Unravelling cell fate decisions through single cell methods and mathematical models

Mircea, M.

2

Phiclust: A clusterability measure for single-cell transcriptomics reveals phenotypic subpopulations

The ability to discover new cell phenotypes by unsupervised clustering of single-cell transcriptomes has revolutionized biology. Currently, there is no principled way to decide whether a cluster of cells contains meaningful subpopulations that should be further resolved. Here we present phiclust (ϕ_{clust}), a clusterability measure derived from random matrix theory, that can be used to identify cell clusters with non-random substructure, testably leading to the discovery of previously overlooked phenotypes.

This chapter is based on **Mircea, M.**, Hochane, M., Fan, X., Chuva de Sousa Lopes, S.M., Garlaschelli, D., Semrau S., Phiclust: a clusterability measure for single-cell transcriptomics reveals phenotypic subpopulations. Genome Biology 23, 18 (2022). https://doi.org/10.1186/s13059-021-02590-x [1]

2.1 Background

Unsupervised clustering methods [2–5] are integral to most single-cell RNA-sequencing (scRNA-seq) analysis pipelines [6], as they can reveal distinct cell phenotypes. Importantly, all existing clustering algorithms have adjustable parameters that have to be chosen carefully to reveal the true biological structure of the data. If the data is over-clustered, many clusters are driven purely by technical noise and do not reflect distinct biological states. If the data is under-clustered, subtly distinct phenotypes might be grouped with others and will thus be overlooked. Furthermore, most analysis pipelines rely on qualitative assessment of clusters based on prior knowledge, which can hinder the discovery of new phenotypes. To assess the quality of a clustering quantitatively and help choose optimal parameters, some measures of clustering quality and clusterability have been proposed [7], most of which are not directly applicable to scRNA-seq data. For example, some existing methods rely on multimodality of the expression matrix, which is not always justified for scRNA-seq data, especially when considering highly dynamic systems. Other methods have input parameters, such as the optimal number of dimensions for dimensionality reduction, that cannot be easily determined. Also, general methods do not explicitly account for uninformative sources of variability, related to cell cycle progression or the stress response, for example, which can be important confounders. In the context of scRNA-seq, one of the most widely used measures is the silhouette coefficient [8]. This measure requires the choice of a distance metric to compute the similarity between cells. Notwithstanding its usefulness, it cannot be excluded that a partition of random noise obtains a high silhouette coefficient, indicating high clustering quality. Other measures based on distance metrics or the fit of probability densities suffer from similar issues and often only provide binary results instead of a quantitative score [9]. A different approach is pursued by ROGUE [10], a recently developed tool to assess clustering quality specifically in scRNA-seq data. ROGUE applies the concept of entropy on a per-gene basis to quantify the mixing of cell types. While a clear improvement over existing methods, ROGUE depends on a challenging step of selecting informative genes to explain the differences between cell types. It also assumes a particular noise distribution and requires the careful choice of an adjustable parameter. Here we present phiclust, a new clusterability measure for scRNA-seq data that addresses some of the shortcomings of existing methods. This measure is based on the angle ϕ between vectors of the noise-free signal and the measured, noisy signal. We consider clusterability to be the theoretically achievable agreement with the unknown ground truth clustering, for a given signal-to-noise ratio. (Below, we will describe in detail how we define "signal" and "noise" in this context.) Importantly, our measure can estimate the level of achievable agreement without knowledge of the ground truth. High clusterability (phiclust close to 1) means that multiple phenotypic subpopulations are present and that clustering algorithms should be able to distinguish them. Low clusterability (phiclust close to 0) means that the noise is too strong for even the best possible clustering algorithm to find any clusters accurately. If phiclust equals 0, the observed variability within a cluster is consistent with random noise. Any subclusters of such a cluster still have a phiclust of 0, which prevents over-clustering of random noise. Instead of assuming a certain noise distribution or relying on a selection of informative genes, our measure can be applied to arbitrary types of random noise and includes all genes in the analysis. This is made

possible by certain universal properties of random matrix theory (RMT) [11], which has been applied successfully in finance [12], physics [13] and recently also scRNA-seq data analysis [14].

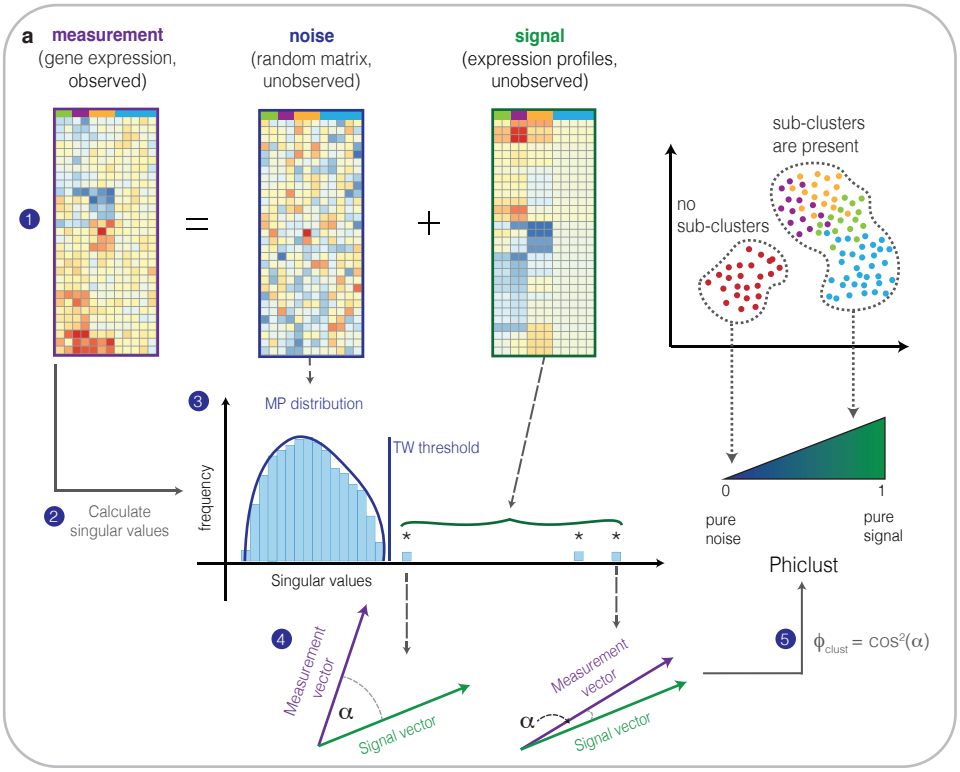


Figure 2.1: Concept of phiclust a Scheme illustrating the rationale behind phiclust.

Below we will use results of RMT on the singular value decomposition (SVD) of a single-cell gene expression matrix, where rows correspond to genes and columns correspond to cells. To get an intuitive understanding of RMT, it is useful to first consider the cell-cell correlation matrix, calculated from the gene expression profiles. We start from the null hypothesis that the data does not contain any structure and is produced by a random process. In the context of single-cell transcriptomics, "structure" means multiple, distinguishable clusters of cells, or phenotypes. RMT can predict, what the correlation matrix looks like, if the entries of the gene expression matrix are samples of random variables that are independent and identically distributed. Trivially, the diagonal elements of this correlation matrix are all equal to 1. The off-diagonal elements are not exactly 0, however, despite the absence of any meaningful structure in the data. Only in the limit of measuring an infinite number of (random) genes would the off-diagonal elements become identically 0 and the correlation matrix would become the identity matrix. In that case, the only eigenvalue of the correlation matrix is 1. RMT describes the properties of a correlation matrix for a finite ratio of cells

and genes. These correlation matrices are, in a sense, distributed "around" the identity matrix, which corresponds to an eigenvalue spectrum distributed around 1. Although the individual entries of the correlation matrix fluctuate from realization to realization, RMT shows that the eigenvalue spectrum is robust (a so-called "self-averaging" property) and an analytical expression for it can be obtained [15]. Likewise, RMT predicts that the singular value distribution of a purely random matrix is closely approximated by the Marchenko-Pastur (MP) distribution. This result holds true irrespective of the distribution of the random variable. This universal property of random matrices allows us to apply RMT to gene expression matrices obtained by scRNA-seq. Of course, any biologically interesting scRNA-seq measurement should contain structure, usually in the form of cell clusters. RMT allows us to regard singular values lying above the MP distribution as evidence for the rejection of the null hypothesis (i.e., the absence of structure in the data). The MP distribution is characterized by sharp upper and lower limits for the singular values of a random matrix, but is strictly valid only in the limit of infinite numbers of genes and cells (while keeping the cell-gene ratio fixed). For finite matrices, the largest and smallest singular values are distributed around those sharp limits, which is described by the Tracy-Widom distribution [16]. As explained above, the presence of structure manifests itself as singular values above the MP distribution (i.e., the prediction for a purely random matrix). Qualitatively, the magnitude of those outlying singular values corresponds to the magnitude of the differences between clusters. We can understand this relationship, if we assume that the measured gene expression matrix is the sum of a random matrix (the "noise") and a matrix of noise-free gene expression profiles (the "signal"), see Fig. 2.1. The bigger the difference in gene expression between phenotypes, the larger the magnitude of the non-zero singular values of the signal matrix. If the number of non-zero singular values (i.e., the rank of the signal matrix) is small compared to the dimensions of the matrix, low-rank perturbation theory [17] is applicable. This theory allows us to calculate the singular values of the measured gene expression matrix from the singular values of the signal matrix. Remarkably, knowledge of the complete signal matrix is not required for this calculation. phiclust is meant to help identify non-random (or deterministic) structure. At the level of a complete data set, for example of a complex tissue, clusters are typically easily discernible. However, if we zoom in on a single cluster, it is much more difficult to decide, whether the variability within that cluster corresponds to meaningful sub-structure (such as the presence of multiple phenotypes) or is consistent with random noise. Below, we will precisely define a notion of clusterability, based on the adjusted rand index, and show that it strongly correlates with phiclust. Furthermore, we will demonstrate that our measure compares favorably to the silhouette coefficient and ROGUE on simulated data and experimental data sets with known ground truth. (See Table S1 for a list of all used simulated and experimental data sets.) Finally, we will apply phiclust to scRNA-seq measurements of complex tissues and obtain new biological insights, which we validate with follow-up measurements.

