ENHANCE: Accurate denoising of single-cell RNA-Seq data

Florian Wagner[1,2], Dalia Barkley[1] & Itai Yanai[1,3]

[1] Institute for Computational Medicine, NYU School of Medicine, New York, USA
[2] Email: florian.wagner@nyu.edu
[3] Email: itai.yanai@nyulangone.org

Single-cell expression measurements are commonly affected by high levels of technical noise, posing challenges for data analysis and interpretation. Here, we propose *ENHANCE*, an algorithm that denoises single-cell RNA-Seq data by first performing nearest-neighbor aggregation and then inferring expression levels from principal components. We benchmark ENHANCE and three previously described methods on simulated data that closely mimic real datasets, and show that ENHANCE provides the best overall denoising accuracy.

Single-cell RNA-Seq (scRNA-Seq) enables the simultaneous measurement of the transcriptome across thousands of cells from complex tissues and entire organs[1–5]. While scRNA-Seq data is unprecedented in its resolution and scale, single-cell expression measurements also exhibit inherently high levels of technical noise[6–9]. This noise can severely impede data analysis, eroding gene correlation structure and precluding the effective visualization of expression patterns in scatter plots and heatmaps[10]. As a result, several methods to computationally remove noise by sharing information across cells and genes have been proposed[10–16]. Here, we describe *ENHANCE* (**E**xpression de**n**oising **h**euristic using **a**ggregation of **n**eighbors and principal **c**omponent **e**xtraction; https://github.com/yanailab/enhance), which represents an intuitive and effective approach for denoising scRNA-Seq data. Using simulated data with realistic technical and biological characteristics, we systematically assess the accuracy of ENHANCE in comparison to three previously described denoising methods.

The vast majority of technical noise in scRNA-Seq data can be understood as a sampling artifact resulting from the detection of only a small fraction of the transcripts in each cell. This sampling noise affects the expression measurement of each gene in a largely independent fashion[6–9]. In contrast, as the co-regulation of genes is a pervasive feature of cell biology, true expression differences between cell subpopulations result in substantial gene-gene correlation structure[17]. We therefore reasoned that principal component analysis (PCA), applied to variance-stabilized data[9], could be a powerful approach for separating biological heterogeneity from technical noise in scRNA-Seq data. By inferring expression levels based upon the leading principal components (PCs), we would retain only true biological differences, while discarding the technical noise captured by higher PCs (**Fig. 1a**). However, we encountered two problems with this approach. First, since technical noise becomes an increasingly dominant factor as the true expression level of a gene decreases, the signal captured by PCA is strongly biased towards highly expressed genes (**Fig. 1b**). Second, the number of PCs that capture biological differences can vary between datasets, and it is unclear how to determine this number in a principled and efficient manner. To address the first problem, we begin with a nearest-neighbor aggregation step, in which the expression profile of each cell is replaced by the aggregate sum of profiles of itself and the *k-1* most similar other cells (neighbors), where *k* is chosen in relation to the average transcript count and

