Short paired-end reads trump long single-end reads for expression analysis

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

John M. Gaspar, 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: tsackton@g.harvard.edu

ABSTRACT

Background: Typical experimental design advice for expression analyses using RNA-seq generally assumes that single-end reads provide robust gene-level expression estimates in a cost-effective manner, and that the additional benefits obtained from paired-end sequencing are not worth the additional cost. However, in many cases (e.g., with Illumina NextSeq and NovaSeq instruments), shorter paired-end reads and longer single-end reads can be generated for the same cost, and it is not obvious which strategy should be preferred. Using publicly available data, we test whether short-paired end reads can achieve more robust expression estimates and differential expression results than single-end reads of approximately the same total number of sequenced bases.

Results: At both the transcript and gene levels, 2x40 paired-end reads unequivocally provide expression estimates that are more highly correlated with 2x125 than 1x75 reads; in nearly all cases, those correlations are also greater than for 1x125, despite the greater total number of sequenced bases for the latter. Across an array of metrics, differential expression tests based upon 2x40 consistently outperform those using 1x75.

Conclusion: Researchers seeking a cost-effective approach for gene-level expression analysis should prefer short paired-end reads over a longer single-end strategy. Short paired-end reads will also give reasonably robust expression estimates and differential expression results at the isoform level.

BACKGROUND

For over a decade, RNA-seq has empowered gene expression analysis. This has led to fundamental advances in our understanding of diverse phenomena, including the evolution of

47 gene regulation across species (1), alternative splicing (2,3), relative importance of cis vs -trans
48 regulation (4,5), genetic underpinnings of heritable disease (6,7), and the genetic architecture of
49 phenotypes in natural populations (8,9). While the cost of sequencing has steadily decreased,
50 making expression analyses feasible even on modest research budgets, maximizing cost
51 efficiency remains high priority. Cost effectiveness is particularly important as studies scale up
52 to hundreds or even thousands of samples. Maximizing performance given a particular study
53 design has already received much attention, with investigations of optimal strategies for quality
54 trimming (10), the effects of quality trimming on expression estimates and differential
55 expression testing (11), optimal sequencing depth (12,13), importance of biological replicates
56 (12,14,15), and the relative performance of expression and differential expression tools and
57 pipelines (16–20). Less attention has focused on how to improve performance given a fixed
58 number of biological replicates and total sequenced bases.

59 While the typical length of sequenced reads has increased as costs have decreased, when
60 an annotated genome is available it remains standard procedure to sequence single end reads to
61 estimate gene expression, typically around 75 base pairs. Yet, this strategy does not leverage
62 information provided by the full length of the library fragments, particularly the greater
63 specificity in read mapping (or pseudo-alignment) that would be provided by sequencing from
64 both ends of a fragment. Given a fixed cost per base for the Illumina NextSeq, HiSeq X Ten, and
65 NovaSeq instruments, paired-end 75bp reads would be twice as expensive as single-end 75bp
66 sequencing, assuming an equivalent number of sequenced fragments. However, a sizeable
67 fraction of the long-range information provided by paired-end sequencing should still be retained
68 by sequencing shorter paired-end reads. In this study, we evaluate whether that intuition is
69 correct, by assessing whether expression estimates and differential expression results obtained

70 with 2x40 paired end sequencing are more consistent with 2x125 paired end sequencing than
71 either 1x75 or 1x125 single-end sequencing. Our particular focus is on 1x75, as the cost of 1x75
72 and 2x40 are the same per read when sequencing using the Illumina NextSeq (75 cycle kit), a
73 common instrument for RNA-seq studies. Given the fixed cost per base for the HiSeq X Ten and
74 NovaSeq instruments, there are analogous choices to make between single-end vs. paired-end
75 sequencing. For example, for the 100bp NovaSeq kit, a sequencing strategy trade-off exists
76 between sequencing 2x50 vs. 1x100 reads. We conduct these analyses with publicly available
77 Illumina sequence data across 12 different SRA (Additional file 1: Table S1). accessions
78 encompassing diverse experimental designs and multiple model organisms with high quality
79 annotated genomes.
80

81 **RESULTS AND DISCUSSION**

82 Our approach to evaluating the performance of short paired-end reads vs. longer single-end ones
83 is to compare expression estimates and differential expression results derived from these
84 strategies to a truth set. Because we cannot know the "truth", we assume that its closest
85 approximation are results obtained from long, paired-end reads. Thus, we define our gold
86 standard as the results obtained from 2x125 paired-end reads. We then trim these data down to
87 two data sets with read lengths of 40 and 75 base pairs, and calculate Spearman's rank
88 correlations between transcripts-per-million (TPM) estimates based upon 2x40, 1x75 and 1x125
89 with the 2x125 gold standard. We present these correlations in two different ways. First, we
90 examine the distributions of those correlation coefficients across alternative sequencing
91 strategies. Second, we plot the correlations obtained using 2x40 over those obtained with 1x75
92 and 1x125.

