BIOINFORMATICS

Vol. 00 no. 00 2016
Pages 1–3

Fast Bootstrapping-Based Estimation of Confidence Intervals of Expression Levels and Differential Expression from RNA-Seq Data

Igor Mandric [1*], Yvette Temate-Tiagueu [1], Tatiana Shcheglova [3], Sahar Al Seesi [2,3], Alex Zelikovsky [1], and Ion I. Măndoiu [2*]

[1]Department of Computer Science, Georgia State University, Atlanta, Georgia, USA
[2]Computer Science and Engineering Department, University of Connecticut, Storrs, CT, USA
[3]Immunology Department, University of Connecticut Health Center, Farmington, CT, USA

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Summary: This note presents IsoEM2 and IsoDE2, new versions with enhanced features and faster runtime of the IsoEM and IsoDE packages for expression level estimation and differential expression. IsoEM2 estimates FPKM and TPM levels for genes and isoforms with confidence intervals through bootstrapping, while IsoDE2 performs differential expression (DE) analysis using the bootstrap samples generated by IsoEM2. Both tools are available with a command line interface as well as a graphical user interface through wrappers for the Galaxy platform.

Availability: The source code of this software suite is available at `https://github.com/mandricigor/isoem2`. The Galaxy wrappers are available at `https://toolshed.g2.bx.psu.edu/view/saharlcc/isoem2_isode2/c6d2dbdf0a4d`

Contact: imandric1@student.gsu.edu, ion@engr.uconn.edu

1 INTRODUCTION

RNA-Seq experiments use high-throughput sequencing to generate both sequence and abundance information about expressed gene isoforms. The two most common applications of RNA-Seq are to quantify gene/isoform expression levels in single samples and identify genes/isoforms that are differentially expressed between samples. Both applications are affected by noise introduced by library preparation and sequencing errors as well as ambiguities in read mapping.

Numerous tools for RNA-Seq quantification have been developed to address these challenges. A comprehensive assessment study (Kanitz *et al.*, 2015) recently compared the most commonly used tools BitSeq (Glaus *et al.*, 2012), CEM (Li and Jiang, 2012), Cufflinks (Trapnell *et al.*, 2010), eXpress (Roberts and Pachter, 2013), IsoEM (Nicolae *et al.*, 2011), MMSEQ (Turro *et al.*, 2011), RSEM (Li and Dewey, 2011), rSeq (Salzman *et al.*, 2011), Sailfish (Patro *et al.*, 2014), Scripture (Guttman *et al.*, 2010), and TIGAR2 (Nariai *et al.*, 2014). The results in (Kanitz *et al.*, 2015) show that IsoEM (Nicolae *et al.*, 2011) has one of the highest accuracies in all experiments (see also Supplementary Table 1) while being orders of magnitude faster than the other best-performing methods.

IsoEM is based on the *Expectation-Maximization (EM)* algorithm. Its probabilistic model takes into account the fragment length distribution (with mean/standard deviations specified by the user or automatically inferred when using paired-end reads) and incorporates base quality scores and strand information (if available). IsoDE (Al Seesi *et al.*, 2014) performs differential gene expression analysis using FPKM/TPM values estimated for bootstrap samples generated by re-sampling alignments. Although bootstrapping is computationally expensive, the high speed of IsoEM makes the running time of IsoDE practical.

Here we introduce IsoEM2, a new version of the IsoEM package that uses bootstrapping to infer confidence intervals for gene and isoform expression level estimates. The accompanying differential expression tool IsoDE2 has also been updated to take advantage of the fast in-memory bootstrapping of IsoEM2, resulting in speedups of over 200× over the original version in (Al Seesi *et al.*, 2014). Compared to the previous versions, the main enhancements are the addition of confidence intervals for FPKM and TPM estimates produced by IsoEM2, the substantially faster running time for performing bootstrapping with IsoDE2, and the development of Galaxy wrappers making both IsoEM2 and IsoDE2 easy to use via a user-friendly web interface.

