

# Projected $t$ -SNE for batch correction

Emanuele Aliverti<sup>1</sup>, Jeff Tilson<sup>2</sup>, Dayne Filer<sup>2,3</sup>, Benjamin Babcock<sup>3,4</sup>, Alejandro Colaneri<sup>3</sup>, Jennifer Ocasio<sup>4,5</sup>, Timothy R. Gershon<sup>4,5,6,7</sup>, Kirk C. Wilhelmsen<sup>2,3,4</sup>, and David B. Dunson<sup>8</sup>

<sup>1</sup>Department of Statistical Sciences, University of Padova, Padova

<sup>2</sup>RENCI, Chapel Hill, NC 27517 USA

<sup>3</sup>Department of Genetics, University of North Carolina School of Medicine, Chapel Hill

<sup>4</sup>Department of Neurology, University of North Carolina School of Medicine, Chapel Hill

<sup>5</sup>UNC Neuroscience Center, University of North Carolina School of Medicine, Chapel Hill

<sup>6</sup>Carolina Institute for Developmental Disabilities, University of North Carolina School of Medicine, Chapel Hill

<sup>7</sup>Lineberger Comprehensive Cancer Center, University of North Carolina School of Medicine, Chapel Hill

<sup>8</sup>Department of Statistical Science, Duke University, Durham

## Abstract

**Motivation:** Low-dimensional representations of high-dimensional data are routinely employed in biomedical research to visualize, interpret and communicate results from different pipelines. In this article, we propose a novel procedure to directly estimate  $t$ -SNE embeddings that are not driven by batch effects. Without correction, interesting structure in the data can be obscured by batch effects. The proposed algorithm can therefore significantly aid visualization of high-dimensional data.

**Results:** The proposed methods are based on linear algebra and constrained optimization, leading to efficient algorithms and fast computation in many high-dimensional settings. Results on artificial single-cell transcription profiling data show that the proposed procedure successfully removes multiple batch effects from  $t$ -SNE embeddings, while retaining fundamental information on cell types. When applied to single-cell gene expression data to investigate mouse medulloblastoma, the proposed method successfully removes batches related with mice identifiers and the date of the experiment, while preserving clusters of oligodendrocytes, astrocytes, and endothelial cells and microglia, which are expected to lie in the stroma within or adjacent to the tumors.

**Availability:** Source code implementing the proposed approach in R and Julia is available at the link [https://github.com/emanuelealiverti/BC\\_tSNE](https://github.com/emanuelealiverti/BC_tSNE), including a tutorial to reproduce the simulation studies.

**Contact:** [aliverti@stat.unipd.it](mailto:aliverti@stat.unipd.it)

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 INTRODUCTION

Recent technological improvements in transcriptome analysis have led to many valuable insights into complex biological systems, with single-cell RNA transcription profiling (scRNAseq) analysis being one of the most popular tools to investigate intricate cellular processes (Hwang et al., 2018). In biostatistical analysis, low-dimensional representations of high-dimensional scRNAseq data are ubiquitous, playing a central role during multiple phases of scientific investigation. For example, visualisation tools are used during normalisation, correction and dimensionality reduction to evaluate success of the pipelines, and in downstream analysis to illustrate results from intermediate procedures such as clustering (e.g. Luecken and Theis, 2019; Lun et al., 2016; Vieth et al., 2019).

A wide variety of methods for linear and non-linear dimensionality reduction and data visualisation are available, with  $t$ -distributed Stochastic Neighbor Embedding ( $t$ -SNE, Maaten and Hinton, 2008) and Uniform Manifold Approximation and Projection (UMAP, McInnes et al., 2018) being of great utility in analyzing scRNAseq data. Such methods allow one to describe the dataset in 2-3 dimensions via graphical representations, highlighting the main structure of the data and preserving relevant properties such as the presence of isolated clusters (Kobak and Berens, 2019). During pre-processing, low-dimensional representations are fundamental for identifying potential issues in the data; for example, inadequate data integration or the presence of batch effects (Luecken and Theis, 2019; Lun et al., 2016). Indeed, without explicit adjustment, variations in the low-dimensional summaries may be driven by nuisance covariates — such as batches due to different devices used for an experiment — instead of the primary factors of scientific interest — such as cell types. In intermediate analyses, such batch effects can limit the utility of the low-dimensional graphical representations in visualizing, interpreting and communicating results from downstream processes conducted at the cell level; for example, clustering, cell annotations or compositional analysis (Wagner et al., 2016).

In a typical workflow, standardized pipelines proceed sequentially, with low-dimensional embeddings estimated after several steps involving normalisation, integration, batch-correction and feature selection from raw data; see Luecken and Theis (2019) and references therein for a recent detailed review. However, such processing might lead to propagation of errors and unreliable representations. For example, over-correction of batch effects might also remove important biological features, and lead to low-dimensional embeddings which are not driven by such biological factors (e.g. Lun et al., 2016). Such an issue will be entirely propagated to downstream processes, leading to low-dimensional embeddings which cannot highlight information on factors of interest and might provide misleading evidence.

Motivated by the above considerations, the focus of this article is on producing batch-corrected modifications of  $t$ -SNE that can be used to remove associations with multiple batches from low-dimensional

embeddings. Our methods are based on linear algebra results and modification of gradient descent optimisation, therefore providing simple and scalable tools in high-dimensional problems. The proposed procedure directly estimates low-dimensional embeddings, which are not driven by systematic batch-effects including batch-correction, and provides a synthetic representation to correctly visualise results from different pipelines.