2.2 Results

2.2.1 Phiclust is derived from first principles and does not have free parameters

To derive phiclust, we considered the measured gene expression matrix as a random matrix perturbed by the unobserved, noise-free gene expression profiles (Fig. 2.1). This is the exact opposite of the conventional approach, which considers random noise as a perturbation to a deterministic signal. Note that, in our approach, the random matrix contains both the biological variability within a phenotype as well as the technical variability (which is due to limited RNA capture and conversion efficiency, for example). Our point of view allows us to leverage well-established results from RMT [14, 18] and perturbation theory [17].

Figure 2.2 illustrates the basic principles that were applied to derive phiclust. RMT predicts that the SVD of a random noise matrix results in normal distributed singular vectors and a distribution of singular values that is closely approximated by the MP distribution, if the matrix is large enough (Fig. 2.2a, left column). Here, we consider the noise-free gene expression profiles of the cells in various phenotypes, as the "signal" that perturbs the random matrix and thus its singular value distribution. Since biological and technical variability are lumped into the random matrix, expression profiles are identical for cells that belong to the same phenotype. For example, in the case of two distinct phenotypes, the signal matrix has only one non-zero singular value (Fig. 2.2a, middle column). The observed (or measured) gene expression matrix is obtained as the sum of the random noise matrix and the noise-free gene expression profiles (2.2a, right column). The singular value distribution of the measured expression matrix has exactly one singular value above the upper limit that the theory predicts for a purely random matrix, the Tracy-Widom (TW) threshold. The outlying singular value and its associated singular vector correspond to the deterministic component of the measured expression matrix. The distribution of the remaining singular values (the "bulk") is still closely approximated by the MP distribution. Importantly, as the perturbation becomes larger, the value of the outlying singular value also increases (Fig. 2.2b). A larger perturbation means more distinct and therefore more easily clusterable phenotypes (compare the singular vectors in the middle row of Figs. 2.2a and b). The basic idea of phiclust is to use the magnitude of the outlying singular values to quantify clusterability.

Due to the universality of RMT, all described principles are independent of the particular noise distribution (see Fig. 2.2a-b for normal distributed noise and Fig. 2.2c-d for Poisson distributed noise). SVD of appropriately preprocessed real data sets therefore leads to singular value distributions with the same shape as obtained in simulations: a bulk closely approximated by the MP distribution and one or multiple outlying values. We found that data preprocessing has to comprise normalization and log-transformation, as well as gene-wise and cell-wise scaling (Fig. 2.3 a-d). SVD of raw data or log-transformed, normalized data typically results in a largest outlying singular value that is much larger than all others (Fig. 2.3 a,b). The corresponding singular vector reflects a global trend in the data and is called "market mode" in the context of stock market analysis [12, 19].

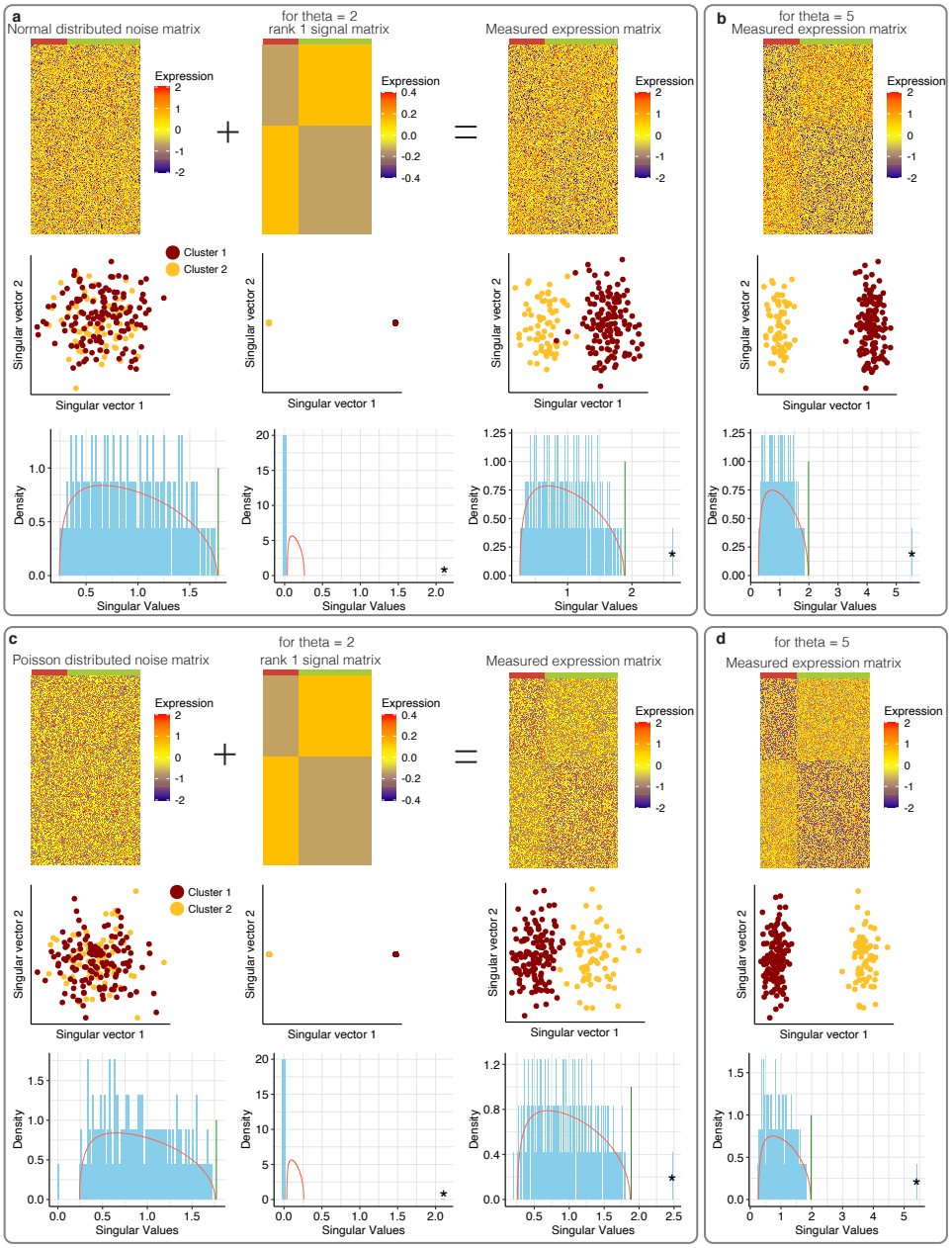


Figure 2.2: Basic principles of random matrix theory and perturbation theory a Top row: heatmaps of a random matrix drawn from a normal distribution, a rank 1 signal matrix with a singular value θ of 2, and the resulting expression matrix. Middle row: Singular vectors of the corresponding matrices. Bottom row: histograms of the corresponding singular values. Red line: MP distribution, green line: TW threshold. b Heatmap, singular vectors and singular values of an expression matrix constructed as in a, except the singular value of the signal matrix was 5. c,d Matrices, singular vectors and singular values obtained as in a and b, but the random matrix was drawn from a Poisson distribution.

Here, we call it "transcriptome mode", since it corresponds to an expression trend that is present across all cells, irrespective of cell type (such as, for example, high expression of particular cytoskeletal genes or essential enzymes and low expression of certain membrane receptors or transcription factors). The transcriptome mode is obviously not informative for clustering. Scaling shifts its singular value to 0, which effectively removes it from further analysis (Fig. 2.3c,d).

We tested for all data sets used in this study, whether the bulk of the singular value distribution of each cluster deviates significantly from the MP distribution after the described preprocessing (Kolmogorov-Smirnov test, Fig. 2.3e). For reasonably large clusters (> 50 cells), we only found one example of a (marginally significant) deviation from the MP distribution.

We next wanted to confirm, for real data, that the remaining outlying singular values reflect the strength of the signal, i.e., differences between the phenotypes. To that end, we extracted the gene expression profiles from two clusters in an experimental single-cell RNA-seq data set and added, as additional signal, a matrix with one non-zero singular value. As to be expected, SVD of the combined data results in one additional singular value, which increases with the strength of the perturbation (Fig. 2.3f-g). See Table S2 for a list of all outlying singular values of experimentally measured expression matrices as well as the corresponding signal matrices. All in all, these tests show that the basic principles of random matrix theory and perturbation theory are applicable to real single-cell RNA-seq data.

So far, we have shown that the values of the outlying singular values are, qualitatively, related to the differences between phenotypes. However, their magnitudes are difficult to interpret. Phiclust is derived from the outlying singular values and can be interpreted as a measure of clusterability, as we will show in the next section. More specifically, phiclust is defined as the squared cosine of the angle between the leading singular vector of the measured gene expression matrix and the corresponding singular vector of the unobserved, noise-free expression matrix. Low-rank perturbation theory is able to predict this angle using only the dimensions of the measured gene expression matrix and its singular values, but without knowledge of the noise-free expression profiles. See Additional File 2 for a detailed derivation. If the noise level is low compared to the signal, this angle will be small, since the measured gene expression matrix is then very similar to the noise-free signal. This would result in phiclust close to 1. As the level of noise increases, for a fixed signal, the singular vectors of the measured expression matrix and the noise-free signal become increasingly orthogonal and phiclust approaches 0. To illustrate the calculation of phiclust, we simulated data sets with realistic noise structure using the Splatter package [20] (Fig. 2.4a,b). As to be expected, increasing the number of genes that are differentially expressed between clusters makes the clusters more easily separable and leads to larger singular values outside of the MP distribution (Fig. 2.4a). By construction, this results in higher values of phiclust (Fig. 2.4b). Please refer to Table S2 for the numerical values of the outlying singular values in the simulated expression matrices as well as the corresponding signal matrix.

