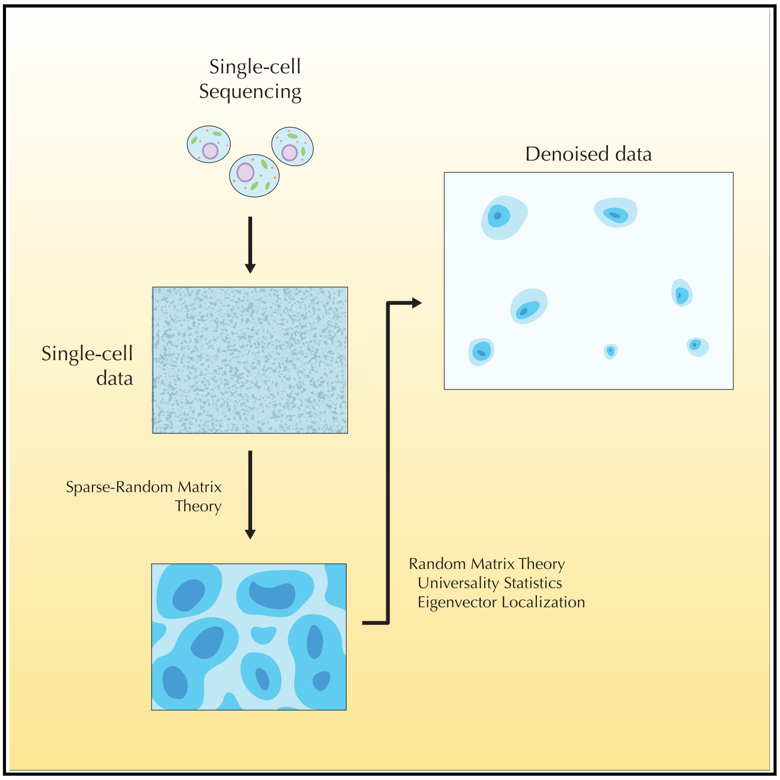
Patterns

A Random Matrix Theory Approach to Denoise Single-Cell Data

Graphical Abstract



Authors

Luis Aparicio, Mykola Bordyuh,
Andrew J. Blumberg, Raul Rabadan

Correspondence

rr2579@cumc.columbia.edu

In Brief

We demonstrate the effectiveness of (sparse) random matrix theory for studying the spectrum of the covariance matrix of single-cell genomic data. We show that single-cell data have a 3-fold structure: a random matrix, a sparsity-induced signal, and a biological signal. Most of the spectrum follows the expectations from random matrix theory (95%), but there exist deviations due to artifacts generated by a sparsity-induced signal (~3%) and a biological signal (~2%).

Highlights

- Sparse random matrix theory provides a suitable framework to study single-cell biology
- Eigenvector localization disentangles sparsity-induced signals from biological signals
- 95% of the information is a random matrix, 3% sparsity-induced signal, and 2% true signal
- The method improves clustering and identification of cell populations

Aparicio et al., 2020, Patterns *1*, 100035
June 12, 2020 © 2020 The Author(s).
https://doi.org/10.1016/j.patter.2020.100035

CellPress

Patterns

CellPress
OPEN ACCESS

Article

A Random Matrix Theory Approach to Denoise Single-Cell Data

Luis Aparicio,[1,2,4] Mykola Bordyuh,[1,2,4] Andrew J. Blumberg,[3] and Raul Rabadan[1,2,5,*]
[1]Department of Systems Biology, Columbia University, New York NY 10032, USA
[2]Department of Biomedical Informatics, Columbia University, New York NY 10032, USA
[3]Department of Mathematics, University of Texas, Austin, TX 78705, USA
[4]These authors contributed equally
[5]Lead Contact
*Correspondence: rr2579@cumc.columbia.edu
https://doi.org/10.1016/j.patter.2020.100035

THE BIGGER PICTURE Single-cell technologies are able to capture information of a biological system cell by cell. Such a level of precision is changing the way we understand complex systems such as cancer or the immune system. However, a major challenge in studying single-cell systems and their underlying biological phenomena is their inherently noisy nature due to their complexity. Random matrix theory is a field with many applications in different branches of mathematics and physics. In the words of one of its developers, the theoretical physicist Freeman Dyson, it describes a "black box in which a large number of particles are interacting according to unknown laws." A complex system with a large number of components (such as genes, biomolecules, or cells) interacting according to unknown laws is the epitome of systems biology. Therefore, random matrix theory looks like a suitable framework to mathematically describe the noise and complexity of gene-cell expression data coming from single-cell biology.

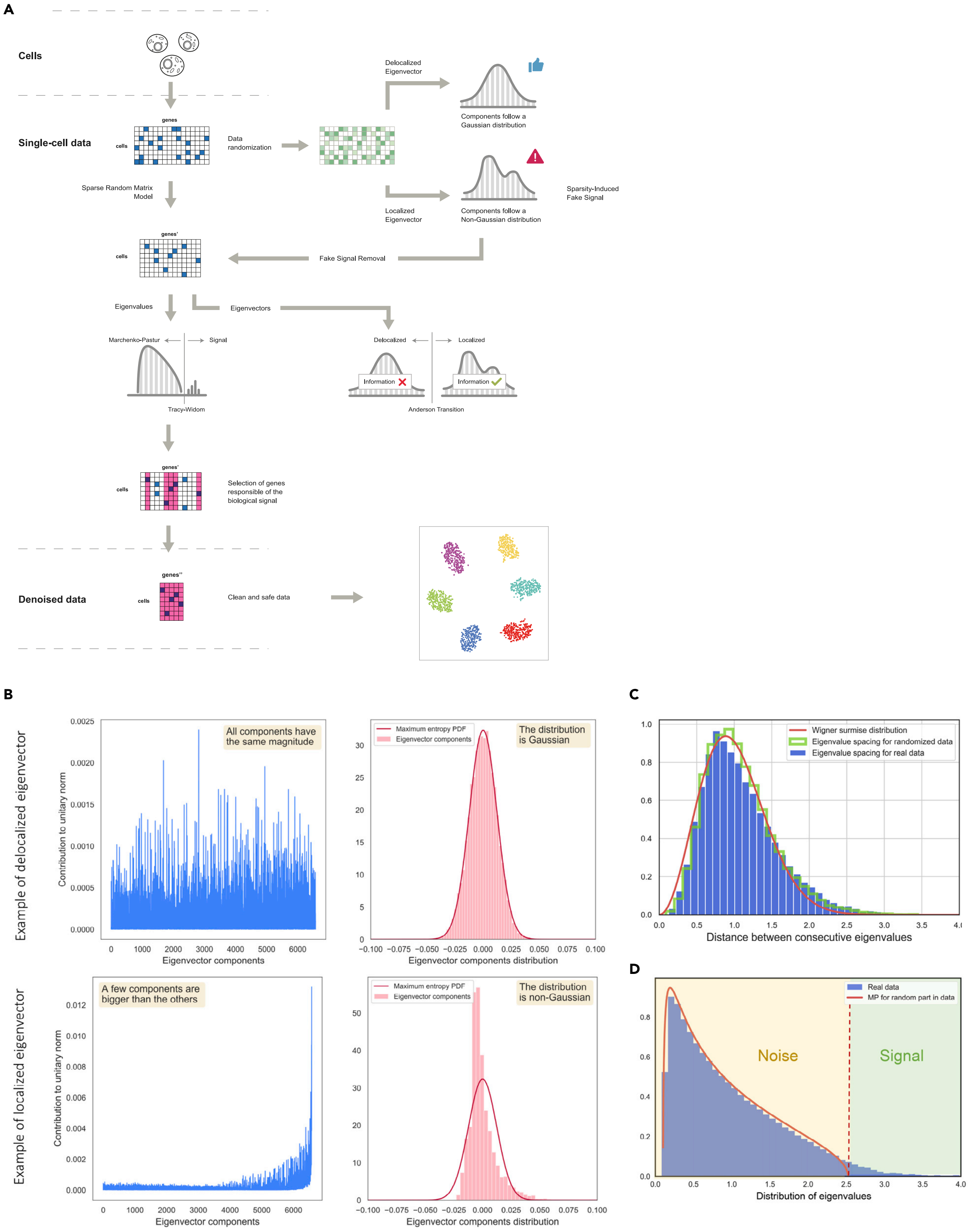
1 **2** 3 4 5 **Proof-of-Concept:** Data science output has been formulated, implemented, and tested for one domain/problem

SUMMARY

Single-cell technologies provide the opportunity to identify new cellular states. However, a major obstacle to the identification of biological signals is noise in single-cell data. In addition, single-cell data are very sparse. We propose a new method based on random matrix theory to analyze and denoise single-cell sequencing data. The method uses the universal distributions predicted by random matrix theory for the eigenvalues and eigenvectors of random covariance/Wishart matrices to distinguish noise from signal. In addition, we explain how sparsity can cause spurious eigenvector localization, falsely identifying meaningful directions in the data. We show that roughly 95% of the information in single-cell data is compatible with the predictions of random matrix theory, about 3% is spurious signal induced by sparsity, and only the last 2% reflects true biological signal. We demonstrate the effectiveness of our approach by comparing with alternative techniques in a variety of examples with marked cell populations.

INTRODUCTION

Characterizing different cellular subtypes in heterogeneous populations and describing their evolution plays a central role in understanding complex systems such as cancer or the immune system. Single-cell technologies offer the opportunity to identify previously unreported cell types and cellular states and explore the relationship between new and known cell states.[1–7] However, there exist several significant biological and technical challenges that complicate the analysis. The first challenge is the lack of a complete quantitative understanding of the different sorts of noise that arise in single-cell measurements, such as intrinsic cell-to-cell variability and spatial and temporal fluctuations within a cell. Moreover, different technologies show biases arising from the process of detecting, amplifying, and sequencing genomic material that significantly vary across different genomic loci. Correctly estimating noise and distinguishing between biological and technical sources of signal is



(legend on next page)

essential for any further analysis, otherwise it is difficult to reliably distinguish states or identify potential variations of a single state. A second complicating factor for single-cell analysis is the sparsity of data (i.e., the large fraction of zero values in the original data matrix), typically caused by the very small amounts of genomic material being amplified.

A number of computational and statistical approaches have been designed to address these challenges.[4,8–13] Imputation methods try to infer the "true" expression levels of missing values from the sample data by empirically modeling the underlying distributions; for instance, using negative binomial plus zero inflation (dropout) for single-cell data. These techniques usually assume that all values are generated by the same distribution (i.e., they assume independent and identically distributed random variables, or i.i.d.). Although there have been efforts to understand the intrinsic stochastic nature of gene expression,[14,15] we currently do not have predictive quantitative models of gene expression. Therefore, it is not clear what the correct distribution is, or whether it is reasonable to make the i.i.d. assumption. Given the lack of a quantitative microscopic description of cell transcription, we would ideally like to have a statistical description of the noise in single-cell data that does not rely on specific details of the underlying distributions of expression.

Universality and Random Matrix Theory in Single-Cell Biology

Historically a similar problem arose in the 1950s in nuclear physics, when the lack of quantitative models of complex nuclei precluded accurate predictions of their energy levels. However, simple theoretical models based on experimental data showed that some observables, such as the spacing between two consecutive energy levels, followed distributions that could be derived from random matrices, i.e., matrices whose entries are independently sampled from a given probability distribution.[16–18] The same distributions were subsequently identified in a variety of complex systems including quantum versions of chaotic systems[19] and patterns of zeros of the Riemann zeta function.[20,21] In this paper, we show that these distributions also appear in the context of single-cell biology and that their properties can be used to denoise single-cell data (Figure 1A).

Random matrix theory (RMT) studies the statistical properties of the eigenvalues and eigenvectors of an ensemble of random matrices. These statistical properties exhibit a phenomenon known as universality, where under mild hypotheses the specific details of the underlying probability distribution generating the entries of the matrix become irrelevant (akin to the central limit theorem).[22,23] Specifically, the observed distributions depend only on the finiteness of the first few moments of the distribution generating the matrix entries.[24–26] RMT universality implies that the density of eigenvalues of covariance matrices obtained from a random matrix follows the Marchenko-Pastur (MP) distribution.[22,27] It also implies that the eigenvectors of a random matrix are delocalized, i.e., their norm is equally distributed across all their components (see Figure 1B and Supplemental Experimental Procedures for an extensive discussion).

