Title

ImpulseDE: detection of differentially expressed genes in time series data using impulse models.

Journal

Bioinformatics, 33(5)

ISSN

1367-4803

Authors

Sander, Jil
Schultze, Joachim L
Yosef, Nir

Publication Date

2017-03-01

DOI

10.1093/bioinformatics/btw665

Peer reviewed

Bioinformatics, 33(5), 2017, 757–759
doi: 10.1093/bioinformatics/btw665
Advance Access Publication Date: 2 December 2016
Applications Note

OXFORD

Gene expression

ImpulseDE: detection of differentially expressed genes in time series data using impulse models

Jil Sander[1], Joachim L. Schultze[1,2] and Nir Yosef[3,*]

[1]Genomics and Immunoregulation, LIMES-Institute, University of Bonn, Bonn, 53115, Germany, [2]Single Cell Genomics and Epigenomics Unit at the University of Bonn and the German Center for Neurodegenerative Diseases, Bonn, Germany and [3]Electrical Engineering and Computer Science, Center for Computational Biology, University of California Berkeley, Berkeley, CA 94720-1776, USA

*To whom correspondence should be addressed.
Associate Editor: Ziv Bar-Joseph

Received on May 11, 2016; revised on September 17, 2016; editorial decision on October 13, 2016; accepted on October 18, 2016

Abstract

Summary: Perturbations in the environment lead to distinctive gene expression changes within a cell. Observed over time, those variations can be characterized by single impulse-like progression patterns. *ImpulseDE* is an R package suited to capture these patterns in high throughput time series datasets. By fitting a representative impulse model to each gene, it reports differentially expressed genes across time points from a single or between two time courses from two experiments. To optimize running time, the code uses clustering and multi-threading. By applying *ImpulseDE*, we demonstrate its power to represent underlying biology of gene expression in microarray and RNA-Seq data.
Availability and Implementation: *ImpulseDE* is available on Bioconductor (https://bioconductor.org/packages/ImpulseDE/).
Contact: niryosef@berkeley.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

When cells are challenged with a certain stimulus, a typical form of transcriptional response is a single impulse-like progress, where the expression of genes goes through an initial change (either rises or drops) and then settles at a second steady-state level (Chechik and Koller, 2009; Yosef and Regev, 2011). This pattern is different from other classes of temporal responses (Bar-Joseph *et al.*, 2012), where genes for example respond in an oscillating modality as classically observed during cell cycle (Yosef and Regev, 2011). Whereas specific methods were introduced for cell cycle data based on Fourier transformations (Kim *et al.*, 2013; Murthy and Hua, 2004), a particular parametric model was developed that captures an impulse-like behavior as a continuous function with six free parameters (Chechik and Koller, 2009). Separately, a method for differential expression analysis termed EDGE was proposed that uses continuous representations of time course data rather than expression levels directly (Storey *et al.*, 2005). By incorporating the more realistic impulse model into a significance analysis process similar to EDGE, we have previously identified genes crucial to T cell function, comparing two time series of gene expression levels (Yosef *et al.*, 2013). Here, we present *ImpulseDE* – a software tool implemented as an R package that conducts this analysis and generalizes upon it. It executes comparative analysis of two time courses or differential expression analysis across time (single time course). Using the impulse model fits based on the available data, it also allows to impute values for unmeasured time points. Additionally, in contrast to other methods being specifically designed for example for RNA-Seq data (Äijö *et al.*, 2014), *ImpulseDE* can be applied on any kind of high throughput gene expression data. We demonstrate this by systematically comparing differential expression analysis results (Soneson and Delorenzi, 2013) to the ones obtained by other approaches and by highlighting canonical genes and functional outcomes being identified by our method. Furthermore, *ImpulseDE* is based on a novel efficient implementation, leading to substantial reduction of running time.

2 Tool description

The *ImpulseDE* pipeline consists of a five-step workflow (Fig. 1A), explained in detail in the supplementary information. The input is

