bioRxiv preprint doi: https://doi.org/10.1101/006015; this version posted June 6, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Polyester: simulating RNA-seq datasets with differential transcript expression

Alyssa C. Frazee [1,3], Andrew E. Jaffe [1,2,3], Ben Langmead [1,3,4] and Jeffrey T. Leek [1,3*]

June 6, 2014

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

* *Correspondence to jtleek@gmail.com*

Abstract

Statistical methods development for differential expression analysis of RNA sequencing (RNA-seq) requires software tools to assess accuracy and error rate control. Since true differential expression status is often unknown in experimental datasets, artificially-constructed datasets must be utilized, either by generating costly spike-in experiments or by simulating RNA-seq data. *Polyester* is an R package designed to simulate RNA-seq data, beginning with an experimental design and ending with collections of RNA-seq reads. The main advantage of *Polyester* is the ability to simulate isoform-level differential expression across biological replicates for a variety of experimental designs at the read level. Differential expression signal can be simulated with either built-in or user-defined statistical models. *Polyester* is available on GitHub at https://github.com/alyssafrazee/polyester.

1 Introduction

RNA sequencing (RNA-seq) experiments have become increasingly popular as a means to study gene expression. There are a range of statistical methods for differential expression analysis of RNA-seq data [6]. The developers of statistical methodology for RNA-seq need to test whether their tools are performing correctly. Often, accuracy tests cannot be performed on real datasets because true gene expression levels and expression differences between populations are usually unknown, and spike-in experiments are costly in terms of both time and money.

bioRxiv preprint doi: https://doi.org/10.1101/006015; this version posted June 6, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Instead, researchers often use computational simulations to create datasets with a known underlying truth. *Polyester* is a new R package for simulating RNA-seq reads. Its main advantage is its built-in functionality for inducing differential expression into the simulated experiment at the gene or isoform level at the sequencing step of the RNA-seq pipeline. In the literature, simulated expression measurements used to evaluate differential expression tools are usually generated directly from a statistical model (e.g. [7], [1]), but these simulated scenarios do not account for variability in expression measurements that arises during upstream steps in RNA-seq data analysis, such as read alignment.

Many existing RNA-seq read simulators were not designed for directly simulating experiments with biological replicates and differential expression. For example, the `rsem-simulate-reads` utility shipped with RSEM [5] requires aligning real sequencing reads to develop a sequencing model before simulating reads and differential expression simulation is not built-in. Neither FluxSimulator [4] nor BEERS [3] have a built-in mechanism for inducing differential expression, nor do they provide a way to define a model for feature expression measurements across replicates. TuxSim has been used to simulate RNA-seq datasets with differential expression [8], but it is not publicly available.

Polyester was created to fulfill the need for a tool to simulate RNA-seq reads for an experiment with replicates and well-defined differential expression. Users can easily simulate small experiments from a few genes or a single chromosome. This can reduce computational time in simulation studies when computationally intensive steps such as read alignment must be performed as part of the simulation. *Polyester* is open-source, cross-platform, and freely available for download at https://github.com/alyssafrazee/polyester.

2 Methods

Polyester takes annotated transcript sequences as input. These can be provided as cDNA sequences FASTA format, or as a GTF file with gene annotations along with full-chromosome DNA sequences. The number of reads for each gene or transcript can be set using *Polyester*'s built-in model for differential expression. The built-in transcript abundance model assumes that the distribution of the number of reads to simulate from each transcript is negative binomial, across biological replicates. The negative binomial model for read counts has been shown to satisfactorily capture biological and technical variability ([1], [7]).

