Check for updates

METHOD ARTICLE

# REVISED netSmooth: Network-smoothing based imputation for single cell RNA-seq [version 3; referees: 2 approved]

Jonathan Ronen, Altuna Akalin

Scientific Bioinformatics Platform, Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine, Berlin, 13125, Germany

**v3** **First published:** 03 Jan 2018, **7**:8 (doi: 10.12688/f1000research.13511.1)
**Second version:** 24 Jan 2018, **7**:8 (doi: 10.12688/f1000research.13511.2)
**Latest published:** 10 Jul 2018, **7**:8 (doi: 10.12688/f1000research.13511.3)

## Abstract

Single cell RNA-seq (scRNA-seq) experiments suffer from a range of characteristic technical biases, such as dropouts (zero or near zero counts) and high variance. Current analysis methods rely on imputing missing values by various means of local averaging or regression, often amplifying biases inherent in the data. We present netSmooth, a network-diffusion based method that uses priors for the covariance structure of gene expression profiles on scRNA-seq experiments in order to smooth expression values. We demonstrate that netSmooth improves clustering results of scRNA-seq experiments from distinct cell populations, time-course experiments, and cancer genomics. We provide an R package for our method, available at: https://github.com/BIMSBbioinfo/netSmooth.

## Keywords

scRNA-seq, single-cell, genomics, imputation, networks

This article is included in the RPackage gateway.

### Open Peer Review

**Referee Status:** ✓ ✓

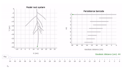
| | Invited Referees<br>1 | <br>2 |
|---|---|---|
| REVISED<br>**version 3**<br>published<br>10 Jul 2018 | | |
| UPDATE<br>**version 2**<br>published<br>24 Jan 2018 | ✓<br>report | ✓<br>report |
| **version 1**<br>published<br>03 Jan 2018 | | |

1 **Fernando J. Calero-Nieto**, Addenbrooke's Hospital, UK
**Fiona Kathryn Hamey**, Addenbrooke's Hospital, UK

2 **Siddharth Dey**, University of California, Santa Barbara, USA

### Discuss this article

Comments (0)



This article is included in the Interactive Figures collection.

**Corresponding author:** Altuna Akalin (altuna.akalin@mdc-berlin.de)

**Author roles: Ronen J**: Data Curation, Formal Analysis, Methodology, Software, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Akalin A**: Conceptualization, Methodology, Resources, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

**Competing interests:** No competing interests were disclosed.

**How to cite this article:** Ronen J and Akalin A. **netSmooth: Network-smoothing based imputation for single cell RNA-seq [version 3; referees: 2 approved]** *F1000Research* 2018, **7**:8 (doi: 10.12688/f1000research.13511.3)

**Grant information:** AA and JR are funded by core funding from Max Delbrück Center, part of Helmholtz Association.
*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

**First published:** 03 Jan 2018, **7**:8 (doi: 10.12688/f1000research.13511.1)

**REVISED** **Amendments from Version 2**

The updated version addresses the suggestions that were made by the reviewers. Some of the text and figures have been improved for clarity. Figures 2, S3 and S6 are amended to include the heatmap color scales. We have re-created Figure 2A to show proper clustering of single-cells. A new panel added to Figure 8 showing adjusted mutual information obtained from random networks vs real network. Figures S9 to S14 added to show performance of CIDR comparison to other imputation methods.

**See referee reports**

## Introduction

Single cell RNA sequencing (scRNA-seq) enables profiling of single cells' transcriptomes at unprecedented throughput and resolution. It has enabled previously impractical, studies of cell type heterogeneity, differentiation, and developmental trajectories[1]. However, the adaptation of RNA sequencing techniques from bulk samples to single cells did not progress without challenges. Typically, only a fraction of a cells transcriptome may be captured by the experiment, leading to so called "drop-out" events where a gene gets a false 0 (or near 0) count in some cell. The dropout rate is related to the population level expression of a gene leading to many false zero counts for lowly expressed genes, and artificially low counts for highly expressed ones[2]. Furthermore, the drop-out rate could be related to the biology of the cell type, as some cell types transcribe fewer genes than others, which will appear as drop-out events[2]. When summed over many samples, transcript counts from single cells resemble those of bulk experiments[3], but across individual cells there is significant variation. This makes analysis more difficult than in bulk RNA sequencing experiments.

Computational methods designed to deal with these issues treat dropout events as missing data points, whose values may be imputed based on non-missing data points (observed measurements). The proportion of 0 counts per gene, a proxy for its technical dropout rate, is a function of the population-wise mean expression of that gene[2,4]. This observation has led researchers to treat 0 counts as dropout candidates to be imputed.

CIDR[5] attempts to impute missing values based on the predicted mean expression of a gene, given its empirical dropout rate (0-count). scImpute[6] estimates dropout likelihoods per gene and per sample, and assigns each gene in each sample a status as a dropout candidate. Genes might be considered likely dropouts even with nonzero expression, and 0-count genes might not be considered likely dropouts, based on their population-wide expression distributions. It then uses a regularized linear model to predict the expression of dropout genes based on the expression of likely non-dropouts in all other cells. MAGIC[7] performs local averaging after building a topological graph of the data, updating the expression value of all genes in all cells to their local neighborhood average.

All of the methods mentioned above use measured information in the data in order to impute the missing information within the same data. As such, they amplify whatever biases are present in a dataset; similar cells pre-imputation will become more similar after imputation, as expression profiles of non-dropout genes will drive similarities in imputed dropped-out genes. Further, all methods except MAGIC only impute unobserved expression events (0s or near 0s), while the dropout phenomenon actually affects the whole transcriptome. Hence, imputation methods for scRNAseq should also adjust non-0 expression measurements in order to recover the true signal.

We present a method, called *netSmooth*, that uses prior knowledge to temper noisy experimental data. RNA sequencing experiments produce counts data as a proxy for gene activity, which is not known a-priori, especially for experiments profiling unknown cell types. However, decades of molecular biology research have taught us much about the principles of gene interaction. Interacting genes are likely to be co-expressed in cells[8,9], and as such, protein-protein interaction (PPI) databases[10,11] describe genes' propensity for co-expression. We developed a graph-diffusion method on PPI networks for smoothing of gene expression values. Each node in the graph (a gene) has an associated gene expression value, and the diffusion presents a weighted averaging of gene expression values among adjacent nodes in the graph, within each cell. This is done iteratively until convergence, strengthening co-expression patterns which are expected to be present. Incorporation of prior data from countless experiments in the preprocessing of scRNA-seq experiments improves resistance to noise and dropouts. Similar network based approaches have been used to extract meaningful information from sparse mutational profiles[12,13], and indirectly on gene expression data by diffusing test statistics on the network to discover regulated gene candidates[14]. We propose diffusion of gene expression values directly on the network as a method for data denoising and imputation. Furthermore, the parameters of this proposed method could be optimized using clustering robustness metrics. We applied our method to a variety of single cell experiments and compared its performance to other selected imputation methods scImpute and MAGIC. These methods represent the latest and divergent ways of imputing the scRNA-seq data.

While we mention CIDR in this review, we do not include comparisons to CIDR in the main text, alongside MAGIC and scImpute, because CIDR uses its own clustering procedure as a part of the imputation workflow. scImpute and MAGIC are agnostic about post-imputation analysis, and therefore we were able to compare them to *netSmooth* using a unified analysis framework (see Methods). For completeness, we include benchmark results of CIDR in the supplement (Figures **??** - **??**).

We also made available an R package providing the necessary functionality to use our method on other data. It is available on GitHub: https://github.com/BIMSBbioinfo/netSmooth, or using Bioconductor: https://bioconductor.org/packages/release/bioc/html/netSmooth.html.

## Results

### Overview of the method

The intuition behind the *netSmooth* algorithm is that gene networks encoding co-expression patterns can be used to smooth scRNA-seq data, pushing its coexpression patterns in a biologically meaningful direction. We demonstrate this using protein-protein interaction networks, which are predictive of coexpression[9]. We produced a PPI graph of high-confidence interactions based on the PPI database STRING[10].

There are 2 inputs to the method: (1) a gene expression matrix, $N$ genes by $M$ cells, and (2) a graph where genes are nodes, and edges indicate genes which are expected to be co-expressed. The edges may be weighed, indicating the strength or direction of a relationship; an edge weight of 2 indicates stronger expected co-expression than an edge weight of 1, and an edge weight of $-1$ indicates negative expected co-expression, such as one gene being a repressor for another. The expression profile of each cell is then projected onto the graph, and a diffusion process is used to smooth the expression values, within each sample, of adjacent genes in the graph (Figure 1). In this way, post-smoothing values of genes represent an estimate of activity levels based on reads aligned to that gene, as well as those aligned to its neighbors in the graph. Thus, a gene with a low read count (possible technical drop-out), whose neighbors in the graph are highly expressed, will get a higher value post smoothing. The rate at which expression values of genes diffuse to their neighbors is degree-normalized, so that genes with many edges will affect their neighbors less than genes with more specific interactions. The diffusion is done using a "random walks with restarts" (RWR) process[13], where a conceptual random walker starts in some node in the graph, and at each iteration moves to a neighboring node with a probability determined by the edge weight between the nodes, or, with some probability, restarts the walk from the original node. The *network-smoothed* value is the stationary distribution of this process. The RWR process has one free parameter, the restart rate. A low value for the restart rate allows diffusion to reach further in the graph; a high restart rate will lead to more local diffusions. For more details see the Methods section.

## Network smoothing improves cell type identification from single-cell RNA-seq

We first assess *netSmooth* on a dataset of 1645 mouse hematopoietic stem/progenitor cells (HSPCs) assayed using flow cytometry as well as scRNA-seq[15]. The cells are FACS-sorted into 12 common HSPC phenotypes. This presents an atlas of the hematopoiesis process at a single cell resolution, showing the differentiation paths taken by E-SLAM HSCs as they differentiate to E, GM, and L progenitors. The authors of this study demonstrate that upon clustering the data, some clusters corresponds to cell types. However, the clusters are not noise free and do not fully recapitulate cell type identity. We obtained clusterings of the cells from the normalized counts, as well as after application of *netSmooth*, MAGIC[7], and scImpute[6], using a robust clustering procedure based on the *clusterExperiment* R package[16] (See Methods). After clustering, we used the edgeR-QLF test[17] to identify genes that are differentially expressed in any of the discovered clusters. Figure 2a,b shows the log-transformed expression values of the 500 most differentially expressed genes, before and after application of *netSmooth*. The column annotations indicate the FACS-sorted cell type of each cell, as well as the cluster assignment obtained from the *netSmooth* R package. The figure suggests that the network-smoothing effect is subtle on the individual genes, as difference between the heatmaps is negligible visually. Figure 2c,d shows the same for the MAGIC and scImpute-preprocessed data, respectively. MAGIC seems to do the strongest transformation to the data, as is also seen in lower dimension embeddings (Figures ??, ??).