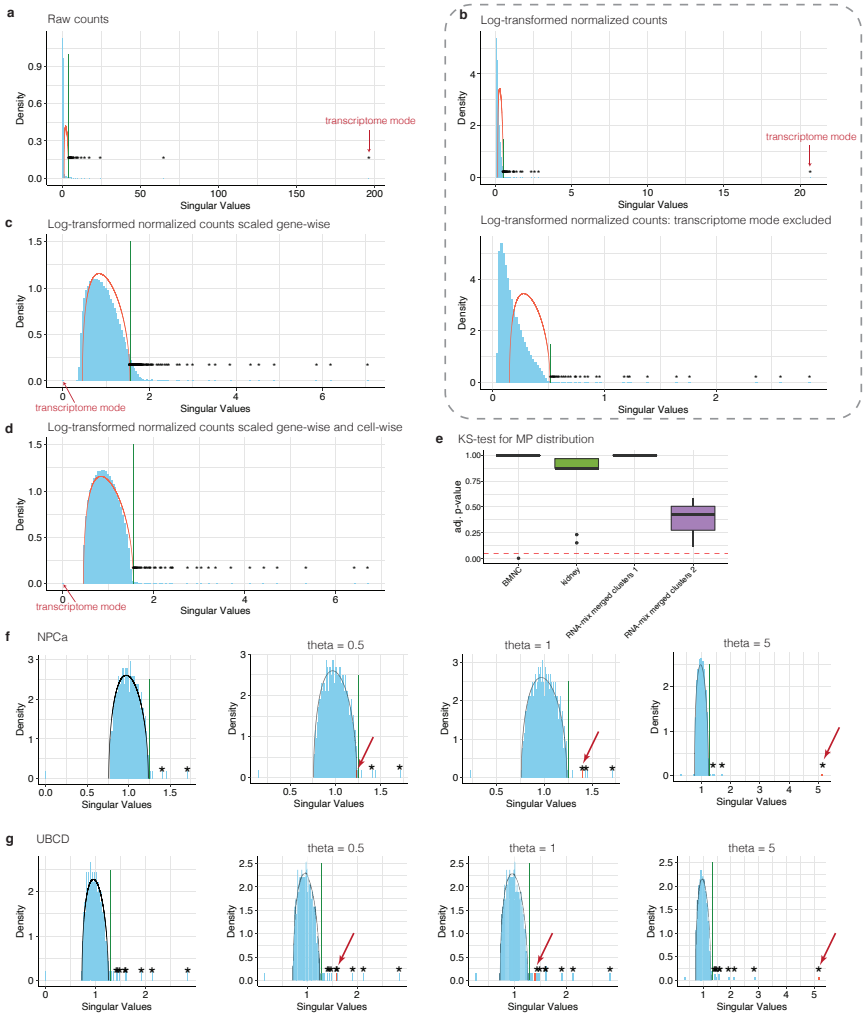


Figure 2.3: Importance of preprocessing for MP fit and effect of perturbation on singular value distribution Singular value (SV) distributions of the fetal kidney single-cell RNA-seq data set after different preprocessing steps. a Raw UMI counts. Arrow indicates transcriptome mode b Log-transformed, normalized UMI counts. Arrow indicates transcriptome mode. Right: Transcriptome mode was excluded. c Log-transformed, normalized data as in b, that was additionally centered gene-wise. The transcriptome mode, visible as the highest singular value in a and b appears close to 0 (indicated by the arrow). d Log-transformed, normalized, and gene-wise standardized data, as in c, that was additionally standardized cell-wise. The SV distribution coincides with the bulk of the MP distribution. This is not a fit: The MP distribution is completely determined by the dimensions of the matrix and has no free parameters. e Kolmogorov-Smirnov (KS) test of a significant difference between the bulk of the singular value distributions and the MP distribution. The boxplot shows the adjusted p-values of the KS test for each cluster per data set. Red dashed line indicates an adjusted p-value of 0.05. f Histogram of singular values for the NPCa cluster of the fetal kidney data set. Left: original values. Rest: Singular values of the NPCa expression matrix plus a rank 1 perturbation with increasing magnitudes of the perturbation (singular value θ of the perturbation = 0.5, 1 or 5). The red arrow indicates the singular value that stems from the additional perturbation. g Histogram of singular values for the UBCD cluster of the fetal kidney data set. Left: original values. Rest: Singular values of UBCD expression matrix plus a rank 1 perturbation with increasing magnitudes of perturbation (singular value θ of the pertrubation = 0.5, 1 or 5). The red arrow indicates the singular value that stems from the additional perturbation.

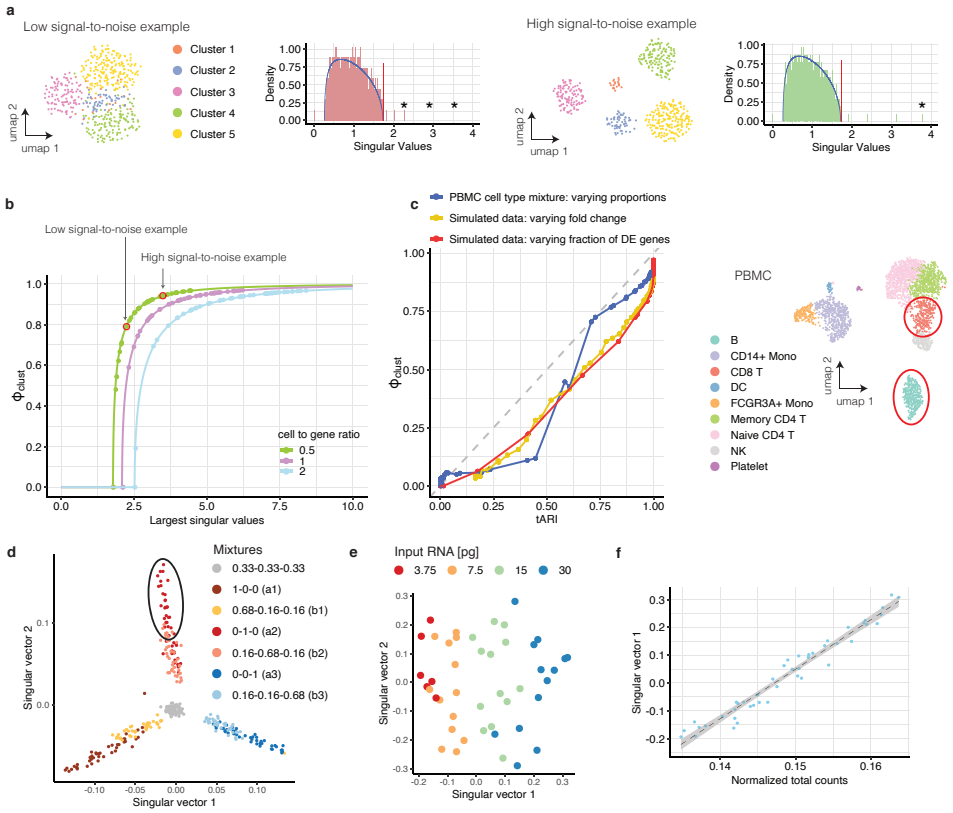


Figure 2.4: Phiclust is a proxy for the theoretically achievable adjusted rand index (tARI). a Singular value distributions of simulated data sets with 5 clusters and different levels of noise; Red: low signal-to-noise, Green: high signal-to-noise. The MP distribution is indicated by a solid blue line, the TW threshold is indicated by a red solid line and significant singular values are highlighted with asterisks. Inserts show UMAPs of the data. The data set with a higher signal-to-noise ratio has more significant singular values and those singular values are bigger. b Value of the largest singular value versus for simulated data. Arrows indicate where the examples from panel b are located. The relationship between the largest singular values and phiclust only depends on the dimensions of the expression matrix. Simulations with different cell-to-gene ratios are shown in different colors. c Phiclust versus theoretically achievable ARI (tARI). Red data points: Simulated data sets with two clusters. The number of differentially expressed (DE) genes was varied, the log fold change between clusters was fixed. Green data points: Simulated data sets with two clusters. The mean log fold change between clusters was varied, the number of differentially expressed genes was fixed. Blue data points: Two synthetic clusters were created by weighted averages of cells from two clusters in the PBMC data set. Cluster weights were varied. The grey dashed line indicates identity. Inset: UMAP of PBMC data set with the two clusters used indicated by red solid circles. d scRNA-seq of mixtures of RNA extracted from three different cell lines. Each data point is a mixture. For each mixture the entries of the first two singular vectors are plotted. Colors indicate different ratios of contributions from the three cell lines. e First two singular vectors of the cluster indicated by a black solid ellipse in f. The amount of mRNA per mixture [pg] is indicated in color. g Normalized total counts per mixture versus first singular vector of the cluster shown in g. Linear regression (dashed line) is used to regress out the correlation with the total counts. Grey area indicates standard deviation.

We would like to stress at this point that phiclust is derived from universal properties of perturbed random matrices, which can be considered first principles. By contrast, many other measures are developed based on empirical observations and justified post hoc by their usefulness. Phiclust is calculated using only the SVD and the dimensions of the expression matrix. Thus, it does not have any free, adjustable parameters, which would have to be chosen by the user or learned from the data.

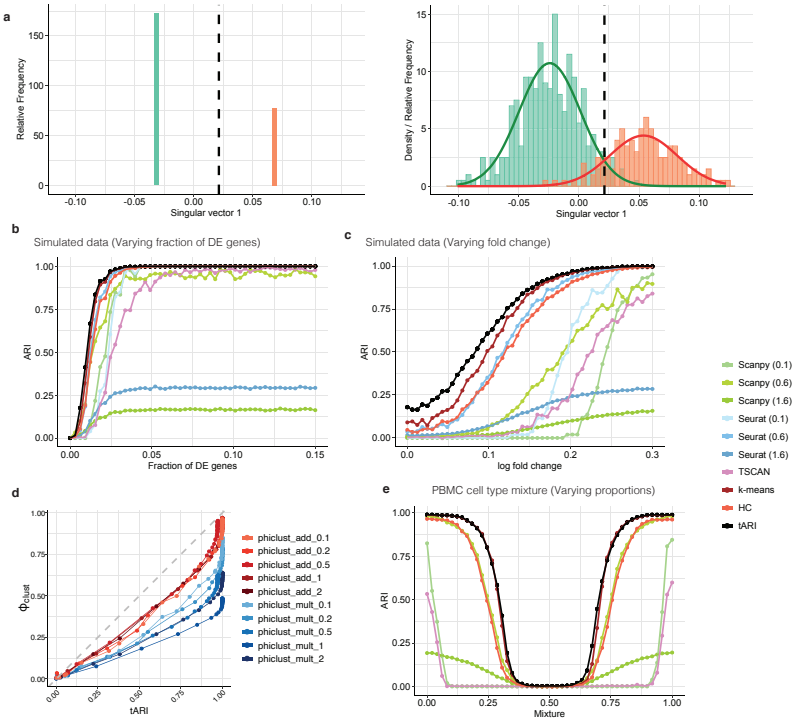


Figure 2.5: An upper limit to the achievable ARI can be estimated using a Bayes classifier. a Left: Histogram of the noise-free singular vector for a scenario with two clusters (or phenotypes) . Only the first singular vector is significant. The dashed line indicates a possible decision boundary. Right: Histogram of the first singular vector in the presence of noise. The color indicates to which simulated (ground truth) cluster the cells belong. Two normal distributions fitted separately to the singular vector entries belonging to the two clusters are shown as solid lines. The Bayesian error rate is estimated from the overlap of these two distributions and used to calculate the theoretical ARI (tARI). The dashed line indicates the optimal decision boundary. b ARI achieved by various clustering methods compared to the ground truth and tARI for simulated data with two clusters. The number of differentially expressed genes was varied. c ARI achieved by various clustering methods compared to the ground truth and tARI for simulated data with two clusters. The mean log fold change between clusters was varied. d tARI versus phiclust for simulated data sets with two clusters and different fractions of DE genes. Red curves: Values of phiclust for additive perturbation at different cell to gene ratios. Blue curves: Values of phiclust for multiplicative perturbation at different cell to gene ratios. Dashed grey line indicates diagonal. e ARI achieved by various clustering methods compared to the ground truth and tARI for PBMC cell type mixtures. Two synthetic clusters were created by weighted averages of cells from two clusters in the PBMC data set (see Fig. 1d). The mixture proportions were varied from 0 to 1. b,c,e The numbers in the legend indicate the resolution parameter used.

