Differential analysis of RNA-seq incorporating quantification uncertainty

Harold Pimentel[1], Nicolas L Bray[2], Suzette Puente[3], Páll Melsted[4] & Lior Pachter[5]

We describe sleuth (http://pachterlab.github.io/sleuth), a method for the differential analysis of gene expression data that utilizes bootstrapping in conjunction with response error linear modeling to decouple biological variance from inferential variance. sleuth is implemented in an interactive shiny app that utilizes kallisto quantifications and bootstraps for fast and accurate analysis of data from RNA-seq experiments.

Many methods have been developed for differential analysis of RNA-seq data[1]. Some of these methods are designed to translate models developed for microarray analysis[2], while others are based on models tailored to RNA-seq[1,3–5]. Differential analysis of RNA-seq experiments requires careful assessment of gene expression variability from a few replicate samples to identify biologically relevant expression differences between conditions[6]. There is ongoing debate on even basic questions such as how to measure gene abundances[1], whether there is sufficient power to test for differences in abundance of individual isoforms[7], and how to best utilize biological replicates[1]. Part of the reason for this uncertainty is the lack of agreed-upon standards for testing and benchmarking RNA-seq methods. In most cases, accuracy claims are based on simulations of read counts from distributions assumed in the models, rather than on simulations of raw reads[2,5,6,8,9]. Such read-count-based simulations typically discount the effects of ambiguously mapping reads and fail to capture both the possibilities for and challenges of isoform-specific differential analysis.

Here we describe a novel approach to RNA-seq differential analysis and a comprehensive benchmarking framework with broader scale and scope than have previously been published. Our approach is implemented along with interactive visualization software that provides crucial transparency in assessing results, and it offers users a convenient tool for exploratory data analysis. Throughout the paper we use the name 'sleuth' to refer to both our statistical method and the software application.

sleuth relies on variance decomposition to identify biological differences in transcript or gene expression (**Fig. 1**). While using a standard strategy of shrinkage to stabilize variance estimates from few samples[2,6], sleuth is able to leverage recent advances in quantification[10] to obtain error estimates that can be used to decouple biological variance from inferential variance before shrinkage (**Fig. 1a**). Variance decomposition is important because of the diversity of variance estimates across genes that arise when quantifying abundances. In one example (**Fig. 1b**,**c**), DESeq2 and voom were run on the same data with featureCounts summaries, and these tools identified a gene as differentially expressed at a false discovery rate (FDR) threshold of 0.10 (reported FDR 8.81×10^{-21} and 5.56×10^{-10}, respectively); whereas sleuth did not find differences between conditions to be significant (reported FDR 0.156) because of the high inferential variance. Some methods have attempted to utilize estimates of quantification errors[6,11,12], but these methods are limited by long run times and lack of robustness to ambiguously mapping reads. By leveraging kallisto's[10] rapid quantification and variance estimation, sleuth overcomes these issues and provides a statistically rigorous, flexible, and efficient tool for RNA-seq analysis.

To test its performance, we compared sleuth with other widely used methods on both simulated and biological data. Our simulation was based on two experimental conditions with three replicates each (see Online Methods). We simulated biological variance (dispersion) according to the negative binomial count model used by DESeq2 (ref. 9). To accurately assess performance, each simulation was repeated 20 times. All programs except sleuth used quantifications inferred from genome alignments (see Online Methods). Full details of the simulation experiments are in **Supplementary Note 1**.

sleuth displayed higher sensitivity than Cuffdiff 2 (ref. 12), DESeq[6], DESeq2 (ref. 9), EBSeq[7], edgeR[8], voom[2]; and sleuth displayed log-fold change[13] in the FDR range of usual interest (0–10%) and beyond, up to FDR 0.3 (**Fig. 2**; **Supplementary Figs. 1** and **2**; and **Supplementary Note 1**). As expected, our simulations found that DESeq2 has more power than DESeq at all relevant FDRs, and that the naïve ranking of genes by log-fold change produces poor results. To control for the effect of different filtering strategies, we ran all programs on a common filtered set of genes and showed that sleuth maintains its sensitivity advantage (**Supplementary Note 1**). Finally, to show the benefit of directly estimating inferential variance, we also demonstrated that sleuth outperforms models in which inferential variance is assumed to be Poisson or zero (**Supplementary Note 1**).