As this dataset has cells with labels independent of the RNAseq (FACS-sorted phenotypes), it presents us with an opportunity to compare the gene expression levels (as measured by RNAseq), to a meaningful phenotypic variable, i.e. the cell type. The cell type discrimination of a clustering result is compared using a cluster purity metric and and the adjusted mutual information (AMI). The cluster purity measures how cell-type specific clusters are by comparing homogeneity of the external labels (FACS-defined cell types), within clusters provided by scRNA-seq data. AMI is a chance-adjusted information theoretic measure of agreement between two labellings. This method accounts for artificially high mutual information between external labels and clusters when there is large number of clusters (See Methods for details on metrics). We also measured number of cells in robust clusters as quantitative metric. The robust clustering

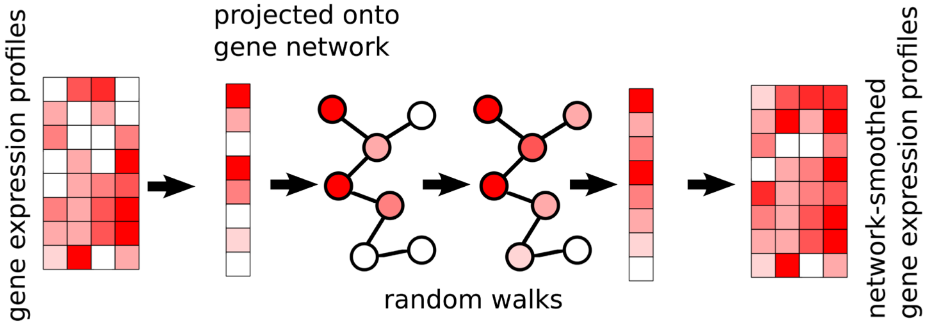


**Figure 1. The *netSmooth* algorithm takes a gene expression profile, and a gene network.** The expression profile of each sample is projected onto the network, where a diffusion process allows genes' expression values to be smoothed by their neighbors'. This is done for each cell independently of others. The end result is a network smoothed gene expression matrix.

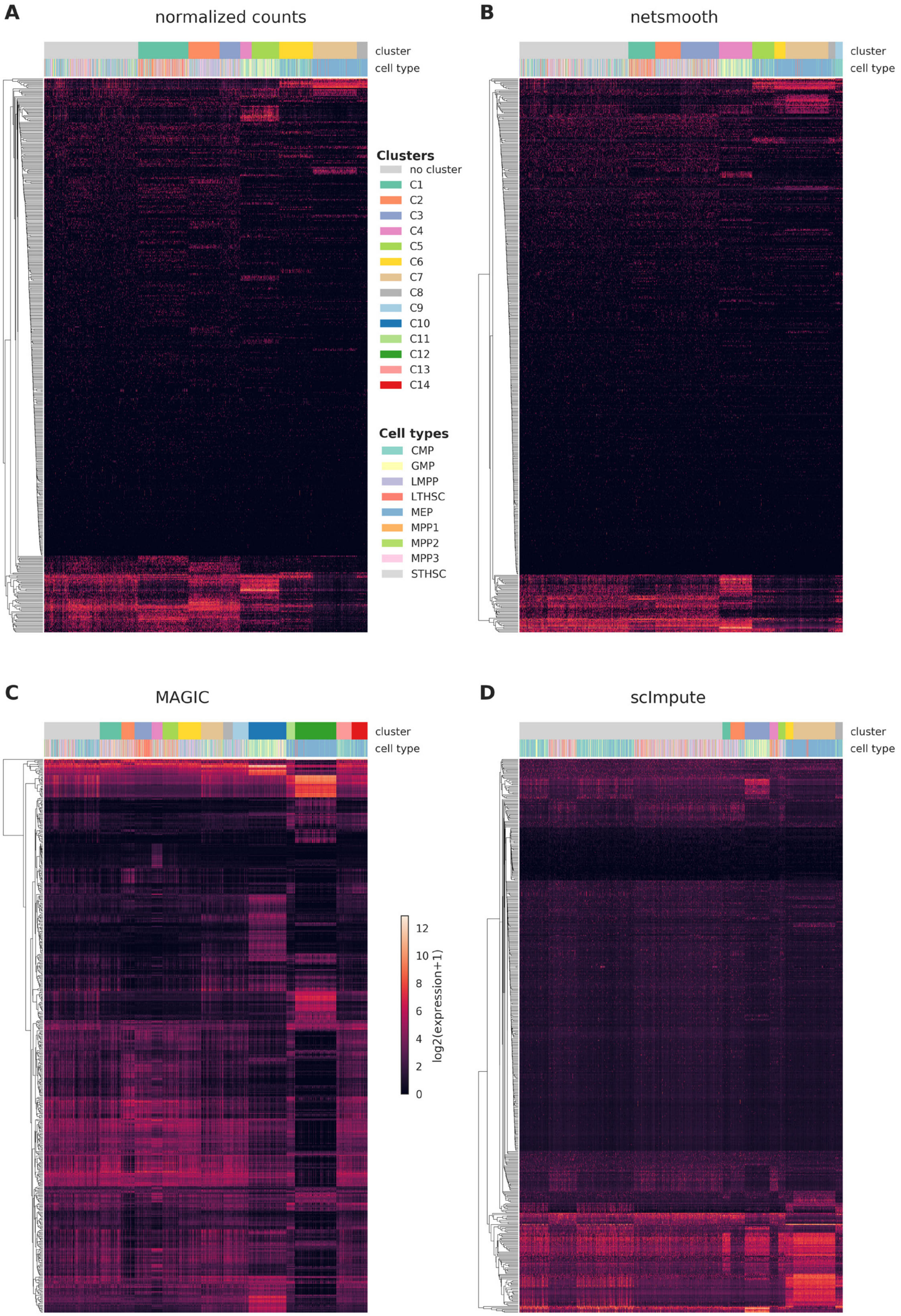


**Figure 2. Cells were clustered using the robust clustering procedure, and the log-transformed expression values of the log-transformed expression values of the 500 most differentially expressed genes (by edgeR-QLF test adjusted P value) in any of the discovered clusters are shown in a heatmap, as well as cluster assignments and FACS-sorted cell types. A**) raw (no imputation), **B**) after application of *netSmooth*, **C**) missing values imputed using MAGIC **D**) missing values imputed using scImpute.

procedure allows cells to be omitted (not be assigned to a cluster) if they cannot be placed in a cluster across multiple clustering methods and/or parameters (See Methods). MAGIC-processed data leads to a larger proportion of cells assigned to robust clusters, while *netSmooth* and scImpute lead to a reduction in the clustering robustness metric (Figure 3a). All three methods are able to discover some novel clusters in the data with high purity (Figure 3b). A closer inspection shows that MAGIC achieves this through a proliferation of small clusters, which are not so far as we can judge meaningful, as evidenced by the lower adjusted mutual information score (Figure 3c). *netSmooth*-preprocessed clusterslead to a higher AMI score, which, while modest, is biologically relevant.

## Network smoothing improves capture of developmental expression patterns

Next, we test *netSmooth* on 269 isolated cells from mouse embryos at different stages of pre-implantation development between oocyte and blastocyst, as well as 5 liver cells and 10 fibroblast cells[18]. The authors of this study demonstrated that lower dimension embeddings capture much of the developmental trajectory (Figure 4a, Figure ??a, ??a). We then applied *netSmooth*, MAGIC, and scImpute. Figure 4b shows the principal component analysis of *netSmooth*-processed data, and Figures 4c and 4d show the PCA plot following application of MAGIC and scImpute, respectively. *netSmooth* and scImpute preserve most of the variance structure of the data, while MAGIC seems to push the data onto a completely different manifold (Figure 4, Figure ??). We used the robust clustering procedure to obtain clusters, and computed the cluster purity and AMI metrics. *netSmooth* enabled the clustering procedure to place more of the samples into robust clusters (Figure 5a), and as in the hematopoiesis case, *netSmooth* is able to assist in identifying the developmental stage or tissue that cells belong to better than the other methods, as evidenced by the higher cluster purities (Figure 5b) and AMI (Figure 5c). scImpute also improves the cluter purity and AMI metrics (Figure 5b,c), and is not easily differentiable from *netSmooth* in the PCA scatter plot (Figure 4). The *netSmooth* results are marginally better, which hints at an equivalence in the recovered signal quality between the two methods, *netSmooth*'s quasiimputation incorporating priors, and scImpute's linear model-based imputation. scImpute achieves this by reducing the overall 0-count genes significantly more than *netSmooth* (Figure ??), which suggests that incorporating priors the way *netSmooth* does can achieve similar results to data-imputation. The smaller change in the proportion of 0-count genes following *netSmooth* than scImpute (Figure ??) shows that *netSmooth* works primarily by smoothing values of genes with measured expression, as opposed to imputing suspected missing counts, which suggests a lesser transformation of the data, such as through application of *netSmooth*, can uncover much of the true signal hidden in the noisy data.

## Network smoothing improves identification of glioblastoma tumors

Finally, we demonstrate applicability of *netSmooth* to cancer research. Patel et al. generated scRNA-seq data of 800 cells from 5 glioblastoma tumors and 2 cell lines[19]. Lower dimension embedding plots show that cells from different tumors or cell lines generally group together, but some are not wholly distinguishable from other tumors (Figure 6a, ??a, ??a). Further, the two cell lines group closer to each other than the other patient samples. After applying *netSmooth* to the data, tumors become easier to distinguish in a lower dimensional embedding (Figure 6b), indicating that *netSmooth* improves assignment of each cell to its tumor, cell line, or clone of origin. Again, scImpute also leads to similar reduced dimension embedding (Figure 6d), while MAGIC distorted the data more than the other methods (Figure 6c). We used the robust clustering procedure before and after *netSmooth*, MAGIC, and scImpute. Only MAGIC increase the clusterabitliy of the data (Figure 7a), but *netSmooth* leads to the most pure clusters, in terms of tumor or cell line of origin (Figure 7b, Figure 7c).

Tumor or cell line of origin is an imperfect proxy for phenotypical variation in cancer cells, because some cells cluster by cell type rather than tumor of origin, demonstrating the heterogeneity in these glioblastoma tumors and similarities across origins[19]. Nevertheless, we chose to compute cluster purity based on the cell origin rather than other labels which might be assigned to them, as it is the only *ground truth* variable that is independent of the RNAseq experiment. Further, cells do group by origin (Figure 6, Figure ??), and identification of origin is an interesting question in its own right in the field of cancer genomics, particularly for heterogeneous tumors such as these.

## Sensitivity to the network

Next, we set out to ensure that the results are not an artifact of the network structure, i.e. that the actual links between genes that we used in the network are important. We expect *netSmooth* not to perform well when using networks with similar characteristics, but where edges do not represent real interactions. To that effect, we constructed 20 random networks by keeping the same graph structure of the real PPI graph, but shuffling the gene names. Thus, these random networks share all the characteristics of the real network (degree distribution, community structure), except for the true identity of the nodes. We then used those networks as inputs to *netSmooth* and ran the benchmarks as before on the hematopoiesis dataset. Using random networks as an input to *netSmooth* gives cluster purities distributed around a mode given by the cluster purities of the raw data, while the cluster purities given from using the real PPI network lie at the extreme edge of the distribution (Figure 8a). Further, most random networks result in fewer samples belonging to robust clusters (Figure 8b). Finally, we also calculated the adjusted mutual information of clusterings resulting from the randomized networks (Figure 8c). Again, most shuffled networks produce worse clusterings, with the real network outperforming all of them, as well as the no-smoothing case. These results demonstrate that it is indeed the information contained in the PPI graph enables netSmooth to transform the gene expression matrix in a more biologically coherent direction, and that the transformation we see can not be explained simply by the network structure.

