Supplementary Information: Impulse model-based differential expression analysis of time course sequencing data

David S. Fischer[a,b,d], Fabian Theis[a,c], and Nir Yosef[d,e]

[a]Institute of Computational Biology, Helmholtz Centre Munich
[b]TUM School of Life Sciences Weihenstephan, Technical University of Munich
[c]Department of Mathematics, Technical University of Munich
[d]Department of Electrical Engineering & Copmuter Science and Center for Computational Biology, University of California, Berkeley CA
[e]Ragon Institute of MGH, MIT& Harvard, Cambridge MA

March 3, 2017

1 Supplementary Notes

1.1 Method comparison on the hESC (Chu) data set

ImpulseDE2 and DESeq2 give globally similar results on the hESC (Chu) data set (Fig. 1A,B). As expected, the q-values for differential expression assigned by ImpulseDE2 are largely upper bounded by the p-values of DESeq2 (Fig. 1B) as they have the same statistical testing power: Both use six parameters for the mean expression model on this data set with six sampled time points. The genes which receive much lower q-values for differential expression by DESeq2 compared to ImpulseDE2 have potentially multi-modal trajectories (Fig. SI11) which are penalized by the impulse model for the expression trajectory in ImpulseDE2.

Gene set enrichment analysis suggests, that these genes which are only called differentially expressed by DESeq2 but not by ImpulseDE2 are indeed not related to the process as this gene set is not enriched in any GO biological process or GO hallmark term (Supplementary Data 3).

In summary, ImpulseDE2 behaves as expected and does not have an advantage in statistical testing power as only six time points were sampled. However, ImpulseDE2 guards against potentially multi-modal trajectories which may represent noise.

The scatter plot of the q-values for differential expression of ImpulseDE2 against edge (Fig. 1C) suggests that there is a group of genes which tend to have low expression mean which are preferentially labeled differentially expressed by edge and the remaining genes which are preferentially labeled differentially expressed by ImpulseDE2 (Fig. SI13, SI12). The gene set with high expression means which has a high proportion of genes only labeled differentially expressed by ImpulseDE2 and not by edge contains trajectories which represent visually convincing cases of differential expression. Genes labeled differentially expressed by edge and not by ImpulseDE2 do contain trajectories which are visually not convincing: Some genes with very low expression counts receive very low q-values by edge (Fig. SI13).

In summary, ImpulseDE2 produces more convincing results than edge on the hESC (Chu) data set.

ImpulseDE2 and ImpulseDE give very different results on this data set (Fig. 5D). Gene set enrichment analysis with the MSigDB hallmark set suggests that the set of genes only called differentially expressed by ImpulseDE2 and not by ImpulseDE is enriched in Myc targets (Supplementary Data 3). Differential Myc activity is to be expected in the context of embryonic stem cell development.

1.2 Method comparison on the erythroid chromatin (Lara-Astiaso) data set

A comparison of the different methods on the erythroid chromatin (Lara-Astiaso) data set gave similar results. The data set has samples from seven cell types (which are treated as time points). Therefore, ImpulseDE2 is not expected to have much higher statistical testing power than DESeq2. Indeed, the q-values assigned by DESeq2 are approximately an upper bound for the q-values assigned by ImpulseDE2 (Fig. 2A,B). There are multiple genes in the lower right half of the p-value scatter plot of Fig. 2 which receive much lower p-values by DESeq2 than by ImpulseDE2. Again, we could visually confirm that multiple of these genes to belong to the potentially multi-modal class but we also observe some apparently non-optimal impulse model fits to uni-modal patterns (Fig. SI15, SI14).