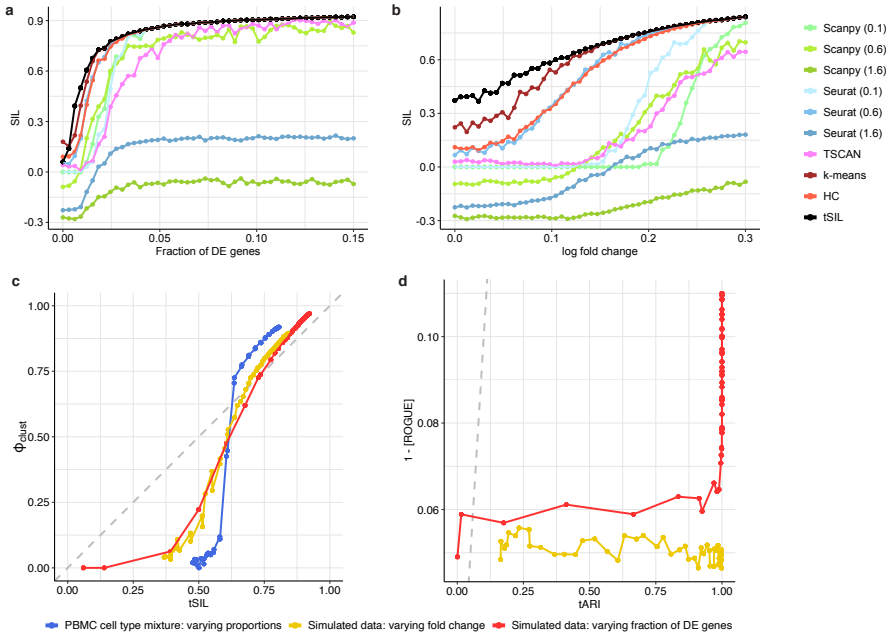


Figure 2.6: An approximate upper limit to the best possible silhouette coefficient and accordance of ROGUE with tARI. a Silhouette coefficient (SIL) achieved by various clustering methods and theoretical SIL (tSIL) for simulated data with two clusters. The number of differentially expressed (DE) genes was varied. b SIL achieved by various clustering methods and tSIL for simulated data with two clusters. The mean log fold change between clusters was varied. a,b The numbers in the legend indicate the resolution parameter used. c tSIL versus phiclust. Red data points: Simulated data sets with two clusters. The number of DE genes was varied, the log fold change between clusters was fixed. Green data points: Simulated data sets with two clusters. The log fold change between clusters was varied, the number of DE genes was fixed. Blue data points: Two synthetic clusters were created by weighted averages of cells from two clusters in the PBMC data set (see Fig. 2.4c). Cluster weights were varied. The Grey dashed line indicates identity. d tARI versus 1 - [ROGUE] score. Red data points: Simulated data sets with two clusters. The number of DE genes was varied, the log fold change between clusters was fixed. Green data points: Simulated data sets with two clusters. The log fold change between clusters was varied, the number of DE genes was fixed. Blue data points: Two synthetic clusters were created by weighted averages of cells from two clusters in a PBMC data set (see Fig. 2.4c). Cluster weights were varied. The Grey dashed line indicates identity.

2.2.2 Phiclust is a proxy for clusterability

To show that phiclust is a proxy for clusterability, we have to make the concept of clusterability more precise and quantifiable. We adopted the Adjusted Rand Index (ARI) [21] as a well-established measure for the agreement between an empirically obtained clustering and the ground truth. Next, we will argue that perfect agreement with the ground truth (ARI = 1) is not achievable in the presence of noise, even with the best conceivable clustering algorithm.

Take, for instance, the simplest possible case of two cell types, A and B. Without any noise (technical or biological), expression profiles within a cell type are identical and the data can be clustered perfectly. Correspondingly, the singular vector of the expression matrix has only two different entries Fig. 2.5a, left). Therefore, it is easy to find a threshold

that discriminates between the two cell types. In the presence of noise, however, there is a chance that the measured expression profile of a cell from cell type A looks more like cell type B and is therefore clustered with other cells from cell type B and vice versa. Correspondingly, the entries of the singular vector are now spread by the noise and can overlap (Fig. 2.5a, right). Even if we use the best possible threshold to discriminate between the two cell types, some cells will be necessarily misclassified, if the distributions overlap. This type of error is unavoidable (or irreducible) and known as Bayes error rate [22] in the context of statistical classification. From the overlap of the singular vector entries, we can calculate the Bayes error rate or, equivalently, a theoretically achievable ARI (tARI, see also Additional File 2). Of course, this is only possible for data with known ground truth. We first used simulated data to show empirically that commonly used clustering methods are not able to exceed the tARI (Fig. 2.5b,c). It therefore quantifies our notion of clusterability: With increased noise, tARI decreases and it is more challenging even for the best conceivable clustering algorithm to distinguish the difference between phenotypes. Importantly, phiclust is strongly correlated with the tARI (Fig. 2.4d) and thus allows us to estimate clusterability without knowing the ground truth.

So far, we have assumed additive noise (i.e., the measured gene expression is the sum of a random matrix and the noise-free expression matrix). Low-rank perturbation theory also makes a prediction for multiplicative noise (i.e., the measured gene expression is the product of a random matrix and the noise-free expression matrix). In that case, phiclust still scales approximately linearly with the tARI, but its dynamic range depends somewhat on the cell-to-gene ratio (Fig. 2.5d). To our knowledge, the noise generating mechanisms at work in scRNA-seq have not been pinpointed comprehensively. Therefore, we will continue to assume additive noise, noting that our approach can be easily adapted to multiplicative noise.

To test the relationship between phiclust and the tARI in experimentally measured data, we used an scRNA-seq data set of peripheral blood mononuclear cells (PBMCs) [23]. We chose two very distinct cell types and created new clusters as weighted, linear combinations of expression profiles from the two cell types. This approach allowed us to precisely control the difference between the newly created clusters, while maintaining the experimentally observed noise structure (Fig. 2.5e). Also for this data, phiclust strongly correlates with the tARI (Fig. 2.4d). As an alternative to the tARI, we also calculated the theoretically achievable silhouette coefficient [8] (tSIL), which considers the distances between the best possible clusters (Fig. 2.6 a-c). For a large range of simulation parameters, the tSIL has a smaller dynamic range than the tARI, which makes it less useful overall for assessing clusterability. In contrast to phiclust, ROGUE [10] does not show collinearity with the tARI (Fig. 2.6d). Therefore, ROGUE seems to implement a notion of clusterability that is distinct from our point of view.

2.2.3 Confounder regression removes unwanted variability

To further characterize the performance of phiclust on experimental data sets with known ground truth, we used a measurement of purified RNA from 3 cell types, mixed at different ratios [24] (Fig. 2.4e). We noticed a significant correlation between the amount of input RNA and the entries of the first singular vector of individual clusters (Fig. 2.4f). This might be explained by lowly expressed genes not being well-represented in the low-input libraries,

and the resulting differences in the expression profiles. In any case, the amount of input RNA seemed to be a confounding factor that could lead to high values of phiclust, even in the absence of meaningful subclusters. Correspondingly, we found a correlation between the singular vector entries and the number of total counts, despite normalization of the data (Fig. 2.4g). This is consistent with the finding that total counts are a confounding factor in scRNA-seq data that cannot be eliminated by normalization using one single scaling factor per cell [23, 25]. Different groups of genes scale differently with the total counts per cell. Therefore, a correlation with the total counts remains even after normalization.

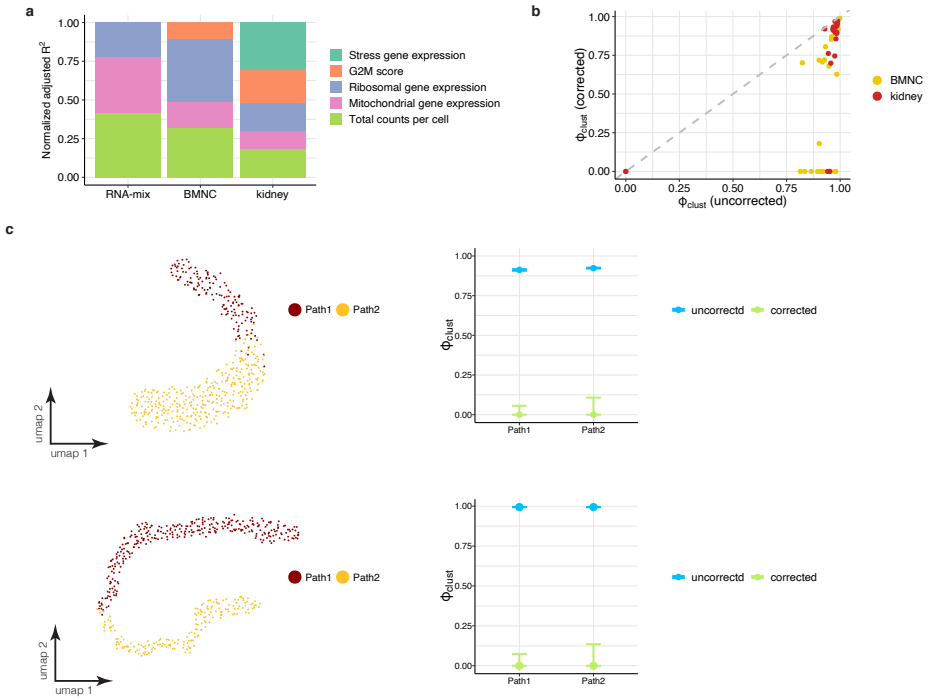


Figure 2.7: Correcting for nuisance parameters and unwanted variability. a Summary of adjusted R2 for several nuisance parameters in all experimental data sets. b Original (uncorrected) phiclust values vs phiclust corrected for the influence nuisance parameters in the BMNC and fetal kidney data sets. Dashed grey line indicates diagonal. c Two examples of differentiation paths with different numbers of differentially expressed genes. Left: UMAPs of simulated data sets with two differentiation paths. Right: Original (uncorrected) values of phiclust and phiclust values corrected by confounder regression using pseudotime as the only confounder.

