, the event of highest frequency across all assemblers appears to be one
491 where there is no variation in CDS among contigs but variable intron retention, due to either

492 transcription of intron sequence or the inclusion of un-spliced pre-mRNA in sequencing libraries
493 (Supplemental Figure 8). Looking across all contigs, regardless of their genomic position of
494 origin, all assemblies contain large numbers of contigs that are nearly identical (Figure 5a,
495 Supplemental Figure 9). For TRINITY assemblies that putatively report "genes" as components,
496 while the redundancy is greater for contigs assembled from the same graph, there are a
497 substantial number of contigs with high similarity that originate from different TRINITY "genes"
498 (Supplementary Figure 10). From different component graphs, Trinity will erroneously assemble
499 haplotypes that differ only by a few SNVs (Figure 5b).
500

501 **3.5 SNP calling and heterozygosity**

502 SNP calls and allele frequencies inferred from them and that are derived from transcriptome
503 assemblies have been used in a variety of contexts, including estimation genetic diversity in
504 evolutionary analyses (Romiguier et al., 2014), population structure (Tepolt & Palumbi, 2015),
505 reconstruction of demographic history and gene flow (De Wit & Palumbi, 2013; Le Gac et al.,
506 2016; Roda, Mendes, Hahn, & Hopkins, 2017), estimation of dN/dS (Cokus, Gugger, & Sork,
507 2015), F_{ST} outlier analysis (Bankers et al., 2017) , and tests for allele-specific expression (Singh
508 et al., 2018). Our ST mapped at extremely high rates to the *Mus* genome (MDC, 99.9%;
509 BALB/c, 99.8%; FRA, 99.8%; DEU, 99.8%; IRN, 99.8%; KZK, 99.7%), such that any
510 discrepancies between genotyping methods cannot be attributed to low alignment efficiency. In
511 principle, redundancy through assembly of haplotypes as distinct contigs should lead to
512 erroneous genotype calls and biased estimates of allele frequency spectra. For example, when
513 two contigs represent haplotypes: at a variant position, the true heterozygous position is instead
514 represented as two separate genotypes, one homozygous for the reference allele and one

515 homozygous for the alternative allele. The opposite problem can happen when paralogous loci
516 are collapsed, leading to a heterozygous genotype call where there would otherwise be two
517 invariant sites, or perhaps a mixture of alleles due to, for example, collapsing of an invariant and
518 a heterozygous position. We compared genotype calls and derived diversity metrics between the
519 map-to-reference and map-to-transcriptome approaches using 15 different metrics, in order to
520 determine whether such genotypic distortions occur in de novo transcriptome assemblies. The
521 map-to-reference (MR) approach involves mapping sequence reads to the genome, then calling
522 genotypes with GATK based upon those genomic alignments. In contrast, the transcriptome
523 assembly-based approach involves aligning reads to the ST collapse of a TRINITY assembly, then
524 calling variants with GATK relative to the ST. Comparisons between MR and ST genotypes are
525 based upon orthology determined by aligning the ST to the reference genome. To see example
526 cases of how particular errors are flagged, see Figure 6.

527 The frequency of errors, in which the allele set recovered by the two methods differed or
528 in which only one of the two methods called a genotype at all, was high, ranging from 30-38%
529 for the wild samples, and surprisingly, ≥ 50% for the inbred samples (Table 3). While the median
530 per-base error rate per ST was 0.2% or less, maximum per ST rates typically exceeded 10%
531 (Supplementary Table 6). For five of six samples, the median number of errors per ST was 2,
532 but, depending upon the sample, as many as 56-187 errors could occur within a single ST
533 (Supplementary Table 6). On a per-sample basis, errors show a modest yet statistically
534 significant bias towards the distal end of a ST (sample-level Wilcoxon 1-sided tests, H_0 =0.5, all
535 tests, $p \leq 0.002$; Supplementary Figure 11). Considering only contigs with an ORF hit to
536 UNIREF90 changed error rates in a predictable direction, e.g. reduced false positives and recall,
537 and increased false negatives (Table 3).

538 In 90-100% of all erroneous ST genotype calls, at least one MR allele was included in the
539 ST allele set (Table 3, Error-MrefAlleleIncluded). False negatives (FN; 24-66%) occurred far
540 more frequently than false positives (FP; 8-61%). Anywhere from 57-75% of FN were due to
541 there being no ST aligned at a genomic position where a MR genotype call was made (Table 3,
542 FN-noalign). Given that the ST mapping rates to the genome were extremely high, these FP
543 instances are primarily due to a failure to assembly transcripts overlapping polymorphic genomic
544 positions. The FP rate includes both sites where there is no genotype call in MR, or where MR is
545 fixed for an alternative allele and ST is polymorphic. Looking more closely, paralog collapsing
546 should lead to FP heterozygous ST calls, and FP ST fixed-alternative genotypes will be due local
547 assembly errors; in the latter case, mis-assembly is the most parsimonious explanation, given that
548 the ST reference sequences were derived from the data used to call genotypes. FP-NAtoHet was
549 5-38 times higher than FP-NAtoFixed, indicating that paralog collapsing plays a larger role in FP
550 than local assembly errors within ST. FP-NAtoHet includes ST indel heterozygotes, which are
551 not unexpected in cases of paralog collapse; The frequencies of cases where MR SNVs are called
552 as indels in ST (Error-Snv2Indel) approach 50% in the wild samples. This class of error as well
553 as the somewhat less frequent occurrence of false positive indel calls (FP-indel) collectively
554 suggest that paralog collapses are a general problem, if not as large a contributor to overall error
555 rates as FN. When genotypes were called in STs, the frequency of calls that were correct (i.e.
556 precision) ranged from 81-90% for the wild samples, but precision for the inbred samples,
557 particularly MDC, was substantially lower. Recall ranged from 73-76% in the wild samples; as
558 with precision, recall for the inbred MDC sample was extremely low.
559 Because read depth influences the ability of tools such as GATK to call genotypes, we
560 compared read depth at homologous positions, comparing MR depth and median ST depth (to

561 account for the possibility of > 1 ST mapping to a genomic position). Concordant genotypes
562 displayed a high degree of concordance in depth-of-coverage (Supplementary Figure 12). In
563 contrast, on average FN sites had very small reductions in ST read depth relative to MR at sites
564 where STs mapped to the genome (Supplementary Figure 13, blue lines). FN at sites with modest
565 loss of read depth is likely caused by multiple interacting factors that are beyond the scope of
566 this study to investigate in detail. In contrast, FN due to failure to assemble transcripts leads to an
567 effective ST read depth of zero, and including these sites leads to a much more dramatic
568 reduction in ST depth relative to MR (Supplementary Figure 13, yellow lines). With the
569 exception of a relatively small number of cases of zero MR coverage, MR and ST read depth
570 were largely co-linear (Supplementary Figure 14); the causes of noise about those correlations,
571 as with FNs, remain unclear.
572 One prediction resulting from such genotyping bias is that heterozygosity rates in ST will be
573 negatively biased relative to MR, because alterative alleles will be partitioned amongst different
574 assembly components (i.e. "genes") and reads overlapping these alleles will preferentially map to
575 their corresponding contigs, leading to separate homozygous reference genotype calls rather than
576 one heterozygous call. Consistent with this prediction, in all but the inbred (MDC, BALB/c)
577 strain samples, we observe reduced heterozygosity in the exonic regions of ST relative to MR
578 (Figure 7). This bias does not result from ST being preferentially assembled in regions of lower
579 heterozygosity, given that, for the wild samples, MR heterozygosity is actually higher in ST-
580 overlapping regions (Supplementary Figure 15). Furthermore, the bias cannot be due to some
581 unique feature of genotype calls in exons, as the same pattern holds when considering all
582 genotype calls (Supplementary Figure 16). Another possible explanation may be lower power to
583 detect low frequency alternative alleles with ST. However, in our five samples consisting of

584 pooled individuals, minor allele frequency at MR heterozygous sites did not vary between
585 correct ST heterozygous calls, FN, and FP (Supplementary Figure 17). The driving factor
586 appears to be FN at heterozygous sites; as already noted, FN are dominated by cases where ST
587 are not assembled (Table 3). An additional contributing factor may be erroneous indel
588 polymorphism genotypes in ST. Given that indel genotypes are harder to call correctly than
589 SNVs, a standard approach is to filter out indel sites. Because we perform the same filtering
590 when estimating heterozygosity, we effectively filter out many true heterozygous SNV positions,
591 further reducing heterozygosity estimates in ST. Paradoxically, FP genotype calls may actually
592 mask the magnitude of this bias.
593
594 **Gene and isoform expression**
595 The compositional deviations of de novo transcriptome assemblies relative to the true set of
596 expressed transcripts identified by mapping reads to a reference annotation results have a net
597 effect on expression estimates. Whether in terms of CPM or TPM, expression estimates derived
598 from the contigs that are the best BLAT hit relative to the reference transcripts are only weakly
599 correlated with map-to-reference estimates for those transcripts, and show positive bias for lowly
600 expressed reference transcripts (Supplementary Figure 18ab, Supplementary table 8). Weighted-
601 adjusted CPM correlations with reference isoform CPM led to virtually no change in
602 performance relative to other metrics (Supplementary tabler 8). Surprisingly, for TRINITY and
603 BINPACKER, contig weighted-adjusted CPM led to a substantially stronger correlation with
604 reference gene level expression (Supplementary Figure 18c, Supplementary Table 8), but fell
605 way below gene level correlations (see below and Table 4). Performance of SHANNON was
606 generally poor, and weighted-adjusted CPM correlation with reference gene expression was

607 similar to or even worse than for other abundance estimators.The practice of using single contigs
608 as proxies for gene-level expression is problematic, as contig-based estimates are also poor
609 estimators of reference gene expression. Since individual contigs typically represent a fraction of
610 the mappable sequence of a set of transcripts belonging to a gene, one expects smaller counts on
611 contigs relative to the full set of mappable positions encompassed by transcripts produced by a
612 gene locus. Given this fact, it is surprising that, for lowly expressed genes, their associated best-
613 hit contigs on average have larger counts (Supplementary Figure 18c). It is also surprising that
614 TPM, a metric that should account for effective length differences, shows a similar pattern of
615 bias (Supplementary Figure 18d).