We propose here to apply this universality phenomenon to identify statistical features of noise present in single-cell biology (Figure 1). In particular, we claim that any single-cell dataset can be modeled as a random matrix (that encodes the noise) plus a low-rank perturbation (which is the signal). As a consequence, we expect the noise of the system follows the distributions predicted by RMT universality. Large deviations from these distributions indicate the presence of a signal that can be further analyzed. At the level of eigenvalues, random deviations from the MP distribution are described by the Tracy-Widom (TW) distribution, which is the probability distribution for fluctuations on the value of the largest eigenvalue of a random matrix (Figure 1C). Similar strategies using TW and MP distributions have been already discussed in previous works.[28–30]

Eigenvector Localization in Single-Cell Biology

One of the main novelties of this work is the application of the eigenvector statistics predicted by RMT to single-cell sequencing. For the eigenvectors, the transition between noise and signal is described by a phase transition: the delocalized eigenvectors give way to localized eigenvectors, i.e., eigenvectors characterized by having their norm concentrated in a few components (Figure 1B and Supplemental Experimental Procedures). In condensed matter physics this phenomenon is known as Anderson localization.[31] In the single-cell context, localization can be interpreted as groups of cells whose gene expression is correlated. An essential feature of the situation is that the distribution of components for delocalized eigenvectors approximates a Gaussian distribution, whereas the localized eigenvector components have a non-Gaussian distribution (Figure 1B and Supplemental Experimental Procedures). Eigenvalues that lie outside of the MP distribution are associated with localized eigenvectors (Figures 4 and 5).

As noted above, single-cell data are often very sparse. Sparsity introduces a subtlety in the analysis because sparse random matrices can present deviations from the eigenvalue distributions predicted by RMT universality and can have localized eigenvectors (Figure 2A). As a consequence, in a sparse dataset the deviations from the MP distribution and the localized eigenvectors will be partially induced by the sparsity. A way to identify

Figure 1. Random Matrix Theory Applications to Single-Cell Sequencing Data

(A) Schematic of the analysis based on random matrix theory (RMT). Single-cell data can be modeled using sparse random matrix theory (sRMT), showing a 3-fold structure: a random matrix, a sparsity-induced signal, and a biological signal. The strategy proposed here is to identify the biological signal using the predictions from sRMT applied to the covariance matrix of the data.

(B) Deviations from the Tracy-Widom (TW) distribution have been associated to the phenomenon of eigenvector localization. Delocalized eigenvectors are randomly distributed in an N sphere, whereas localized eigenvectors are localized along some directions in the N sphere. Localization can be identified as deviations in components of the eigenvectors from the expected distribution, which is approximately Gaussian in high dimensions. If we think of the components of the eigenvector as a random variable, its probability density function (PDF) (the Gaussian) corresponds to a maximum entropy PDF.

(C) The Wigner surmise distribution captures the spacing between eigenvalues of Wishart matrix across single-cell RNA-sequencing experiments.

(D) Departures from universal distributions predicted by RMT indicate interesting potential biological signals. In red is the non-parametric Marchenko-Pastur (MP) distribution. Deviations from universality can be found by analyzing the larger eigenvalues in relation to the expected TW distribution.

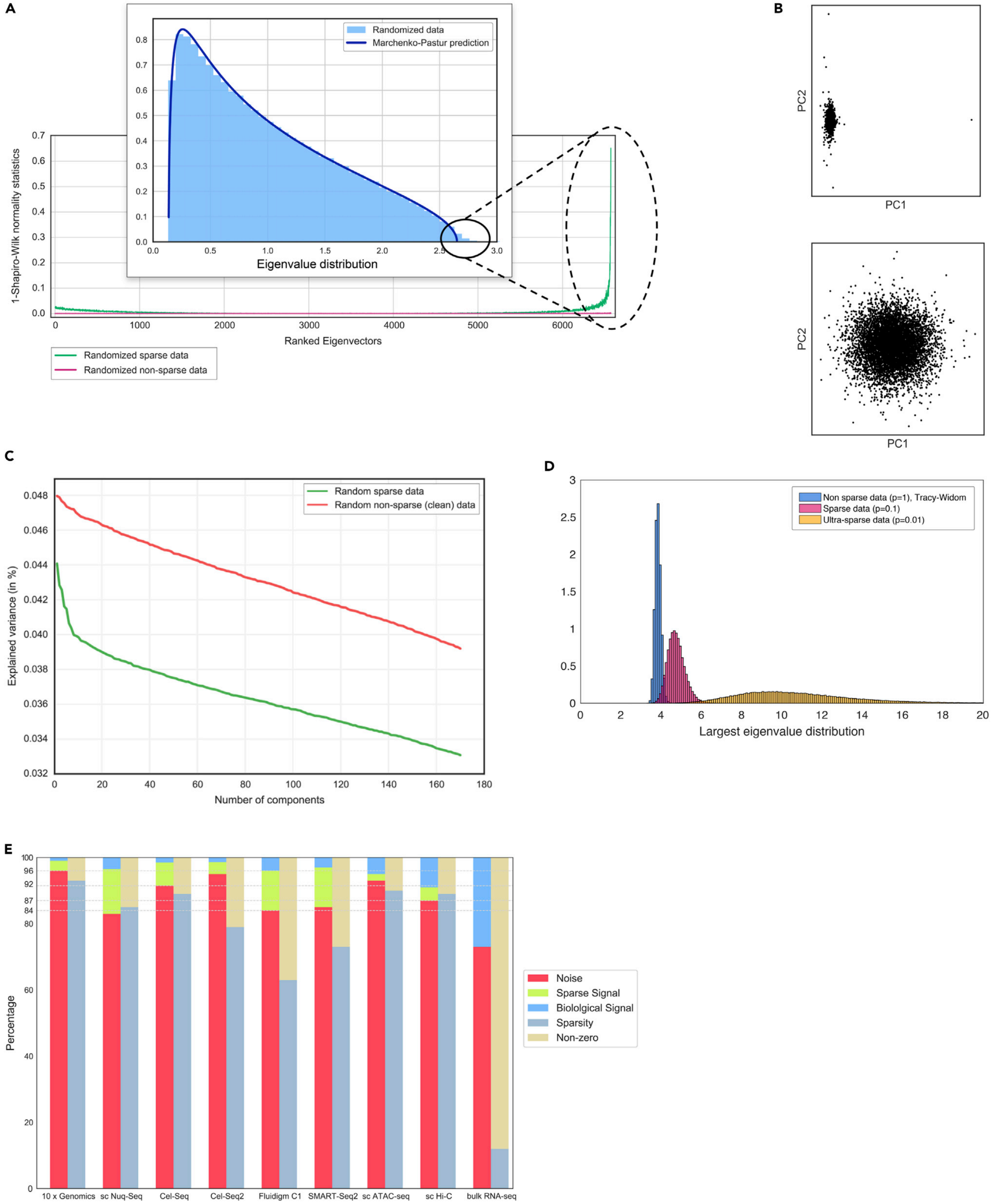


Figure 2. Sparse Random Matrices and Sparsity-Induced Eigenvector Localization

(A) Randomized sparse dataset, corresponding to PBMCs in Kang et al.,[32] where there exist deviations from MP distribution at the eigenvalue level, and presence of localized eigenvectors.

(legend continued on next page)

the effects of sparsity is to randomize the dataset (permuting the cell labels for each gene independently) and observe that, although the entire dataset is now uncorrelated, it still might show localized eigenvectors and potentially also significant deviations in the eigenvalues from the MP prediction (Figure 2A) due to sparsity. We can therefore conclude that single-cell data can be thought of as decomposing into three parts: a random matrix, a sparsity-induced non-biological signal, and a biological signal. To distinguish the biological signal from the sparsity-induced eigenvectors, we propose a feature selection method that discards the features (genes) that are responsible for the localized eigenvectors in the randomized case. This method increases the power for identifying potentially interesting biological signals (Figures 3, 4, and 5). Our approach leads directly to an estimate of the latent space. We show that this procedure is better able than alternative techniques to capture marked single-cell clusters across a variety of datasets (Figure 6).

RESULTS

Quasi-Universality of Single-Cell Sequencing Data

We observed that the distribution of spacing between two consecutive eigenvalues of the sample covariance matrix in different single-cell RNA-sequencing experiments[35–40] resembles the Wigner surmise distribution conjectured by Wigner in 1955[18] in the study of the difference between resonant peaks in slow neutron scattering (Figure 1D). This observation prompted us to investigate the connection between RMT and the spectra of single-cell data, guided by the hypothesis that departures from RMT universality distributions indicate potential biological signals (Figures 1A–1C). We observed that across single-cell datasets these deviations amount to 5% of eigenvalues (Figure 2E). We also demonstrate (Supplemental Experimental Procedures) that the level of localization can be identified as deviations from normality in the distribution of eigenvector components (Figure 2A). Alternatively, localization can be detected using Shannon entropy (Figures S3A–S3C) or by the inverse participation ratio (IPR) (Figures S3D–S3F).

Sparsity-Induced Eigenvector Localization

Single-cell data are usually sparse. Thus, we investigated how sparsity could induce deviations from RMT universality (Figure 2A). By introducing zeros in a random matrix with entries generated with Gaussian or Poisson distributions, we observed deviations in the fluctuations of the eigenvalues from the TW distribution (Figure 2D). A similar phenomenon has been reported in the context of sparse random matrix ensembles, a generalization of RMT to the setting of random matrices with a significant fraction of zero entries. It has been shown[24,41–43] that for the case of sparse Wishart random matrices, the density distribution of eigenvalues deviates from MP and some eigenvectors become localized. We show that this phenomenon can be observed in sparse single-cell data. To this end, we randomized a 95% sparse cell-gene expression matrix corresponding to 6,573 human peripheral blood mononuclear cells (PBMCs) from Kang et al.[32] and analyzed the statistics of its eigenvalues and eigenvectors. Although the bulk of the eigenvalue density seems to follow an MP distribution, it is easily seen that deviations on the upper edge appear. Using a normality test we detected localization in the corresponding eigenvectors (Figures 2A, 4A, and 5A). Eigenvector localization due to sparsity generates artifacts that could potentially be interpreted as true signal in standard application of principal component analysis (PCA). For instance, the highest components of sparse random data show a bias toward the first component (Figure 2B). Another effect of sparsity is the generation of an artifactual "elbow" in randomized sparse data (Figure 2C). Therefore, the first step in our algorithm is to suppress these effects by removing genes that introduce spurious effects due to sparsity. We identify such genes in terms of deviation from normality after random projection (Supplemental Experimental Procedures).

Feature Selection and Application to Single-Cell Transcriptomic Datasets

In this section we explain the application of the RMT analysis to two marked single-cell datasets: 6,573 human PBMCs from Kang et al.[32] (Figure 4) and 3,005 mouse cortex cells from Zeisel et al.[34] (Figure 5). The first step is to remove the sparsity-induced signal. Figures 4A and 5A show the normality test for the eigenvectors before (blue line) and after (red line) removing the sparsity-induced signal. There is a substantial number of eigenvectors that become delocalized once the genes responsible for the sparsity are trimmed out (Supplemental Experimental Procedures). Once the sparsity-induced signal has been removed, the second step in the algorithm is to detect the part of the dataset that corresponds to a random matrix. We first compute the Wishart matrix and then use gradient descent to find the MP distribution in the eigenvalue distribution (Figures 4B and 5B; Supplemental Experimental Procedures). At the same time, the analysis of the normality of the eigenvectors (red line in Figures 4A and 5A) provides an estimate of the amount of information contained in each eigenvector. As mentioned before, the components of delocalized eigenvectors follow a Gaussian distribution; Figures 4A and 5A show the Gaussian profile of each eigenvector through a normality test (Shapiro-Wilk). Interestingly, even some of the eigenvectors corresponding to eigenvalues outside the MP distribution are delocalized and hence do not carry information. A similar argument can be made in terms of other eigenvector features, such as Shannon entropy or IPR (Figure S3).

(B) The localization phenomenon due to sparsity can bias the lower-dimensional representations (*up*). Eliminating the genes that cause eigenvector localization in the randomized dataset generates a more homogeneous distribution in the lower-dimensional representation (*down*), reflecting the random nature of the data.

(C) The effects of sparsity can also be appreciated in the classical elbow plots: sparsity can introduce an artifactual elbow in randomized data.

(D) Deviations from TW distributions can be easily seen in sparse matrices. In this case, 100-by-100 random matrices are drawn a mixture of a normal and a Dirac-delta at zero. Similar results are obtained with other sparse distributions.

(E) Departures from universality amount to near 5% of eigenvalues. However, most of these can be explained by the sparsity of data, suggesting that Sparse Random Matric Theory can provide a better model to understand single-cell sequencing data. Truly potential biological signal amounts to only ~2% of eigenvalues.