Several approaches are available in the literature for batch-correction and data integration, covering a wide range of methods which encompass linear modelling via Empirical-Bayes (Johnson et al., 2007), canonical correlation analysis (Butler et al., 2018) and Mutual Nearest Neighbors (MNN, Haghverdi et al., 2018); see Büttner et al. (2019) and references therein for a recent comparison, and the `scater` package (McCarthy et al., 2017) for a convenient implementation. Differently from routine corrections for scRNAseq data, our approach is not targeted to correct the entire set of high-dimensional data, but only its low-dimensional representation obtained via  $t$ -SNE. Therefore, the proposed approach directly relates to the framework of “removal of unwanted variations” (RUV. See Grün and van Oudenaarden, 2015; Risso et al., 2014; Leek and Storey, 2007), where interest is on measuring latent variables which are not affected by batch-effects and experimental conditions, but are only driven by relevant biological factors.

Specifically, we introduce a novel modification of  $t$ -SNE to integrate batch correction into estimation of low-dimensional embeddings. Such an approach is not intended as a substitute to the canonical pipelines for downstream analysis; which, for example, focus on estimating clusters in the  $k$ -NN graph of the PCA subspace (e.g. Wolf et al., 2019). Instead, the proposed contribution serves as a parallel tool to provide a robust visualisation of scRNAseq data, which is less subject to propagation of errors and can be used to validate results from different pipelines, or to identify potential pitfalls. Although there is some evidence that clustering in the  $t$ -SNE subspace can provide insights on the community structure of the data (Linderman and Steinerberger, 2019), such a procedure is generally not recommended in the analysis of scRNAseq data and is beyond the scope of the current article; see Kobak and Berens (2019) for further discussion. The proposed algorithm allows joint correction for multiple batches and leverages two different adjustments, to handle both linear and non-linear effects. Linear correction is achieved adapting the strategy of Aliverti et al. (2018), while correction for non-linear effects leverages a projection step during  $t$ -SNE optimisation, related to the locally linear correction implemented in Haghverdi et al. (2018). A full implementation of the method is publicly available in a mixed R and Julia implementation, at the link <https://github.com/emanuelealiverti/BC.tSNE>.

## 2 METHODS

### 2.1 Notation & problem formulation

Consider a data matrix  $\mathbf{X} = \{\mathbf{x}_i^T\}_{i=1}^n$  with observations  $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})$ . In many biological applications, the number of features  $p$  is tremendously large and it is of interest to provide accurate low-dimensional representation of such high dimensional data. Dimensionality reduction techniques focus on finding low-dimensional counterparts  $\mathbf{y}_i = (y_{i1}, \dots, y_{iq})$  of each  $\mathbf{x}_i$ , preserving as much structure as possible with  $q \ll p$  components; generally,  $q = 2$  or  $q = 3$  for the ease of graphical visualisation. Original observations  $\mathbf{x}_i$  can potentially lie in a complex and highly non-linear manifold; for example, wrapped spaces such as rolls (e.g Lee and Verleysen, 2005). In contrast, the desired low-dimensional embeddings lie on a standard  $q$ -dimensional Euclidean space, and  $\mathbf{y}_i$  determines the position of observation  $i$  in such an embedded space.

Low-dimensional representations are constructed in order to preserve some specific structure of the original data; some examples include preserving Euclidean distances (Multidimensional Scaling, Kruskal, 1964), variances (Principal Component Analysis), neighborhood graphs (Local Linear Embedding and Isomap, Roweis and Saul, 2000; Tenenbaum et al., 2000) or local similarities among points (Stochastic Neighbor Embedding, Hinton and Roweis, 2003). Many methods estimate an explicit function between the original data and their embeddings; for example, the PCA solution is a linear combination of the columns of  $\mathbf{X}$ . More recently, focus has shifted to obtaining  $\mathbf{y}_i$  without explicitly defining such a map, thus allowing a greater flexibility and range of application. In this article, we focus on the  $t$ -SNE methodology for dimensionality reduction and data visualisation (Maaten and Hinton, 2008).  $t$ -SNE attempts to find low-dimensional representations that preserve local similarities among data points, with similarities parameterized as conditional probabilities of belonging to the same local neighborhood. In the following paragraphs, we review the standard formulation of  $t$ -SNE before introducing our adjustments for batch effects.

### 2.2 Standard $t$ -SNE algorithm

In the original input space,  $t$ -SNE defines dissimilarities among points as symmetric probabilities  $p_{ij} = (p_{i|j} + p_{j|i})/2n$ , with

$$p_{i|j} = \frac{\exp(-0.5\|\mathbf{x}_i - \mathbf{x}_j\|^2/\sigma_i^2)}{\sum_{k,k \neq i} \exp(-0.5\|\mathbf{x}_i - \mathbf{x}_k\|^2/\sigma_i^2)}. \quad (1)$$

Equation 1 can be interpreted as the probability that point  $i$  picks  $j$  as its neighbor, under a Gaussian kernel centered at  $\mathbf{x}_i$  and with standard deviation equal to  $\sigma_i$ . The intuition behind the introduction of  $p_{ij}$  comes from averaging  $p_{i|j}$  and  $p_{j|i}$  to reduce the relative impact of outliers and define a symmetric dissimilarity metric (Maaten and Hinton, 2008). The parameter  $\sigma_i^2$  determines the width of the Gaussian kernel and, indirectly, the number of local neighbors associated with each point  $i$ , with

$i = 1, \dots, n$ . Defining  $\sigma_i^2$  is a primary step in producing  $t$ -SNE embeddings, with the selection determining the *perplexity* of the resulting distribution (Maaten and Hinton, 2008; Hinton and Roweis, 2003). Large values of  $\sigma_i^2$  correspond to a larger number of local neighbors and greater perplexity, while default values of perplexity range in the interval  $\{10, 50\}$  (Maaten and Hinton, 2008). Embeddings generally show robustness to moderate changes in perplexity (Maaten, 2014).