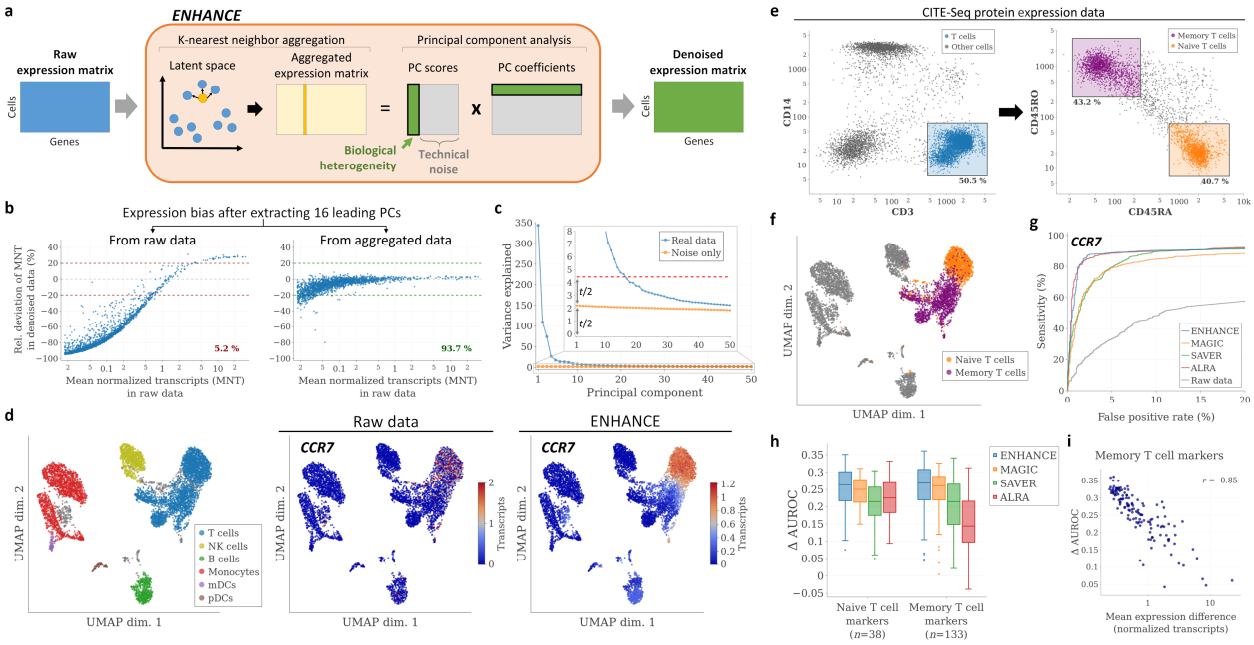

Figure 1. Outline and validation of the ENHANCE denoising algorithm. a Schematic overview of the ENHANCE algorithm. **b** Per-gene expression bias after extracting the leading PCs from a PBMC dataset (4,334 cells), before and after K-nearest neighbor aggregation, with *k*=58. The percent of expressed genes with a bias of 20% or less (relative to the raw data) is indicated. For better readability, only a random subset of 2,000 genes is shown. **c** Variance explained for the first 50 principal components (PCs), for the same PBMC dataset as well as for simulated data containing only technical noise. The threshold (*t*; red dotted line) for determining significant PCs is twice the amount of variance explained by the first PC of the simulated data. **d** Left: UMAP visualization and cell type assignments (**Supplementary Fig. 1**) for transcriptome data from a CITE-Seq PBMC dataset (7,666 cells). Unassigned cells are shown in gray. Middle and right: Expression profile of *CCR7* in the raw data and after denoising with ENHANCE. To improve readability, the color scale for raw *CCR7* expression values was clipped at 2. **e** "Gating" strategy for identifying naïve and memory T cell populations in the same dataset based on protein expression data. **f** Gating results overlaid on top of the UMAP visualization shown in (d). **g** ROC curves showing performance of CCR7 expression as a marker for distinguishing between naïve and memory T cells. **h** Improvement in AUROC scores for ENHANCE and three other denoising methods, for all genes with AUROC > 0.6 in the raw data. **i** Correlation between improvement in AUROC score and absolute expression difference between naïve and memory T cells.

the number of cells in the dataset (**Methods**). Aggregation reduces the noise levels, particularly for lowly expressed genes, and therefore results in a substantial mitigation of the PCA expression bias (**Fig. 1b**). To address the second problem, we developed a simulation approach to estimate the maximum amount of technical noise that a single PC can capture. We then only retain "significant" PCs that capture at least twice this amount, to ensure that most of their signal represents biological differences (**Fig 1c**). Our resulting denoising algorithm (**Fig. 1a** and **Methods**) does not require any parameter tuning and executes in under two minutes on the datasets examined in this study.

To validate our approach, we decided to take advantage of a CITE-Seq dataset of human peripheral blood mononuclear cells (PBMCs), which comprises single-cell expression measurements for the transcriptome as well as a panel of cell surface proteins. We applied ENHANCE to the single-cell mRNA measurements and observed a dramatic effect on the expression profile of individual genes such as the naïve T cell marker *CCR7* (**Fig. 1d**). To determine whether the denoised expression profiles more accurately represented the true expression levels than the raw profiles, we used the protein expression data to independently define naïve and memory subsets of T cells (**Fig 1e, f**). Using these subsets, we

found that unlike in the raw data, *CCR7* expression levels in the denoised data predicted naïve T cell identity with high sensitivity and specificity. We compared the ENHANCE results for *CCR7* to those of MAGIC[10], SAVER[11], and ALRA[16], and found that only ALRA provided a comparable improvement (**Fig. 1g** and **Methods**). We next expanded our analysis to all genes with an AUROC > 0.6 in the raw data, and found that on average, ENHANCE led to the best AUROC improvement, followed by MAGIC and SAVER (**Fig. 1h**). The success of ENHANCE depended on both nearest-neighbor aggregation and PC extraction, as omitting either step resulted in smaller increases in AUROC scores (**Supplementary Fig. 2**). Denoising provided the most benefit for genes with small absolute expression differences between the cell types, underscoring the dramatic impact of noise on the T cell expression profiles (**Fig. 1i**).