93 Kallisto transcript-level TPM estimates made with 2x40 are always more highly
94 correlated with those made with the 2x125 gold standard than those made with data trimmed to
95 1x75 reads (Fig. 1a,c), and perform better than estimates made with data trimmed to 1x125 for
96 all but a few samples from one SRA accession (Fig. 1a,c,e). A similar performance advantage for
97 the 2x40 strategy compared to the 1x75 strategy is observed at the gene level (Fig. 1 b,d,f),
98 although there is one accession where 1x125 performs better than 2x40.
99 RSEM-based TPM estimates show an overall pattern generally consistent with that
100 observed with kallisto (Fig. 2). A handful of samples for which 2x40 performed less well than
101 both 1x75 and 1x125 had low bowtie2 alignment rates to the reference transcripts (for the 2x125
102 libraries, overall alignment rates of 37.1-51.0%), suggesting that library quality issues were
103 responsible, but raising a broader issue concerning whether variation in alignment rates between
104 strategies could explain the relative performance of 2x40 to 1x75 and 1x125.
105 Kallisto does not conduct formal sequence alignment, so for each sample-strategy
106 combination we summed the kallisto transcript-level counts as a proxy for the number of
107 informative reads used in estimating expression. The summed kallisto counts for 2x40 have a
108 weak correlation with performance (with respect to TPM) relative to 1x75 the transcript level,
109 and a moderate correlation at the gene level (Fig 3a,b); the relative differences in counts between
110 2x40 and 1x75 appear uncorrelated with the relative performance of 2x40 and 1x75 (Fig. 3c,d).
111 A similar pattern is observed with respect to 1x125 (Additional file 2: Fig. S1). For RSEM-
112 derived expression estimates, we found no impact of the percent of reads uniquely aligning on
113 the relative performance of 2x40, but at the gene level found that relative differences in overall
114 alignment rate had some effect on the robustness of TPM estimates of 2x40 relative to both 1x75
115 and 1x125 (Additional file 3: Fig. S2,S3). For both kallisto and RSEM, while there are no

116 obvious taxonomic effects there are clear experiment-specific effects, manifested as differences
117 in relative 2x40 performance among accessions derived from the same species. Similar to the
118 above results based upon expression correlations, looking across all accessions, for most samples
119 root-mean-square error of expression estimates based upon 1x75 and 1x125 was greater than for
120 2x40 (Fig. 4).

121 We next evaluated the extent to which differences in performance with respect to
122 estimating expression levels translated into downstream effects on tests of differential
123 expression. Within each SRA accession and for each pair of conditions and for each sequencing
124 strategy, we conducted Wald tests with, sleuth, limma-voom, and DESeq2. Similar to our
125 evaluation of expression estimate robustness, we defined "true" differential expression signals as
126 those recovered with 2x125, given a false discovery rate (FDR) of 0.01. We then calculated
127 seven performance metrics on a per-accession basis for each differential expression method -
128 sequencing strategy combination.

129 Sequencing strategy clearly impacts downstream differential expression analyses.
130 Looking across all three differential expression methods, at both the transcript and gene level, for
131 the majority of accessions false negative rates are lower for 2x40 compared to 1x75 (Figure 5;
132 Additional file 4: Fig. S4, S5; Additional file 5: Table S2). The empirical false discovery rate—
133 that is, for a strategy being evaluated, the proportion of putative differentially expressed features
134 with a Benjamini-Hochberg adjusted p-value below our chosen FDR threshold of 0.01for which
135 the adjusted p-value is >0.01 with 2x125—was also lower for 2x40 compared to both 1x75 and
136 x125, although it was consistently higher than 0.01, the threshold used to classify Wald tests as
137 significant, including the 2x125 analyses with which we defined true positives and negatives
138 (Figure 5; Additional file 4: Fig. S4, S5; Additional file 5: Table S2). FDR control was poorer at

139 the transcript than the gene level, and, as previously reported, DESeq2 did not control FDR as
140 well as limma-voom and sleuth (18). At both the transcript and gene level, AUC is greater for
141 2x40 than 1x75 for all but a handful of method-accession combinations (Figure 5, Additional file
142 4: Fig. S4, S5; Additional file 5: Table S2). Additional metrics indicate a similar overall
143 performance advantage of 2x40 compared to 1x75 (Additional file 4: Table S2). For all metrics,
144 the short paired-end strategy outperforms 1x125 for the majority of method-accession
145 combinations, although the magnitude of performance advantage of 2x40 compared to 1x125 is
146 smaller than when compared with 1x75 smaller (Figure 5, Additional file 4: Fig. S4,S5;
147 Additional file 5: Table S2).
148
149 **CONCLUSIONS**
150 When designing a gene expression study using bulk RNA-seq, given a fixed budget, more robust
151 expression estimates and differential expression test results can be obtained from short paired-
152 end reads compared to longer single-end reads. In other words, the extra information provided by
153 reads pairs during the mapping or pseudoalignment phase outweighs any potential penalty of
154 short reads. In almost all cases, short paired-end reads trump single-end reads that sequence
155 approximately the same number of bases. Short paired-end reads provide an additional benefit in
156 that they can identify a large fraction of differentially expressed isoforms, for the cost of gene-
157 level analysis with longer single-end reads. From a practical perspective, neither NextSeq nor
158 NovaSeq sequencing kits distinguish between single and paired-end libraries. In the experience
159 of support staff at the Bauer sequencing core facility at Harvard, one can get 92 bp from the
160 NextSeq 75 bp kit, such that 2x40 is a viable sequencing design. Alternatively, the smallest
161 NovaSeq kit is for 100 cycles, which can be used to sequence paired 50bp reads, allowing one to