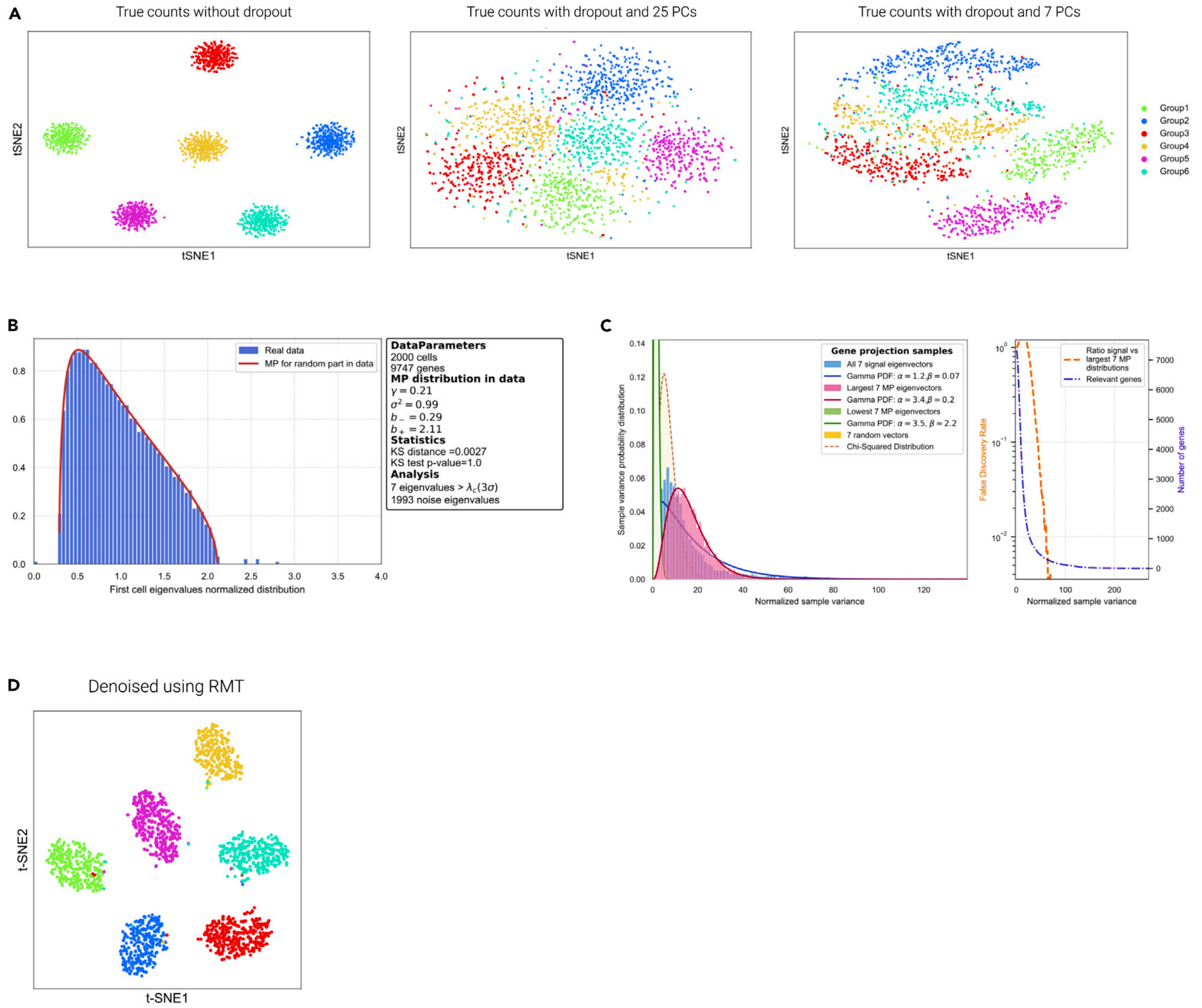


Figure 3. Application to Simulations of Single-Cell and Comparison with Standard PCA

(A) t-SNE representation of a six-cell population single-cell simulation using Splatter[33] for the cases with and without noise associated with dropout effects, and for different selection of principal components after applying a standard PCA technique. The colors correspond to the label of each group of cells simulated, and no clustering has been performed.

(B) MP prediction and identification of the relevant components.

(C) Selection of features (genes) responsible for signal.

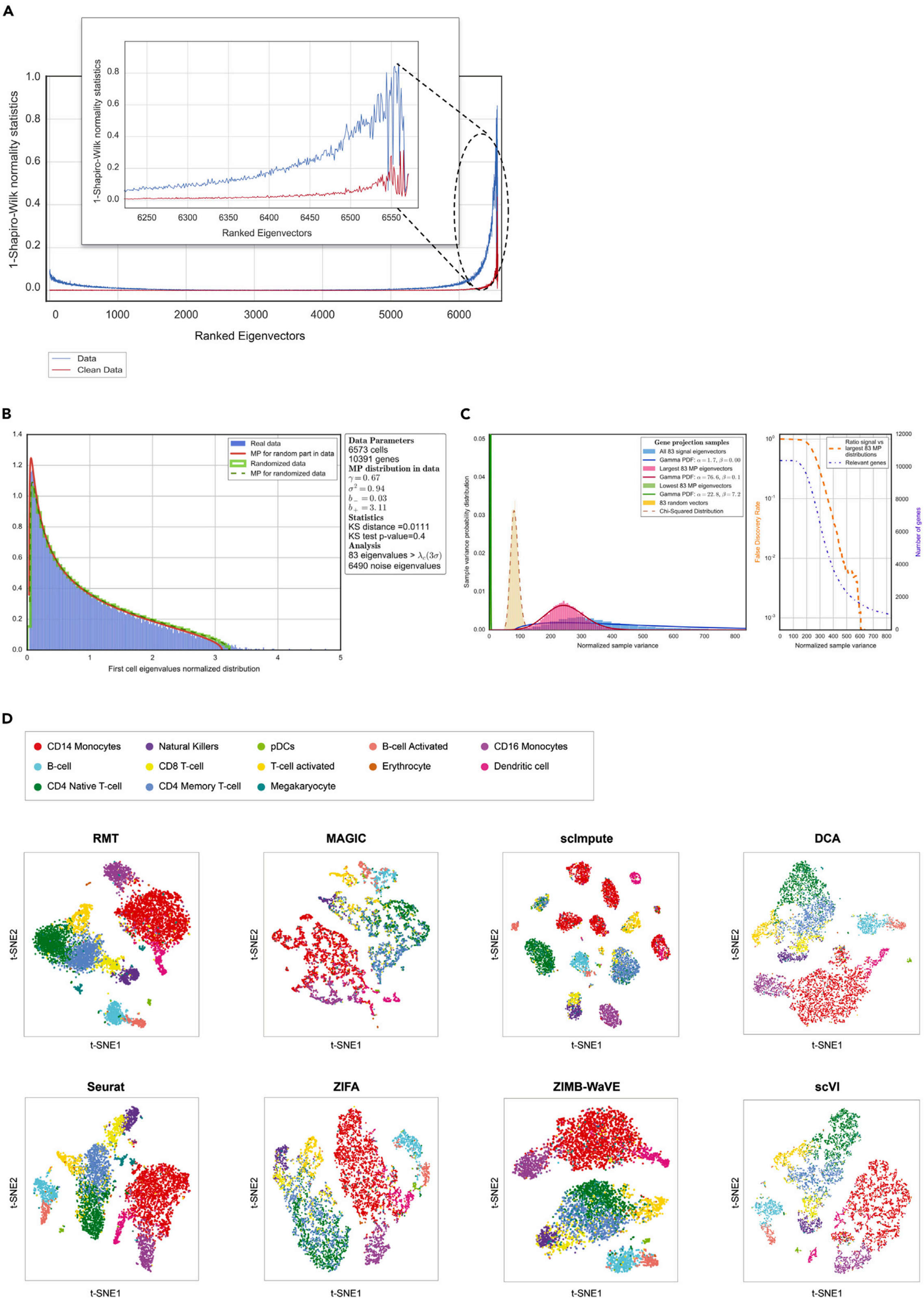
(D) t-SNE representation after results after processing through the RMT.

The strategy of our analysis is to detect and remove these delocalized (non-informative) eigenvectors. Filtering the eigenvectors complements and improves the analysis based only on the eigenvalues.

The third step consists in projecting the dataset onto the eigenvectors that carry signal and also onto different subsets of the eigenvectors that correspond to eigenvalues in the MP distribution (Supplemental Experimental Procedures). Using a chi-squared test for the variance of each gene projected onto the signal and noise eigenvectors, we use a false discovery rate to evaluate which genes are responsible for signal or noise (Figures 4C and 5C). The end result is a selection of features (genes) and a projection of the dataset onto the signal directions. Finally, Figures 4D and 5D (see also Figure S4) show t-distributed stochastic neighbor embedding (t-SNE) representations to visualize in two dimensions the latent space after denoising using our approach. The colors represent the cell populations described in Kang et al.[32] and Butler et al.[11] for human PBMCs, and for marked mouse cortex cell populations described in Zeisel et al.[34] In the same figures we also show a comparison with other methods used to denoise single-cell datasets based on imputation and zero-inflated dimensionality reduction.

Biological Interpretation

We have performed a gene set enrichment analysis on the genes that the algorithm selects as responsible for the biological signal.



(legend on next page)

Using a hypergeometric test on reported biological processes in our mouse brain dataset, the top pathways in the signal gene list correspond to specific brain functions (transmission across chemical synapses, q value = 1.4×10^{-23}, Neural system q value = 2.8×10^{-23}) while in the PBMC dataset the top pathways correspond to immune-system-related processes (immune system q value = 1.9×10^{-32}, cytokine signaling in immune system q value = 2.0×10^{-21}). On the other hand, taking the genes that were not selected by our algorithm, the most significant pathways are associated with generic biological processes (S-phase q value = 4.1×10^{-12}; cell-cycle q value = 2.0×10^{-11}). These results support the contention that eigenvector localization can be used to identify biological processes that are specific to independent cell populations within each experiment.

Simulations and Comparison of Alternative Approaches

We now proceed to evaluate the performance of the algorithm for the identification of potential relevant biological signals. We first perform a single-cell RNA-sequencing simulation of six cell populations using Splatter[33] (Supplemental Experimental Procedures). Figure 3A shows a t-SNE representation of a simulation without and with noise associated with dropout effects. Here, 25 and 7 principal components have been selected. The colors correspond to each cell group simulated (no clustering has been performed). Figure 3D shows the result after our algorithm, and Figures 3B and 3C the associated MP statistics. The first example illustrates the challenge of identifying structures based on t-SNE plots before performing the algorithm (Figure 3A); in contrast, after the algorithm has been applied, we see clearly separated clusters (Figure 3D).

We now perform a comparison with some published algorithms in terms of cell-phenotype cluster resolution. We again use the datasets from Kang et al.[32] (human PBMCs) and Zeisel et al.[34] (mouse cortex) described in the previous section. As explained in the previous sections, these references together with Butler et al.[11] have cells already labeled by phenotype. We claimed in previous sections that our method is able to remove system noise such that the cell-phenotype clusters are better resolved. This noise is partially generated due to the missing values in single-cell experiments. For this reason, we compare the two main approaches in the field that address this: imputation (MAGIC[8] and scImpute[10]) and zero-inflated dimensionality reduction (ZIFA[13] and ZIMB-WaVE[9]). We also perform a comparison with non-linear neural networks methods: scVI[44] and DCA.[45] For completeness, we also compare the raw data with a selection of genes based on higher variance (top 300 genes) and with Seurat.[46] The comparison is performed using the knowledge of cell phenotypes in the studies by Butler et al., Kang et al., and Zeisel et al.[11,32,34] and by computing the mean silhouette score in the reduced space, whereby higher values would indicate a better (less noisy) cell-phenotype cluster resolution. In Figures 6A–6D we represent the mean silhouette score as a function of the latent space number of dimensions for 13 PBMC phenotypes described in Butler et al.[11](Figure 6A) and for 7 (Figure 6B), 15 (Figure 6C), and 26 (Figure 6D) marked mouse cortex cell populations described in Zeisel et al.[34] We have selected the 1,500 most signal-like genes using RMT and we can observe how RMT outperforms other methods in the identification of known marked populations. Notice also how this becomes more dramatic as we increase the number of populations. Although this exercise is done with known populations in order to give a comparative quantitative measure, from Figures 6A–6D we can also conclude that RMT method is a suitable one to better disentangle cell populations by noise removal and hence to find new potential cell populations. Moreover, the performance advantage of the RMT method increases with the dimension of the latent space. This last feature is particularly interesting since in the future, the number of required dimensions in the latent space for an accurate analysis is expected to grow due to continuing improvements in resolution and the number of cells that can be measured.