We next aimed to systematically compare the accuracies of ENHANCE, MAGIC, SAVER, and ALRA in a simulation study. To ensure that the results were representative of the methods' performance on real scRNA-Seq data, our goal was to generate artificial datasets whose biological and technical sources of variation mirrored those of real datasets. To overcome the limitations of previously described simulation approaches[10,11,13–15,18] (**Supplementary Note 1**), we used the output of a denoising method as the ground truth, and then simulated realistic efficiency and sampling noise (**Methods**). We applied this approach to simulate data based on a renal cell carcinoma biopsy sample[19], which contained a heterogeneous set of populations from the tumor microenvironment (**Fig. 2a**) . To avoid biasing the analysis towards any single denoising method, we performed three separate simulations, Sim-ENHANCE, Sim-MAGIC, and Sim-SAVER, based on the outputs of ENHANCE, MAGIC and SAVER, respectively. To test whether each simulated dataset captured the biological differences present in the real dataset, we performed clustering on the real dataset and overlaid the cluster assignments onto an independent t-SNE visualization of the simulated dataset. We observed a nearly perfect congruence between clusters in the real and the simulated datasets (**Fig. 2a** and **Supplementary Fig. 3**). We next assessed the technical characteristics of the simulated data by comparing the means, standard deviations, and fraction of zero values for each gene with those in the real data, and again found near-perfect agreements (**Fig. 2b** and **Supplementary Fig. 3**). Thus, the simulated datasets recapitulated the cell type differences present in the real dataset, while simultaneously exhibiting the characteristic noise profile of scRNA-Seq data. At the same time, the simulated datasets were far from a carbon copy of the real data, as between 68-70% of the total variance in the simulated datasets constituted randomly generated noise (**Methods**).