162 leverage that instrument's much smaller cost per sequenced base. Overall, our findings suggest
163 that for a given number of sequencing cycles, the most efficient design for RNA-seq expression
164 analysis is sequencing shorter paired reads, rather than longer single-end reads.
165
166 **METHODS**
167
168 **Datasets**
169 From NCBI's Short Read Archive, we downloaded paired-end RNA-seq reads from 12 different
170 studies (Table 1), representing model organisms with well-annotated reference genomes: mouse
171 (*Mus musculus*), human (*Homo sapiens*), yeast (*Saccharomyces cerevisiae*), fly (*Drosophila*
172 *melanogaster*), and worm (*Caenorhabditis elegans*). In selecting data sets, we required minimum
173 pre-processing read lengths to be ≥ 125 bp, and experimental designs to contain ≥ 2 experimental
174 conditions.
175
176 **Short-read processing**
177 We removed adapters and trimmed reads with NGmerge (21). For each sample, we trimmed
178 reads to 125, 75, and 40bp using awk. For downstream analyses, we used 2x125, 2x40,1x125,
179 and 1x75 trimmed data sets. Single-end data sets were comprised of the first read from the
180 original paired reads. Reads were trimmed from the distal ends, such that trimming mimicked
181 how sequencing to different lengths would be carried out on an Illumina sequencer.
182
183 **Expression analyses**

184 For each sample-sequencing strategy combination, we estimated transcript-level expression with
185 kallisto (22). We designated abundance estimates derived from 2x125 as our gold standard
186 against which to benchmark all other strategies. We used Spearman's rank correlations to
187 quantify the degree of concordance between alternative sequencing strategies and the gold
188 standard in terms of transcript per million (TPM). We performed similar analyses at the gene
189 level by aggregating transcript-level TPMs using the R package Tximport (19). For yeast, which
190 has no alternative splicing, analyses were only conducted at the transcript level. As an additional
191 estimate of the correspondence between alternative sequencing strategies with 2x125 we also
192 calculated root-mean-square error (RMSE).

193 Because a large proportion of RNA-seq applications involve tests of differential
194 expression (DE), we compared differential expression test results between the gold standard
195 2x125 and alternative strategies. Specifically, for each experiment we conducted all possible
196 pairwise Wald tests of transcript-level DE with sleuth (18) version 0.30.0, and used p-value
197 aggregation (https://pachterlab.github.io/sleuth_walkthroughs/pval_agg/analysis.html) for gene-
198 level analyses. To compare performance of alternative strategies relative to the gold standard, we
199 calculated false negative and false positive rates, defining significant tests as those with
200 Benjamini-Hochberg adjusted p-values ≤ 0.01.

201 In order to evaluate whether our findings were pipeline-specific, perhaps related to
202 particular features of tools such as kallisto that rely on pseudo-alignment, we also performed
203 expression analyses with an alternative pipeline, involving estimation of expression levels with
204 RSEM (23) version 1.2.29 based upon alignments with bowtie2 version 2.3.2, (24). When SRA
205 metadata enabled us to determine that libraries were stranded, we included the relevant RSEM
206 command line arguments for restricting bowtie2 alignments to the correct read pair

207 configuration. Using these expression estimates, we conducted differential expression tests with
208 two different methods: limma-voom (25) version 3.38.3 and DESeq2 (26) version 1.22.2. For
209 limma-voom, we only conducted tests where, for each gene or transcript,there were ≥ 2 samples
210 with counts per million (CPM) ≥ 1. For limma-voom, we employed TMM normalization, and the
211 voomWithQualityWeights option. Sleuth and limma-voom have previously been demonstrated to
212 be top performers that correctly control FDR (22), and while DESeq2 has been shown to not
213 control FDR as well as these two methods (22), we included it because it is a commonly used
214 differential expression method. For all differential expression analyses, we calculated
215 performance metrics for Wald tests where, at the specified false discovery rate (FDR) threshold
216 of 0.01, there were ≥ 50 significantly differentially expressed features in all sequencing
217 strategies. Metrics were calculated relative to 2x125, i.e. significant and non-significant results
218 obtained with 2x125 were treated as true positives and negatives, respectively. Those metrics
219 were the empirical false discovery rate, false positive rate, false negative rate, sensitivity,
220 specificity, precision, and area under the curve (AUC) based upon receiver-operator statistics.
221 These metrics were calculated with python scripts available at our associated github repository
222 (see below). As the multiple-test adjusted p-values are the standard metric with which to
223 determine whether particular transcripts or genes are differentially expressed, for AUC
224 calculations we used these values as the response variable relative to the "true" classification of
225 differences derived from tests based upon the 2x125 data. All reference genomes and gtf-
226 formatted annotations against which to estimate expression were downloaded from Ensembl, and
227 transcripts for kallisto-based analyses were extracted using the *extract-reference-transcripts*
228 utility in RSEM.
229

DECLARATIONS

Availability of data and materials

The datasets analyzed during the current study and which support the conclusions of this article are publicly available Illumina short read data available on NCBI's Short Read Archive (https://www.ncbi.nlm.nih.gov/sra). Specific accessions are listed in Additional File 1: Table S1. Command lines for executing NGmerge, kallisto, RSEM, sleuth, and limma-voom analyses are available on GitHub at https://github.com/harvardinformatics/rnaseq_readlength_assessment.

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Funding

This study was funded by the Harvard University Faculty of Arts and Sciences.