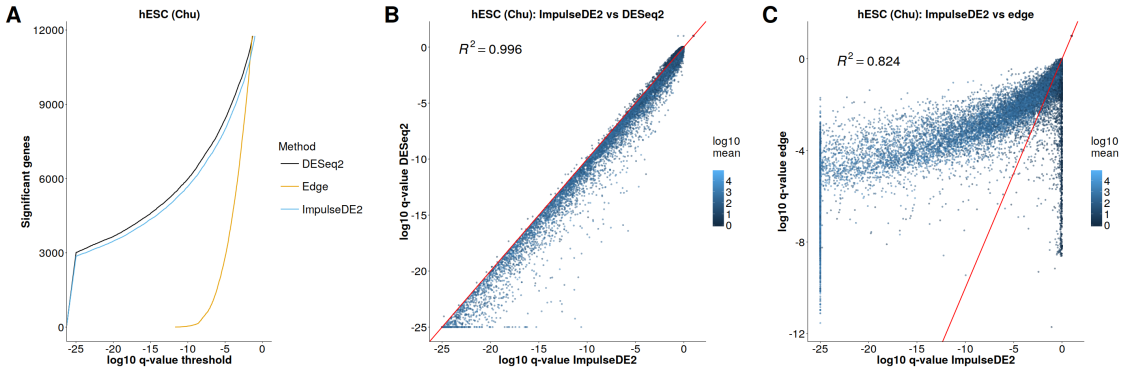

Figure 1: **ImpulseDE2 performance on Chu et al. RNA-seq. A** Number of significantly differentially expressed genes as a function of the significance threshold. **B** Correlation plot of the inferred differential expression Benjamini-Hochberg corrected p-values for all genes between ImpulseDE2 and DESeq2. **C** Correlation plot of the inferred differential expression Benjamini-Hochberg corrected p-values for all genes between ImpulseDE2 and edge. R^2 shown are Pearson correlations coefficients.

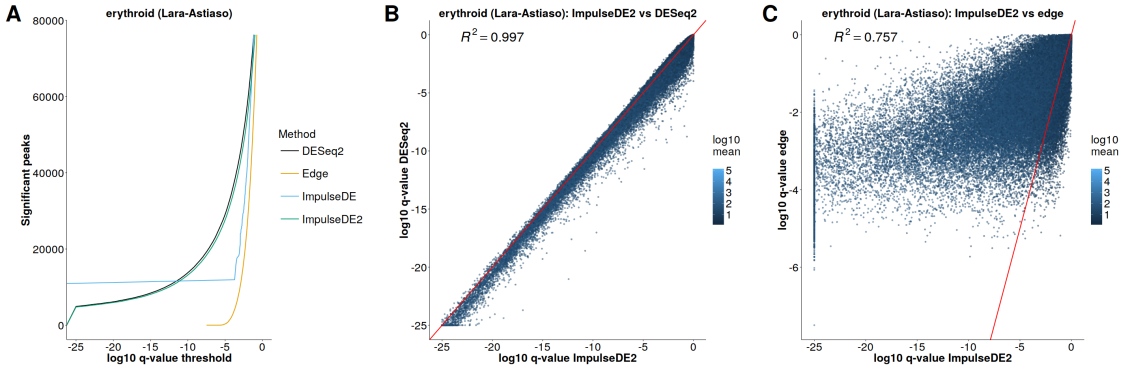

Figure 2: **ImpulseDE2 performance on iChIP hematopoiesis data. A** Number of significantly differentially expressed iChIP peaks as a function of the significance threshold. **B** Correlation plot of the inferred differential expression Benjamini-Hochberg corrected p-values for all iChiP peaks between ImpulseDE2 and DESeq2. **C** Correlation plot of the inferred differential expression Benjamini-Hochberg corrected p-values for all iChiP peaks between ImpulseDE2 and edge. R^2 shown are Pearson correlations coefficients.

Gene set enrichment analysis shows that the gene sets only called differentially expressed by DESeq2 and by ImpulseDE2 respectively both are enriched in MSigDB hallmark terms related to immune system signalling (Supplementary Data 5).

The assigned p-values of ImpulseDE2 and edge also differ strongly on the erythroid chromatin (Lara-Astiaso) data set (Fig. 2A,C). Again, we observe that genes which are only labeled differentially expressed by edge and not by ImpulseDE2 tend to have low mean expression while genes which are only labelled differentially expressed by ImpulseDE2 and not by edge are visually convincing cases of differential expression (Fig. SI16, SI17).

Gene set enrichment analysis supports the hypothesis that the genes selectively labeled differentially expressed are related to the underlying process of hematopoesis: Genes only labeled differentially expressed by ImpulseDE2 and not by edge are enriched in several MSigDB hallmark sets related to hematopesis, such as heme-metabolism and several immune cell signalling sets (Supplementary Data 5).

The assigned p-values of ImpulseDE2 and ImpulseDE also differ strongly (Fig. 5D). Similar to edge, genes which are only labeled differentially expressed by ImpulseDE and not by ImpulseDE2 tend to have low mean expression while genes which are only labeled differentially expressed by ImpulseDE2 and not by ImpulseDE are visually convincing cases of differential expression (Fig. SI18, SI19).

Genes only labeled differentially expressed by ImpulseDE2 and not by ImpulseDE are in enriched several MSigDB hallmark sets related to hematopesis, such as heme-metabolism and several immune cell signalling sets (Supplementary Data 5). Genes only labeled differentially expressed by ImpulseDE and not by ImpulseDE2 are not enriched in MSigDB hallmark sets related to immune cell signalling (Supplementary Data 5).

