

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

David S. Fischer, Fabian J. Theis, Nir Yosef

June 20, 2018

Contents

1	The log-likelihood ratio test of ImpulseDE2 produces a chi-squared distributed deviance	1
1.1	Proofs: The null models are nested within the alternative models	1
1.1.1	Case-only differential expression analysis	1
1.1.2	Case-control differential expression analysis	3
2	ImpulseDE2 algorithm	3
2.1	Hyper-parameter estimation	4
2.2	Model Fitting	4
2.3	Differential expression analysis	5
3	Global gene-expression profile heatmaps	5
4	Gene set enrichment analysis	6
5	Count matrix generation	6
5.1	myeloid (Sykes)	6
5.2	hESC (Chu)	6
5.3	erythroid chromatin (Lara-Astiaso)	6
5.4	LPS (Jovanovic)	7
5.5	estrogen (Baran-Gale)	7
5.6	Plasmodium (Broadbent)	7
5.7	Drosophila (Graveley)	7
6	Reference method parameters	9
6.1	Standard settings	9
6.2	DESeq2 settings adapted to sequencing time course data	10
7	Data simulation	10
7.1	Method benchmaring	10
7.1.1	Simulation of Fig. 2a: Varying the number of time points sampled	10
7.1.2	Simulation of Fig. 2b: Varying the standard deviation of a randomly changing expression profile	10
7.2	Distribution of the deviance computed by ImpulseDE2	10

1 The log-likelihood ratio test of ImpulseDE2 produces a chi-squared distributed deviance

For the log-likelihood ratio test to yield a chi-squared distributed deviance, the full and reduced model have to be nested which we show here analytically. We also provide empirical evidence based that the deviance is indeed chi-square distributed on data simulated under the null model 7.

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

1.1.1 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.

