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 reads; in nearly all cases, those correlations are greater than for 1x125, despite the greater total number of sequenced bases for the latter. In addition, differential expression false negative rates are consistently lower for 2x40 compared to 1x125.

Conclusion: Researchers seeking a cost-effective approach for gene-level expression analysis should prefer short paired-end reads over a longer single-end strategy. False negative rates for gene-level differential expression tests with 2x40 relative to 2x125 are in fact quite modest. In addition, short paired-end reads will also give reasonably robust expression estimates and differential expression results at the isoform level.

BACKGROUND

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

60 While the typical length of sequenced reads has increased as costs have decreased, when
61 an annotated genome is available it remains standard procedure to sequence single end reads to
62 estimate gene expression, typically around 75 base pairs. Yet, this strategy does not leverage
63 information provided by the full length of the library fragments, particularly the greater
64 specificity in read mapping (or pseudo-alignment) that would be provided by sequencing from
65 both ends of a fragment. Given a fixed cost per base for a given Illumina instrument, paired-end
66 75bp reads would be twice as expensive as single-end 75bp sequencing, assuming an equivalent
67 number of sequenced fragments. Intuitively, however, a sizeable fraction of the long-range
68 information provided by paired-end sequencing should still be retained by sequencing shorter

69 paired-end reads. In this study, we evaluate whether that intuition is correct, by assessing
70 whether expression estimates and differential expression results obtained with 2x40 paired end
71 sequencing are more consistent with 2x125 paired end sequencing than either 1x75 or 1x125
72 single-end sequencing. Our particular focus is on 1x75, as the cost of 1x75 and 2x40 are
73 approximately the same per read when sequencing using the Illumina NextSeq, a common
74 instrument for RNA-seq studies. As with NextSeq, NovaSeq sequencing kits do not distinguish
75 between single and paired-end libraries, and are configured a particular number of bases. Thus,
76 for the 100bp NovaSeq kit, a similar sequencing strategy trade-off exists between sequencing
77 2x50 vs. 1x100 reads. We conduct these analyses with publicly available Illumina sequence data
78 across 12 different SRA (Additional file 1: Table S1). accessions encompassing diverse
79 experimental designs and multiple model organisms with high quality 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 examine relative differences in these correlations in the form of "ratio

92 plots" that have, on the x-axis, the ratio of the 1x125 expression correlation with 2x125 to that of
93 2x40, and on the y-axis, the ratio of the expression correlation between 1x75 and 2x125 to that of
94 2x40. In this manner, we can visualized in bivariate space, the frequency of cases where 2x40
95 outperforms both 1x75 and 1x125, where 2x40 only outperforms 1x75 (or 1x125), or where 2x40
96 is inferior to both single-end strategies.

97 Kallisto transcript-level TPM estimates made with 2x40 are always more highly
98 correlated with those made with the 2x125 gold standard than those made with data trimmed to
99 1x75 reads (Fig. 1a,c), and perform better than estimates made with data trimmed to 1x125 for
100 all but a few samples from one SRA accession (Fig. 1c). A similar performance advantage for
101 the 2x40 strategy compared to the 1x75 strategy is observed at the gene level (Fig. 1 b,d).
102 RSEM-based TPM estimates show a generally consistent overall pattern (Fig. 2), although at the
103 transcript level four samples have higher performance (better correlations with estimated from
104 untrimmed data) for one or both single-end strategies than the 2x40 strategy. These samples are
105 widely dispersed in the rank correlation ratio plots (Fig. 2c); two of the four samples have an
106 aberrant pattern where TPM estimates from 1x75 reads more highly correlated with estimated
107 from 2x125 reads than with estimates from 1x125 reads, despite having 67% more nucleotides
108 sequenced per sample. Upon closer inspection, we discovered that samples comprising this
109 unusual pattern had low bowtie2 alignment rates to the reference transcripts (for the 2x125
110 libraries, overall alignment rates of 37.1-51.0%), suggesting library quality issues were
111 responsible.