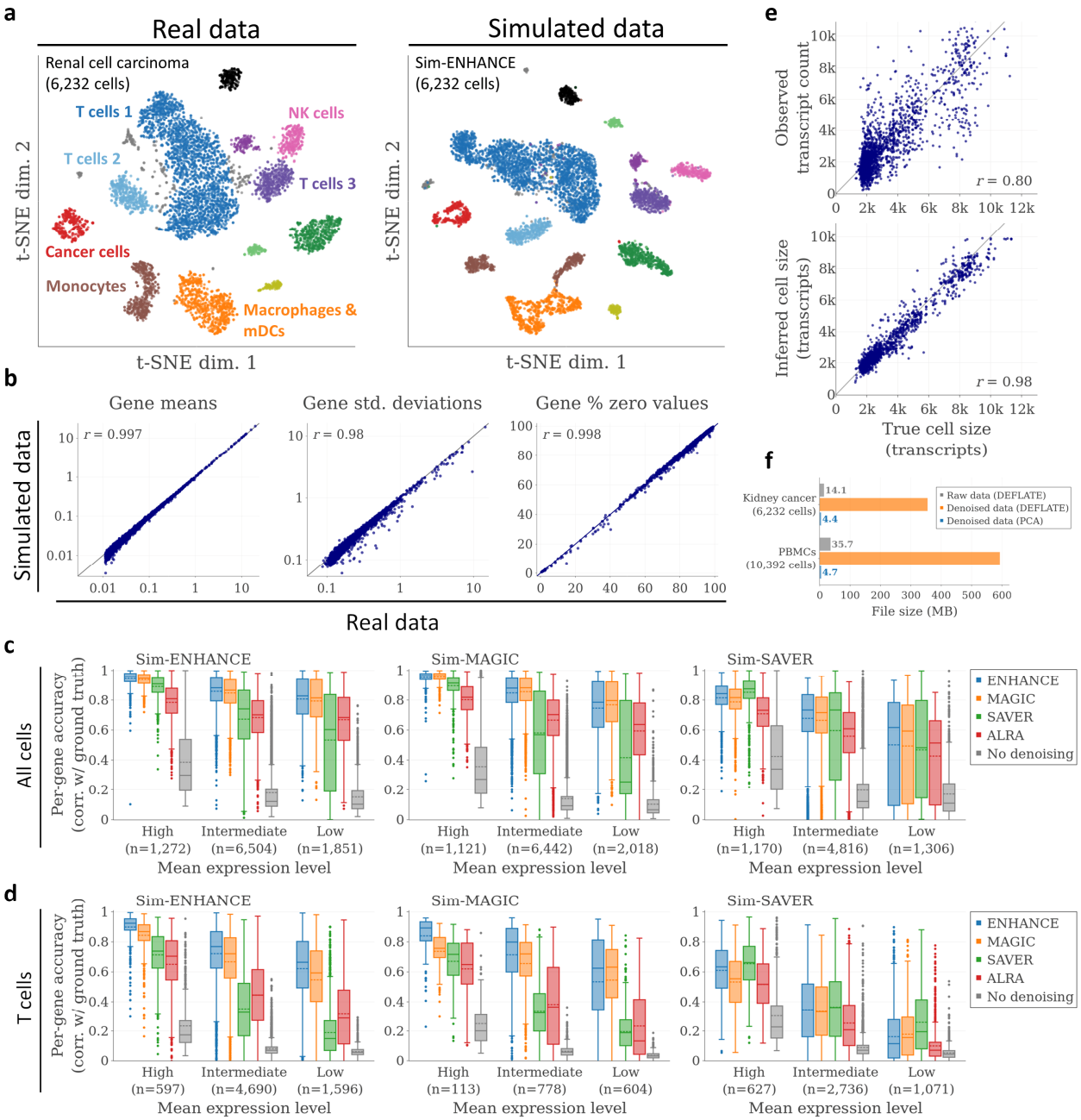

Figure 2. Benchmarking of ENHANCE and three other denoising methods on simulated data. a Left: t-SNE visualization and clustering results of kidney cancer dataset (6,232 cells). Right: Independent t-SNE visualization of the simulated data, overlaid with the clustering results from the raw data. **b** Comparison of gene means, standard deviations and proportion of zero values between real and simulated data. **c** Box plots showing gene-wise correlations between ground truth and denoised data, for simulated kidney cancer data generated using three different simulation methods. **d** Box plots as in (c), after selecting only T cells from the ground truth and simulated datasets. **e** Comparison of true, observed, and ENHANCE-inferred cell sizes for the simulated kidney cancer dataset generated using Sim-ENHANCE (each dot is a cell; only a random subset of 2,000 cells is shown). **f** Comparison of file sizes for different ways storing the raw and denoised data in compressed form.

For each simulation, we used the ground truth data to identify variable genes (**Methods**), and then assessed the correlation between their expression patterns in the denoised data and the ground truth. We grouped genes by expression level, and found that as expected, denoising accuracies were generally the lowest for lowly expressed genes (**Fig. 2c**). In both Sim-ENHANCE and Sim-MAGIC, SAVER and ALRA were clearly outperformed by ENHANCE and MAGIC, which exhibited similar performance. In Sim-SAVER, SAVER outperformed the other methods on highly expressed genes, but not on genes with intermediate and low expression. We observed that MAGIC tended to oversmooth gene expression patterns (**Supplementary Fig. 4**), while SAVER failed to accurately denoise many lowly expressed genes and genes with relatively low variability (**Supplementary Fig. 5**). Next, we aimed to examine the ability of each method to recover expression differences between closely related cell types. To this end, we took advantage of the fact that the kidney cancer dataset contained three different population of T cells, and repeated our correlation analysis after selecting only the T cells from the data. The results showed that ENHANCE outperformed MAGIC in both Sim-ENHANCE and Sim-MAGIC, on genes with high and intermediate expression (**Fig 2d**). We found again that the accuracy of ENHANCE depended on both cell aggregation and PC extraction (**Supplementary Fig. 6**). To assess the robustness of our findings, we repeated our analyses on technical replicates, and obtained the same results (**Supplementary Fig. 7**). Additionally, we performed a second benchmark study using simulated PBMC data, and obtained similar results (**Supplementary Fig. 8**). In summary, our analyses demonstrated that ENHANCE exhibited the best overall denoising accuracy, and that it specifically outperformed the other methods in recovering expression differences between closely related cell types.

While the other denoising methods only generate normalized expression profiles, ENHANCE also aims to remove efficiency noise and infer differences in cell size. In Sim-ENHANCE, we used bootstrapping in conjunction with the observed cell sizes to simulate efficiency noise. Using this approach, we found that ENHANCE was able to accurately infer cell sizes (**Fig. 2e**). Finally, we noticed that since the denoised datasets were no longer sparse, they were not amenable to efficient compression and required more than ten times more disk space than the raw data. However, as the output of ENHANCE can be represented as the scaled product of two much smaller matrices containing PC coefficients and scores, we found that it was possible to reduce the amount of disk space required by about two orders of magnitude, thus allowing for the convenient storage and exchange of ENHANCE results (**Fig. 2f**). In conclusion, ENHANCE represents an intuitive and effective approach to denoising scRNA-Seq data. By simulating realistic scRNA-Seq datasets, we demonstrated significant differences in accuracy between ENHANCE and three popular denoising methods, with ENHANCE exhibiting the best overall performance. More generally, our work demonstrates how PCA can be used to learn an accurate low-dimensional representation of cell types from scRNA-Seq data, which can provide the basis for other machine learning tasks such as classification[20,21]. Future research may be directed at further improving the resolution of ENHANCE, for example by defining a less conservative criterion for identifying significant PCs.