616 Expression estimates derived from de novo transcriptome assemblies are far more robust
617 at the gene level, when one is able to group contigs into genes via annotation (Figure 8, Table 4).
618 It is reassuring that the correlation between the methods is quite strong, although there is sizeable
619 positive bias at lower expression levels for TPM, which is substantially reduced for CPM.
620 Furthermore, for many genes there are substantial deviations from estimates obtained with
621 mapping to the reference. Bias and deviations from the map-to-reference estimates are reduced
622 when one rescales CPMs to account for differences in effective length estimates (Figure 8c).
623 This indicates that effective length (EL) differences—that is, the difference in the number of start
624 sites for read alignments, given a particular sequenced fragment size, between a full-length
625 transcript, and a set of transcript sub-sequences represented by contigs—are part of the problem,
626 and in particular, that effective length is not properly estimated on the fragmented gene models
627 that are typical in denovo transcriptome assemblies. Correlations between reference gene
628 expression and gene-level summing across assembly contigs obtained with KALLISTO were very
629 similar to those obtained with RSEM (Table 5). Interestingly, adjusting for effective length had

630 little effect on KALLISTO CPM, and in some cases reduced the correlation between methods. This
631 suggests that KALLISTO may be doing a better job at estimating effective length on transcriptome
632 assemblies, such that it's inferior performance relative to RSEM must be due to other reasons.
633 As demonstrated earlier, high gene coverage is most frequently obtained for highly
634 expressed genes, but transcriptome assembly-derived expression estimates can still deviate
635 substantially from the map-to-reference gold standard at high coverage (Supplementary figure
636 21). Such deviation is well explained by the disparity in EL, with estimates showing the least
637 deviation where EL values are identical or nearly so (Figure 9), EL disparities persist at high
638 expression levels, even if those disparities are not as variable as at lower expression levels
639 (Figure 9). There are positive and negative trends in bias for CPM (Figure 9a) and TPM (Figure
640 9b), respectively, with the clearest bias pattern observed for TPM. Adjusting CPM for effective
641 length differences effectively eliminates biases that would otherwise be produced by differences
642 in EL (Figure 9c). Nevertheless, negative bias at lower expression levels and large deviations
643 from map-to-reference expression estimates are ubiquitous.
644 Biases in expression estimates derived from transcriptome assemblies might not have
645 large effects on downstream differential expression analyses if biases are consistent across
646 samples. While the pooled *Mus* data sets we used did not permit us to look at inter-individual
647 variation amongst individual libraries used to build a single assembly, we looked across the wild
648 *Mus* datasets to see if biases in gene-level expression were concordant, i.e. were biases for a
649 particular gene always of similar magnitude and in the same direction. We focus on results for
650 TRINITY assemblies of wild *Mus* samples, as TRINITY consistently showed stronger correlations
651 with map-to-reference based estimates, such that our findings are conservative with respect to
652 assemblers showing weaker correlations. Reassuringly, log-fold change of transcriptome

653 assembly vs. map-to-reference expression estimators (LFC_{METHOD}) showed broad patterns of
654 moderate concordance at the gene level indicating that gene-level features, influence the
655 magnitude of bias (Supplementary Figure 22). Inter-sample correlations were, for CPM, 0.625-
656 0.727; TPM, 0.790-0.830; and wt.-adj. CPM, 0.605-0.671. These correlations are weaker for
657 CPM and wt.-adj. CPM, in part, because biases are reduced overall, compressing LFC_{METHOD}
658 toward zero. Despite these broad correlations, a number of genes may show excessive bias. We
659 measured this as a bias effect size that indicates how much greater a difference in expression
660 would be detected using a denovo transcriptome approach, relative to mapping to reference
661 annotation. Comparing the FRA and DEU samples, for CPM, TPM, and wt-adj-CPM, 5.24%,
662 5.16% and 3.35% of genes showed bias effect sizes ≥ 5, with more than 2% of genes showing an
663 order of magnitude of difference between methods of ≥ 10 for CPM and TPM (Figure 10).
664 Correcting for effective length differences substantially reduced the frequency of such bias
665 outliers, but did not eliminate them entirely.
666
667 **DISCUSSION**
668 De novo transcriptome assemblies have been an important entry point for functional and
669 population genomics in non-model organisms, but the two key assumptions underlying their
670 application are consistently violated, albeit in unexpected ways. Below, we explain how
671 particular compositional aspects of de novo transcriptome assemblies may impact the power of
672 evolutionary inference.
673 While some genes not assembled by one algorithm are by another, dozens of genes will
674 consistently fail to get assembled. While this seems like a modest fraction of genes, this occurs
675 only in the idealized case where one can perform nucleotide-level searches against a transcript

676 database from a closely related organism, or in our case, the same organism. Using a common
677 approach that filters out lowly expressed contigs, and only retains those with a hit to a protein
678 database, we found that hundreds of protein coding genes (from 1-2%) would not be recovered
679 by any assembly method, and some assemblers missed several hundred to over 1,000 genes.
680 Thus, while almost all expressed genes may be in the assembly somewhere, annotation
681 challenges amplify missingness. Clearly, combining outputs from multiple assembly
682 algorithms—a non-trivial challenge in its own right—will not prevent missing a sizeable fraction
683 of genes. The efficiency of annotation will be further exacerbated by mis-assembly errors, such
684 as those revealed by indel errors within SuperTranscripts. As *Mus* proteins are well represented
685 in UNIREF90, our finding is overly optimistic relative to the typical predicament molecular
686 ecologists find themselves in. Annotation problems should be greatest for more rapidly evolving
687 genes as they will be more divergent from currently archived protein and nucleotide sequences.
688 Previous studies have confirmed that increasing divergence to annotation databases increases
689 annotation errors (Hornett & Wheat, 2012; Ockendon et al., 2016; Vijay et al., 2013), and these
690 issues should be more precisely considered with respect to both taxon divergence and the
691 evolutionary rate of genes within pathways of interest.
692 The difficulty in assembling genes that show less conservation is due, in part, to their
693 being lowly expressed. Whether due to the difficulty assembling contigs from few reads, or
694 polymorphism in more rapidly evolving genes, the fact remains that genes tied to phenotypes of
695 interest may be missing from transcriptome assemblies. To the extent that expression level drives
696 missingness, conserved, lowly expressed genes such as transcription factors that produce large
697 phenotypic effects may also be missing from transcriptome assemblies more often than expected
698 by chance. Molecular ecologists must not confuse expression level with the capacity to drive

699 phenotypic evolution and should find means to estimate the probability that genes of interest are
700 likely to be recovered in a de novo transcriptome assembly.

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

712 The nucleotide composition of contigs sheds light on how genes may wind up missing
713 from a de novo transcriptome assembly. We found a relatively high frequency of non-coding
714 bases in assembly contigs, which by definition, places constraints on the utility of standard
715 protein database searches. To the extent that an underlying CDS is divided into a series of
716 contigs, power to discover a protein homolog will be reduced; for contigs dominated by non-
717 coding bases containing only short strings of CDS, successful ORF prediction and annotation is
718 unlikely.

719 We expect that missingness both actual and annotation-based, combined with the high
720 frequency of sequence reads dominated by UTR bases, will also reduce power for differential
721 expression studies in a previously unappreciated way. Practically speaking, one throws away all

722 reads that map to unannotated contigs in addition to unmapped reads. Rules of thumb for
723 sequencing depth sufficient for differential expression analyses are largely derived from map-to-
724 reference studies, and do not consider this additional source of annotation-conditional filtering.
725 Loss of coverage depth is most problematic for evolutionary studies of non-model organisms,
726 given that they frequently employ small numbers of biological replicates (Todd, Black, &
727 Gemmell, 2016) which places an additional constraint upon statistical power. We suggest that
728 researchers employing transcriptome assemblies determine the number of reads per sample that
729 remain after those mapping to unannotated contigs have been discarded and consider additional
730 sequencing if read counts far well below desired targets. A second possibility is to "rescue"
731 unannotated contigs, by finding ways to cluster them with annotated contigs. Standard clustering
732 approaches such as CD-HIT will not perform well in this context, as they operate on sequence
733 similarity. By definition, a contig derived mostly from UTR will share little sequence with an
734 annotated contig comprised mostly of coding sequence. Underlying this type of phenomenon is
735 the fact that assembly graphs will often get broken into pieces deriving from different functional
736 compartments. For example, a short repeat in a UTR can split a transcript into an exclusively
737 UTR contig, and a contig comprised of a few UTR bases plus the down or upstream CDS.
738 Recently developed methods that leverage information from equivalence classes obtained from
739 pseudo-alignments of paired-end reads, such as CORSET (Davidson & Oshlack, 2014) and
740 GROUPER (Davidson & Oshlack, 2014), may help mitigate the loss of unannotated contigs. In
741 particular, GROUPER utilizes "orphan" read pairs that map to different contigs, which should
742 enable clustering of contigs representing transcript fragments that share little or no sequence.
743 Integration of this type of approach with annotation-based clustering is a promising avenue that
744 requires further investigation.

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

759 Even when annotation-based groupings of contigs are performed in an optimal setting,
760 such as our analyses of *Mus* transcriptomes, there is substantial noise in gene-based expression
761 estimates derived from transcriptome assemblies relative to a map-to-reference based approach,
762 and there is positive bias at lower expression levels. Furthermore, the practice of selecting a
763 single best hit contig to represent a gene seems ill-advised given that correlations between contig
764 and map-to-reference gene-level expression are poor, and subject not only to positive bias at
765 lower expression levels, but negative bias at higher levels. Incorrect estimation of effective
766 length metrics used in RNA abundance calculations play a substantial role in the observed
767 patterns of bias. Nevertheless, it may be surprising to researchers that biases persist for standard

768 length-adjusted metrics such as TPM. CPM has less bias compared to TPM. Bias and noise can
769 be reduced by rescaling CPM to account for differences in effective length, such that the rescaled
770 CPM more faithfully recapitulates map-to-reference based expression estimates, even if
771 substantial noise persists. Adjusting expression estimates to account for effective length
772 differences can be beneficially used in two different applications. First, one could conduct
773 differential expression estimates across different transcriptome assemblies when it is necessary
774 to generate different assemblies for different conditions, such as divergent species or sub-species.
775 Because one of the first steps differential expression tools execute is to normalize by library size,
776 one can simply apply the correction directly on the counts themselves. CPM and count scale are
777 directly proportional, with the former effectively being the latter multiplied by a sample-specific
778 constant, such that all benefits we report on weighted-adjusted CPM will apply to similarly
779 transformed couts. Adjusting counts across a set of species would require treating one species as
780 the "reference", and rescale counts for other species relative to it. Outside of the realm of formal
781 tests of differential expression, we suggest that wt.-adj.-CPM could be useful in phylogenetic
782 comparative analyses of expression at the gene level. One of the big challenges of these methods
783 is accounting for length differences between transcript/contig sets belonging to orthologous
784 genes, and the adjusted CPM appears to substantially outperform other metrics in this regard. We
785 should note that our adjustment method is relatively simple, at that it may be possible to adopt
786 more sophisticated approaches that leverage such information as sequence composition and
787 complexity for additional improvements.