253 **Authors' contributions**

254 TBS and AHF designed the experiments, AHF and JMG analyzed the data, and AHF wrote the
255 paper with TBS.

256 **Acknowledgments**

257 The computations in this paper were run on the Odyssey cluster supported by the FAS Division
258 of Science, Research Computing Group at Harvard University. We thank Claire Reardon,
259 director of the Bauer Core at Harvard University for helpful discussions during the study design
260 phase of our study, and for feedback on preliminary results. We thank Emma White at the Bauer
261 Core for details concerning available sequencing read lengths and kit sizes for Illumina
262 instruments. We also thank members from the Harvard FAS Informatics Group for feedback on
263 preliminary results produced by this study. This work was conducted on the traditional territory
264 of the Wampanoag and Massachusett peoples.

265

266 **REFERENCES**

267 1. Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, et al. The evolution of
268 gene expression levels in mammalian organs. Nature. 2011 Oct 19;478(7369):343–8.

269 2. Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, et al. The evolutionary
270 landscape of alternative splicing in vertebrate species. Science. 2012 Dec
271 21;338(6114):1587–93.

272 3. Frumkin I, Yofe I, Bar-Ziv R, Gurvich Y, Lu Y-Y, Voichek Y, et al. Evolution of intron splicing
273 towards optimized gene expression is based on various Cis- and Trans-molecular
274 mechanisms. PLOS Biol. 2019 Aug 23;17(8):e3000423.

275 4. Crowley JJ, Zhabotynsky V, Sun W, Huang S, Pakatci IK, Kim Y, et al. Analyses of allele-
276 specific gene expression in highly divergent mouse crosses identifies pervasive allelic
277 imbalance. Nat Genet. 2015 Apr;47(4):353–60.

278 5. Verta J-P, Jones FC. Predominance of cis-regulatory changes in parallel expression
279 divergence of sticklebacks. de Meaux J, Tautz D, editors. eLife. 2019 May 15;8:e43785.

280 6. Castel SE, Cervera A, Mohammadi P, Aguet F, Reverter F, Wolman A, et al. Modified
281 penetrance of coding variants by cis-regulatory variation contributes to disease risk. Nat
282 Genet. 2018;50(9):1327–34.

283 7. Gandal MJ, Zhang P, Hadjimichael E, Walker RL, Chen C, Liu S, et al. Transcriptome-wide
284 isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. Science. 2018 Dec
285 14;362(6420):eaat8127.

286 8. Campbell-Staton SC, Cheviron ZA, Rochette N, Catchen J, Losos JB, Edwards SV. Winter
287 storms drive rapid phenotypic, regulatory, and genomic shifts in the green anole lizard.
288 Science. 2017 Aug 4;357(6350):495–8.

289 9. Mallarino R, Henegar C, Mirasierra M, Manceau M, Schradin C, Vallejo M, et al.
290 Developmental mechanisms of stripe patterns in rodents. Nature. 2016 24;539(7630):518–
291 23.

292 10. MacManes MD. On the optimal trimming of high-throughput mRNA sequence data. Front
293 Genet [Internet]. 2014 Jan 31 [cited 2019 Sep 16];5. Available from:
294 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3908319/

295 11. Williams CR, Baccarella A, Parrish JZ, Kim CC. Trimming of sequence reads alters RNA-Seq
296 gene expression estimates. BMC Bioinformatics. 2016 Feb 25;17:103.

297 12. Tarazona S, García-Alcalde F, Dopazo J, Ferrer A, Conesa A. Differential expression in RNA-
298 seq: a matter of depth. Genome Res. 2011 Dec;21(12):2213–23.

299 13. Liu Y, Zhou J, White KP. RNA-seq differential expression studies: more sequence or more
300 replication? Bioinforma Oxf Engl. 2014 Feb 1;30(3):301–4.

301 14. Todd EV, Black MA, Gemmell NJ. The power and promise of RNA-seq in ecology and
302 evolution. Mol Ecol. 2016 Mar;25(6):1224–41.

303 15. Robles JA, Qureshi SE, Stephen SJ, Wilson SR, Burden CJ, Taylor JM. Efficient experimental
304 design and analysis strategies for the detection of differential expression using RNA-
305 Sequencing. BMC Genomics. 2012 Sep 17;13(1):484.

306 16. Fonseca NA, Marioni J, Brazma A. RNA-Seq Gene Profiling - A Systematic Empirical
307 Comparison. PLOS ONE. 2014 Sep 30;9(9):e107026.

308 17. Kanitz A, Gypas F, Gruber AJ, Gruber AR, Martin G, Zavolan M. Comparative assessment of
309 methods for the computational inference of transcript isoform abundance from RNA-seq
310 data. Genome Biol. 2015 Jul 23;16:150.

311 18. Pimentel H, Bray NL, Puente S, Melsted P, Pachter L. Differential analysis of RNA-seq
312 incorporating quantification uncertainty. Nat Methods. 2017 Jul;14(7):687–90.

313 19. Soneson C, Love MI, Robinson MD. Differential analyses for RNA-seq: transcript-level
314 estimates improve gene-level inferences. F1000Research. 2015 Dec 30;4:1521.

315 20. Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of
316 RNA-seq data. BMC Bioinformatics. 2013 Mar 9;14(1):91.

317 21. Gaspar JM. NGmerge: merging paired-end reads via novel empirically-derived models of
318 sequencing errors. BMC Bioinformatics. 2018 Dec 20;19(1):536.

319 22. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq
320 quantification. Nat Biotechnol. 2016 May;34(5):525–7.

321 23. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or
322 without a reference genome. BMC Bioinformatics. 2011;12:323.

323 24. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012
324 Apr;9(4):357–9.

325 25. Law CW, Chen Y, Shi W, Smyth GK. voom: precision weights unlock linear model analysis
326 tools for RNA-seq read counts. Genome Biol. 2014;15:R29.

