Flexible isoform-level differential expression analysis with Ballgown

Alyssa C. Frazee[1], Geo Pertea[2,3],Andrew E. Jaffe[1,3,4], Ben Langmead[1,2,3,5], Steven L. Salzberg[1,2,3,5], & Jeffrey T. Leek[1,3*]

March 2014

1. *Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health*
2. *McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine*
3. *Center for Computational Biology, Johns Hopkins University*
4. *Lieber Institute for Brain Development, Johns Hopkins Medical Campus*
5. *Department of Computer Science, Johns Hopkins University*

* *Correspondence to jtleek@gmail.com*

Abstract

We have built a statistical package called *Ballgown* for estimating differential expression of genes, transcripts, or exons from RNA sequencing experiments. *Ballgown* is designed to work with the popular *Cufflinks* transcript assembly software and uses well-motivated statistical methods to provide estimates of changes in expression. It permits statistical analysis at the transcript level for a wide variety of experimental designs, allows adjustment for confounders, and handles studies with continuous covariates. *Ballgown* provides improved statistical significance estimates as compared to the *Cuffdiff2* differential expression tool included with *Cufflinks*. We demonstrate the flexibility of the *Ballgown* package by re-analyzing 667 samples from the GEUVADIS study to identify transcript-level eQTLs and identify non-linear artifacts in transcript data. Our package is freely available from: `https://github.com/alyssafrazee/ballgown`

A key advantage of RNA sequencing (RNA-seq) over hybridization-based technologies such as microarrays is that RNA-seq makes it possible to reconstruct complete gene structures, including multiple splice variants, from raw RNA-seq reads without relying on previously established annotations [17, 29, 8]. The price for this flexibility is a dramatically larger quantity of raw data [20] and much greater computational cost associated with assembly and quantification of transcript expression[25]. The most widely used pipeline for transcript assembly, quantification, and differential expression analysis is the Tuxedo suite, which aligns reads with *Bowtie* and *Tophat2* [10], assembles transcripts with *Cufflinks* [29] and performs differential expression analysis with *Cuffdiff2* [28]. This suite has been used in many influential projects [16, 32, 15], including the ENCODE [5] and modENCODE [9] consortium projects.

We have developed software called *Tablemaker* that takes a GTF file and a set of BAM files and generates a set of linked, tab-delimited text files. These text files contain the structure of assembled transcripts, mappings from exons and splice junctions to transcripts, and expression data measured by FPKM (Fragments Per Kilobase of transcript per Million reads sequenced) and by average per-base coverage (Figure 1a, Supplementary Material: Tablemaker output files). The *Ballgown* package can then be used to read these data into an easy-to-access and analyze *R* object for downstream analysis (Figure 1b), and *Ballgown* includes a flexible linear model framework for differential expression analysis (Supplementary Material: Data and Notation, statistical methods for detecting differential expression). It is also possible to link BAM [14] files to the *Ballgown* object and use them to plot the read-level coverage for transcripts of interest. *Ballgown* can work with any assembly tool that outputs assembled transcripts and expression estimates in the same format as *tablemaker* output (Supplementary Material: Tablemaker output files).

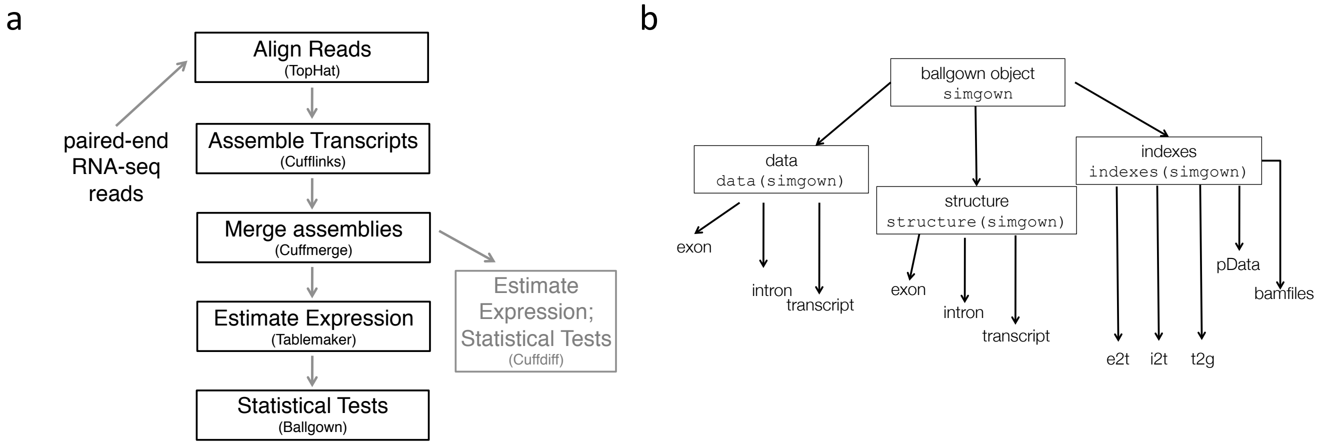

Figure 1: **The Ballgown pipeline. a.** To use *Ballgown* run the standard steps in the *Cufflinks* pipeline until *Cuffmerge*. Then run *Tablemaker* on each sample to calculate per sample FPKMs and average coverage numbers. These tables can then be loaded into `R` using the *Ballgown* package. **b.** The *Ballgown* package loads data into an object with linked tables for expression levels of exons, introns and transcripts. The object also loads information about the exon, intron and transcript structures and corresponding indexes for matching structures to expression and phenotype data to genomic measurements.