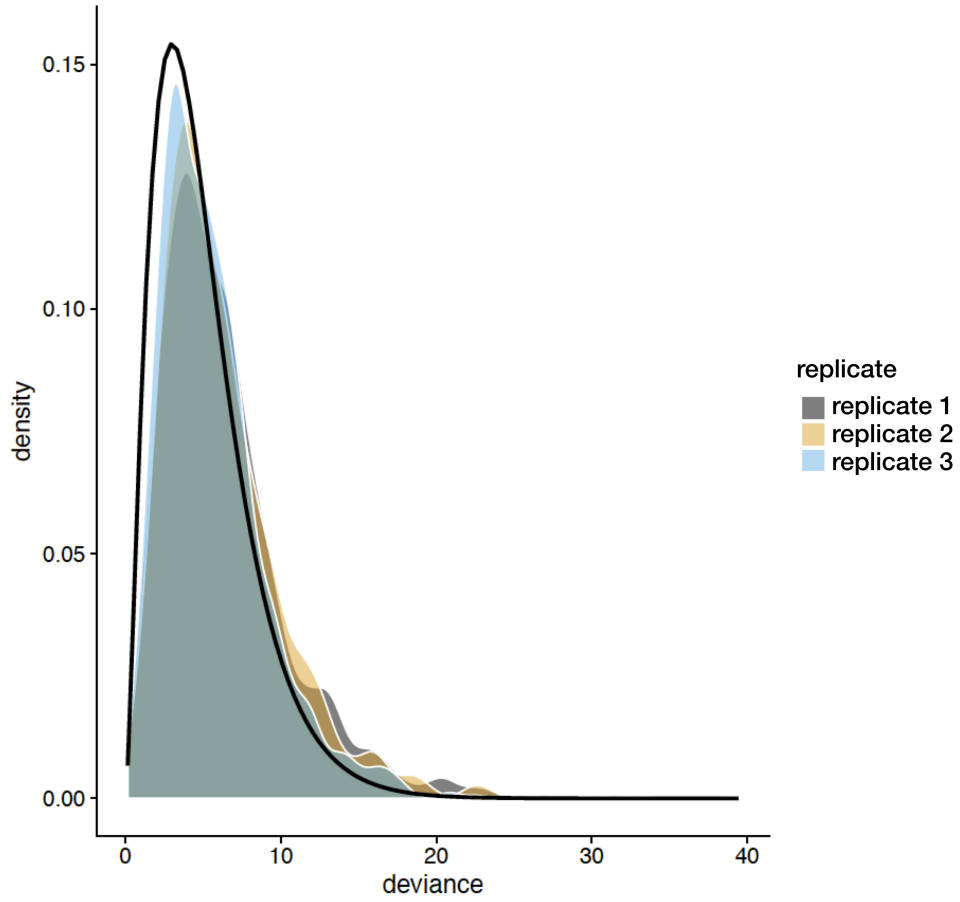


Figure 7: **Distribution of deviance of ImpulseDE2 on data simulated under null model with reference chi-square distribution.** The three shaded densities correspond to three separate simulations, the black line is the chi-square distribution with the corresponding number of degrees of freedom. We describe this simulation in the Supp. methods sec. 7.2.

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

Proof Hypothesis 1:

$$f(x) = \frac{1}{h_1} (h_0 - (h_1 - h_0) \frac{1}{1 + e^{-\beta(x-t_1)}}) * (h_2 - (h_1 - h_2) \frac{1}{1 + e^{\beta(x-t_2)}}) \quad (1)$$

$h_0 = \underline{h_1 = h_2} \quad h_0$

1.1.2 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 ImpulseDE2 algorithm

Input Matrix of count data (N genes \times J samples) and table specifying meta data of samples.

1. **Hyper-parameter estimation:** Run DESeq2 and save dispersion parameters. Catch high variance dispersion parameter outliers. Compute size factors.
2. **Model fitting:** If performing case-control differential expression analysis: Split data set into three data sets with only samples contained in case condition ("case"), only samples contained in control condition ("control") and with all samples ("combined") and iterate over data sets. Parallelize over genes within a data set.
 - a) **Constant model:** Optimize constant model parameters based negative binomial likelihood with the Broyden-Fletcher-Goldfarb-Shanno algorithm (BFGS) (up to I iterations).
 - b) **Sigmoid model:** Optional, only fit if transients trajectories are to be identified. **Initialization:** Initialize two sigmoid models for each gene based on a increase and a decrease model. **Optimization:** Optimize sigmoid model parameters of both initializations based on negative binomial likelihood with BFGS with the dispersion parameter estimated by DESeq2 (up to I iterations). **Fit selection:** Keep sigmoid model fit with highest log-likelihood.
 - c) **Impulse model:** **Initialization:** Initialize two impulse models for each gene based on a peak and a valley model (explained below in Model fitting). **Optimization:** Optimize impulse model parameters of both initializations based on negative binomial likelihood with BFGS with the dispersion parameter estimated by DESeq2 (up to I iterations). **Fit selection:** Keep impulse model fit with highest log-likelihood.
3. **Differential expression analysis:** Perform log-likelihood ratio test. Compute p-values from χ^2 -distributed deviance and perform Benjamin-Hochberg correction.

Output Benjamini-Hochberg corrected p-values for differential expression for each gene. If selected, a classification of differentially expressed genes into transiently and monotonously changing expression trajectories.

The algorithmic complexity the run time dominating impulse fitting step of ImpulseDE2 is $\mathcal{O}(I * N * J)$. The run time of ImpulseDE2 is approximately linear in the number of genes and samples. Therefore, we expect ImpulseDE2 to scale well to large data sets. The number of numerical fitting iterations necessary to reach convergence may however differ between data sets and increase with the number of samples leading to above linear complexity in the number of samples. ImpulseDE2 is parallelized over genes in the run-time dominating model fitting steps (constant, impulse and sigmoid model).

2.1 Hyper-parameter estimation

Hyper-parameter estimation precedes the mean model fitting step in ImpulseDE2. During hyper-parameter estimation, sample-specific normalization factors (size factors) and gene-specific dispersion parameters are estimated. The estimation of the mean model (impulse, sigmoid and constant) is conditioned on the hyper-parameters.

Dispersion parameters Dispersion parameters have been previously regularized by smoothing gene-wise point estimators of genes according to a genome-wide trend in an empirical Bayes scheme [1]. Such a smoothing breaks the independence of gene-wise expression models. To maintain independence between the gene-wise expression models, we estimate the dispersion parameters as hyper-parameters prior to fitting the mean model. The dispersion parameters are estimated with DESeq2 [16] treating time as a categorical variable. ImpulseDE2 catches high-dispersion outliers in genes which contain zero count observations which are not regularized through empirical Bayes smoothing by DESeq2 and replaces the raw dispersion estimate with the a-posteriori estimate from DESeq2. This outlier handling results in much lower regularized variance estimates of the outlier genes and does not affect the remaining genes.