112 More generally, a number of different library-specific factors could potentially associate
113 with variation in the relative performance of a 2x40 strategy. One important faction is simply the
114 overall number of informative reads in an experiment. To estimate this, we summed the kallisto

115 transcript-level counts for each sample. We find that as the number of informative reads
116 increases, the relative performance of a 2x40 strategy (compared to 1x125) improves (Fig 3). We
117 did not perform the same analysis for the 1x75 strategy, as using 2x40 reads always
118 outperformed 1x75 reads at the transcript level with kallisto. The correlation between total
119 number of sequenced reads and relative performance is slightly weaker ($\rho = 0.46$, $p = 4.4 \times 10^{-}$
120 ^{10}), indicating that read mappability in addition to read number, is important for robustness of
121 expression estimates for short PE reads. While there are no obvious taxonomic effects, there are
122 clear library-specific effects, manifested as differences in relative 2x40 performance among
123 accessions derived from the same species.

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

130 Sequencing strategy clearly impacts downstream differential expression analyses. Using
131 either the kallisto-sleuth or RSEM-limma pipeline, at both the transcript and gene level, false
132 negative rate relative to the 2x125 gold standard was elevated in 1x75 relative to 2x40 (Fig. 4),
133 while false positive rates remained relatively consistent (Additional file 1: Table S1). With few
134 exceptions, false negative rate was very similar between 2x40 and 1x125, or slightly lower in the
135 former.

136

137 **CONCLUSIONS**

138 When designing a gene expression study using bulk RNA-seq, given a fixed budget, more robust
139 expression estimates and differential expression test results can be obtained from short paired-
140 end reads compared to longer single-end reads. In other words, the extra information provided by
141 reads pairs during the mapping or pseudoalignment phase outweighs any potential penalty of
142 short reads. In almost all cases, short paired-end reads trump single-end reads that sequence
143 approximately the same number of bases. Short paired-end reads provide an additional benefit in
144 that they can identify a large fraction of differentially expressed isoforms, for the cost of gene-
145 level analysis with longer single-end reads. From a practical perspective, neither NextSeq nor
146 NovaSeq sequencing kits distinguish between single and paired-end libraries. In the experience
147 of support staff at the Bauer sequencing core facility at Harvard, one can get 80 bp from the
148 NextSeq 75 bp kit, such that 2x40 is a viable sequencing design. Alternatively, the smallest
149 NovaSeq kit is for 100 cycles, which can be used to sequence paired 50bp reads, allowing one to
150 leverage that instrument's much smaller cost per sequenced base. Overall, our findings suggest
151 that for a given number of sequencing cycles, the most efficient design for RNA-seq expression
152 analysis is allocating 50% of the cycles to paired end reads, instead of 100% of the cycles to a
153 longer, single-end read.
154
155 **METHODS**
156
157 **Datasets**
158 From NCBI's Short Read Archive, we downloaded paired-end RNA-seq reads from 12 different
159 studies (Table 1), representing model organisms with well-annotated reference genomes: mouse
160 (*Mus musculus*), human (*Homo sapiens*), yeast (*Saccharomyces cerevisiae*), fly (*Drosophila*

161 *melanogaster*), and worm (*Caenorhabditis elegans*). In selecting data sets, we required minimum
162 pre-processing read lengths to be ≥ 125 bp, and experimental designs to contain ≥ 2 experimental
163 conditions.
164
165 **Short-read processing**
166 We removed adapters and trimmed reads with NGmerge (16–20). For each sample, we trimmed
167 reads to 125, 75, and 40bp using awk. For downstream analyses, we used 2x125, 2x40,1x125,
168 and 1x75 trimmed data sets. Single-end data sets were comprised of R1 from the original paired
169 reads.
170
171 **Expression analyses**
172 For each sample-sequencing strategy combination, we estimated transcript-level expression with
173 kallisto (22). We designated abundance estimates derived from 2x125 as our gold standard
174 against which to benchmark all other strategies. We used spearman's rank correlations to
175 quantify the degree of concordance between alternative sequencing strategies and the gold
176 standard in terms of transcript per million (TPM). We performed similar analyses at the gene
177 level by aggregating transcript-level TPMs using the R package Tximport (19).
178 Because the majority of RNA-seq applications involve tests of differential expression
179 (DE), we compared differential expression test results between the gold standard 2x125 and
180 alternative strategies. Specifically, for each experiment we conducted all possible pairwise Wald
181 tests of transcript-level DE with sleuth (18) version 0.30.0, and used p-value aggregation
182 (https://pachterlab.github.io/sleuth_walkthroughs/pval_agg/analysis.html) for gene-level
183 analyses. To compare performance of alternative strategies relative to the gold standard, we