Statistical significance comparisons

Cuffdiff2 is designed for two-class differential expression analysis. It has been observed that *Cuffdiff2* produces conservatively biased statistical results when evaluating differential expression between two groups [7, 28]. To confirm this result, we collected *Cuffdiff2* output from InSilico DB [3] for two experiments with sufficient sample sizes for differential expression analysis (Supplementary Section: Data Analyses, InSilico DB Analysis). The first experiment [11] compared lung adenocarcinoma ($n = 12$) and normal control samples ($n = 12$) in nonsmoking female patients. The second experiment [31] compared cells at five developmental stages. We analyzed the data from two stages: embryonic stem cells ($n = 34$) and pre-implantation blastomeres ($n = 78$). We compared only transcripts with average FPKM greater than one across all samples within a study to avoid test results from transcripts with little or no observed expression.

Comparing transcript expression between either tumor and normal samples or between developmental cell types should show strong differential expression signals, given the sample size and distinct phenotypes. In the cancer versus normal comparison, there were 4454 transcripts with an average FPKM greater than one. *Cuffdiff2* identified 1 transcript as differentially expressed at the FDR 5% level, while *Ballgown*'s F-test identified 2178. When comparing developmental phenotypes, there were 12,469 assembled transcripts with average FPKM greater than one, and *Cuffdiff2* identified 0 differentially expressed transcripts versus *Ballgown*'s 7236. These results on large scale studies suggest that *Cuffdiff2*'s statistical significance estimates of differential expression at the isoform level show a strong conservative bias (Figure 2a,2b).

To confirm this result, we created an open-source tool called *polyester* for generating simulated RNA-seq reads from experiments with biological replicates and transcript-level differential expression (Supplementary Material: Simulation studies). We simulated a differential expression experiment with $n = 10$ samples in each of two groups, from $m = 2,745$ annotated transcripts on human chromosome 22 from the Ensembl [6] annotation (GRCh37 build, v74). We set 274 transcripts to be differentially expressed with a fold change of 6 between groups, with an equal number of transcripts differentially expressed in each direction. In the simulated data, *Cuffdiff2* showed the same strong conservative bias, calling 0 transcripts differentially expressed (controlling FDR at the 5% level), compared to 80 using *Ballgown* 's F-test (Supplementary Material: Simulation studies, Model fitting in simulated data). Accordingly, the p-value distributions showed similar patterns to those we observed in the adenocarcinoma and developmental cell datasets (Figure 2c). *Ballgown* also produced a more accurate ranking of transcripts for differential expression than *Cuffdiff2*: 78 of the top 100 transcripts called differentially expressed were truly differentially expressed for *Ballgown* versus 63 for *Cuffdiff2*, a 23% increase in truly differentially expressed genes (Figure 2d). We further investigated the source of the conservative bias of *Cuffdiff2* and found that when we sampled reads with equal probability from each transcript, ignoring transcript length, *Cuffdiff2* produced accurate measures of statistical significance (Supplmentary Figure 1). This result suggests that the conservative bias may be due to transcript length normalization in the *Cuffdiff2* software.