Table 1 provides a feature-based comparison of the tools included in the assessment of (Kanitz *et al.*, 2015) and the subsequently published Kallisto (Bray *et al.*, 2016). IsoEM2 offers a broad range of features and achieves one of the highest accuracies (Supplementary Table 1). It is also significantly faster than the other best-performing methods with the exception of Kallisto. On real datasets with over 100M read pairs, the HISAT2/IsoEM pipeline requires just over 1 hour to perform *both* read alignment and RNA-Seq quantification with 200 bootstraps using 16 CPU cores (Supplementary Table 3). Although the alignment-free Kallisto is 5-10× faster, its confidence intervals are substantially less reliable than those generated by IsoEM2 (see Tables 2 and 3).

*To whom correspondence should be addressed.

Table 1. Feature-based comparison of state-of-the-art RNA-Seq quantification tools. In the reference row, G stands for genome and T for transcriptome.

Feature \ Tool	IsoEM2	Kallisto	BitSeq	CEM	Cufflinks	eXpress	MMSEQ	RSEM	rSeq	Sailfish	Scripture	TIGAR2
Alignment free	✗	✓	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗
Reference	G/T	T	T	G	G	T	T	G/T	T	T	G	T
Confidence intervals	✓	✓	✗	✗	✓	✓	✗	✓	✗	✗	✗	✗
Indels	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓
Integrated DE	✓	✓	✓	✗	✓	✓	✓	✓	✗	✓	✗	✗
GUI	✓	✗	✗	✗	✓	✗	✗	✗	✗	✓	✗	✗
Multi-threading	✓	✓	✓	✗	✓	✓	✗	✓	✗	✓	✗	✗
Frag. length distribution	✓	✓	✗	✗	✓	✓	✓	✓	✗	✗	✗	✗
Sequence bias	✓	✓	✓	✓	✓	✓	✓	✗	✗	✓	✗	✓

2 SOFTWARE FEATURES

2.1 IsoEM2

IsoEM2 takes as input aligned RNA-Seq reads in (compressed) SAM format and outputs FPKM and TPM estimates of gene and isoform expression levels. Unlike the original implementation in (Nicolae *et al.*, 2011), IsoEM2 computes confidence intervals for the estimates using the bootstrap method (Efron and Efron, 1982). In each run IsoEM2 generates N bootstrap estimates by in-memory re-sampling of the compatible read alignments. For each genomic feature (gene or isoform) and given confidence level $C \in (0, 1)$, the confidence interval $[c_{low}, c_{hi}]$ is computed from the N bootstrap estimates $\mathcal{B} = \{b_1, \ldots, b_N\}$ by setting c_{low} and c_{hi} equal to the k-th smallest, respectively k-th largest element of $\mathcal{B}$, where $k = \lfloor N(1 - C)/2 \rfloor$. By default IsoEM2 uses $C = 0.95$ and $N = 200$, but these settings can be changed by the user. IsoEM2 generates four tab delimited output files for gene/isoform FPKM/TPM estimates. Each file includes a point estimate and the confidence interval for each feature. Additionally, it generates a compressed archive containing the bootstrap estimates used to compute the confidence intervals; these archives can be used for DE analysis using IsoDE2.

Besides the command-line version, IsoEM2 is also available with a user-friendly GUI through a Galaxy wrapper (Supplementary Figure 1). The wrapper can be downloaded from the Galaxy Tool Shed and installed on any local installation of Galaxy. The Galaxy tool is designed to work with both single-end and paired-end Illumina RNA-Seq reads as well as single-end Ion Torrent reads. It takes as input unaligned RNA-Seq reads and it maps them to a transcriptome reference selected by the user through the wrapper interface. The aligned reads are then automatically processed by IsoEM2. In addition to IsoEM2, the wrapper needs HISAT2 (Kim *et al.*, 2015) to be installed on the Galaxy server.

2.2 IsoDE2

IsoDE2, which is an extension of IsoDE (Al Seesi *et al.*, 2014), performs differential expression (DE) analysis using bootstrap samples generated by IsoEM2. To test for differential expression, the bootstrap expression level estimates generated for the two conditions by IsoEM2 are paired and used to compute for each gene a set of fold change estimates. A confident fold change f is then computed for a user-specified significance level under the null hypothesis that fold changes obtained from bootstrap estimates are equally likely to be greater or smaller than f. For details on the format of IsoDE2 output files see Supplementary data.

3 EXPERIMENTAL RESULTS

We conducted experiments to assess both the running time and the accuracy of confidence intervals of the updated IsoEM2/IsoDE2 suite and of the newly published Kallisto (Bray *et al.*, 2016). We only included Kalisto in this comparison since IsoEM was already shown to dominate in accuracy and/or running time the methods included in the comparative assessment of (Kanitz *et al.*, 2015).