788 We do not test here datasets amenable to differential expression analyses, and a counter-
789 argument to our findings of bias and noise is that neither should have a substantial effect, so long
790 as bias and deviations from map-to-reference expectations are consistent across replicates and

791 conditions. While we did find broad congruence in bias across samples, large differences in bias
792 among samples were frequent enough such that they will likely elevate false positive rates for
793 differential expression tests derived from de novo transcriptome assemblies. This effect can be
794 minimized by utilizing metrics that adjust for effective length differences. Finally, our results
795 clearly indicate that isoform-level analyses should be avoided.

796 By far the most unexpected findings from our study were the extremely high genotyping
797 error rates derived from transcriptome assemblies, specifically ST representations of TRNITY
798 assemblies. False negatives were due, in part, to the absence of an ST aligned to a putatively true
799 MR polymorphic site, which is not unexpected given the known fragmented nature of many *de*
800 *novo* transcriptome assemblies. As our mapping rates of ST to genomes was extremely high,
801 such FN are due to either the failure to assembly any contigs at all for a gene (i.e. missing genes),
802 or the incomplete coverage of genes not extending over some fraction of polymorphic sites. That
803 genotyping errors were biased towards the distal end of contigs suggest that assembly graphs are
804 getting broken at hard to assemble features containing low complexity and/or increased
805 polymorphism. When such graph-splitting occurs, the intervening polymorphic sites will, by
806 definition, not be covered by ST, leading to downwardly biased estimates of heterozygosity in *de*
807 *novo* assemblies. False negatives may have also been caused by the slightly reduced read depth
808 at homologous positions in read alignments to ST. That the reduction in read depth in the latter,
809 was, on average, small suggests that other factors may play a role. It may be that genotype
810 inference methods such at the GATK HAPLOTYPE CALLER that leverage haplotypic information,
811 may have poorer performance when iterating over alignments to relatively short assembly
812 contigs. Another possibility is that assembly errors may generate patterns that get flagged by
813 standard genotyping filters. In contrast, false positives appear to be driven by paralog collapses

814 given the bias towards heterozygous false positive sites. In the absence of a genome sequence, it
815 is hard to correct FN biases for any particular transcriptome assembly. In contrast, sophisticated
816 filtering methods might permit the learning of sequence features that produce false positives, and
817 to derive metrics that could mitigate the false positive rate.

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

829 Even in the face of a high occurrence of false positive heterozygous genotype calls, false
830 negatives with respect to true heterozygous sites led to an overall negative bias in estimates of
831 heterozygosity estimates derived from ST relative to the gold standard MR approach, most
832 prominently in the wild *Mus* samples. This finding strongly suggests that the underlying site-
833 frequency spectrum derived from an ST assembly will be biased in ways that distort other
834 population genetic summary statistics typically reported in molecular ecological investigations
835 that utilize genotypes called from transcriptome assemblies. More broadly, the error rates we
836 document suggest that several lines of transcriptome-assembly based research are problematic.

837 First, erroneous heterozygous calls, when added to the redundancy problems described above,
838 will substantially reduce power and elevate false positive rates in allele-specific expression
839 analyses. Second, levels of genome-wide patterns of genetic diversity derived from
840 transcriptome assemblies will likely be negatively biased, at least for diploid species. The
841 distinction concerning ploidy is an important one because, in the absence of a genome or
842 karyotypic information, there can be no a priori expectation of how events such as whole-
843 genome or even chromosome-level duplication will impact genotyping inference. Thus, both
844 integration of diversity metrics across transcriptome assemblies with differing, unknown levels
845 of ploidy, and comparing transcriptome and genome assembly-derived metrics will confound
846 statistical method with biology and will lead to spurious inferences. Even when applied solely to
847 diploid transcriptome assemblies, as haplotypes will get assembled into different ST, the
848 expectation is that negative bias will be greater for more heterozygous genomes, such that
849 estimated levels of diversity across assemblies will artificially appear more similar than they are
850 in actuality, impeding researchers' ability to detect underlying patterns of variation.
851 While it is possible that a combination of algorithmic and annotation-based clustering
852 techniques can be employed to better collapse assembly redundancy, further research on this
853 topic is necessary. Whether or not more robust clustering techniques ameliorate bias, naïve
854 application of genotyping to ST derived from Trinity component "genes" will lead to high false
855 positive and false negative rates. Finally, the distortion of allele frequencies derived from
856 transcriptome assemblies strongly suggests that they will inject bias into analyses of population
857 structure, demographic inference, and efforts to detect positive selection.
858
859 **Conclusion**

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

882 analyses modes that suffer from fundamental biases, mitigating biases where possible for others,
883 and, always, placing appropriate caveats given the biases that will inevitably remain.
884

885 **ACKNOWLEDGMENTS**

886 The computations in this paper were run on the Odyssey cluster supported by the FAS Division
887 of Science, Research Computing Group at Harvard University. We thank members of the FAS
888 Informatics Group for discussion. We also thank Nadia Davidson and Alicia Oshlack for
889 feedback on an earlier version of this manuscript. This work was conducted on the traditional
890 territory of the Wampanoag and Massachusett peoples.
891

892 **COMPETING INTERESTS**

893 The authors have indicated that no competing interests exist.
894

895 **DATA AVAILABILITY**

896 Upon acceptance for publication, all transcriptome assemblies, SuperTranscript assemblies, and
897 vcf-format genotype files will be made available on Dryad.
898

899 **AUTHOR CONTRIBUTIONS**

900 The research project was designed by A. H. F., M. C., and T. B. S. Analyses were conducted by
901 A. H. F. with input from M.C. and T. B. S. A. H. F. wrote the paper with T. B. S. All authors
902 read and approved the final manuscript
903

904 **REFERENCES**

905 Austin, C. M., Tan, M. H., Harrisson, K. A., Lee, Y. P., Croft, L. J., Sunnucks, P., … Gan, H. M.
906 (2017). De novo genome assembly and annotation of Australia's largest freshwater fish,
907 the Murray cod (Maccullochella peelii), from Illumina and Nanopore sequencing read.
908 *GigaScience*, *6*(8), 1–6. doi: 10.1093/gigascience/gix063
909 Bankers, L., Fields, P., McElroy, K. E., Boore, J. L., Logsdon, J. M., & Neiman, M. (2017). Genomic
910 evidence for population-specific responses to co-evolving parasites in a New Zealand
911 freshwater snail. *Molecular Ecology*, *26*(14), 3663–3675. doi: 10.1111/mec.14146
912 Berens, A. J., Hunt, J. H., & Toth, A. L. (2015). Comparative transcriptomics of convergent
913 evolution: different genes but conserved pathways underlie caste phenotypes across
914 lineages of eusocial insects. *Molecular Biology and Evolution*, *32*(3), 690–703. doi:
915 10.1093/molbev/msu330
916 Birol, I., Jackman, S. D., Nielsen, C. B., Qian, J. Q., Varhol, R., Stazyk, G., … Jones, S. J. M. (2009).
917 De novo transcriptome assembly with ABySS. *Bioinformatics*, *25*(21), 2872–2877. doi:
918 10.1093/bioinformatics/btp367
919 Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq
920 quantification. *Nature Biotechnology*, *34*(5), 525–527. doi: 10.1038/nbt.3519
921 Cokus, S. J., Gugger, P. F., & Sork, V. L. (2015). Evolutionary insights from de novo transcriptome
922 assembly and SNP discovery in California white oaks. *BMC Genomics*, *16*, 552. doi:
923 10.1186/s12864-015-1761-4
924 Davidson, N. M., Hawkins, A. D. K., & Oshlack, A. (2017). SuperTranscripts: a data driven
925 reference for analysis and visualisation of transcriptomes. *Genome Biology*, *18*(1). doi:
926 10.1186/s13059-017-1284-1

927 Davidson, N. M., & Oshlack, A. (2014). Corset: enabling differential gene expression analysis for
928 de novoassembled transcriptomes. *Genome Biology*, *15*(7), 410. doi: 10.1186/s13059-
929 014-0410-6
930 De Wit, P., & Palumbi, S. R. (2013). Transcriptome-wide polymorphisms of red abalone (Haliotis
931 rufescens) reveal patterns of gene flow and local adaptation. *Molecular Ecology*, *22*(11),
932 2884–2897. doi: 10.1111/mec.12081
933 Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., … Regev, A.
934 (2011a). Full-length transcriptome assembly from RNA-Seq data without a reference
935 genome. *Nature Biotechnology*, *29*(7), 644–652. doi: 10.1038/nbt.1883
936 Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., … Regev, A.
937 (2011b). Full-length transcriptome assembly from RNA-Seq data without a reference
938 genome. *Nature Biotechnology*, *29*(7), 644–652. doi: 10.1038/nbt.1883
939 Harr, B., Karakoc, E., Neme, R., Teschke, M., Pfeifle, C., Pezer, Ž., … Tautz, D. (2016). Genomic
940 resources for wild populations of the house mouse, Mus musculus and its close relative
941 Mus spretus. *Scientific Data*, *3*, 160075. doi: 10.1038/sdata.2016.75
942 Hornett, E. A., & Wheat, C. W. (2012). Quantitative RNA-Seq analysis in non-model species:
943 assessing transcriptome assemblies as a scaffold and the utility of evolutionary
944 divergent genomic reference species. *BMC Genomics*, *13*, 361. doi: 10.1186/1471-2164-
945 13-361
946 Kannan, S., Hui, J., Mazooji, K., Pachter, L., & Tse, D. (2016). Shannon: An Information-Optimal
947 de Novo RNA-Seq Assembler. *BioRxiv*, 039230. doi: 10.1101/039230