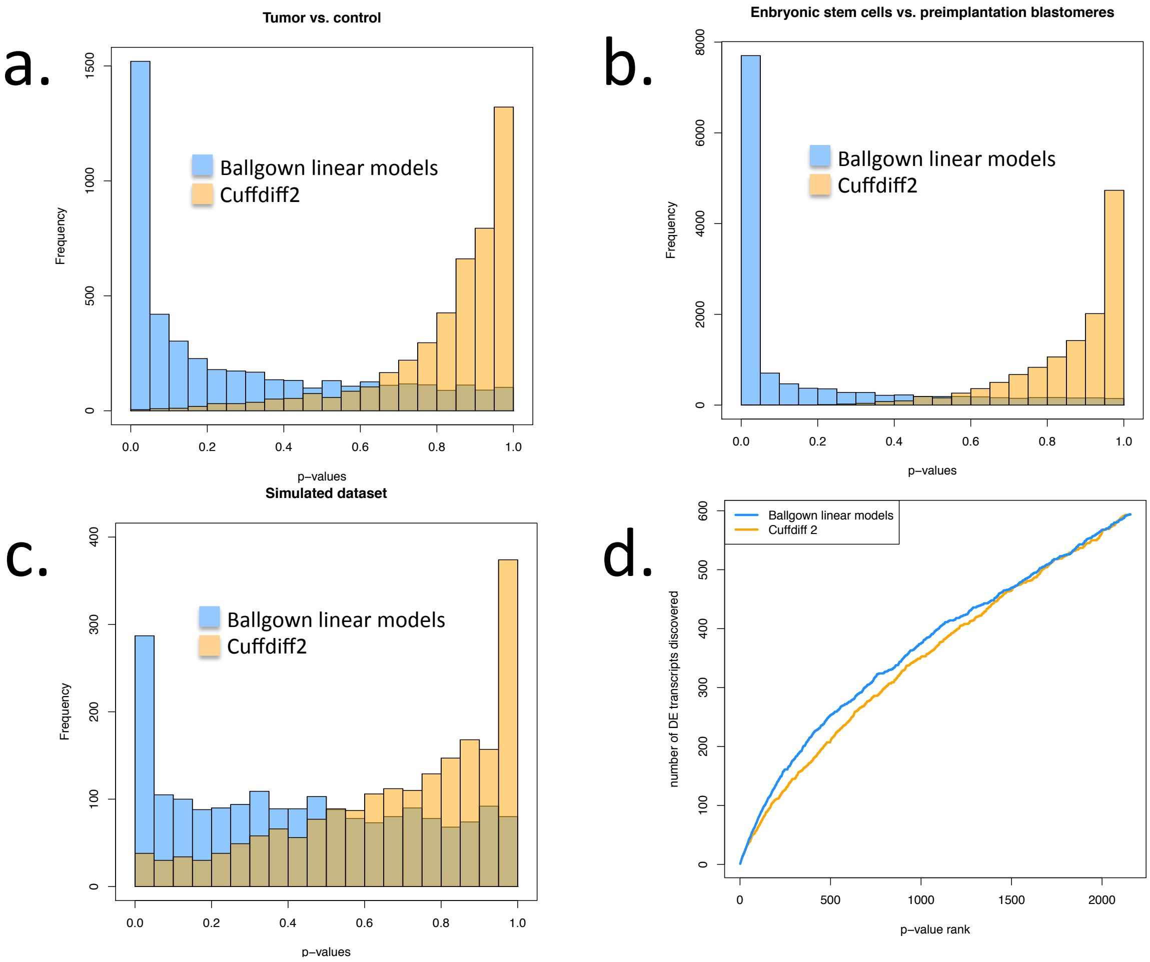

Figure 2: **Comparison of statistical significance for** *Cuffdiff2* **and** *Ballgown* **a.** Histograms of p-values from the comparison of 12 lung adenocarcinomas and 12 normal controls from female patients who never smoked (*Ballgown* in blue, *Cuffdiff2* in orange). **b.** Histograms of p-values from the comparison of 78 pre-implantation blastomere samples and 34 embryonic stem cell samples (*Ballgown* in blue, *Cuffdiff2* in orange). **c.** Histograms of p-values from a simulated data set of 2,745 transcripts in 10 cases and 10 controls. 10% of transcripts were simulated to be differentially expressed at a fold change of 6. We observe the same strong conservative bias as in the two reanalyzed studies. **d.** A plot of the ranking of transcripts from most to least differentially expressed based on p-value (x-axis) versus the number of truly differentially expressed transcripts (y-axis), using the simulated dataset. Among the top 100 transcripts ranked by each method for differential expression, 63 are truly differentially expressed for *Cuffdiff2* and 78 are for *Ballgown*.

1 Flexibility of statistical models

The main advantage of *Ballgown* over *Cuffdiff2* is the added flexibility to compare any nested set of models for differential expression or to apply standard differential expression tools in *Bioconductor*, such as the *limma* package [24] (Supplementary Material: Data and Notation, statistical methods for detecting differential expression). To demonstrate *Ballgown*'s flexiblity, we performed two analyses that are not possible with *Cuffdiff2*: modeling continuous covariates and eQTL.

Analysis of quantitative covariates

In the first analysis, we treated RNA Integrity Number (RIN) [22] as a continuous covariate [26] and used *Ballgown* 's modeling framework to discover transcripts in the GEUVADIS dataset [12] whose expression levels were significantly associated with RIN (Supplementary Material: Data Analysis). Of 43,622 assembled transcripts with average FPKM above 0.1, 19,118 showed a significant effect ($q < 0.05$) of RIN on expression, using a natural cubic spline model for RIN and adjusting for population and library size [18]. The populations included in the study were Utah residents with Northern and Western European ancestry (CEU), Yoruba in Ibadan, Nigeria (YRI), Toscans in Italy (TSI), British in England and Scotland (GBR), and Finnish in Finland (FIN).

A previous analysis of the GEUVADIS data modeled variation in RNA-quality as a linear effect [1]. We fit this model and identified an enrichment of transcripts that showed positive correlation between FPKM values and RNA-quality as expected (Supplementary Figure 2). To investigate the impact of using a more flexible statistical model to detect artifacts, we tested whether a 3rd-order polynomial fit for RIN on transcript expression was significantly better than simply including a linear term for RIN after adjusting for population. We found that the cubic fit was significantly better than the linear fit ($q < 0.05$) for 1,450 transcripts (Figure 3), suggesting that simple linear adjustment for confounding variables such as RNA quality might not be sufficient to capture unwanted sources of variation in transcript data.