184 calculated false negative and false positive rates, defining significant tests as those with
185 Benjamini-Hochberg adjusted p-values ≤ an FDR of 0.01.
186 In order to evaluate whether our findings were pipeline-specific, perhaps related to
187 particular features of tools such as kallisto that rely on pseudo-alignment, for 5 of the 12 SRA
188 accessions, we performed expression analyses with an alternative pipeline, involving estimation
189 of expression levels with RSEM (23) version 1.2.29 based upon alignments with bowtie2 version
190 2.3.2, (24) and differential expression testing with limma-voom (25) version 3.38.3. For limma-
191 voom, we only conducted tests where, for each gene or transcript,there were ≥ 2 samples with
192 counts per million (CPM) ≥ 1. In addition, we employed TMM normalization, and the
193 voomWithQualityWeights option. Sleuth and limma-voom have previously been demonstrated to
194 be top performers that correctly control FDR (22). When SRA metadata enabled us to determine
195 that libraries were stranded, we included the relevant RSEM command line arguments for
196 restricting bowtie2 alignments to the correct read pair configuration. For both sleuth and limma-
197 voom based analyses, we only calculated performance metrics for Wald tests where, at the
198 specified false discovery rate (FDR) of 0.01, there were ≥ 50 significantly differentially
199 expressed features. All reference genomes and gtf-formatted annotations against which to
200 estimate expression were downloaded from Ensembl, and transcripts for kallisto-based analyses
201 were extracted using the *extract-reference-transcripts* utility in RSEM.
202
203 **CODE AVAILABILITY**
204 Command lines for executing NGmerge, kallisto, RSEM, sleuth, and limma-voom analyses are
205 available on GitHub at https://github.com/harvardinformatics/rnaseq_readlength_assessment.
206

ACKNOWLEDGMENTS

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

AUTHORS' CONTRIBUTIONS

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

REFERENCES

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

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

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

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

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

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

238 7. Gandal MJ, Zhang P, Hadjimichael E, Walker RL, Chen C, Liu S, et al. Transcriptome-wide
239 isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. Science. 2018 Dec
240 14;362(6420):eaat8127.

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

244 9. Mallarino R, Henegar C, Mirasierra M, Manceau M, Schradin C, Vallejo M, et al.
245 Developmental mechanisms of stripe patterns in rodents. Nature. 2016 24;539(7630):518–
246 23.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