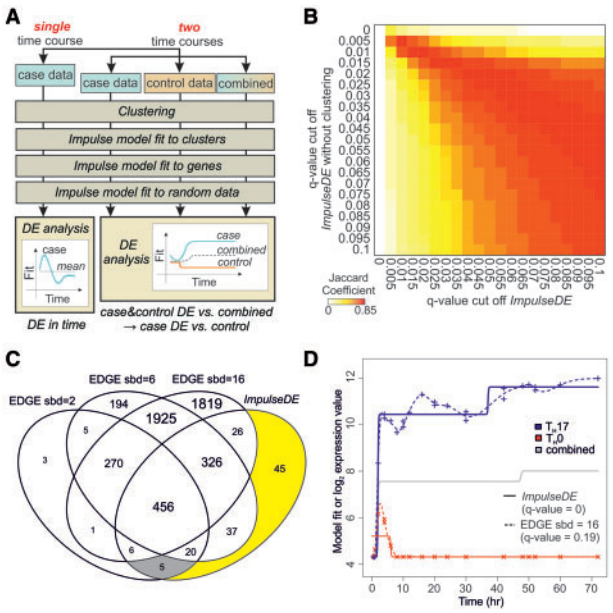

Fig. 1. Performance of *ImpulseDE*. (**A**) Analysis workflow. (**B**) Heatmap of overlapping differentially expressed (DE) genes between *ImpulseDE* applied with and without clustering for *q*-values ranging from 0 to 0.1. Overlaps were determined using the Jaccard-Coefficient. Colors range from white (no) to red (large overlap). (**C**) Venn Diagram showing the qualitative partitioning of genes into distinct groups of overlapping probesets (identified as DE using a *q*-value cutoff of 0.01) between four approaches. (**D**) Impulse model (thick) and EDGE based natural cubic spline (thin dashed lines) fits for gene Il21. The latter is based on a spline basis dimension (sbd) of 16. Expression values and model fits are on $\log_2$-scale. Combined data refers to the union of the T_H17 and T_H0 datasets, where the class affiliations (case or control) were ignored

either one time course (differential behavior over time) or two (comparative analysis) with at least six time points. The genes are first grouped into a limited number of clusters (step 1), and then the parameters of the impulse model are fit to the mean expression profile of each cluster (optimization problem, step 2). The three best sets of parameters (determined by minimizing the sum of squared error; SSE), are finally used as starting points to fit the model to each gene separately (step 3). To determine the significance of rejecting the null hypothesis, random sampling (Storey *et al.*, 2005) and the bootstrap (Efron and Tibshirani, 1994) are used (step 4). The resulting *P*-values are FDR-corrected (*q*-value) to account for multiple testing (Benjamini and Hochberg, 1995) enabling the identification of significantly differentially expressed genes (step 5). Additionally, our implementation makes use of multi-threading to further reduce running time.

3 Case study

As a proof of principle, *ImpulseDE* was applied to a microarray dataset (GSE43955) of T_H0 (control) and T_H17 (case) T cells observed at 18 different time points (0.5–72 h), focusing on a pre-filtered list (supplementary information) of 7526 probesets (Yosef *et al.*, 2013). The complete run took about 5 hours performed on a Desktop Computer (Intel(R) Core(TM) i7 Processor with 32 GB RAM) utilizing 6 cores and default options otherwise. We identified 921 probesets as differentially expressed (DE) between T_H17 and T_H0 using a *q*-value cutoff of 0.01. We then evaluated the performance of *ImpulseDE* in multiple contexts. First, we compared to the results obtained from the original much slower version missing the clustering step, where every gene is fit separately instead using the same number of iterations. While the running time was about 7 fold longer (supplementary information), the improvements of fitting accuracies were marginal (Supplementary Fig. S1A–C), justifying the value of the much faster clustering procedure. In line with this, decreased *q*-value cut offs compensated the discrepancies in numbers of overlapping DE genes (Fig. 1B).

Second, we applied EDGE (https://github.com/jdstorey/edge; Storey *et al.*, 2005) as an alternative method on the microarray data using 2, 6 and 16 as three different spline basis dimensions (spd), the optimal discovery procedure (Storey, 2007) and all default options otherwise. Based on the same *q*-value cutoff of 0.01, we found an overlap of 456 probesets identified as DE by all three EDGE approaches as well as *ImpulseDE* (Fig. 1C). However, we also observed several differences. 45 probesets (marked in yellow in Fig. 1C) were exclusively called as DE by *ImpulseDE*, including the canonical T_H17 genes Il21 (Figs. 1D, S2A) and Socs3 (Supplementary Fig. S2B) (Ciofani *et al.*, 2012; Yosef *et al.*, 2013). Furthermore, the classical T_H17 marker genes Il17a (Supplementary Fig. S2C) and Il9 (Supplementary Fig. S2D) (Ciofani *et al.*, 2012; Yosef *et al.*, 2013) were detected by both *ImpulseDE* and EDGE spd = 2 (marked in grey in Fig. 1C), where the latter however clearly seemed to underfit the data. In line with this and as expected from the varying numbers of parameters, we observed that the spd-16 model resulted in a better fit to the time course data compared to *ImpulseDE*, and that the latter provided a better fit than the lower-dimensionality spline functions (Supplementary Fig. S3A, B). Clearly, the accuracy of fits to the time course data might imply overfitting (Fig. 1D), as may be discerned by the non-smooth profiles of the spd-16 model. To test that, we compared the ability of all methods to impute expression data for missing measurements. We used *ImpulseDE* and according spline models as used in EDGE to fit time courses with missing data, each time hiding a segment of four consecutive time points. We then measured the model's accuracy as the SSE between the imputed values and the true but hidden measurements for each block of time points separately. In terms of imputation accuracies, *ImpulseDE* performed best for 9 out of 13 time point blocks compared to all three EDGE methods (Supplementary Fig. S3C, D), supporting its ability to model underlying dynamic behavior.

Additionally, we ranked the genes by their normalized dispersion over time (Fano Factor) across the T_H17 data (considered as a single time course by ignoring T_H0). Although the ranking was overall consistent with *ImpulseDE* and the EDGE approaches (Supplementary Fig. S4A–D), the simplistic time-agnostic method missed genes with an obvious temporal trend that is captured by *ImpulseDE* and EDGE (Supplementary Fig. S4E).