948 Kent, W. J. (2002). BLAT--the BLAST-like alignment tool. *Genome Research*, *12*(4), 656–664. doi:
949 10.1101/gr.229202. Article published online before March 2002
950 Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory
951 requirements. *Nature Methods*, *12*(4), 357–360. doi: 10.1038/nmeth.3317
952 Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature*
953 *Methods*, *9*(4), 357–359. doi: 10.1038/nmeth.1923
954 Le Gac, M., Metegnier, G., Chomérat, N., Malestroit, P., Quéré, J., Bouchez, O., … Chapelle, A.
955 (2016). Evolutionary processes and cellular functions underlying divergence in
956 Alexandrium minutum. *Molecular Ecology*, *25*(20), 5129–5143. doi: 10.1111/mec.13815
957 Li, B., & Dewey, C. N. (2011a). RSEM: accurate transcript quantification from RNA-Seq data with
958 or without a reference genome. *BMC Bioinformatics*, *12*, 323. doi: 10.1186/1471-2105-
959 12-323
960 Li, B., & Dewey, C. N. (2011b). RSEM: accurate transcript quantification from RNA-Seq data with
961 or without a reference genome. *BMC Bioinformatics*, *12*, 323. doi: 10.1186/1471-2105-
962 12-323
963 Li, B., Fillmore, N., Bai, Y., Collins, M., Thomson, J. A., Stewart, R., & Dewey, C. N. (2014).
964 Evaluation of de novo transcriptome assemblies from RNA-Seq data. *Genome Biology*,
965 *15*(12). doi: 10.1186/s13059-014-0553-5
966 Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Genome Project Data
967 Processing, S. (2009). The Sequence Alignment/Map format and SAMtools.
968 *Bioinformatics (Oxford, England)*, *25*(16), 2078–2079. doi:
969 10.1093/bioinformatics/btp352;

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

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

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

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

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

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

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*, *20*(9), 1297–1303. doi: 10.1101/gr.107524.110

991 Mi, H., Muruganujan, A., Casagrande, J. T., & Thomas, P. D. (2013). Large-scale gene function
992 analysis with the PANTHER classification system. *Nature Protocols*, *8*(8), 1551–1566. doi:
993 10.1038/nprot.2013.092
994 Michael, T. P., Jupe, F., Bemm, F., Motley, S. T., Sandoval, J. P., Lanz, C., … Ecker, J. R. (2018).
995 High contiguity Arabidopsis thaliana genome assembly with a single nanopore flow cell.
996 *Nature Communications*, *9*(1), 541. doi: 10.1038/s41467-018-03016-2
997 Nowoshilow, S., Schloissnig, S., Fei, J.-F., Dahl, A., Pang, A. W. C., Pippel, M., … Myers, E. W.
998 (2018). The axolotl genome and the evolution of key tissue formation regulators.
999 *Nature*, *554*(7690), 50–55. doi: 10.1038/nature25458
1000 Ockendon, N. F., O'Connell, L. A., Bush, S. J., Monzón-Sandoval, J., Barnes, H., Székely, T., …
1001 Urrutia, A. O. (2016). Optimization of next-generation sequencing transcriptome
1002 annotation for species lacking sequenced genomes. *Molecular Ecology Resources*, *16*(2),
1003 446–458. doi: 10.1111/1755-0998.12465
1004 Pankey, M. S., Minin, V. N., Imholte, G. C., Suchard, M. A., & Oakley, T. H. (2014). Predictable
1005 transcriptome evolution in the convergent and complex bioluminescent organs of squid.
1006 *Proceedings of the National Academy of Sciences*, *111*(44), E4736–E4742. doi:
1007 10.1073/pnas.1416574111
1008 Peona, V., Weissensteiner, M. H., & Suh, A. (2018). How complete are "complete" genome
1009 assemblies?—An avian perspective. *Molecular Ecology Resources*, *18*(6), 1188–1195.
1010 doi: 10.1111/1755-0998.12933
1011 Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic
1012 features. *Bioinformatics*, *26*(6), 841–842. doi: 10.1093/bioinformatics/btq033

1013 Roda, F., Mendes, F. K., Hahn, M. W., & Hopkins, R. (2017). Genomic evidence of gene flow
1014 during reinforcement in Texas Phlox. *Molecular Ecology*, *26*(8), 2317–2330. doi:
1015 10.1111/mec.14041
1016 Romiguier, J., Gayral, P., Ballenghien, M., Bernard, A., Cahais, V., Chenuil, A., … Galtier, N.
1017 (2014). Comparative population genomics in animals uncovers the determinants of
1018 genetic diversity. *Nature*, *515*(7526), 261–263. doi: 10.1038/nature13685
1019 Schwentner, M., Combosch, D. J., Pakes Nelson, J., & Giribet, G. (2017). A Phylogenomic
1020 Solution to the Origin of Insects by Resolving Crustacean-Hexapod Relationships. *Current*
1021 *Biology*, *27*(12), 1818-1824.e5. doi: 10.1016/j.cub.2017.05.040
1022 Siepel, A. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast
1023 genomes. *Genome Research*, *15*(8), 1034–1050. doi: 10.1101/gr.3715005
1024 Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015).
1025 BUSCO: assessing genome assembly and annotation completeness with single-copy
1026 orthologs. *Bioinformatics*, *31*(19), 3210–3212. doi: 10.1093/bioinformatics/btv351
1027 Singh, R., Jones, T., Wai, C. M., Jifon, J., Nagai, C., Ming, R., & Yu, Q. (2018). Transcriptomic
1028 analysis of transgressive segregants revealed the central role of photosynthetic capacity
1029 and efficiency in biomass accumulation in sugarcane. *Scientific Reports*, *8*(1), 4415. doi:
1030 10.1038/s41598-018-22798-5
1031 Smith-Unna, R., Boursnell, C., Patro, R., Hibberd, J. M., & Kelly, S. (2016). TransRate: reference-
1032 free quality assessment of de novo transcriptome assemblies. *Genome Research*, *26*(8),
1033 1134–1144. doi: 10.1101/gr.196469.115

1034 Soneson, C., Love, M. I., & Robinson, M. D. (2015). Differential analyses for RNA-seq: transcript-
1035 level estimates improve gene-level inferences. *F1000Research*, *4*, 1521. doi:
1036 10.12688/f1000research.7563.1
1037 Song, L., & Florea, L. (2015). Rcorrector: efficient and accurate error correction for Illumina
1038 RNA-seq reads. *GigaScience*, *4*, 48. doi: 10.1186/s13742-015-0089-y
1039 Tepolt, C. K., & Palumbi, S. R. (2015). Transcriptome sequencing reveals both neutral and
1040 adaptive genome dynamics in a marine invader. *Molecular Ecology*, *24*(16), 4145–4158.
1041 doi: 10.1111/mec.13294
1042 Todd, E. V., Black, M. A., & Gemmell, N. J. (2016). The power and promise of RNA-seq in ecology
1043 and evolution. *Molecular Ecology*, *25*(6), 1224–1241. doi: 10.1111/mec.13526
1044 Veilleux, H. D., Ryu, T., Donelson, J. M., van Herwerden, L., Seridi, L., Ghosheh, Y., … Munday, P.
1045 L. (2015). Molecular processes of transgenerational acclimation to a warming ocean.
1046 *Nature Climate Change*, *5*(12), 1074–1078. doi: 10.1038/nclimate2724
1047 Velotta, J. P., Wegrzyn, J. L., Ginzburg, S., Kang, L., Czesny, S., O'Neill, R. J., … Schultz, E. T.
1048 (2017). Transcriptomic imprints of adaptation to fresh water: parallel evolution of
1049 osmoregulatory gene expression in the Alewife. *Molecular Ecology*, *26*(3), 831–848. doi:
1050 10.1111/mec.13983
1051 Vijay, N., Poelstra, J. W., Künstner, A., & Wolf, J. B. W. (2013). Challenges and strategies in
1052 transcriptome assembly and differential gene expression quantification. A
1053 comprehensive *in silico* assessment of RNA-seq experiments. *Molecular Ecology*, *22*(3),
1054 620–634. doi: 10.1111/mec.12014

1055 Wu, T. D., & Watanabe, C. K. (2005). GMAP: a genomic mapping and alignment program for
1056 mRNA and EST sequences. *Bioinformatics*, *21*(9), 1859–1875. doi:
1057 10.1093/bioinformatics/bti310
1058 Zook, J. M., Catoe, D., McDaniel, J., Vang, L., Spies, N., Sidow, A., … Salit, M. (2016). Extensive
1059 sequencing of seven human genomes to characterize benchmark reference materials.
1060 *Scientific Data*, *3*. doi: 10.1038/sdata.2016.25
1061

1062 **TABLES**

1063 **Table 1**. Counts of missing genes by assembler for three *Mus* samples

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

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

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

1064

1065

1066 **Table 2.** For filtered assemblies, median genomic alignment-based transcript per million (TPM)

1067 and phastCons scores for ENSEMBL genes that are recovered by all de novo transcriptome

1068 assembly methods (All), recovered by 1-2 of the 3 methods (Some), and recovered by no

1069 assembler (None). Flagged tests are significant after Bonferroni correction; * P < 0.005, ** P

1070 <0.0005, and *** P < 1.0 x 10^{-10}.

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

1071

1072 **Table 3**. Selected performance metrics comparing genotypes derived from SuperTranscript
1073 collapse of TRINITY de novo transcriptome assemblies with those derived from a map-to-
1074 reference genome approach. See **METHODS** for information on metric calculations. Values in
1075 parentheses are after converting SuperTranscripts with no protein hit to UNIREF90 to NA. The
1076 full table of metrics is available as Supplementary Table 7.