282

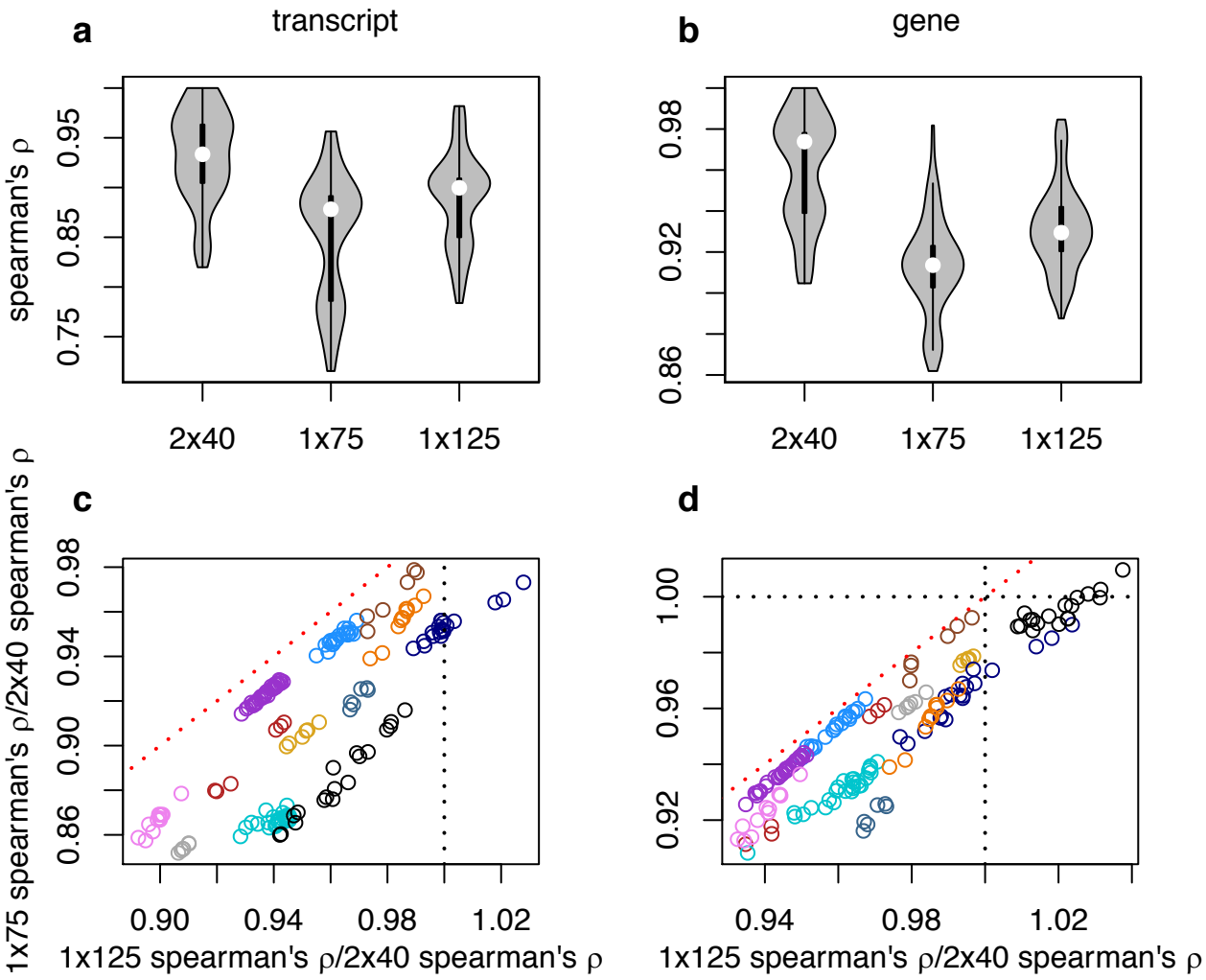

Fig. 1. Spearman rank correlations for kallisto-derived transcripts per million (TPM) between the gold standard paired-end 2x125 strategy and alternative strategies for **(a)** transcript and **(b)** gene-level inference. Comparison of correlations with 2x125 in the form of ratios for **(c)** transcript and **(d)** gene levels. Points above the red dotted line and to the right of the black dotted line indicate samples (colored by SRA accession) where 1x75 and 1x125, respectively, are more highly correlated with the gold standard that 2x40.