Expression quantitative trait locus analysis

To demonstrate the flexibility of using the post-processed *Ballgown* data for differential expression compared to *Cuffdiff2*, we next performed an eQTL analysis of the 464 non-duplicated GEUVADIS samples across all populations (Supplementary Material: Data Analyses, eQTL analysis). We filtered to transcripts with an average FPKM across samples greater than 0.1 and removed SNPs with a minor allele frequency less than 5%, resulting in 7,072,917 SNPs and 44,140 transcripts. We constrained our analysis to search for *cis*-eQTLs where the genotype and transcript pairs were within 1000kb of each other resulting in 218,360,149 SNP-transcript pairs. To adjust for potential confounding factors, we adjusted for the first three principal components of the genotype data [19] and the first three principal components of the observed transcript FPKM data [13]. The analysis was performed in 2 hours and 3 minutes on a standard Desktop computer using the MatrixEQTL package [23].

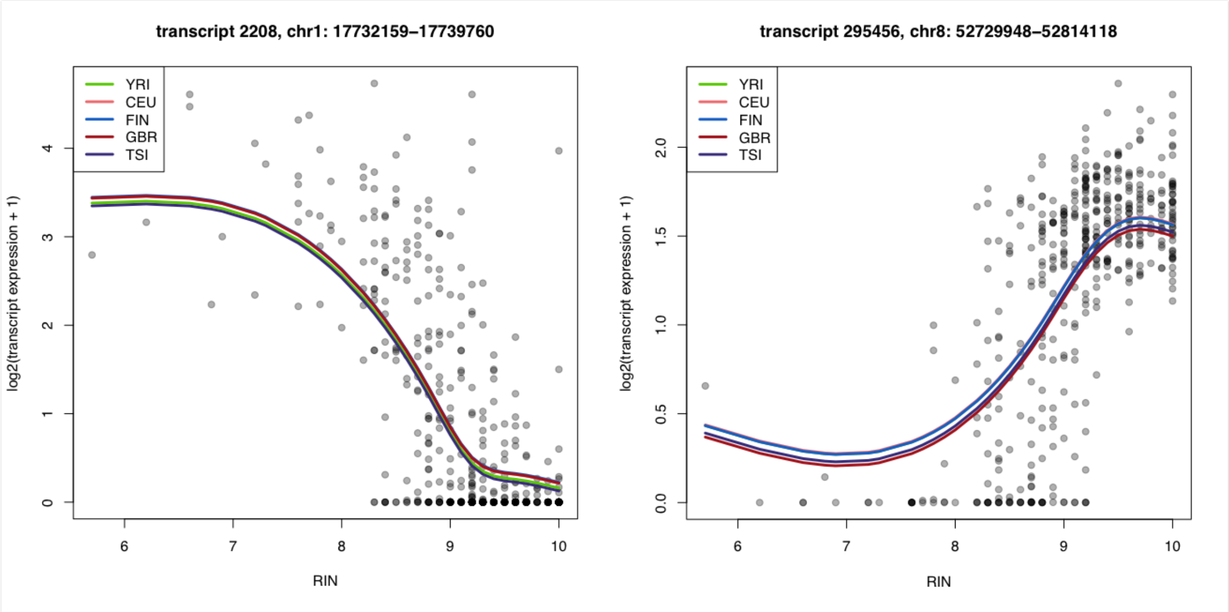

Figure 3: **Non-linear effects of RNA quality on transcript expression** These transcripts (FDR < 0.001) and 1,448 others showed a relationship with RNA quality (RIN) that was significantly better captured by a non-linear trend with three degrees of freedom than a standard linear model. Colored lines shown are predicted values from a natural cubic spline fit and represent predictions for the specified population, assuming average library size.