Since FDR control is fundamental for identifying differentially expressed genes in experiments with few replicates, we carefully

[1]Department of Computer Science, University of California, Berkeley, Berkeley, California, USA. [2]Innovative Genomics Institute and Department of Molecular & Cell Biology, University of California, Berkeley, Berkeley, California, USA. [3]Department of Statistics, University of California, Berkeley, Berkeley, California, USA. [4]Faculty of Industrial Engineering, Mechanical Engineering and Computer Science, University of Iceland, Reykjavík, Iceland. [5]Division of Biology and Biological Engineering, Caltech, Pasadena, California, USA. Correspondence should be addressed to L.P. (lpachter@caltech.edu).

RECEIVED 27 JANUARY; ACCEPTED 4 MAY; PUBLISHED ONLINE 5 JUNE 2017; CORRECTED AFTER PRINT 23 AUGUST 2017; DOI:10.1038/NMETH.4324

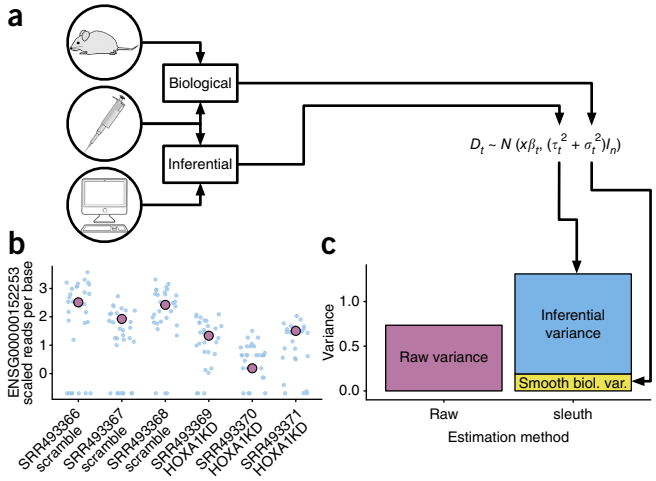

Figure 1 | Overview of sleuth. (**a**) sleuth models different sources of variance to predict differentially expressed transcripts and genes. Biological variance (biol. var.) results from differences in RNA content between replicates and from stochastic biochemistry during library preparation, while inferential variance arises from random sequencing and computational analysis of reads. See Online Methods for description of terms. (**b**) Results for an example gene after running kallisto on RNA-seq data from Trapnell *et al.*[12] generated from human lung fibroblasts transfected with scrambled siRNA (scramble condition) and HOXA1 siRNA (HOXA1KD condition). DESeq2 and voom identify the gene as differentially expressed, but high inferential variance causes sleuth to find no difference. Red dots, point estimates. Blue dots, results for bootstrap samples to assess inferential variance. (**c**) The between-sample raw variance leads to a small estimated biological variance that fails to account for uncertainty introduced when quantifying the samples.

examined the accuracy with which methods self-report their FDR. Other than sleuth and voom, all methods underestimated their FDR; several methods reported an FDR of 0.01, when the true FDR was greater than 0.1. While sleuth overestimates the FDR, this error is conservative; i.e., fewer genes are reported, yet they are highly enriched for being differentially expressed.

To test the accuracy of FDR estimation on biological data, we repeated an experiment from the DESeq2 paper[9] (see Online Methods). Using the Bottomly data set[14], which contains more than ten replicates from two mice strains, we randomly selected two sets of three samples (in sets of 20 replicates) and used differential expression results from the remaining samples as the 'truth' while ensuring that batch types were equally represented across replicates. Each method was tested to see how well it could (i) recapitulate its own results using a smaller data set and (ii) control the FDR, as assessed by comparing to its own results in the remaining high-replicate samples. As in the simulation, sleuth and voom demonstrate a superior ability to estimate their FDR accurately (**Fig. 3a**).