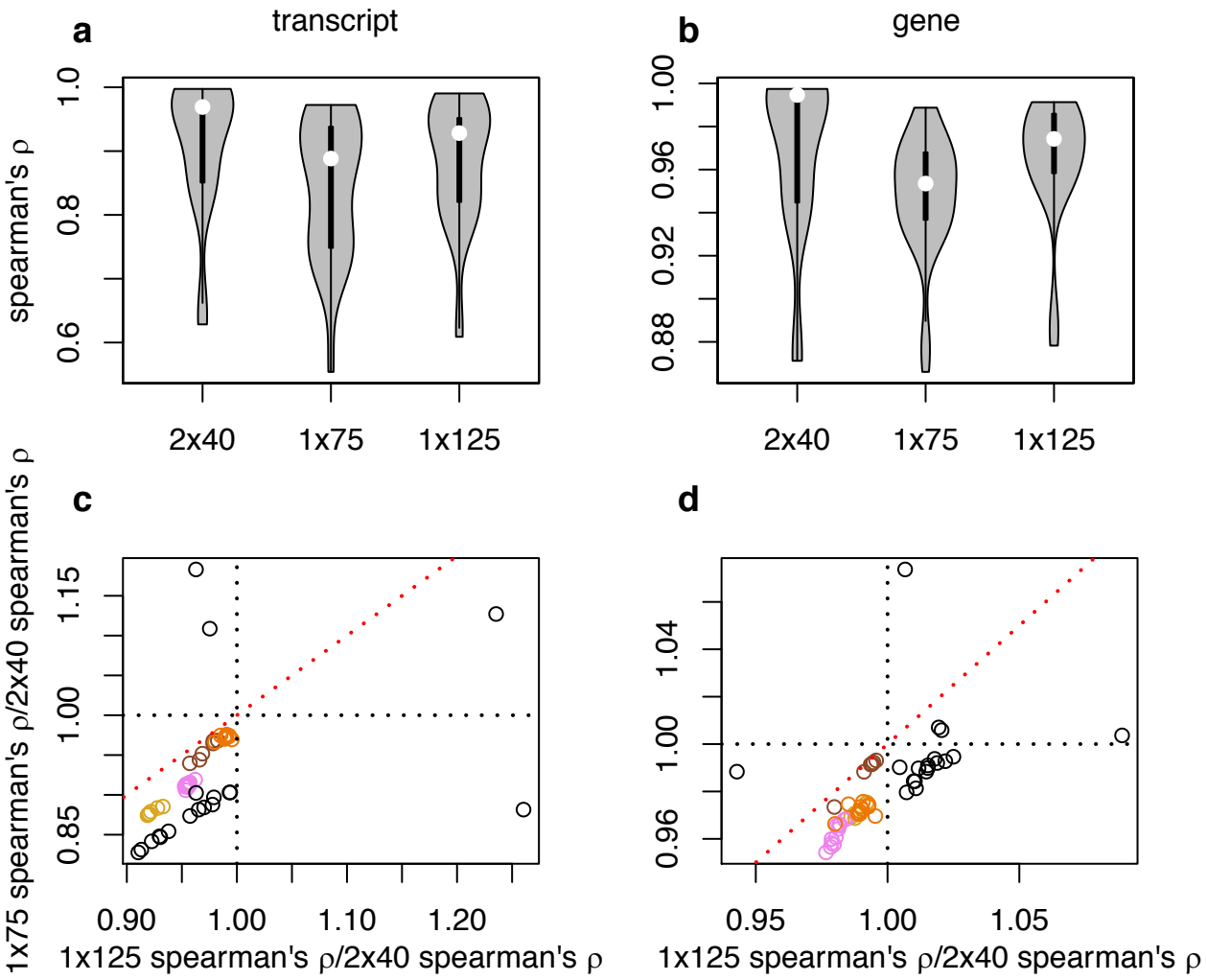

Fig. 2. Spearman rank correlations for RSEM-derived transcripts per million (TPM) between the gold standard paired-end 2x125 strategy and alternative strategies for **(a)** transcript and **(b)** gene-level inference. Comparison of correlations with 2x125 in the form of ratios for **(c)** transcript and **(d)** gene levels. Points above the red dotted line and to the right of the black dotted line indicate samples (colored by SRA accession) where 1x75 and 1x125, respectively, are more highly correlated with the gold standard that 2x40.

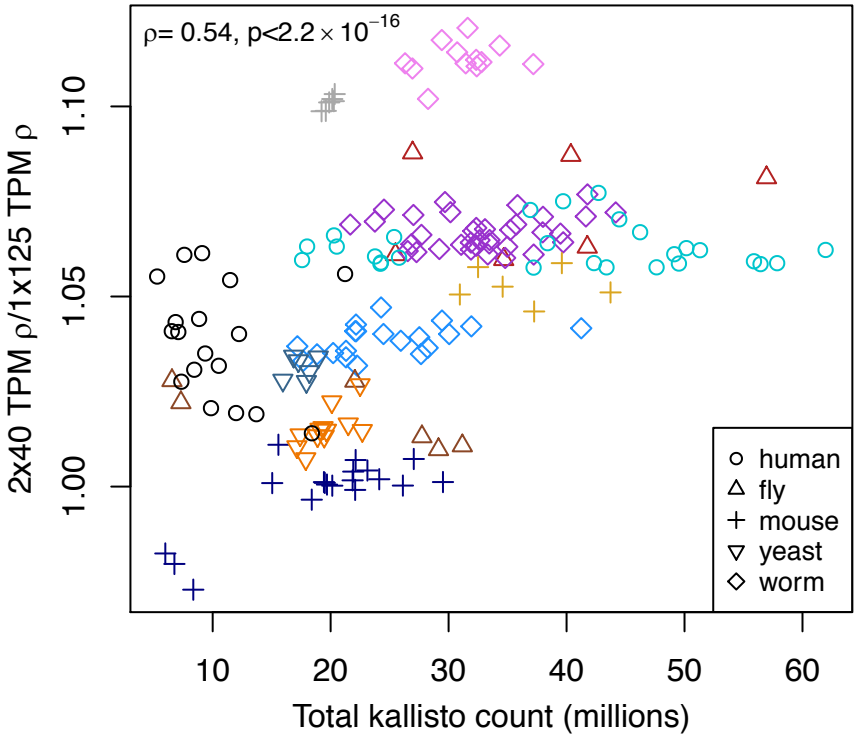

Fig. 3. Ratio of the Spearman rank correlation between 2x40 transcript-level TPM and 2x125 TPM to that between 1x125 and 2x125, as a function of the sum of kallisto counts across all transcripts for a given library.