Size factors The normalization factor \tilde{s}_j of sample j is the median of the ratios of observed counts within this sample to the geometric means κ_i of the genes i [16]. Size factors are computed on the subset of genes which do not contain zero observations I_0 .

$$\forall i \in I_0 : \kappa_i = \left(\prod_{j=1}^J x_{i,j} \right)^{\frac{1}{n_{i,\cdot}}}, \quad (2)$$

where $n_{i,\cdot}$ is the number of observations for gene i .

$$\tilde{s}_j = \text{median}(\{ \frac{x_{i,j}}{\kappa_i} \}_{i \in I_0}) \quad (3)$$

The median ratio of observed counts to the geometric mean over all genes I has been previously used as a sample normalization factor because it is less sensitive to regions with high mean and high variance count distributions than the sequencing depth [16].

2.2 Model Fitting

Constant model The constant mean parameter is estimated as a maximum likelihood estimator based on a negative binomial log-likelihood with the dispersion factor as the hyper-parameter inferred using DESeq2. The optimization is performed with the BFGS algorithm.

$$\begin{aligned} \{\mu_i^{MLE}, C_i^{MLE}\} &= \arg \max_{\{\mu, C\}} \sum_{j=1}^J \log \mathcal{L}_{NB}(x_{i,j} | \mu = \\ &\mu * \exp(\langle X, C \rangle) * \tilde{s}_j, \phi = \tilde{\phi}_i) \end{aligned} \quad (4)$$

where μ_i is the constant mean parameter of gene i and C_i is the set of batch correction coefficients for gene i in log space which correspond to the model matrix X .

Impulse model

1. **Initialization:** Each gene model is initialized twice based on a peak and a valley model. Firstly, one expression level is estimated for each time point by averaging all size factor-corrected samples. If a batch structure is given, the expression means are corrected for batch effects based on batch factor estimates. The batch factors are estimated as the ratio of the mean size factor-corrected expression level of the samples of a given batch of a given gene and all samples of this gene. Initial (h0) and steady state (h2) are initialized as the expression level at the first and the last time point. The transition state is estimated as the highest (peak) or lowest (valley) expression level between first and last time point ('extremum'). To estimate the transition time parameters t1 and t2, expression gradients are locally linearly approximated between adjacent time points. t1 is initialized as the average of the two time coordinates associated with the maximal (peak) or minimal (valley) linear gradient approximation out of the time points before the estimated 'extremum'. t2 is initialized as the average of the two time coordinates associated with the minimal (peak) or maximal (valley) linear gradient approximation out of the time points after the estimated 'extremum'.

2. **Optimisation:** The cost function for the fit is the negative binomial log-likelihood of the data, the value of the impulse model is the mean parameter. The optimization is performed with the BFGS algorithm.

$$\{\mathcal{P}_i^{MLE}, C_i^{MLE}\} = \arg \max_{\{\mathcal{P}, C\}} \sum_{j=1}^J \log \mathcal{L}_{NB}(x_{i,j} | \mu) \quad (5)$$

$$\mu = f_{\text{Impulse}}(t_j | \mathcal{P}_i) * \exp((X, C)) * \tilde{s}_j, \phi = \tilde{\phi}_i$$

where $\mathcal{P}_i = \{h_0, h_1, h_2, t_1, t_2, \beta\}$ is the set of parameters of the impulse model fit to gene i and C_i is the set of batch correction coefficients for gene i in log space which correspond to the model matrix X . $f_{\text{Impulse}}(t | \mathcal{P}_i)$ is the impulse model with the parameters \mathcal{P}_i evaluated at time point t . ImpulseDE2 fits the amplitude parameters of the impulse model and batch factors in log space and imposes a log-likelihood sensitivity boundary at 10^{-10} below which changes in the parameters do not affect the log-likelihood which constraints the value of the impulse function to be positive. Thereby, the negative binomial mean parameter is guaranteed to be larger than zero.

