1   **Compositional Data Analysis is necessary for simulating and analyzing RNA-Seq data**

2   Warren A. McGee[1], Harold Pimentel[2,3], Lior Pachter[4] and Jane Y. Wu[1,*]

3   [1] Northwestern University, Department of Neurology (Chicago, IL, United States); [2] Stanford University,

4   Department of Genetics (Stanford, CA, United States); [3] Howard Hughes Medical Institute (Stanford, CA,

5   United States); [4] California Institute of Technology, Division of Biology and Biological Engineering (Pasadena,

6   CA, United States)

7   Corresponding author: Jane Y. Wu (jane-wu@northwestern.edu)

8   **Keywords**

11 **Abstract**

12 *Seq techniques (e.g. RNA-Seq) generate *compositional* datasets, i.e. the number of fragments sequenced is

13 not proportional to the sample's total RNA content. Thus, datasets carry only relative information, even though

14 absolute RNA copy numbers are of interest. Current normalization methods assume most features do not

15 change, which can lead to misleading conclusions when there are many changes. Furthermore, there are few

16 real datasets and no simulation protocols currently available that can directly benchmark methods when many

17 changes occur.

18

19 We present **absSimSeq**, an R package that simulates compositional data in the form of RNA-Seq reads. We

20 compared **absSimSeq** with several existing tools used for RNA-Seq differential analysis: sleuth, DESeq2,

21 edgeR, limma, sleuth and ALDEx2 (which explicitly takes a compositional approach). We compared the

22 standard normalization of these tools to either "compositional normalization", which uses log-ratios to anchor

23 the data on a set of negative control features, or RUVSeq, another tool that directly uses negative control

24 features.

25

26 Our analysis shows that common normalizations result in reduced performance with current methods when

27 there is a large change in the total RNA per cell. Performance improves when spike-ins are included and used

28 with a compositional approach, even if the spike-ins have substantial variation. In contrast, RUVSeq, which

29 normalizes count data rather than compositional data, has poor performance. Further, we show that previous

30 criticisms of spike-ins did not take into consideration the compositional nature of the data. We demonstrate

31 that **absSimSeq** can generate more representative datasets for testing performance, and that spike-ins should

32 be more frequently used in a compositional manner to minimize misleading conclusions in differential

33 analyses.

34

35 **Author Summary**

36   A critical question in biomedical research is "Is there any change in the RNA transcript abundance when

37   cellular conditions change?" RNA Sequencing (RNA-Seq) is a powerful tool that can help answer this

38   question, but two critical parts of obtaining accurate measurements are (A) understanding the kind of data that

39   RNA-Seq produces, and (B) "normalizing" the data between samples to allow for a fair comparison. Most tools

40   assume that RNA-Seq data is count data, but in reality it is "compositional" data, meaning only

41   percentages/proportions are available, which cannot directly answer the critical question. This leads to

42   distorted results when attempting to simulate or analyze data that has a large global change.

43

44   To address this problem, we designed a new simulation protocol called *absSimSeq* that can more accurately

45   represent RNA-Seq data when there are large changes. We also proposed a "compositional normalization"

46   method that can utilize "negative control" features that are known to not change between conditions to anchor

47   the data. When there are many features changing, this approach improves performance over commonly used

48   normalization methods across multiple tools. This work highlights the importance of having negative controls

49   features available and of treating RNA-Seq data as compositional.

Introduction

High-throughput methods, including RNA-Seq, are frequently used to determine what features—genes, transcripts, protein isoforms—change in abundance between different conditions [1]. Importantly, though, researchers ultimately care about the absolute abundance of RNA transcripts. In other words, is there a change in the number of RNA molecules in a cell when the conditions change? However, current techniques are limited to reporting relative abundances of RNA molecules: the proportion of fragments generated by a sequencer that contain a given sequence [2-5]. This means that RNA-Seq is inherently compositional data, where relative proportions are the only information available, yet those are being used to draw conclusions about the absolute abundance of features [2-5] (see Note 1 in **Supporting Information**). Several studies have raised the alarm on ways in which interpretation of the results can be distorted if RNA-Seq data are not properly treated as compositional [2-6].

The first statistical problem in an RNA-Seq analysis lies in determining the origin of the fragments generated. There are two classes of tools available to solve this problem: (1) tools that use traditional alignments to determine the exact genomic location (**tophat2**, **bwa**, **STAR**, **HISAT2**, etc.) (reviewed in [7]); there are other tools that take these traditional alignments and estimate exon-, transcript-, or gene-level expression levels (reviewed in [8]); (2) tools that probabilistically estimate transcript sets that are compatible with producing the corresponding fragments using pseudoalignment and quantify the levels of transcript expression (**kallisto**, **salmon**, **sailfish**) [9-11].

The second statistical problem, the focus of this paper, is to compare the differences in samples collected under different experimental conditions (e.g. comparing cancer cells with control cells; comparing wild-type cells with mutant cells). We will refer to this second step as "differential analysis." A number of tools are available for differential analyses (**DESeq2**, **edgeR**, **limma-voom**, etc), using continuous or count data (reviewed in [1,12]). One recently developed tool, **sleuth**, utilizes the bootstraps produced by the quasi-mapping tools to estimate the technical variation introduced by the inferential procedure [13].

74   It is a recognized need to normalize and transform the data before conducting differential analyses. Multiple

75   strategies have been developed to meet this need, including quantile normalization [14], the trimmed mean of

76   M-values (**TMM**) method used by **edgeR** [15], the median ratio method used by **DESeq** and **DESeq2** [16], and

77   the **voom** transformation used by **limma** [17]. In addition, multiple units are used when modeling and reporting

78   RNA-Seq results [18], including counts [16,19], CPM [17], FPKM [20], and TPM [21]. Importantly, all of these

79   strategies, even those that are focused just on the counts for each feature, utilize units that are really proportions,

80   which belies the fact that RNA-Seq data are compositional [5,22] (see Note 1 in **Supporting Information**).

81   Furthermore, all of these normalization strategies assume that the total RNA content does not change

82   substantially across the samples (see [23] for a review). This assumption allows users to leap from the inherently

83   relative information contained in the dataset to the RNA copy number changes in the population under study

84   without quantifying the actual RNA copy numbers. However, there are biological contexts where this assumption

85   is not true [4,24,25], and it is unclear how much change can occur before distorting results when the datasets

86   are not considered as compositional during analyses.

87   If information is available about negative controls (e.g. spike-ins, validated reference genes), then such

88   information could be used to anchor the data. This has been done in several studies, where the use of spike-ins

89   led to a radically different interpretation of the data compared to the standard pipeline [4,24,26]. In one study,

90   the RUVg approach was designed to use this reference information to normalize RNA-Seq data, as part of the

91   RUVSeq R package [27]. There have been recommendations to include spike-ins as part of the standard

92   protocol [22,24,28]. However, Risso et al. observed significant variation in the percentage of reads mapping to

93   spike-ins, as well as discordant global behavior between spike-ins and genes [27]. While spike-ins are often

94   used in single-cell RNA-seq applications, they are not routinely used in bulk RNA-Seq experiments.

95   John Aitchison developed an approach to compositional data with the insight that ratios (or log-transformed

96   ratios, called "log-ratios") capture the relative information contained in compositional data [29]. There are three

97   requirements for any analytical approach to compositional data: scale invariance, subcompositional coherence,

98   and permutation invariance (see Note 2 in **Supporting Information**) [30]. It was recently demonstrated that

99   correlation, a widely used measure of association in RNA-Seq analysis, is subcompositionally incoherent and

100   may lead to meaningless results, and an alternative called "proportionality" was proposed [4]. A tool was

101   previously developed to apply compositional data analysis to differential analysis, called ALDEx2 [3]. However,

102   ALDEx2 is not well-suited for utilizing the bootstraps generated by the pseudoalignment tools and is unable to

103   detect any differentially expressed features when there are less than five replicates [31]. Therefore, it is

104   necessary to develop a compositional approach for other tools.

105   Two tools are commonly used to simulate RNA-Seq: polyester and RSEM-sim [21,32]. These tools require the

106   input of estimated counts per transcript and the expected fold changes between groups. However, without

107   considering the data as compositional, protocols used to simulate RNA-Seq data result in the total read counts

108   being confounded with the condition, such that one condition will have on average a greater depth compared to

109   the other condition (for example, see Supplementary Table S2 of [13]). A protocol that simulates many changing

110   features and at the same time yields similar sequencing depth per sample, is lacking, but could be done using

111   principles from compositional data analysis. This challenge, along with the one above, both motivated the

112   present work.

113   Here, we present **absSimSeq**, a protocol to simulate RNA-Seq data using concepts from compositional data

114   analysis. This protocol allows us to directly model large global shifts in RNA content while still maintaining

115   equivalent sequencing depths per sample. Further, we developed a normalization approach that uses negative

116   control features (e.g. spike-ins) with log-ratios, which we call "compositional normalization". We created an

117   extension of sleuth, called **sleuth-ALR**, to use compositional normalization, both to predict candidate reference

118   genes and to normalize the data. We also adapted already available methods to implement compositional

119   normalization for other differential analysis tools. We then used absSimSeq to benchmark performance of

120   differential analysis tools in the setting of either a small or large change to the total RNA content. Within each

121   setting, we compared the current normalization approaches versus either compositional normalization or the

122   RUVg approach with spike-ins.

123   When there was only a small change in total RNA content, all tested tools had similar performance on simulated

124   data, whereas sleuth, sleuth-ALR and limma had the best performance on real data. However, when there was

125   a large change in the RNA content, either up or down, all tools had much better performance if compositional

126   normalization with spike-ins was used. When analyzing a well-characterized real dataset which had a large

127   change in total RNA content per cell [4,26], only compositional normalization with a validated reference gene

128   was able to capture the overall decrease in RNA transcription. Surprisingly, RUVg had poor performance, and

129   the choice of normalization approach had a greater impact on performance than the choice of tool. Furthermore,

130   both of the concerns about spike-ins raised by Risso and colleagues [27] are actually the expected

131   consequences of how compositional data behaves between samples, though they do raise concerns about the

132   proper protocol for including spike-ins.

133   In summary, we provide **absSimSeq** as a resource to generate simulated RNA-Seq datasets that more

134   accurately reflect the behavior of real datasets. This will help future development of RNA-Seq analysis when

135   testing performance. Furthermore, our work suggests that using compositional normalization with spike-ins or

136   validated reference genes is essential for differential analyses of RNA-seq data. When such information is

137   missing, it raises major concerns about the limitations of drawing conclusions from the inherently compositional

138   data of RNA-Seq and other "omics" techniques.

139

140   **Results**

141   *Simulation of RNA transcript copy numbers, normalization, and performance of different tools*

142   To test how changes in the total RNA content can affect performance of differential analysis tools, we developed

143   **absSimSeq** to simulate RNA-Seq data (**Fig 1**). Because the experimental step of generating a library from

144   samples in a real RNA-Seq experiment generates a compositional dataset, this protocol directly simulates that

145   step, producing simulated compositional data. Using this protocol, we carried out three simulation studies, each

146   having five experiments. In each study, current normalization methods were compared to compositional

147   normalization and **RUVg**. A set of highly expressed spike-ins was used as the set of negative control features

148   for compositional normalization and RUVg (see methods for details).