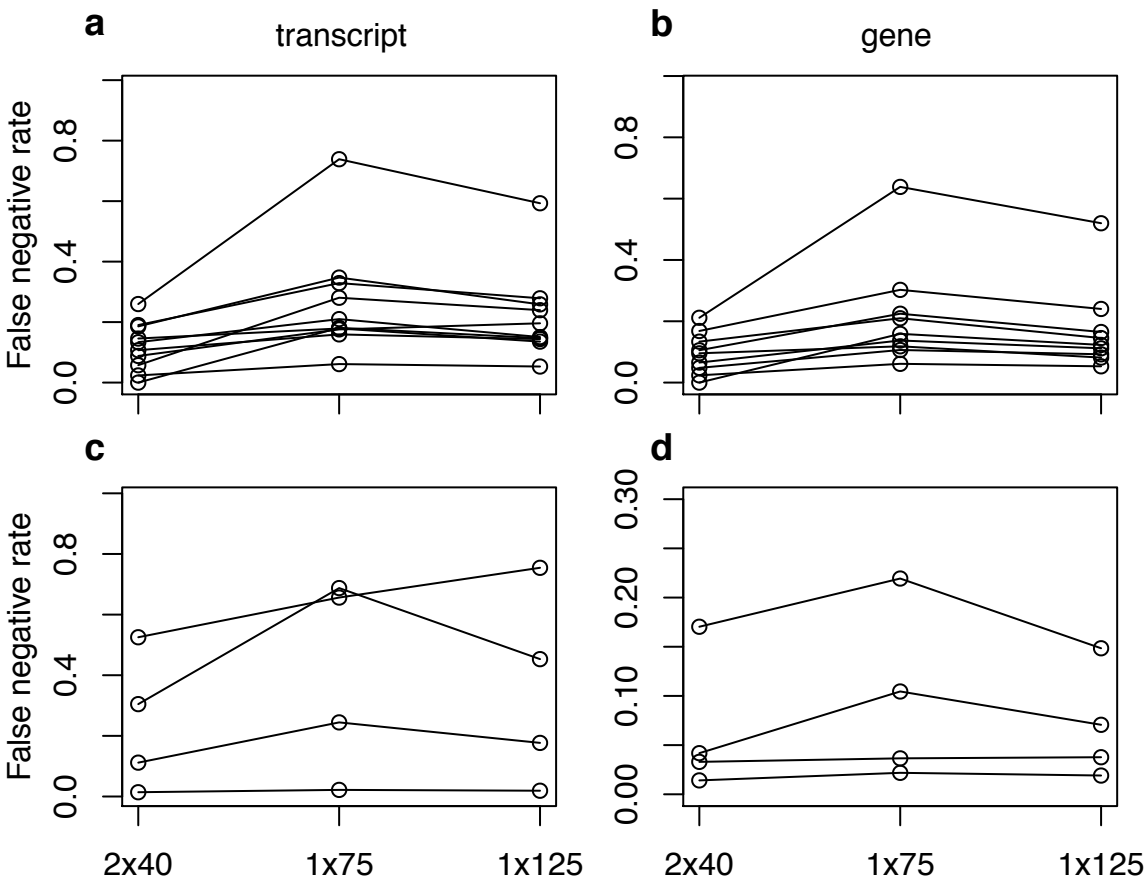

Fig. 4. Violin plots of false negative rate at FDR = 0.01 for differential expression Wald tests, along with median trend "reaction norms" per data set across sequencing strategy for sleuth transcript **(a)** and **(b)** gene, and limma voom **(c)** transcript and **(d)** gene-level inference.