Dissimilarity among points  $\mathbf{y}_i$  in the embedded space is defined through the kernel of a  $t$ -distribution with 1 degree of freedom, setting

$$q_{ij} = \frac{(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_{k, k \neq i} (1 + \|\mathbf{y}_i - \mathbf{y}_k\|^2)^{-1}}. \quad (2)$$

The  $t$ -SNE embeddings  $\mathbf{y}_i$  are selected minimizing the Kullback-Leibler divergence between  $p_{ij}$  and  $q_{ij}$ ; note that  $p_{ij}$  does not depend on  $\mathbf{y}$  and is a fixed value given the input data. Let  $\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_n)$  and highlight the dependency of  $q_{ij}$  on the embeddings  $\mathbf{y}$  in Equation 2 as  $q_{ij}(\mathbf{y})$ . Formally,  $t$ -SNE is the solution to the following optimisation problem.

$$\operatorname{argmin}_{\mathbf{y}} \{L(\mathbf{y})\} = \operatorname{argmin}_{\mathbf{y}} = \left\{ \sum_{i=2}^n \sum_{j=1}^i p_{ij} \log \frac{p_{ij}}{q_{ij}(\mathbf{y})} \right\} \quad (3)$$

The objective function  $L(\mathbf{y})$  can be optimized through gradient methods. Indeed, the partial derivative of the loss functions in Equation 3 with respect to  $\mathbf{y}_i$  is equal to

$$\begin{aligned} \nabla L(\mathbf{y}_i) &= \frac{\partial L}{\partial \mathbf{y}_i} = 4 \sum_{\substack{j=1 \\ j \neq i}}^n (p_{ij} - q_{ij}) q_{ij} Z(\mathbf{y}_i - \mathbf{y}_j); \\ Z &= \sum_{l \neq k} (1 + \|\mathbf{y}_l - \mathbf{y}_k\|^2)^{-1}; \end{aligned} \quad (4)$$

see Maaten and Hinton (2008, Appendix A) for the complete derivation.

Therefore, the generic gradient descent step with momentum correction for updating  $\mathbf{y}_i$  at iteration  $t + 1$  sets

$$\mathbf{y}_i^{(t+1)} = \mathbf{y}_i^{(t)} + \eta^{(t)} \nabla L(\mathbf{y}_i^{(t)}) + \alpha^{(t)} (\mathbf{y}_i^{(t)} - \mathbf{y}_i^{(t-1)}), \quad (5)$$

with  $\eta^{(t)}$  indicating the learning rate and  $\alpha^{(t)}$  the momentum term; see Maaten and Hinton (2008) for practical advices on the choice of such functions.

### 2.3 Batch-corrected $t$ -SNE

Let  $\mathbf{Z}$  denote an additional variable which contains batch information. We refer to the proposed method as BC- $t$ -SNE (Batch-Corrected  $t$ -SNE) in the sequel. When the number of features  $p$  is extremely large and when it exceeds the number of observations  $n$ , direct application of  $t$ -SNE on the raw data  $\mathbf{X}$

can be challenging, computationally inefficient, and lead to poor results. Therefore, it is generally advised to perform a preliminary dimensionality reduction, and then apply  $t$ -SNE over such reduced representation to improve the results (Maaten, 2014). For example, default software implementation estimates  $t$ -SNE embeddings on the first  $k$  principal components, with  $k$  in the range [30 – 50] (e.g. Krijthe et al., 2018). Reducing the dimensionality from  $p$  to  $k$  speeds up computation and reduces noise without affecting local similarities among observations (Maaten and Hinton, 2008).

The first step of BC- $t$ -SNE is motivated by the above considerations, and focuses on processing the data with the approach introduced in Aliverti et al. (2018) to explicitly obtain the optimal low-rank approximation of a matrix  $\mathbf{X}$  in Frobenius norm under an orthogonality constraint between such approximation and the batch variable  $\mathbf{Z}$ . Therefore, the method removes linear effects between the reduced data and the variables in  $\mathbf{Z}$  with minimal information loss. Such an approach is based on computing the residuals from a multivariate regression among the left singular vectors of  $\mathbf{X}$  and  $\mathbf{Z}$ , and is therefore comparable with standard PCA in terms of computational requirements, providing a practical alternative to perform dimensionality reduction while simultaneously achieving batch removal. Although such a procedure is optimal in removing the linear influence of batches, effects beyond linearity might still affect  $t$ -SNE embeddings. In practical applications such higher order effects are often small in magnitude, and second order adjustment often lead to satisfactory results in terms of batch-correction (e.g. Aliverti et al., 2018). However, since the  $t$ -SNE embeddings  $\mathbf{y}_i$  are a complex non-linear functional of the original  $\mathbf{x}_i$ , inclusion of higher-order constraints provides a reasonable conservative choice.