149 **Fig 1. AbsSimSeq, A novel simulation protocol to model compositional RNA-Seq data**. All RNA-Seq

150 experiments convert copy numbers per cell to relative abundances because of the selection step and because

151 the depth of sequencing is arbitrary with respect to the total RNA present (top panel; see Note 1 in **Supporting**

152 **Information**). The **absSimSeq** protocol simulates that conversion process (bottom diagram), with the key, novel

153 steps highlighted in red. It takes the mean empirical relative abundances (in TPM units) from any dataset. It then

154 makes a conceptual leap by assuming those values are copy numbers per cell. Then, the fold changes are

155 simulated on these copy numbers for the experimental condition. After that, in the crucial step of the protocol,

156 the copy numbers are re-normalized back to relative abundances to simulate what happens in the RNA-Seq

157 experiment. From there, the expected reads per transcripts are calculated using relative abundances and the

158 median of the estimated effective lengths from the original dataset. These are then submitted to the polyester R

159 package for a negative binomial simulation.

160 In the first study ("small"), only a small fraction of features were changed (5% of all transcripts), and the total

161 RNA content was similar between experimental groups (<2% change) (**S1 Table** and **S2 Table**). This study was

162 intended to simulate an experiment that fulfills the assumption of the current normalization methods. Under these

163 conditions, all tools tested performed similarly whether using their current normalization approaches or using

164 compositional normalization (**Fig 2A**; **S2 Fig**, panel A).

165 **Fig 2**. **Compositional normalization markedly improves performance when there is a large**

166 **compositional change.** The copy numbers were modeled using the estimated average abundances from the

167 GEUVADIS Finnish women samples (N = 58). Each of the three studies consists of five simulations under

168 specified global conditions: (A) the "small" group has 5% of the transcriptome (~10K transcripts) simulated as

169 differentially expressed, with the total average copy numbers per cell in each group roughly equal, (B) the

170 "down" group has 20% of the transcriptome (~40K transcripts) simulated as differentially expressed, with 90%

171 of the transcripts down-regulated, and (C) the "up" group has the same number of transcripts changing as the

172 "down" group, but 90% of the transcripts are up-regulated. The compositional normalization methods (solid

173 lines) used a set of highly expressed spike-ins to illustrate, especially in the "down" and "up" groups, what

174 happens to the performance of tools when there is a large shift in the global copy numbers per cell. Average

175   false discovery rate across the simulations within each group (n = 5) is shown on the x-axis, and average

176   sensitivity is shown on the y-axis. The FDR range between 0 and 0.25 is shown. Note that edgeR+RUVg is not

177   shown because it always had an FDR above 0.25. See **S2 Fig** for the full range.

178   In the second set of studies ("down" and "up"), many features were differentially expressed (20% of all

179   transcripts), resulting in a large change in the composition, with the total RNA content decreased by ~33% or

180   increased ~2.8-fold, respectively (**S1 Table** and **S2 Table**). Under such conditions, compositional normalization

181   led to greatly improved performance for all tools compared to these tools using their current normalization

182   methods (**Fig 2B-C**; **S2 Fig**, panels B-C). In contrast to current normalization methods and compositional

183   normalization, the **RUVg** approach from **RUVSeq** resulted in the worst performance for edgeR and DESeq2 in

184   all three studies (**Fig 2**), even though it used the same set of spike-ins as compositional normalization.

185   It is worth noting that, among the tools tested, sleuth and ALDEx2 performed the best when there were large

186   compositional changes in the data (**Fig 2B-C**); ALDEx2 uses the IQLR transformation, which is a compositional

187   approach designed to be robust in the presence of large changes to the composition. We also observed that for

188   each transformation used by ALDEx2, it performed almost identically regardless of statistical test used (**S3 Fig**).

189   Finally, sleuth-ALR had similar performance whether TPMs or estimated counts were modeled, or if the Wald

190   test or the likelihood ratio test was used (**S4 Fig**).

191

192   *Performance is not degraded by significant variation in individual spike-ins*

193   A previous study reported that spike-ins had significant variation [27], raising a concern about their utility for

194   normalization. In particular, they observed significant variation both between and within groups. In our simulated

195   data, spike-ins were modeled similarly as other features, with over-dispersed variation, drawing from a negative

196   binomial distribution. When estimating the percentage of transcript fragments mapping to spike-ins per sample,

197   we also detected significant variation across samples (**Fig 3**). Importantly, in the "up" and "down" studies, there

198   were systematic differences between groups, similar to what was observed in the MAQC-III study and in the

199   zebrafish study previously analyzed [27]. These systematic differences between groups are expected given the

200   compositional nature of the data (see Note 3 in **Supporting Information** and Discussion). We further observed

201   a large spectrum of estimated fold changes across individual spike-ins, with a systematic asymmetry in the

202   distribution of fold changes in the experiments from the "up" and "down" studies (**S5 Fig**). Despite the significant

203   variation of spike-ins, individually and collectively, spike-ins led to greatly improved performance when used for

204   normalization (**Fig 2**). Consistent with previous work [33,34], our results suggest that the ratio information

205   contained in spike-ins are collectively robust to variation, and that spike-ins can be used for sample-wise

206   normalization.

207   **Fig 3. Spike-ins have significant within-group and between-group variation, despite improved**

208   **performance when used for normalization.** Previous work expressed a concern about variation observed in

209   spike-ins between samples. In each experiment, the 92 ERCC spike-ins from Mix 1 were simulated to have no

210   change in copy numbers between the two conditions, as well as to have over-dispersed variation between

211   samples, drawing from a negative binomial distribution. Plotted here is the percentage of all fragments that

212   map to spike-ins, compared to the total number of fragments from the sample, in **(A)** the "small" study, with

213   <2% change in the total RNA in each condition; **(B)** the "down" study, with a ~33% decrease in total RNA in

214   the experimental condition; and **(C)** the "up" study, with a ~2.8-fold increase in total RNA in the experimental

215   condition. The dotted line represents the expected percentage of fragments mapping to spike-ins in the control

216   group. Across all experiments, there is significant within-group variation; in the "down" and "up" studies, there

217   is also significant between-group variation. The latter is to be expected given the compositional nature of the

218   data (see Note 3 in **Supporting Information**).

219

220   ***Sleuth-ALR** has best self-consistency and negative control performance among compositional normalization*

221   *methods*

222   To confirm that compositional normalization performs similarly to current methods in the context of real data, we

223   repeated the analyses using data from the original study on sleuth [13]. The first test was the "self-consistency"

224   test using data from [35]. We reasoned that a tool should provide consistent results from an experiment, whether

225   a few samples per group are sequenced (in this case, n = 3 per group), or more samples per group are used (n

226   = 7-8), as measured by the "true positive rate" (TPR) and "false discovery rate" (FDR). In this experiment, true

227   positives were defined as hits identified in both the smaller and larger datasets, and false positives were defined

228   as hits identified by the smaller dataset but not by the larger dataset. Among all tools using compositional

229   normalization, sleuth and sleuth-ALR with Wald test showed the best balance between the TPR and FDR (**Fig**

230   **4**; **S6 Fig**). Limma-voom and sleuth/sleuth-ALR with the likelihood ratio test had the lowest FDR at the cost of a

231   lower TPR. DESeq2 and edgeR both had higher FDR, and on average slightly lower TPR compared to sleuth-

232   ALR. In contrast, the Welch and Wilcoxon statistics in ALDEx2 were unable to identify any hits in any of the

233   "training" datasets, consistent with a recent benchmarking study [31] (data not shown), suggesting that they

234   have greatly reduced power with less data (N = 3 samples per group). ALDEx2's "overlap" statistic was able to

235   identify hits, but this led to the worst consistency (i.e. highest false discovery rate) among the tools tested. Finally,

236   while DESeq2 with RUVg had similar performance to DESeq2 with compositional normalization, edgeR with

237   RUVg had among the worst consistency.

238   **Fig 4. sleuth-ALR Wald has best balance of self-consistency between less and more data from same**

239   **dataset.** Depicted is the Bottomly et al self-consistency test at the isoform level, with **(A)** the false discovery rate

240   at three specified levels, and **(B)** the relative sensitivity as compared to sleuth-ALR with the Wald test. This

241   extends the test from the original sleuth paper [13]. A large dataset is split into a small "training" dataset (3

242   samples per group), and larger "validation" datasets. A "false discovery" in this test is defined as a hit identified

243   in the "training" dataset but not in the larger "validation" dataset at the given FDR level, and a "true positive" in

244   this test is a hit identified in both datasets at that FDR level. A tool performs well in this test if it can identify the

245   same hits with less data, as well as control the "false discovery rate" at the specified FDR level. The full dataset

246   was split twenty times. Note that the number above each tool in panel A is the number of "training" datasets out

247   of twenty that identified at least one hit at the specified FDR level. See **S6 Fig** for the results at the FDR levels

248   of 0.01 and 0.05.

249   Next, we tested the performance of compositional normalization on a negative control dataset, where there are

250   no expected differentially expressed features. We repeated the null resampling experiment from the sleuth paper

251   [13] using the GEUVADIS Finnish women dataset (n = 58) [36]. Six samples were randomly selected (stratified

252   by lab to minimize technical variation) and split into two groups, with the expectation of finding no hits. We found

253   that sleuth-ALR with the likelihood ratio test performed similarly to sleuth and limma-voom (median number of

254   false positives < 5) (**Fig 5**), and sleuth-ALR with the Wald test also showed good false positive control (median

255   number of false positives = 10). In contrast, DESeq2 and edgeR with compositional normalization showed higher

256   numbers of false positives (median of 71 and 66, respectively), and the "overlap" statistic for ALDEx2 showed

257   the highest number of false positives (median of >5000 at the 0.1 FDR level).

258   **Fig 5. sleuth-ALR and limma perform best on the GEUVADIS null dataset.** Depicted is the null experiment

259   at the isoform level **(A)**. This also extends the test from the original sleuth paper [13]. The data were from the

260   lymphoblastoid cells of 58 Finnish women, a relatively homogeneous population, taken from the GEUVADIS

261   project [36]. Data from six women were resampled from the larger dataset, stratifying by lab to minimize technical

262   variation, and then randomly split into two groups to simulate a "null experiment". The number of false positives,

263   defined as any hits, are reported here based on twenty rounds of resamplings. A tool performs well in this

264   experiment by minimizing the number of hits reported. ALDEx2 used the IQLR transformation; all "C.N." methods

265   and sleuth-ALR used compositional normalization; all "RUVg" methods used RUVg for normalization.

266

267   *Performance of compositional normalization on a dataset with a global decrease in transcription*

268   To compare different tools in a real dataset with a large compositional change, we used the "yeast starvation

269   dataset" [26]. In this dataset, yeast cells were starved of a nitrogen source, inducing them to enter a reversible

270   quiescent state without active cell division [37]. Absolute copy numbers per cell were estimated for each mRNA

271   by being normalized to a collection of 49 mRNAs that were quantified using NanoString nCounter [38]. Our re-

272   analysis clearly shows a large global decrease in RNA content (**Fig 6A**), with ~95% of genes decreasing in copy

273   numbers per cell in the starvation group versus control, confirming that the dataset has a large compositional

274   change (**Fig 6B**). On the contrary, analyses using previously developed methods failed to identify this pattern of

275 changes in gene expression, only reporting equivalent numbers of hits up- and down-regulated transcripts

276 (**Table 1**).