DISCUSSION

In this paper, we demonstrate the effectiveness of (sparse) RMT for studying the spectrum of the covariance matrix of single-cell genomic data. We have shown that single-cell data shows a 3-fold structure: a random matrix, a sparsity-induced signal, and a biological signal. We also show that while most of the spectrum follows the expectations from RMT (95%), there exist deviations due to artifacts generated by a sparsity-induced signal (~3%) and due to a biological signal (~2%). The large contribution of the random component to the spectral properties of the covariance matrix of single-cell expression data could be due to the stochastic nature of gene expression at single-cell level, as has been studied in a variety of biological contexts.[14,47]

We have introduced a method to denoise single-cell sequencing data studying eigenvalue and eigenvector properties based on RMT. This method uses RMT universality properties of eigenvalue distributions, e.g., the TW and MP distributions, and extends it to the study of the eigenvector properties, based on the localization/delocalization phase transition. This method is also able to select genes responsible for potentially interesting biological signals. The algorithm provides a powerful

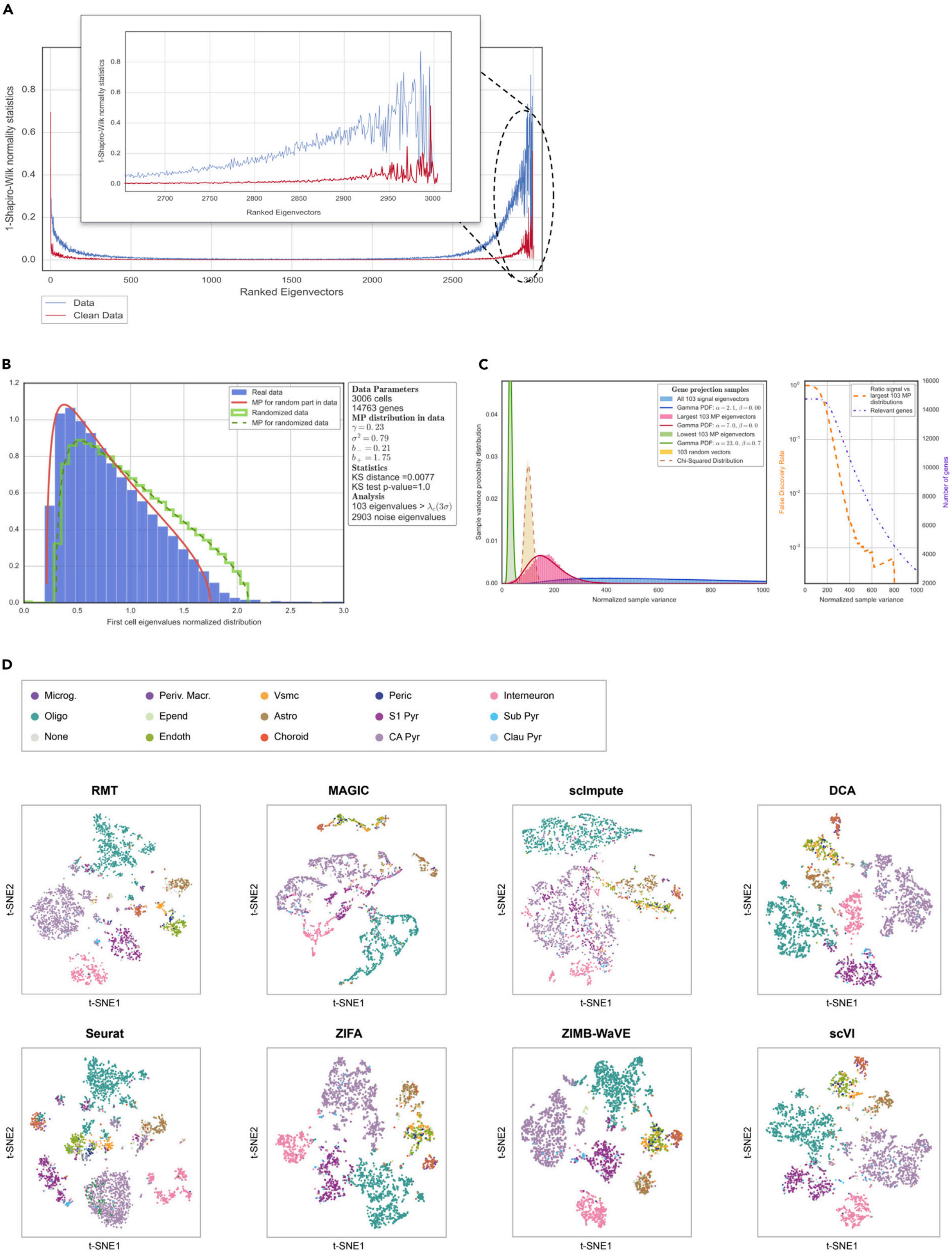
Figure 4. Application to PBMC Single-Cell Expression

(A) Localization properties of the eigenvectors in a single-cell dataset of PBMCs.[32] The blue line represents the system dominated by sparsity and the red line corresponds to the system after removing sparsity. This figure also shows how some eigenvectors corresponding to eigenvalues out of MP distribution are delocalized (red line) and therefore do not carry any information.

(B) MP prediction and identification of relevant components.

(C) Study of the chi-squared test for the variance (normalized sample variance) in signal and noise gene projections. In the left panel, the distributions correspond to a projection of genes into the 83 signal eigenvectors (corresponding to the 83 eigenvalues of A) and the projection into the 83 lowest and 83 largest MP eigenvectors. There is also a projection into 83 random vectors. Finally, the lines show how gamma functions can fit the distributions discussed. The right panel shows the number of relevant genes in terms of the test discussed above, together with a false discovery rate. Higher values for the chi-squared test for variance indicate that the genes are less responsible for the signal.

(D) Comparison of the t-SNE representation for different public algorithms. This case corresponds to 13 different PBMC phenotypes sequenced in Kang et al.[32] and described in Butler et al.[11]



(legend on next page)

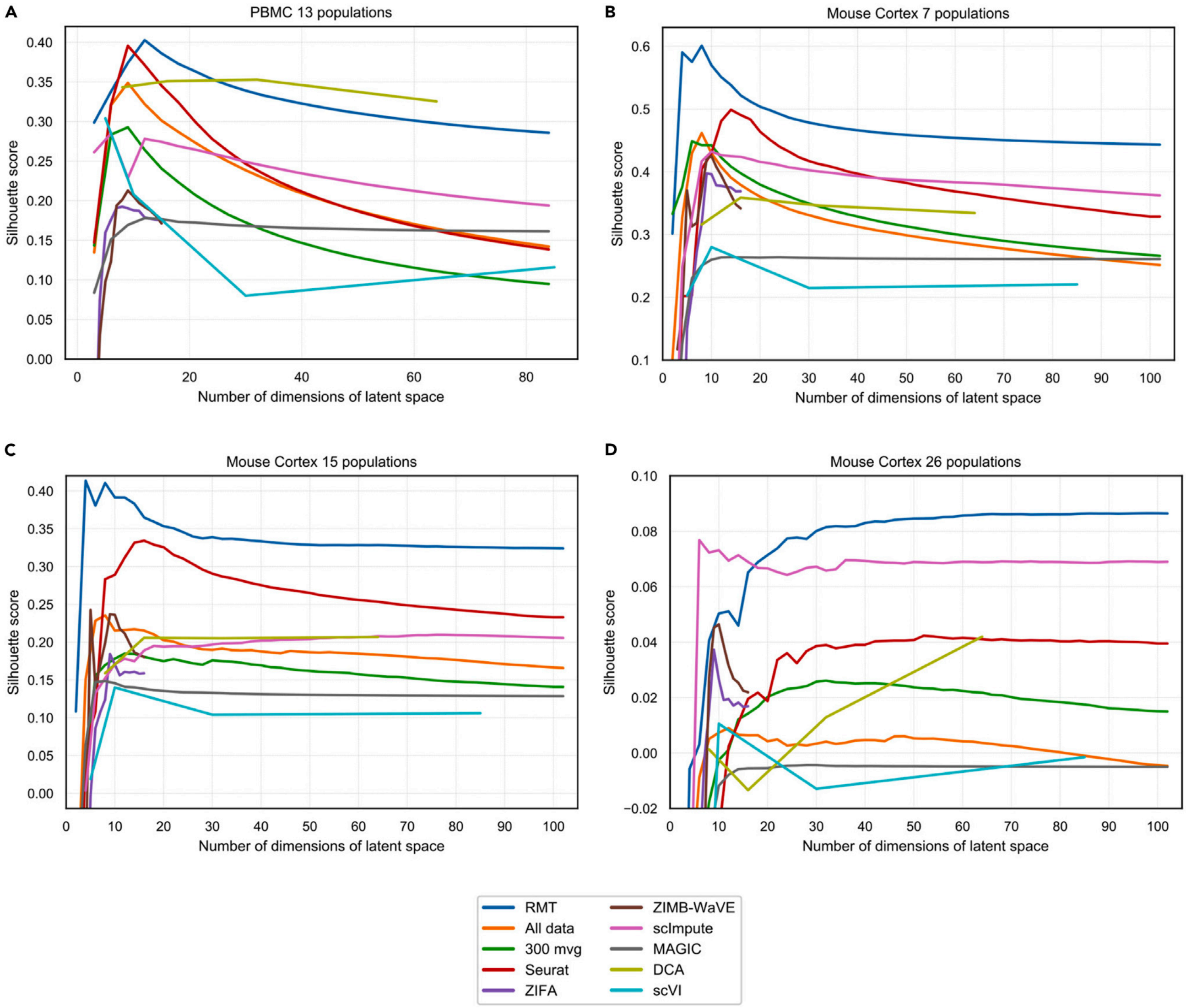


Figure 6. Comparison of Alternative Approaches for Single-Cell Analysis

(A) Mean silhouette score for different methods as a function of the number of dimensions of the latent space for the case of 13 PBMC cell phenotypes described in Butler et al.[11]

(B–D) Mean silhouette score for different methods as a function of the reduced space number of dimensions for the case of 7 (B), 15 (C), and 26 (D) mouse cortex cell phenotypes described in Zeisel et al.[34]

tool to identify this signal and produce a low-rank representation of single-cell data that may be used for further interpretation. Additionally, we should point out that the universality we observed in Wishart/covariance matrices is also observable in the spectra of graph Laplacians (including sparse graphs[48]) and kernel random matrices,[49] which are used in other single-

Figure 5. Application to Mouse Cortex Single-Cell Expression

(A) Localization properties of the eigenvectors in a single-cell dataset of PBMCs.[32] The blue line represents the system dominated by sparsity and the red line corresponds to the system after removing sparsity. This figure also shows how some eigenvectors corresponding to eigenvalues out of MP distribution are delocalized (red line) and therefore do not carry any information.

(B) MP prediction and identification of relevant components.

(C) Study of the chi-squared test for the variance (normalized sample variance) in signal and noise gene projections. In the left panel, the distributions correspond to a projection of genes into the 103 signal eigenvectors (corresponding to the 103 eigenvalues of A) and the projection into the 103 lowest and 103 largest MP eigenvectors. There is also a projection into 103 random vectors. Finally, the lines show how gamma functions can fit the distributions discussed. The right panel shows the number of relevant genes in terms of the test discussed above together with a false discovery rate. Higher values for the chi-squared test for variance indicate that the genes are less responsible for the signal.

(D) Comparison of the t-SNE representation for different methods and algorithms. This case corresponds to 15 different mouse cortex cell phenotypes described in Zeisel et al.[34]

cell analytic techniques, suggesting that the approach followed here could be applied more broadly. The code for the algorithm is publicly available on https://rabadan.c2b2.columbia.edu/html/randomly/.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.patter.2020.100035.

ACKNOWLEDGMENTS

We want to thank Carlos Aparicio, Francesco Brundu, Andrew Chen, Tim Chu, Oliver Elliott, Ioan Filip, Karen Gomez, Antonio Iavarone, Zhaoqi Liu, and Richard Wolff. We would like to especially thank Ivan Corwin for helpful discussions about mathematical aspects and implications of this work. This work was funded in part by R01CA185486-01, R01CA179044-01A1, U54CA209997, U54CA193313, and a Chan-Zuckerberg pilot grant.