## Using other networks with netSmooth

In addition to using an unweighed (where all edge weights are 1), undirected (where all edge weights are positive) network

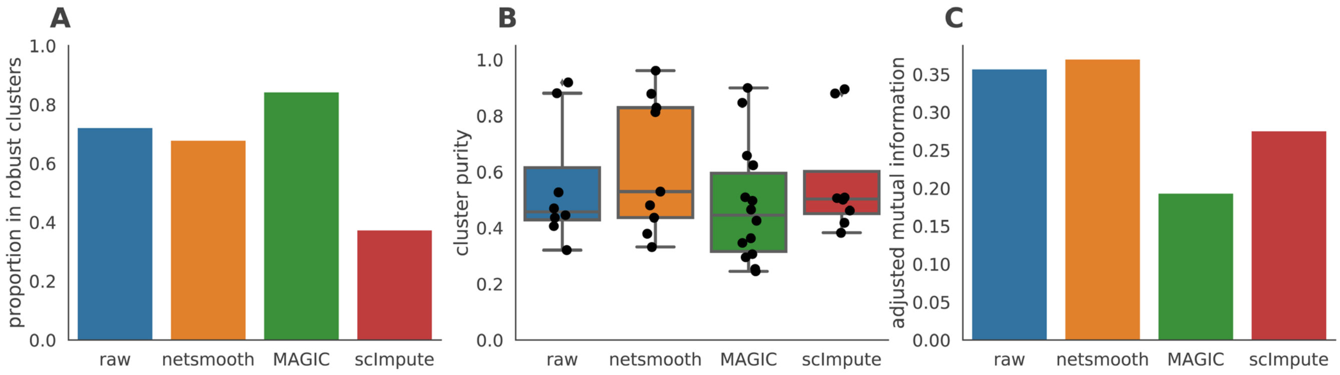


**Figure 3. Hematopoiesis clustering metrics. A**) The proportion of cells which were assigned to robust clusters. **B**) cluster purity (proportion of dominant cell type) for the robust clusters. *netSmooth* produces the most pure clusters in terms of cell types. **C**) AMI of the clustering results obtained after application of each of the methods. Only *netSmooth* increases the AMI between the clustering and the cell types. The online version of this figure is interactive.

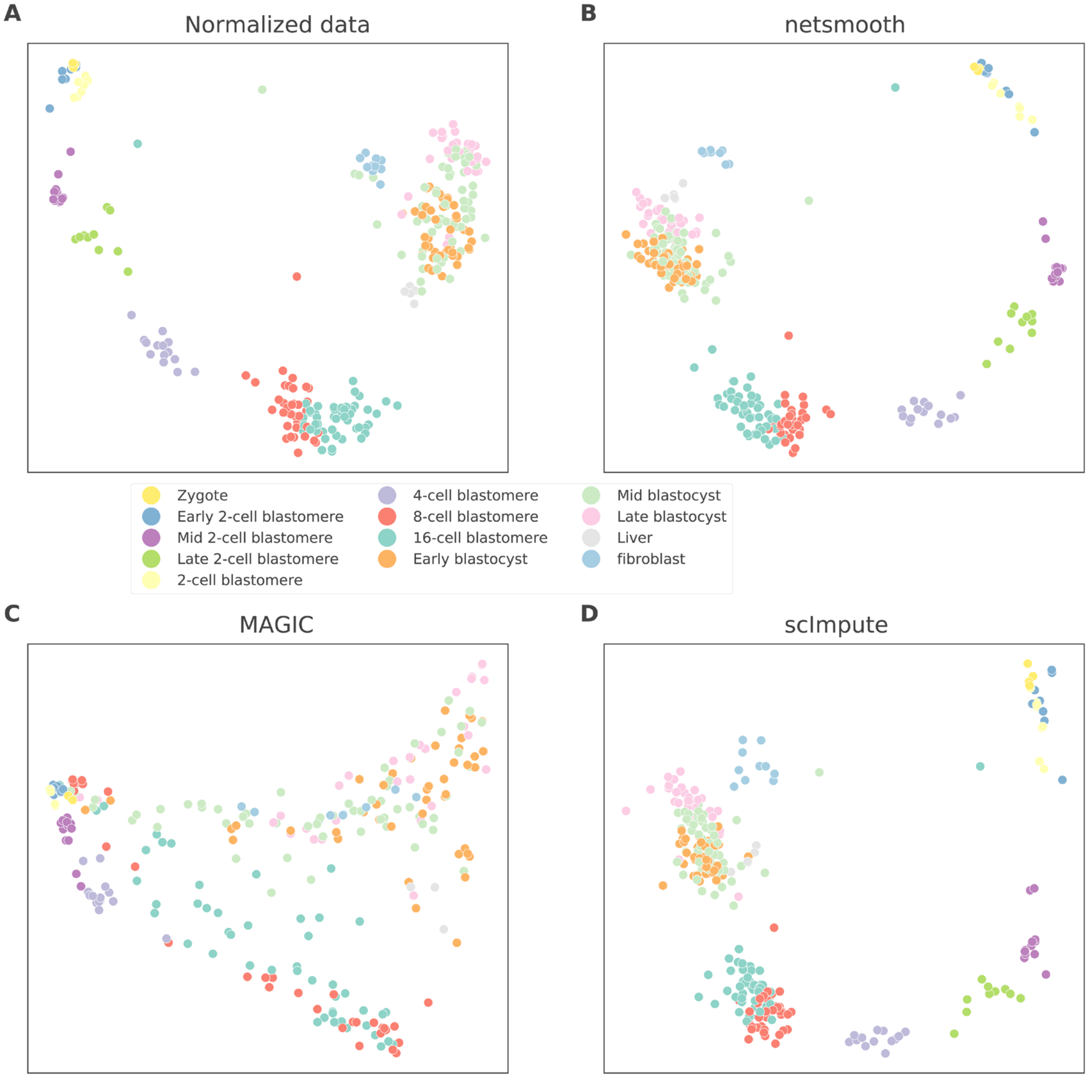


**Figure 4.** 2D PCA plots of the embryonic development dataset **A**) no preprocessing, **B**) after application of *netSmooth,* **C**) after imputing missing values with scImpute, and **D**) after application of MAGIC. The online version of this figure is interactive.

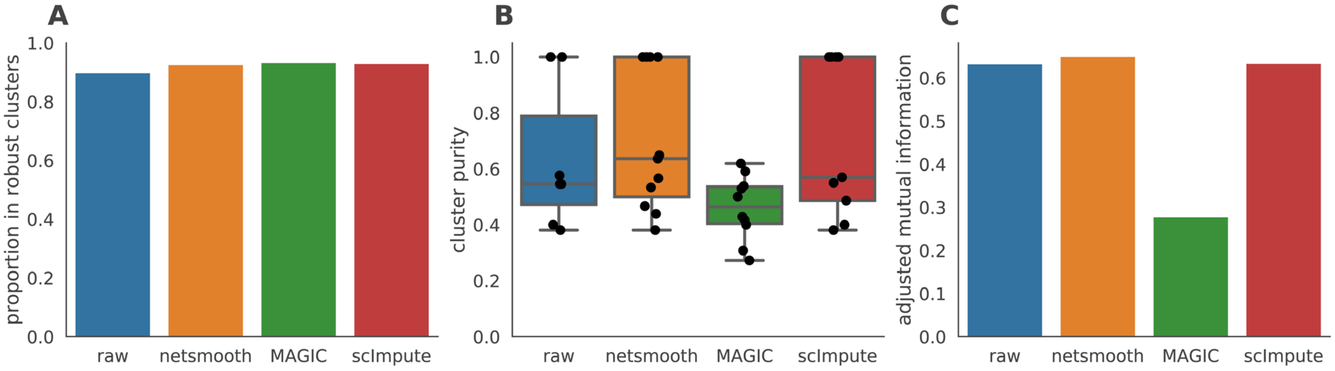


**Figure 5. The Embryonic development dataset. A**) The proportion of cells which were assigned to robust clusters. All three methods lead to better clusterability, with MAGIC having the strongest effect. **B**) cluster purity (proportion of dominant cell type) for the robust clusters. *netSmooth* produces the most pure clusters in terms of cell types. **C**) Adjusted mutual information of clusterings and cell types. Only *netSmooth* increases the AMI over the non-preprocessed data. The online version of this figure is interactive.

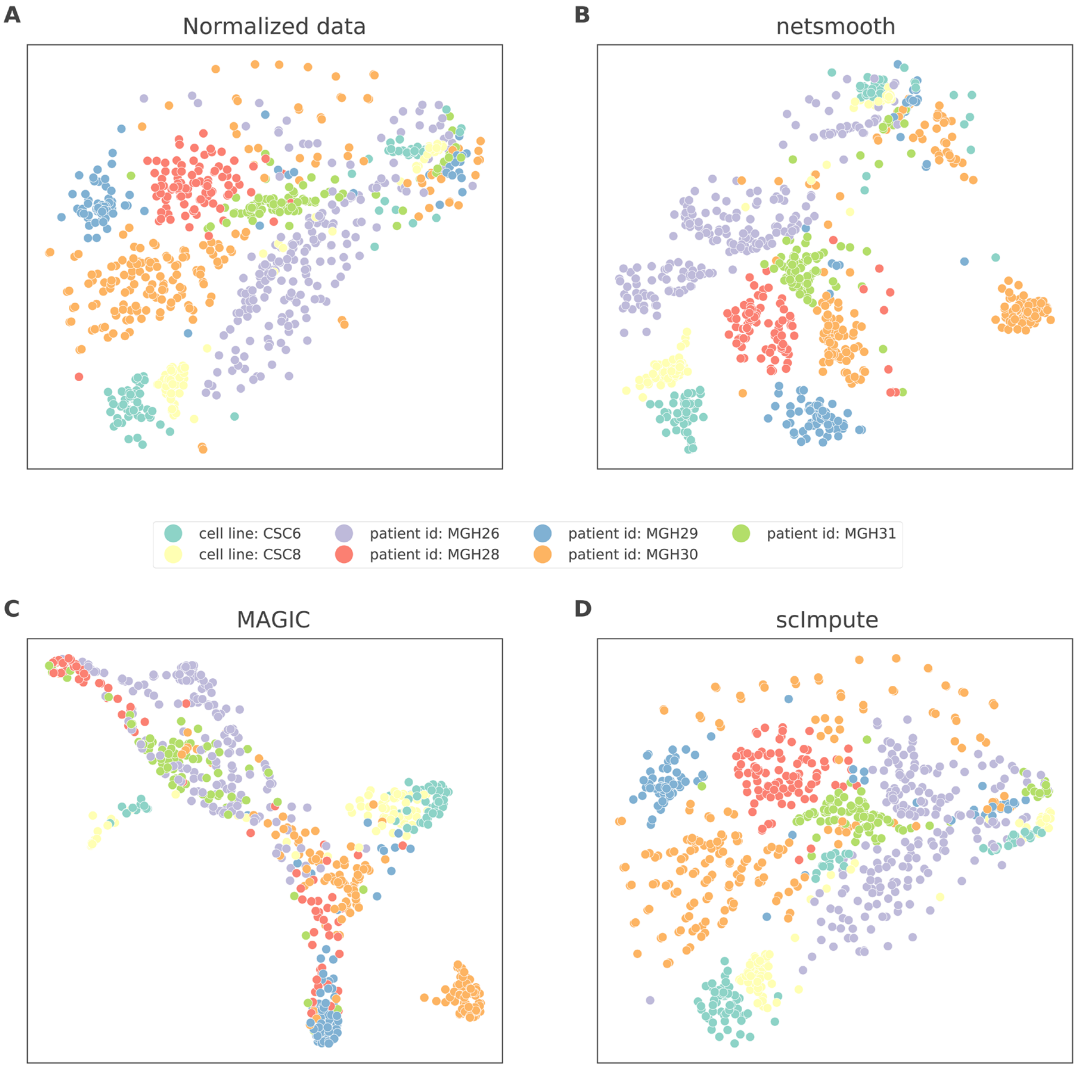


**Figure 6.** t-SNE plots of the glioblastoma dataset **A**) no preprocessing, **B**) after application of *netSmooth*, **C**), using MAGIC, and **D**) after application of scImpute. The online version of this figure is interactive.