277 **Fig 6. A yeast starvation study shows a large global decrease in RNAs. (A)** A violin plot showing the

278 distribution of absolute counts in control yeast cells (pink) and yeast cells starved of their nitrogen source for 24

279 hours (green). The data were from Marguerat et al [26]. The absolute counts were estimated by normalizing

280 RNA-Seq data to a panel of reference genes whose copy numbers were quantified using the NanoString

281 nCounter assay. As can be observed, there is a global decrease in the RNA present. **(B)** a comparison of log2

282 fold changes calculated using a standard RNA-Seq pipeline (the example shown here in gray is kallisto + sleuth),

283 and the log2 fold changes calculated using compositional normalization (sleuth-ALR, shown in purple) or directly

284 from the estimated absolute counts (in blue). As shown, the vast majority of genes were observed to be

285 downregulated when estimating from the absolute counts. The standard RNA-Seq approach misses this global

286 shift, but compositional normalization is able to identify it.

287 Next, we examined the importance of using negative control features with compositional normalization. We first

288 analyzed the data using the gene with the most constant proportion across all samples (as measured by

289 coefficient of variation), rqc1 (Pombase: SPAC1142.01). In this context, compositional normalization missed the

290 global pattern of down-regulation; instead, it reported a similar numbers of hits compared to other tools using

291 current normalization methods, both up-regulated and down-regulated (**Table 1**). We then selected a gene with

292 approximately constant expression (opt3, Pombase: SPCC1840.12) as the denominator for compositional

293 normalization. Compared to current methods, compositional normalization using a validated reference gene was

294 able to identify the global decrease in transcription observed by previous analyses of the data [4,26] (**Table 1**).

295 All tools tested (ALDEx2, DESeq2, edgeR, limma-voom, and sleuth-ALR) were able to identify a similar number

296 of hits when using compositional normalization. In contrast, while RUVg was also given the same validated

297 reference gene, surprisingly it had greatly reduced power and was unable to capture the global pattern of down-

298 regulation.

299

Tool	Up-Regulated Genes	Down-Regulated Genes
ALDEx2 ALR overlap	**719**	**4496**
ALDEx2 IQLR overlap	2716	2677
DESeq2	2424	2415
DESeq2 + C.N.	**632**	**4344**
DESeq2 + RUVg	**698**	**1001**
edgeR	2621	2536
edgeR + C.N.	**582**	**4290**
edgeR + RUVg	**109**	**90**
limma	2603	2527
limma + C.N.	**573**	**4332**
sleuth	2529	2554
sleuth-ALR (trend)	2692	2225
sleuth-ALR	**614**	**4356**
Absolute Counts	**522**	**5751**

300 **Table 1. Only compositional normalization (C.N.) accurately reflects global decrease in the yeast**

301 **starvation study.** This table shows the number of hits identified by each tool using default settings and

302 kallisto-calculated estimated counts and abundances. "Sleuth-ALR trend" used rqc1 (Pombase:

303 SPAC1142.01) as a denominator; this gene had the most consistent abundance (TPM value) across all

304 samples. The compositional normalization methods (all tools in red: "ALDEx2 ALR overlap"; "sleuth-ALR"; all

305 "+C.N." tools) used opt3 (Pombase: SPCC1840.12) as a denominator; this gene was considered a "validated

306 reference gene". "RUVg" for edgeR and DESeq2 (in blue) also used opt3 as a negative control gene. Only

307 compositional normalization methods, using opt3, were able to accurately reflect the severe global decrease

308 observed in the data, as shown by the number of genes showing down-regulation of the absolute counts. Note

309 that ALDEx2 Welch and Wilcoxon statistics yielded <5 significant hits.

310

Discussion

312 Compositional normalization performed similarly to current normalization methods when there was only a small

313 change to total RNA (**Fig 2A; Fig 4; Fig 5**). This similarity is expected because, in this scenario, it is valid to

314 assume that most features are not changing. However, when there was a large change in global RNA,

315   compositional normalization using negative control features (spike-ins) had much better performance compared

316   to the current normalization methods, for both simulated data (**Fig 2B-2C**) and real data (**Table 1**). In this case,

317   the assumption held by the current methods is violated, and this is likely what greatly reduced their performance.

318   Further, although the IQLR transformation in ALDEx2 was designed to be robust to large changes in global RNA,

319   it only modestly improved performance. This indicates that at least some of the features it selected and assumed

320   to be unchanging were indeed changing, in both the simulated data and the real data.

321   The worse performance of current normalization methods is likely related to how fold changes behave in the

322   absolute case versus the relative case. Current normalization methods assume that most features are not

323   changing; one could equivalently assume that the total RNA per cell is unchanged [23]. Thus, if the total RNA

324   content changes, current methods will anchor the data on whatever this global change is. This results in a shift

325   in the observed distribution of fold changes (see Supplementary Figure S16 of [4]). Extreme changes will still be

326   observed to have the same direction, but there will be a group of features that are changing less dramatically

327   than the global change that will be observed to have the wrong sign, and many unchanged features will appear

328   to be changing.

329   Interestingly, the choice of normalization method had a much greater impact on performance than the choice of

330   differential analysis tool, validating a finding from an older study [14]. This was true both for the simulated data

331   (**Fig 2B-2C**) and for the yeast dataset (**Table 1**). In both cases, compositional normalization clearly outperformed

332   current methods when analyzing a dataset with substantial changes to the total RNA content, whether simulated

333   or real. However, this was only true when negative control features were used (**Table 1**). This indicates that the

334   choice of tool is much less important than the choice of normalization and the availability of negative control

335   features (spike-ins, validated reference gene) to properly anchor the data.

336   Surprisingly, RUVg had poor performance in both the simulated data and the experimental yeast dataset. It is

337   unclear why this occurred, other than the likely possibility that it treats the data as count data rather than

338   compositional data. More work would need to be done to see if RUVg could be modified to more accurately

339   capture the global trend in the data from negative control features.

340   In our simulations, the total RNA content was either decreased by 33% or tripled, respectively. The overall

341   change in the "up" study was less extreme than what was observed after c-Myc overexpression [24,39]. In that

342   context, the researchers found a general transcriptional activation that was not captured by the traditional

343   analysis of the RNA-Seq data, and required cell number normalization using spike-ins to see the overall trend

344   of increasing gene expression; the total RNA content increase observed by RNA-Seq was ~5.5-fold (see Table

345   S2 of [24]). The overall change in the "down" group was less extreme than that observed after the Marguerat et

346   al dataset, which observed an 88% decrease total RNA content when using normalized RNA-Seq data. How

347   often large shifts occur in real datasets is unclear because of how infrequently spike-ins or validated reference

348   genes are used when generating data. Future work should determine more carefully how drastic composition

349   changes need to be before performance starts to degrade for methods which assume that most features are not

350   changing.

351   *Results from Bottomly et al. self-consistency test and GEUVADIS null experiment*

352   Sleuth-ALR had the best self-consistency (**Fig 5**), and sleuth-ALR and limma had the best performance in the

353   negative control dataset (**Fig 6**). ALDEx2 was unable to identify any hits using three samples per group with the

354   standard statistical methods (Wilcoxon and Welch), and its "overlap" statistic showed a very high FDR, indicating

355   that its results were inconsistent between the "training" and larger "validation" datasets. This indicates that

356   ALDEx2 may not perform well when there are few replicates per group. While this manuscript was in preparation,

357   a recent benchmarking study came to the same conclusion [31]. This behavior is likely due to the fact that the

358   algorithm of ALDEx2 does not include any shrinkage of the variance. Variance shrinkage has been demonstrated

359   to improve performance when there are few replicates [17,40,41]. Interestingly, though, all three statistics used

360   by ALDEx2 had similar performance on the simulated data (**S3 Fig**), and the "overlap" statistic identified a similar

361   set of hits in the real dataset as other compositional normalization methods (**Table 1**), suggesting that the

362   "overlap" statistic may have utility in small datasets despite poor self-consistency or poor control of false positives

363   in a negative control dataset. Future work could explore how to improve ALDEx2 performance for smaller

364   datasets.

365

The lack of real datasets with verified global changes

367 It was difficult to identify an example of a real dataset, with clear-cut evidence of substantial changes to the total

368 RNA content, that was also amenable to re-analysis using our pipeline. We were unable to re-analyze the

369 previous data measuring the impact of c-Myc overexpression [24] because the RNA-Seq dataset did not have

370 technical or biological replicates. We were also unable to re-analyze the selective growth assay used in the

371 ALDEx2 paper [3] because the raw data, which is necessary for our pipeline, was not publicly available. Other

372 datasets have used spike-ins, but had no other confirmatory data on the absolute copy numbers to confirm if the

373 spike-ins accurately captured the global trend or not. This dearth of bulk RNA-Seq datasets with verified global

374 changes speaks to how much the problem of neglecting to treat bulk RNA-Seq as compositional data has gone

375 unrecognized in the community.

376

How to choose a denominator for compositional normalization and interpret the results

378 When using compositional normalization, regardless of which denominator is chosen, the interpretation of

379 differential expression and fold-changes is "the change of feature X with respect to the denominator". Although

380 all transformations are permutation invariant and therefore any chosen denominator will produce mathematically

381 equivalent results [29], the choice of denominator has important implications for the interpretation of the results

382 and for the downstream validation experiments.

383 If an experimenter has information about absolute copy numbers per cell in their experiment, they can readily

384 use that information with compositional normalization. For example, if spike-ins are included proportional to the

385 number of cells, as recommended in the c-Myc study [24], those spike-ins can be used as the denominator. If

386 one or more reference genes are validated, as was done with the Yeast Starvation study [26], then a reference

387 gene known to be approximately constant under the experimental conditions can be used. In principle, if qPCR

388 is used to validate differential analysis results in this scenario, a predicted reference gene after using spike-ins

389   or the validated reference gene used for compositional normalization would be the best choice for a reference

390   gene.

391   What about experiments that do not have spike-ins or validated reference genes? Spike-ins have only slowly

392   been adopted as a part of *Seq protocols [28]. It has further been extensively documented that reference genes

393   are frequently not properly validated [42], and that expression of commonly used reference genes could change

394   dramatically under certain circumstances [43,44]. There have been several techniques to identify reference

395   genes using RNA-Seq data [45-47]. Importantly, these techniques all find a feature that has an approximately

396   constant proportion throughout all of the samples. However, researchers are usually attempting to identify a

397   reference gene with approximately constant absolute copy numbers per cell throughout. In order to draw this

398   conclusion, the techniques must make the same assumption that standard RNA-Seq analysis tools make, i.e.

399   that the global RNA content remains constant in all samples, or that only a few features are differentially

400   expressed. If many features are changes, features identified by these tools will only reflect the global change

401   (up or down), rather than being approximately constant in absolute copy numbers per cell.

402   None of the compositional normalization methods solve this problem (for an example, see sleuth-ALR with the

403   "trend" feature compared to the other methods in **Table 1**), because no tool can *in principle* solve this problem

404   without access to external information. As described in a recent review article [5], no approach can formally

405   recapitulate the absolute data, and only approaches that are using truly constant features can adequately anchor

406   the data to accurately estimate the true changes in the absolute data. In most datasets without spike-ins or

407   validated reference genes, it is unknown if there is a significant change in the total RNA per cell. Thus, all that

408   is left is how one feature behaves relative to another feature.