More generally, there are various experimental and biological factors that drive artefactual or irrelevant variability in single-cell RNA-seq data [23, 26]. We therefore introduced a regression step that removes the influence of any nuisance variables, such as the number of total counts per cell, ribosomal gene expression, mitochondrial gene expression or cell cycle phase (see also Additional File 2). More specifically, we first regress the entries of a singular vector on one or multiple confounders. The fraction of variance explained by all confounder is then given by the adjusted R2 (coefficient of determination) of the linear

regression. Since the squared singular values can also be interpreted as the amount of variance explained, we correct them by multiplying with 1 − the adjusted R2 found in the confounder regression. (See Table S2 for a list of the uncorrected and corrected singular values for both simulated and experimental expression matrices.) The corrected singular values are then used to calculate phiclust.

Interestingly, the relative influence of the confounders considered in this study varied substantially between data sets (Fig. 2.7 a). For example, cell stress is a relevant confounder only in the kidney data set. This is likely related to the cell dissociation procedure, which is necessarily more aggressive for kidney tissue, compared to the other samples: bone marrow mononuclear cells (BMNCs) and purified RNA, extracted from cell cultures. Total counts and ribosomal gene expression explain most of the artefactual variance in BMNCs. This might be explained by high variability in the metabolic state of the cells. In Table S2 we list the R2 values of each considered confounder for each cluster. For real scRNA-seq data sets, confounder regression can lead to a significant reduction of phiclust (Fig. 2.7b, see Table S2 for the numerical values.) It is therefore an important part of the method.

Confounder regression can also help to analyze data sets that are not made up of regular clusters but contain irregularly shaped continua of gene expression. For example, in developmental and stem cell biology we commonly observe differentiation paths, which are large clusters with gradually changing expression profiles. Uncorrected phiclust values are high for such paths, which suggests meaningful subpopulations (2.7c,d). Depending on the biological question, it might in fact be desirable to cluster differentiation paths, for example, to separate a stem cell state from a differentiated cell type. For other applications, it could be preferable to treat a differentiation path as one cluster. In that case we can use pseudotime approaches [27] to infer a temporal order of the gene expression profiles and use the inferred pseudotime in the confounder regression. If all observed variability is explained by developmental dynamics, phiclust is reduced to 0 and thus no sub-clustering is suggested (2.7c,d).

2.2.4 Phiclust has high sensitivity for the detection of substructure

After correction for unwanted variability, we compared the performance of phiclust with other clusterability measures in the RNA mixture data set (Fig. 2.4e). Phiclust successfully indicated the presence or absence of subclusters for all tested combinations of the 7 original mixtures (Fig. 2.8). By contrast, ROGUE only indicated the presence of substructure when the merged clusters were very clearly distinguishable (Fig. 2.8 b,c). The silhouette coefficient was qualitatively similar to phiclust but its dynamic range was much smaller (Fig. 2.8, middle row). This might become critical in the case of highly similar phenotypes, which is precisely where phiclust might have an advantage. An example for this can be seen in Fig. 2.8b: the silhouette coefficients in the pure cluster are very similar to the merged clusters (which were composed of two original clusters). To compare phiclust with the silhouette coefficient in more detail, we carried out additional simulations (Fig. 2.9). First, we simulated 3 clusters and subsequently merged two of them. While phiclust clearly distinguished the merged cluster from the pure cluster, the silhouette coefficients were similar for both. Increasing the fraction of genes that are differentially expressed between the merged cluster increased the difference in silhouette coefficient, but only

gradually (Additional file 1: Fig. S7b). By contrast, phiclust jumped to values close to 1 for the merged cluster for very small fractions of differentially expressed genes (around 0.03). It is therefore the more sensitive measure. The silhouette coefficient strongly depends on the number of principal components used in dimensionality reduction (Fig. 2.9c), as well as the metric for distances between expression profiles (Fig. 2.9d). Phiclust does not depend on such user-defined parameters.

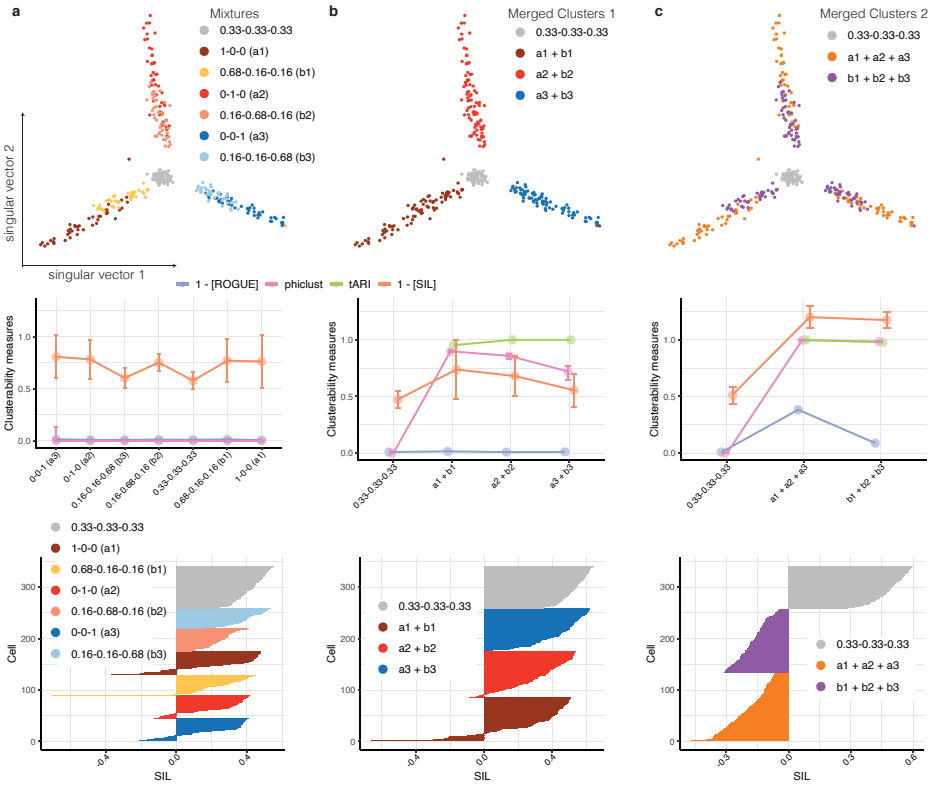


Figure 2.8: Phiclust outperforms other measures on experimental data. Clusters of mixtures of RNA extracted from three different cell lines were merged in different ways to vary the amount of variability in each merged cluster. Top: first two singular vectors of RNA mixture data. Colors indicate different ratios of contributions from the three cell lines. Middle: The values of phiclust (rose), 1 – silhouette coefficient [SIL] (orange), tARI (green) and 1 - [ROGUE] (blue) for each corresponding cluster. For the calculation of the error bars, see Methods. Bottom: Bar plot of silhouette coefficients for each cell, sorted by cluster. a Original RNA mixture. b Merged clusters. Red: 0-1-0 merged with 0.16-0.68-0.16. Blue: 0-0-1 merged with 0.16-0.16-0.68. Green: 1-0-0 merged with 0.68-0.16-0.16. c Violet: merged cluster contains mixtures 0.68-0.16-0.16, 0.16-0.68-0.16 and 0.16-0.16-0.68. Orange: merged cluster contains mixtures 1-0-0, 0-1-0, and 0-0-1.