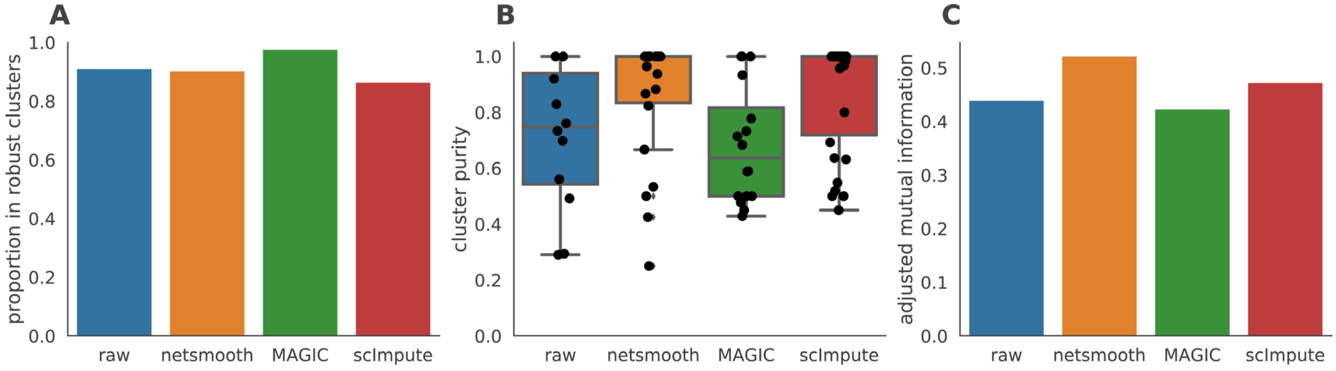


**Figure 7. Imputation performance for the glioblastoma dataset. A**) The proportion of cells which were assigned to robust clusters. *netSmooth*, MAGIC, and scImpute all increased the proportion of cells that are assigned to robust clusters, with MAGIC leading, *netSmooth* in second place, and scImpute in third. **B**) cluster purity (proportion of dominant cell type) for the robust clusters. *netSmooth* produces the most pure clusters in terms of tumor or cell line of origin. **C**) AMI of the clustering results obtained after application of each of the methods. The online version of this figure is interactive.

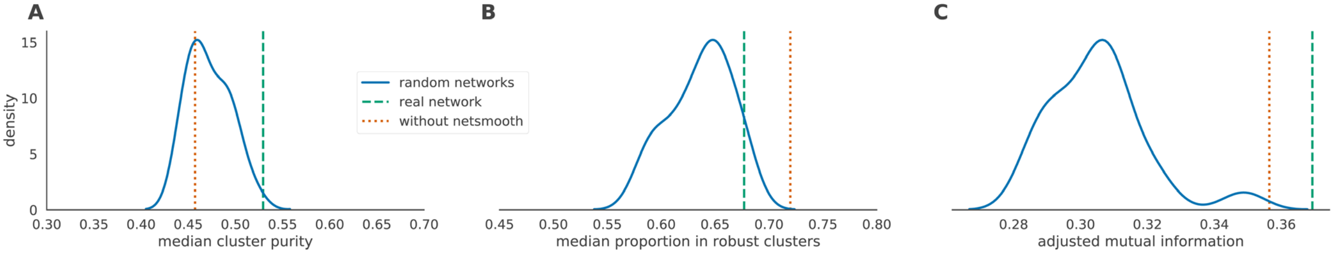


**Figure 8. Performance of *netSmooth* with randomized networks. A**) The median cluster purity achieved with the random networks. The real network outperforms the random ones, which result in cluster purities distributed around the purity given without using *netSmooth*. **B**) The proportion of samples assigned to robust clusters using the random networks as well as the real one. While all networks result in fewer samples robustly clustered (in the hematopoiesis dataset), the real network outperforms most random networks. **C**) The Adjusted Mutual Information achieved with the randomized networks. Most random networks produce clusterings with a worse AMI than using no network-smoothing. netSmooth with the real network structure produces the clustering result with the best AMI. The online version of this figure is interactive.

from string-db, we constructed other gene networks and used them as inputs to *netSmooth*. We created a directed gene network from only those edges in string-db which are marked as activating or inhibiting[i]. We set the edge weights of the activating interactions to +1, and −1 for the inhibiting interactions, allowing gene expression values to be adjusted downwards for genes whose known antagonists are highly expressed. After smoothing, we set all negative smoothed expression values to 0. We also constructed a gene network from string-db using only genes that are known to demonstrate cell-type specific expression. In order to obtain a list of genes with such cell-type specific expression patterns from the *Expression Atlas*[20], we used only the genes which show a cell-type specific expression with a mean TPM of at least 1 in some cell type, and used the subset of string-db network containing those genes as an input to *netSmooth*. Both of those modified graphs perform similarly to the undirected graph from string-db (Figure 9, Figure ??a, Figure ??b), demonstrating that *netSmooth* is able to use priors from different types of experiments in order to improve clustering of scRNA-seq.

We also considered other sources for the gene network. We constructed a gene network from HumanNet[21], a functional gene network where edges denote interactions between two genes. We constructed a smoothing graph by taking all edges from HumanNet, and producing a graph where all edge weights are set to 1. We then used this graph as an input to *netSmooth* on the glioblastoma dataset. It performs similarly to the network from string-db (Figure 10, Figure ??c), demonstrating

[i] Most interactions in string-db do not specify the direction, or nature of the interaction

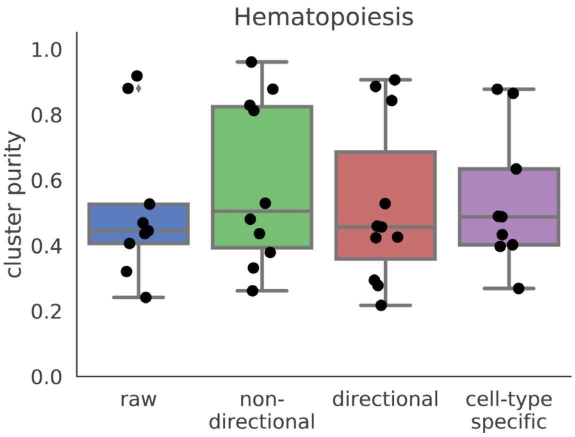


**Figure 9. Cluster purities after applying *netSmooth* with different input networks.** Raw refers to no smoothing, non-directional is the same as the results shown in previous sections. Directional refers to a gene network where inhibitory relationships have negative edge weights, and cell-type specific refers to a gene network of only genes which are known to have cell-type specific expression patterns. The online version of this figure is interactive.

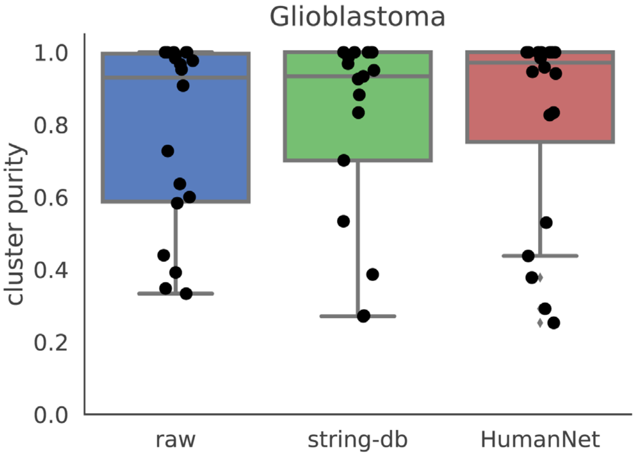


**Figure 10. Cluster purities after applying *netSmooth* with different input networks.** Raw refers to no smoothing, string-db is the same as the results shown in previous sections, and HumanNet refers to a gene network constructed from the HumanNet database. The online version of this figure is interactive.

that other sources for gene interactions may also be used by *netSmooth* to improve clustering results of scRNA-seq.

As more scRNAseq experiments are published, context-specific networks will be made possible to create, potentially improving *netSmooth*'s performance. The networks we have shown above have links between genes which are known in a general context, but scRNAseq experiments might uncover previously unknown cell-type specific gene interactions, which could contribute to the information uncovered by network smoothing. Here, we have demonstrated that even the general context networks we have used are able to assist in identifying specific cell types from noisy scRNAseq datasets.

### Optimizing the smoothing parameters by cluster robustness

The *netSmooth* algorithm, given a gene network, has one free parameter - the restart rate of the random walker, $(1 - \alpha)$. Alternatively, $\alpha$ is the complement of the restart rate. An $\alpha = 0$ indicates a perfect restart rate and consequently no smoothing; an $\alpha = 1$ corresponds to a random walk without restarts. Intermediate values for $\alpha$ result in increasing levels of smoothing; the value of $\alpha$ determines how far random walks will go on the graph before restarting, or how far along the network a gene's influence is allowed to reach (See Methods). It is tempting to optimize $\alpha$ with respect to the variable the experiment sets out to measure, e.g. cluster purity. For instance, in the embryonic development dataset, we would choose $\alpha = 0.4$ as the value that produces the highest cluster purity (Figure 11b). However, in many experiments the identity of the samples is not known a-priori. Therefore, we propose a data driven workflow to pick a sensible value for $\alpha$.

One such data-driven statistic is the proportion of samples assigned to robust clusters; following application of *netSmooth*, the robust clustering procedure is able to assign more samples to statistically robust clusters. For two of the three datasets, picking the $\alpha$ that gives the highest proportion of cells in robust clusters, also gives the clusters with the highest purity index (Figure 12). Importantly, this metric is entirely data-driven and does not require external labels, making it feasible for any scRNA-seq study. The results in the previous sections all use the value of $\alpha$ picked to optimize proportion in robust clusters.

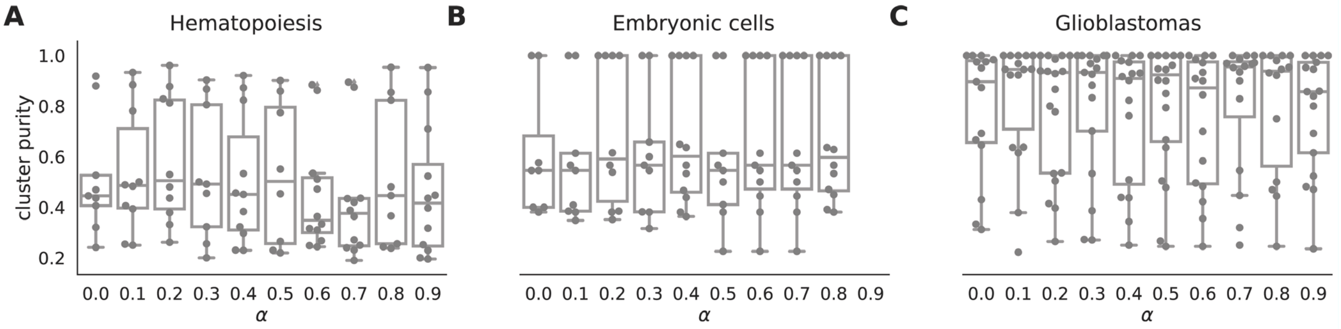


**Figure 11. boxplots of cluster purity for clusters obtained by the robust clustering procedure following application of *netSmooth* with different values of *α*.** *α* = 0 is equivalent to not using *netSmooth* at all. The procedure is robust to alpha, that is, most values of alpha produce more robust clusters. **A**) HSPCs, **B**) embryonic cells, **C**) glioblastomas. The online version of this figure is interactive.