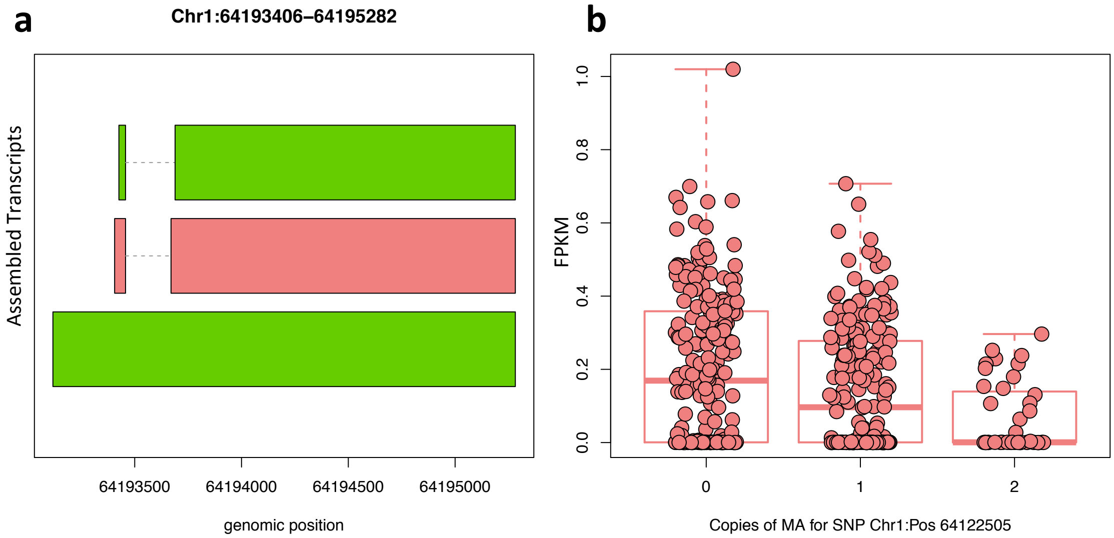

Figure 4: **An assembled transcript that does not overlap previously annotated transcripts but shows a significant eQTL. a.** A diagram of the transcript structures for the three assembled transcripts at this locus. The green transcripts had an average FPKM < 0.1, the red transcript had a significant eQTL (FDR $< 1\%$). **b.** A boxplot of the FPKM values for the middle transcript from panel **a.** showing a consistent and statistically significant eQTL.

Visual inspection of the distribution of statistically significant results and corresponding QQ-plot indicated that our confounder adjustment was sufficient to remove major sources of bias (Supplementary Figure 3). We identified significant eQTL at the FDR 1% level for 17,276 transcripts overlapping 10,624 unique Ensembl-annotated genes. We calculated a global estimate of the number of null hypotheses and estimated that 5.8% of SNP-transcript pairs showed differential expression. 57% and 78% of transcript-SNP pairs significant at FDR of 1% appeared in the list of significant transcript eQTL identified in the original analysis of the EUR and YRI populations individually. 14% of eQTL pairs were identified for transcripts that did not overlap Ensembl annotated transcripts (Figure 4).

Computational time comparison

Next we investigated the computational efficiency of our approach compared to the standard *Cufflinks* pipeline. *Tophat* and *Cufflinks* can be run on each sample separately, but *Cuffdiff2* must be run on all samples simultaneously. While *Cuffdiff2* can make use of many cores on a single computer, is not parallelizable across computers. It has been noted that *Cuffdiff2* can take weeks or longer to run on experiments with a few hundred samples. This issue has led consortia and other groups to rely on unpublished software for transcript abundance estimation[1, 4].

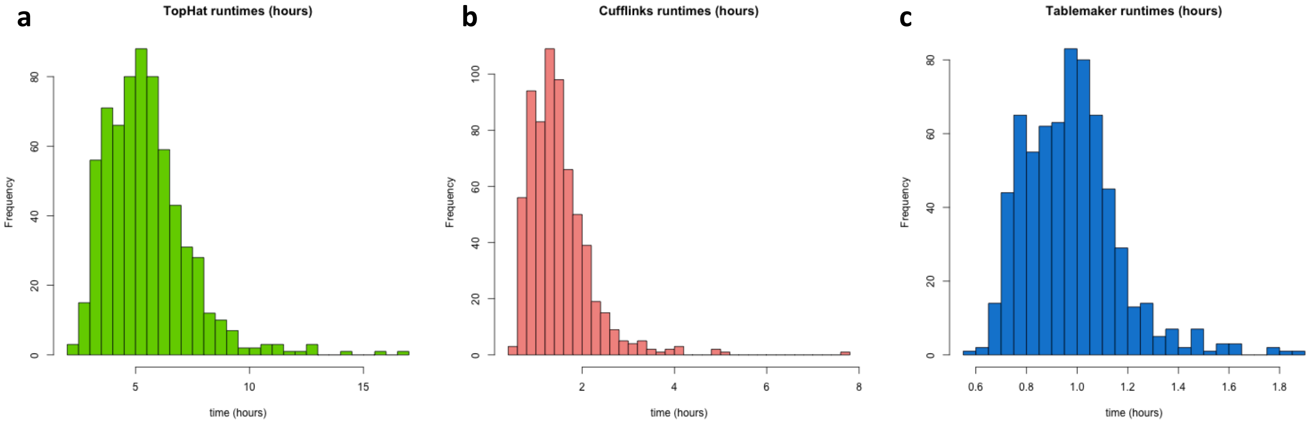

Figure 5: **Timing results for the 667 GEUVADIS samples at each stage of the pipeline. a.** Timing (in hours) for each sample to run through *TopHat2*. **b.** Timing (in hours) for each sample to run through *Cufflinks*. **c.** Timing (in hours) for each sample to run through *Tablemaker*.