3.1 Runtime Comparison

The running time of IsoEM2 is much smaller compared to the bootstrapping step (called IsoBoot) of the old IsoDE. This is achieved by implementing the re-sampling in IsoEM2 based on internal data structures representing connected components of the read-isoform compatibility graph (Nicolae *et al.*, 2011). To assess the runtime improvement, we used two mouse retina RNA-Seq datasets from (Karunakaran *et al.*, 2016) with ∼100M unaligned read pairs each. On each dataset, generating 200 bootstrap samples with IsoEM2 has a speed-up of over 200× compared to IsoBoot (Supplementary Table 2). Although Kallisto is 5-10× faster (Supplementary Table 3), the HISAT2/IsoEM pipeline remains very practical, requiring just over 1 hour to perform read alignment and RNA-Seq quantification with 200 bootstraps using 16 CPU cores.

3.2 Accuracy Comparison

To assess the accuracy of gene/isoform expression level estimates we computed the Pearson correlation with the known ground truth. To assess the quality of confidence intervals we used the percentage of genes for which confidence intervals contained the known ground truth. Since Kallisto does not output explicit confidence intervals, we ran it with the "-B 200" option to generate 200 bootstrap estimates and computed confidence intervals using the approach described in Section 2 for IsoEM2.

Table 2. Gene expression level estimation accuracy on simulated RNA-Seq datasets with 1M-10M single-end reads from Kanitz *et al.* (2015).

.

Number of reads		1M	3M	10M
All genes				
Pearson correlation	IsoEM2	0.995	0.996	0.996
	Kallisto	0.84	0.84	0.84
Confidence interval coverage for C=95%	IsoEM2	0.94	0.95	0.94
	Kallisto	0.80	0.78	0.78
Genes with non-zero ground-truth				
Pearson correlation	IsoEM2	0.96	0.98	0.98
	Kallisto	0.96	0.98	0.98
Confidence interval coverage for C=95%	IsoEM2	0.74	0.77	0.72
	Kallisto	0.33	0.27	0.38

Table 3. Transcript expression level estimation accuracy on simulated RNA-Seq datasets with 1M-10M single-end reads from Kanitz *et al.* (2015).

Number of reads		1M	3M	10M
All isoforms				
Pearson correlation	IsoEM2	0.98	0.98	0.98
	Kallisto	0.89	0.89	0.89
Confidence interval coverage for C=95%	IsoEM2	0.95	0.95	0.94
	Kallisto	0.89	0.86	0.82
Isoforms with non-zero ground truth				
Pearson correlation	IsoEM2	0.90	0.94	0.96
	Kallisto	0.90	0.94	0.96
Confidence interval coverage for C=95%	IsoEM2	0.59	0.64	0.61
	Kallisto	0.44	0.38	0.28

Tables 2 and 3 give Pearson correlations and confidence interval coverages for gene, respectively isoform expression level estimates obtained by IsoEM2 and Kallisto on datasets with 1M-10M simulated single-end reads from (Kanitz *et al.*, 2015). The confidence interval coverage for $C = 95\%$ reports how frequently the 95% CI estimated by IsoEM2 or Kallisto contains the true gene expression value. The accuracy metrics are computed both over the subset of genes/isoforms with non-zero ground truth, as in (Kanitz *et al.*, 2015), and over all genes/isoforms. We note that, although Kallisto has similar Pearson correlations to IsoEM2 over the genes and isoforms with non-zero truth, its Pearson correlation is significantly lower than that of IsoEM2 when including isoforms with zero ground-truth. More importantly, for all considered sets of genes and isoforms, the coverage of 95%-confidence intervals computed by Kallisto is substantially lower than that of IsoEM2.

4 CONCLUSION

In this note we presented the IsoEM2/IsoDE2 suite for RNA-Seq gene and isoform expression level estimation and differential expression analysis. The main feature of these tools is the fast non-parametric computation of confidence intervals and identification of DE genes based on bootstrapping. Although when including alignment time IsoEM2 is 5-10$\times$ slower than the alignment free tool Kallisto, the confidence intervals computed by IsoEM2 are substantially more accurate than those generated by Kallisto. We hope that the improved accuracy combined with the ease of use provided by the Galaxy interface will make the IsoEM2/IsoDE2 suite a preferred choice for the analysis of RNA-Seq datasets.