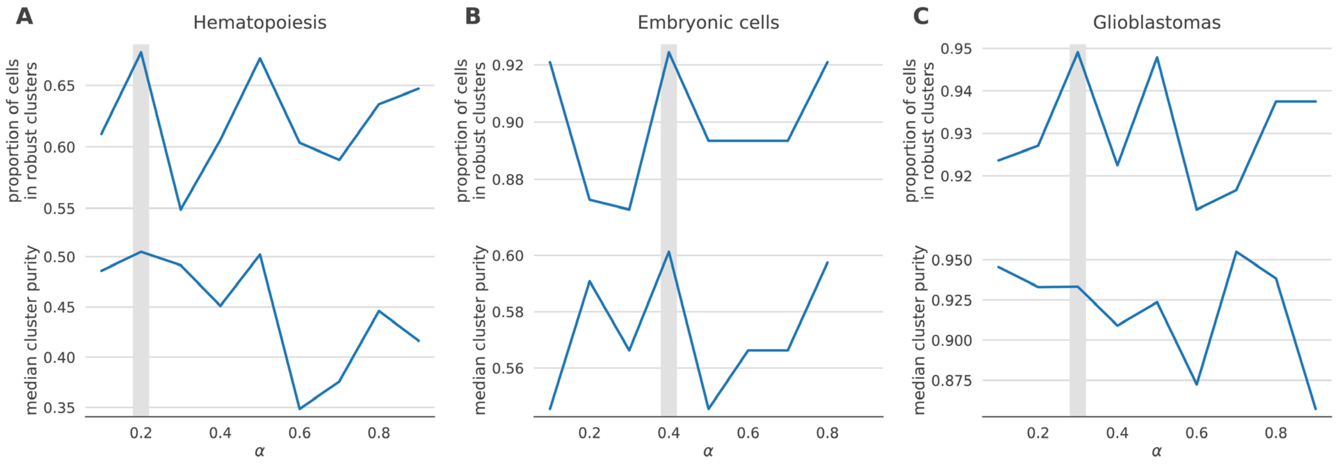


**Figure 12. the proportion of cells in robust clusters, and cluster purity for those robust clusters, for a range of alpha values, shows that picking the alpha with the highest proportion in robust clusters also picks the alpha with the highest cluster purity. A**) hematopoietic stem/progenitor cells **B**) embryonic cells, **C**) glioblastomas. The online version of this figure is interactive.

The *netSmooth R package*[22] provides an alternative way to optimize *α* in the absence of true labels, by optimizing the entropy in a 2D embedding of the data. See Methods for details.

## Discussion

Single cell RNA sequencing technology provides whole-genome transcriptional profiles at unprecedented throughput and resolution. However, high variance and dropout events that happen in all current scRNA-seq platforms complicate the interpretation of the data. Methods that treat 0 counts as missing values and impute them based on nonzero values in the data may amplify biases in the data.

We presented *netSmooth* as a preprocessing step for scRNA-seq experiments, overcoming these challenges by the use of prior information derived from protein-protein interactions or other molecular interaction networks. We demonstrated that network smoothing assists in several standard analyses that are common in scRNA-seq studies. This procedure enhances cell type identification in hematopoiesis; it elucidates time series data and assists identification of the developmental stage of single cells. Finally, it is also applicable in cancer, improving identification of tumor of origin for glioblastomas. In addition, we showed that network smoothing parameter can be optimized by cluster robustness metrics, providing a workflow when there are no other external labels to distinguish cells. We demonstrated that *netSmooth* can use prior information from different sources in order to achieve this. We compared *netSmooth* with scImpute, a statistical genome-wide imputation method, and MAGIC, a genome-wide data smoothing algorithm, and demonstrated that while scImpute and MAGIC reduce the drop-out phenomenon more than *netSmooth* does, *netSmooth* outperforms them in amplifying the biological/technical variability ratio. *netSmooth* provides clusters that are more homogeneous and have higher adjusted mutual information (AMI) with respect to cell types. Although, in some cases data processed by MAGIC produces more robust clusters, the clusters returned after

MAGIC processing do not have higher AMI or cluster purity. Higher robustness achieved by MAGIC processing might be due to the fact that the algorithm reinforces local structures too much in the data and producing artificially similar expression profiles between cells. Comparisons to CIDR (Figures **??** - **??**) also show inferior performance to *netSmooth*.

In most of the benchmarks we ran, scImpute shows similar performance to *netSmooth*, while the former relies on other data points in order to impute missing data, and the latter performs a quasi-imputation based on priors from other experiments. Our analysis shows that *netSmooth* affects the drop-out rate less than scImpute, while uncovering slightly more of the biological signal. This happens across the different overall drop-out rates in the 3 experiments we profiled, indicating that *netSmooth* can achieve better results, with less obtrusive transformations of the data, then the imputation methods, across a range of experimental conditions.

Finally, *netSmooth* is a versatile algorithm that may be incorporated in any analysis pipeline for any experiment where the organism in question has a high quality PPI network available. Although not shown, the algorithm is applicable to any omics data set that can be constructed as a genes-by-samples matrix, such as proteomics, SNPs and copy number variation. In addition, most of the computational load of network smoothing can be done “off-line”. As such it scales well with the number of cells, which is likely to increase in future scRNA-seq experiments. We have made available an R package to that end, which is available on GitHub: https://github.com/BIMSBbioinfo/netSmooth, and Bioconductor: https://bioconductor.org/packages/release/bioc/html/netSmooth.html.

## Methods and data

### The data sets

The hematopoiesis dataset[15] was obtained from the Gene Expression Omnibus[23]. The embryonic[18] and glioblastoma[19] datasets were obtained from *conquer*[24], a repository of uniformly processed scRNA-seq datasets. We have made the datasets available, see Table 1.

### The random walks with restarts process

The *netSmooth* algorithm takes a graph $G = \{V, E\}$ where $V = \{gene_i\}$ is the set of genes, and $E = \{(i \rightarrow j)\}$ is the set of edges between genes. The edge weights are degree-normalized, so that each gene’s outgoing edges’ weights sum to 1. We then define a process of random walk with restarts as in 13, on the PPI graph, where a conceptual random walker starts on a node in the graph (a gene/protein) and at each step walks to an adjacent node with the probability determined by the $\alpha$ times the edge weight. Further, at each step, there is a probability of $(1 - \alpha)$ that the walker restarts to its original node.

**Table 1. Datasets and availability.**

| Dataset | URL |
|---|---|
| Hematopoiesis | https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE81682 |
| Embryonic cells | http://imlspenticton.uzh.ch/robinson_lab/conquer/data-mae/GSE45719.rds |
| Glioblastoma | http://imlspenticton.uzh.ch/robinson_lab/conquer/data-mae/GSE57872.rds |

Mathematically, given a graph defined by an adjacency matrix $A_{[MxM]}$, where $A_{ij}$ is the edge weight between gene $i$ and gene $j$ (and 0 for unconnected genes), and a vector $f_{[Mx1]}$, where $f_i^t$ is the probability that the walker is at node $i$ at step $t$, the process is defined by

$$f^{t+1} = \alpha A f^{t} + (1 - \alpha) f^{0}.$$

This process is convergent, and the stationary distribution is given by

$$f^{\infty} = (1 - \alpha)(I - \alpha A)^{-1} f^{0}.$$

Hence, the random walk with restarts process is a diffusion process defined on the PPI graph, or through the diffusion kernel (smoothing kernel)

$$K_A^{\alpha} = (1 - \alpha)(I - \alpha A)^{-1}$$

where $(1 - \alpha)$ is the restart probability, and $A$ is the (column normalized) adjacency matrix of the PPI graph. Consequently, we define the *network-smoothed* expression profile

$$E_{sm} = K_A^{\alpha} E,$$

where $E_{[MxN]}$ is the normalized count values of the $M$ genes in the $N$ cells.

### The clustering procedure

Clustering analysis features prominently in scRNA-seq analyses; whether recapitulating known results or discovering new cell types, clustering cells by their gene expression profiles is commonly used to identify distinct populations. While some approaches directly take into account the zero-inflation of scRNA-seq data[5], other studies use traditional methods[18]. There is no standard method for clustering single cell RNAseq data, as different studies produce data with different topologies, which respond differently to the various clustering algorithms.

In order to avoid optimizing different clustering routines for the different datasets we benchmark on, we have implemented a robust clustering routine based on *clusterExperiment*[ii] [16], a framework for robust clustering based on consensus clustering of clustering assignments obtained from different clustering algorithms, different parameters for these algorithms, and different views of the data. The different views are different reduced dimensionality projections of the data based on different techniques. Thus, no single clustering result will dominate the data, and only cluster structures which are robust to

[ii] Version 1.4.0, available from Bioconductor https://bioconductor.org/packages/release/bioc/html/clusterExperiment.html

different analyses will prevail. The procedure we implemented using the framework is as follows:

1. Perform different dimensionality reduction techniques on the data
   - PCA on the 500 most variable genes
     - with 5 components
     - with 15 components
     - with 50 components
   - Alternatively to PCA, t-SNE on the 500 most variable genes
     - with 2 dimensions
     - with 3 dimensions
   - Select the most variable genes
     - 100 most variable genes
     - 500 most variable genes
     - 1000 most variable genes
2. On each reduced dimension view of the data, perform PAM clustering with K ranging from 5 to 10
3. Calculate the co-clustering index for each pair of samples (the proportion of times the samples are clustered together, in the different clustering results based on the different reduced dimensions and clustering parameters above)
4. Find a consensus clustering from the co-clustering matrix. This is done by constructing a dendrogram using average linkage, and traversing down the tree until a block with a self-similarity of at least 0.6, and a minimum size of 20 samples emerges. (instead of using `cutree`).
5. Perform hierarchical clustering of the cluster medioids, with similarities based on expression of the 500 most variable genes
6. Perform a DE analysis between clusters that are adjacent in the hierarchy from (5), and merge them if the proportion of genes that are found to be significantly differentially expressed between them (adjP < .05) is less than than 0.1.

Using only the 500 most variable genes insures the biological variation will dominate the technical variation, and enhances the reproducibility of t-SNE[25].

Importantly, samples that at step (4) don't have a high enough affinity to any emerging cluster, will not be assigned to any cluster. The clustering is performed using the `clusterExperiment::clusterSingle` and `clusterExperiment::clusterMany` functions, the consensus clustering is obtained using the `clusterExperiment::combineMany` function, and the cluster merging (steps 5 and 6) using the `clusterExperiment::makeDendrogram` and `clusterExperiment::mergeClusters` functions. For more details, see 16.