In summary, ImpulseDE2 gives overall similar results to DESeq2 on the erythroid chromatin (Lara-Astiaso) data set and and gives more convincing differential expression labels than edge and ImpulseDE.

1.3 Proofs: The null models are nested within the alternative models

1.3.1 Case-only differential expression analysis

In the case of differential expression analysis over time within one condition, the null model may take the form of any real valued constant function and the alternative model is any impulse model [eq. 1].

Hypothesis 1: Any real valued constant function lies within parameter space of impulse model f.

Proof Hypothesis 1:

$$\begin{aligned} f(x) &= \frac{1}{h_1}(h_0 - (h_1 - h_0)\frac{1}{1+e^{-\beta(x-t_1)}}) \\ &\quad * (h_2 - (h_1 - h_2)\frac{1}{1+e^{\beta(x-t_2)}}) \\ &\overset{h_0=h_1=h_2}{=} h_0 \end{aligned} \tag{1}$$

1.3.2 Case-control differential expression analysis

In case-control differential expression analysis, the null model is an impulse fit to the combined data set with all samples and the alternative model are separate impulse fits to the sample sets of case and control condition.

Hypothesis 2: For any one impulse model f_0 there exists a parameterization of two impulse models f_a and f_b such that $f_a = f_0$ and $f_b = f_0$.

Proof Hypothesis 2: If both impulse models have the same parameters ($f_a = f_b$), then $f_a = f_0$ and $f_b = f_0$.

2 Supplementary Methods

2.1 Example batch correction

We explain batch correction in the following using an example:

sample	time	condition	patient	processing_batch
s1	0	case	A	batch1
s2	2	case	A	batch1
s3	4	case	A	batch1
s4	8	case	A	batch2
s5	0	case	B	batch2
s6	2	case	B	batch1
s7	4	case	B	batch1
s8	8	case	B	batch2
s9	0	case	C	batch1
s10	2	case	C	batch1
s11	4	case	C	batch2
s12	8	case	C	batch2

Consider the scenario above in which there are RNA-seq time course samples from three patients (A, B and C) at four time points (0,2,4,12) each which were processed in two batches ($batch1$ and $batch2$). All come from condition case: We are only interested whether a gene changes over time (case-only differential expression analysis). To correct for the confounding variables patient and processing_batch, one can assign batch correction factors to each level ("batch") in each confounding variable. The batch correction factor corresponding to the first batch in each confounding variable (patient: A, processing_batch: $batch1$) receive a batch correction factor of one and the following batches are scaled accordingly. An observation (a gene expression count) from sample s6 would accordingly be corrected with the batch factors of patient B and processing_batch $batch1$:

$$\begin{aligned}\prod_{l}^{L} b_{i,l(j=\text{s1})} &= b_{i,\text{patient(s1)}} * b_{i,\text{processing\_batch(s1)}} \\ &= b_{i,B} * b_{i,batch1}\end{aligned} \tag{2}$$

2.2 Global gene-expression profile heatmaps

All heatmaps show z-scores of all differentially expressed genes. The expression matrix underlying the z-score profiles is the mean of the DESeq2 size factor corrected samples of a gene and a time point. Only the case condition samples were chosen from the LPS (Jovanovic) data set. Differentially expressed genes were selected without any constraints on the expression trajectory with DESeq2. Differentially expressed genes were clustered with K-means based on their z-score profiles. Clusters were ordered by peak time. The following DESeq2 adjusted p-value thresholds were used: eythroid (Lara-Astasio) [2] (q=1e-5), LPS (Jovanovic) [1] (q=1e-2), myeloid (Sykes) [3] (distance: z-scores, q=1e-5), hESC (Chu) [4] (q=1e-5), estrogen (Baran-Gale) [5] (q=1e-5), Plasmodium (Broadbent) [6] (shown are all lncRNAs in the data set).

2.3 Gene set enrichment analysis

We performed gene set enrichment analysis against gene sets from MSigDB [7]: The H hallmark gene set [8], the C2 curated gene set collection [7], the C3.tft transcription factor target sets from TRANSFAC 7.4 [9], the C5.mf GO molecular function sets [10], the C5.bp GO biological process sets [10] and the C7 immunological signatures gene sets [11]. The gene sets are deposited in MSigDB as HGNC identifiers. We mapped all identifiers from the the data sets to HGNC identifiers with biomaRt [12].

We define genes as differentially called it they do not received the same differential expression label (significant: yes or no) by both ImpulseDE2 and by a reference method . Significance is evaluated based on false-discovery rate corrected p-values at a common threshold of $1e-2$ (or $1e-5$ for erythroid chromatin (Lara-Astiaso) with more than 100,000 peaks). We used the the Benjamini-Hochberg false-discovery rate correction for all methods to make the results comparable.