409   When one feature is used, there is a clear advantage to compositional normalization versus current methods

410   because there is a clear interpretation of the results (i.e. how features are behaving relative to this feature), and

411   because there would be a clear choice of reference gene for any qPCR validation downstream (for example,

412   spp1 would be used in the yeast starvation study). Any other choice for qPCR reference gene would likely yield

413   discordant results. Importantly, though, identifying a feature with approximately constant proportion, in the

414   absence of information about the overall changes in RNA content, can still help experimenters identify important

415   biology. This is analogous to the approach taken by Gene Set Enrichment Analysis (GSEA) [48]. Its "competitive"

416   null hypothesis leads to the identification of gene sets or pathways that are behaving differently with respect to

417   the general trend of expression changes across the whole genome [49]. GSEA's approach has led to uncovering

418   interesting biology in the past, as demonstrated by how frequently it has been used and cited.

419   In the context of RNA-Seq differential analysis, most datasets will be restricted to this option, and thus

420   experiments will be forced to sacrifice knowledge about the absolute copy numbers for an interpretation of the

421   data anchored to whatever the global change is. This should alarm researchers conducting these experiments

422   to recognize the limitations of the current methodologies. This should also push the community to call for

423   technical innovation and standardization that will make more widespread both the use of spike-ins for

424   normalization, and the validation of reference genes specific to the experiment at hand. Furthermore, this issue

425   regarding the compositional nature of the data is not limited to RNA-Seq, but to many if not all high-throughput

426   ("omics") techniques (see Note 4 in **Supporting Information**).

427

428   _Concerns about the utility of spike-ins_

429   The authors of RUVg [27] made two observations that raised concerns about the utility of spike-ins. However,

430   when interpreting the spike-in abundances through the lens of compositional data analysis, the observed

431   behavior of spike-ins is precisely what would be expected (See Note 3 in **Supporting Information**). In particular,

432   systematic variation between conditions in the spike-in abundances is expected if there is a global change (**Fig**

433   **3**; **S5 Fig**; **S4 Table**). This was observed in the yeast dataset, with the validated reference gene have a greatly

434   increased abundance in the nitrogen starved cells versus control cells.

435   However, their other observation raises valid concerns about the current protocol for using spike-ins. They

436   observed a global discrepancy between spike-ins and the rest of the genes when comparing two control libraries

437   to each other (see Figure 4d of [27]). This could be partially explained by dropout effects, but is most likely due

438   to differences in non-poly-adenylated RNA expression (especially rRNA) between the samples. The way that

439   the spike-ins were added in their experiment (adding an equal amount to approximately equal aliquots of the

440   total RNA) causes the spike-ins to also be subject to compositional changes (**S5 Table**). For bulk RNA-Seq

441   experiments, the standard protocol adds spike-ins to equal amounts of RNA after isolation and selection (poly-

442   A selection or rRNA-depletion), but if there are changes in the excluded RNAs, this protocol impedes the ability

443   of the spike-ins to accurately capture the true fold changes of the RNAs under consideration. In contrast, the

444   approach advocated by Lovén et al [24] was to add spike-ins before RNA isolation, in proportion to the number

445   of cells. With this approach, the spike-ins can, in principle, accurately capture the behavior of the genes, even

446   when there are non-poly-adenylated RNA changes (**S6 Table**). There are challenges with using spike-ins in

447   complex tissues [23], and there may be technical biases that affect spike-ins differently from endogenous RNAs,

448   but further work must clarify this. What is certain is that future work with spike-ins absolutely must keep in mind

449   the compositional nature of the data being generated, and protocols for bulk RNA-Seq may need to be revised

450   to improve the chance of spike-ins accurately anchoring the data to copy numbers.

451

452   *Conclusions*

453   In summary, simulating RNA-Seq data using a compositional approach more closely aligns with the kind of data

454   being generated in RNA-Seq. Compositional normalization using negative control features yields a significant

455   improvement over previous methods, in that it performs best in experimental contexts where the composition

456   changes substantially. Importantly, this method can still be safely used in contexts where the compositional

457   changes are unknown. There is much potential to extend the principles of compositional data analysis to other

458   "omics" approaches, since they all generate compositional data; one intriguing possibility is a normalization free

459   method that examines "differential proportionality" [50]. However, our results from simulation and from real

460   datasets demonstrate that, without access to spike-ins or to validated reference features, a researcher is limited

461   in what conclusions can be drawn from *Seq data because of the compositional nature of the data. This work

462   also makes a strong case for there to be more effort to improve and standardize the use of spike-in controls and

463   validated reference features in all "omics" experiments.

464

Materials and methods

466   *absSimSeq approach to simulating RNA-Seq data*

467   See **Fig 1** for a summary diagram of our protocol for **absSimSeq**. When generating RNA-Seq data, the key

468   experimental step which requires a compositional approach is when the actual changes in the RNA content are

469   sampled using an equal but arbitrary amount of RNA by the library preparation process, resulting in a dataset of

470   proportions. To simulate this shift from count data to compositional data, the **absSimSeq** protocol starts with a

471   set of transcripts and their TPMs, either defined by the user or estimated from real data. It then conceptually

472   shifts from considering transcripts per million (a proportion) to considering copy numbers per cell, i.e. the number

473   of transcripts present in each cell (the absolute count unit of interest). It then simulates the fold changes expected

474   to occur between groups directly on the copy numbers, which may or may not result in a substantial change in

475   the total RNA per cell. The next key step is then converting these new expected copy numbers back to TPMs to

476   represent the expected proportion of each transcript that would be present in an equal aliquot taken from each

477   group. These new TPMs are then converted to expected counts per transcript based on their lengths and the

478   desired library sequencing depth, and those expected counts, along with user-defined or estimated parameters

479   for variance within each group, are then submitted to the R package **polyester** to simulate an RNA-Seq

480   experiment.

481   **AbsSimSeq** also has the option to add spike-ins to the simulated experiment. In our studies, the ERCC ExFold

482   Spike-in mixes are used to define which sequences are included and in what proportions. The user can define

483   what percentage of the transcripts should be coming from spike-ins and which mix to use.

484   *Simulation of copy numbers for this study*

485   To model a simulated dataset after real data, we took an approach modified from Patro et al [10] and Pimentel

486   et al [13]. To estimate the mean and variance for the control group for our simulation, we wished to use a

487   population without expected biological changes within the group. We thus used as a proxy the largest

488   homogeneous population in the GEUVADIS data set, a set of 58 lymphoblastoid cell lines taken from Finnish

489   women. We estimated transcript abundances using kallisto and human Gencode v. 25 transcripts (Ensembl v.

490   87), and then estimated negative binomial parameters (the Cox-Reid dispersion parameter) using DESeq2. We

491   next took the mean TPMs from this dataset, for input into **absSimSeq**.

492   Three simulation studies were performed, with five simulation experiments in each study. The "small" study was

493   intended to simulate experiments where there was no substantial change in the copy numbers per cell per group.

494   The "down" and "up" studies were designed to simulate experiments where there was a large compositional

495   shift, with the total copy numbers either decreasing or increasing.

496   To simulate differential expression, we first applied a filter where the transcript had to have a TPM value of at

497   least 1. We then randomly and independently assigned each filtered transcript as either not changing (i.e. fold-

498   change of 1) or differentially expressed, using a Bernoulli trial with varying probability of success (5% of all

499   transcripts for the "small" study; 20% for the "down" and "up" studies). For each differentially expressed

500   transcript, a truncated normal distribution was used to simulate the fold change, with a mean of 2-fold, a standard

501   deviation of 2, and a floor of 1.5-fold. A Bernoulli trial was then used to choose either up-regulation or down-

502   regulation with varying probability of success (70% down for the "small" study, chosen to produce roughly equal

503   total RNA in each group; 90% down for the "down" study; 90% up for the "up" study).

504   The estimated null distribution and the simulated fold changes thus defined the mean copy numbers per cell for

505   the control group and the experimental group, respectively. These copy numbers were then converted back to

506   TPM. Because TPM is proportional to the estimated counts divided by effective length [13], the TPMs were

507   multiplied by the effective lengths and then normalized by the sum to get the expected proportion of reads per

508   transcript per condition. This was then multiplied by a library size of 30 million reads to get the expected reads

509   per transcript per condition. This matrix of expected reads and the Cox-Reid dispersion parameters estimated

510   from the GEUVADIS dataset were used as input for the **polyester** package [32] to simulate 5 samples in each

511   group, with a random variation of about 2-3% introduced into the exact sequencing depth used. The dispersion

512   parameters for the spike-ins was set to the median dispersion of all transcript that had a mean TPM within 5%

513   of the TPM for the spike-in.

514   **S1 Table** summarizes the simulation parameters and the number of transcripts that are differentially expressed,

515   and **S2 Table** shows the average global copy numbers per cell per condition for each of the fifteen runs. Note

516   that the experimental group in the "up" study had a ~2.8-fold increase, on average, in the total RNA copy

517   numbers per cell. This is less than the 5.5-fold increase in total mRNA observed after over-expressing the

518   oncogene c-Myc (See the normalized data in Table S2 in Lovén et al., 2012)). The experimental group in the

519   "down" study had a ~33% decrease in the total RNA copy numbers per cell. This is less than the decrease in

520   total RNA observed in the yeast dataset (See Supplementary Table S2 from [26]).

521

522   *Implementing a compositional approach for differential analysis tools: the Log-ratio transformation*

523   To allow tools to use negative control features, like spike-ins or validated reference genes, in a compositional

524   manner, we present a method that uses what is called the "additive log-ratio" (ALR) transformation. This was

525   proposed by John Aitchison to address problems analyzing compositional data. He demonstrated that any

526   meaningful function of compositions must use ratios [29]. Further, he proposed the use of log-ratios to avoid the

527   statistical difficulty arising from using raw ratios. ALR is the simplest of the transformations proposed by Aitchison

528   and others in the field of Compositional Data Analysis. In ALR, if there are D components in a composition, then

529   the D-th component is used as the denominator for all of the other D-1 components.

530   Formally, if T is a set of D transcripts, then $x = \{x_t\}_{t \subset T}$ defines the relative abundance of the t-th transcript in

531   the composition, with $\sum_{t=1}^{D} x_t = C$, where C is some arbitrary constant (e.g. 1 million for TPMs). These relative

532   abundances are proportional to the units commonly used in RNA-Seq (RPKM, TPM, etc) [18]. The ALR

533   transformation takes a component to be used as the denominator, analogous to the "reference gene" used in

534   qPCR experiments (see "How to interpret the results" below). This forms a new set of $D - 1$ transformed log-

535   ratios,

536
$$\log_2\frac{x_1}{x_D},\log_2\frac{x_2}{x_D},...,\log_2\frac{x_{D-1}}{x_D}$$

537

538 that can then be used for downstream statistical analyses. If one wishes to use a collection of features (e.g. a

539 panel of validated reference genes; a pool of spike-ins), then the geometric mean, $g(x)$, of those multiple

540 features can be used on all D components:

541
$$\log_2\frac{x_1}{g(x)},\log_2\frac{x_2}{g(x)},...,\log_2\frac{x_D}{g(x)}$$

542 If this is a subset of features, this is called the "multi-additive log-ratio" (malr) [31]. If the geometric mean of all

543 of the features are used, this is the "centered log-ratio" (CLR) transformation introduced by Aitchison and is used

544 in the default mode of ALDEx2 [3]. These are all options available for use in sleuth-ALR.