We compared each component of the pipeline in terms of computational time on the simulated dataset with 20 samples and 2,745 transcripts. The *Tophat2* - *Cufflinks* -*Tablemaker*-*Ballgown* pipeline was fastest, taking about 3 minutes per sample for *Tablemaker*, 7 seconds to load transcript data into R and less than 1 second for differential expression analysis. This is faster than the recently published *Tophat2* - *Cufflinks* -*Cuffquant* -*Cuffdiff2* pipeline [27], which required about 4 minutes per sample for *Cuffquant*, 23 minutes for differential expression analysis with *Cuffdiff2*. The *Ballgown* -*tablemaker* pipeline was also substantially faster than directly running *Cufflinks* -*Cuffdiff2* , where the *Cuffdiff2* step took about 75 minutes. For all these pipelines, *Tophat2* took about 2 hours per sample and *Cufflinks* about 5 minutes per sample. All possible multicore processes (*Tophat2* , *Cufflinks* , *Cuffdiff2* , *Cuffquant*, *Tablemaker*) were run on 4 cores.

We also calculated the per-sample distribution of processing times for each step in the *Tophat2* - *Cufflinks* - *Tablemaker* pipeline for all 667 samples in the GEUVADIS study [12] (Figure 5a-c). *Tablemaker* took a median of 0.97 hours per sample (IQR 0.24 hours) on a standard 4 core computer; this calculation can be parallelized across samples. By contrast, *Cuffdiff2* would take months to perform this analysis on a standard 4 core computer. *Ballgown* multiclass differential expression analysis between the CEU ($n = 162$), YRI ($n = 163$), FIN ($n = 114$), GBR ($n = 115$) and TSI ($n = 93$) samples for 334,206 transcripts took 42 minutes on a single core Desktop computer.

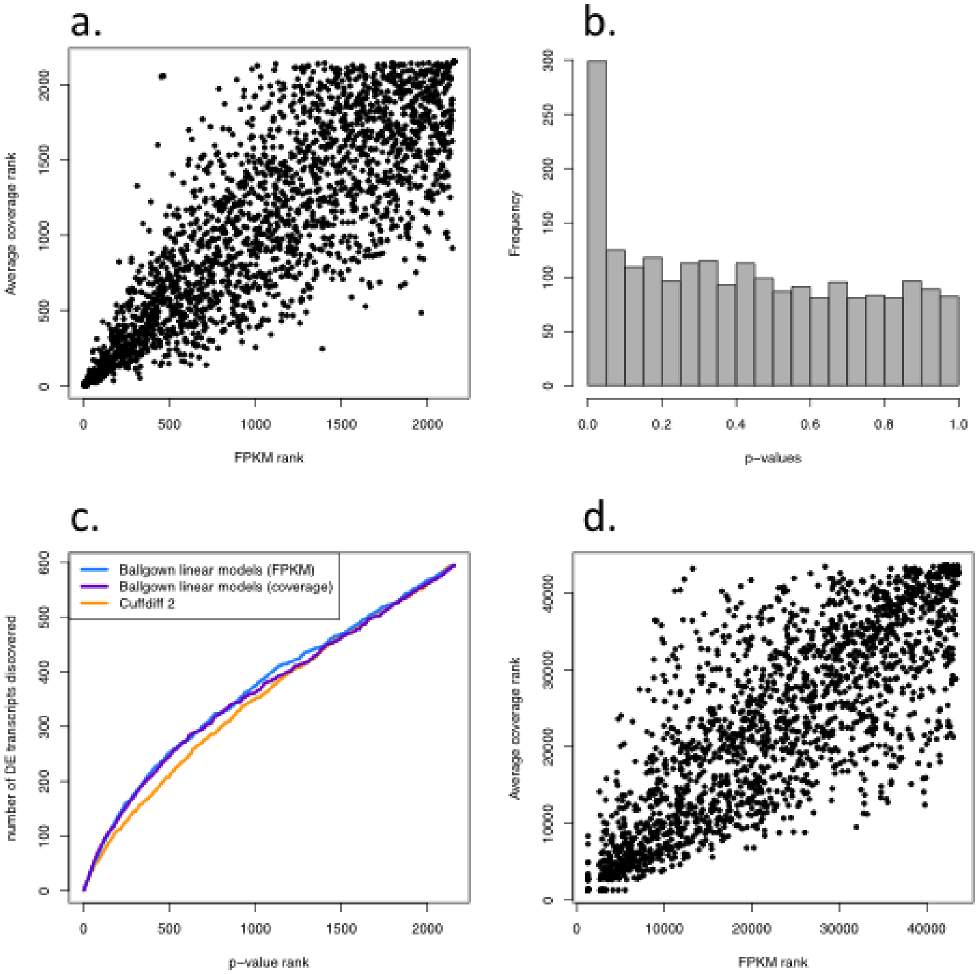

Figure 6: **Using average per-base coverage as transcript expression measurement instead of FPKM. a.** Differential expression ranks in the simulated dataset using FPKM (x-axis) vs. using average coverage (y-axis). **b.** Distribution of p-values from differential expression tests between 10 cases and 10 controls, using average coverage as the expression measurement. This distribution is very similar to the distribution observed when using FPKM as the expression measurement (Figure 2c). **c.** Similar to Figure 2d, ranking of transcripts from most differentially expressed to least (x-axis) versus the number of truly differentially expressed transcripts (y-axis), using the simulated dataset. Among the top 100 transcripts ranked by each method for differential expression, 63 are truly differentially expressed for *Cuffdiff2* , 78 are for *Ballgown* (FPKM), and 82 are for *Ballgown* (average coverage). **d.** Differential expression ranks in the GEUVADIS dataset using FPKM (x-axis) vs. using average coverage (y-axis) to analyze differential expression based on RIN value.