AUTHOR CONTRIBUTIONS

L.A., M.B., and R.R. developed the application of localization and sparse RMT concepts into single-cell biology. L.A. and M.B. have developed the RMT-based algorithm and applied it to the single-cell datasets described in the main text and methods under the supervision of R.R. L.A., M.B., and R.R. wrote the manuscript. A.J.B. provided valuable mathematical insights and strategies and helped during the writing of the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing financial interests.

Received: January 15, 2020
Revised: March 5, 2020
Accepted: April 2, 2020
Published: May 4, 2020

REFERENCES

1. Patel, A.P., Tirosh, I., Trombetta, J.J., Shalek, A.K., Gillespie, S.M., Wakimoto, H., Cahill, D.P., Nahed, B.V., Curry, W.T., Martuza, R.L., et al. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science *344*, 1396–1401.
2. Bintu, L., Yong, J., Antebi, Y.E., McCue, K., Kazuki, Y., Uno, N., Oshimura, M., and Elowitz, M.B. (2016). Dynamics of epigenetic regulation at the single-cell level. Science *351*, 720–724.
3. Cao, J., Packer, J.S., Ramani, V., Cusanovich, D.A., Huynh, C., Daza, R., Qiu, X., Lee, C., Furlan, S.N., Steemers, F.J., et al. (2017). Comprehensive single-cell transcriptional profiling of a multicellular organism. Science *357*, 661–667.
4. Rizvi, A.H., Camara, P.G., Kandror, E.K., Roberts, T.J., Schieren, I., Maniatis, T., and Rabadan, R. (2017). Single-cell topological RNA-seq analysis reveals insights into cellular differentiation and development. Nat. Biotechnol. *35*, 551–560.
5. Azizi, E., Carr, A.J., Plitas, G., Cornish, A.E., Konopacki, C., Prabhakaran, S., Nainys, J., Wu, K., Kiseliovas, V., Setty, M., et al. (2018). Single-cell map of diverse immune phenotypes in the breast tumor microenvironment. Cell *174*, 1293–1308.e36.
6. Cusanovich, D.A., Reddington, J.P., Garfield, D.A., Daza, R.M., Aghamirzaie, D., Marco-Ferreres, R., Pliner, H.A., Christiansen, L., Qiu, X., Steemers, F.J., et al. (2018). The cis-regulatory dynamics of embryonic development at single-cell resolution. Nature *555*, 538–542.
7. Farrell, J.A., Wang, Y., Riesenfeld, S.J., Shekhar, K., Regev, A., and Schier, A.F. (2018). Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. Science *360*, https://doi.org/10.1126/science.aar3131.
8. van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A.J., Burdziak, C., Moon, K.R., Chaffer, C.L., Pattabiraman, D., et al. (2018). Recovering gene interactions from single-cell data using data diffusion. Cell *174*, 716–729.e27.
9. Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S., and Vert, J.P. (2018). A general and flexible method for signal extraction from single-cell RNA-seq data. Nat. Commun. *9*, 284.
10. Li, W.V., and Li, J.Y.J. (2018). An accurate and robust imputation method scImpute for single-cell RNA-seq data. Nat. Commun. *9*, 997.
11. Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat. Biotechnol. *36*, 411–420.
12. Stegle, O., Teichmann, S.A., and Marioni, J.C. (2015). Computational and analytical challenges in single-cell transcriptomics. Nat. Rev. Genet. *16*, 133–145.
13. Pierson, E., and Yau, C. (2015). ZIFA: dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome Biol. *16*, https://doi.org/10.1186/s13059-015-0805-z.
14. Elowitz, M.B., Levine, A.J., Siggia, E.D., and Swain, P.S. (2002). Stochastic gene expression in a single cell. Science *297*, 1183–1186.
15. Peccoud, J., and Ycart, B. (1995). Markovian modeling of gene-product synthesis. Theor. Popul. Biol. *48*, 222–234.
16. Dyson, F.J. (1962). Statistical theory of energy levels of complex systems .1. J. Math. Phys. *3*, 140–&.
17. Dyson, F.J. (1962). A brownian-motion for eigenvalues of a random matrix. J. Math. Phys. *3*, 1191.
18. Wigner, E.P. (1955). Characteristic vectors of bordered matrices with infinite dimensions. Ann. Math. *62*, 548–564.
19. Bohigas, O., Giannoni, M.J., and Schmit, C. (1984). Characterization of chaotic quantum spectra and universality of level fluctuation laws. Phys. Rev. Lett. *52*, 1–4.
20. Odlyzko, A.M. (1987). On the distribution of spacings between zeros of the zeta-function. Math. Comput. *48*, 273–308.
21. M.L. Mehta, ed. (2004). Pure and Applied Mathematics*142* (Academic Press).
22. Tracy, C.A., and Widom, H. (1993). Level-spacing distributions and the airy kernel. Phys. Lett. B *305*, 115–118.
23. Tao, T., and Vu, V. (2010). Random matrices: universality of local eigenvalue statistics up to the edge. Commun. Math. Phys. *298*, 549–572.
24. Mirlin, A.D., and Fyodorov, Y.V. (1991). Universality of level correlation-function of sparse random matrices. J. Phys. A Math Gen. *24*, 2273–2286.
25. Ben Arous, G., and Peche, S. (2005). Universality of local eigenvalue statistics for some sample covariance matrices. Commun. Pur Appl. Math. *58*, 1316–1357.
26. Pillai, N.S., and Yin, J. (2014). Universality of covariance matrices. Ann. Appl. Probab. *24*, 935–1001.
27. Marchenko, V.A.P., and Pastur, L.A. (1967). Distribution of eigenvalues for some sets of random matrices. Math. USSR Sb. *72*, 457–483.
28. Kiselev, V.Y., Kirschner, K., Schaub, M.T., Andrews, T., Yiu, A., Chandra, T., Natarajan, K.N., Reik, W., Barahona, M., Green, A.R., and Hemberg, M. (2017). SC3: consensus clustering of single-cell RNA-seq data. Nat. Methods *14*, 483–486.
29. Klein, A.M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D.A., and Kirschner, M.W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell *161*, 1187–1201.
30. Shekhar, K., Lapan, S.W., Whitney, I.E., Tran, N.M., Macosko, E.Z., Kowalczyk, M., Adiconis, X., Levin, J.Z., Nemesh, J., Goldman, M., et al. (2016). Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. Cell *166*, 1308–1323.e30.
31. Anderson, P.W. (1958). Absence of diffusion in certain random lattices. Phys. Rev. *109*, 1492–1505.

32. Kang, H.M., Subramaniam, M., Targ, S., Nguyen, M., Maliskova, L., McCarthy, E., Wan, E., Wong, S., Byrnes, L., Lanata, C.M., et al. (2018). Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. Nat. Biotechnol. *36*, 89–94.

33. Zappia, L., Phipson, B., and Oshlack, A. (2017). Splatter: simulation of single-cell RNA sequencing data. Genome Biol. *18*, https://doi.org/10.1186/s13059-017-1305-0.

34. Zeisel, A., Muñoz-Manchado, A.B., Codeluppi, S., Lönnerberg, P., La Manno, G., Juréus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., et al. (2015). Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science *347*, 1138–1142.

35. Lawlor, N., George, J., Bolisetty, M., Kursawe, R., Sun, L., Sivakamasundari, V., Kycia, I., Robson, P., and Stitzel, M.L. (2017). Single-cell transcriptomes identify human islet cell signatures and reveal cell-type-specific expression changes in type 2 diabetes. Genome Res. *27*, 208–222.

36. Grun, D., Muraro, M.J., Boisset, J.C., Wiebrands, K., Lyubimova, A., Dharmadhikari, G., van den Born, M., van Es, J., Jansen, E., Clevers, H., and de Koning, E.J.P. (2016). De novo prediction of stem cell identity using single-cell transcriptome data. Cell Stem Cell *19*, 266–277.

37. Muraro, M.J., Dharmadhikari, G., Grün, D., Groen, N., Dielen, T., Jansen, E., van Gurp, L., Engelse, M.A., Carlotti, F., de Koning, E.J., and van Oudenaarden, A. (2016). A single-cell transcriptome atlas of the human pancreas. Cell Syst. *3*, 385–394.e3.

38. Ramani, V., Deng, X., Qiu, R., Gunderson, K.L., Steemers, F.J., Disteche, C.M., Noble, W.S., Duan, Z., and Shendure, J. (2017). Massively multiplex single-cell Hi-C. Nat. Methods *14*, 263–266.

39. Nestorowa, S., Hamey, F.K., Pijuan Sala, B., Diamanti, E., Shepherd, M., Laurenti, E., Wilson, N.K., Kent, D.G., and Göttgens, B. (2016). A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. Blood *128*, e20–e31.

40. Buenrostro, J.D., Wu, B., Litzenburger, U.M., Ruff, D., Gonzales, M.L., Snyder, M.P., Chang, H.Y., and Greenleaf, W.J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. Nature *523*, 486–490.

41. Fyodorov, Y.V., and Mirlin, A.D. (1991). Localization in ensemble of sparse random matrices. Phys. Rev. Lett. *67*, 2049–2052.

42. Evangelou, S.N., and Economou, E.N. (1992). Spectral density singularities, level statistics, and localization in a sparse random matrix ensemble. Phys. Rev. Lett. *68*, 361–364.

43. Rodgers, G.J., and Bray, A.J. (1988). Density of states of a sparse random matrix. Phys. Rev. B Condens Matter *37*, 3557–3562.

44. Lopez, R., Regier, J., Cole, M.B., Jordan, M.I., and Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. Nat. Methods *15*, 1053–1058.

45. Eraslan, G., Simon, L.M., Mircea, M., Mueller, N.S., and Theis, F.J. (2019). Single-cell RNA-seq denoising using a deep count autoencoder. Nat. Commun. *10*, 390.

46. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F., and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. Nat. Biotechnol. *33*, 495–U206.

47. Kaern, M., Elston, T.C., Blake, W.J., and Collins, J.J. (2005). Stochasticity in gene expression: from theories to phenotypes. Nat. Rev. Genet. *6*, 451–464.

48. Tran, L.V., Vu, V.H., and Wang, K. (2013). Sparse random graphs: eigenvalues and eigenvectors. Random Struct. Algor. *42*, 110–134.

49. El Karoui, N. (2010). The spectrum of kernel random matrices. Ann. Stat. *38*, 1–50.

PATTER, Volume *1*

Supplemental Information

A Random Matrix Theory

Approach to Denoise Single-Cell Data

Luis Aparicio, Mykola Bordyuh, Andrew J. Blumberg, and Raul Rabadan

METHODS

Introduction to eigenvalue statistics in covariance random matrices

Given a $N \times P$ matrix X, each column is independently drawn from a distribution with mean zero and variance σ, the corresponding Wishart matrix is defined as

$$W = \frac{1}{P} XX^T$$

The eigenvalues λ_i and normalized eigenvectors ψ_i of the Wishart matrix where $i = 1, 2, \dots N$ are given by the following relation:

$$W\psi_i = \lambda_i \psi_i$$

If X happens to be a random matrix (a matrix whose entries x_{ij} are randomly sampled from a given distribution) then W becomes a random covariance matrix and the properties of its eigenvalues and eigenvectors are described by Random Matrix Theory (RMT). In the case of the random distribution being normal with mean 0 and variance 1, one can refer to this as a Wishart ensemble. One of the most interesting properties RMT is the so-called *universality* of the eigenvalue local and global statistics. The global statistics consist of the study of eigenvalue distribution of $\mathcal{O}(N)$ number of eigenvalues. On the contrary, local statistics study the behavior of a small number of eigenvalues, like the distribution of distances between neighboring ordered eigenvalues, the distribution of the largest and smallest eigenvalues and the correlation functions. A property is called universal if it only depends on the symmetry properties that define the ensemble and not on specific details of the underlying probability distribution beyond the first few moments. Universality properties arise both at the local and global scales in the limit $N \to \infty,\ P \to \infty,\ \gamma = \frac{N}{P}$ fixed .