545 Log-ratio transformations are undefined if either the numerator or denominator is zero. **Sleuth-ALR** has

546 implemented an imputation procedure for handling zeros that minimizes any distortions on the dependencies

547 within the composition. See Note 5 of **Supporting Information** for more details.

548

549 *How to choose a denominator for compositional normalization and how to interpret the results*

550 The proposed interpretation of the results generated by **sleuth-ALR** is simple: whatever the denominator is, the

551 results show how all the other features change relative to that feature or features. For example, if GAPDH is

552 selected as the reference feature, the results show how every other gene is changing relative to GAPDH.

553 Thus, if there are one or more features which are known *a priori* to be negative controls—a validated reference

554 gene, a pool of spike-ins—these are natural choices for use as a denominator in **sleuth-ALR**, either using a

555 single feature, or using the geometric mean of multiple features. Since the copy numbers of these features are

556 expected to be constant between samples, there is now an anchor for the relative proportions between samples.

557 If negative controls are unavailable, we propose identifying one or more features that have the most consistent

558 proportion across all samples. For **sleuth-ALR**, we chose the coefficient of variation as the metric to measure

559 consistency in proportion. Without external information, though, it is unknown if this "consistent feature" is indeed

560 also consistent in copy numbers. If there is a global change in RNA content, this feature would represent the

561 average change. All current normalization methods assume there is no such change, and so make corrections

562 to the data to remove any perceived change; they are therefore mathematically equivalent to this proposed

563 approach (see [5] for a full mathematical proof). However, this approach has the advantage of making explicit

564 the implicit and necessary interpretation (how are features changing relative to the selected reference

565 feature(s)?). It also provides a feature or set of features that can be used as a reference gene for follow-up

566 validation.

567

568 *How sleuth-ALR fits into the current sleuth pipeline*

569 See **S1 Fig** for the pipeline and how it compares to the current pipeline. In the current sleuth pipeline, estimated

570 counts of transcripts from kallisto or salmon are first normalized by the DESeq2 median-of-ratios method [16],

571 and then transformed on the natural log scale with a 0.5-fragment offset to prevent taking the logarithm of zero

572 and to reduce the variability of low-abundance transcripts [13,17]. With the additive log-ratio transformation, the

573 size factor is replaced by the estimated expression of the chosen denominator, and the offset is replaced by the

574 imputation procedure. Once a denominator is chosen, zero values are imputed, the ratios between each feature

575 and the denominator is calculated, and the data is then transformed on a log scale, which can then be used

576 directly in the sleuth model. Our implementation simply replaces the current normalization and transformation

577 functions with the function provided by the **sleuth-ALR** package. For ease of interpretation, the modeling can

578 be done on TPMs directly (using the *which_var* argument for *sleuth_fit*), though the previous choice of modeling

579 the estimated counts can also be used.

580

581 *Compositional approach for the other tools*

582   **ALDEx2** has an explicitly compositional approach, and it solves the imputation problem by simulating bootstraps

583   on the data using the Dirichlet-multinomial distribution, which will never yield zero values. It then calculates

584   estimated statistics by examining the differences between groups within each bootstrap. Its default option is to

585   use the CLR transformation, but it has several options for other choices of denominator. A recent paper

586   examined **ALDEx2**'s performance using these different options [31]; and its results suggested that the IQLR

587   ("interquartile range log-ratio transformation") provided the best balance of performance across real and

588   simulated datasets, with respect to accuracy and computer time. This transformation uses the subset of all

589   features that, after using the CLR transformation, have a sample-wise variance in the interquartile range.

590   Theoretically, this transformation is robust to many features changing either up or down. This and the CLR

591   transformation were used as the normalization methods tested in our study. It can also take a predefined subset

592   of features and uses the geometric mean of those features within each sample; this was the approach taken

593   when utilizing spike-ins for compositional normalization.

594   **DESeq2** uses the median-of-ratios method [16]. If one wishes to calculate a single size factor to normalize each

595   sample, it is calculated by the *estimateSizeFactors* function. This function has a *controlGenes* option, which

596   allows the user to define a set of features that are expected to have constant expression across all samples. A

597   recent review demonstrated that the DESeq2 size factor is mathematically equivalent to the compositional log-

598   ratio proposed by Aitchison [5]. If a dataset has known negative control features (e.g. spike-ins), these can be

599   used to calculate a DESeq2 size factor similar to what is calculated by **sleuth-ALR** or **ALDEx2**. For **DESeq2**,

600   **edgeR**, and **limma**, we calculated DESeq2 size factors using the *estimateSizeFactors* function with designated

601   negative control features (spike-ins for the simulated data; a validated reference gene for the yeast starvation

602   dataset).

603

604   *Pipeline to analyze simulations*

605   The simulated data (FASTA files) were analyzed by **kallisto** for downstream use by all of the tools tested. Spike-

606   ins from ERCC Spike-in Mix 1 were included for the simulations (2% of the total RNA), and so were used as the

607   set of features known to have constant expression between samples. Previous studies observed that only highly

608   expressed spike-ins had consistent ratios across samples [33,34]. Thus, we selected spike-ins that had an

609   average log2 concentration of at least 3 between both mixes. This filter results in a set of 47 spike-ins that were

610   used for compositional normalization and for **RUVg** from **RUVSeq**. **RUVg** was used with **DESeq2** and **edgeR**

611   to test its ability to use spike-in information using its own approach.

612   Filtering is an important issue for managing the accuracy of estimation. Different pipelines make different

613   decisions about what features to filter. To allow the tools to be compared fairly, the same set of filtered transcripts

614   were tested in all tools, defined by those transcripts that passed the standard sleuth filter of having at least 5

615   estimated counts in at least half of the samples. DESeq2's default functionality to use independent filtering and

616   Cooks' outlier filtering did not significantly impact its performance on the simulated data (data not shown), so

617   these were left on.

618

619   *Experiments from the original sleuth paper*

620   To see if compositional normalization would produce similar results with fewer replicates, we repeated the self-

621   consistency experiment as described in the sleuth paper [13]. Briefly, we used the Bottomly et al dataset [35],

622   randomly split the 21 samples into a small training dataset consisting of 3 samples in each condition, and a large

623   validation dataset consisting of remaining samples. The "truth" set of features was defined by the hits identified

624   in the larger validation dataset. This was repeated 20 times. At each of three FDR levels (0.01, 0.05, 0.1), we

625   compared the smaller dataset against the larger dataset, and plotted the estimated FDR and sensitivity relative

626   to sleuth-ALR. Since spike-ins were not used in this experiment, and it is unknown if there was any significant

627   change in the total RNA between the groups, the denominator for compositional normalization was chosen

628   based on which feature had the lowest coefficient of variation across the whole dataset. Zfp106-201

629   (ENSMUST00000055241.12 in Ensembl v. 87) was used as the denominator for **sleuth-ALR** in all datasets.

630   This was also used for RUVg. The IQLR transformation was used for ALDEx2, and otherwise the current

631   normalization methods from the original sleuth paper were used.

632   To test the performance of compositional normalization when analyzing a negative control dataset, we also

633   repeated the null resampling experiment as described in the sleuth paper [13]. Briefly, we used the Finnish

634   samples from the GEUVADIS dataset, and randomly subsampled the data into twenty null experiments with 3

635   samples in two groups. This subsampling was stratified by lab to minimize technical variability that may have

636   occurred between labs. Because of the homogeneous population and minimized technical variation, the

637   expectation is that there would be zero differentially expressed features. The null experiments were analyzed,

638   and the number of false positives was plotted at the transcript-level and gene-level. The same denominator was

639   used for compositional normalization in **sleuth-ALR** across all twenty of the null experiments: SRSF4-201

640   (ENST00000373795.6) for the transcript-level and SRSF4 (ENSG00000116350) for the gene-level. This

641   transcript and gene were determined to have the respective lowest coefficient of variation across all of the

642   samples used for this experiment.

643

644   *Pipeline to analyze yeast dataset*

645   To test the different tools and normalization approaches on a real dataset, we chose a well-characterized "yeast

646   starvation" dataset [26]. In this dataset, yeast were cultured in two conditions: (1) freely proliferating using

647   Edinburgh Minimal Medium; (2) the same medium without a nitrogen source (NH_4Cl), resulting in the cells

648   reversibly arresting into a quiescent state. Two samples from each condition were processed for poly-A selected

649   RNA-Seq or for total RNA (no selection or depletion step). A collection of 49 mRNAs were selected for absolute

650   quantification using the Nanostring nCounter, which uses a fluorescent tagging protocol to digitally count mRNA

651   molecules without the need for RNA purification. The results were normalized to external RNA controls to

652   estimate copy numbers per cell of each mRNA. We used the absolute counts summarized in Supplementary

653   Table S2 of [26] as a basis for selecting opt3 (Pombase: SPCC1840.12.1) as the gene with the smallest

654   coefficient of variation for estimated absolute counts among all samples. This can be considered a validated

655   reference gene. Thus, it was used with methods utilizing negative control features (**sleuth-ALR**, **RUVg**, and

656   other tools using **DESeq2**'s estimateSizeFactors with controlGenes argument) to normalize the data. **Sleuth-**

657   **ALR** was also tested using rqc1 (Pombase: SPAC1142.01.1), which was selected as having the smallest

658   coefficient of variation for raw abundances (TPM values) across all the samples. This gene represents the

659   "average global trend" or "average global change" in the data, as discussed in "How to choose a denominator"

660   section above.

661   To re-analyze the RNA-Seq data, we downloaded the *Schizosaccharomyces pombe* genome cDNA FASTA file

662   from ftp.ensemblgenomes.org (Fungi release 37). This was used as the reference for generating the kallisto

663   index. Each tool was then run using default settings.

664

665   **Declarations**

666   *Availability of data and code*

667   The yeast starvation dataset was taken from Marguerat et al [26] from ArrayExpress at accession E-MTAB-

668   1154, and the absolute counts were taken from Supplementary Table S2 from [26]. The GEUVADIS Finnish

669   data can be found at ArrayExpress using accession E-GEUV-1, using the samples with the population code

670   "FIN" and sex "female". The Bottomly et al data [35] can be found on the Sequence Read Archive (SRA) using

671   the accession SRP004777. Human annotations were taken from Gencode v. 25 and Ensembl v. 87, mouse

672   annotations were taken from Gencode v. M12 and Ensembl v. 87, and yeast annotations were taken from

673   Ensembl Genomes Fungi release 37. The code and vignette for **absSimSeq** can be found on GitHub at

674   www.github.com/warrenmcg/absSimSeq, the code and vignette for using sleuth-ALR can be found at

675   www.github.com/warrenmcg/sleuth-ALR, and the full code to reproduce the analyses in this paper can be found

676   at www.github.com/warrenmcg/sleuthALR_paper_analysis. Here are the versions of each of the software used:

677   **kallisto** v. 0.44.0, **limma** v. 3.34.9, **edgeR** v. 3.20.9, **RUVSeq** 1.12.0, and **DESeq2** 1.18.1; the version of

678   **polyester** used is a forked branch that modified version 1.14.1 with significant speed improvements (found here:

679   www.github.com/warrenmcg/polyester); the version of **sleuth** used is a forked branch that modified version

680   0.29.0 with speed improvements and modifications to allow for **sleuth-ALR** (found here:

681   www.github.com/warrenmcg/sleuth/tree/speedy_fit); the version of **ALDEx2** used is a forked branch that

682   modified version 1.10.0 to make some speed improvements and to fix a bug that prevented getting effects if the

683   ALR transformation with one feature was used (found here: www.github.com/warrenmcg/ALDEx2). All R code

684   was run using R version 3.4.4, and the full pipeline was run using snakemake.