327 26. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-
328 seq data with DESeq2. Genome Biol. 2014 Dec 5;15(12):550.

329

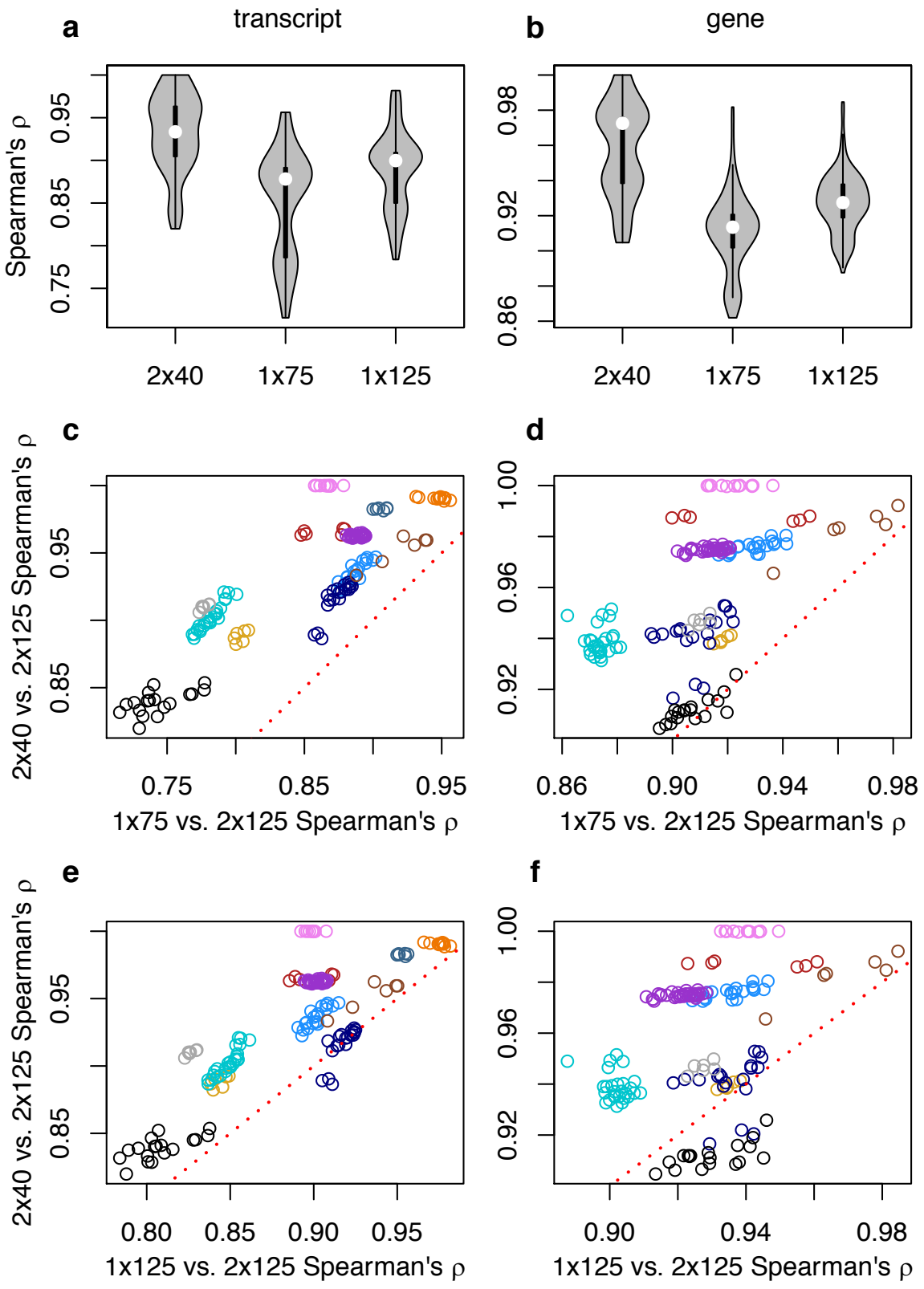

Fig. 1. Spearman's rank correlations for kallisto-derived transcripts per million (TPM) between the gold standard paired-end 2x125 strategy and alternative strategies. Violin plots of **(a)** transcript and **(b)** gene-level inference. Comparison of correlations with 2x125 between 2x40 and 1x75 for **(c)** transcript and **(d)** genes, and between 2x40 and 1x125 for **(e)** transcript and **(f)** genes. For **c-**f, points above the red dotted line are samples where estimates of expression from 2x40 is more highly correlated with the gold standard than the contrasted single-end strategy.