We tested over-representation of these differentially called gene sets in the MSigDB data sets with a hypergeometric test and manually assessed the meaning of the over-represented gene sets at a q-value of 0.05 for the individual experimental settings. The full results with all significantly over-represented gene

sets can be found in the Supplmentary Tables 1,2,3,4,5.

We mapped the gene identifiers to HGNC as follows:
We mapped the ensemble transcript identifiers of the LPS (Jovanonvic) data set to homologous HGNC identifiers (human) (biomaRt mmusculus_gene_ensembl: ensembl_transcript_id to hsapiens_homolog_associated_gene_name).
We mapped the Ensemble gene identifiers (mouse) of the myeloid differentiation RNA-seq data set (Sykes et al. [3]) directly to HGNC identifiers (human) (biomaRt mmusculus_gene_ensembl: mgi_symbol to hsapiens_homolog_associated_gene_name).
The published identifiers of the human embryonal stem cell RNA-seq data set (differentiation of human embryonal stem cells to definite endoderm, Chu et al. [4]) are HGNC (human) and we used these directly.
We mapped iChIP peaks (mouse) of the erythroid lineage iChIP peak data set (Lara-Astasio et al. [2]) to MGI identifiers (mouse) with GREAT [13] with the following default settings: "GREAT version 3.0.0 Species assembly: mm9 Association rule: Basal+extension: 5000 bp upstream, 1000 bp downstream, 1000000 bp max extension, curated regulatory domains included". Then, we mapped the MGI identifiers (mouse) to Ensembl gene identifiers (mouse) (biomaRt mmusculus_gene_ensembl: mgi_symbol to ensembl_gene_id). Then, we mapped the Ensembl gene identifiers (mouse) to homologous HGNC identifiers (human) (biomaRt mmusculus_gene_ensembl: ensembl_gene_id to hsapiens_homolog_associated_gene_name).

2.4 Count matrix generation

myeloid (Sykes) We used the published expected count matrix for all analysis.
hESC (Chu) We used the published expected count matrix for all analysis.
erythroid chromatin (Lara-Astiaso) We aligned reads from fastq files with bowtie2 [14] against GRCm38.p3. Peaks were called with MACS2 [15] (callpeak –format BAM -g mm –broad, no input control published) on the merged alignments of all samples at each time point. The seven cell states of the erythroid lineage contained in the data set are in developmental order [2] which corresponds to the ordering discussed in the paper: 1 - Long Term Hematopoietic Stem Cell (LT-HSC) (1 sample), 2 - Short Term Hematopoietic Stem Cell (ST-HSC) (3 samples), 3 - Multipotent Progenitor (MPP) (2 samples), 4 - Common Myeloid Progenitor (CMP) (3 samples), 5 - Megakaryocitic erythroyd progenitor (MEP) (1 sample), 6 - Erythrocytes A (Ery A) (2 samples), 7 - Erythrocytes B (Ery B) (2 samples). The peak files were then merged across time to give a background set of peaks. A count matrix was created based on the number of overlapping reads within each sample with each background peak.
LPS (Jovanovic) We created an expected count matrix with kallisto [16] based on a mm10 index build with kmers of length 29 base pairs.
estrogen (Baran-Gale) We used the published expected count matrix for the heatmap.
Plasmodium (Broadbent) We used the published expected FPKM matrix for the heatmap.

All count matrices were normalized by DESeq2 size factors and log transformed for ImpulseDE and edge.

2.5 Reference method parameters

The model formula for DESeq2 were:
Case-only single batch (Time versus 1),
Case-only multiple batches (Time + Batch versus Batch),
Case-control single batches (Condition + Condition:Time versus Time)
Case-control multiple batches standard model with one batch factor per batch (Condition + Condition:Time + Batch versus Time + Batch).
Case-control multiple batches ImpulseDE2-like model with one batch factor per batch and condition which only makes a difference if batches are present in both conditions (Condition + Condition:Time + Condition:Batch versus Time + Batch). We use this ExtraBatch model in case-control analysis of the LPS (Jovanovic) data set: The interaction term Condition:Batch implies that for each gene, one constant batch correction factor is fit to the sample groups A_case, A_ctrl, B_case and B_ctrl (Fig. 4). Note that in the standard batch correction setting, a constant batch correction factor would only be fit to the sample group A and B.
DESeq2 was run on expected count matrices.