## Choice of dimensionality reduction technique in the clustering procedure

In step (1) above, we cluster cells in a lower dimension embedding using either PCA[26] or t-SNE[27], in a dataset-dependent manner. Different single cell datasets respond better to different dimensionality reduction techniques which are better able to tease out the biological cluster structure of the data. In order to pick the right technique algorithmically, we compute the entropy in a 2D embedding. We obtained 2D embeddings from the 500 most variable genes using either PCA or t-SNE, binned them in a 20x20 grid, and computed the entropy using the `discretize` and `entropy` functions in the *entropy* R package[iii] [28]. The entropy in the 2D embedding is a measure for the information captured by it. For the clustering procedure, we pick the embedding with the highest information content. For the hematopoiesis and glioblastoma datasets, this is t-SNE, while for the embryonic development dataset it is PCA (Table 2). This method may be used to pick any dimensionality reduction technique other than the ones mentioned here, which might be more suitable for other analyses.

## Cluster purity and adjusted mutual information

The cluster purity metric displayed above refers to the proportion of the samples in a cluster which are of the dominant cell type in that cluster. The purity for cluster $i$ is given by

$$Purity_i = \frac{\sum_{j \in C_i} \begin{cases} 1, & \text{if } label_j = \text{dom}_i \\ 0, & \text{otherwise} \end{cases}}{n_i}$$

where $C_i$ = {$j$|cell$_j$ ∈ cluster$_i$}, $label_j$ is the cell type of $cell_j$, $n_i = |C_i|$ is the number of cells in cluster $i$, and

$$\text{dom}_i = \underset{l}{\arg\max} \sum_{j \in C_i} \begin{cases} 1, & \text{if } label_j = l \\ 0, & \text{otherwise} \end{cases}$$

is the dominant cell type in cluster $C_i$.

In addition to the cluster purity metric, we computed the Adjusted Mutual Information (AMI)[29], an information theoretic measure of clustering accuracy which accounts for true positives (two cells of the same type in the same cluster) being caused by chance. The AMI between a clustering $C$ and the true labels $L$ is given by

$$AMI(L,C) = \frac{MI(L,C) - E[MI(L,C)]}{max(H(L),H(C)) - E[MI(L,C)]},$$

where $M\ I(a, b)$ is the mutual information between labellings $a$ and $b$, $H(a)$ is entropy of clustering $a$, and $E[\cdot]$ denotes the expectation.

We do not compare the clusterings using the Rand index, as that measure penalizes for so-called *false negatives* (two cells of the same cell type but in different clusters), which is

**Table 2. Entropy in 2D lower dimension embeddings.**

| Dataset | PCA Entropy | t-SNE Entropy |
|---|---|---|
| Hematopoiesis | 4.96 | 5.03 |
| Embryonic cells | 4.09 | 3.94 |
| Glioblastoma | 4.87 | 5.06 |

[iii] Version 1.2.1, available from CRAN: https://cran.r-project.org/web/packages/entropy/index.html

undesirable as cells from the same cell type might be rightly split into several clusters when a novel cell type is identified.

### Construction of the smoothing kernel

The PPI graph from which the diffusion kernel was derived was constructed using data from string-db[10]. For each pair of proteins, string-db provides a *combined interaction score*, which is a score indicating how confident we can be in the interaction between the proteins, given the different kinds of evidence string-db collates. We subset the links to only those above the 90th percentile of combined interaction scores, only keeping the 10% most confident interactions. For mouse that is 1,020,816 interactions among 17013 genes. For human, 852,722 interactions among 17467 genes.

### MAGIC and scImpute parameters

For all the results presented in this paper, scImpute was run using the default parameters (`drop_thre = 0.5`). For MAGIC, we used values for the diffusion time parameter ($T = \{1, 2, 4, 8, 16\}$). Unlike *netSmooth*, for MAGIC the proportion of samples in robust clusters and the cluster purities were anti-correlated; thus we picked the one that gave the best cluster purities as the best MAGIC parameter. The chosen T values are given in Table 3. We used MAGIC version 0.1[iv] and scImpute version 0.0.2[v].

**Table 3. Optimal diffusion time values for MAGIC.**

| Dataset | Optimal T |
|---|---|
| Hematopoiesis | 1 |
| Embryonic cells | 4 |
| Glioblastoma | 2 |

### the *netSmooth* R package

The analysis for this paper was done using the companion *netSmooth* R-package[22], which is available online: https://github.com/BIMSBbioinfo/netSmooth.

The *netSmooth* R package was included in the 3.7 release of Bioconductor: https://bioconductor.org/packages/release/bioc/html/netSmooth.html and was developed and tested under R version 3.5.

Archived code at time of publication: https://doi.org/10.5281/zenodo.1119064.

License: GPLv3.

### Author contributions

AA conceptualized the project, AA and JR conceived of the algorithm together. All the analysis and software development was done by JR, who also wrote the initial draft of manuscript with input from AA. AA supervised the writing, software development and analysis. JR wrote R package with input and code review and contributions from AA.

### Competing interests

The authors declare none.

### Grant information

AA and JR are funded by core funding from Max Delbrück Center, part of Helmholtz Association.

### Acknowledgements

We would like to thank Vedran Franke, Bora Uyar and Brendan Osberg for valuable comments and input for the development of this manuscript.

## Supplementary material

Supplementary figures S1–S14.

Click here to access the data.

[iv] Available from GitHub: https://github.com/pkathail/magic.

[v] Available from GitHub: https://github.com/Vivianstats/scImpute.

### References

1. Wagner A, Regev A, Yosef N: **Revealing the vectors of cellular identity with single-cell genomics.** *Nat Biotechnol.* 2016; **34**(11): 1145–1160.
   PubMed Abstract | Publisher Full Text | Free Full Text
2. Kharchenko PV, Silberstein L, Scadden DT: **Bayesian approach to single-cell differential expression analysis.** *Nat Methods.* 2014; **11**(7): 740–742.
   PubMed Abstract | Publisher Full Text | Free Full Text
3. Wu AR, Neff NF, Kalisky T, *et al.*: **Quantitative assessment of single-cell RNA-sequencing methods.** *Nat Methods.* 2014; **11**(1): 41–46.
   PubMed Abstract | Publisher Full Text | Free Full Text
4. Pierson E, Yau C: **ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis.** *Genome Biol.* 2015; **16**: 241.
   PubMed Abstract | Publisher Full Text
5. Lin P, Troup M, Ho JW: **CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data.** *Genome Biol.* 2017; **18**(1): 59.
   PubMed Abstract | Publisher Full Text | Free Full Text
6. Li WV, Li JJ: **scimpute: Accurate and robust imputation for single cell rna-seq data.** *bioRxiv.* 2017.
   Publisher Full Text
7. van Dijk D, Nainys J, Sharma R, *et al.*: **Magic: A diffusion-based imputation method reveals gene-gene interactions in single-cell rna-sequencing data.** *bioRxiv.* 2017.
   Publisher Full Text
8. Bhardwaj N, Lu H: **Correlation between gene expression profiles and protein-protein interactions within and across genomes.** *Bioinformatics.* 2005; **21**(11):

2730–2738.
PubMed Abstract | Publisher Full Text

9. Fraser HB, Hirsh AE, Wall DP, *et al.*: **Coevolution of gene expression among interacting proteins.** *Proc Natl Acad Sci U S A.* 2004; **101**(24): 9033–9038.
PubMed Abstract | Publisher Full Text | Free Full Text

10. Szklarczyk D, Morris JH, Cook H, *et al.*: **The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible.** *Nucleic Acids Res.* 2017; **45**(D1): D362–D368.
PubMed Abstract | Publisher Full Text | Free Full Text

11. Lee I, Blom UM, Wang PI, *et al.*: **Prioritizing candidate disease genes by network-based boosting of genome-wide association data.** *Genome Res.* 2011; **21**(7): 1109–1121.
PubMed Abstract | Publisher Full Text | Free Full Text

12. Hofree M, Shen JP, Carter H, *et al.*: **Network-based stratification of tumor mutations.** *Nat Methods.* 2013; **10**(11): 1108–1115.
PubMed Abstract | Publisher Full Text | Free Full Text

13. Vandin F, Upfal E, Raphael BJ: **Algorithms for detecting significantly mutated pathways in cancer.** *J Comput Biol.* 2011; **18**(3): 507–522.
PubMed Abstract | Publisher Full Text

14. Dørum G, Snipen L, Solheim M, *et al.*: **Smoothing gene expression data with network information improves consistency of regulated genes.** *Stat Appl Genet Mol Biol.* 2011; **10**(1): pii: /j/sagmb.2011.10.issue-1/sagmb.2011.10.1.1618/sagmb.2011.10.1.1618.xml.
PubMed Abstract | Publisher Full Text

15. Nestorowa S, Hamey FK, Pijuan Sala B, *et al.*: **A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation.** *Blood.* 2016; **128**(8): e20–31.
PubMed Abstract | Publisher Full Text | Free Full Text

16. Purdom E, Risso D: **clusterExperiment: Compare Clusterings for Single-Cell Sequencing.** 2017; R package version 1.2.0.

17. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.** *Bioinformatics.* 2010; **26**(1): 139–140.
PubMed Abstract | Publisher Full Text | Free Full Text

18. Deng Q, Ramsköld D, Reinius B, *et al.*: **Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells.** *Science.* 2014; **343**(6167): 193–196.
PubMed Abstract | Publisher Full Text

19. Patel AP, Tirosh I, Trombetta JJ, *et al.*: **Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma.** *Science.* 2014; **344**(6190): 1396–1401.
PubMed Abstract | Publisher Full Text | Free Full Text

20. Petryszak R, Keays M, Tang YA, *et al.*: **Expression Atlas update--an integrated database of gene and protein expression in humans, animals and plants.** *Nucleic Acids Research.* 2016; **44**(D1): D746–D752.
PubMed Abstract | Publisher Full Text | Free Full Text

21. Lee I, Blom UM, Wang PI, *et al.*: **Prioritizing candidate disease genes by network-based boosting of genome-wide association data.** *Genome Res.* 2011; **21**(7): 1109–1121.
PubMed Abstract | Publisher Full Text | Free Full Text

22. Ronen J, Akalin A: **netSmooth: Net-work smoothing for scRNAseq.** 2018; R package version 1.0.
Reference Source

23. Edgar R, Domrachev M, Lash AE: **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.** *Nucleic Acids Res.* 2002; **30**(1): 207–210.
PubMed Abstract | Publisher Full Text | Free Full Text

24. Soneson C, Robinson MD: **Bias, robustness and scalability in differential expression analysis of single-cell rna-seq data.** *bioRxiv.* 2017.
Publisher Full Text

25. McCarthy DJ, Campbell KR, Lun AT, *et al.*: **Scater: pre-processing, quality control, normalization and visualization of single-cell rna-seq data in r.** *Bioinformatics.* 2017; **33**(8): 1179–1186.
PubMed Abstract | Publisher Full Text | Free Full Text

26. Hastie T, Friedman J, Tibshirani R: **The Elements of Statistical Learning.** *Springer Series in Statistics.* Springer New York Inc., New York, NY, USA, 2001.
Publisher Full Text

27. van der Maaten LJP, Hinton GE: **Visualizing high-dimensional data using t-sne.** 2008.
Reference Source

28. Hausser J, Strimmer K: **entropy: Estimation of Entropy, Mutual Information and Related Quantities.** R package version 1.2.1. 2014.
Reference Source

29. Vinh NX, Epps J, Bailey J: **Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance.** *J Mach Learn Res.* 2010; **11**: 2837–2854.
Reference Source

# Open Peer Review

## Current Referee Status: ✓ ✓

Version 2

Referee Report 14 February 2018

doi:10.5256/f1000research.14981.r30156

✓ **Siddharth Dey**

Center for Bioengineering, University of California, Santa Barbara, Santa Barbara, CA, 93106-5080, USA

In this manuscript, the authors have developed a new computational method to reduce technical biases that result from dropout events in single-cell mRNA sequencing experiments, a problem that particularly affects genes that are expressed at low levels. While single-cell mRNA sequencing has revolutionized our understanding of several biological systems in the last few years, the relatively low efficiency of amplifying small quantities of mRNA from a single cell results in dropout events that bias downstream analysis. To reduce this technical bias, the authors use information from protein-protein interaction maps to smoothen transcript counts across the entire dataset. Several groups are working on imputation based methods in single-cell mRNA-seq, and this manuscript presents an exciting approach to reduce technical noise. It would be helpful if the authors could clarify and discuss the points below in greater detail:

1. Does the smoothening process bias against genes or gene networks that are not well represented in the protein-protein interaction network? One of the striking features of sc mRNA-seq is that it can identify the expression of specific genes that were previously not associated with a particular cell-type. Would this be impacted by netSmooth and can the authors provide examples from the datasets they have analyzed that netSmooth still retains these observations?