The second step of BC- $t$ -SNE adjustment can be better motivated by introducing some details on gradient descent, which can be interpreted as an optimisation to minimize the linearisation of the loss function (Kullback-Leibler divergence for  $t$ -SNE) around the current estimates, including a smoothing penalty that penalises abrupt changes (e.g. Hastie et al., 2015). To see that, consider the gradient descent step in Equation 5, setting without loss of generality  $\alpha^{(t)} = 0$ . The following alternative representation holds.

$$\mathbf{y}_i^{(t+1)} = \operatorname{argmin}_{\mathbf{y}_i \in \mathbb{R}^q} \left\{ L(\mathbf{y}_i^{(t)}) + \langle \nabla L(\mathbf{y}_i^{(t)}), \mathbf{y}_i - \mathbf{y}_i^{(t)} \rangle - \frac{1}{\eta^{(t)}} \|\mathbf{y}_i - \mathbf{y}_i^{(t)}\|_2^2 \right\}. \quad (6)$$

This view of gradient descent facilitates the introduction of further constraints. Indeed, this aim is achieved by restricting the solution  $\mathbf{y}_i \in \mathcal{C}$ , with  $\mathcal{C}$  denoting a constrained region of the original space  $\mathbb{R}^q$ . Such a constraint can be easily imposed by performing a standard gradient step, and then projecting the result back into the constrained set  $\mathcal{C}$ , leading to a procedure referred to as *projected* gradient descent; see, for example, Hastie et al. (2015, Sec 5.3.2) for further details. The choice of the constrained set  $\mathcal{C}$  covers a central role in the optimisation, since the projection should be computed

**Algorithm 1:** Batch-corrected- $t$ -SNE with projected gradient

Apply OG (Aliverti et al., 2018) to extract the first  $k$  components of  $\mathbf{X}$  and remove linear batch effects. Denote the  $n \times k$  reduced and adjusted matrix as  $\hat{\mathbf{X}}$  **for**  $i = 1, \dots, n$  **do**

- Perform binary search to find the value  $\sigma_i^2$  that achieves desired level of *perplexity* (Maaten and Hinton, 2008)

**end**

Compute the pairwise similarities  $p_{ij}$  in Equation 1 from  $\hat{\mathbf{X}}$  and  $\{\sigma_i^2\}_{i=1}^n$  **for**  $t = 1, \dots, T$  **do**

- Compute affinities  $q_{ij}$  defined in Equation 2 **for**  $i = 1, \dots, n$  **do**
  - Update  $\mathbf{y}_i^{(t+1)}$  (the  $i$ -th row of  $\mathbf{Y}^{(t+1)}$ ) as
 
$$\mathbf{y}_i^{(t+1)} = \mathbf{y}_i^{(t)} + \eta^{(t)} \nabla L(\mathbf{y}_i^{(t)}) + \alpha^{(t)} (\mathbf{y}_i^{(t)} - \mathbf{y}_i^{(t-1)})$$

**end**

Compute  $\boldsymbol{\beta}^{(t+1)} \leftarrow \text{solve}(\mathbf{Z}^\top \mathbf{Z}, \mathbf{Z}^\top \mathbf{Y}^{(t+1)})$  Compute projected gradient update, setting

$$\tilde{\mathbf{Y}}^{(t+1)} = \mathbf{Y}^{(t+1)} - \mathbf{Z} \boldsymbol{\beta}^{(t+1)}$$

**end**

**Output:** Return  $\tilde{\mathbf{Y}}^{(T)}$ .

easily in order to make the method practical in high-dimensional applications. With this motivation in mind, we propose a computationally simple solution and restrict  $\mathbf{Y} = \{\mathbf{y}_i^\top\}_{i=1}^n$  such that it is orthogonal with the subspace spanned by the columns of  $\mathbf{Z}$ . This constraint can be easily imposed with linear regression, computing at each iteration a projected gradient step which constructs an update  $\tilde{\mathbf{Y}}^{(t+1)}$  that projects the unconstrained solution  $\mathbf{Y}^{(t+1)}$ , making it orthogonal with the batch variables  $\mathbf{Z}$ . Pseudo-code illustrating the method is reported in Algorithm 1.

### 3 SIMULATION STUDY

A simulation study is conducted to evaluate the performance of the proposed method on artificial scRNAseq data. Artificial single-cell RNA sequencing data were generated with the BioConductor library `splatter`, which provides an interface to create complex datasets with several realistic features (Zappia et al., 2018). Specifically, a dataset consisting of  $p = 10000$  genes measured over  $n = 800$  cells was generated with 4 batch effects and 4 different cell types. A complete tutorial to reproduce the artificial data and simulation study in R and julia is available at the link [https://github.com/emanuelealiverti/BC\\_tSNE](https://github.com/emanuelealiverti/BC_tSNE).

The focus of the simulation is on assessing the success of BC- $t$ -SNE at removing unwanted associations while retaining information of the scientific factors of interest, which correspond to cell types in this particular example. The adjusted approach is also compared with a standard implementation of  $t$ -SNE, available with R package `Rtsne` (Krijthe, 2015), and with routine methods for batch-correction. In particular, we apply the recently proposed MNN (Haghverdi et al., 2018) and Harmony (Korsunsky

et al., 2019) methods for batch-correction, available through the R packages `batchelor` and `harmony`. In order to properly compare the methods, parameters of BC-*t*-SNE were fixed to the default configuration of the package `Rtsne`, which corresponds to setting the number of iterations  $T = 1000$ , a value of perplexity equal to 30 and  $\eta^{(t)} = 200, t = 1, \dots, T$  and  $\alpha^{(t)} = 0.5$  for  $t < 250$  and  $\alpha^{(t)} = 0.8$  for  $t \geq 250$ ; see also [Maaten and Hinton \(2008\)](#).