ACKNOWLEDGEMENTS

This work was supported in part by a GSU Molecular Basis of Disease Fellowship to IM and NSF awards 1564899, 1564936, 1618347, and 16119110 to IIM and AZ.

REFERENCES

Al Seesi, S., Tiagueu, Y. T., Zelikovsky, A., *et al.* (2014). Bootstrap-based differential gene expression analysis for RNA-Seq data with and without replicates. *BMC genomics*, **15**(Suppl 8), S2.

Bray, N., Pimentel, H., Melsted, P., *et al.* (2016). Near-optimal RNA-Seq quantification. *Nature Biotechnology*, **34**, 525–527.

Efron, B. and Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*, volume 38. SIAM.

Glaus, P., Honkela, A., and Rattray, M. (2012). Identifying differentially expressed transcripts from RNA-seq data with biological variation. *Bioinformatics*, **28**(13), 1721.

Guttman, M., Garber, M., Levin, J. Z., *et al.* (2010). Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature Biotechnology*, **28**, 503–510.

Kanitz, A., Gypas, F., Gruber, A. J., *et al.* (2015). Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data. *Genome Biology*, **16**(1), 1–26.

Karunakaran, D., Seesi, S. A., Banday, A., *et al.* (2016). Network-based bioinformatics analysis of spatio-temporal RNA-Seq data reveals transcriptional programs underpinning normal and aberrant retinal development. *BMC Genomics*, **17(Suppl 5):495**, 477–492.

Kim, D., Langmead, B., and Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*, **12**, 357–350.

Li, B. and Dewey, C. N. (2011). RSEM: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC Bioinformatics*, **12**(1), 323.

Li, W. and Jiang, T. (2012). Transcriptome assembly and isoform expression level estimation from biased RNA-Seq reads. *Bioinformatics*, **28**(22), 2914.

Nariai, N., Kojima, K., Mimori, T., *et al.* (2014). TIGAR2: sensitive and accurate estimation of transcript isoform expression with longer RNA-Seq reads. *BMC Genomics*, **15**(10), S5.

Nicolae, M., Mangul, S., Măndoiu, I. I., *et al.* (2011). Estimation of alternative splicing isoform frequencies from RNA-Seq data. *Algorithms for Molecular Biology*, **6**(1), 1.

Patro, R., Mount, S. M., and Kingsford, C. (2014). Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nature Biotechnology*, **32**, 462–464.

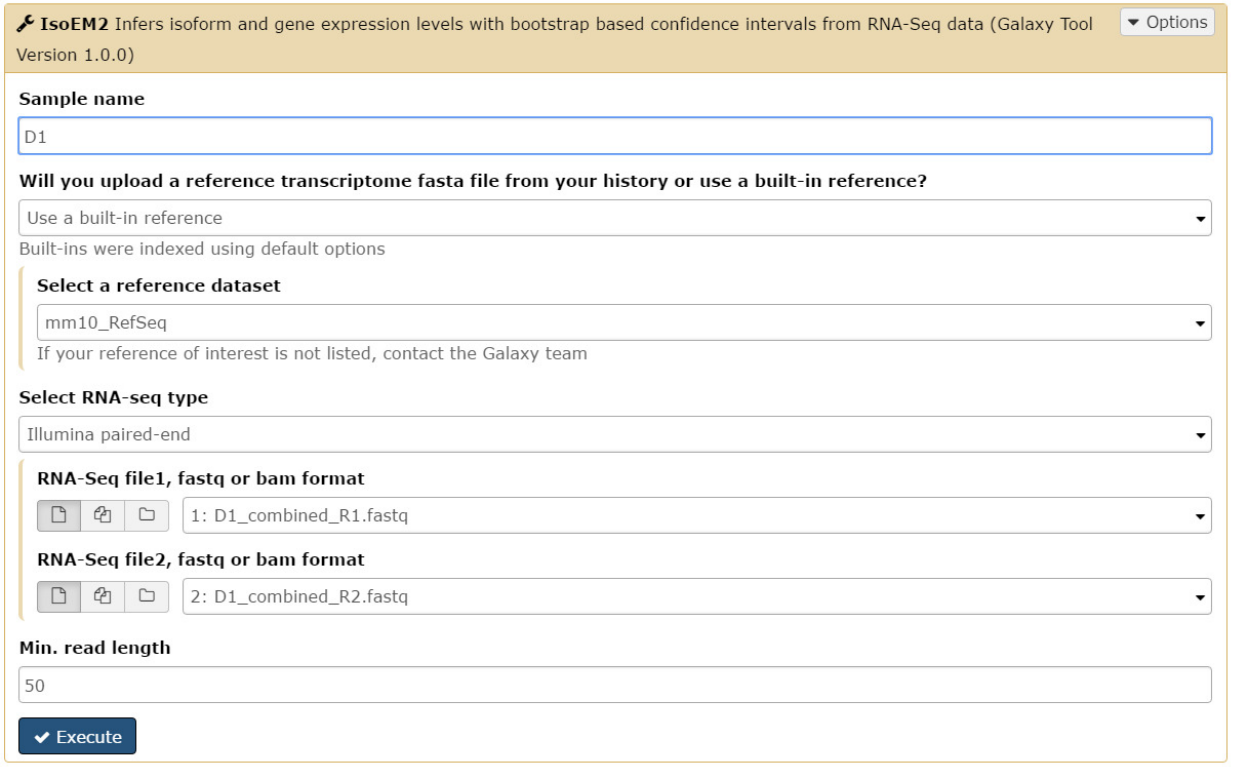
Roberts, A. and Pachter, L. (2013). Streaming fragment assignment for real-time analysis of sequencing experiments. *Nature Methods*, **10**, 7173.