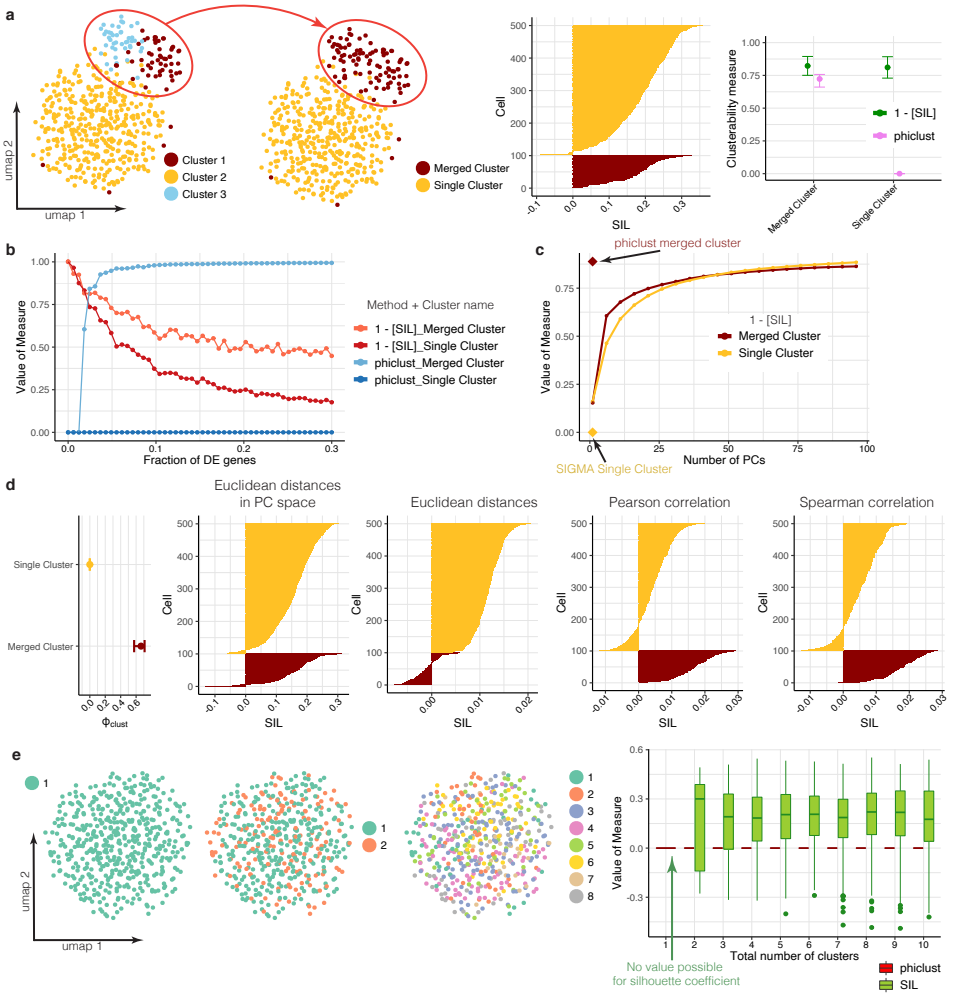


Figure 2.9: Phiclust outperforms the silhouette coefficient on simulated data. a Left: UMAP of 3 simulated clusters. Two clusters were merged to one, resulting in two clusters in total. Middle: Bar plot of silhouette coefficients for each cell, sorted by cluster. Right: phiclust value and average silhouette coefficient for each cluster. b Simulation of clusters as shown in a with different fractions of differentially expressed genes. Phiclust (blue) and average silhouette coefficient (red) for merged cluster and single cluster. c Simulation of clusters as shown in a. The average silhouette coefficient per cluster was calculated for Euclidean distances using different numbers of principal components. Value of phiclust for each cluster is indicated by diamond-shaped data points. d Simulation of clusters as shown in a. Leftmost graph: Value of phiclust for each cluster. Other graphs: Bar plot of silhouette coefficients for each cell, sorted by cluster, calculated with the following distance metrics: Euclidean distances in principal component space, Euclidean distances in the original space, Pearson correlation and Spearman correlation. e Simulation of a cluster consistent with random noise. Clustering was performed with k-means to obtain 2 to 10 clusters. Left: UMAP of simulated data with 0, 2 and 8 clusters. Right: Boxplot of the values of phiclust and the silhouette coefficient for each k-means clustering.

Most importantly, the silhouette coefficient cannot answer the question, whether an identified cluster contains meaningful substructure, as it requires partitioning into at least 2 sub-clusters. We simulated a cluster without any substructure and all variability was purely random (Fig.2.9e). The silhouette coefficient was maximal for a k-means clustering with k=2, which might prompt a user to conclude (wrongly) that there are 2 sub-clusters present. Phiclust, which does not require any further partitioning of the cluster, was 0, indicating correctly that the observed variability was consistent with random noise. All in all, these comparisons indicate that phiclust is a sensitive measure, which detects differences between highly similar phenotypes.

2.2.5 Genes responsible for the detected substructure can be identified

In full analogy to the reasoning outlined so far, our approach can also be used to characterize variability in gene space, for which we defined the conjugate measure g-phiclust (see Additional File 2 for the derivation). Above, we considered only the right singular vectors, where each entry corresponds to a cell in the data set. We therefore also call them "cell-singular vectors". In the simplest case of well separated clusters, entries in the cell singular vectors indicate the membership of a cell in a cluster or a group of clusters. For the left singular vectors, each entry corresponds to a gene. Therefore, we also call them "gene-singular vector". The squared cosine of the angle between the leading gene-singular vector in the measured gene expression matrix and the corresponding gene-singular vector of the noise-free signal matrix is g-phiclust . As for phiclust, data sets with higher signal-to-noise ratios are characterized by higher values of g-phiclust (Fig. 2.10a). "Signal" and "noise" are defined exactly as above: "noise" is a random matrix and the "signal" is a low-rank matrix consisting of noise-free expression profiles, where the strength of the signal (or difference between the clusters) corresponds to the magnitude of the non-zero singular values. A g-phiclust close to 0 would indicate that all observed differential gene expression can be explained by random noise. Larger values of g-phiclust indicate less overlap of the gene expression profiles between phenotypes. We therefore expect to find a bigger number of significantly differentially expressed (DE) genes and/or larger fold changes between phenotypes. We confirmed by simulations that genes with larger absolute entries in a gene-singular vector contribute more to the differences between the clusters separated along the corresponding cell-singular vector (Fig. 2.10b-d): For example, if two clusters (A and B) are separated along a cell-singular vector and cells in cluster A are characterized by positive entries, the genes with large positive entries in the corresponding gene-singular vector will be mostly expressed in cluster A. We call these "variance driving" genes. Our approach thus not only identifies relevant substructure in a cell cluster but can also reveal the genes responsible for it. In contrast to differential expression tests, the variance driving genes can be obtained before clustering and might help the user interpret the observed variability and make an informed decision on whether it is useful to sub-cluster the data. If the variance driving genes have enriched biological features (such as being involved in the same signaling pathway or cellular function), we can take that as additional evidence for biologically meaningful sub-population.

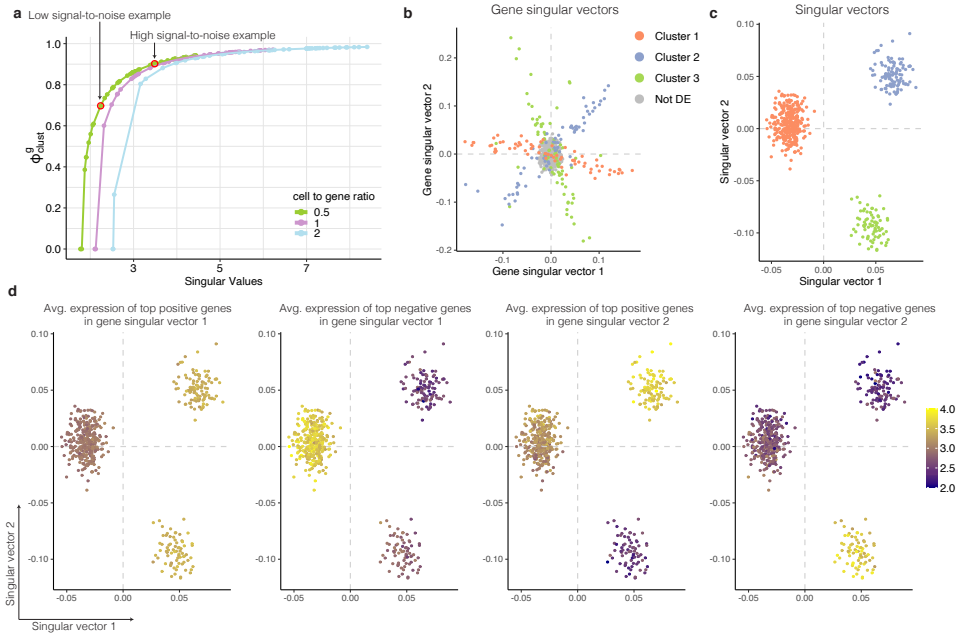


Figure 2.10: Variance-driving genes identified in gene singular vectors coincide with differentially expressed genes in a simulated data set. Genes with high absolute values in the gene singular vector contribute the most to the variability. a Value of the largest singular value versus the squared cosine of the angle between the gene singular vector of the signal matrix and the gene singular vector of the measured expression matrix (g-phiclust) in simulated data. Arrows indicate examples shown in Figure 1b. b First two gene-singular vectors. Differentially expressed genes of each cluster are indicated by color. c First two (cell-)singular vectors for the simulated data set shown in panel b. Dashed grey lines indicate the 0 value on each of the axes. Cell clusters are indicated by color. d First two singular vectors as in c. Dashed grey lines indicate the 0 value on each of the axes. The average log-transformed expression of the top 1% genes driving the variance is indicated by color. The 4 panels show, respectively, from left to right: genes corresponding to the highest values in gene singular vector 1, genes corresponding to the lowest values in gene singular vector 1, genes corresponding to the highest values in gene singular vector 2, and genes corresponding to the lowest values in gene singular vector 2.

2.2.6 Application of phiclust to a BMNC data set drives the discovery of biologically meaningful sub-clusters.

The most important application of phiclust, in our opinion, is to prioritize clusters for further sub-clustering and follow-up studies. For a complex tissue with dozens of clusters, it is not feasible to sub-cluster all of them and try to validate all resulting subpopulations. This is particularly inefficient, if many subclusters are in fact just driven by random noise. High values of phiclust nominate those clusters that likely have deterministic structure and are therefore worthwhile to be scrutinized experimentally in more detail. To demonstrate the application of phiclust and g-phiclust, we analyzed scRNA-seq measurements of complex tissues. In a data set of bone marrow mononuclear cells (BMNCs) [28] we calculated phiclust for the clusters reported by the authors (Fig. 2.11a,b).