Using our simulated data, we also performed a consistency experiment from the DESeq2 paper[9] to test whether methods produce similar results with less data. Results from simulated data recapitulate the results from biological data, thus validating the reliability of the consistency experiment (**Fig. 3b**).

Additionally, we performed a negative-control experiment comparing two groups of randomly selected female Finnish samples from the GEUVADIS data set[15], for which no biologically meaningful differential expression is expected between groups. sleuth and voom found very few false positives, whereas other methods generated many (sleuth and voom are the only methods with a median of less than 5 false positives at all FDR ranges tested at both the gene and isoform level, whereas the next best method has 95; **Supplementary Figs. 3** and **4**).

While RNA-seq is standard for gene-level differential analysis, there has been debate about its suitability and power at the isoform level. We and others previously demonstrated that isoform-level differential analysis can highlight interesting differential splicing and promoter usage[12], but the significance and reliability of such results have been contested[13].

To examine the power and accuracy of isoform-level differential analysis, we repeated the gene-level analysis at the transcript level (**Fig. 2b**) using kallisto quantifications as the input for each program. We confirm previous findings that the increased testing

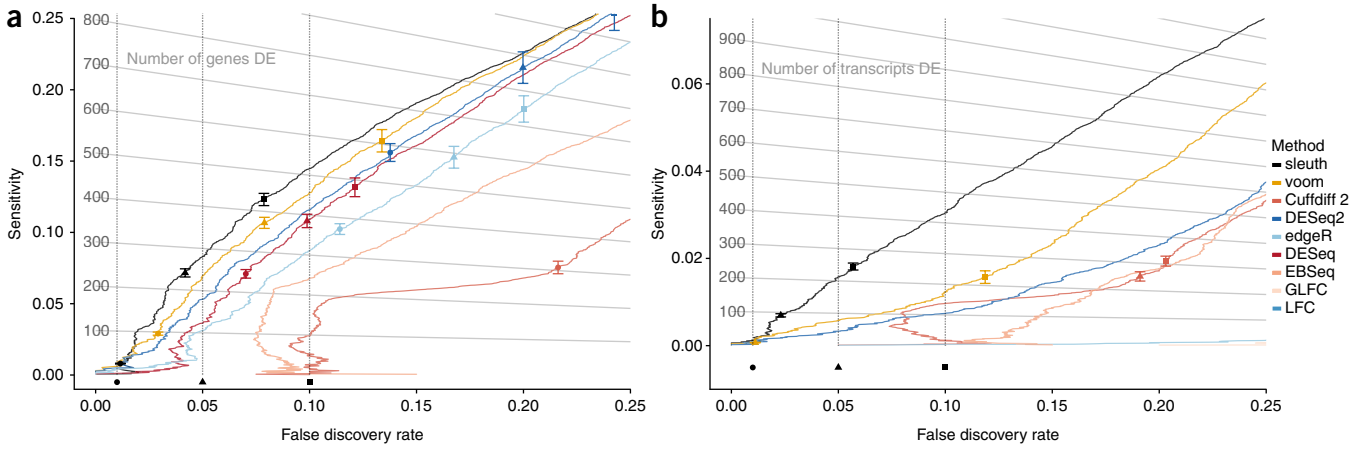

Figure 2 | Sensitivity and false discovery rates of differential expression methods. (**a**,**b**) Sensitivity versus FDR curve for each program on simulated data ('effect from experiment' simulation; see Online Methods), showing the ranking of all genes (**a**) or transcripts (**b**) passing its filter. Circles, triangles and squares represent rankings at an FDR of 0.01, 0.05, and 0.10, respectively. Ideally, each symbol would lie directly above the corresponding symbol on the *x*-axis indicating true FDR. Error bars, 2 s.d. ($n = 20$). Each isoline represents an indicated number of genes (or transcripts) that are called differentially expressed; intersection with a curve indicates a program's performance when looking at that number of top-ranked genes. FDR lines were averaged over 20 replications of the simulation.