2. There are several sc mRNA-seq methods (for example, CEL-Seq, Smart-Seq etc.) that are currently used by different labs. These methods have different features, such as, full-length transcripts or 3’ end sequencing, and the possibility of employing unique molecule identifiers. How do the 3 computational methods compared in this manuscript work on different experimental techniques?

3. For most of the example datasets used in this manuscript, scImpute shows very similar performance to netSmooth on all 3 metrics used to compare the methods. Can the authors discuss how these two methods, while using different approaches, achieve similar performance. Are there conditions/datasets where one method would perform better than the other?

4. The proportion of cells in robust clusters seems to be very sensitive to the choice of the free parameter in netSmooth (Figure 12). Further, in contrast to a statement in the text (last paragraph on page 10), the value of the free parameter that gives the highest proportion of cells in robust clusters does not correspond to the highest median cluster purity in the glioblastoma dataset. This high sensitivity to alpha can potentially pose a challenge. Can the authors comment on this? Could the authors propose alternate strategies for picking the optimal alpha value.

**Is the rationale for developing the new method (or application) clearly explained?**

Yes

**Is the description of the method technically sound?**

Yes

**Are sufficient details provided to allow replication of the method development and its use by others?**

Yes

**If any results are presented, are all the source data underlying the results available to ensure full reproducibility?**

Yes

**Are the conclusions about the method and its performance adequately supported by the findings presented in the article?**

Yes

***Competing Interests:*** No competing interests were disclosed.

**I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

Reader Comment 27 Jun 2018

**Altuna Akalin**,

We thank Dr. Dey for valuable comments. We tried to address them as demonstrated below and we are in the process of uploading a new version with changes.

***1. Does the smoothening process bias against genes or gene networks that are not well represented in the protein-protein interaction network? One of the striking features of sc mRNA-seq is that it can identify the expression of specific genes that were previously not associated with a particular cell-type. Would this be impacted by netSmooth and can the authors provide examples from the datasets they have analyzed that netSmooth still retains these observations?***

Dr. Dey point out that we have demonstrated netSmooth using gene networks derived from all sorts of general context experiments, while scRNAseq experiments might reveal cell-type-specific gene interactions. We appreciate this comment, and have added a paragraph discussing this to the text. We would mostly like to underline that, until such context-specific networks may be constructed, we have demonstrated netSmooth's applicability using general context networks.

***2. There are several sc mRNA-seq methods (for example, CEL-Seq, Smart-Seq etc.) that are currently used by different labs. These methods have different features, such as, full-length transcripts or 3' end sequencing, and the possibility of employing unique molecule identifiers. How do the 3 computational methods compared in this manuscript work on different experimental techniques?***

Dr. Dey raised interesting questions about the performance of the different imputation methods we compared, coupled with different scRNAseq methods (Smart-Seq, CEL-Seq, etc.). While this is a highly relevant question, we feel it is beyond the scope of this study, as answering that question would require obtaining relevant datasets with appropriate ground-truth labels, and using each relevant technique

***3. For most of the example datasets used in this manuscript, scImpute shows very similar performance to netSmooth on all 3 metrics used to compare the methods. Can the authors discuss how these two methods, while using different approaches, achieve similar performance. Are there conditions/datasets where one method would perform better than the other?***

We agree that there is a similar performance of scImpute and netSmooth, although netSmooth is slightly better in our metrics when looking at all the data sets. Our analysis shows that netSmooth affects the drop-out rate less than scImpute, while uncovering slightly more of the biological signal. This happens across the different overall drop-out rates in the 3 experiments we profiled, indicating that netSmooth can achieve better results with less obtrusive transformations of the data than the imputation methods, across a range of experimental conditions.

***4. The proportion of cells in robust clusters seems to be very sensitive to the choice of the free parameter in netSmooth (Figure 12). Further, in contrast to a statement in the text (last paragraph on page 10), the value of the free parameter that gives the highest proportion of cells in robust clusters does not correspond to the highest median cluster purity in the glioblastoma dataset. This high sensitivity to alpha can potentially pose a challenge.***

We provide also a different way of picking alpha parameter in the R package, which is based on 2D entropy. This way we can pick alpha that optimizes the entropy in 2D PCA embedding.

***Competing Interests:*** No competing interests were disclosed.

Referee Report 05 February 2018

**doi:**10.5256/f1000research.14981.r30157

✓ **Fernando J. Calero-Nieto** , **Fiona Kathryn Hamey**

Wellcome Trust and MRC Cambridge Stem Cell Institute & Cambridge Institute for Medical Research, Addenbrooke's Hospital, Cambridge, UK

In this manuscript, the authors describe a method for imputing values to overcome the problem of technical dropouts in single-cell RNA-seq datasets. As stated by the authors, the problem is well known and caused by technical limitations that affect low and high expressed genes. The approach discussed in the manuscript uses prior knowledge about protein interactions in order to smooth the expression values between pairs of genes encoding interacting proteins, thus reducing the number of zero values and altering the expression values of detected genes in each cell and influencing clustering and visualisation results.

The proposed fundament is interesting and certainly worth exploring. However, there are a few considerations when using this type of approach: 1) the possible enhancement of known relationships to the detriment of the discovery of previously unknown ones; 2) Inferring dropouts using pre-known interactions could result in the overestimation of the expression of certain genes; 3) the gene relationships tend to be very cell-type specific so networks and PPIs should be different from cell type to cell type. In relation to the latter point, it is very interesting that the method is flexible enough to accept networks constructed from different sources.

This manuscript compares the netSmooth algorithm to two existing approaches: Magic and scImpute. Overall, netSmooth presents an approach to smooth scRNA-seq data, which may prove useful in noisy datasets affected heavily by dropouts. However, there are several aspects of the manuscript that we found unclear or feel warrant further discussion.

- An important point is how the performance of these type of methods can be assessed. The authors decided to use a combination of clustered heatmaps, robustness and purity of the clustering, including a measurement of the correspondence between the 2 of them (AMI). Important downsides of this method are: a) external annotation of the datasets is required, which is not always available; b) robustness of the clustering seems to be strongly affected by the processing of the data, as the authors show in relation to MAGIC; c) the purity of the clustering could be strongly biased by the size of the clusters, since small clusters could have a greater chance to get a higher score. It would be interesting if the authors could comment on how their metrics are affected by the number of robust clusters identified. For example, could identifying more small clusters in the dataset have an effect in increasing the median cluster purity? And if so, is this a reliable measure for comparing between algorithms.
- It would be useful to include colour bars for the heatmaps. It should also be mentioned what scale the data is plotted using e.g. is it linear or log-transformed and is the scale comparable between all of the processed datasets? Additionally, some panels (for example 2C and 2D) have more clusters than there are colors shown in the cluster color key next to panel A, so the keys should be changed to match the data.
- Figures showing the cluster purity are quite confusing. From the legend and methods, we understood that each point on the boxplot represents the purity for one of the clusters displayed in the clustered heatmaps. Yet in figure 2A, for example, there are 4 robust clusters found in the raw data, but 8 points for the raw clusters in figure 2B. It seems either that there is an error in one of the plots, or that we have misunderstood the cluster purity metric in which case it needs to be more clearly explained.
- For continuity purposes, it would be better that either PCA or tSNE visualisations were shown for all dataset comparisons in the main figures instead of alternation between clustered heatmaps, PCA and tSNE.
- In the introduction section, the authors mention the imputation programme CIDR along with scImpute and MAGIC, yet only compare the performance of netSmooth to scImpute and MAGIC. The authors should either include benchmarking against CIDR on the three datasets, or discuss why this is not appropriate.
- When discussing applying netSmooth to the haematopoietic data, the authors state that “Figure 2a,b shows that after network-smoothing, we are able to identify clusters with a more pronounced differential expression profile. Further, many more of the genes identified as differentially expressed between the clusters (without smoothing) seem to have low and uninformative expression values overall.” However, from visual inspection of Figure 2 there appears to be very little difference in the expression levels of the differentially expressed genes in the two heatmaps,

or in the number of genes with low expression levels. We are unsure exactly what the authors mean by “more pronounced differential expression profile” as it is hard to see a difference in the heatmaps.

- The authors state that “Only MAGIC is able to increase the proportion of cells in this dataset which fall into robust clusters (Figure 3a), but only *netSmooth* leads to more biologically meaningful clusters, in terms of purity and AMI (Figures 3b,c), demonstrating that *netSmooth* can assist in cell type identification, and outperformed both MAGIC and scImpute in this task.” The increase in AMI in 3B is marginal compared to the raw data, and the proportion in robust clusters is higher for raw data than for netSmooth. There is also no clear improvement in the visualizations of figures S1 and S2 between netSmooth and raw data. Combined with the heatmaps in figure 2, we didn’t feel that there was compelling evidence that netSmooth was useful in cell type identification, and therefore this statement should be toned down.
- In figure 4, it is hard to see how either netSmooth or scImpute offer an improved visualization compared to the raw data. This is backed up by very similar metric scores in Figure 5A and 5C between the raw, netSmooth and scImpute bars. Therefore the statement “Although MAGIC and scImpute reduce the 0-count genes further than *netSmooth* (Figure S1), they do not add as much clarity to the developmental stage signal inherent in the data.” appears to overstate how well netSmooth performs on this dataset in comparison to the other two algorithms.
- In several places the text references the wrong supplementary figures. For example, in the sentence “Although MAGIC and scImpute reduce the 0-count genes further than *netSmooth* (Figure S1)” the authors appear to be actually referring to Figure S5.
- In Figure S5, it should be clarified what is plotted. The legend needs to be changed to make it clear what this is showing. Is this the proportion of zero genes *per cell* in each dataset? Also, the data in this figure suggests that this method has a stronger effect on the expression of genes that are already expressed more than in the removal of zeros. The authors should comment on this in the main manuscript.
- When applying netSmooth to the tumor data, the authors assess the ability of their algorithm on the extent to which it separates cells from different samples, stating “it is also applicable in cancer, improving identification of tumor of origin for glioblastomas.” In fact, many researchers are actually interested in removing this effect in order to be able to compare similar cell types between different patients. Is it possible that netSmooth is actually enhancing “batch effect” in this dataset? It would be interesting to see whether netSmooth increases technical (rather than biological) batch effect in another dataset where a strong biological batch effect is not expected.
- When assessing the importance of the PPI network structure, the authors calculate clustering metrics for randomly permuted networks. Can the authors comment on the fact some random networks have better cluster purity than the real network? Also, why do the authors not show AMI for these random clusters when it is often used to support the success of netSmooth compared to other approaches (e.g. in the haematopoiesis dataset)?
- When discussing the parameter selection the authors state that “in the embryonic development dataset, we would choose alpha= 0.7 as the value that produces the highest cluster purity”. But in figure 11B it is actually alpha= 0.4 that has the highest cluster purity. It would be interesting if the authors commented in why there are at least 2 alpha values that get very similar maximum values for each dataset. Also, the figure would benefit from including alpha=0 values to compare with raw data.