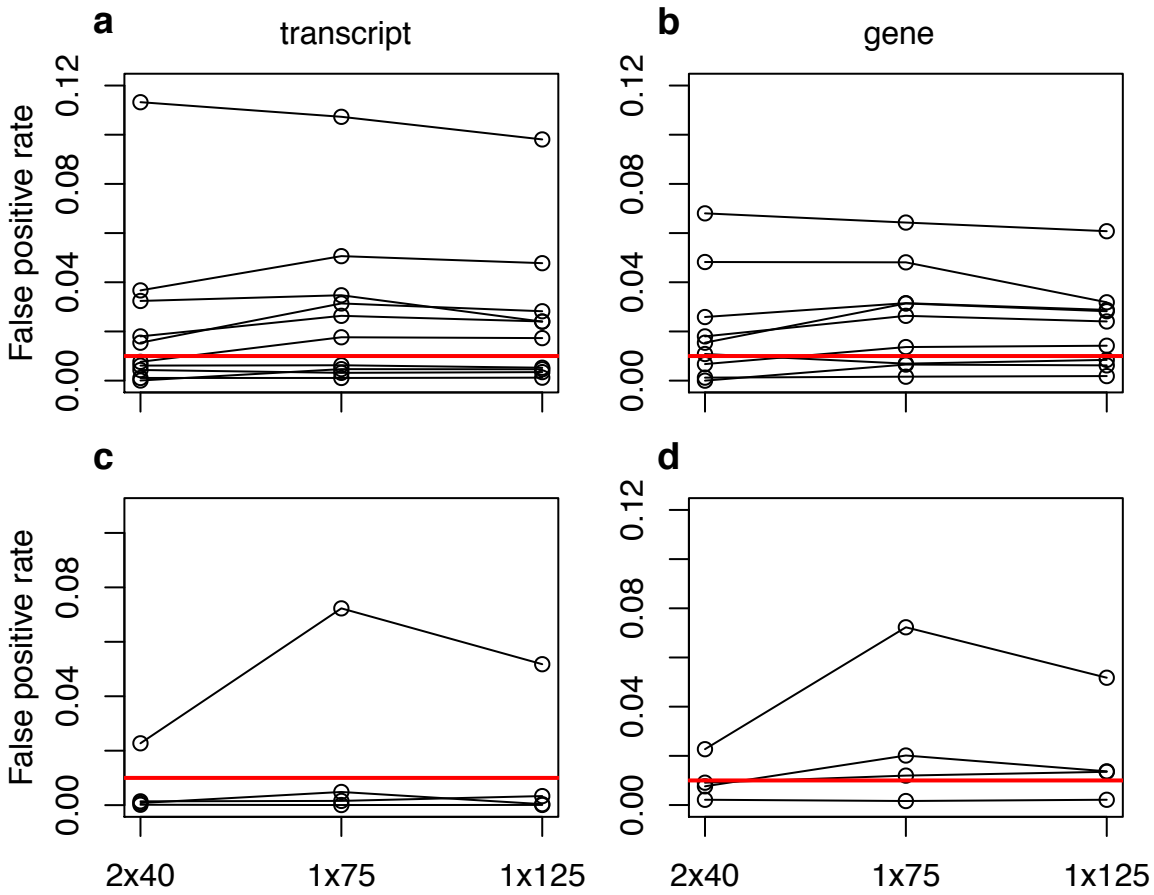

Fig. S1. Violin plots of false positive rate at FDR = 0.01 (red horizonal line) for differential expression Wald tests, along with median trend "reaction norms" per data set across sequencing strategy for sleuth transcript **(a)** and **(b)** gene, and limma voom **(c)** transcript and **(d)** gene-level inference.

Table S1. SRA accessions used in this study.

Accession	Organism	Instrument	Paired-end read length	Conditions	Biological replicates per condition	Kallisto	RSEM
SRP133853	*H. sapiens*	HiSeq 2500	125	6	3	x	x
SRP115815	*H. sapiens*	HiSeq 2500	125-126	9	3	x	
SRP105271	*M. musculus*	HiSeq 2000	125	5	3-4	x	
SRP143508	*M. musculus*	HiSeq X Ten	150	2	3	x	
SRP096374	*M. musculus*	HiSeq 4000	150	2	3	x	x
ERP017328	*D. melanogaster*	HiSeq 2500	126	2	3	x	
SRP128516	*D. melanogaster*	HiSeq 4000	151	2	3	x	x
SRP089981	*C. elegans*	NextSeq 500	151	4	5	x	
SRP092256	*C. elegans*	HiSeq 2500	126	7	5	x	
SRP129557	*C. elegans*	HiSeq 3000	126	4	3	x	x
SRP133093	*S. cerevisiae*	HiSeq 2000	151	4	3	x	x
SRP142501	*S. cerevisiae*	HiSeq X Ten	150	2	3	x	