	Sample					
Metric	**MDC**	**BALB/c**	**FRA**	**DEU**	**IRN**	**KZK**
Error rate	0.870 (0.892)	0.500 (0.477)	0.359 (0.453)	0.376	0.336	0.301
Error-MrefAlleleIncluded	1.000 (1.000)	0.898 (0.898)	0.969 (0.971)	0.977	0.969	0.959
False negative (FN)	0.662 (0.754)	0.389 (0.412)	0.260 (0.396)	0.265	0.246	0.239
FN-noalign	0.728 (0.868)	0.745 (0.881)	0.585 (0.860)	0.585	0.572	0.668
FN-het	0.623 (0.752)	0.742 (0.885)	0.259 (0.451)	0.259	0.242	0.268
False positive (FP)	0.609 (0.558)	0.179 (0.110)	0.129 (0.091)	0.147	0.115	0.078
FP-NAtoFixed	0.118 (0.094)	0.053 (0.051)	0.027 (0.019)	0.028	0.026	0.024
FP-NAtoHet	0.882 (0.906)	0.919 (0.928)	0.916 (0.923)	0.909	0.917	0.915
FP-indel	0.188 (0.196)	0.063 (0.048)	0.184 (0.146)	0.225	0.194	0.070
Recall	0.334 (0.243)	0.609 (0.587)	0.735 (0.602)	0.731	0.750	0.757
Precision	0.176 (0.161)	0.735 (0.826)	0.829 (0.856)	0.805	0.849	0.897

1077

1078 **Table 4.** Correlation between map-to-reference and map-to-denovo-transcriptome expression at
1079 the gene level, for both RSEM and KALLISTO, based upon CPM, TPM, and weighted-adjusted
1080 CPM. See equation in methods for details regarding the last of these. Correlations exclude
1081 expressed reference genes where there are not associated denovo assembly contigs, i.e. genes
1082 that are missing from the assembly. Correlations are calculated for entries with TPM > 0 for
1083 either map-to-reference or map-to-denovo assembly. Bold indicates strongest correlation across
1084 metrics and estimation algorithms for a particular sample-assembler combination. Bold italics
1085 indicate the strongest correlation across all estimators and assembly algorithms for a particular
1086 sample.

1087

Sample	**CPM**		**TPM**		**Weighted EL-adjusted CPM**	
	RSEM	**kallisto**	**RSEM**	**kallisto**	**RSEM**	**kallisto**
MDC						
Trinity	0.853	0.858	0.756	0.675	***0.901***	0.823
Shannon	0.629	**0.833**	0.567	0.662	0.680	0.784
BinPacker	0.824	0.825	0.796	0.754	**0.877**	0.783
BALB/c						
Trinity	0.869	0.881	0.818	0.819	***0.897***	0.863
Shannon	0.819	0.859	0.790	0.815	0.854	0.826
BinPacker	0.835	0.846	0.801	0.803	0.870	0.824
Massif Central						
Trinity	0.850	0.861	0.809	0.790	**0.903**	0.841
Shannon	0.849	0.857	0.812	0.785	***0.906***	0.819
BinPacker	0.815	0.821	0.808	0.786	0.872	0.793
Germany						
Trinity	0.848	0.859	0.804	0.788	**0.903**	0.838
Shannon	0.847	0.855	0.812	0.784	***0.905***	0.813
BinPacker	0.811	0.818	0.808	0.784	0.873	0.789
Iran						
Trinity	0.853	0.862	0.806	0.785	***0.906***	0.842
Shannon	0.847	0.854	0.804	0.777	**0.904**	0.817
BinPacker	0.809	0.816	0.799	0.773	0.869	0.787
Kazakhstan						
Trinity	0.852	0.862	0.810	0.791	**0.903**	0.841
Shannon	0.852	0.858	0.814	0.793	***0.907***	0.824
BinPacker	0.816	0.821	0.809	0.788	0.872	0.795

1088

1089 **FIGURE LEGENDS**

1090 **Figure 1.** Coverage of individual *Mus* reference transcripts and genes by the TRINITY assembly.
1091 Coverage of all expressed reference sequences by the unfiltered assembly is defined by BLAT
1092 alignments with a minimum percent identity of 95% for **(a)** individual contigs with the highest
1093 scoring alignment, **(b)** summed coverage of all BLAT hits and **(c)** gene-level coverage as the
1094 expression-weighted average of summed coverage across isoforms. Comparable coverage
1095 estimates for only protein-coding transcripts/genes of minimum length 200bp, by contigs with
1096 TPM ≥ 1 and an open-reading frame alignment to UNIREF90 are represented by **(d-f)**. Blue lines
1097 indicate smoothed, fitted trends on the coverage-TPM relationship, while purple lines indicate
1098 the empirical cumulative frequency distribution on TPM.

1099

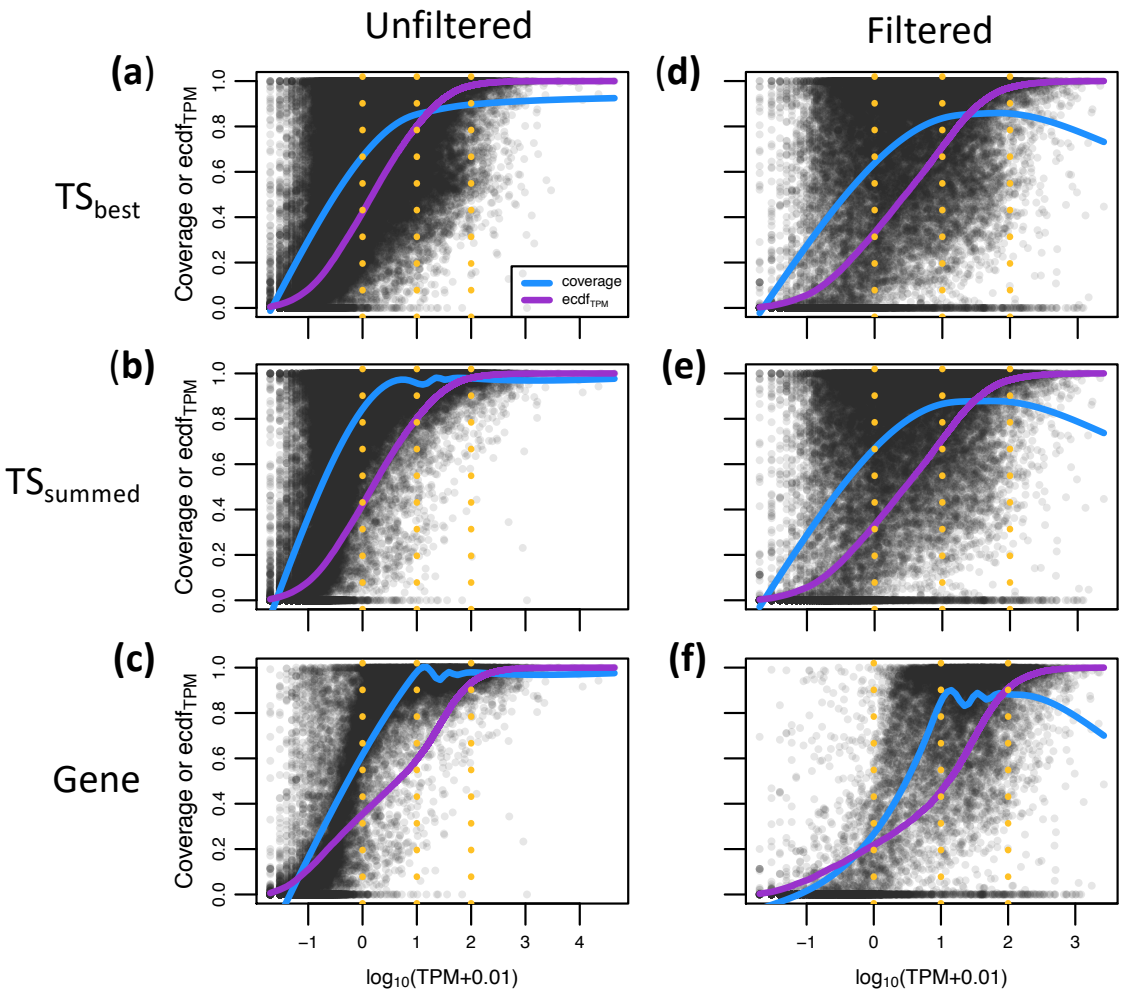
1100 **Figure 2.** Violin plots of **(a)** TPM and **(b)** phastCons scores for protein coding genes with a
1101 minimum transcript length of 200bp, which are recovered in all three transcriptome assemblers
1102 (TRINITY, SHANNON, BINPACKER), 1-2 of three assemblers, or that are missing from all three
1103 assemblers. Missingness is determined using filtered transcriptome assemblies, that only include
1104 contigs with TPM ≥ 1 and an open-reading frame alignment to UNIREF90.