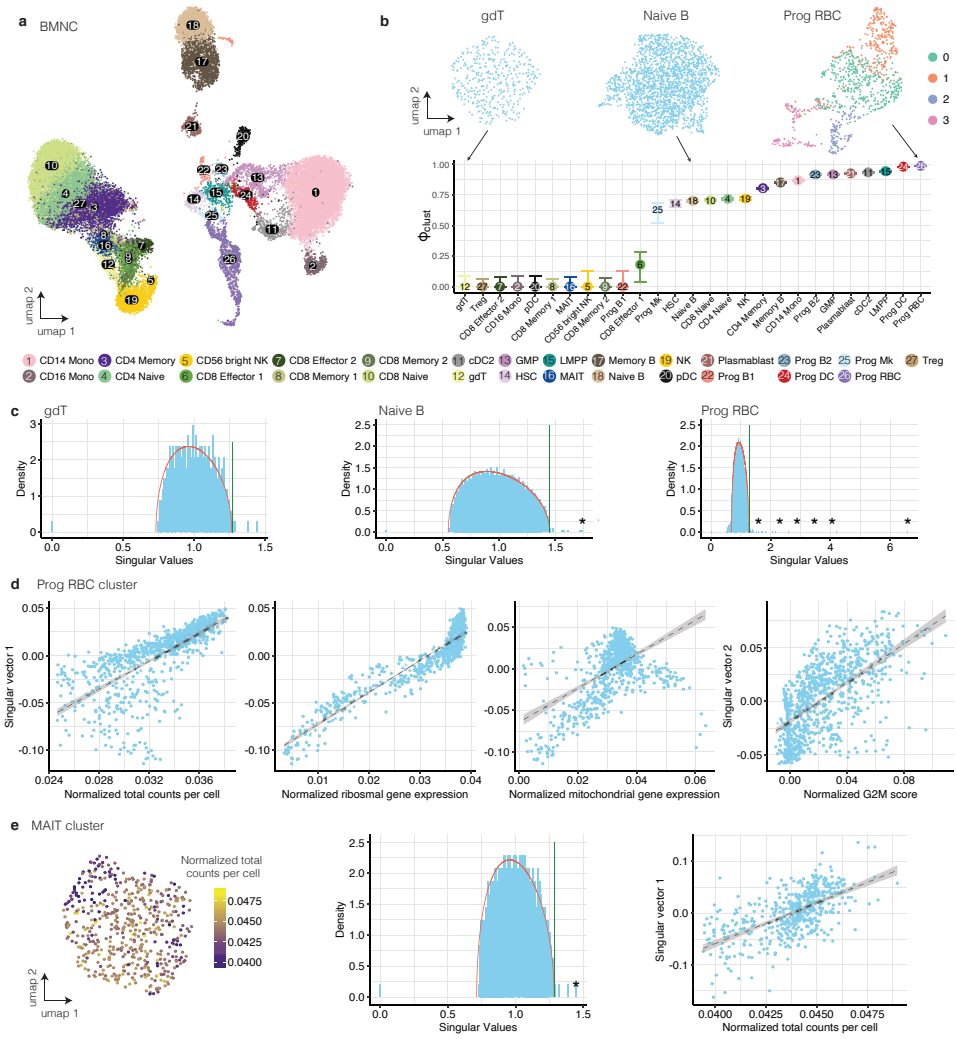
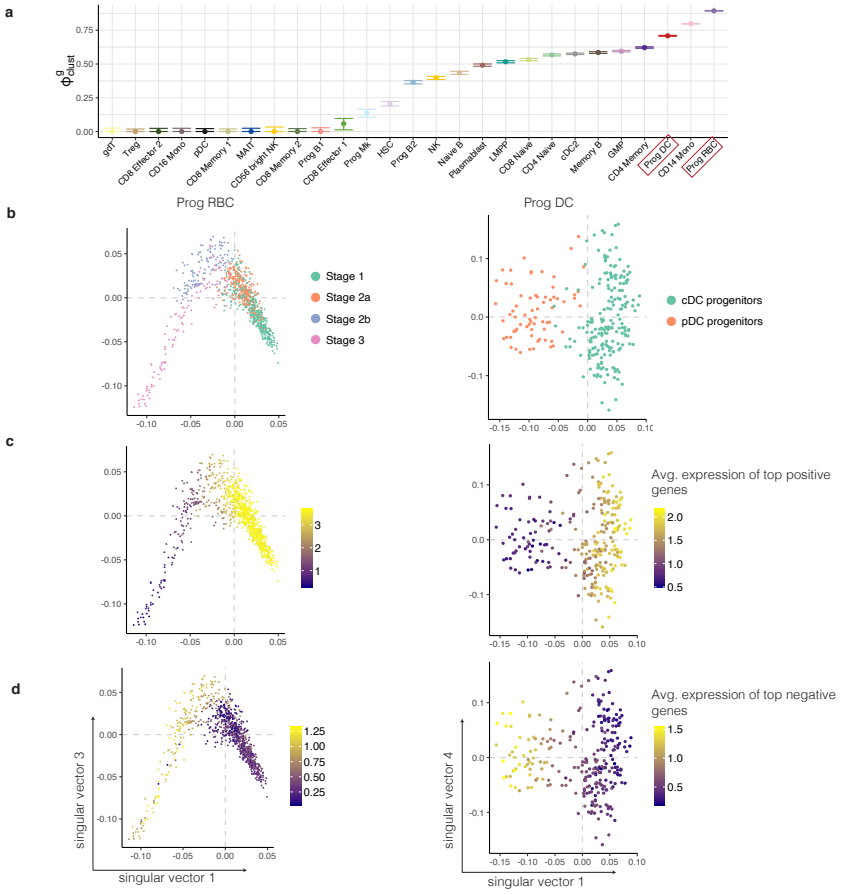


Figure 2.11: Application of phiclust to a BMNC data set drives the discovery of biologically meaningful sub-clusters. a UMAP of BMNC data set. b Phiclust for the BMNC data set. Error bars indicate the uncertainty obtained by resampling the noise. Inset: UMAP of clusters with low, intermediate, and high values of phiclust. c Singular value distribution, MP distribution (red line) and TW threshold (green line) of clusters with low, intermediate, and high values of phiclust. Significant singular values are highlighted with asterisks. In the gdT cluster, the singular vectors corresponding to the outlying singular values had normal distributed entries and were thus not significant. d First three graphs: First singular vector of the red blood cell progenitor cluster in the BMNC data set versus normalized total counts per cell, normalized expression of ribosomal genes, and normalized expression of mitochondrial genes. Rightmost graph: Second singular vector versus normalized G2M score. The dashed line indicates the linear regression and the grey area indicates the standard deviation. e Left: UMAP of the MAIT cell cluster in BMNC data set. The color indicates the normalized total counts per cell. Middle: singular value distribution, MP distribution (red line) and TW threshold (green line) for the MAIT cell cluster. The only significant singular value is indicated by an asterisk. Right: Normalized total counts per cell versus the singular vector associated with the significant singular value (here: first singular vector) in the MAIT cluster.



top positive genes Stage1 / Stage2a	top negative genes Stage2b / Stage3
RPS2	HBA2
RPL37A	HBM
RPLP1	SNCA
PTMA	BPGM
RPS3	HBD
RPL27A	HBA1
RPLP0	SLC4A1
RPL3	ALAS2
EEF1A1	TMCC2
RPS16	GYPB
RPL14	IFIT1B
RPL8	PDZK1IP1
RPL13A	SLC25A37
RPS18	TSPO2
RPS6	HEMGN
RPS9	KRT1
RPS23	HBB
RPS5	LGALS3
RPL36	BNIP3L
RPS17	NCOA4

top positive genes cDC progenitors	top negative genes pDC progenitors
LYZ	IGLL1
S100A6	IRF8
S100A4	JCHAIN
LST1	TCF4
COTL1	ADA
CSTA	MZB1
TSPO	CCDC50
FTL	UGCG
MNDA	FAM129C
TMSB10	SCT
S100A10	NUCB2
S100A11	PLAC8
S100A9	ITM2C
CEBPD	IRF7
SPI1	STMN1
ANXA1	SPIB
RNASE2	SOX4
SH3BGRL3	APP
RPL8	C12orf75
ANXA2	TSPAN13

Figure 2.12: Congruence between variance-driving genes and differentially expressed genes between sub-clusters in a BMNC data set. (Caption on the next page)

For all clusters, except the red blood cell (RBC) progenitor cluster, the bulk of the singular value distribution was well-described by the MP distribution. (In the RBC progenitors, we found several singular values below the lower limit of the MP distribution. These outliers did not influence the further analysis since we are only interested in singular vectors above the upper limit.) The first cell-singular vectors of all clusters were significantly correlated with several confounding factors (see Fig. 2.11d for RBC progenitors and Fig. 2.11e for MAIT cells). After correction for these confounding factors, phiclust corresponded well with a visual inspection of the cluster UMAPs (Fig. 2.11b): Where obvious clusters were present, phiclust was highest, while homogeneous, structure-less clusters resulted in a phiclust of 0. Reassuringly, many progenitor cell types received a high phiclust (indicating possible substructure) in agreement with the known higher variability in these cell types. Ranking existing clusters by g-phiclust resulted in a very similar order (Fig. 2.12a).
To confirm the presence of relevant substructure, we subclustered the two original clusters with the highest phiclust (Fig. 2.12 b-e). In the RBC progenitors, we identified 4 subsets that correspond to different stages of differentiation, ranging from erythroid precursors to highly differentiated RBCs, as identified previously [29]. In the dendritic cell (DC) progenitor cluster, two subclusters were identified, which correspond to precursors of classical or plasmacytoid DCs, respectively [30]. For both examples, the variance-driving genes found in the gene-singular vectors were localized to their corresponding clusters (Fig. 2.12 c,d) and overlapped strongly with differentially expressed genes found after subclustering (see Table S3).

Figure 2.12: Congruence between variance-driving genes and differentially expressed genes between sub-clusters in a BMNC data set. (Figure on the previous page) a g-phiclust for each cluster in the BMNC data set. b Singular vectors of the two clusters from the BMNC data set with the highest phiclust. The color indicates sub-clustering. Dashed grey lines indicate the 0 value on each of the axes. c Singular vectors of clusters shown in panel a with color indicating the average log-transformed gene expression of genes with the 1% highest values in the first gene singular vector. d Singular vectors of clusters shown in panel a with color indicating the average log-transformed gene expression of genes with the 1% lowest values in the first gene singular vector. e Genes driving the variance in the two clusters shown in b. These genes have the 20 highest/lowest values in the first gene singular vector respectively. In blue: top 20 upregulated genes based on differential expression (DE) test between the sub-clusters using findMarkers (from scran R package).

2.2.7 Phiclust reveals subpopulations in a fetal human kidney data set that can be confirmed experimentally

As a second example of our approach we analyzed a fetal human kidney data set we published previously [31]. In our original analysis, we were forced to merge several clusters, since we were unsure if we the observed variability was just noise. We hence wanted to use phiclust to find previously overlooked subpopulations. As for BMNCs, phiclust corresponded well with a qualitative assessment of cluster variability (Fig. 2.13 a): Clusters with visible sub-clusters had the highest values of phiclust. Ordering the clusters by g-phiclust resulted in a similar ranking as phiclust (Fig. 2.15a). Subclustering of the cluster with the highest phiclust, ureteric bud/collecting duct (UBCD), revealed a subset of cells with markers of urothelial cells (UPK1A, KRT7) (Fig. 2.13b, Fig. 2.15 b-e). Immunostaining of these two genes, together with a marker of the collecting system (CDH1), in week 15 fetal human kidney sections confirmed the presence of the urothelial subcluster (Fig. 2.14a, Fig. 2.16a).

Another cell type we did not find in our original analysis, are the parietal epithelial cells (PECs). They could now be identified within the SSBpr cluster (S-shaped body proximal precursor cells) (Fig. 2.13b, Fig. 2.15 b-e).