685

689

690   *Author's Contributions*

691   WAM conceived the idea, designed the approach, and wrote the software for sleuth-ALR and absSimSeq. WAM

692   and HP wrote the code for the analysis pipeline. JYW and LP provided supervision. WAM and JYW wrote the

693   manuscript.

694

695   **Acknowledgments**

697

698   **Competing Interests**

699   The authors declare no competing financial interests.

700

701   **References**

702   1.   Huang H-C, Niu Y, Qin L-X. Differential Expression Analysis for RNA-Seq: An Overview of Statistical

703   Methods and Computational Software. Cancer Informatics. SAGE PublicationsSage UK: London,

704   England; 2015;14: 57–67. doi:10.4137/CIN.S21631

705   2.   Lovell D, Müller W, Taylor J, Zwart A, Helliwell C. Proportions, Percentages, PPM: Do the Molecular

706   Biosciences Treat Compositional Data Right? In: Pawlowsky-Glahn V, Buccianti A, editors.

707   Compositional Data Analysis. Chichester, UK: John Wiley & Sons, Ltd; 2011. pp. 191–207.

708   doi:10.1002/9781119976462.ch14

709   3.   Fernandes AD, Reid JN, Macklaim JM, McMurrough TA, Edgell DR, Gloor GB. Unifying the analysis of

710   high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and

711   selective growth experiments by compositional data analysis. Microbiome 2014 2:1. BioMed Central;

712   2014;2: 15. doi:10.1186/2049-2618-2-15

713   4.   Lovell D, Pawlowsky-Glahn V, Egozcue JJ, Marguerat S, Bähler J. Proportionality: A Valid Alternative

714   to Correlation for Relative Data. PLoS Computational biology. Public Library of Science; 2015;11:

715   e1004075. doi:10.1371/journal.pcbi.1004075

716   5.   Quinn TP, Erb I, Richardson MF, Crowley TM. Understanding sequencing data as compositions: an

717   outlook and review. Wren J, editor. Bioinformatics (Oxford, England). 2018;34: 2870–2878.

718   doi:10.1093/bioinformatics/bty175

719   6.   Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome Datasets Are Compositional:

720   And This Is Not Optional. Front Microbiol. Frontiers; 2017;8: 57. doi:10.3389/fmicb.2017.02224

721   7.   Zhao Q-Y, Gratten J, Restuadi R, Li X. Mapping and differential expression analysis from short-read

722   RNA-Seq data in model organisms. Quant Biol. Higher Education Press; 2016;4: 22–35.

723   doi:10.1007/s40484-016-0060-7

724   8.   Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of

725   best practices for RNA-seq data analysis. Genome Biology. 2016;17: 13. doi:10.1186/s13059-016-

726   0881-8

727   9.   Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nature

728   Biotechnology. 2016;34: 525–527. doi:10.1038/nbt.3519

729   10.   Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware

730   quantification of transcript expression. Nature Methods. 2017;14: 417–419. doi:10.1038/nmeth.4197

731   11.   Patro R, Mount SM, Kingsford C. Sailfish enables alignment-free isoform quantification from RNA-seq

732   reads using lightweight algorithms. Nature Biotechnology. Nature Publishing Group; 2014;32: 462–464.

733   doi:10.1038/nbt.2862

734   12.   Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq

735   data. BMC Bioinformatics. 2013;14: 91. doi:10.1186/1471-2105-14-91

736   13.   Pimentel H, Bray NL, Puente S, Melsted P, Pachter L. Differential analysis of RNA-seq incorporating

737   quantification uncertainty. Nature Methods. 2017;14: 687–690. doi:10.1038/nmeth.4324

738   14.   Bullard JH, Purdom E, Hansen KD, Dudoit S. Evaluation of statistical methods for normalization and

739   differential expression in mRNA-Seq experiments. BMC Bioinformatics. BioMed Central; 2010;11: 94.

740   doi:10.1186/1471-2105-11-94

741   15.   Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-

742   seq data. Genome Biology. 2010;11: R25. doi:10.1186/gb-2010-11-3-r25

743   16.   Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biology.

744   BioMed Central; 2010;11: R106. doi:10.1186/gb-2010-11-10-r106

745   17.   Law CW, Chen Y, Shi W, Smyth GK. voom: Precision weights unlock linear model analysis tools for

746   RNA-seq read counts. Genome Biology. 2014;15: R29. doi:10.1186/gb-2014-15-2-r29

747   18.   Pachter L. Models for transcript quantification from RNA-Seq. 2011. arXiv:1104.3889v2

748   19.   Robinson MD, Smyth GK. Moderated statistical tests for assessing differences in tag abundance.

749   Bioinformatics (Oxford, England). Oxford University Press; 2007;23: 2881–2887.

750   doi:10.1093/bioinformatics/btm453

751   20.   Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and

752   quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell

753   differentiation. Nature Biotechnology. Nature Publishing Group; 2010;28: 511–515.

754   doi:10.1038/nbt.1621

755   21.   Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a

756   reference genome. BMC BIoinformatics. BioMed Central; 2011;12: 1. doi:10.1186/1471-2105-12-323

757   22.   Chen K, Hu Z, Xia Z, Zhao D, Li W, Tyler JK. The Overlooked Fact: Fundamental Need for Spike-In

758   Control for Virtually All Genome-Wide Analyses. Molecular and Cellular Biology. American Society for

759   Microbiology; 2015;36: 662–667. doi:10.1128/MCB.00970-14

760   23.   Evans C, Hardin J, Stoebel DM. Selecting between-sample RNA-Seq normalization methods from the

761   perspective of their assumptions. Briefings in Bioinformatics. Oxford University Press; 2018;19: 776–

762   792. doi:10.1093/bib/bbx008

763   24.   Lovén J, Orlando DA, Sigova AA, Lin CY, Rahl PB, Burge CB, et al. Revisiting Global Gene Expression

764   Analysis. Cell. 2012;151: 476–482. doi:10.1016/j.cell.2012.10.012

765   25.   Shippy R, Fulmer-Smentek S, Jensen RV, Jones WD, Wolber PK, Johnson CD, et al. Using RNA

766   sample titrations to assess microarray platform performance and normalization techniques. Nature

767   Biotechnology. 2006;24: 1123–1131. doi:10.1038/nbt1241

768   26.   Marguerat S, Schmidt A, Codlin S, Chen W, Aebersold R, Bähler J. Quantitative analysis of fission

769   yeast transcriptomes and proteomes in proliferating and quiescent cells. Cell. 2012;151: 671–683.

770   doi:10.1016/j.cell.2012.09.019

771   27.   Risso D, Ngai J, Speed TP, Dudoit S. Normalization of RNA-seq data using factor analysis of control

772   genes or samples. Nature Biotechnology. Nature Publishing Group; 2014;32: 896–902.

773   doi:10.1038/nbt.2931

774   28.   Hardwick SA, Deveson IW, Mercer TR. Reference standards for next-generation sequencing. Nature

775   Reviews Genetics. Nature Research; 2017;18: 473–484. doi:10.1038/nrg.2017.44

776   29.   Aitchison J. The single principle of compositional data analysis, continuing fallacies, confusions and

777   misunderstandings and some suggested remedies. Daunis-i-Estadella J, Martín-Fernández JA, editors.

778   Proceedings of CoDAWork'08, The 3rd Compositional Data Analysis Workshop. Universitat de Girona.

779   Departament d'Informàtica i Matemàtica Aplicada; 2008.

780   30.   van den Boogaart KG, Tolosana-Delgado R. Fundamental Concepts of Compositional Data Analysis.

781   Analyzing Compositional Data with R. Berlin, Heidelberg: Springer, Berlin, Heidelberg; 2013. pp. 13–

782   50. doi:10.1007/978-3-642-36809-7_2

783   31.   Quinn TP, Crowley TM, Richardson MF. Benchmarking differential expression analysis tools for RNA-

784   Seq: normalization-based vs. log-ratio transformation-based methods. BMC Bioinformatics. BioMed

785   Central; 2018;19: 274. doi:10.1186/s12859-018-2261-8

786   32.   Frazee AC, Jaffe AE, Ben Langmead, Leek JT. Polyester: simulating RNA-seq datasets with differential

787   transcript expression. Bioinformatics (Oxford, England). 2015;17. doi:10.1093/bioinformatics/btv272

788   33.   SEQC MAQC-III Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and

789   information content by the Sequencing Quality Control Consortium. Nature Biotechnology. Nature

790   Publishing Group; 2014;32: 903–914. doi:10.1038/nbt.2957

791   34.   Munro SA, Lund SP, Pine PS, Binder H, Clevert D-A, Conesa A, et al. Assessing technical

792   performance in differential gene expression experiments with external spike-in RNA control ratio

793   mixtures. Nature Communications. 2014;5: 5125. doi:10.1038/ncomms6125

794   35.   Bottomly D, Walter NAR, Hunter JE, Darakjian P, Kawane S, Buck KJ, et al. Evaluating Gene

795   Expression in C57BL/6J and DBA/2J Mouse Striatum Using RNA-Seq and Microarrays. Zhuang X,

796   editor. PLoS ONE. Public Library of Science; 2011;6: e17820. doi:10.1371/journal.pone.0017820

797   36.   Lappalainen T, Sammeth M, Friedländer MR, 't Hoen PAC, Monlong J, Rivas MA, et al. Transcriptome

798   and genome sequencing uncovers functional variation in humans. Nature. 2013;501: 506–511.

799   doi:10.1038/nature12531

800   37.   Yanagida M. Cellular quiescence: are controlling genes conserved? Trends in Cell Biology. Elsevier

801   Current Trends; 2009;19: 705–715. doi:10.1016/j.tcb.2009.09.006

802   38.   Kulkarni MM. Digital Multiplexed Gene Expression Analysis Using the NanoString nCounter System.

803   Hoboken, NJ, USA: John Wiley & Sons, Inc; 2001. pp. 25B.10.1–25B.10.17.

804   doi:10.1002/0471142727.mb25b10s94

805   39.   Lin CY, Lovén J, Rahl PB, Paranal RM, Burge CB, Bradner JE, et al. Transcriptional Amplification in

806   Tumor Cells with Elevated c-Myc. Cell. Elsevier; 2012;151: 56–67. doi:10.1016/j.cell.2012.08.026

807   40.   Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data

808   with DESeq2. Genome Biology. BioMed Central Ltd; 2014;15: 31. doi:10.1186/s13059-014-0550-8

809   41.   Wu H, Wang C, Wu Z. A new shrinkage estimator for dispersion improves differential expression

810   detection in RNA-seq data. Biostatistics. Oxford University Press; 2013;14: 232–243.