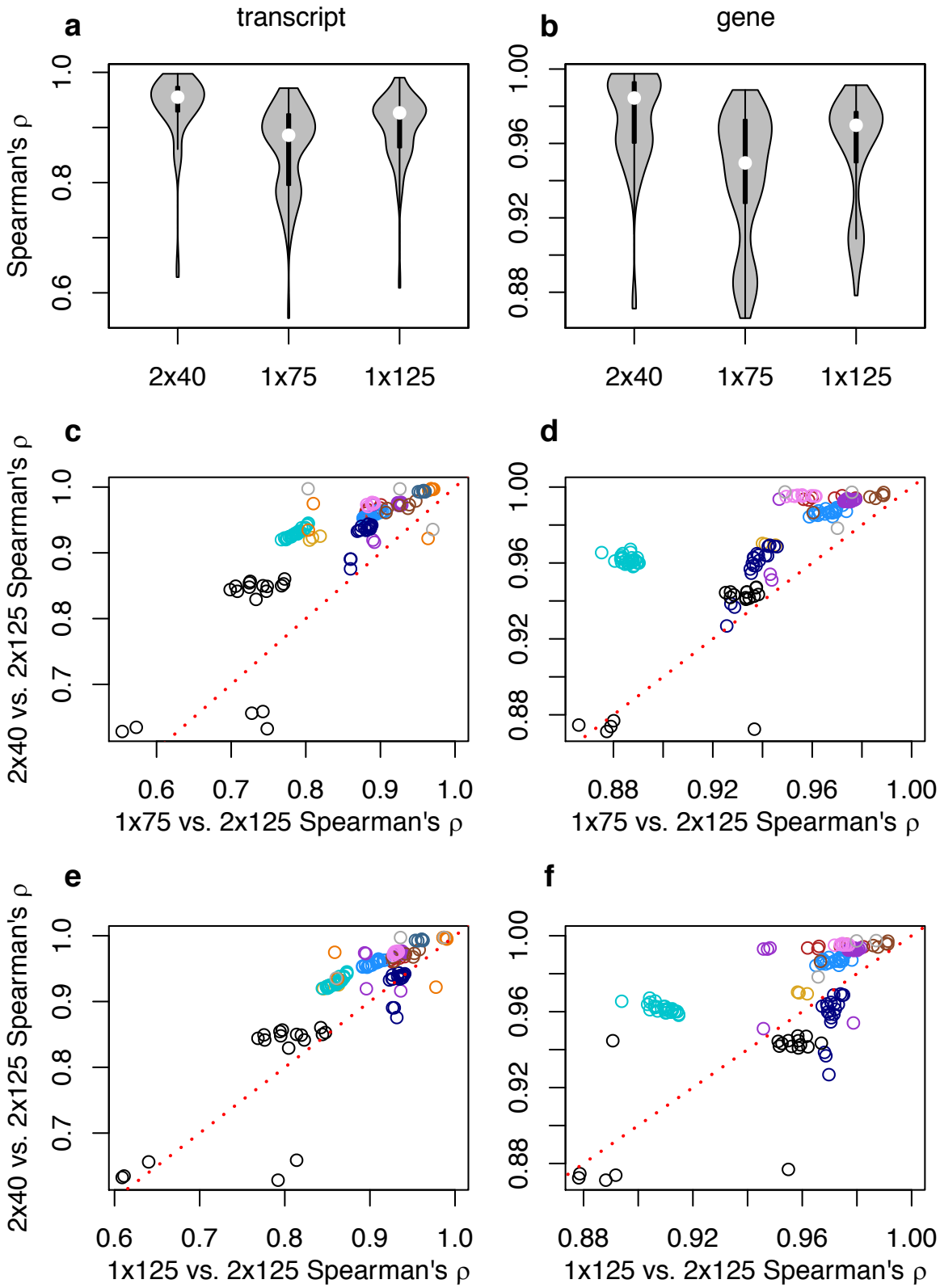

Fig. 2. Spearman's rank correlations for RSEM-derived transcripts per million (TPM) between the gold standard paired-end 2x125 strategy and alternative strategies. Violin plots of **(a)** transcript and **(b)** gene-level inference. Comparison of correlations with 2x125 between 2x40 and 1x75 for **(c)** transcript and **(d)** genes, and between 2x40 and 1x125 for **(e)** transcript and **(f)** genes. For **c**-f, points above the red dotted line are samples where estimates of expression from 2x40 is more highly correlated with the gold standard than the contrasted single-end strategy.