Salzman, J., Jiang, H., and Wong, W. H. (2011). Statistical modeling of RNA-Seq data. *Statistical Science*, **26**(1), 62–83.

Trapnell, C., Williams, B. A., Pertea, G., *et al.* (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, **28**, 511–515.

Turro, E., Su, S.-Y., Goncalves, A., *et al.* (2011). Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. *Genome Biology*, **12**(2), R13.

Supplementary Data for "Fast Bootstrapping-Based Estimation of Confidence Intervals of Expression Levels and Differential Expression from RNA-Seq Data"

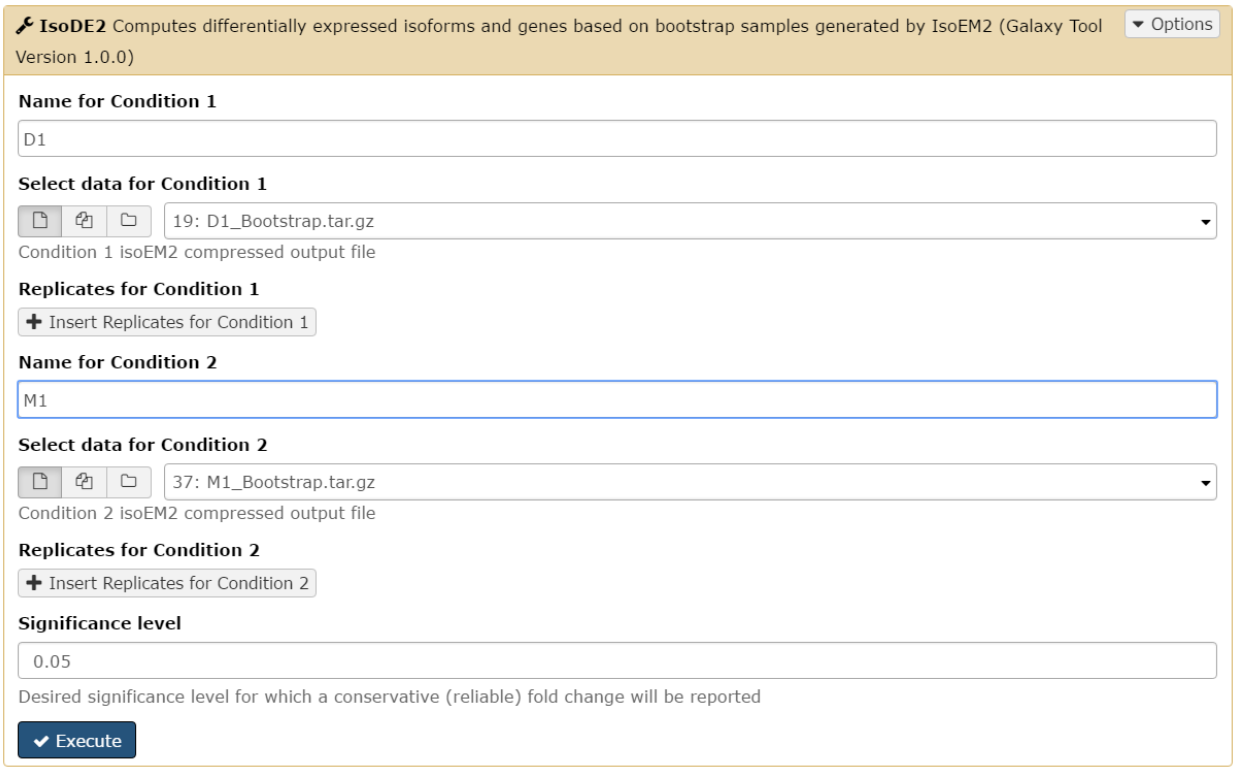


Supplementary Figure 1: User interface for IsoEM2 on Galaxy

IsoDE2 input/output description

As for IsoEM2, IsoDE2 is made available both through a command line interface and a Galaxy GUI (Supplementary Figure 2). This requires the user to provide one or more (if replicates are available) IsoEM2 bootstrapping archives for each condition along with a desired significance level α. By default, IsoDE2 generates four tab-delimited output files containing DE results based on gene/isoform FPKM/TPM estimates. The four files have identical format with the following fields:

- Gene/isoform ID
- Confident log2(FC): the base 2 logarithm of the largest condition 2 vs condition 1 fold change of gene/isoform FPKM/TPM estimates supported by the bootstrap samples at a significance level of α