Comparison of average coverage and FPKM for differential expression

There are two major classes of statistical methods for differential expression analysis of RNA-seq: those based on RPKMs or FPKMs, as exemplified by *Cufflinks*, and those based on counting the reads overlapping specific regions, as exemplifed by *DESeq* [2] and *edgeR* [21]. *Tablemaker* produces both FPKM estimates from *Cufflinks* and average coverage of each exon, intron, and transcript (Supplementary Materials: Tablemaker output files). We used our simulated dataset to investigate the impact of using average coverage as the transcript expression measurement, compared to using FPKM, as was done in our previous analyses. To do this comparison, we re-ran the same *Ballgown* model as in our simulation study (Figure 2), but used average coverage as the expression measurement. The differential expression rankings were highly correlated when using either FPKM or average coverage (Figure 6a). The p-value distribution using average coverage (Figure 6b) was similar to the p-value distribution using FPKM (Figure 2c), and the ranking accuracy of the transcript ranks was almost the same, whether average coverage or FPKM was used (Figure 6c). We also observed correlated ranks between the differential expression results by RIN value in the GEUVADIS dataset (Figure 6d). These results confirm the expected result: in differential expression analyses, count-based and FPKM-based (length-normalized) expression measurements perform similarly. *Ballgown* allows users to perform analyses with whatever expression measurement is available in their dataset, so other expression measurements, such as transcripts per million (TPM) [30] could also be explored within our framework.

The *Ballgown* R package includes functions for interactive exploration of the transcriptome assembly, visualization of transcript structures and feature-specific abundances for each locus, and post-hoc annotation of assembled features to annotated features. Direct availability of feature-by-sample expression tables makes it easy to apply alternative differential expression tests or to evaluate other statistical properties of the assembly, such as dispersion of expression values across replicates or genes. The *tablemaker* preprocessor writes the tables directly to disk and they can be loaded into *R* with a single function call. The *Ballgown* , *tablemaker* and *polyester* software are available from GitHub (Supplementary Material: Software), and code and data from the analyses presented here are in the process of being uploaded to GitHub (Supplementary Material: Scripts and Data).

Acknowledgements

The authors would like to thank Peter A.C. 't Hoen and Tuuli Lappalainen for providing QC data and assistance with contacting ArrayExpress. JTL, GP, and BL were partially supported by 1R01GM105705 and AF is supported by a Hopkins Sommer Scholarship.

References

[1] Peter AC't Hoen, Marc R Friedländer, Jonas Almlöf, Michael Sammeth, Irina Pulyakhina, Seyed Yahya Anvar, Jeroen FJ Laros, Henk PJ Buermans, Olof Karlberg, Mathias Brännvall, et al. Reproducibility of high-throughput mrna and small rna sequencing across laboratories. *Nature biotechnology*, 2013.

[2] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome biol*, 11(10):R106, 2010.

[3] Alain Coletta, Colin Molter, Robin Duqué, David Steenhoff, Jonatan Taminau, Virginie De Schaetzen, Stijn Meganck, Cosmin Lazar, David Venet, Vincent Detours, et al. Insilico db genomic datasets hub: an efficient starting point for analyzing genome-wide studies in genepattern, integrative genomics viewer, and r/bioconductor. *Genome Biol*, 13(11):R104, 2012.

[4] Manolis Dermitzakis, Gad Getz, Kristin Ardle, Roderic Guigo, and for the GTEx consortium. Response to: "gtex is throwing away 90% of their data". http://liorpachter.wordpress.com/2013/10/31/response-to-gtex-is-throwing-away-90-of-their-data/, 2013.

[5] Sarah Djebali, Carrie A Davis, Angelika Merkel, Alex Dobin, Timo Lassmann, Ali Mortazavi, Andrea Tanzer, Julien Lagarde, Wei Lin, Felix Schlesinger, et al. Landscape of transcription in human cells. *Nature*, 489(7414):101–108, 2012.

[6] Paul Flicek, M Ridwan Amode, Daniel Barrell, Kathryn Beal, Konstantinos Billis, Simon Brent, Denise Carvalho-Silva, Peter Clapham, Guy Coates, Stephen Fitzgerald, et al. Ensembl 2014. *Nucleic acids research*, 42(D1):D749–D755, 2014.