The global statistics are determined by the calculation of the eigenvalue density or empirical density of states:

$$\rho(\lambda) = \langle \frac{1}{N} \sum_{i=1}^{N} \delta(\lambda - \lambda_i) \rangle$$

which for the Wishart matrices converges in the limit $N \to \infty,\ P\ \to \infty,\ \gamma = \frac{N}{P} \leq 1$ to the so-called Marchenko-Pastur (MP) distribution:

$$\rho_{MP}(\lambda) = \frac{1}{2\pi\gamma\sigma^2} \frac{\sqrt{(a_+ - \lambda)(\lambda - a_-)}}{\lambda} \mathbb{I}_{[a_-, a_+]}$$

where

$$a_{\pm} = \sigma^2 (1 \pm \sqrt{\gamma})^2$$

If $N \to \infty,\ P\ \to \infty,\ \gamma = \frac{N}{P} > 1$, then the Marchenko-Pastur (MP) distribution has a delta function centered at zero:

$$\rho_{MP}(\lambda) = \frac{1}{2\pi\gamma\sigma^2} \frac{\sqrt{(a_+ - \lambda)(\lambda - a_-)}}{\lambda} \mathbb{I}_{[a_-, a_+]} + (1 - \frac{1}{\gamma}) \delta(0)$$

where $\mathbb{I}_{[a_-, a_+]}$ means that the distribution has support in the closed interval $[a_-, a_+]$. The parameter σ represents the variance of the probability distribution that generates each element in the random matrix ensemble. In Supplementary Figure 2 there is a graphical representation of the eigenvalue density with two regimes for the ordered eigenvalues: the bulk and edges (largest and smallest eigenvalues). The emergence of MP density is already a form of universality, because the density of eigenvalues is asymptotically the same regardless of the details of the probability distribution of the individual matrix elements. If the first two moments are fixed to be 0 and 1 and the distribution has a sufficient number of finite moments, the universality property of the local statistics of the covariance matrix is satisfied without requiring that the entries of the underlying random matrix be i.i.d. [1,2].

The local statistics of eigenvalues are based on Wigner's original observation concerning the distribution of the distances (gaps) between consecutive eigenvalues, colloquially known as the Wigner Surmise:

$$P(s) \approx \frac{\pi s}{2} e^{\left(-\frac{\pi}{4}s^2\right)} ds$$

where $s = \left(\lambda_j - \lambda_{j-1}\right)/D$ and D is the mean spacing among eigenvalues.

Local universality has been shown in two flavors [1-4]:

- Universality of local eigenvalue statistics:
    1. Bulk universality is the celebrated Wigner-Dyson-Gaudin-Mehta conjecture. It says that regardless of the probability distribution, the n-point correlation functions are described by the sine kernel:

    $$K\left(\frac{\lambda_i}{N}, \frac{\lambda_j}{N}\right) = \frac{\sin \pi(\lambda_i - \lambda_j)}{\pi(\lambda_i - \lambda_j)}$$

    Notice that the kernel does not factorize and therefore shows the strong correlation among eigenvalues. On the other hand, given that the distribution of gaps can be computed from the correlation functions [5], bulk universality also explains the Wigner Surmise-like universality.

    2. Largest eigenvalue universality: the behavior of the largest eigenvalue a_+ is such that

    $$Prob(\lambda_{max} < t) \approx F_\beta(N^{2/3}(t - a_+))$$

where F_β in the Wishart case is a function known as the Tracy-Widom distribution. A similar argument can be made for the smallest eigenvalue a_-.

- Eigenvector delocalization: this property implies that the norm of the eigenvectors ψ_i is equally distributed among all their components α:

$$\left|\psi_i^{(\alpha)}\right| \sim \frac{1}{\sqrt{N}}$$

Introduction to eigenvector localization

Let us now consider the case of perturbed random matrices, i.e. matrices that contain a random part plus a perturbation that partially breaks the randomness in some direction. The spike model of Johnstone provides a simple example where a finite rank perturbation is added to a large random matrix [6]. If the perturbation is below a certain critical value the largest eigenvalue follows the Tracy-Widom distribution of the unperturbed matrix. However, if the perturbation is larger than the critical value, the largest eigenvalue may separate from the bulk MP distribution and present Gaussian fluctuations [7]. This is the so-called BBP (from Baik-Ben Arous-Peche) phase transition. Subsequent works, like [8], have shown that there is no universality in the fluctuation pattern for eigenvalues that separate from the bulk—i.e. the parameters of the distribution of the fluctuation depends on the perturbation features. A similar phase transition was found at the eigenvector level by [9]. Eigenvectors associated with the eigenvalues out of the bulk get localized: the norm gets concentrated in a small number of coordinates containing information about the original perturbation. In Figure 1B (left panels), we show an example of localized and delocalized eigenvectors. The x-axis represents the order of the components, and the y-axis the squares of the components' values $\left(\psi_i^{(\alpha)}\right)^2$. Notice that the eigenvectors are normalized, such that

$$\sum_{\alpha=1}^{N}\left(\psi_i^{(\alpha)}\right)^2 = 1$$

The delocalized -localized phase transition is an example of the famous Anderson localization phase transition that was first observed in quantum disordered systems [10]. It has been observed that, in the delocalized phase, the eigenvalue statistics are governed by RMT.

Interestingly, the distribution of components for delocalized eigenvectors can be easily estimated by considering them as vectors on a unit sphere of dimension N-1. The distribution of their components is then given by

$$f(\psi) = (1-\psi^2)^{\frac{N-3}{2}}$$

that in case of large N approximates a Gaussian distribution with mean zero and 1/N variance (see Figure 1B)

$$f(\psi) \sim \frac{N}{\sqrt{2\pi}} e^{\left(\frac{-N\psi^2}{2}\right)}$$

This can be made more precise in terms of information theory, by noticing that delocalized eigenvectors do not carry any information, while localized vectors are correlated with the original perturbation. This insight proves very useful when attempting to distinguish between noise and signal in any complex system, biological systems in particular. The noise in the data will correspond to the part of the spectrum that can be described in terms of RMT.

In Figures 4A and 5A we implement a Shapiro-Wilk normality test on the coordinates of the eigenvector from two test datasets, allowing us to separate signal from noise. In addition, as a mean to confirm the results, we perform two alternative statistical tests. The first test is based on information theory and its results are shown in Supplementary Figures 3B and 3C. We have compared the Shannon entropy for each eigenvector with randomized data. The second test is based on applications to financial data [11] and the result can be seen in Supplementary Figures 3E and 3F. In this case we are calculating the inverse of the Inverse Participation Ratio (IPR):

$$IPR_i = \sum_{\alpha=1}^{N} \left(\psi_i^{(\alpha)}\right)^4$$

The inverse of the IPR quantifies the number of eigenvector components that contribute significantly. The results are equivalent to those obtained with the other statistical tests.

Sparse Random Matrix ensembles and sparsity induced localization

Sparse Random Matrix Theory (sRMT) ensembles are a class of random matrices with a fraction of non-zero elements, p. In contrast to the RMT's presented in the first section of the Methods, sRMT's exhibit localized eigenvectors, a phenomenon that we call here sparsity induced localization.

The universality properties of covariance sRMT have been studied in [12], [13], [14] and recently in the context of sparse covariance matrices [15]. The local statistics of eigenvalues preserve the bulk and largest eigenvalue universalities. The main difference with non-sparse RMT is the global statistics [16] and the presence of localized eigenvectors [17,18]. Regarding the global statistics, there could be significant deviations from the original MP and Tracy-Widom distributions, depending on the fraction of non-zero values p (Figure 2D). There has been a considerable recent interest in understanding universality properties of sparse random graphs with a fixed sparsity parameter [14,19,20] and it would be interesting to extend this line of work to distributions of sparsity in the covariance matrix. [15]. In Figures 2D we are using sparse ensembles from a mixture of Gaussian distribution, Dirac delta distribution centered at zero and Poisson distribution applied to the randomized dataset [21].

The presence of localized eigenvectors is a very important feature. In the bottom panel of Figure 2A, the correlation between the MP deviations and the presence of localized eigenvectors is shown, by using the Gaussianity test discussed above. In Supplementary Figure 3, we evaluated the localization of eigenvectors using two other tests: Shannon entropy and IPR, previously described. The three tests show sparsity induced localization. This sparsity induced localization is not associated with any biologically relevant information. Sparsity induced localization can introduce artifacts as outliers in PCA and artifactual elbow plots (Figure 2B and 2C). We have also performed a comparison between the sparse dataset and the one after removing sparsity in Figures 2A, 2B and 2C applied to the randomized dataset [21]. Colors distinguish between sparse and clean data. The same comparison has been performed for the original datasets [21] (Figure 4) and [22] (Figure 5).

Algorithm description for denoising of single-cell data

We outline three major steps in the denoising of single-cell data algorithm on the example of PBMC dataset by Kang et al. [21], and illustrate in the Figure 1.

- **Preprocessing**

The goal of preprocessing is to remove genes that create artifacts due to the sparse nature of the data. Gene expression values for each cell were divided by the total number of transcripts and multiplied by 10^6. These values were then log2 transformed adding 1.0 to each value in order to handle with zero entries. After, the single-cell data matrix X is Z-score normalized, such that every gene has mean 0 and standard deviation 1. A randomized matrix is obtained via random permutation of cells for every gene independently, to destroy potential correlations. We project the expression of each gene onto the eigenvector basis of the randomized matrix. To assess normality, we evaluated several related methods: we used the Kolmogorov-Smirnov test, the Anderson-Darling test, and the Shapiro-Wilk test, all providing similar results. In this manuscript, we used the Shapiro-Wilk statistics, comparing to genes that express less than a certain number of transcripts (7 transcripts by default). Genes that have Shapiro-Wilk statistic higher than the minimum statistic of the sparse genes with less that 7 transcripts are considered to be abnormal and are removed from the further analysis. Alternatively, as the p-value is a monotonic function of the Shapiro-Wilk statistic, one can impose an equivalent cut-off on p-value, correcting for multiple hypotheses.

- **Marchenko-Pastur parameter estimation**

After the identification and removal of abnormal genes, the Wishart matrix of the preprocessed data is constructed, and a full set of eigenvalues and eigenvectors is computed using standard Singular Value Decomposition

(SVD) algorithms. The full set of eigenvalues of the Wishart matrix is required to estimate the parameters of the MP distribution. Gradient descent iterative search is implemented to find an optimal fit of the MP distribution with the eigenvalues of the randomized matrix as an initial step in the iterative process. Eigenvalues that fit the MP distribution are regarded as consistent with the noise. Eigenvalues above Tracy-Widom critical eigenvalue are considered to be associated with biological signal.

- **Gene Selection**

To select genes that are the most consistent with biological signal, we analyze the variance of every gene projected onto the signal eigenvectors identified in the previous step and compare it to the largest variance that can be attributed to noise. We project genes onto four subsets of equal size of the eigenvectors of the Wishart matrix in question: signal eigenvectors that correspond to the eigenvalues above the Tracy-Widom critical eigenvalue; eigenvectors right below the critical eigenvalue of Tracy-Widom distribution; eigenvectors of lower spectrum of MP distribution; and an equal number of eigenvectors in the bulk of MP distribution spectrum.

Our goal is to infer the maximum and minimum variance that genes can have due to noise. We select the most variant genes across signal eigenvectors versus noise eigenvectors. Note that these genes are different from the most variant genes across all the eigenvectors in general.

Eigenvectors in the bulk of MP distribution spectrum are considered to be the most compatible with noise. The variance distribution of the genes projected onto these eigenvectors can be modeled using standard χ^2 distribution. The variance distribution of genes projected onto the eigenvectors corresponding to the set of largest eigenvalues of the MP distribution has standard deviation larger than that of χ^2 distribution. The variance distribution of genes projected onto the eigenvectors corresponding to the set of lowest eigenvalues of the MP distribution has standard deviation smaller than that of χ^2 distribution. These variance distributions can be modelled using Gamma distributions. We estimate the parameters of the Gamma distributions and χ^2 distribution using standard maximum likelihood estimation procedure. To select genes, we compare the variance of genes across signal eigenvectors and the right spectrum of MP distribution. We establish the False Discovery Rate FDR (0.001 by default). Genes that have a ratio of variance across the subset of MP eigenvectors right below the Tracy-Widom cut-off (largest variance associated with noise) and across the subset of signal eigenvectors below the FDR are selected.

As a result of the denoising algorithm, the eigenvalues compatible with noise are nullified and genes that have variance across signal eigenvectors compatible with noise are removed (controlled by the free parameter FDR).

Wishart matrix statistics simulations

Single-cell RNAseq simulated dataset (Figure 3) was generated using the Splatter R package. A mean expression level for each gene is simulated using a gamma distribution. The negative binomial distribution is used to generate a count for each cell based on these means, with a fixed dispersion parameter.

The simulation was done for 2000 cells with 10000 genes for Library size (Location, Scale, Norm) 6 groups of cells with the following proportions (0.1, 0.2, 0.3, 0.2, 0.1, 0.1). The following Splatter parameters were used: Mean (Rate, Shape): (0.79, 9.58): (10, 0.69, False). Exprs outliers (Probability, Location, Scale): (0.02, 4.62, 0.91), Diff expr (Probability, Down Prob, Location, Scale): (0.1, 0.5, 0.1, 0.4), BCV (Common DISP, DOF): (0.19, 38.8), Dropout (Midpoint, shape) : (-0.085, -1.14), Paths (From, Length, Skew, Non-linear, Sigma Factor) : (0, 100, 0.5, 0.1, 0.8).