Sigmoid model: Fit as the impulse model but with the sigmoid instead of the impulse model.

$$\mu(t) = f_{\text{sigmoid}}(t) = h_0 + (h_2 - h_0) \frac{1}{1 + e^{-\beta(t-t_1)}} \quad (6)$$

$$= f_{\text{Impulse}}(t | h_1 = h_2)$$

where the amplitude parameters are $h_0 = f_{\text{Impulse}}(t \rightarrow \infty)$ and $h_2 = f_{\text{Impulse}}(t \rightarrow \infty)$ (steady state expression), t_1 is the state transition time, β is the slope parameter of the sigmoid function.

Initializations are based on a monotonous increase and a monotonous decrease model.

2.3 Differential expression analysis

Standard differential expression analysis Model selection between alternative and null model is based on a log-likelihood ratio test. P-values for differential expression are computed based on the χ^2 -distributed deviance and are Benjamin-Hochberg corrected [3]. Note that the deviance is only χ^2 -distributed if the null model is nested within the alternative model which is given in both case-only and case-control differential expression analysis (proof in Supplementary Notes).

Identification of transiently activated genes Again, model selection between alternative and null model is based on a log-likelihood ratio test. P-values for differential expression are computed based on the χ^2 -distributed deviance and are Benjamin-Hochberg corrected [3]. Transiently regulated genes are defined as genes which have Benjamin-Hochberg corrected p-value of the comparison impulse model against sigmoid model below a significance threshold. Moreover, transiently regulated genes are additionally required to not have a monotonous impulse model fit. A fit is defined as monotonous if the largest and the smallest fitted value of all observed time points correspond at the first and the last time point respectively. Genes with monotonous expression trajectories are defined as genes which do not meet one of the above two conditions but which have a FDR corrected p-value of the the comparison sigmoid model against constant model below a significance threshold. Up- and down- regulation is defined based on the impulse model fits: Up-regulated transient genes have a peak-like trajectory, down-regulated transient genes have a valley like trajectory, permanently up-regulated genes have a monotonously increasing trajectory, permanently down-regulated genes have a monotonously decreasing trajectory.

3 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: Drosophila (Graveley) [11] ($q=1e-5$), erythroid (Lara-Astasio) [14] ($q=1e-5$), LPS (Jovanovic) [12] ($q=1e-2$), myeloid (Sykes) [20] (distance: z-scores, $q=1e-5$), hESC (Chu) [6] ($q=1e-5$), estrogen (Baran-Gale) [2] ($q=1e-5$), Plasmodium (Broadbent) [5] (shown are all lncRNAs in the data set).

4 Gene set enrichment analysis

We performed gene set enrichment analysis against gene sets from MSigDB [19]: The H hallmark gene set [15], the C2 curated gene set collection [19], the C3.tft transcription factor target sets from TRANSFAC 7.4 [17], the C5.mf GO molecular function sets [7], the C5.bp GO biological process sets [7] and the C7 immunological signatures gene sets [10]. The gene sets are deposited in MSigDB as HGNC identifiers. We mapped all identifiers from the data sets to HGNC identifiers with biomaRt [8].

We define genes as differentially called if 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 Benjamin-Hochberg 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 Benjamin-Hochberg 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 overrepresented gene sets at a q-value of 0.05 for the individual experimental settings.

We mapped the gene identifiers to HGNC as follows:

We mapped the ensemble transcript identifiers of the LPS (Jovanovic) 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. [20]) 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. [6]) are HGNC (human) and we used these directly.

We mapped iChIP peaks (mouse) of the erythroid lineage iChIP peak data set (Lara-Astiaso et al. [14]) to MGI identifiers (mouse) with GREAT [18] 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).

5 Count matrix generation

5.1 myeloid (Sykes)

We used the published expected count matrix for all analysis [20] (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE84874>).

5.2 hESC (Chu)

We used the published expected count matrix for all analysis [6] (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE75748>).

5.3 erythroid chromatin (Lara-Astiaso)

We aligned reads from fastq files [14] (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE59636>) with bowtie2 [13] against GRCm38.p3. Peaks were called with MACS2 [9] (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 [14] 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 - Megakaryocytic erythroid 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.

5.4 LPS (Jovanovic)

We created an expected count matrix with kallisto [4] based on a mm10 index build with kmers of length 29 base pairs [12] (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE59784>).