Figure 1 compares results from unadjusted *t*-SNE and the proposed method, respectively in the upper and lower panels; both approaches are estimated over a  $k = 30$  reduced components. Results for the unadjusted case confirm the presence of strong batch effects. Indeed, cells are divided into 4 main clusters corresponding to the different batches, denoted with different point shapes. Within each cluster, smaller groups of cells of the same type are present; however, it is clear that the main clusters are driven by batch information instead of cell types. Therefore, results from the upper panel of Figure 1 do not allow us to properly identify regions of the space of partitions which are consistent with the factors of scientific interest. The bottom panels of Figure 1 illustrate results for BC-*t*-SNE, Harmony and MNN and show that, after adjustment, the effect of unwanted batches is effectively removed from the *t*-SNE embeddings. Indeed, different point shapes are uniformly spread across the four main clusters, which now correspond to the different cell types denoted with different colors. From visual inspection, all the competitors achieve satisfactory results in terms of removing batch effects while preserving cell types, with BC-*t*-SNE highlighting the presence of different clusters more distinctly than the competitors. Such preliminary findings are quantitatively evaluated in Table 1, where the ability of the methods in removing batches and preserving cell types is evaluated in terms of silhouette coefficients, using the `scone` software ([Cole et al., 2019](#)), kBET test metric ([Büttner et al., 2019](#)), average LISI score ([Korsunsky et al., 2019](#)) and PC regression using `scater` ([McCarthy et al., 2017](#)). All the measures have been normalised and rescaled in  $[0 - 1]$ , with 0 indicating perfect separation across groups and 1 perfect integration; note that the interpretation of such metrics is different depending on the partitioning under investigation. Specifically, good performance in terms of batch effect removal corresponds to large values of the proposed metrics, while adequate preservation of cell types is associated with small values (e.g. [Korsunsky et al., 2019](#)). Table 1 indicates that all the methods achieve good performance in terms of batch-removal, with BC-*t*-SNE being most accurate in terms of kBET, average LISI and highly competitive in terms of rescaled silhouette coefficient. Coherency with Figure 1, the second half of Table 1 shows that BC-*t*-SNE outperforms the competitors in terms of conservation of cell types.

## 4 APPLICATION

### 4.1 Dataset description

Medulloblastoma is among the most frequent malignant brain tumors in children. Recent studies have observed that the Sonic Hedgehog (SHH) signaling pathway is hyperactivated in 30% of human



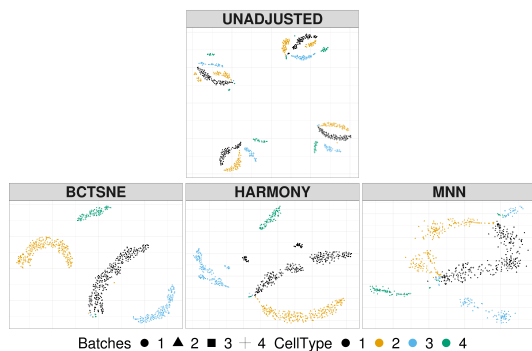


Figure 1: Simulation study. The color of points varies according to cell types, while shapes vary with batch groups. Upper plot shows the unadjusted  $t$ -SNE coordinates, while results after adjustment are reported in the bottom panels.

Table 1: Simulation study. Evaluation of batch removal and cell types preservation. Best performance is reported in boldface.

		SIL	kBET	LISI	PcReg
<b>Batches</b>	BC- $t$ -SNE	0.983	<b>0.999</b>	<b>0.741</b>	0.000
	Harmony	0.978	0.997	0.733	0.000
	MNN	<b>0.984</b>	0.921	0.668	0.000
<b>Cell types</b>	BC- $t$ -SNE	<b>0.428</b>	<b>0.294</b>	<b>0.011</b>	1.000
	Harmony	0.473	0.314	0.014	1.000
	MNN	0.689	0.999	0.043	1.000

medulloblastoma, therefore stimulating novel studies in this direction (Zurawel et al., 2000; Ellison et al., 2011). Activation of the SHH pathway, which stimulates proliferation of granule cell neurons during cerebellar development, has been used to create genetically engineered mice for scientific purposes, with the SmoM2 process being a routinely used pipeline (Rubin and de Sauvage, 2006). Specifically, SmoM2 mice have a transgenic mutated Smo allele which was originally isolated from a tumor and can be engineered to be not expressed until acted upon by Cre recombinase (Mao et al., 2006; Helms et al., 2000; Machold and Fishell, 2005). These mice are mated with genetically engineered matches that express Cre recombinase in cerebellar granular neuron progenitors, leading to descendants which develop medulloblastoma with 100% frequency by postnatal day 12.