Specifically, define Y_{ijk} as the number of reads simulated from replicate i, experimental condition j, and transcript k ($i = 1, ..., n_j$, $j = 1, 2$, and $k = 1, ..., N$, where n_j is the number of replicates in condition j and N is the total number of transcripts provided). *Polyester* assumes $Y_{ijk} \sim \text{NegativeBinomial}(\text{mean} = \mu_{jk}, \text{dispersion} = r_{jk})$. In this negative binomial parameterization, $E(Y_{ijk}) = \mu_{jk}$ and $\text{Var}(Y_{ijk}) = \mu_{jk} + \frac{\mu_{jk}^2}{r_{jk}}$, so each transcript's expression variance is quadratically related to its baseline mean expression. The user can provide μ_{jk} for each transcript k and experimental group j. In particular, the user can relate transcript k's length to μ_{jk}.

By default, $r_{jk} = \frac{\mu_{jk}}{3}$, which means $\text{Var}(Y_{ijk}) = 4\mu_{jk}$. The user can adjust r_{jk} on a per-transcript basis as needed, to explore different mean/variance expression models. Differential

bioRxiv preprint doi: https://doi.org/10.1101/006015; this version posted June 6, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

expression can be set by providing a fold change for each transcript. Initially, a baseline mean μ_k is provided for each transcript, and μ_{1k} and μ_{2k} are set to μ_k. Then, if fold change λ is provided, μ_{1k} and μ_{2k} are adjusted: if $\lambda > 1$, $\mu_{1k} = \lambda\mu_k$, and if $\lambda < 1$, $\mu_{2k} = \frac{1}{\lambda}\mu_k$.

Alternatively, *polyester* allows users to individually specify the number of reads to generate from each transcript, for each sample. This provides the capability to simulate complex experimental designs, such as timecourse experiments or multi-group studies, and to explore the effects of a wide variety of experimental parameters on differential expression results.

After the transcripts have been defined and each transcript's abundance in the simulated experiment has been determined, *polyester* simulates the RNA sequencing process, described in detail in [6], beginning at the fragmentation step: first, all transcripts present are broken into short fragments. Fragment lengths are drawn from a normal distribution, with mean μ_{fl} and standard deviation σ_{fl}. By default, $\mu_{fl} = 250$ nucleotides and $\sigma_{fl} = 25$. Fragment locations are chosen at random. *Polyester* simulates reads from an unstranded protocol, so each fragment is reverse-complemented with probability 0.5: it is equally likely that a cDNA fragment originated from either the Watson or Crick strand.

After fragmentation is the *sequencing* step: the first R nucleotides are read from each fragment. If paired-end reads are desired, the last R nucleotides are also read from that fragment. The default read length is $R = 100$. *Polyester* assumes a uniform sequencing error model, so an incorrect nucleotide is read from the fragment with probability p_e (default 0.005). Simulated reads are automatically written to FASTA files on disk. The FASTA files identify which reads originated from which transcripts, facilitating assessment of downstream alignment accuracy.

3 Example Analysis

To demonstrate a use case for *polyester*, we simulated a small differential expression experiment and tested the accuracy of Ballgown's statistical models [2] at detection of this differential expression. The built-in read count model was used, and for each transcript k on human chromosome 22 (hg19 build, $N = 918$), μ_k was set to length(transcript$_k$)/5, which corresponds to approximately 20x coverage for 100-bp reads. We randomly chose 75 transcripts to have $\lambda = 3$ and 75 to have $\lambda = 1/3$; the rest had $\lambda = 1$. For $n_j = 7$ replicates in each group j, we simulated paired-end reads from 250-base fragments ($\sigma_{fl} = 25$), with error probability 0.005. Simulated reads were aligned to hg19 with TopHat 2.0.11 [9], and Cufflinks 2.2.1 [10] was used to obtain transcript-specific expression estimates. We then ran transcript-level differential expression tests using Ballgown [2], and by matching each Cufflinks transcript to the closest transcript in the annotation set used to generate the reads, we were able to examine Ballgown's accuracy using an ROC curve (Figure 1).