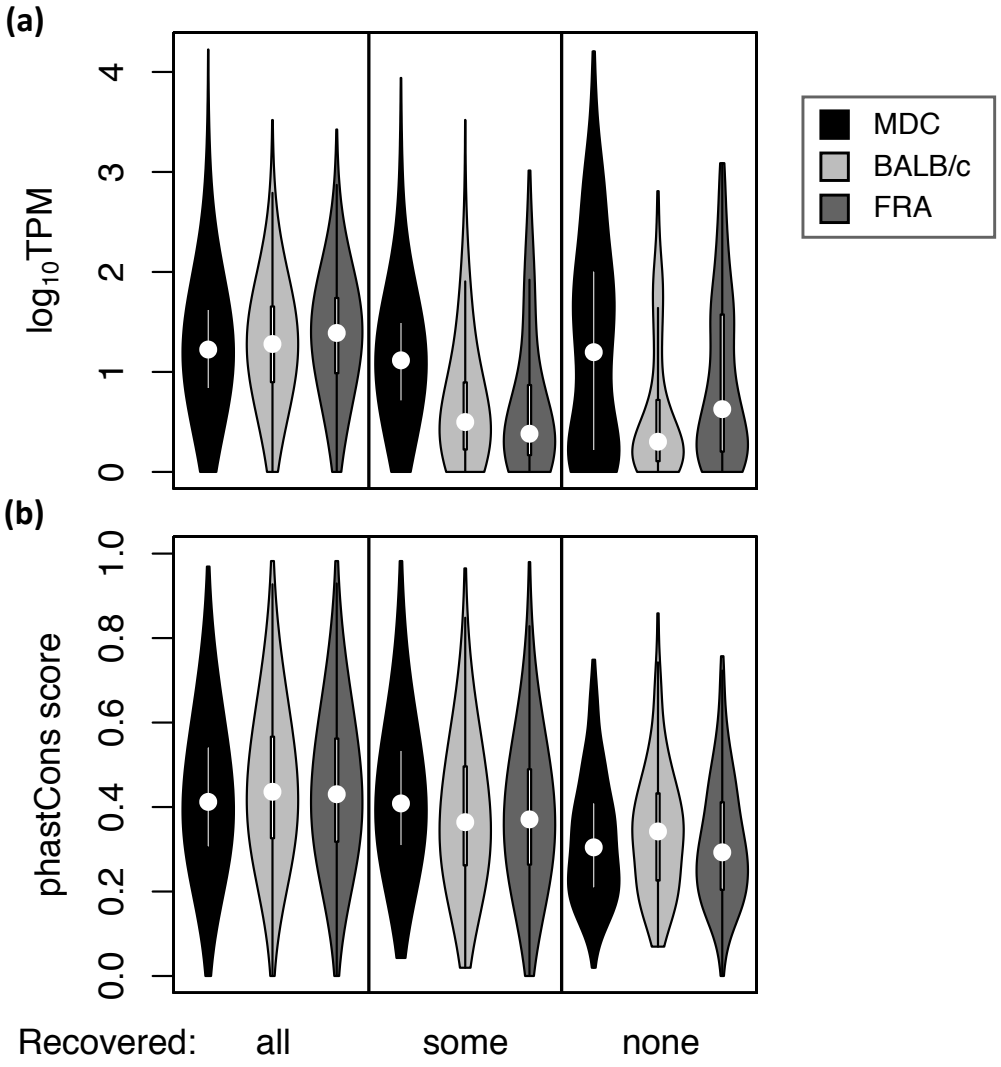
1105

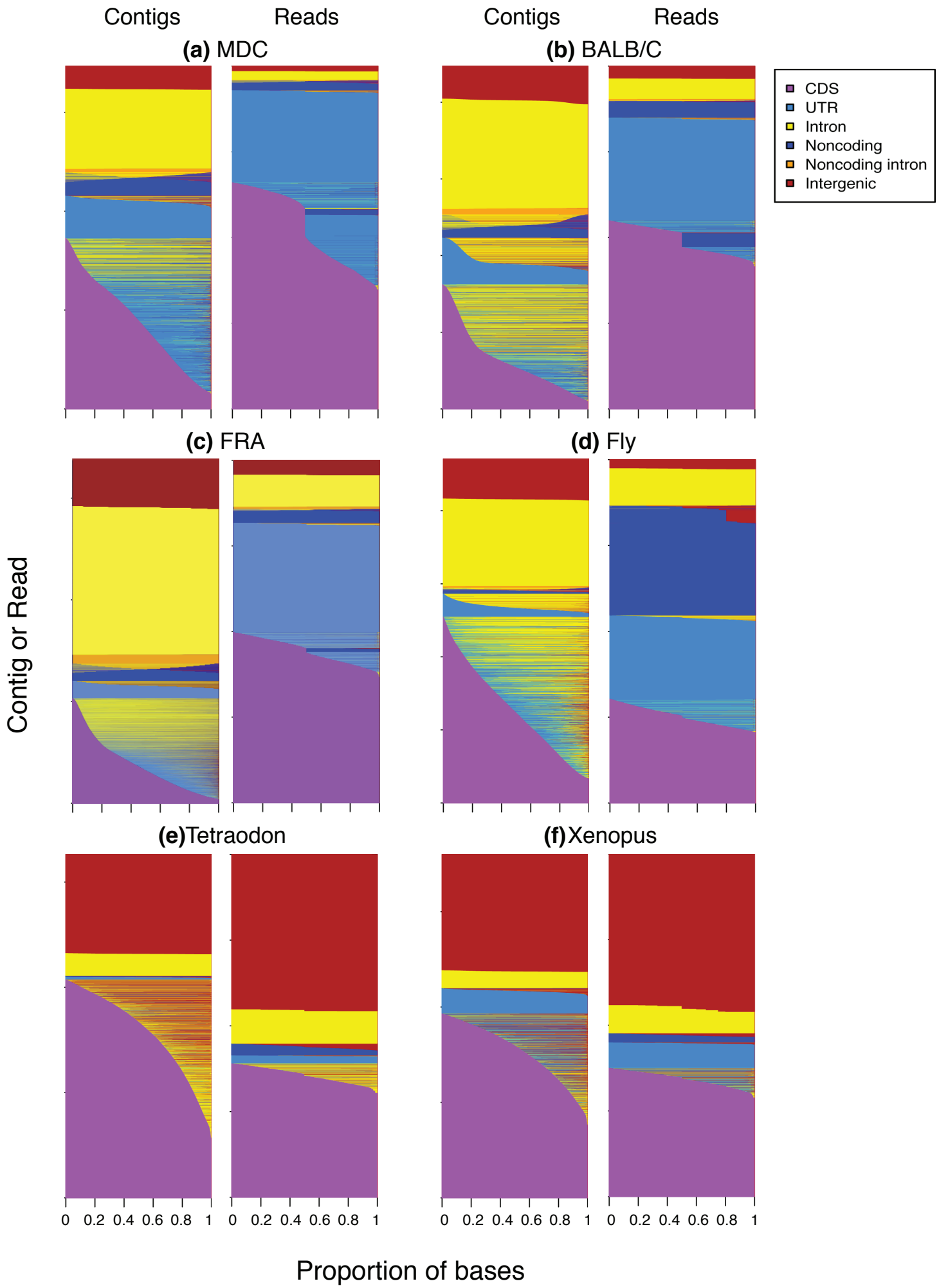
1106 **Figure 3.** Base composition by TRINITY contig (left panels), and sequence reads (right panels)
1107 for **(a)** *M. musculus* dendritic cells, **(b)** *M. musculus* BALB/c strain brain, **(c)** wild outcrossed *M.*
1108 *musculus* brain, **(d)** *D. melanogaster* head, **(e)** *T. nigroviridis* brain, and **(f)** *X. tropicalis* kidney.
1109 For each panel, each y-axis row represents a single contig or read. Sequence panels for reads are
1110 based upon 100,000 randomly sampled mapped reads. Intersections of contigs and reads with
1111 annotations are based upon mapping with GMAP and HISAT2, respectively (see **Methods** for

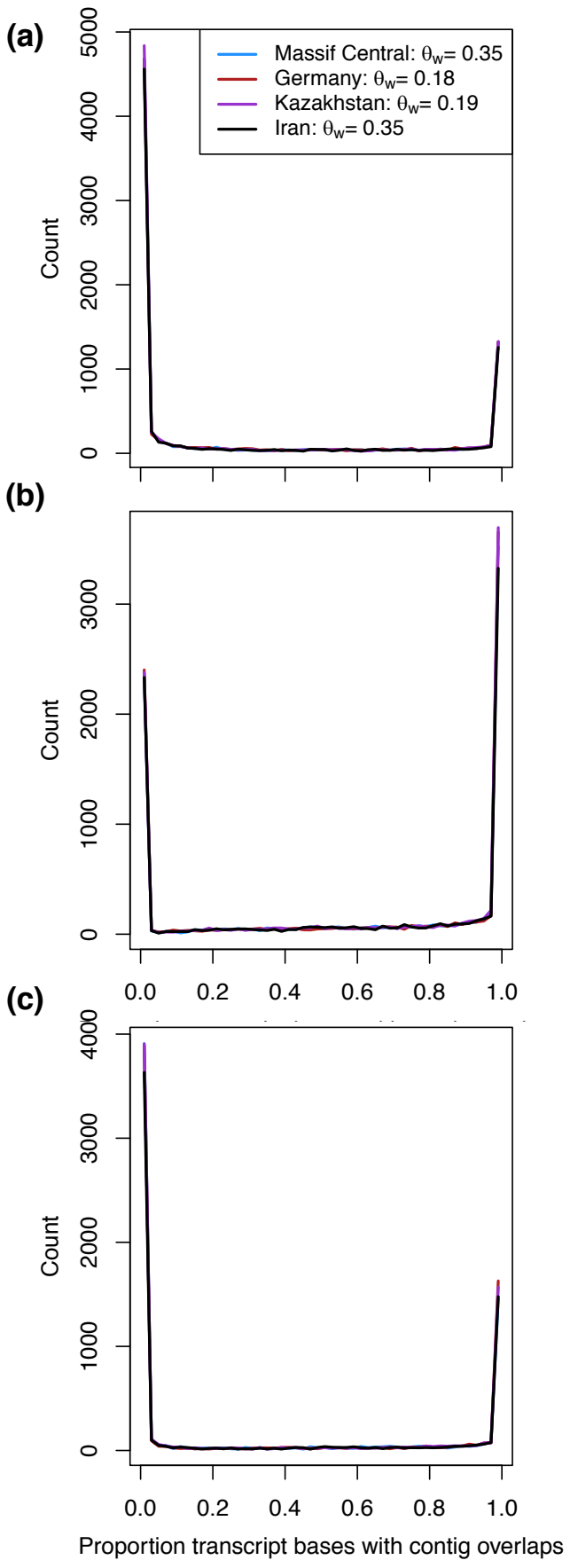
1112 details). For visualization purposes, sequence proportions are arranged from the bottom to the
1113 top of a panel, sorted in descending order, consecutively by CDS, UTR, non-coding, non-coding
1114 intron, intron, and intergenic proportions.
1115
1116 **Figure 4.** Frequency of proportional overlaps between contigs, based upon BLAT all-by-all hits,
1117 for known *Mus* single isoform genes, for three 8-individual pools of wild *M. m. domesticus* (Iran;
1118 Massif Central, France and Germany) and one 8-individual pool of *M. m. musculus*
1119 (Kazakhstan), for **(a)** TRINITY, **(b)** SHANNON and **(c)** BINPACKER. Estimates of Watterson's θ are
1120 those reported from whole genome sequence data in Harr et al. (2016).
1121
1122 **Figure 5. (a)** For FRA TRINITY assembly, hexbin 2D histogram of the proportion of contig
1123 lengths comprising BLAT pairwise alignments resulting from all-by-all searches. **(b)** Example
1124 alignment to the *Mus* genome, to a protein-coding transcript of the GDA gene, of nearly
1125 identically sized (1220 bp) FRA TRINITY contigs. These contigs have a BLAT-based 99.8 %
1126 identity in an alignment equal to the entirety of the contig length, and which are assembled from
1127 different TRINITY components (i.e. "genes"), as indicated by the red box around their labels. The
1128 contigs differ by only two adjacent SNPs (A and T), and also share fixed C (far left) and G (far
1129 right) differences relative to the reference genome. Thus, different Trinity components have
1130 assembled different haplotypes.
1131
1132 **Figure 6.** Examples of map-to-reference and SuperTranscript derived genotype combinations
1133 that produce particular error classes.
1134