Comparison to other techniques

When comparing with other methods, we are normalizing normalized the data using $log_2(1 + TPM)$ for ZIFA [23], and scImpute [24]. ZIMB-WaVE [25] does not need any normalization. For Seurat [26], we are using the normalization.method = "LogNormalize" with scale.factor = 10000 and finding variable genes using x.low.cutoff = 0.0125, x.high.cutoff = 13, y.cutoff = -10.5. For scImpute, after comparing several combinations, we have decided to use parameters k = 11 and t = 0.5 for the case of dataset [22]. For the dataset [21], we are using k=13 and t = 0.5. MAGIC [27] uses its own normalization and we are using the following parameters: number of PCA dimensions = 20, k = 10 and k_a = 30 (authors recommend 3 times k). Regarding DCA [28], we have used the indications given in the tutorial online, i.e. selection of genes above min_counts=. We used mode='latent' to calculate the latent space and to select the bottle neck dimension hidden_size=(64,dim, 64). For scVI [29] we have used the indication given in the tutorial online, we have made a selection of the 10000 most variant genes and to change the latent space dimension the flag n_latent=dim in VAE class.

For Figures 6A-6D, we have calculated points inside the first 80 dimensions of the latent space for the dataset [21] and first 100 for the dataset [22]. With ZIFA and ZIMB-WaVE, we have used the preprocessing that they indicate by default, and we have calculated the first 15 components due to time limitations. In the y-axis of these figures, we are computing the mean silhouette coefficient [30] for each cell. The silhouette coefficient for a specific cell is given by:

$$s = \frac{b - a}{\max{(a, b)}}$$

where the a is the mean distance between a cell and all the other cells of the same class (the class is defined by the phenotype labels provided in [21,22,31]). Parameter b is the mean distance between a cell and all other cells in the next nearest cluster.

For Supplementary Figures 4, 5 and Supplementary 4, we are using t-SNE representation after using the corresponding method, where we have selected the optimal number of principal components according to Figures 6A-6D.

t-SNE representation

The t-SNE representation were obtained using the default parameters, which are: Learning rate = 1000, Perplexity = 30 and Early exaggeration = 12.

Performance

The most computationally intensive part of our approach relates to the identification of the full set of eigenvectors and eigenvalues. We compared different off-the-shelf SVD approaches [32]. Arpack implementation of standard SVD scales as $\mathcal{O}(N \times P \times k)$, where N is the number of cells, P is number of genes, and k is the number of dimensions. One can see that tThe computational complexity scales as $\mathcal{O}(N^2)$ in our case. One can take advantage of the randomized implementations of SVD [21,22,31] that scale as $\mathcal{O}(N \times P \times \log(k) + (N + P) \times k^2)$, provided one avoids computing all the dimensions and restricts to a small number k. In that case, one canthe cost scales linearly with the number of cells $\mathcal{O}(N)$. Sparse SVD does not provide additional benefits in our scenario, since we sacrifice the sparse inputs, by imposing a Z-score normalization. MP curve fitting converges in a couple of iterations on average and does not present a computational burden.

Data and code availability

Code for the algorithm and denoising pipeline is publicly available on https://rabadan.c2b2.columbia.edu/html/randomly/ .

In Figures 2A, 4 we are using Kang et al. [21] GSE96583 with their labels and a second set of labels (those referred as control by the authors) from Butler et al. [31]. For Figure 5, we used Zeisel et al. dataset GSE60361 and annotation from [22]. For Figure 2E, we are using the following datasets:

- Count matrix for the SMART-Seq2 data set [33] was obtained under the GEO accession number GSE81682.
- Count matrices for CelSeq and CelSeq2 data sets were obtained from accession numbers GSE81076 and GSE86469 correspondingly.

- Count matrix for Fluidigm C1 technology was obtained from accession number GSE86469.
- Hi-C data was obtained from accession number GSE84290.
- Data for the ATAC-seq data was obtained under the accession number GSE65360.
- Raw Nuq-seq data was obtained from the Gene Expression Omnibus with accession number GSE84371.
- Data for bulk RNAseq GBM was obtained from TCGA firehose portal (illuminahiseq_rnaseqv2-RSEM_genes(MD5), http://firebrowse.org/?cohort=GBM&download_dialog=true).
- For the 10x platform and human PBMC dataset, the data was obtained from 10x genomics (https://support.10xgenomics.com/single-cell-gene-expression/datasets/2.1.0/t_4k).

For Supplementary Figure 1, we are using the following datasets:

- Murine embryonic stem cell (mESC) differentiation was obtained from the NCBI Gene Expression Omnibus (GEO) database, with accession number GSE94883.
- High grade glioma dataset was taken from GSE103224.
- Kang et al. [21].
- Pancreas islet single-cells dataset GSE84133.

REFERENCES

1 Pillai, N. S. & Yin, J. (2012). Edge Universality of Correlation Matrices. Ann Stat 40, 1737-1763.

2 Pillai, N. S. & Yin, J. (2014). Universality of Covariance Matrices. Ann Appl Probab 24, 935-1001.

3 Tao, T. & Vu, V. (2012). Random Covariance Matrices: Universality of Local Statistics of Eigenvalues. Ann Probab 40, 1285-1315.

4 Erdos, L., Schlein, B., Yau, H. T. & Yin, J. (2012). The local relaxation flow approach to universality of the local statistics for random matrices. Ann I H Poincare-Pr 48, 1-46.

5 Bourgade, P., Erdos, L. & Yau, H. T. (2014). UNIVERSALITY OF GENERAL beta-ENSEMBLES. Duke Math J 163, 1127-1190.

6 Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. Ann Stat 29, 295-327.

7 Baik, J., Ben Arous, G. & Peche, S. (2005). Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. Ann Probab 33, 1643-1697.

8 Capitaine, M., Donati-Martin, C. & Feral, D. (2009). The Largest Eigenvalues of Finite Rank Deformation of Large Wigner Matrices: Convergence and Nonuniversality of the Fluctuations. Ann Probab 37, 1-47.

9 Benaych-Georges, F. & Nadakuditi, R. R. (2011). The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. Adv Math 227, 494-521.

10 Anderson, P. W. (1958). Absence of Diffusion in Certain Random Lattices. Phys Rev 109, 1492-1505.

11 Plerou, V. *et al.* (2002). Random matrix approach to cross correlations in financial data. Phys Rev E 65.

12 Huang, J. Y., Landon, B. & Yau, H. T. (2015). Bulk universality of sparse random matrices. J Math Phys 56.

13 Mirlin, A. D. & Fyodorov, Y. V. (1991). Universality of Level Correlation-Function of Sparse Random Matrices. J Phys a-Math Gen 24, 2273-2286.

14 Lee, J. O. & Schnelli, K. (2018). Local law and Tracy-Widom limit for sparse random matrices. Probab Theory Rel 171, 543-616.

15 Hwang, J. Y., Lee, J. O. & Schnelli, K. (2019). Local Law and Tracy-Widom Limit for Sparse Sample Covariance Matrices. Ann Appl Probab 29, 3006-3036.

16 Rodgers, G. J. & Bray, A. J. (1988). Density of states of a sparse random matrix. Phys Rev B Condens Matter 37, 3557-3562.

17 Fyodorov, Y. V. & Mirlin, A. D. (1991). Localization in ensemble of sparse random matrices. Phys Rev Lett 67, 2049-2052.

18 Evangelou, S. N. & Economou, E. N. (1992). Spectral Density Singularities, Level Statistics, and Localization in a Sparse Random Matrix Ensemble. Physical Review Letters 68, 361-364.

19 Yukun He, A. K., Matteo Marcozzi (2018). Local law and complete eigenvector delocalization for supercritical Erdős-Rényi graphs. arXiv:1808.09437

20 Erdos, L., Knowles, A., Yau, H. T. & Yin, J. (2013). Spectral Statistics of Erdos-Renyi Graphs I: Local Semicircle Law. Ann Probab 41, 2279-2375.

21 Kang, H. M. *et al.* (2018). Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. Nat Biotechnol 36, 89-94.

22 Zeisel, A. *et al.* (2015). Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science 347, 1138-1142.

23 Pierson, E. & Yau, C. (2015). ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome Biol 16.

24 Li, W. V. & Li, J. Y. J. (2018). An accurate and robust imputation method scImpute for single-cell RNA-seq data. Nat Commun 9.

25 Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S. & Vert, J. P. (2018). A general and flexible method for signal extraction from single-cell RNA-seq data. Nat Commun 9.

26 Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. Nature Biotechnology 33, 495-U206.

27 van Dijk, D. *et al.* (2018). Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. Cell 174, 716-+.

28 Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S. & Theis, F. J. (2019). Single-cell RNA-seq denoising using a deep count autoencoder. Nat Commun 10, 390.

29 Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. Nat Methods 15, 1053-1058.

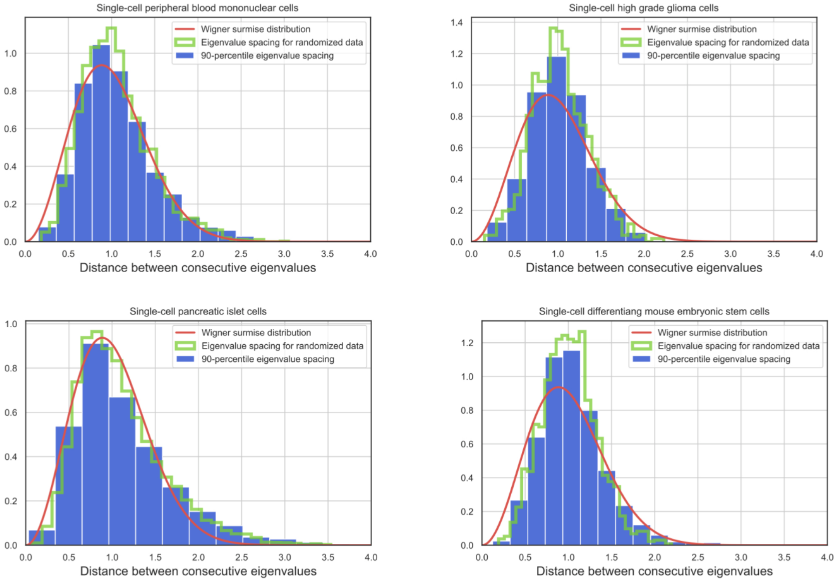
30 Rousseeuw, P. J. (1987). Silhouettes - a Graphical Aid to the Interpretation and Validation of Cluster-Analysis. J Comput Appl Math 20, 53-65.

31 Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nature Biotechnology 36, 411-+.

32 Halko, N., Martinsson, P. G. & Tropp, J. A. (2011). Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions. Siam Rev 53, 217-288.