This example is just one of many ways *polyester* can be used to explore the effects of analysis choices on differential expression results. Data and scripts used in the example analysis are available at https://github.com/alyssafrazee/polyester_code.

bioRxiv preprint doi: https://doi.org/10.1101/006015; this version posted June 6, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

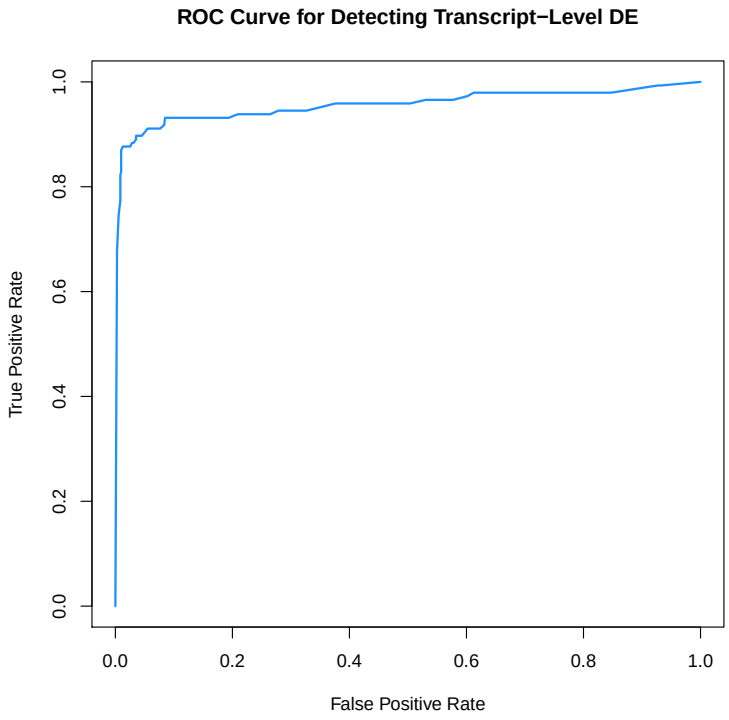

Figure 1: ROC curve for detecting transcript-level differential expression simulated with *polyester*. The pre-set differential expression was accurately detected using *Ballgown*'s differential expression tests.

bioRxiv preprint doi: https://doi.org/10.1101/006015; this version posted June 6, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Acknowledgements

JL and BL are supported by NIH R01 GM105705. AF is supported by a Hopkins Sommer Scholarship. AJ is supported by the Lieber Institute for Brain Development

References

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

[2] Alyssa C Frazee, Geo Pertea, Andrew E Jaffe, Ben Langmead, Steven L Salzberg, and Jeffrey T Leek. Flexible isoform-level differential expression analysis with Ballgown. biorXiv doi: http://dx.doi.org/10.1101/003665.

[3] Gregory R Grant, Michael H Farkas, Angel D Pizarro, Nicholas F Lahens, Jonathan Schug, Brian P Brunk, Christian J Stoeckert, John B Hogenesch, and Eric A Pierce. Comparative analysis of RNA-seq alignment algorithms and the RNA-seq unified mapper (RUM). *Bioinformatics*, 27(18):2518–2528, 2011.

[4] Thasso Griebel, Benedikt Zacher, Paolo Ribeca, Emanuele Raineri, Vincent Lacroix, Roderic Guigó, and Michael Sammeth. Modelling and simulating generic RNA-seq experiments with the flux simulator. *Nucleic acids research*, 40(20):10073–10083, 2012.

[5] Bo Li and Colin N Dewey. RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinformatics*, 12(1):323, 2011.

[6] Alicia Oshlack, Mark D Robinson, Matthew D Young, et al. From RNA-seq reads to differential expression results. *Genome Biology*, 11(12):220, 2010.

[7] 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.

[8] 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, 2013.

[9] Cole Trapnell, Lior Pachter, and Steven L Salzberg. TopHat: discovering splice junctions with RNA-seq. *Bioinformatics*, 25(9):1105–1111, 2009.

[10] 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.