5.5 estrogen (Baran-Gale)

We used the published expected count matrix for the heatmap [2] (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE78167>).

5.6 Plasmodium (Broadbent)

We used the published expected FPKM matrix for the heatmap [5] (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE57439>).

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

5.7 Drosophila (Graveley)

We downloaded the .sra files and extracted .fastq files from GEO with the following SRR IDs, mapped to developmental stages and time points:

sample name	SRR	time
embryos, 0-2 hr	SRR1197337	0
embryos, 0-2 hr	SRR1197370	0
embryos, 0-2 hr	SRR767606	0
embryos, 2-4 hr	SRR1197368	2
embryos, 2-4 hr	SRR1197336	2
embryos, 2-4 hr	SRR767626	2
embryos, 4-6 hr	SRR1197338	4
embryos, 4-6 hr	SRR767609	4
embryos, 6-8 hr	SRR767610	6
embryos, 6-8 hr	SRR1197333	6
embryos, 8-10 hr	SRR1197335	8
embryos, 8-10 hr	SRR767615	8
embryos, 10-12 hr	SRR767616	10
embryos, 10-12 hr	SRR1197334	10
embryos, 10-12 hr	SRR1197367	10
embryos, 12-14 hr	SRR1197369	12
embryos, 12-14 hr	SRR1197332	12
embryos, 12-14 hr	SRR767613	12
embryos, 14-16 hr	SRR767618	14
embryos, 14-16 hr	SRR1197331	14
embryos, 16-18 hr	SRR1197330	16
embryos, 16-18 hr	SRR1197365	16
embryos, 16-18 hr	SRR767605	16
embryos, 18-20 hr	SRR767622	18
embryos, 18-20 hr	SRR1197363	18
embryos, 18-20 hr	SRR1197327	18
embryos, 20-22 hr	SRR1197329	20
embryos, 20-22 hr	SRR1197364	20
embryos, 20-22 hr	SRR767620	20
embryos, 22-24 hr	SRR1197366	22
embryos, 22-24 hr	SRR767625	22
embryos, 22-24 hr	SRR1197328	22
L1 stage larvae	SRR1197325	42
L1 stage larvae	SRR767624	42
L1 stage larvae	SRR1197426	42
L2 stage larvae	SRR1197425	66
L2 stage larvae	SRR767623	66
L2 stage larvae	SRR1197324	66
L3 stage larvae, 12 hr post-molt	SRR1197326	83
L3 stage larvae, 12 hr post-molt	SRR767627	83
L3 stage larvae, 12 hr post-molt	SRR1197424	83
pupae, 12 hr after WPP	SRR767608	132
pupae, 12 hr after WPP	SRR1197422	132
pupae, 12 hr after WPP	SRR1197289	132
pupae, 24 hrs after WPP	SRR1197288	144
pupae, 24 hrs after WPP	SRR1197421	144
pupae, 24 hrs after WPP	SRR767604	144
pupae, 2 days after WPP	SRR1197287	168
pupae, 2 days after WPP	SRR767614	168
pupae, 2 days after WPP	SRR1197420	168
pupae, 3 days after WPP	SRR1197419	192
pupae, 3 days after WPP	SRR767611	192
pupae, 3 days after WPP	SRR1197285	192
pupae, 4 days after WPP	SRR1197416	216
pupae, 4 days after WPP	SRR1197286	216
pupae, 4 days after WPP	SRR767612	216

adult male, 1 day after eclosion	SRR1197315	240
adult male, 1 day after eclosion	SRR1197415	240
adult male, 1 day after eclosion	SRR767619	240
adult male, 5 days after eclosion	SRR1197468	336
adult male, 5 days after eclosion	SRR1197431	336
adult male, 5 days after eclosion	SRR1197417	336
adult male, 5 days after eclosion	SRR1197435	336
adult male, 5 days after eclosion	SRR1197432	336
adult male, 5 days after eclosion	SRR1197429	336
adult male, 5 days after eclosion	SRR1197469	336
adult male, 5 days after eclosion	SRR1197436	336
adult male, 5 days after eclosion	SRR1197430	336
adult male, 5 days after eclosion	SRR1197316	336
adult male, 5 days after eclosion	SRR767607	336
adult male, 30 days after eclosion	SRR1197391	936
adult male, 30 days after eclosion	SRR1197311	936
adult male, 30 days after eclosion	SRR1197413	936

Note that there all correspond to RNA-seq experiments. We prepared count matrices from the .fastqs with kallisto quant using a 31mer index prepared from a *Drosophila melanogaster* reference transcriptome obtained from ftp://ftp.ensembl.org/pub/release-90/fasta/drosophila_melanogaster/cdna/Drosophila_melanogaster.BDGP6.cdna.all.fa.gz.