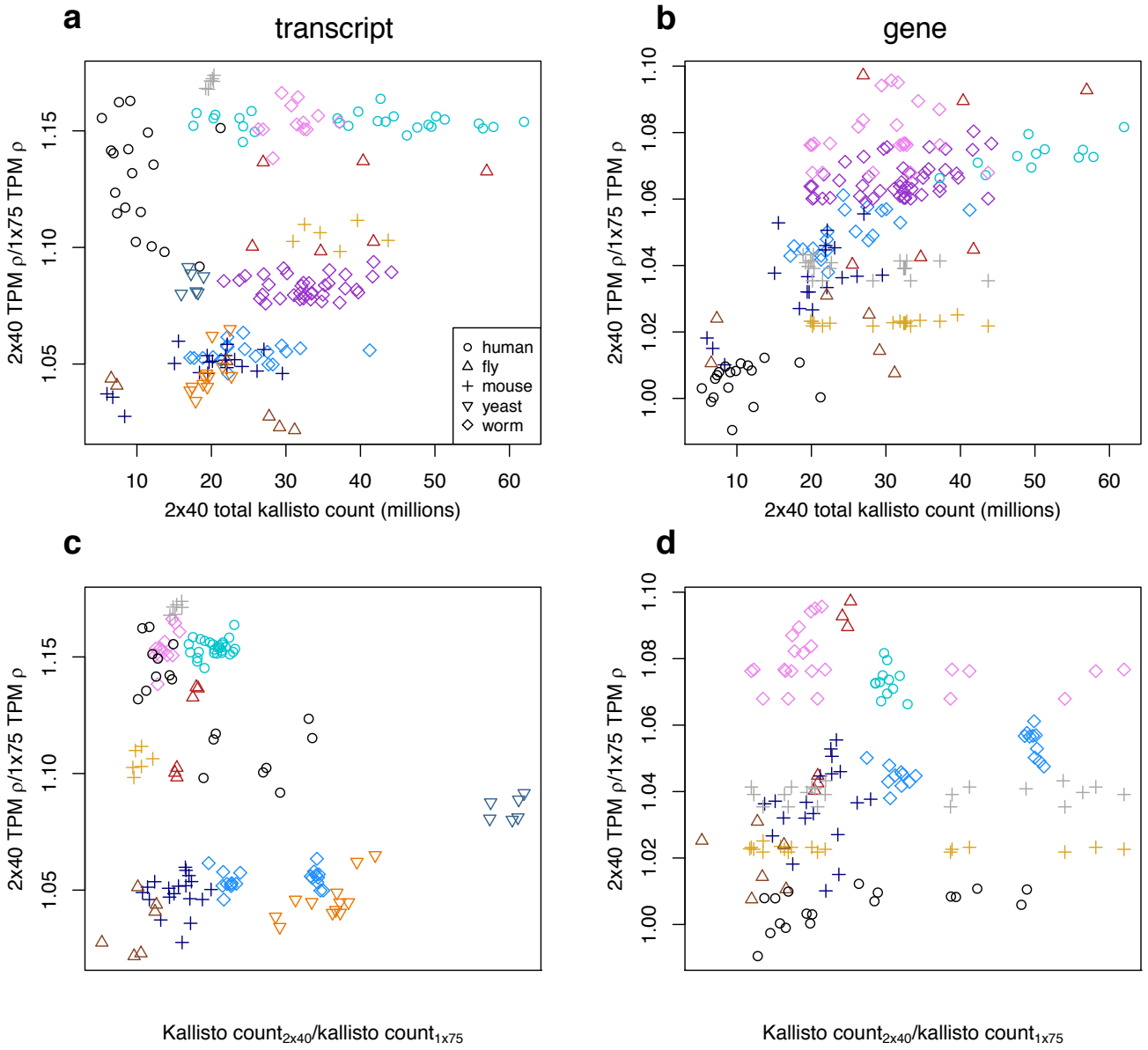

Fig. 3. Effects of kallisto pseudo-alignment efficiency on differences in TPM estimate Spearman's rank correlations between 2x40 vs. 1x75. Efficiency of 2x40 is presented as the total kallisto count for 2x40 for **(a)** transcripts and **(b)** genes, and contrasted with the efficiency of 1x75 in the form of count ratios for **(c)** transcripts and **(d)** genes.