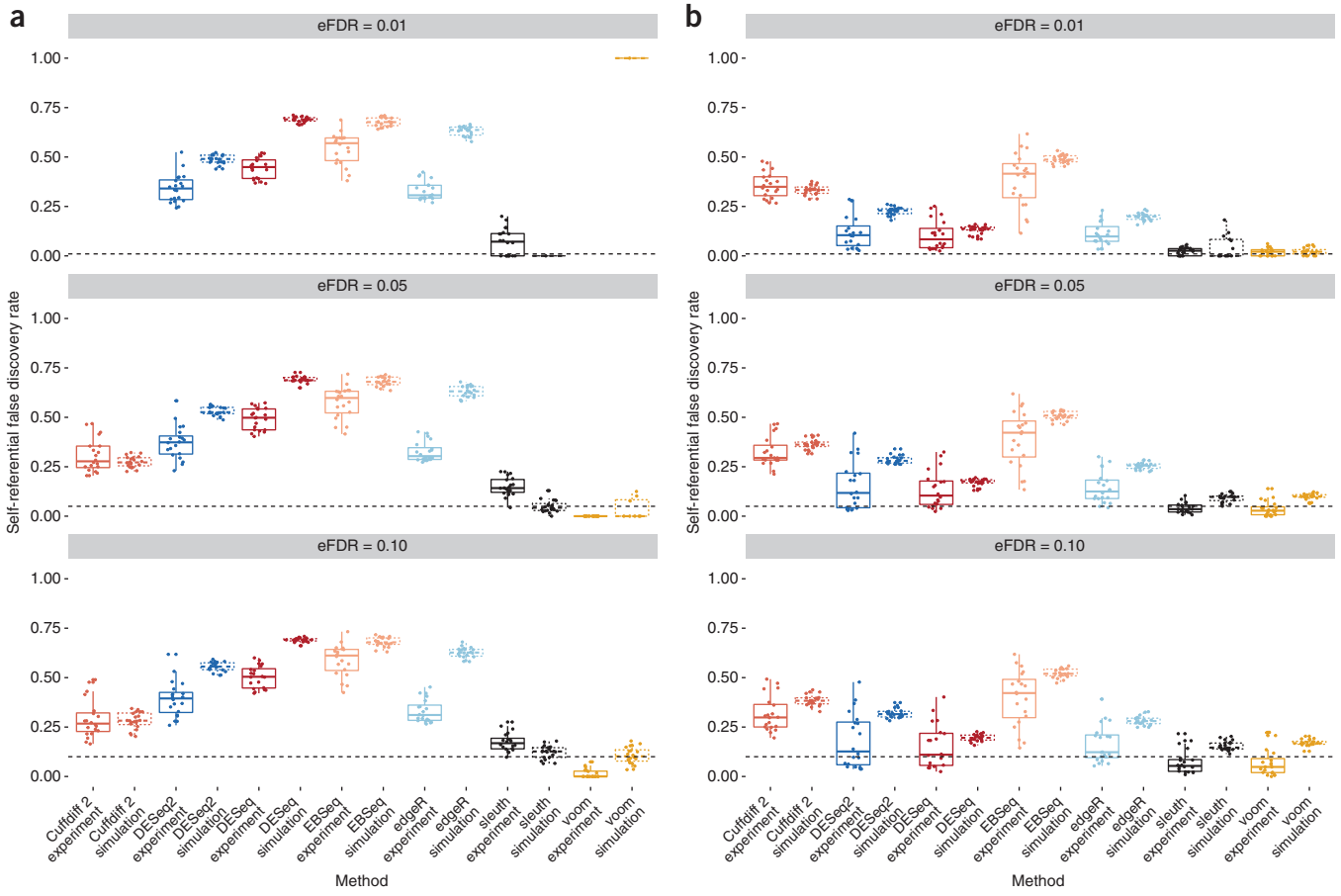

Figure 3 | Self-consistency of differential expression methods when using less data. (**a**,**b**) The estimated versus true FDR for the Bottomly data set[14] and our simulation are shown at the (**a**) isoform and (**b**) gene level. "Experiment" refers to the Bottomly data set; "simulation" refers to our simulations mimicking the Bottomly resampling experiment. The panels from top to bottom display the true FDR for each program when it estimates the FDR as 0.01, 0.05, and 0.10, respectively. Each point represents the FDR of a method on a single experiment. Boxplots indicate median with 25th and 75th percentile hinges, and whiskers extending to the smallest/largest value no less/more than 1.5 × interquartile range (IQR) from the median. Dashed horizontal line represents the target FDR.