811   doi:10.1093/biostatistics/kxs033

812   42.   Bustin SA, Benes V, Garson JA, Hellemans J, Huggett J, Kubista M, et al. The MIQE Guidelines:

813   Minimum Information for Publication of Quantitative Real-Time PCR Experiments. Clinical Chemistry.

814   Clinical Chemistry; 2009;55: 611–622. doi:10.1373/clinchem.2008.112797

815   43.   Barber RD. GAPDH as a housekeeping gene: analysis of GAPDH mRNA expression in a panel of 72

816   human tissues. Physiological Genomics. 2005;21: 389–395. doi:10.1152/physiolgenomics.00025.2005

817   44.   Rubie C, Kempf K, Hans J, Su T, Tilton B, Georg T, et al. Housekeeping gene variability in normal and

818   cancerous colorectal, pancreatic, esophageal, gastric and hepatic tissues. MCP. 2005;19: 101–109.

819   doi:10.1016/j.mcp.2004.10.001

820   45.   Bin Zhuo, Emerson S, Chang JH, Di Y. Identifying stably expressed genes from multiple RNA-Seq data

821   sets. PeerJ. PeerJ Inc; 2016;4: e2791. doi:10.7717/peerj.2791

822   46.   Van L T Hoang, Tom LN, Quek X-C, Tan J-M, Payne EJ, Lin LL, et al. RNA-seq reveals more

823   consistent reference genes for gene expression studies in human non-melanoma skin cancers. PeerJ.

824   PeerJ Inc; 2017;5: e3631. doi:10.7717/peerj.3631

825   47.   Yim AK-Y, Wong JW-H, Ku Y-S, Qin H, Chan T-F, Lam H-M. Using RNA-Seq Data to Evaluate

826   Reference Genes Suitable for Gene Expression Studies in Soybean. Amancio S, editor. PLoS ONE.

827   Public Library of Science; 2015;10: e0136343. doi:10.1371/journal.pone.0136343

828   48.   Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set

829   enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles.

830   PNAS. 2005;102: 15545–15550. doi:10.1073/pnas.0506580102

831   49.   Maciejewski H. Gene set analysis methods: statistical models and methodological differences.

832   Briefings in Bioinformatics. 2014;15: 504–518. doi:10.1093/bib/bbt002

833   50.   Erb I, Quinn T, Lovell D, Notredame C. Differential Proportionality - A Normalization-Free Approach To

834   Differential Gene Expression. bioRxiv. 2018. doi:10.1101/134536

835   **Supplemental Information Captions**

836   **Supplemental Notes**

837   **Supporting Information.** Contains Supplementary Notes 1-5. Note 1 argues that *Seq datasets are

838   compositional datasets. Note 2 discusses the three requirements of techniques for analyzing compositional

839   data. Note 3 discusses RUVg and the Compositional Behavior of Spike-ins. Note 4 discusses extending the

840   compositional approach to other high-throughput methods. Note 5 discusses how sleuth-ALR handles zeros.

841   **Supplemental Figure Legends**

842   **S1 Fig. The sleuth-ALR approach for compositional normalization.** Under the sleuth model, an

843   observation is modeled as having some error associated with it that is due to the inferential procedure. The

844   true value is modeled as a linear combination of covariates and biological noise. In the original sleuth model

845   (shown in the bottom left), the estimate for the noisy observation was the estimated counts for feature i in

846   sample j, normalized by the DESeq2 size factor. This and other current normalization methods attempt to

847   translate purely relative information to inferences about absolute changes, but only by assuming no change to

848   the total RNA content. The proposed sleuth-ALR estimate (shown on the bottom right) is an example of how to

849   use compositional normalization. It first focuses on abundances (TPMs) rather than estimated counts, and

850   second normalizes the abundances by a "reference feature". This avoids having to assume only a few features

851   change, but at the cost of not translating to inferences about absolute changes unless the chosen feature is a

852   validated reference gene or spike-in.

853   **S2 Fig. A full-range view of the simulation results, accompanying Fig 2.** This shows the full range of FDR

854   and sensitivity for the three simulation studies: **(A)** "small" (5% DE; roughly equal copy numbers in each group);

855   **(B)** "down" (20% DE; ~33% decrease in copy numbers in the experimental group); and **(C)** "up" (20% DE; ~2.8-

856   fold increase in copy numbers in the experimental group). This shows that (1) compositional normalization has

857   similar or superior performance throughout the full range of sensitivities and FDR, and (2) RUVg has poor

858   performance, especially when combined with edgeR.

859   **S3 Fig. ALDEx2 performs similarly in simulations regardless of which statistical method is used**. With

860   the same simulation studies described in **Fig 2**, the performance of ALDEx2 was compared using the Welch t-

861   test (the recommended statistic by the developers), the non-parametric Wilcoxon test, or the reported "overlap"

862   statistic. The overlap statistic is the posterior probability of the effect size being 0 or the opposite direction as

863   what is reported, given the Dirichlet bootstrap samples observed. These three statistics perform approximately

864   similarly no matter which transformation is used: CLR, IQLR, or "denom" (aka ALR, the same as used in sleuth-

865   ALR). This remains true across all three studies: **(A)** "small" (5% DE; roughly equal copy numbers in each group);

866   **(B)** "down" (20% DE; 33% decrease in copy numbers in the experimental group); and **(C)** "up" (20% DE; 2.8-

867   fold increase in copy numbers in the experimental group). This is important because the Welch and Wilcoxon

868   statistics were the ones recommended by the developers but have poor performance when there are few

869   samples.

870   **S4 Fig. sleuth and sleuth-ALR perform similarly regardless of which statistical method or data unit is**

871   **used.** With the same simulation studies described in **Fig 2**, the performance of sleuth and sleuth-ALR was

872   compared when using the Wald test or the likelihood ratio test (LRT), as well as when using TPMs or

873   estimated counts for modeling. All combinations perform similarly within each tool across all three studies: **(A)**

874   "small" (5% DE; roughly equal copy numbers in each group); **(B)** "down" (20% DE; 33% decrease in copy

875   numbers in the experimental group); and **(C)** "up" (20% DE; 2.8-fold increase in copy numbers in the

876   experimental group).

877   **S5 Fig. Spike-ins show a broad range of fold changes and systematic differences in studies with large**

878   **shifts, accompanying Fig 3.** Using the "ground truth" counts from **polyester,** the log2 fold change was

879   calculated for all spike-ins, and then separated by their concentration in the ERCC Mixes, with "high"

880   expression spike-ins having an average log2 concentration of at least 3 attomoles between both mixes (N = 47

881   out of 92). These were the spike-ins used for normalization in **Fig 2**. Shown are boxplots of the spike-in fold

882   changes in each experiment across the three studies: (A) "small" (approximately constant total RNA); (B)

883   "down" (~33 decrease in total RNA); and (C) "up" (~2.8-fold increase in total RNA). Low-expression spike-ins

884   tend to have a broad range of fold changes, and the high-expression spike-ins tend to have a systematic bias

885   in fold changes in the "down" and "up" studies. For reference, the red dotted line in each run indicates the

886   "ideal" fold change for a spike-in, if it precisely matches the reciprocal of the change in copy numbers between

887   the control and experimental conditions; the blue and gold dotted lines indicate the fold change between

888   conditions of the DESeq2 median-of-ratios and the sleuth-ALR geometric mean of high-expression spike-ins,

889   respectively, suggesting that both are generally good approximations of the "ideal" fold change, and thus are

890   good denominators for normalization.

891   **S6 Fig. The False Discovery Rate and Relative sensitivity for the Bottomly self-consistency test at**

892   **additional FDR levels.** This accompanies **Fig 4** in the main text. Shown here are the **(A)** False Discovery

893   Rate, and **(B)** relative sensitivity (% change) at the FDR levels of 0.01 and 0.05.

894   **S7 Fig. Effect of imputation value on bootstrap variation**. This depicts the summary of bootstraps variation

895   for AAGAB-207 (ENST00000561452.5) within each sample of run #6. The true fold change for AAGAB-207 copy

896   numbers is an 82% decrease. The recommended strategy in Compositional Data Analysis for imputing zero

897   values is to choose a value smaller than the smallest observed value; however, because of the extremely small

898   estimated abundances, this results in a very large variation in the bootstraps within each sample **(A)**. This occurs

899   when at least one bootstrap reports an estimated abundance of zero. Our recommendation is to follow the

900   strategy of previous tools, and choose a larger value to impute. Panel (B) shows the reduction in bootstrap

901   variation after choosing 0.01 for the imputation. The wide variation observed in (A) resulted in a non-significant

902   q-value (0.450), whereas the stabilized variation observed in (B) resulted in a significant q-value (0.047).

903   **S8 Fig. Effect of imputation on overall simulation performance**. This depicts the full sensitivity versus false

904   discovery rate curve for different choices of imputation value, as compared to standard sleuth as well as the

905   recommended strategy of choosing a value smaller than the smallest observed value (here depicted as "sleuth-

906   ALR counts" for A-C and "sleuth-ALR TPM" for D-F). **(A)** and **(D)** show the results for the "small" simulation

907   group (5% DE; <2% change in copy numbers per cell); **(B)** and **(E)** show the results for the "down" simulation

908   group (20% DE; 33% decrease in overall copy numbers per cell); **(C)** and **(F)** show the results for the "up"

909   simulation group (20% DE; 2.8-fold increase in overall copy numbers per cell). There is improved performance

910 of using imputation versus no imputation, and there are only minor differences in performance in all three studies

911 among any of the choices for imputation values except 0.1 TPM impute value, which is very high (roughly

912 equivalent to a count imputation of 3), in the "up" study.

913

914 **Supplemental Tables Legends**

915 **S1 Table. Summary of Parameters for Simulation Studies.** For each of the fifteen simulation runs, shown are

916 the parameters to establish the number of differentially expressed (DE) transcripts, as well as the number that

917 are up-regulated versus down-regulated. Only transcripts with a TPM of at least 1 were used to simulated

918 differential expression, but the probability of differential expression was determined by the total number of

919 transcripts (~200K). The proportion of up-regulated transcripts for the "small" study was tuned to result in similar

920 total RNA content in both conditions. The random number generator seed was chosen solely on the basis of

921 yielding consistency in total RNA content across each run within a study. Also shown are the actual number of

922 DE transcripts present in the set of filtered transcripts used by all tools for each run.

923

924 **S2 Table. Total RNA Content Per Cell Per Condition for all Simulation Runs.** For each of the fifteen

925 simulation runs, shown are the total RNA copy numbers per cell for each condition. Also reported are the average

926 change in copy numbers between the two conditions for each study.

927

928 **S3 Table. Doubling the copy numbers per cell results in the same composition.** Depicted is a toy example

929 of a simple cell with five genes of varying abundances. After an experimental manipulation, each gene has

930 exactly double the copy numbers per cell compared to the control condition. This results in the same relative

931 abundances, and therefore the same composition.

932

933 **S4 Table. Spike-in abundances change with large compositional shifts but still accurately capture fold**

934 **changes.** Depicted is another toy example using the same cell with five genes. In this case, there are large

935 changes in the mRNA genes, but no change in the rRNA gene. Spike-ins added in equal amounts both before

936   and after RNA isolation, show changes in their abundances, and therefore would have changes in the

937   percentage of reads mapping to them. Despite this change in their abundance, spike-ins accurately capture

938   the true fold changes.

939

940   **S5 Table. Spike-in abundances change discordantly when non-poly-adenylated RNA changes.** Depicted

941   is another toy example using the same cell with five genes. In this experiment, there is a small change in the

942   rRNA, but no changes to the mRNA genes. Spike-ins were added in equal amounts after RNA isolation, to

943   simulate the protocol used in the zebrafish dataset. Because of the unobserved rRNA change, the spike-ins

944   are affected by the compositional change and show discordant fold changes when compared to the mRNA

945   genes. Normalizing the mRNA genes to the spike-ins results in artefactually elevated fold changes. Thus, the

946   discrepancy observed in the zebrafish dataset can be explained by unobserved changes in the rRNA.