1135 **Figure 7.** Heterozygosity per base for exonic biallelic SNV variants called relative to the *Mus*
1136 reference genome using the GATK best practices pipeline versus those obtained from
1137 supertranscripts constructed from TRINITY assemblies using scripts provided with the TRINITY v.
1138 2.6.5 package. See **Methods** for details concerning how exonic coordinates were obtained for
1139 SuperTranscripts.
1140
1141 **Figure 8.** For the FRA data set, comparison of gene-level expression summed across BLAT
1142 alignments of TRINITY contigs to reference transcripts (grouped by gene), to reference gene
1143 expression in terms of **(a)** CPM, **(b)** TPM, and **(c)** weighted-adjusted CPM that accounts for
1144 differences in effective length estimates between methods. For CPM, TPM, and weighted-
1145 adjusted CPM, 0.05, 0.01 and 0.05 are added to values prior to log-transformation, in order to
1146 visualize instances where expression = 0 for one of the methods being compared. The red line
1147 represents unity, the blue line is the smoothed fit to data. We exclude all cases where an
1148 expressed reference gene has no TRINITY contigs aligned to it, such that it is missing from the de
1149 novo transcriptome assembly.
1150
1151 **Figure 9.** For the FRA data set, for genes with non-zero expression in both transcriptome
1152 assembly and map-to-reference, log-fold change between TRINITY and map-to-reference
1153 expression estimates as a function of difference in effective length (EL) for **(a)** CPM, **(b)** TPM,
1154 and **(c)** weighted-adjusted CPM.
1155
1156 **Figure 10.** Frequency of bias effect sizes for gene-level expression estimated by CPM, TPM,
1157 and weighted-adjusted CPM for a TRINITY-based comparison of the FRA and DEU samples.

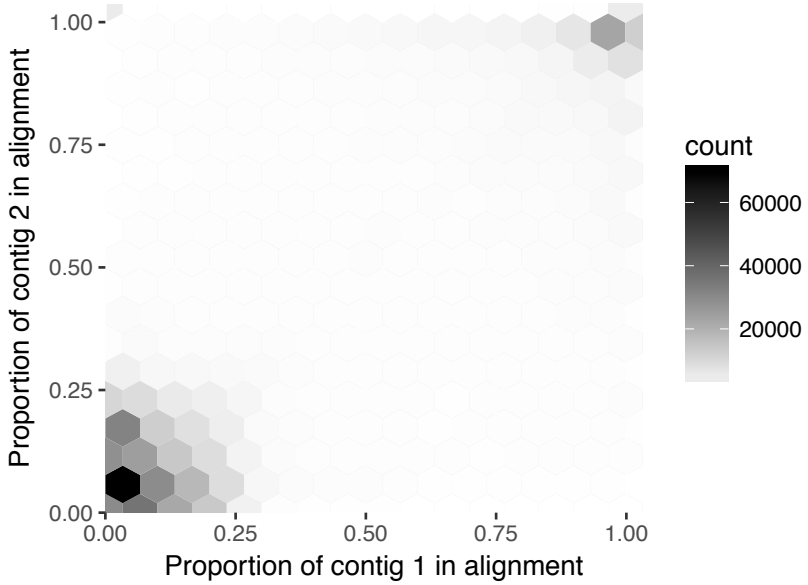

Unfiltered
Filtered
(a)
(b)
(c)
(d)
(e)
(f)
TS best
TS summed
Gene
Coverage or ecdf TPM
coverage
ecdf TPM
log10(TPM+0.01)


(a)
log10TPM
0
1
2
3
4
MDC
BALB/c
FRA
(b)
phastCons score
0.0
0.2
0.4
0.6
0.8
1.0
Recovered:
all
some
none

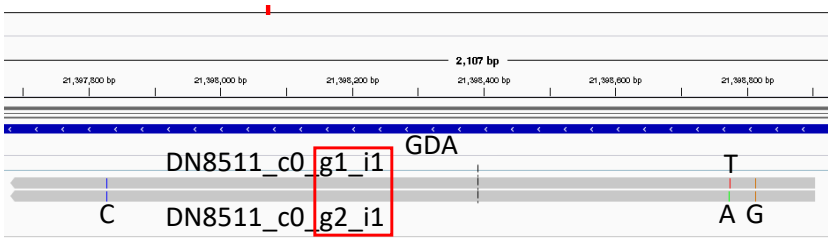

Contigs
Reads
Contigs
Reads
(a) MDC
(b) BALB/C
(c) FRA
(d) Fly
(e)Tetraodon
(f)Xenopus
CDS
UTR
Intron
Noncoding
Noncoding intron
Intergenic
Contig or Read
0 0.2 0.4 0.6 0.8 1
0 0.2 0.4 0.6 0.8 1
0 0.2 0.4 0.6 0.8 1
0 0.2 0.4 0.6 0.8 1
Proportion of bases


(a)
(b)
(c)
Massif Central: θw= 0.35
Germany: θw= 0.18
Kazakhstan: θw= 0.19
Iran: θw= 0.35
Count
0
1000
2000
3000
4000
5000
0.0
0.2
0.4
0.6
0.8
1.0
Proportion transcript bases with contig overlaps

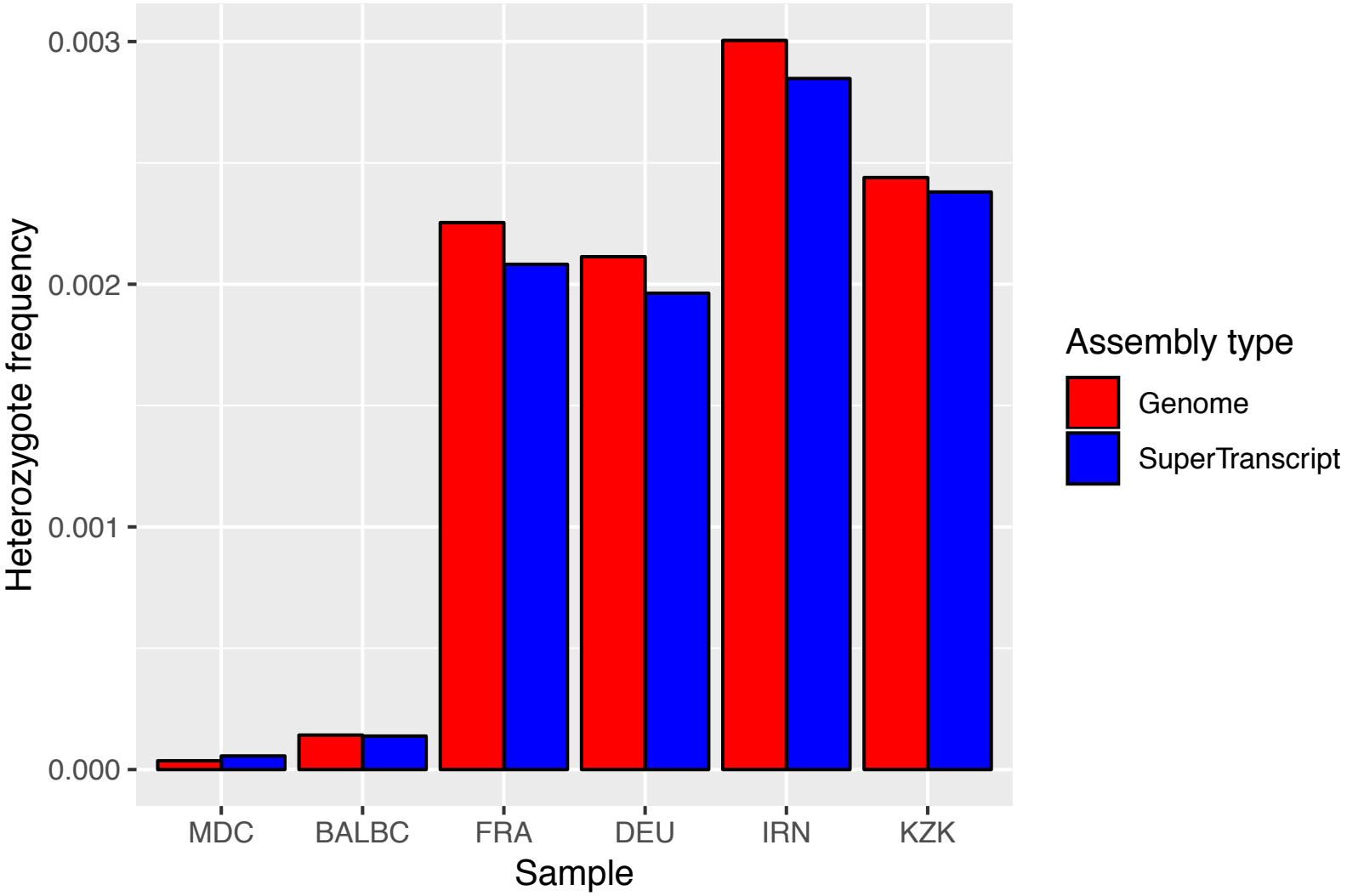
(a)