Methods

A detailed description of the methods is provided in the **Online Methods**. Python and R implementations of ENHANCE can be found at https://github.com/yanailab/enhance and https://github.com/yanailab/enhance-R, respectively.

References

1. Svensson, V., Vento-Tormo, R. & Teichmann, S. A. Exponential scaling of single-cell RNA-seq in the past decade. *Nat. Protoc.* **13**, 599–604 (2018).
2. Baslan, T. & Hicks, J. Unravelling biology and shifting paradigms in cancer with single-cell sequencing. *Nat. Rev. Cancer* **17**, 557–569 (2017).
3. Papalexi, E. & Satija, R. Single-cell RNA sequencing to explore immune cell heterogeneity. *Nat. Rev. Immunol.* **18**, 35–45 (2018).
4. Ofengeim, D., Giagtzoglou, N., Huh, D., Zou, C. & Yuan, J. Single-Cell RNA Sequencing: Unraveling the Brain One Cell at a Time. *Trends Mol. Med.* **23**, 563–576 (2017).
5. Griffiths, J. A., Scialdone, A. & Marioni, J. C. Using single-cell genomics to understand developmental processes and cell fate decisions. *Mol. Syst. Biol.* **14**, e8046 (2018).
6. Grün, D., Kester, L. & van Oudenaarden, A. Validation of noise models for single-cell transcriptomics. *Nat. Methods* **11**, 637–640 (2014).
7. Ziegenhain, C. *et al.* Comparative Analysis of Single-Cell RNA Sequencing Methods. *Mol. Cell* **65**, 631-643.e4 (2017).
8. Wang, J. *et al.* Gene expression distribution deconvolution in single-cell RNA sequencing. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E6437–E6446 (2018).
9. Wagner, F., Yan, Y. & Yanai, I. K-nearest neighbor smoothing for high-throughput single-cell RNA-Seq data. *bioRxiv* (2017). doi:10.1101/217737
10. van Dijk, D. *et al.* Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell* **174**, 716-729.e27 (2018).
11. Huang, M. *et al.* SAVER: gene expression recovery for single-cell RNA sequencing. *Nat. Methods* **15**, 539–542 (2018).
12. Li, W. V. & Li, J. J. An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nat. Commun.* **9**, 997 (2018).
13. Talwar, D., Mongia, A., Sengupta, D. & Majumdar, A. AutoImpute: Autoencoder based imputation of single-cell RNA-seq data. *Sci. Rep.* **8**, 16329 (2018).
14. Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S. & Theis, F. J. Single-cell RNA-seq denoising using a deep count autoencoder. *Nat. Commun.* **10**, 390 (2019).
15. Gong, W., Kwak, I.-Y., Pota, P., Koyano-Nakagawa, N. & Garry, D. J. DrImpute: imputing dropout events in single cell RNA sequencing data. *BMC Bioinformatics* **19**, 220 (2018).
16. Linderman, G. C., Zhao, J. & Kluger, Y. Zero-preserving imputation of scRNA-seq data using low-rank approximation. *bioRxiv* 397588 (2018). doi:10.1101/397588
17. Crow, M. & Gillis, J. Co-expression in Single-Cell Analysis: Saving Grace or Original Sin? *Trends Genet. TIG* **34**, 823–831 (2018).
18. Andrews, T. S. & Hemberg, M. False signals induced by single-cell imputation. *F1000Research* **7**, 1740 (2018).
19. Young, M. D. *et al.* Single-cell transcriptomes from human kidneys reveal the cellular identity of renal tumors. *Science* **361**, 594–599 (2018).
20. Trapnell, C. Defining cell types and states with single-cell genomics. *Genome Res.* **25**, 1491–1498 (2015).
21. Wagner, F. & Yanai, I. Moana: A robust and scalable cell type classification framework for single-cell RNA-Seq data. *bioRxiv* 456129 (2018). doi:10.1101/456129