required for isoform-level analysis decreases sensitivity in comparison to that of gene-level analysis. However, we find that sleuth can still control the FDR at the isoform level while calling many differentially expressed isoforms. Interestingly, while the power to discover differentially expressed features is lower, our simulations show that, at a given FDR, the total number of differential features detected is fairly similar to that when performing gene-level analysis (see isolines in **Fig. 2**; **Supplementary Note 1**). Moreover, for simulations in which isoform abundances change independently between conditions, we find that sleuth outperforms other methods. The same is also true for the correlated-effect simulation (see Online Methods, **Supplementary Note 1**). In addition, we tested BitSeq[5] on a single sample, as its run time was prohibitive on the entire simulation set. We found BitSeq performed well overall, although sleuth outperformed BitSeq when the true FDR was less than 0.12 (**Supplementary Note 1**). To demonstrate that the improvement of sleuth's performance arises from its model rather than from its use of kallisto's quantifications, we ran sleuth for one replicate of our simulation using RSEM quantifications for the original data along with manually performed bootstraps, and we saw almost identical performance (**Supplementary Figs. 5** and **6**). We also used tximport[16] to test the result of using kallisto quantifications to estimate gene abundances for differential analysis with other programs, and we found that sleuth remained superior to other methods (**Supplementary Figs. 7** and **8**).

Our results show that by accounting for uncertainty in quantifications, sleuth is more accurate than previous approaches at both the gene and isoform levels. Crucially, the estimated FDRs reported by sleuth are well controlled and reflect the true FDRs, making the predictions of sleuth reliable and useful in practice.

The sleuth workflow was designed to be simple, interpretable, and fast. The model was chosen in part for its tractability, and the Shiny visualization framework was chosen for its portability (**Supplementary Note 2**). The modularity of the algorithm also makes it easy to explore improvements and extensions, such as analysis of more general transcript groups and different shrinkage and normalization schemes to improve performance. As a result, when coupled with kallisto, which has dramatically reduced run times for quantification based on the use of pseudoalignment, sleuth is a quick, accurate, and versatile tool for the analysis of RNA-seq data.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

H.P. and L.P. were partially supported by NIH grant nos. R01 DK094699 and R01 HG006129. We thank D. Li, A. Tseng, and P. Sturmfels for help with implementing some of the interactive features in sleuth.

AUTHOR CONTRIBUTIONS

H.P. led the development of the sleuth statistical model and was assisted by S.P., N.L.B., P.M., and L.P. The method comparison and testing framework was designed by H.P., N.L.B., P.M., and L.P. The interactive sleuth live software was designed and implemented by H.P., as was the sleuth R package. H.P. automated production of the results. H.P., N.L.B., P.M., and L.P. analyzed results and wrote the paper.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Conesa, A. *et al. Genome Biol.* **17**, 13 (2016).
2. Law, C.W., Chen, Y., Shi, W. & Smyth, G.K. *Genome Biol.* **15**, R29 (2014).
3. Li, B., Ruotti, V., Stewart, R.M., Thomson, J.A. & Dewey, C.N. *Bioinformatics* **26**, 493–500 (2010).
4. Trapnell, C. *et al. Nat. Biotechnol.* **28**, 511–515 (2010).
5. Glaus, P., Honkela, A. & Rattray, M. *Bioinformatics* **28**, 1721–1728 (2012).
6. Anders, S. & Huber, W. *Genome Biol.* **11**, R106 (2010).
7. Leng, N. *et al. Bioinformatics* **29**, 1035–1043 (2013).
8. Robinson, M.D. & Smyth, G.K. *Bioinformatics* **23**, 2881–2887 (2007).
9. Love, M.I., Huber, W. & Anders, S. *Genome Biol.* **15**, 550 (2014).
10. Bray, N.L., Pimentel, H., Melsted, P. & Pachter, L. *Nat. Biotechnol.* **34**, 525–527 (2016).
11. Turro, E., Astle, W.J. & Tavaré, S. *Bioinformatics* **30**, 180–188 (2014).
12. Trapnell, C. *et al. Nat. Biotechnol.* **31**, 46–53 (2013).
13. Teng, M. *et al. Genome Biol.* **17**, 74 (2016).
14. Bottomly, D. *et al. PLoS One* **6**, e17820 (2011).
15. Lappalainen, T. *et al. Nature* **501**, 506–511 (2013).
16. Soneson, C., Love, M.I. & Robinson, M.D. *F1000Res.* **4**, 1521 (2016).