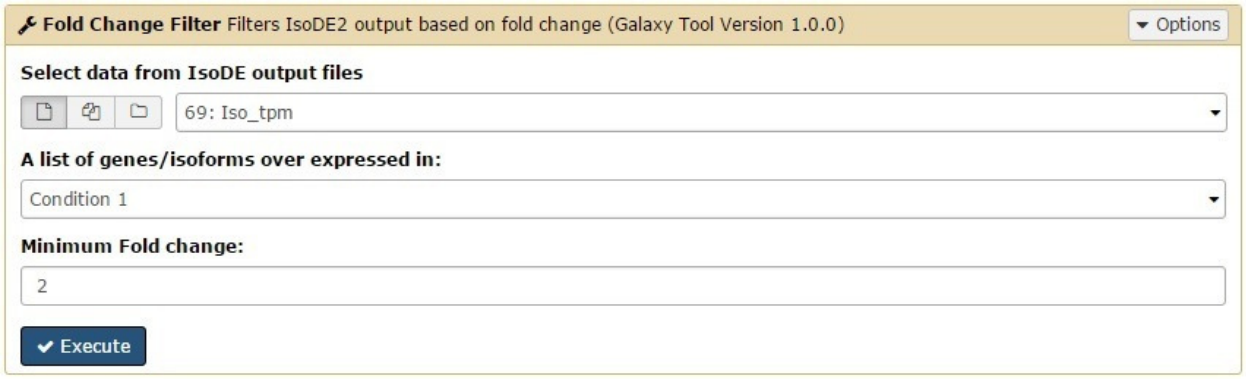


Supplementary Figure 2: User interface for IsoDE2 on Galaxy

(see [3] for details on the model used to compute bootstrap support). Positive values represent over-expression in condition 2, negative values representing over-expression in condition 1, and zero values indicate that no significant change was detected.

- Single run log2(FC): the base 2 logarithm of the ratio between expression levels estimated by IsoEM2 for condition 2 and condition 1 (or the mean estimates in case replicates are provided for the two conditions).
- Condition 1/2 FPKM/TPM: expression level estimated for conditions 1 and 2 (mean values in case of replicates).

The subset of genes with confident fold change above a user specified threshold can be selected from the output of IsoDE2 using a separate Galaxy filtering tool (Supplementary Figure 3).



Supplementary Figure 3: User interface for IsoDE2 fold change filter on Galaxy

Supplementary Table 1: Pearson correlation between estimated and ground truth gene/isoform expression levels on simulated RNA-Seq datasets with 1M-10M single-end reads from [1]. As in [1], correlation is computed only over the genes/isoforms with non-zero simulated expression. Highest value for each dataset is typeset in bold.