[7] Alyssa C Frazee, Sarven Sabunciyan, Kasper D Hansen, Rafael A Irizarry, and Jeffrey T Leek. Differential expression analysis of rna-seq data at single-base resolution. *Biostatistics*, page kxt053, 2014.

[8] Manfred G Grabherr, Brian J Haas, Moran Yassour, Joshua Z Levin, Dawn A Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, et al. Full-length transcriptome assembly from rna-seq data without a reference genome. *Nature biotechnology*, 29(7):644–652, 2011.

[9] Brenton R Graveley, Angela N Brooks, Joseph W Carlson, Michael O Duff, Jane M Landolin, Li Yang, Carlo G Artieri, Marijke J van Baren, Nathan Boley, Benjamin W Booth, et al. The developmental transcriptome of drosophila melanogaster. *Nature*, 471(7339):473–479, 2011.

[10] Daehwan Kim, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L Salzberg. Tophat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*, 14(4):R36, 2013.

[11] Sang Cheol Kim, Yeonjoo Jung, Jinah Park, Sooyoung Cho, Chaehwa Seo, Jaesang Kim, Pora Kim, Jehwan Park, Jihae Seo, Jiwoong Kim, et al. A high-dimensional, deep-sequencing study of lung adenocarcinoma in female never-smokers. *PloS one*, 8(2):e55596, 2013.

[12] Tuuli Lappalainen, Michael Sammeth, Marc R Friedländer, Peter AC't Hoen, Jean Monlong, Manuel A Rivas, Mar Gonzàlez-Porta, Natalja Kurbatova, Thasso Griebel, Pedro G Ferreira, et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 2013.

[13] Jeffrey T Leek and John D Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS genetics*, 3(9):e161, 2007.

[14] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, et al. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[15] Ryan Lister, Eran A Mukamel, Joseph R Nery, Mark Urich, Clare A Puddifoot, Nicholas D Johnson, Jacinta Lucero, Yun Huang, Andrew J Dwork, Matthew D Schultz, et al. Global epigenomic reconfiguration during mammalian brain development. *Science*, 341(6146), 2013.

[16] Ryan Lister, Mattia Pelizzola, Yasuyuki S Kida, R David Hawkins, Joseph R Nery, Gary Hon, Jessica Antosiewicz-Bourget, Ronan O'Malley, Rosa Castanon, Sarit Klugman, et al. Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature*, 471(7336):68–73, 2011.

[17] Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature methods*, 5(7):621–628, 2008.

[18] Joseph N Paulson, O Colin Stine, Héctor Corrada Bravo, and Mihai Pop. Differential abundance analysis for microbial marker-gene surveys. *Nature methods*, 2013.

[19] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–909, 2006.

[20] Brent G Richter and David P Sexton. Managing and analyzing next-generation sequence data. *PLoS computational biology*, 5(6):e1000369, 2009.

[21] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

[22] Andreas Schroeder, Odilo Mueller, Susanne Stocker, Ruediger Salowsky, Michael Leiber, Marcus Gassmann, Samar Lightfoot, Wolfram Menzel, Martin Granzow, and Thomas Ragg. The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC molecular biology*, 7(1):3, 2006.

[23] Andrey A Shabalin. Matrix eqtl: ultra fast eqtl analysis via large matrix operations. *Bioinformatics*, 28(10):1353–1358, 2012.

[24] Gordon K Smyth. Limma: linear models for microarray data. In *Bioinformatics and computational biology solutions using R and Bioconductor*, pages 397–420. Springer, 2005.

[25] Lincoln D Stein et al. The case for cloud computing in genome informatics. *Genome Biol*, 11(5):207, 2010.

[26] John D Storey, Wenzhong Xiao, Jeffrey T Leek, Ronald G Tompkins, and Ronald W Davis. Significance analysis of time course microarray experiments. *Proceedings of the National Academy of Sciences of the United States of America*, 102(36):12837–12842, 2005.

[27] C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen, and J. L. Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.*, Mar 2014.

[28] Cole Trapnell, David G Hendrickson, Martin Sauvageau, Loyal Goff, John L Rinn, and Lior Pachter. Differential analysis of gene regulation at transcript resolution with rna-seq. *Nature biotechnology*, 31(1):46–53, 2012.

[29] Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5):511–515, 2010.

[30] Günter P Wagner, Koryu Kin, and Vincent J Lynch. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory in Biosciences*, 131(4):281–285, 2012.

[31] Liying Yan, Mingyu Yang, Hongshan Guo, Lu Yang, Jun Wu, Rong Li, Ping Liu, Ying Lian, Xiaoying Zheng, Jie Yan, et al. Single-cell rna-seq profiling of human preimplantation embryos and embryonic stem cells. *Nature structural & molecular biology*, 2013.

[32] Robert S Young, Ana C Marques, Charlotte Tibbit, Wilfried Haerty, Andrew R Bassett, Ji-Long Liu, and Chris P Ponting. Identification and properties of 1,119 candidate lincrna loci in the drosophila melanogaster genome. *Genome biology and evolution*, 4(4):427–442, 2012.