ONLINE METHODS

Model. We consider an additive response error model in which the total between-sample variability has two additive components, (i) 'biological variance' that arises from differences in expression between samples as well as from variability due to library preparation and (ii) 'inferential variance,' which includes differences arising from computational inference procedures in addition to measurement 'shot noise' arising from random sequencing of fragments. The model is an extension of the general linear model where the total error has two additive components. Given a design matrix x, we assume a general linear model for the (unknown) abundance Y_{ti} of transcript t in sample i in terms of fixed-effect parameters beta and 'noise' epsilon.

$$Y_{ti} \mid x_i = x_i^T \beta_t + \varepsilon_{ti} \quad (1) \tag{1}$$

While the Y_{ti} are not directly observed, (normal) perturbations of Y_{ti} constitute the observed random variables D_{ti}:

$$D_{ti} \mid y_{ti} = y_{ti} + \zeta_{ti} \tag{2}$$

With some further assumptions, one can derive that the D_{ti} values are normally distributed (see **Supplementary Note 2**) and the variance, which is the key to performing differential analysis, can be interpreted as the sum of biological (ε_{ti}) and inferential (ξ_{ti}) variance.

Testing for differential expression. In comparing samples to identify differentially expressed genes or transcripts, sleuth applies the likelihood ratio test—where the full model contains labels for the samples, and the reduced model ignores labels. Underlying the test is an estimate of the variances $V(D_{ti})$, where t ranges over the transcripts and i over the samples. The estimate for $V(D_{ti})$ used in sleuth is

$$\hat{V}(D_{ti}) = \max(\tilde{\sigma}_t^2, \hat{\sigma}_t^2) + \hat{\tau}_t^2 \tag{3}$$

where $\hat{\tau}_t^2$ is the estimate of the inferential variance obtained from bootstrapping, $\hat{\sigma}_t^2$ the raw biological variance, and $\tilde{\sigma}_t^2$ a shrinkage-based estimator of the biological variance. For details of how the individual variance estimates were obtained, see **Supplementary Note 2**.

Simulations. A null distribution for transcript abundances was learned from the largest homogeneous population in the GEUVADIS data set, 59 samples of lymphoblastoid cell lines derived from Finnish females[15], a proxy for a homogeneous set of samples. We estimated transcript-level abundances with kallisto, then we estimated parameters for negative binomial distributions (using the Cox–Reid dispersion estimator) to model count distributions using DESeq2.

After the null distribution was constructed, expression features (isoforms or genes, depending on the type of simulation) were chosen to be differentially expressed. Transcripts with less than five estimated counts on average across the GEUVADIS samples were marked as too rare to be simulated as differentially expressed. A gene was assumed to pass the filter if at least one of its constituent transcripts passed the filter. In each simulation, 20% of the features that passed the filter were chosen to be differentially expressed at random. If the simulation had unequal size factors, random size factors were chosen from the set {1/3, 1, 3} such that the geometric mean equaled 1, similar to the simulation procedure in DESeq2. However, unlike the DESeq2 simulation procedure, our size factors were chosen at random. Counts were generated from the negative binomial distribution, after which reads were simulated using the RSEM simulator[4]. This resulted in about 30 million 75-base-pair paired-end reads per sample for a total of 13.8 billion reads overall (see tables in **Supplementary Note 1** for exact counts). Three types of simulations were performed.