Data used in this section come from 5 mice at postnatal day 12 created using such a pipeline and analysed under different sessions. Specifically, mice 1, 2 (Females) on July 2<sup>nd</sup>, mouse 3 (Male) on July 25<sup>th</sup> and mice 4, 5 (Males) on August 18<sup>th</sup>. Tumors were dissociated and individual cells co-encapsulated in a microfluidics chamber with primer-coated beads in oil-suspended droplets. All primers on each bead contained a bead-specific bar code and an unique molecular identifier (UMI), followed by an oligo-dT sequence, while mRNAs were captured on the oligo-dT, reverse-transcribed and amplified. Libraries were generated using the Drop-seq protocol V3.1 (Macosko et al., 2015). Following standard sequencing and preprocessing procedures, individual transcripts were identified by the UMI

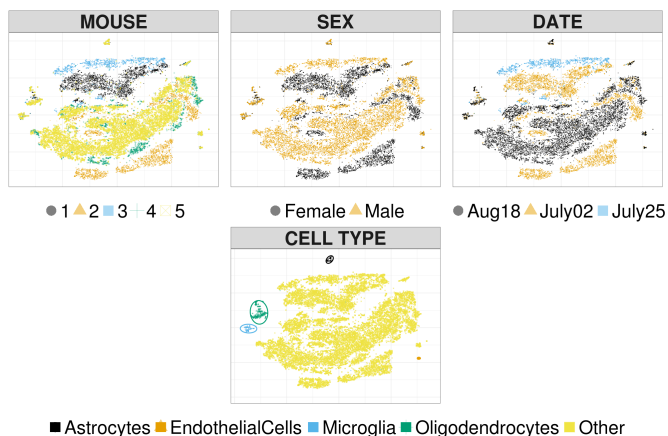


Figure 2: Unadjusted  $t$ -SNE coordinates. Points and shapes vary with batches.

bar code, with cell identity inferred from the bead-specific bar codes. Analysis on the normalised data has been restricted to cells with more than 500 detected genes. Furthermore, outlier cells with more than 4 standard deviations above the median number of genes were excluded from the analysis, along with UMIs and mitochondrial content per cell in order to address the common problems of gene drop out, unintentional cell-cell multiplexing and premature cell lysis (Vladoiu et al., 2018). The resulting pre-processed expression matrix  $\mathbf{X}$  consists of  $p = 16680$  genes measured over  $n = 17746$  different cells; for each mouse, the number of valid cells was 3381, 3402, 1454, 1647 and 8062, respectively.

The focus of our analysis is on evaluating if the proposed BC- $t$ -SNE method provides a robust data representation, which is successful at removing batch effects without affecting biological information of interest; see Ocasio et al. (2019) for an analysis involving cell-annotations on the same dataset.

## 4.2 Results

The presence of batch effects is investigated via unadjusted  $t$ -SNE embeddings, estimated over the first  $k = 50$  principal components; larger numbers of principal components resulted in less structured embeddings, and are not reported. The first row of Figure 2 highlights systematic differences with batch membership, while the second row shows information on cell type. Empirical results confirm the presence of strong batch effects, with respect to mouse identifier (first row, first column), sex of the mouse (first row, second column) and date of the experiment (first row, third column). For example, cells from mouse 1 form a cluster which is clearly distinct from the others. As expected, we observe some overlap between batch variables due to the experimental design. The second row of Figure 2 highlights differences across cell types, confirming that unadjusted  $t$ -SNE produces isolated clusters which are in agreement with the cell types indicated in Ocasio et al. (2019).

Figure 3 refers to adjusted  $t$ -SNE coordinates, estimated with Algorithm 1 using the same settings described in the simulation study. We compare BC- $t$ -SNE with the same approaches used in the

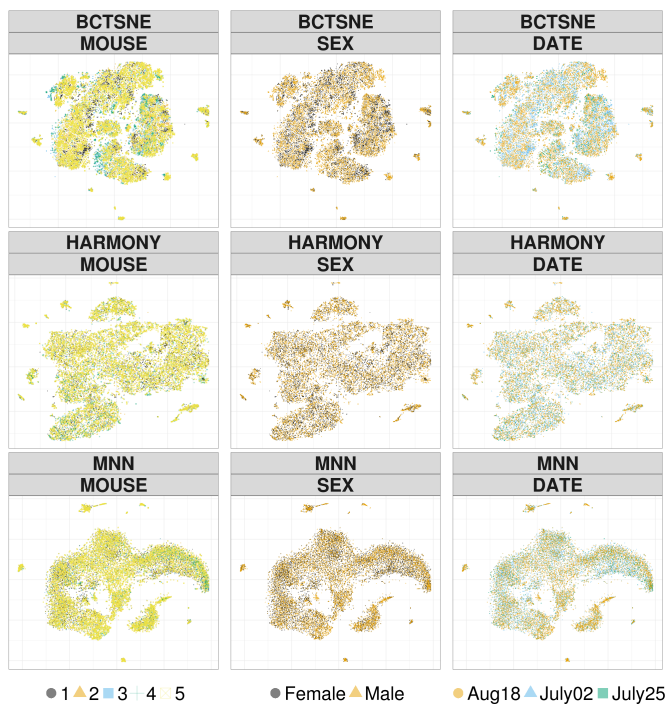


Figure 3:  $t$ -SNE coordinates after correction. Points and shapes vary with batches.

simulation studies. The first row of Figure 3 refers to BC- $t$ -SNE; second and third to Harmony and MNN, respectively. Results suggest a satisfactory performance in terms of batch effect removal for all the methods considered. Indeed, BC- $t$ -SNE embeddings from Figure 3 show no evidence of systematic variation with any of the batch variables under investigation. Such batches are marked by the color and shape of points in Figure 3, showing that the batches are spread homogeneously across the embedded space after adjustment.