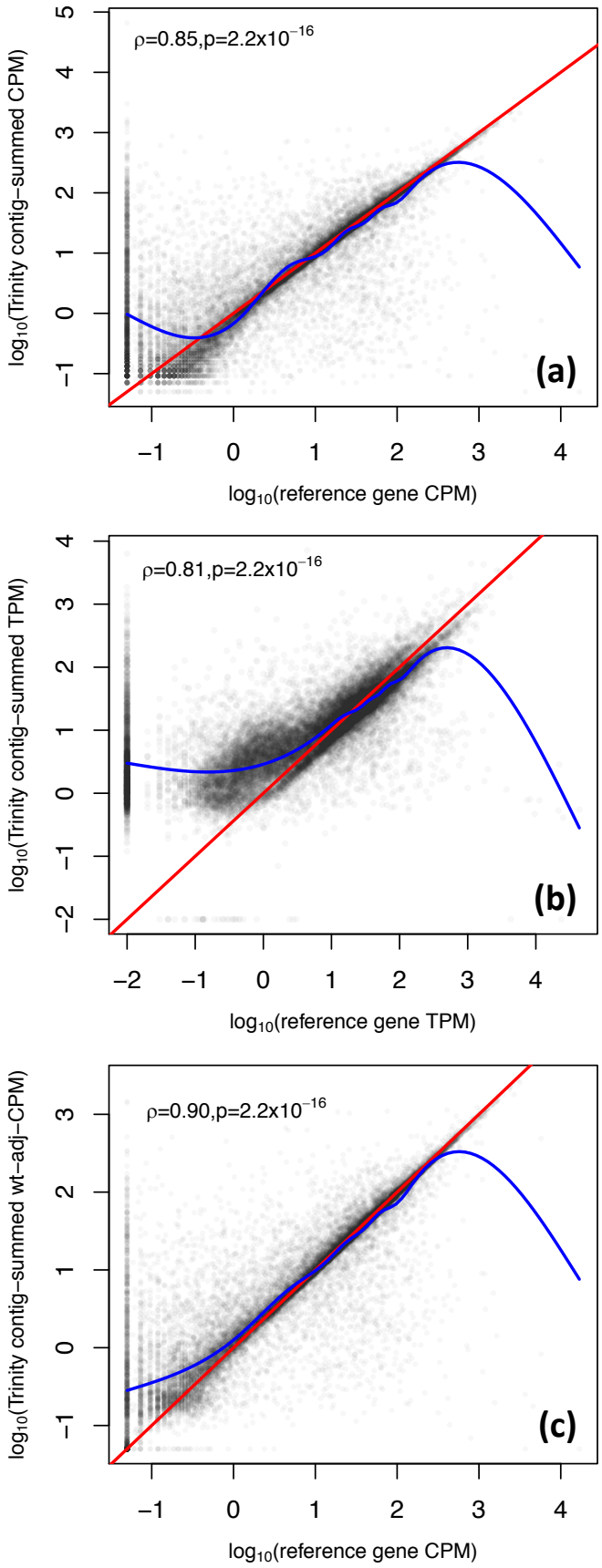

Proportion of contig 2 in alignment
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Proportion of contig 1 in alignment
count
60000
40000
20000

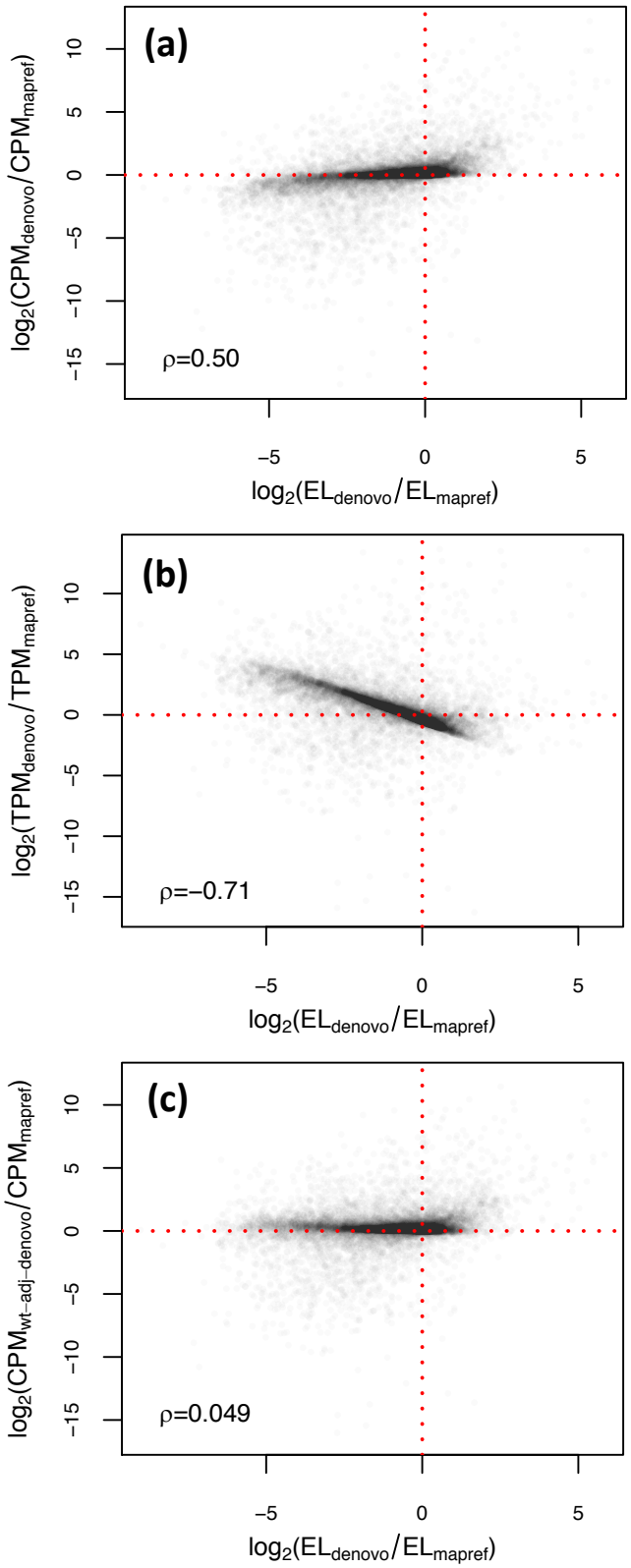
(b)

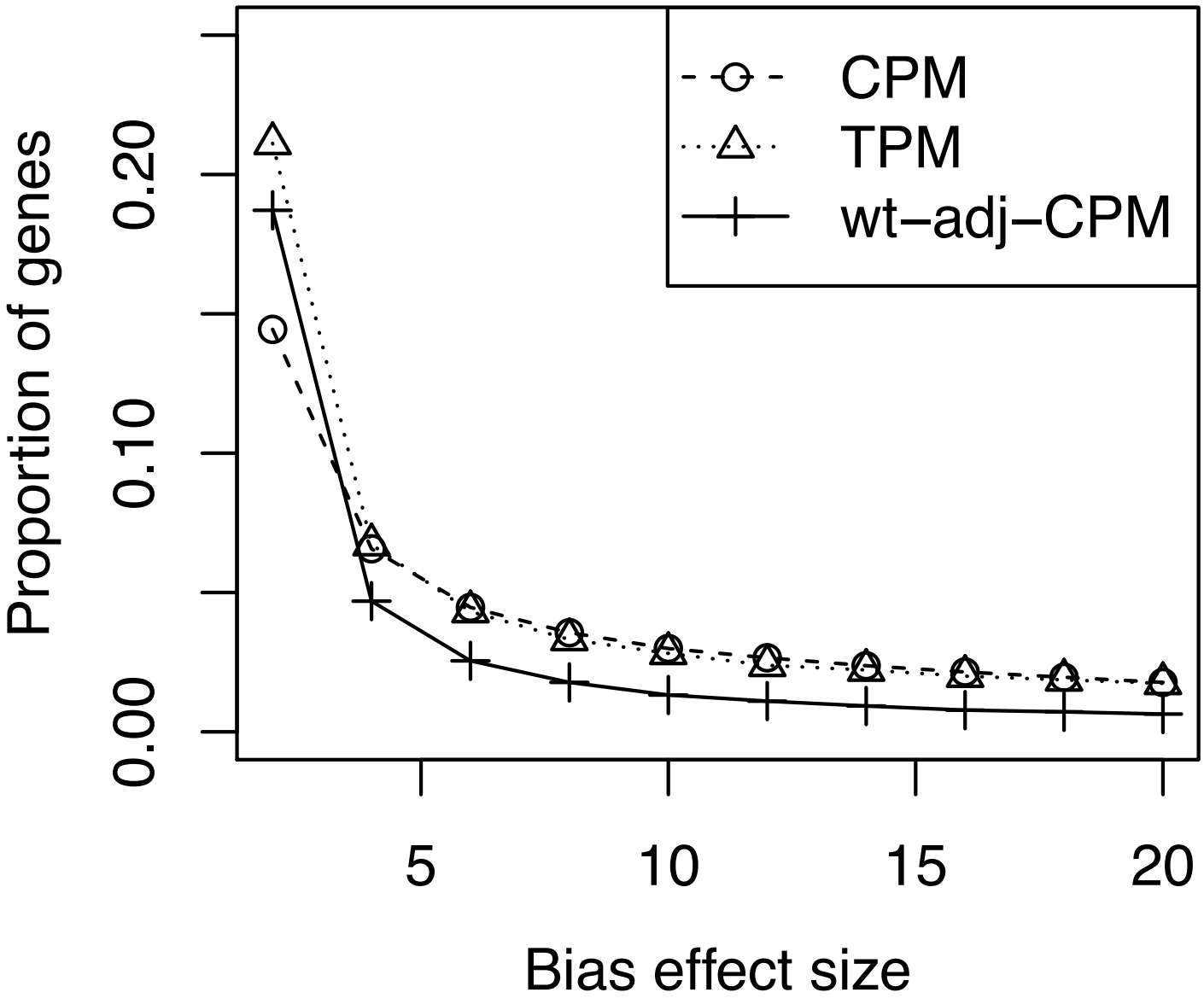

2,107 bp
GDA
DN8511_c0_g1_i1
DN8511_c0_g2_i1
C
T
A G

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


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


ρ=0.85,p=2.2x10^−16
(a)
log10(Trinity contig−summed CPM)
log10(reference gene CPM)
ρ=0.81,p=2.2x10^−16
(b)
log10(Trinity contig−summed TPM)
log10(reference gene TPM)
ρ=0.90,p=2.2x10^−16
(c)
log10(Trinity contig−summed wt−adj−CPM)
log10(reference gene CPM)


(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