Independent effect simulation. Isoforms across the transcriptome were chosen to be differentially expressed at random. The simulations were generated with equal size factors. Effect sizes were chosen from a truncated normal distribution, such that the minimum absolute fold change for differential transcripts or genes was 1.5.

Correlated effect simulation. Genes (instead of isoforms) were randomly chosen to be differentially expressed. A direction (sign) for each effect size was chosen at random, then all the effects were simulated from a truncated normal with minimum absolute fold change 1.5. The simulation used random unequal size factors generated as described above.

Effect from experiment. To mimic the types of changes seen in real experiments, fold changes were learned from Trapnell *et al.*[12] from the set of transcripts that either DESeq2 or sleuth found to be differentially expressed at FDR 0.05. Genes were chosen at random to be differentially expressed. The null mean counts were used to determine the rank of each transcript relative to its parent gene. These ranks were matched between the Trapnell data set, and the null distribution was learned from the GEUVADIS data set.

Self-consistency experiment. In order to validate whether methods would produce similar results with less data, we performed an experiment similar to those of Love *et al.*[9]. For each iteration, we randomly selected three samples from condition C57BL/6J and three samples from condition DBA/2J and ran each tool. The 'truth' set was established by using the remaining samples to identify differentially expressed genes or transcripts using that program. For each FDR level (0.01, 0.05, 0.10), we compared the results from the smaller data set with those of the larger data set for each tool. The FDR was then computed and plotted in **Figure 3**.

Data processing and software notes. Sleuth may be downloaded at http://pachterlab.github.io/sleuth. For isoform analyses, all quantification was performed using kallisto version 0.42.4. For gene-level analyses, HISAT2 was used to align reads to human genome GRCh38 and mouse genome GRCm38 for all programs other than sleuth (which used kallisto). Quantifications for Cuffdiff 2 were performed using Cufflinks. Remaining quantifications were done using featureCounts. Ensembl release 80 was used for human analyses, and release 84 was used for mouse analyses. The following R programs were used to compile the results: sleuth 0.28.1, BitSeq 1.16.0, DESeq 0.24.0, DESeq2 1.12.0, EBSeq 1.12.0, edgeR 3.14.0, and limma-voom 3.28.2. When testing programs at the isoform level, kallisto 0.42.4 was used to obtain quantifications. Cuffdiff 2.21 was used with alignments from HISAT2 2.0.1 (ref. 17). Subread (featureCounts) 1.5.0 (ref. 18) was used

with alignments from HISAT2 to get raw gene counts. BitSeq was provided alignments from Bowtie 1.1.2 (ref. 19). All analyses in the paper are fully reproducible through the Snakemake system[20], available at https://github.com/pachterlab/sleuth_paper_analysis and in the **Supplementary Software**.

Data availability statement. The Bottomly data set is available at the NCBI Gene Expression Omnibus (GSE26024, accession nos. SRR099223–SRR099243). The Trapnell *et al.* data set (**Fig. 1b**) is available at the NCBI Gene Expression Omnibus (GSE37704, accession nos. SRR493366–SRR493371). The GEUVADIS data set is available at the European Nucleotide Archive (accession no. ERP001942).

17. Kim, D., Langmead, B. & Salzberg, S.L. *Nat. Methods* **12**, 357–360 (2015).
18. Liao, Y., Smyth, G.K. & Shi, W. *Bioinformatics* **30**, 923–930 (2014).
19. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. *Genome Biol.* **10**, R25 (2009).
20. Köster, J. & Rahmann, S. *Bioinformatics* **28**, 2520–2522 (2012).

doi:10.1038/nmeth.4324

Erratum: Differential analysis of RNA-seq incorporating quantification uncertainty

Harold Pimentel, Nicolas L Bray, Suzette Puente, Páll Melsted & Lior Pachter
Nat. Methods **14, 687–690 (2017); published online 5 June 2017; corrected after print 23 August 2017**

In the version of this article initially published, the final term in equation (2) in the Online Methods was incorrectly specified as x_{ti}. The correct term is ζ_{ti}. Also, the two callouts to **Supplementary Note 3** in the Online Methods section were incorrect and should have referred to **Supplementary Note 2**. These errors have been corrected in the HTML and PDF versions of the article.