Gene expression

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

Jil Sander¹, Joachim L. Schultze¹,² and Nir Yosef³,*

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

*To whom correspondence should be addressed.

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 (Âijo 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
As a proof of principle, ImpulseDE was applied to a microarray dataset (GSE43955) of T_{17}0 (control) and T_{17}17 (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_{17}17 and T_{17}0 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 436 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_{17}17 genes Il21 (Figs. 1D, S2A) and Socs3 (Supplementary Fig. S2B) (Ciofani et al., 2012; Yosef et al., 2013). Furthermore, the classical T_{17}17 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_{17}17 data (considered as a single time course by ignoring T_{17}0). 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. S5).

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 Nlk1b1, 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 downstream 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**