947

948   **S6 Table. Spike-ins must be added before RNA isolation to accurately capture true fold changes.**

949   Depicted is a final toy example using the same cell with five genes. In this experiment, there are large and

950   varying changes to both the rRNA and mRNA genes. Spike-ins were added in equal amounts both before and

951   after RNA isolation. Only the spike-in added before RNA isolation can accurately capture the fold changes of

952   the mRNA genes; the spike-in added after is itself affected by the compositional shift of the simultaneous

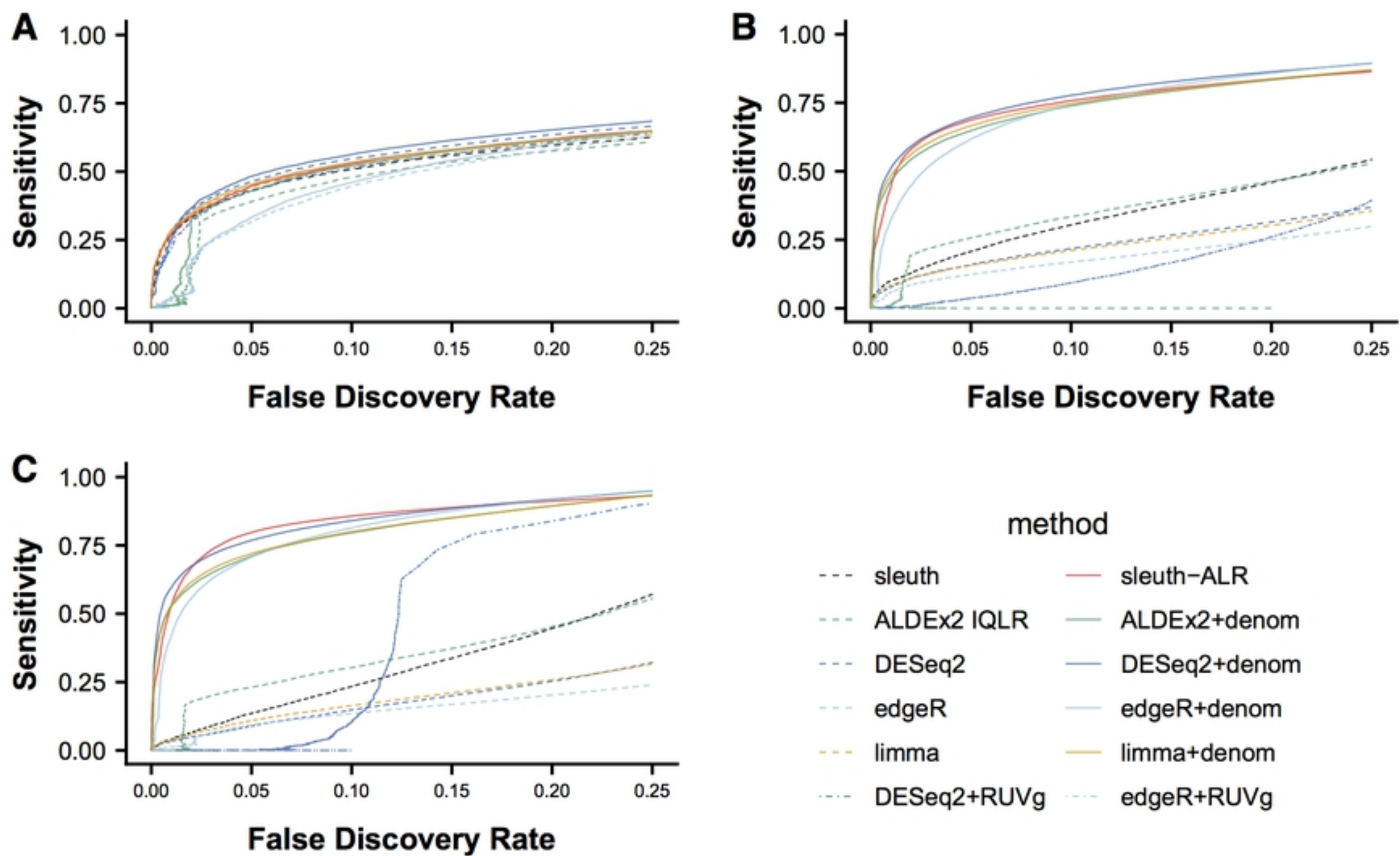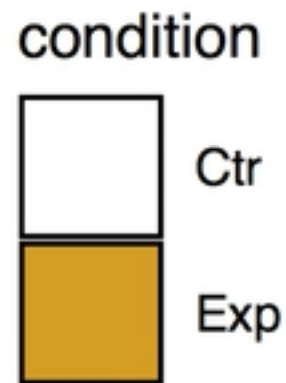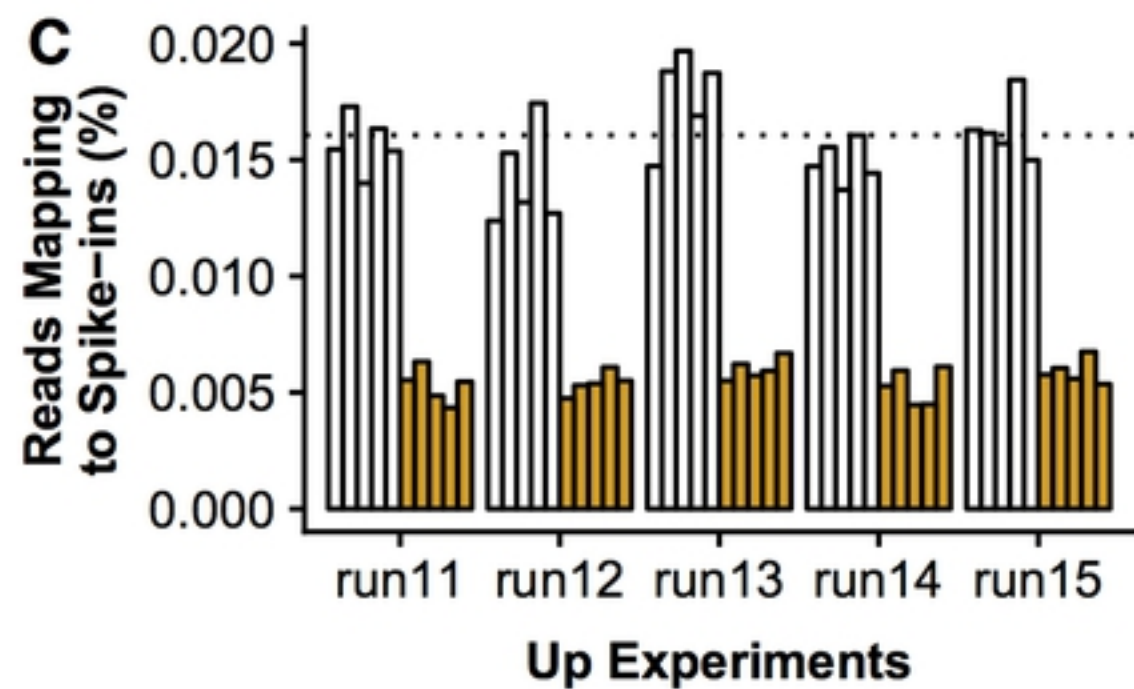953   changes in both the rRNA and mRNA genes.
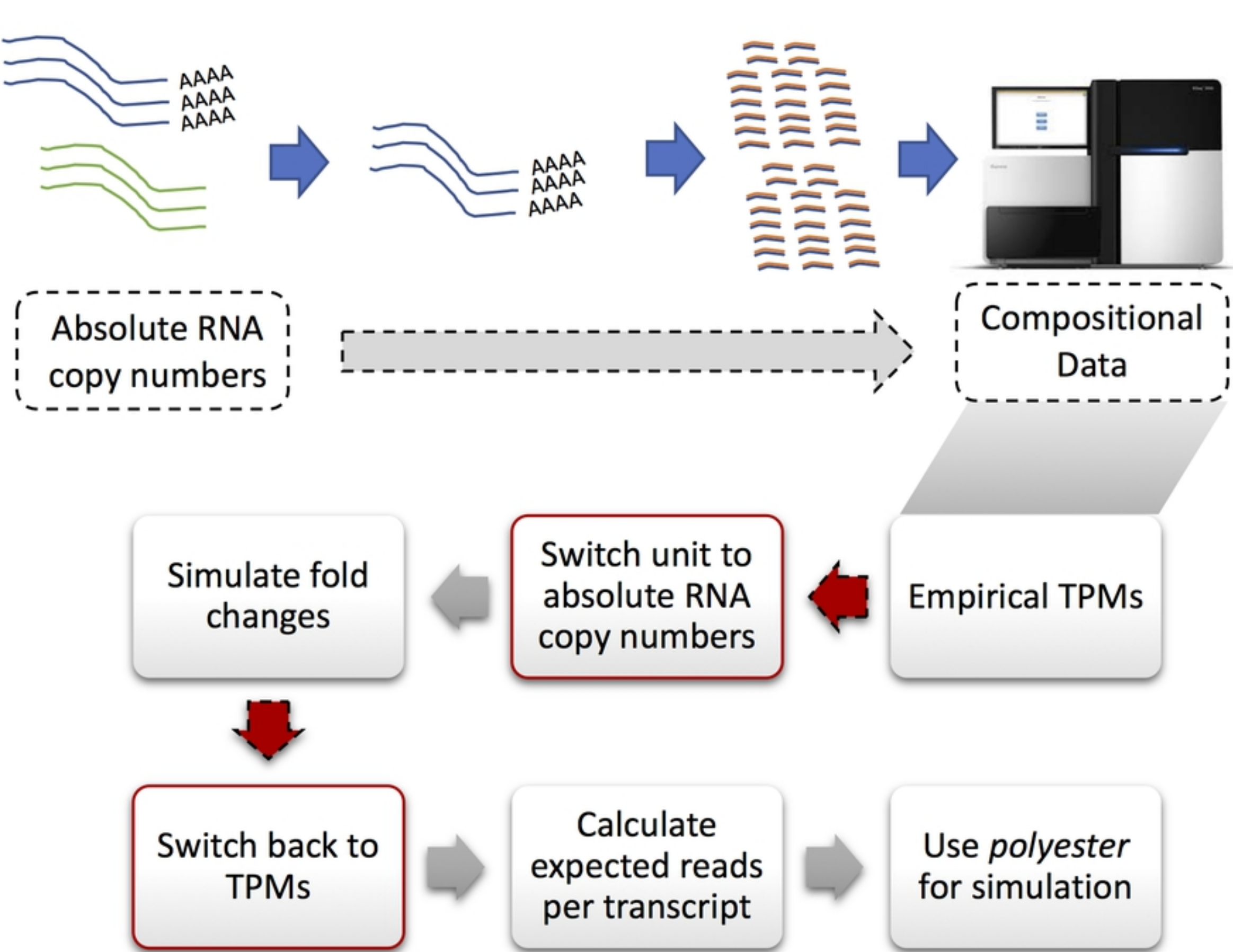
Fig 1

Fig 2

Fig 3

Fig 4

Fig 5

Fig 6