	IsoEM2	Kallisto	BitSeq	CEM	Cufflinks	eXpress	MMSEQ	RSEM	rSeq	Sailfish	Scripture	TIGAR2
Pearson correlation over isoforms with non-zero ground truth												
1M reads	**0.90**	**0.90**	0.89	0.89	0.76	**0.90**	**0.90**	**0.90**	0.87	0.89	0.75	**0.90**
3M reads	**0.94**	**0.94**	**0.94**	0.92	0.77	0.93	0.92	0.93	0.90	0.93	0.76	0.93
10M reads	0.96	0.96	**0.97**	0.94	0.77	0.95	0.93	0.95	0.92	0.95	0.77	0.95
Pearson correlation over genes with non-zero ground truth												
1M reads	**0.96**	**0.96**	0.90	0.93	0.94	**0.96**	0.95	**0.96**	0.93	0.96	0.89	**0.96**
3M reads	**0.98**	**0.98**	0.96	0.95	0.95	0.97	0.97	**0.98**	0.94	**0.98**	0.90	**0.98**
10M reads	**0.98**	**0.98**	**0.98**	0.95	0.96	**0.98**	0.97	**0.98**	0.95	**0.98**	0.91	**0.98**

Supplementary Table 2: Runtime comparison of IsoEM2 with the bootstrapping step of IsoDE, called IsoBoot, on two mouse retina RNA-Seq datasets from [2]. IsoDE2 takes longer than the DE step of IsoDE due to the increased number of performed analyses (genes/isoforms and FPKM/TPM). The time reported for both IsoDE and IsoDE2 is for processing 200 bootstrap samples. Experiments were conducted on a Dell PowerEdge R815 server with quad 2.5GHz 16-core AMD Opteron 6380 processors and 256Gb of RAM.

	Dataset	
Metric	**P0-CE**	**E16-CE**
Raw read pairs	117.4M	99.3M
Mapped read pairs	66.08M	65.83M
IsoBoot runtime (200 bootstraps)	682,074 sec.	499,473 sec
IsoEM2 runtime (200 bootstraps)	2,256 sec.	2,197 sec.
Speedup	**302**×	**227**×
IsoDE run-time (200 bootstraps)	43 sec.	
IsoDE2 run-time (200 bootstraps)	232 sec.	

Supplementary Table 3: Runtime comparison of the HISAT2/IsoEM2 piped commands with Kallisto on the P0-CE and E16-CE mouse retina RNA-Seq datasets from [2]. Here, we include the HISAT2 mapping time in the comparison since Kallisto starts from unmapped reads, while IsoEM2 needs alignments. The runtime of IsoEM2 without read mapping is reported in Supplementary Table 2; however, note that HISAT2 and IsoEM2 computations are partly overlapped since we pipe the HISAT2 output directly to the input of IsoEM2. Experiments were conducted on a Dell PowerEdge R815 server with quad 2.5GHz 16-core AMD Opteron 6380 processors and of 256Gb RAM.

Tool	**Dataset**									
	P0-CE					**E16-CE**				
	# threads					**# threads**				
	1	**2**	**4**	**8**	**16**	**1**	**2**	**4**	**8**	**16**
Kallisto	5779	2,865	1,450	696	332	4,493	2,408	1,396	720	428
HISAT2/IsoEM2	27,482	14,891	9,544	5,486	4,639	24,744	13,859	8,366	4,904	3,629

References

[1] Alexander Kanitz, Foivos Gypas, Andreas J Gruber, Andreas R Gruber, Georges Martin, and Mihaela Zavolan. Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data. *Genome Biology*, 16(1):1–26, 2015.

[2] D.K.P. Karunakaran, S. Al Seesi, A.R. Banday, M. Baumgartner, A. Olthof, C. Lemoine, I.I. Mandoiu, and R.N. Kanadia. Network-based bioinformatics analysis of spatio-temporal RNA-Seq data reveals transcriptional programs underpinning normal and aberrant retinal development. *BMC Genomics*, 17(Suppl 5):495:477–492, 2016.

[3] Sahar Al Seesi, Yvette T Tiagueu, Alexander Zelikovsky, and Ion I Măndoiu. Bootstrap-based differential gene expression analysis for RNA-Seq data with and without replicates. *BMC genomics*, 15(Suppl 8):S2, 2014.