6 Reference method parameters

6.1 Standard settings

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 + Time + Condition:Time versus Time)

Case-control multiple batches standard model with one batch factor per batch (Condition + Time + 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 + Time + 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. 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 DESeq2 based on natural cubic splines with four degrees of freedom (DESeq2splines) were:

Case-only single batch (1 + Spline1 + Spline2 + Spline3 + Spline4 versus 1),

Case-only multiple batches (1 + Spline1 + Spline2 + Spline3 + Spline4 + Batch versus 1+Batch),

Case-control single batches (1 + Condition + Spline1 + Spline2 + Spline3 + Spline4 + Condition:Spline1 + Condition:Spline2 + Condition:Spline3 + Condition:Spline4 versus 1 + Spline1 + Spline2 + Spline3 + Spline4)

Case-control multiple batches standard model with one batch factor per batch (1 + Condition + Spline1 + Spline2 + Spline3 + Spline4 + Condition:Spline1 + Condition:Spline2 + Condition:Spline3 + Condition:Spline4 + Batch versus 1 + Spline1 + Spline2 + Spline3 + Spline4 + 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 (1 + Condition + Spline1 + Spline2 + Spline3 + Spline4 + Condition:Spline1 + Condition:Spline2 + Condition:Spline3 + Condition:Spline4 + Batch + Condition:Batch versus 1 + Spline1 + Spline2 + Spline3 + Spline4 + Batch). The coefficients Spline1, Spline2, Spline3 and Spline4 are the component names of the degree four natural cubic spline basis.

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.

For limma based on a natural cubic spline model with 4 degrees of freedom, the coefficients tested in the differential expression test with topTable() were:
Case-only single batch ("TimeSpline1", "TimeSpline2", "TimeSpline3", "TimeSpline4"),
Case-only multiple batches ("TimeSpline1", "TimeSpline2", "TimeSpline3", "TimeSpline4"),
Case-control single batches ("control", "control:TimeSpline1", "control:TimeSpline2", "control:TimeSpline3", "control:TimeSpline4")
Case-control two batches ("control", "control:BatchB", "control:TimeSpline1", "control:TimeSpline2", "control:TimeSpline3", "control:TimeSpline4"), where TimeSpline are the time-dependent spline coefficients, control and BatchB the control and second batch adjustment coefficients.

6.2 DESeq2 settings adapted to sequencing time course data

We attached example code for the usage of DESeq2 on temporal data in Supp. data 3.

7 Data simulation

7.1 Method benchmarking

Simulations were performed with the simulateDataSetImpulseDE2() function of ImpulseDE2. We first simulated hidden expression trajectories, then imposed library depth effects and then imposed negative binomial noise.

7.1.1 Simulation of Fig. 2a: Varying the number of time points sampled

We simulated 1000 constant genes and 500 genes with the following parametric expression profiles: linear, sigmoid and impulse on time courses with 5 to 12 time points sampled in three replicates.

We used simulateDataSetImpulseDE2() with the following settings: (vecTimePointsA = rep(seq(1, vecNTPbyRun[i]), 3), vecTimePointsB = NULL, vecBatchesA = NULL, vecBatchesB = NULL, scaNConst = 1000, scaNImp = 500, scaNLin = 500, scaNSig = 500, scaNRand = 0, scaMumax = 1000, scaSDEExpressionChange = 0.5, scaSDRand = NULL, scaMuSizeEffect = 1, scaSDSizeEffect = 0.1, scaMuBatchEffect = NULL, scaSDBatchEffect = NULL, scaSeedInit = i) where vecNTPbyRun contains the number of time points to simulate and ranges from 5 to 12 which are each simulated three times and i is the index of the simulation.