Following the metrics used in the simulations, Table 2 quantitatively evaluates the success in removing batch effects. Results indicate that BC- $t$ -SNE achieves a performance which is highly competitive with the other approaches. Focusing, for example, on mouse identifiers, the normalised silhouette coefficient suggests that BC- $t$ -SNE removes batches more effectively than MNN and Harmony; similar conclusions hold also when kBET is considered. According to iLISI, instead, the baseline data adjustment methods perform better than BC- $t$ -SNE. This result is not surprising, since such approaches optimize objective functions which are directly related to the iLISI metric (Korsunsky et al., 2019).

Lastly, it is important to investigate that after removing batch effects, clusters of cell types associated with medulloblastoma are preserved in the low-dimensional coordinates, and similar cells are close in the embedded space. Figure 4 shows results for BC- $t$ -SNE and the baseline methods. The bulk of cells in a large central cluster are from tumors having markers within a range of differentiation states, ranging from proliferative, undifferentiated cells expressing the SHH-pathway transcription factor Gli1, to cells in successive states of CGN differentiation, marked by sequential expression of markers

Table 2: Evaluation of batch removal.

		SIL	kBET	LISI	pcR
BC- <i>t</i> -SNE	Sex	0.995	0.166	0.659	1.000
	Date	0.980	0.235	0.457	1.000
	Mouse	0.975	0.829	0.368	1.000
Harmony	Sex	0.997	0.299	0.846	1.000
	Date	0.987	0.421	0.540	1.000
	Mouse	0.975	0.854	0.498	0.999
MNN	Sex	0.999	0.219	0.844	1.000
	Date	0.996	0.392	0.546	1.000
	Mouse	0.958	0.794	0.502	0.999

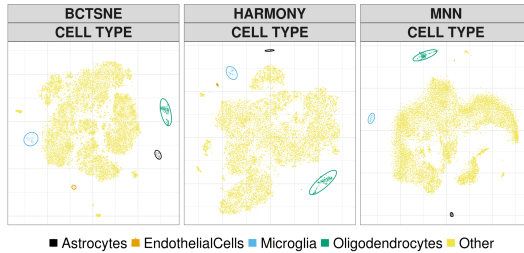


Figure 4: *t*-SNE coordinates after adjustment. Points and shapes vary with cell types.

Ccnd2, Barhl1, Cntn2, Rbfox3, and Grin2b (Ocasio et al., 2019). Clusters of cells surrounded with colored ellipses correspond to endothelial cells, microglia, oligodendrocytes and astrocytes, which are common in the stroma within or adjacent to the tumors. Empirical findings suggest that such clusters are correctly preserved after adjustment; see also Table 3 for a quantitative evaluation.

## 5 DISCUSSION

In this article we have introduced BC-*t*-SNE, a novel modification of *t*-SNE which allows for correction for multiple batch effects during estimation of the low-dimensional embeddings. The proposed approach has demonstrated good performance in simulation studies and on an application involving mouse medulloblastoma, where unwanted variations are successfully removed without removing information on differences across cell types.

A possible extension for future development involves adapting the proposed procedure to more efficient optimisation of the *t*-SNE loss function, to overcome the computational constraints encountered with large  $n$  (Maaten, 2014). One way to address such an issue involves the use of alternative gradient methods, with methods based on stochastic gradients being popular in the literature.

Table 3: Evaluation of cell-type preservation.

	SIL	kBET	iLISI	pcR
BC- <i>t</i> -SNE	0.052	0.000	0.000	0.000
MNN	0.056	0.000	0.000	0.000
Harmony	0.031	0.000	0.000	0.000

## FUNDING

The work of Emanuele Aliverti and David B. Dunson was partially funded by the grant “Fair predictive modelling” from the Laura & John Arnold Foundation. We thank the UNC CGBID Histology Core supported by P30 DK 034987, the UNC Tissue Pathology Laboratory Core supported by NCI CA016086 and UNC UCRF, and the UNC Neuroscience Center Confocal and Multiphoton Imaging and bioinformatics cores supported by The Eunice Kennedy Shriver National Institute of Child Health and Human Development (U54HD079124) and NINDS (P30NS045892). J.O. was supported by NINDS (F31NS100489). T.R.G. was supported by NINDS (R01NS088219, R01NS102627, R01NS106227) and by the UNC Department of Neurology Research Fund. T.R.G., K.W., and B.B. were supported by a TTSA grant from the NCTRACS Institute, which is supported by the National Center for Advancing Translational Sciences (NCATS), National Institutes of Health, through Grant Award Number UL1TR002489.

## REFERENCES

- Aliverti, E., Lum, K., Johndrow, J. E., and Dunson, D. B. (2018). Removing the influence of a group variable in high-dimensional predictive modelling. *arXiv preprint arXiv:1810.08255*.
- Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, 36(5):411.
- Büttner, M., Miao, Z., Wolf, F. A., Teichmann, S. A., and Theis, F. J. (2019). A test metric for assessing single-cell rna-seq batch correction. *Nature Methods*, 16(1):43.
- Cole, M. B., Risso, D., Wagner, A., DeTomaso, D., Ngai, J., Purdom, E., Dudoit, S., and Yosef, N. (2019). Performance assessment and selection of normalization procedures for single-cell rna-seq. *Cell Systems*, 8(4):315–328.
- Ellison, D. W., Dalton, J., Kocak, M., Nicholson, S. L., Fraga, C., Neale, G., Kenney, A. M., Brat, D. J., Perry, A., Yong, W. H., et al. (2011). Medulloblastoma: clinicopathological correlates of shh, wnt, and non-shh/wnt molecular subgroups. *Acta Neuropathologica*, 121(3):381–396.