As another proof of principle, we applied *ImpulseDE* on a RNA-Seq dataset of primary dendritic cells (Jovanovic *et al.*, 2015) stimulated with LPS (case) or with a mock stimulus (control) covering 6 time points (GSE59784). The same options and cutoffs were used as mentioned above. Of 3,147 TPM (transcripts per million) normalized (Li *et al.*, 2010) and filtered genes (Jovanovic *et al.*, 2015), we identified 1499 to be DE between LPS and mock stimulation. Among those we found canonical LPS response genes (Shalek *et al.*, 2014; Torri *et al.*, 2010), including Nfkb1, Stat5a, Cd38, Cd40, Tap1 and Map2k1 (Supplementary Fig. S5A). Additionally, Gene Ontology Enrichment Analysis (GOEA) based on the up- (395) and down-regulated (1104) genes identified by *ImpulseDE* confirmed functions typically associated with LPS stimulation (Supplementary Fig. S5B), including for example the response to lipopolysaccharide (LPS) or bacterium, as well as induced immune and inflammatory responses (Granucci *et al.*, 1999).

The representation of temporal expression profiles as continuous impulse functions has already proven useful to describe the kinetics of down-stream processes such as protein expression (Yosef *et al.*, 2013) and RNA degradation (Rabani *et al.*, 2014). Importantly, *ImpulseDE* can be applied on any type of temporal data exhibiting an impulse-like behavior, including for example changes in chromatin accessibility and histone marks (Lara-Astiaso *et al.*, 2014, Weiner *et al.*, 2015). Therefore, *ImpulseDE* provides differential expression analysis, imputation and modeling for a broad range of high throughput datasets.

Funding

NY was supported by grants U01 MH105979 and U01 HG007910 from the National Institute of Health (NIH). The work was supported by SFB704 to JLS.

Conflict of Interest: none declared.

References

Äjiö,T. *et al.* (2014) Methods for time series analysis of RNA-seq data with application to human Th17 cell differentiation. *Bioinformatics*, **30**, 113–120.

Bar-Joseph,Z. *et al.* (2012) Studying and modelling dynamic biological processes using time-series gene expression data. *Nat. Rev. Genet.*, **13**, 552–564.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **57**, 289–300.

Chechik,G. and Koller,D. (2009) Timing of gene expression responses to environmental changes. *J. Comput. Biol.*, **16**, 279–290.

Ciofani,M. *et al.* (2012) A validated regulatory network for Th17 cell specification. *Cell*, **151**, 289–303.

Efron,B. and Tibshirani,R.J. (1994) An Introduction to the Bootstrap. *Chapman & Hall/CRC*, Monographs on Statistics & Applied Probability 57.

Granucci,F. *et al.* (1999) Early events in dendritic cell maturation induced by LPS. *Microb. Infect.*, **1**, 1079–1084.

Jovanovic,M. *et al.* (2015) Dynamic profiling of the protein life cycle in response to pathogens. *Science*, **347**, 1259038.

Kim,J. *et al.* (2013) A method to identify differential expression profiles of time-course gene data with Fourier transformation. *BMC Bioinformatics*, **14**, 310.

Lara-Astiaso,D. *et al.* (2014) Chromatin state dynamics during blood formation. *Science*, **345**, 943–949.

Li,B. *et al.* (2010) RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, **26**, 493–500.

Lim,H.W. *et al.* (2008) Human Th17 cells share major trafficking receptors with both polarized effector T cells and FOXP3+ regulatory T cells. *J. Immunol.*, **180**, 122–129.

Murthy,K.R.K. and Hua,L.J. (2004). Improved Fourier transform method for unsupervised cell-cycle regulated gene prediction. In: *Proc. IEEE Comp. Sys. Bioinformatics Conf. Proceeding*, pp. 194–203.

Rabani,M. *et al.* (2014) High-resolution sequencing and modeling identifies distinct dynamic RNA regulatory strategies. *Cell*, **159**, 1698–1710.

Shalek,A.K. *et al.* (2014) Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature*, **510**, 363–369.

Soneson,C. and Delorenzi,M. (2013) A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*, **14**, 91.

Storey,J.D. (2007) The optimal discovery procedure: a new approach to simultaneous significance testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **69**, 347–368.

Storey,J.D. *et al.* (2005) Significance analysis of time course microarray experiments. *Proc. Natl. Acad. Sci. U. S. A.*, **102**, 12837–12841.

Torri,A. *et al.* (2010) Gene expression profiles identify inflammatory signatures in dendritic cells. *Plos One*, **5**, e9404.

Weiner,A. *et al.* (2015) High-resolution chromatin dynamics during a yeast stress response. *Mol. Cell*, **58**, 371–386.

Yosef,N. and Regev,A. (2011) Impulse control: temporal dynamics in gene transcription. *Cell*, **144**, 886–896.

Yosef,N. *et al.* (2013) Dynamic regulatory network controlling TH17 cell differentiation. *Nature*, **496**, 461–468.