7.1.2 Simulation of Fig. 2b: Varying the standard deviation of a randomly changing expression profile

We simulated 2000 constant genes on time courses with 6 time points sampled in two replicates with increasingly strong random signal over time.

We used simulateDataSetImpulseDE2() with the following settings: (vecTimePointsA = rep(seq(1,6),2), vecTimePointsB = NULL, vecBatchesA = NULL, vecBatchesB = NULL, scaNConst = 0, scaNImp = 0, scaNLin = 0, scaNSig = 0, scaNRand = 2000, scaMumax = 1000, scaSDEExpressionChange = 0.5, scaSDRand=vecSDbyRun[i], scaMuSizeEffect = 1, scaSDSizeEffect = 0.1, scaMuBatchEffect = NULL, scaSDBatchEffect = NULL, scaSeedInit = i) where vecSDbyRun contains the standard deviation of the normal distribution from which random underlying expression means are drawn for each time point and gene, which contains the values (0, 0.2, 0.4, 0.6, 0.8, 1, 1.2, 1.4, 1.6, 1.8, 2) which are each simulated three times and i is the index of the simulation.

7.2 Distribution of the deviance computed by ImpulseDE2

We included this section to show that the deviance computed in the log-likelihood ratio test of ImpulseDE2 is indeed chi-square distributed.

We simulated three separate data sets with simulateDataSetImpulseDE2() function of ImpulseDE2: We simulated 1000 constant genes over a time course sampled ten time at time one to ten in duplicate. We used simulateDataSetImpulseDE2() with the following settings: (vecTimePointsA = rep(seq(1,10), 2), vecTimePointsB = NULL, vecBatchesA = NULL, vecBatchesB = NULL, scaMumax = 500, scaNConst = 1000, scaNImp = 0, scaNLin = 0, scaNSig = 0, scaNRand = 0, scaSDEExpressionChange = 0.5, scaSDRand

= NULL, scaMuSizeEffect = 1, scaSDSizeEffect = 0.1, scaMuBatchEffect = NULL, scaSDBatchEffect = NULL, scaSeedInit = i,) with i is the index of the simulation.