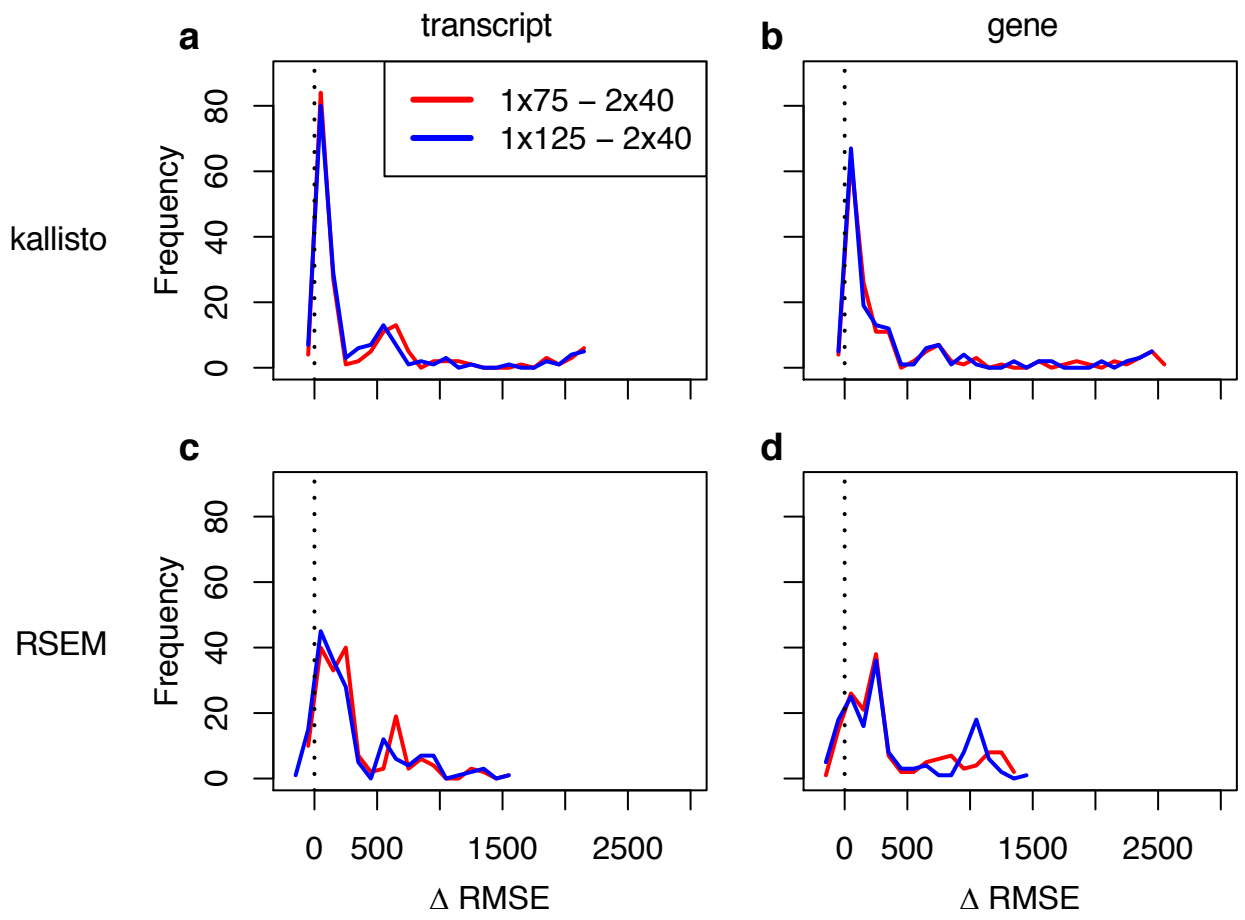

Fig. 4. Histograms of differences in root-mean-square error (RMSE) between 2x40 and either 1x75 or 1x125, where RMSE is calculated using 2x125 as the gold standard. RMSE differences for kallisto for **(a)** transcripts and **(b)** genes, and RSEM for **(c)** transcripts and **(d)** genes. The vertical black dashed line indicates the threshold above which RMSE for either 1x75 or 1x125 is greater than that for 2x40.

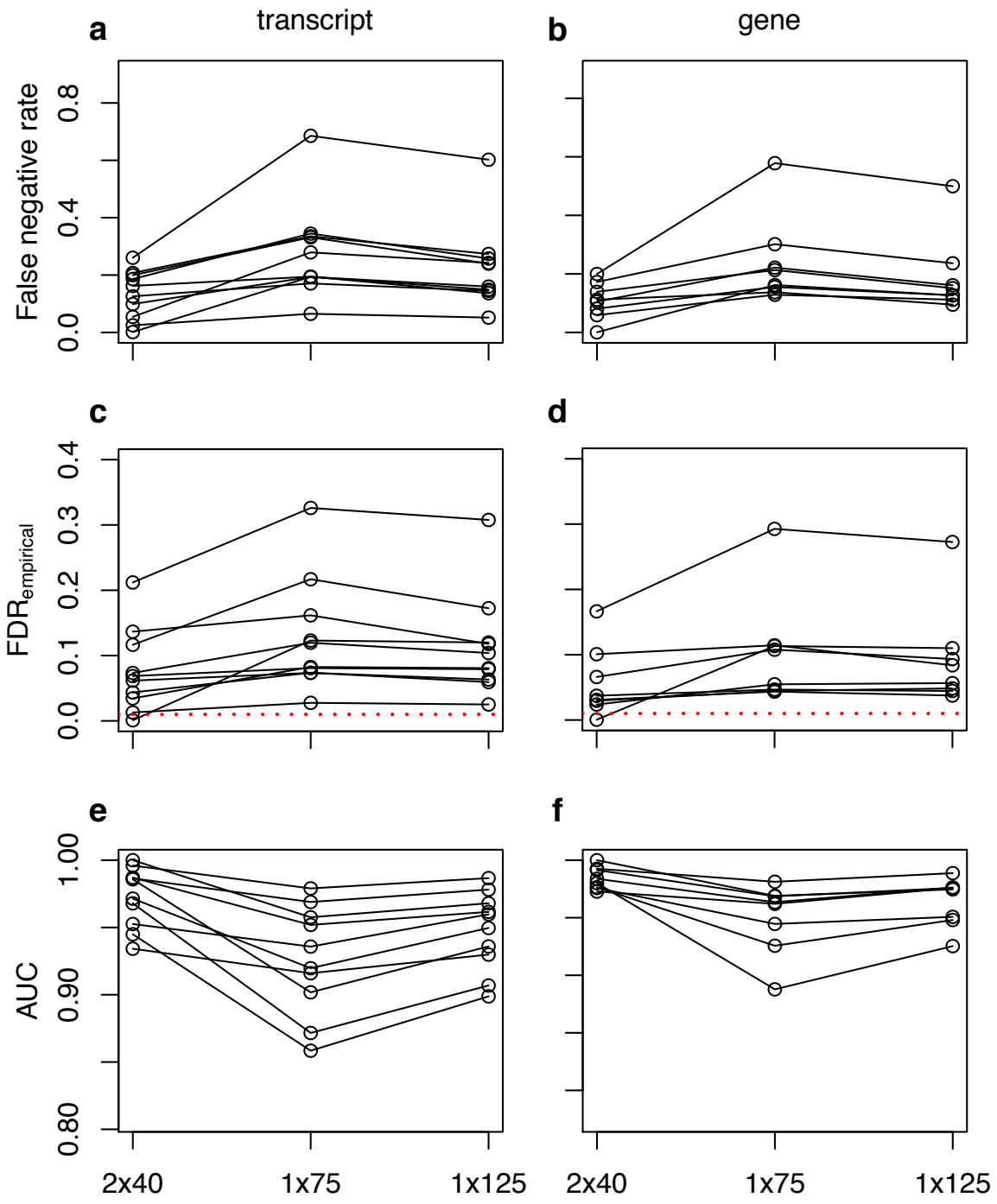

Fig. 5. Variation across SRA accessions in **(a,b)** false negative rate, **(c,d)** empirical false discovery rate, and **(e,f)** AUC between 2x40, 1x75 and 1x125 sequencing strategies based upon differential expression testing with sleuth. Differences are plotted as means for Wald tests within accessions, for **(a,c,e)** transcripts and **(b,d,f)** genes.