33 Nestorowa, S. *et al.* (2016). A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. Blood 128, E20-E31.

Suplementary Figure 1



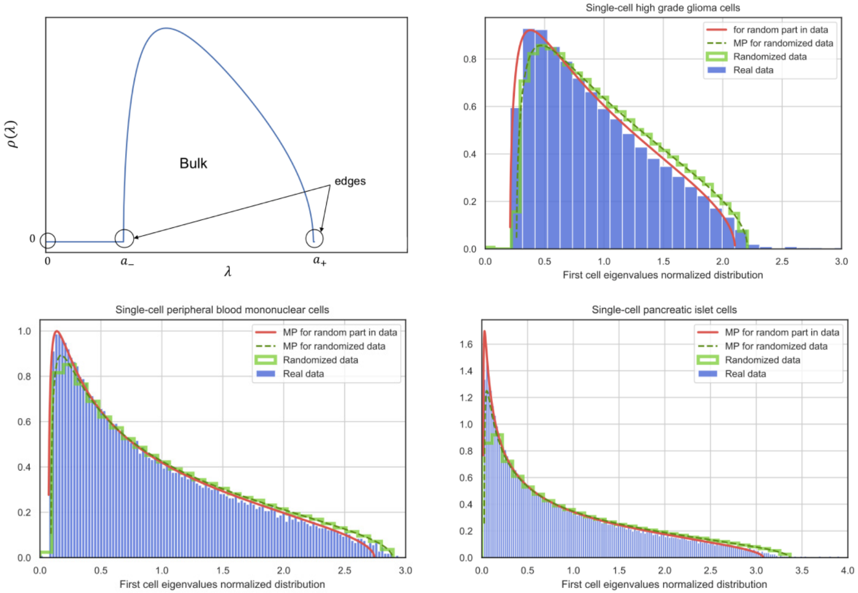
Single-cell peripheral blood mononuclear cells

Single-cell high grade glioma cells

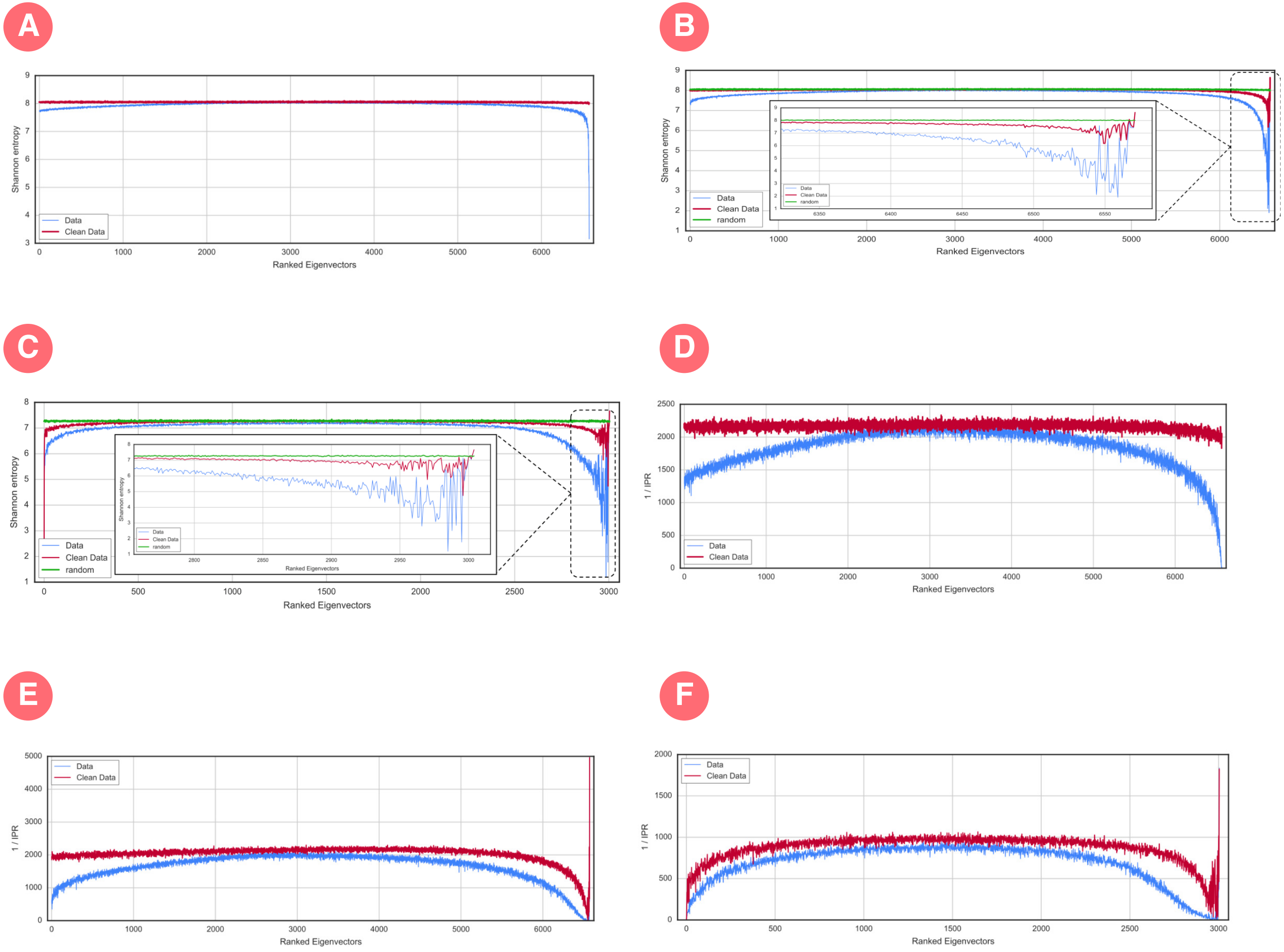
Single-cell pancreatic islet cells

Single-cell differentiang mouse embryonic stem cells

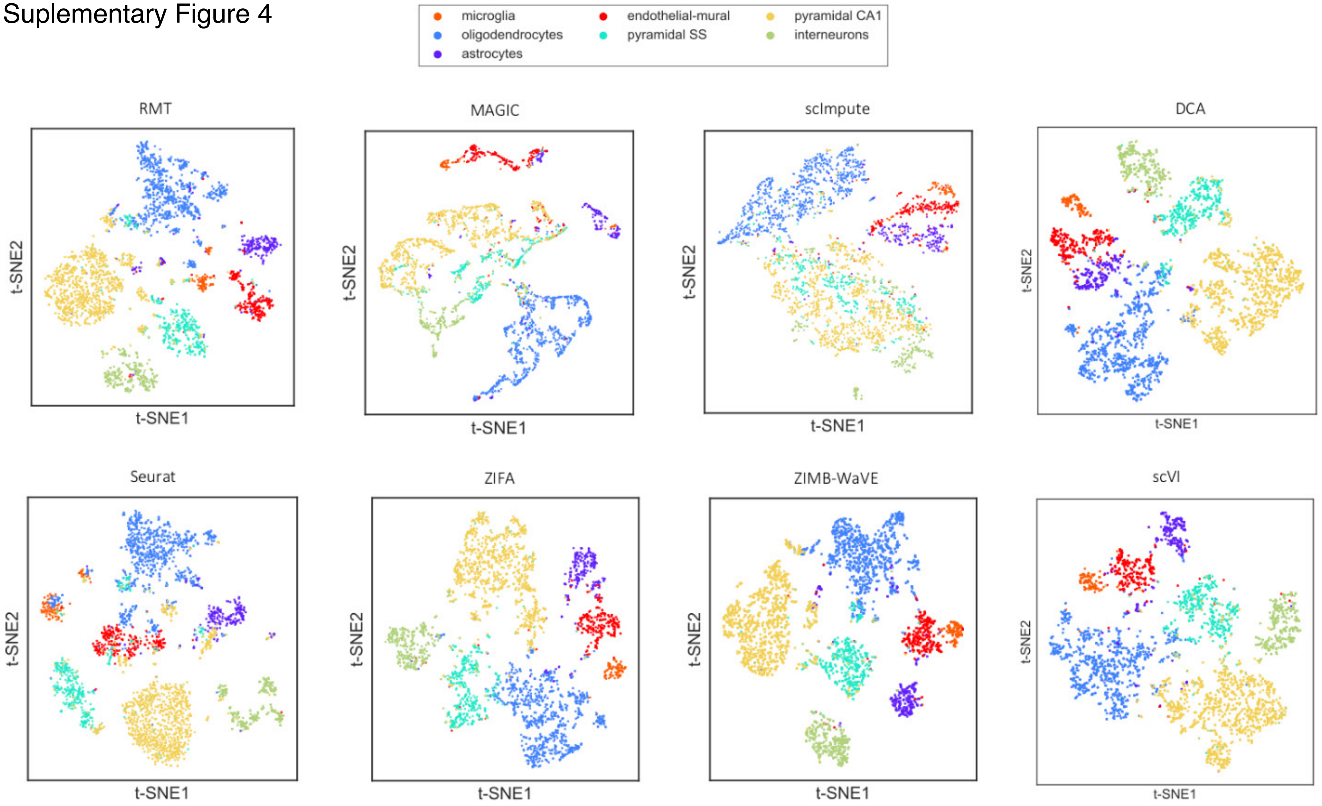
Suplementary Figure 2



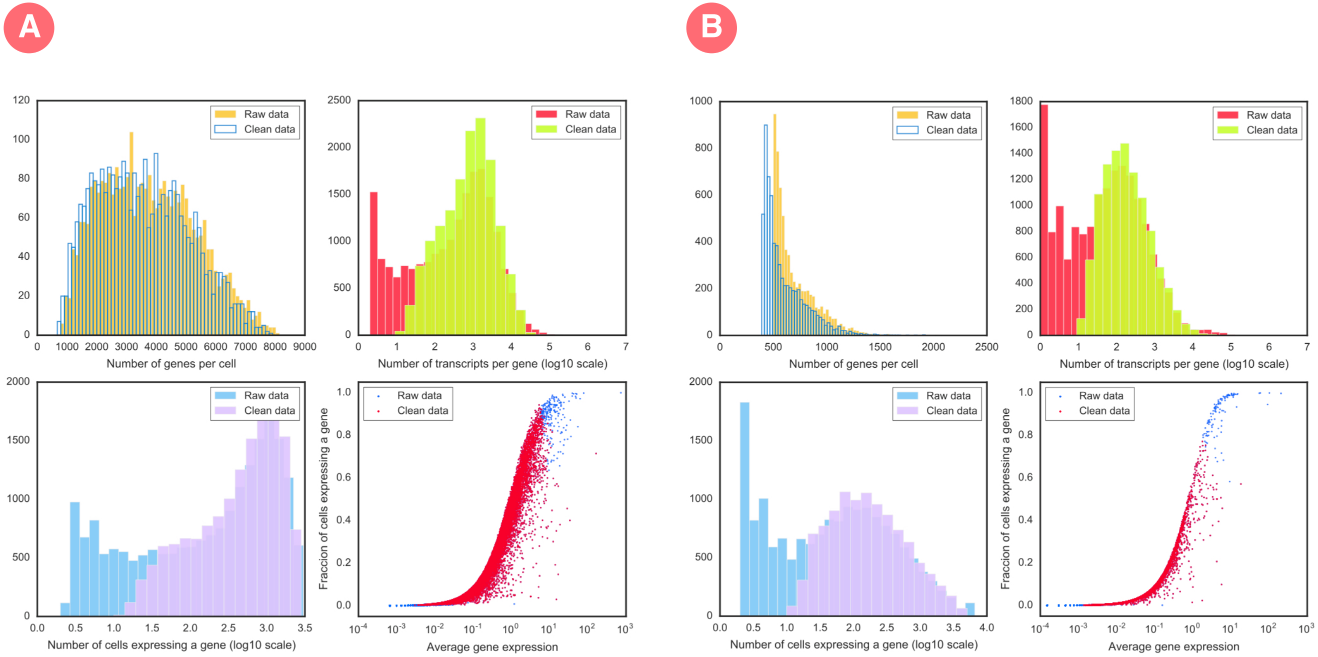
Suplementary Figure 3



Suplementary Figure 4



Suplementary Figure 5



SUPPLEMENTARY FIGURE LEGENDS

Supplementary Figure 1. Distribution of spacing between consecutive Wishart matrix eigenvalues across experiments, and comparison with Wigner surmise.

Supplementary Figure 2. Schematic representation of the Marchenko-Pastur distribution and distribution of density of eigenvalues for the Wishart matrix across experiments and comparison with Marchenko-Pastur distribution.

Supplementary Figure 3. **A)** Calculation of Shannon Entropy for the randomized Kang et al. [21,22,31] dataset. This is another way of expressing the same phenomenon in Figure 2A. When the system is sparse (blue) there are eigenvectors whose entropy decreases. That is a sign of information contained in these eigenvectors. **B)** Calculation of the Shannon entropy for the eigenvectors of the PBMC [21,22,31] dataset and **C)** eigenvectors in a single-cell dataset of mouse cortex cells [21,22,31]. The blue (red) line corresponds to the system before (after) cleaning the sparsity. For completeness, a comparison with the non-sparse randomized dataset (green line) is plotted. **D)** Calculation of the Inverse participation ratio (IPR) for the randomized Kang et al. [21,22,31] dataset. The participation ratio indicates the number of cell covariates that take part for each eigenvector. When sparsity is non-negligible in the random system (blue), there are eigenvectors which have fewer cell covariates. **E)** Calculation of the Inverse participation ratio (IPR) for the PBMC cells, Kang et al. [21,22,31] and **F)** for the mouse cortex cells [21,22,31].

Supplementary Figure 4. Comparison of the t-SNE representation for different public algorithms. This case corresponds to 7 different PBMC cell-phenotypes sequenced in [21] and described in [31].

Supplementary Figure 5. Comparison of some statistics before and after cleaning the sparsity. From left to right and from up down: Number of genes per cell, Number of transcripts per gene, Number of cells expressing a gene and Ratio of cells expressing each gene versus the average gene expression. The analysis for **A)** the case of the PBMC [21] dataset and **B)** the case of mouse cortex cells [21,22,31].