References

- [1] S Anders, W Huber, Anders S, and Huber W. Differential expression analysis for sequence count data. *Genome Biol*, 11(10):R106, 2010.
- [2] Jeanette Baran-gale, Jeremy E Purvis, and Praveen Sethupathy. An integrative transcriptomics approach identifies miR-503 as a candidate master regulator of the estrogen response in MCF-7 breast cancer cells. pages 1–12, 2016.
- [3] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing, 1995.
- [4] Nicolas L Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, 34(5):525–527, 2016.
- [5] Kate M Broadbent, Jill C Broadbent, Ulf Ribacke, Dyann Wirth, John L Rinn, and Pardis C Sabeti. Strand-specific RNA sequencing in *Plasmodium falciparum* malaria identifies developmentally regulated long non-coding RNA and circular RNA. *BMC genomics*, 16(1):454, 2015.
- [6] Li-Fang Chu, Ning Leng, Jue Zhang, Zhonggang Hou, Daniel Mamott, David T. Vereide, Jee Choi, Christina Kendziorski, Ron Stewart, and James A. Thomson. Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome Biology*, 17(1):173, 2016.
- [7] The Gene Ontology Consortium. Gene ontology consortium: Going forward. *Nucleic Acids Research*, 43(D1):D1049–D1056, 2015.
- [8] Steffen Durinck, Yves Moreau, Arek Kasprzyk, Sean Davis, Bart De Moor, Alvis Brazma, and Wolfgang Huber. BioMart and Bioconductor: A powerful link between biological databases and microarray data analysis. *Bioinformatics*, 21(16):3439–3440, 2005.
- [9] Jianxing Feng, Tao Liu, Bo Qin, Yong Zhang, and Xiaole Shirley Liu. Identifying ChIP-seq enrichment using MACS. *Nature Protocols*, 7(9):1728–1740, 2012.
- [10] Jernej Godec, Yan Tan, Arthur Liberzon, Pablo Tamayo, Sanchita Bhattacharya, Atul J Butte, Jill P Mesirov, and W Nicholas Haining. Compendium of Immune Signatures Identifies Conserved and Species-Specific Biology in Response to Inflammation. *Immunity*, 44(1):194–206, 2016.
- [11] Brenton R. Graveley, Angela N. Brooks, Joseph W. Carlson, Michael O. Duff, Jane M. Landolin, Li Yang, Carlo G. Artieri, Marijke J. Van Baren, Nathan Boley, Benjamin W. Booth, James B. Brown, Lucy Cherbas, Carrie A. Davis, Alex Dobin, Renhua Li, Wei Lin, John H. Malone, Nicolas R. Mattiuzzo, David Miller, David Sturgill, Brian B. Tuch, Chris Zaleski, Dayu Zhang, Marco Blanchette, Sandrine Dudoit, Brian Eads, Richard E. Green, Ann Hammonds, Lichun Jiang, Phil Kapranov, Laura Langton, Norbert Perrimon, Jeremy E. Sandler, Kenneth H. Wan, Aaron Willingham, Yu Zhang, Yi Zou, Justen Andrews, Peter J. Bickel, Steven E. Brenner, Michael R. Brent, Peter Cherbas, Thomas R. Gingeras, Roger A. Hoskins, Thomas C. Kaufman, Brian Oliver, and Susan E. Celniker. The developmental transcriptome of *Drosophila melanogaster*. *Nature*, 471(7339):473–479, 2011.
- [12] Marko Jovanovic, Michael S Rooney, Philipp Mertins, Dariusz Przybylski, Nicolas Chevrier, Rahul Satija, Edwin H Rodriguez, Alexander P Fields, Schraga Schwartz, Raktima Raychowdhury, Maxwell R Mumbach, Thomas Eisenhaure, Michal Rabani, Dave Gennert, Diana Lu, Toni Delorey, Jonathan S Weissman, Steven A Carr, Nir Hacohen, and Aviv Regev. Immunogenetics. Dynamic profiling of the protein life cycle in response to pathogens. *Science (New York, N. Y.)*, 347(6226):1259038, 2015.
- [13] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with Bowtie 2. *Nat Methods*, 9(4):357–359, 2012.
- [14] D. Lara-Astiaso, A. Weiner, E. Lorenzo-Vivas, I. Zaretsky, D. A. Jaitin, E. David, H. Keren-Shaul, A. Mildner, D. Winter, S. Jung, N. Friedman, and I. Amit. Chromatin state dynamics during blood formation. *Science*, 345(6199):943–9, 2014.

- [15] Arthur Liberzon, Chet Birger, Helga Thorvaldsdottir, Mahmoud Ghandi, Jill P. Mesirov, and Pablo Tamayo. The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Systems*, 1(6):417–425, 2015.
- [16] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, 2014.
- [17] V. Matys, E. Fricke, R. Geffers, E. Gößling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. E. Kel, O. V. Kel-Margoulis, D. U. Kloos, S. Land, B. Lewicki-Potapov, H. Michael, R. Münch, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele, and E. Wingender. TRANSFAC®: Transcriptional regulation, from patterns to profiles. *Nucleic Acids Research*, 31(1):374–378, 2003.
- [18] Cory Y McLean, Dave Bristor, Michael Hiller, Shoa L Clarke, Bruce T Schaar, Craig B Lowe, Aaron M Wenger, and Gill Bejerano. GREAT improves functional interpretation of cis-regulatory regions. *Nature biotechnology*, 28(5):495–501, 2010.
- [19] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael a Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, and Jill P Mesirov. 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, 2005.
- [20] David B Sykes, Youmna S Kfoury, François E Mercier, Mathias J Wawer, Jason M Law, Mark K Haynes, Timothy A Lewis, Amir Schajnovitz, Esha Jain, Dongjun Lee, Hanna Meyer, Kerry A Pierce, Nicola J Tolliday, Anna Waller, Steven J Ferrara, Ashley L Eheim, Detlef Stoeckigt, Katrina L Maxcy, Julien M Cobert, Jacqueline Bachand, Brian A Szekely, Siddhartha Mukherjee, Larry A Sklar, Joanne D Kotz, Clary B Clish, Ruslan I Sadreyev, Paul A Clemons, Andreas Janzer, Stuart L Schreiber, and David T Scadden. Inhibition of Dihydroorotate Dehydrogenase Overcomes Differentiation Blockade in Acute Myeloid Leukemia. *Cell*, pages 171–186, 2016.