The model formula for edge were:
Case-only single batch (ns(Time, df=4, intercept=FALSE) versus 1),
Case-only multiple batches (ns(Time, df=4, intercept=FALSE) + Batch versus Batch),
Case-control single batch (Condition + ns(Time, df=4, intercept=FALSE) + (Condition):ns(Time, df=4, intercept=FALSE) versus ns(Time, df=4, intercept=FALSE)),

Case-control multiple batches (Condition + ns(Time, df=4, intercept=FALSE) + (Condition):ns(Time, df=4, intercept=FALSE) + Batch versus ns(Time, df=4, intercept=FALSE) + Batch).
Edge was run on DESeq2 size factor normalized and log transformed data.
Where less $n < 5$ time points were modeled in the simulations, $df = n - 1$ was chosen as the degrees of freedom of the natural cubic splines used in edge.

2.6 Data simulation

Simulations were performed with the simulateDataSetImpulseDE2() function of ImpulseDE2.

We first simulated hidden expression trajectories, then imposed library depth and batch effects and then imposed negative binomial noise.

We drew the constant trajectory expression level, the initial amplitudes of the sigmoid and impulse trajectories and initial and final expression level of the linear model from a uniform distribution on the interval [1e-5,1000]. We computed the remaining amplitudes of the sigmoid and impulse model as the product of the initial amplitude and a scaling factor drawn from a normal distribution with mean one and standard deviation one. We simulated random trajectories by drawing a constant baseline expression trajectory from a uniform distribution on [1e-5,1000] and scaled each observation with a factor drawn from a normal distribution with mean one and the standard deviation as states in the plot. If case-control analysis was performed, the initial expression level and the function form were kept for samples from case and control condition but the final expression levels/amplitudes were generated separately for each condition. We drew size factor (library depth) for each sample from a normal distribution with mean one and standard deviation 0.1 and bounded them by 0.1 and 10. If confounders were used, we drew batch factor (library depth) for each batch and gene from a normal distribution with mean one and standard deviation as indicated in the plots and bounded them by 0.1 and 10. We multiplied all hidden expression trajectory values by corresponding size and batch factors. We lower bounded all expression levels at 1e-5.

We drew one dispersion parameter for each gene from a normal distribution with mean one and standard deviation 0.1 and multiplied them by 10 and lower-bounded them at 1. We drew observed count data from a negative binomial distribution with mean as the scaled hidden expression level and dispersion parameter as before.

References

[1] Jovanovic M, et al. (2015) Immunogenetics. Dynamic profiling of the protein life cycle in response to pathogens. *Science (New York, N.Y.)* 347(6226):1259038.

[2] Lara-Astiaso D, et al. (2014) Chromatin state dynamics during blood formation. *Science* 345(6199):943–9.

[3] Sykes DB, et al. (2016) Inhibition of Dihydroorotate Dehydrogenase Overcomes Differentiation Blockade in Acute Myeloid Leukemia. *Cell* pp. 171–186.

[4] Chu LF, et al. (2016) Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome Biology* 17(1):173.

[5] Baran-gale J, Purvis JE, Sethupathy P (2016) An integrative transcriptomics approach identifies miR-503 as a candidate master regulator of the estrogen response in MCF-7 breast cancer cells. pp. 1–12.

[6] Broadbent KM, et al. (2015) Strand-specific RNA sequencing in Plasmodium falciparum malaria identifies developmentally regulated long non-coding RNA and circular RNA. *BMC genomics* 16(1):454.

[7] Subramanian A, et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* 102(43):15545–15550.

[8] Liberzon A, et al. (2015) The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Systems* 1(6):417–425.

[9] Matys V, et al. (2003) TRANSFAC®: Transcriptional regulation, from patterns to profiles. *Nucleic Acids Research* 31(1):374–378.

[10] Blake JA, et al. (2015) Gene ontology consortium: Going forward. *Nucleic Acids Research* 43(D1):D1049–D1056.

[11] Godec J, et al. (2016) Compendium of Immune Signatures Identifies Conserved and Species-Specific Biology in Response to Inflammation. *Immunity* 44(1):194–206.

[12] Durinck S, et al. (2005) BioMart and Bioconductor: A powerful link between biological databases and microarray data analysis. *Bioinformatics* 21(16):3439–3440.

[13] McLean CY, et al. (2010) GREAT improves functional interpretation of cis-regulatory regions. *Nature biotechnology* 28(5):495–501.

[14] Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9(4):357–359.

[15] Feng J, Liu T, Qin B, Zhang Y, Liu XS (2012) Identifying ChIP-seq enrichment using MACS. *Nature Protocols* 7(9):1728–1740.

[16] Bray NL, Pimentel H, Melsted P, Pachter L (2016) Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology* 34(5):525–527.