**Is the rationale for developing the new method (or application) clearly explained?**

Yes

**Is the description of the method technically sound?**

Yes

**Are sufficient details provided to allow replication of the method development and its use by others?**
Yes

**If any results are presented, are all the source data underlying the results available to ensure full reproducibility?**
Yes

**Are the conclusions about the method and its performance adequately supported by the findings presented in the article?**
Partly

***Competing Interests:*** No competing interests were disclosed.

***Referee Expertise:*** Haematopoiesis, single-cell technologies

**We have read this submission. We believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

Reader Comment 27 Jun 2018
**Altuna Akalin**,

We are thankful for these valuable comments by Drs. Hamey and Calero-Nieto. We responded to the comments and changed the text and figures wherever necessary.

***1. An important point is how the performance of these type of methods can be assessed. The authors decided to use a combination of clustered heatmaps, robustness and purity of the clustering, including a measurement of the correspondence between the 2 of them (AMI). Important downsides of this method are: a) external annotation of the datasets is required, which is not always available; b) robustness of the clustering seems to be strongly affected by the processing of the data, as the authors show in relation to MAGIC; c) the purity of the clustering could be strongly biased by the size of the clusters, since small clusters could have a greater chance to get a higher score. It would be interesting if the authors could comment on how their metrics are affected by the number of robust clusters identified. For example, could identifying more small clusters in the dataset have an effect in increasing the median cluster purity? And if so, is this a reliable measure for comparing between algorithms.***

The reviewers raised the accurate point that the cluster purity metric may be biased towards clusterings with larger numbers of clusters. For instance, at the edge case, a clustering which assigns a unique cluster to each sample, will score 100% on cluster purity. In order to address this issue, we also computed the adjusted mutual information (AMI) for each clustering, which is a chance-adjusted metric which is not biased in the same way. We expand on this both in the results section under "Network smoothing improves cell type identification from single-cell RNA-seq", and in the methods section.

***2.It would be useful to include colour bars for the heatmaps. It should also be mentioned what scale the data is plotted using e.g. is it linear or log-transformed and is the scale comparable between all of the processed datasets? Additionally, some panels (for example 2C and 2D) have more clusters than there are colors shown in the cluster color key next to panel A, so the keys should be changed to match the data.***
The Reviewers note that the heatmap figures (Figures 2, S3 and S6) are not clear about what the value is in the heatmap, nor do they include a colorbar to gauge whether expression values in the different methods are comparable. We agree that this was a shortcoming and have amended those figures and legends to reflect what is plotted more accurately.

***3. Figures showing the cluster purity are quite confusing. From the legend and methods, we understood that each point on the boxplot represents the purity for one of the clusters displayed in the clustered heatmaps. Yet in figure 2A, for example, there are 4 robust clusters found in the raw data, but 8 points for the raw clusters in figure 2B. It seems either that there is an error in one of the plots, or that we have misunderstood the cluster purity metric in which case it needs to be more clearly explained.***
The reviewers correctly point out a mistake in the plot where figure 2A only shows 4 robust clusters, where there are in fact 8. We have re-created and corrected the error, and thank the reviewers for pointing out our mistake.

***4. For continuity purposes, it would be better that either PCA or tSNE visualisations were shown for all dataset comparisons in the main figures instead of alternation between clustered heatmaps, PCA and tSNE.***
The reviewers point out that we alternate between using PCA and t-SNE for scatter plots of the different datasets, and express a wish for consistency with the visialuzations throughout the paper. While we understand the desire for consistency in the visual information presented, we wish to stress that this was done on purpose. Different scRNAseq datasets respond differently to the different dimensionality reduction techniques, and it is standard practice in the community to try more than one and then pick the best one ad-hoc. We present as a part of the netSmooth R package a way to automate this step using the entropy of 2D embeddings. We expand on this in more detail in the Methods section, under "Choice of dimensionality reduction technique in the clustering procedure".

***5. In the introduction section, the authors mention the imputation programme CIDR along with scImpute and MAGIC, yet only compare the performance of netSmooth to scImpute and MAGIC. The authors should either include benchmarking against CIDR on the three datasets, or discuss why this is not appropriate.***
The reviewers pointed out that while we included CIDR, a method for imputation and clustering of scRNAseq in our introduction, we did not benchmark it against our method. The reason for this omission was that CIDR uses a built-in clustering procedure as a part of the imputation workflow. We chose to compare to MAGIC and scImpute, which are agnostic to the clustering procedure, in order to have an apples-to-apples comparison of imputation methods using the same post-imputation analysis. However we agree with the reviewers that the omission may have been glaring, and have included benchmarks that were possible. We were not able to compare them on the cluster robustness metric, as CIDR assigns all samples to clusters, and does now have a notion of robust clusters.

***6. When discussing applying netSmooth to the haematopoietic data, the authors state that “Figure 2a,b shows that after network-smoothing, we are able to identify clusters with a***

***more pronounced differential expression profile. Further, many more of the genes identified as differentially expressed between the clusters (without smoothing) seem to have low and uninformative expression values overall." However, from visual inspection of Figure 2 there appears to be very little difference in the expression levels of the differentially expressed genes in the two heatmaps, or in the number of genes with low expression levels. We are unsure exactly what the authors mean by "more pronounced differential expression profile" as it is hard to see a difference in the heatmaps.***

We agree with the reviewers that the claim about the more pronounced differential expression pattern in the heatmap was unsubstantiated, and have accordingly changed the text to point out that the difference in the heatmaps is negligible

***7. The authors state that "Only MAGIC is able to increase the proportion of cells in this dataset which fall into robust clusters (Figure 3a), but only netSmooth leads to more biologically meaningful clusters, in terms of purity and AMI (Figures 3b,c), demonstrating that netSmooth can assist in cell type identification, and outperformed both MAGIC and scImpute in this task." The increase in AMI in 3B is marginal compared to the raw data, and the proportion in robust clusters is higher for raw data than for netSmooth. There is also no clear improvement in the visualizations of figures S1 and S2 between netSmooth and raw data. Combined with the heatmaps in figure 2, we didn't feel that there was compelling evidence that netSmooth was useful in cell type identification, and therefore this statement should be toned down.***

The reviewers point out that the difference in benchmark scores shown in Figure 3 represent only a modest improvement over the raw data, and that the statement about the improvement gained from applying netSmooth should be toned down. We've taken this advice and updated the text to make more modest claims.

***8. In figure 4, it is hard to see how either netSmooth or scImpute offer an improved visualization compared to the raw data. This is backed up by very similar metric scores in Figure 5A and 5C between the raw, netSmooth and scImpute bars. Therefore the statement "Although MAGIC and scImpute reduce the 0-count genes further than netSmooth (Figure S1), they do not add as much clarity to the developmental stage signal inherent in the data." appears to overstate how well netSmooth performs on this dataset in comparison to the other two algorithms.***

The reviewers suggest that several of the statements about results, in Figures 2, S2, S3, and 4, over-state the performance of netSmooth relative to the other methods we compared it to. We have toned down several of the statements, and hope the reviewers will find the current text more acceptable.

***9. In several places the text references the wrong supplementary figures. For example, in the sentence "Although MAGIC and scImpute reduce the 0-count genes further than netSmooth (Figure S1)" the authors appear to be actually referring to Figure S5.***

Thank you. This is corrected

***10. In Figure S5, it should be clarified what is plotted. The legend needs to be changed to make it clear what this is showing. Is this the proportion of zero genes per cell in each dataset? Also, the data in this figure suggests that this method has a stronger effect on the expression of genes that are already expressed more than in the removal of zeros.***

***The authors should comment on this in the main manuscript.***
We have changed the legend in order to make this figure more clear. We also appreciate the note from the reviewers about netSmooth having a stronger effect on nonzero genes than the dropouts, compared with other imputation methods, and have added a short discussion of this, under the section "Network smoothing improves capture of developmental expression patterns".

***11. When applying netSmooth to the tumor data, the authors assess the ability of their algorithm on the extent to which it separates cells from different samples, stating "it is also applicable in cancer, improving identification of tumor of origin for glioblastomas." In fact, many researchers are actually interested in removing this effect in order to be able to compare similar cell types between different patients. Is it possible that netSmooth is actually enhancing "batch effect" in this dataset? It would be interesting to see whether netSmooth increases technical (rather than biological) batch effect in another dataset where a strong biological batch effect is not expected.***
The reviewers note that in the Glioblastoma case, we benchmark the methods' ability to identify cells' tumor of origin, while researchers might in fact be interested in the opposite - removing this effect in order to compare cells between cell types. The reviewers are correct and Patel et. al. (the originators of the data) note that these heterogenous tumors consist of different cell types (Pro-neural, Neural, Mesenchymal, and Classical). Identifying such subsets across tumors is an interesting question for researchers of tumor heterogeneity, and we believe that netSmooth might in fact assist in identifying such groups, in this case by making clear which part of the expression signature is patient specific, and which one owed to cross-tumor signatures.

***12. When assessing the importance of the PPI network structure, the authors calculate clustering metrics for randomly permuted networks. Can the authors comment on the fact some random networks have better cluster purity than the real network? Also, why do the authors not show AMI for these random clusters when it is often used to support the success of netSmooth compared to other approaches (e.g. in the haematopoiesis dataset)?***
Reviewers point out that some random networks produces better cluster purity. Actually, figure 8 demonstrates that using the real network in netSmooth leads to cluster purity in the extreme edge of the distribution of random networks. Certainly, with enough random perturbations, a random network can be constructed that will outperform a real network. We demonstrate that in spite of this, the real network scores significantly above expectation in each of the metrics, which demonstrates that the real network holds useful information.

The reviewers point out that when we run the benchmarks using randomized networks in order to demonstrate that the true network structure is important to the results, we only showed the cluster purity and proportion in robust clusters. We agree with the reviewers that the AMI metric also belongs in this context, and have added the AMI to that plot as well.

***13. When discussing the parameter selection the authors state that "in the embryonic development dataset, we would choose alpha= 0.7 as the value that produces the highest cluster purity". But in figure 11B it is actually alpha= 0.4 that has the highest cluster purity. It would be interesting if the authors commented in why there are at least 2 alpha values that get very similar maximum values for each dataset. Also, the figure would benefit from including alpha=0 values to compare with raw data.***

The reviewers noted a mistake in the text referring the the value of alpha which results in the optimal cluster purity in figure 11. We have corrected the misprint.

***Competing Interests:*** No competing interests were disclosed.