- Grün, D. and van Oudenaarden, A. (2015). Design and analysis of single-cell sequencing experiments. *Cell*, 163(4):799–810.
- Haghverdi, L., Lun, A. T., Morgan, M. D., and Marioni, J. C. (2018). Batch effects in single-cell rna-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology*, 36(5):421.
- Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC.
- Helms, A. W., Abney, A. L., Ben-Arie, N., Zoghbi, H. Y., and Johnson, J. E. (2000). Autoregulation and multiple enhancers control math1 expression in the developing nervous system. *Development*, 127(6):1185–1196.
- Hinton, G. E. and Roweis, S. T. (2003). Stochastic neighbor embedding. In *Advances in neural information processing systems*, pages 857–864.
- Hwang, B., Lee, J. H., and Bang, D. (2018). Single-cell rna sequencing technologies and bioinformatics pipelines. *Experimental & Molecular Medicine*, 50(8):96.
- Johnson, W. E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127.
- Kobak, D. and Berens, P. (2019). The art of using t-sne for single-cell transcriptomics. *Nature Communications*, 10(1):1–14.
- Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P.-r., and Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with harmony. *Nature Methods*, pages 1–8.
- Krijthe, J., van der Maaten, L., and Krijthe, M. J. (2018). Package rtsne.
- Krijthe, J. H. (2015). *Rtsne: T-Distributed Stochastic Neighbor Embedding using Barnes-Hut Implementation*. R package version 0.15.
- Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27.
- Lee, J. A. and Verleysen, M. (2005). Nonlinear dimensionality reduction of data manifolds with essential loops. *Neurocomputing*, 67:29–53.
- Leek, J. T. and Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*, 3(9):e161.

- Linderman, G. C. and Steinerberger, S. (2019). Clustering with t-sne, provably. *SIAM Journal on Mathematics of Data Science*, 1(2):313–332.
- Luecken, M. D. and Theis, F. J. (2019). Current best practices in single-cell rna-seq analysis: a tutorial. *Molecular Systems Biology*, 15(6).
- Lun, A. T., McCarthy, D. J., and Marioni, J. C. (2016). A step-by-step workflow for low-level analysis of single-cell rna-seq data with bioconductor. *F1000Research*, 5.
- Maaten, L. v. d. (2014). Accelerating t-sne using tree-based algorithms. *The Journal of Machine Learning Research*, 15(1):3221–3245.
- Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605.
- Machold, R. and Fishell, G. (2005). Math1 is expressed in temporally discrete pools of cerebellar rhombic-lip neural progenitors. *Neuron*, 48(1):17–24.
- Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214.
- Mao, J., Ligon, K. L., Rakhlin, E. Y., Thayer, S. P., Bronson, R. T., Rowitch, D., and McMahon, A. P. (2006). A novel somatic mouse model to survey tumorigenic potential applied to the hedgehog pathway. *Cancer Research*, 66(20):10171–10178.
- McCarthy, D. J., Campbell, K. R., Lun, A. T., and Wills, Q. F. (2017). Scater: pre-processing, quality control, normalization and visualization of single-cell rna-seq data in r. *Bioinformatics*, 33(8):1179–1186.
- McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.
- Ocasio, J., Babcock, B., Malawsky, D., Weir, S. J., Loo, L., Simon, J. M., Zylka, M. J., Hwang, D., Dismuke, T., Sokolsky, M., et al. (2019). scrna-seq in medulloblastoma shows cellular heterogeneity and lineage expansion support resistance to shh inhibitor therapy. *Nature Communications*, 10(1):1–17.
- Risso, D., Ngai, J., Speed, T. P., and Dudoit, S. (2014). Normalization of rna-seq data using factor analysis of control genes or samples. *Nature Biotechnology*, 32(9):896.
- Roweis, S. T. and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326.

- Rubin, L. L. and de Sauvage, F. J. (2006). Targeting the hedgehog pathway in cancer. *Nature Reviews Drug Discovery*, 5(12):1026.
- Tenenbaum, J. B., De Silva, V., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323.
- Vieth, B., Parekh, S., Ziegenhain, C., Enard, W., and Hellmann, I. (2019). A systematic evaluation of single cell rna-seq analysis pipelines. *bioRxiv*, page 583013.
- Vladoiu, M. C., El-Hamamy, I., Donovan, L. K., Farooq, H., Holgado, B. L., Ramaswamy, V., Mack, S. C., Lee, J. J., Kumar, S., Przelicki, D., et al. (2018). Childhood cerebellar tumors mirror conserved fetal transcriptional programs. *bioRxiv*, page 350280.
- Wagner, A., Regev, A., and Yosef, N. (2016). Revealing the vectors of cellular identity with single-cell genomics. *Nature Biotechnology*, 34(11):1145.
- Wolf, F. A., Hamey, F. K., Plass, M., Solana, J., Dahlin, J. S., Göttgens, B., Rajewsky, N., Simon, L., and Theis, F. J. (2019). Paga: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biology*, 20(1):59.
- Zappia, L., Phipson, B., and Oshlack, A. (2018). Exploring the single-cell rna-seq analysis landscape with the scrna-tools database. *PLoS Computational Biology*, 14(6):e1006245.
- Zurawel, R. H., Allen, C., Chiappa, S., Cato, W., Biegel, J., Cogen, P., de Sauvage, F., and Raffel, C. (2000). Analysis of ptch/smo/shh pathway genes in medulloblastoma. *Genes, Chromosomes and Cancer*, 27(1):44–51.