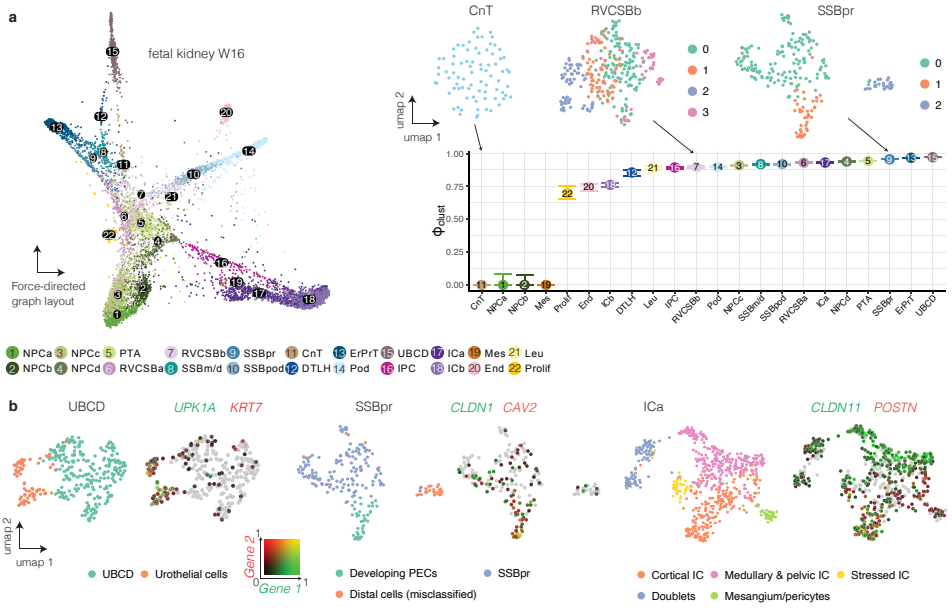


Figure 2.13: Phiclust reveals subpopulations in a human fetal kidney data a Force-directed graph layout and phiclust for the fetal kidney data set. Error bars indicate the uncertainty obtained by resampling the noise. Inset: UMAP of clusters with low, intermediate, and high values of phiclust. b UMAPs of the UBCD, SSBpr, and ICa clusters. Left: Colors indicate sub-clusters. Right: Colors indicate the log-normalized gene expression of the two indicated genes. One gene follows the red color spectrum, the other gene the green color spectrum. Absence of color indicates low expression in both genes, yellow indicates co-expression of both genes.

To reveal these cells in situ, we stained for AKAP12 and CAV2, which were among the top differentially expressed genes in this subcluster (Table S4), together with CLDN1, a known marker of PECs, and MAFB, a marker of the neighboring podocytes (Fig. 2.13d, Fig. 2.16b). Next to the PECs and proximal tubule precursor cells, SSBpr also contained a few cells that were misclassified in the original analysis, indicating the additional usefulness of phiclust as a means to identify clustering errors.
Further analysis of a cluster of interstitial cells (ICa) revealed multiple subpopulations (Fig. 2.13b, Fig. 2.15 b-e). Immunostaining showed that a POSTN-positive population is found mostly in the cortex, often surrounding blood vessels, whereas a SULT1E1-positive population is located in the inner medulla and papilla, often surrounding tubules (Fig. 2.14c, Fig. 2.16c). CLDN11, another gene identified by analysis of the gene-singular vectors (Fig. 2.15b-e), was found mostly in the medulla, but also in the outermost cortex. A more detailed, biological interpretation of the results can be found in Additional File 3.

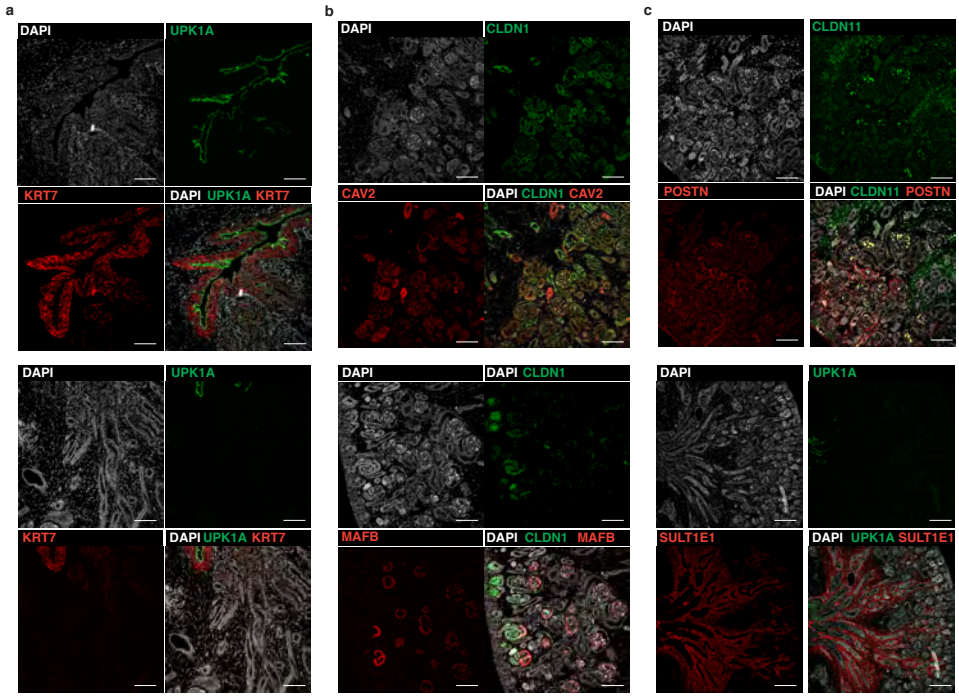
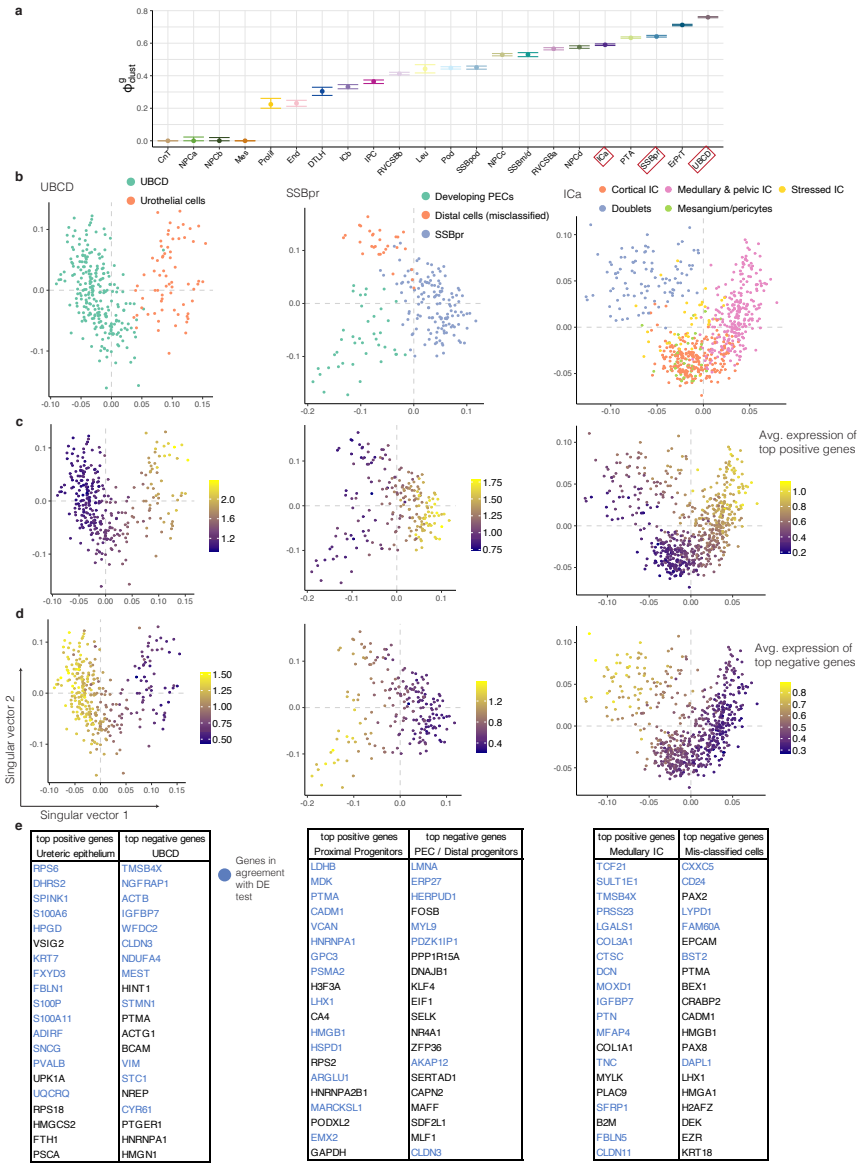


Figure 2.14: Subpopulations in a human fetal kidney data set revealed by phiclust can be confirmed experimentally.c-e Immunostainings of week 15 fetal kidney sections. c UPK1A and KRT7 are expressed in the urothelial cells of the developing ureter (upper panel) and absent from the tubules in the adjacent inner medulla (lower panel). d PECs express CLDN1 and CAV2 (upper panel), as well as CLDN1 at the capillary loop stage and later stages (lower panel). MAFB staining is found in podocytes and their precursors in the SSB (lower panel). e CLDN11 and POSTN are expressed in interstitial cells visualized by immunostaining (upper panel). SULT1E1 is expressed in the interstitial cells surrounding the ureter (marked by UPK1A) and the tubule in the inner medulla (lower panel). Scale bars: 100 µm.



top positive genes Ureteric epithelium	top negative genes UBCD
RPS6	TMSB4X
DHRS2	NGFRAP1
SPINK1	ACTB
S100A6	IGFBP7
HPGD	WFDC2
VSIG2	CLDN3
KRT7	NDUFA4
FXYD3	MEST
FBLN1	HINT1
S100P	STMN1
S100A11	PTMA
ADIRF	ACTG1
SNCG	BCAM
PVALB	VIM
UPK1A	STC1
UQCRQ	NREP
RPS18	CYR61
HMGCS2	PTGER1
FTH1	HNRNPA1
PSCA	HMGN1

top positive genes Proximal Progenitors	top negative genes PEC / Distal progenitors
LDHB	LMNA
MDK	ERP27
PTMA	HERPUD1
CADM1	FOSB
VCAN	MYL9
HNRNPA1	PDZK1IP1
GPC3	PPP1R15A
PSMA2	DNAJB1
H3F3A	KLF4
LHX1	EIF1
CA4	SELK
HMGB1	NR4A1
HSPD1	ZFP36
RPS2	AKAP12
ARGLU1	SERTAD1
HNRNPA2B1	CAPN2
MARCKSL1	MAFF
PODXL2	SDF2L1
EMX2	MLF1
GAPDH	CLDN3

top positive genes Medullary IC	top negative genes Mis-classified cells
TCF21	CXXC5
SULT1E1	CD24
TMSB4X	PAX2
PRSS23	LYPD1
LGALS1	FAM60A
COL3A1	EPCAM
CTSC	BST2
DCN	PTMA
MOXD1	BEX1
IGFBP7	CRABP2
PTN	CADM1
MFAP4	HMGB1
COL1A1	PAX8
TNC	DAPL1
MYLK	LHX1
PLAC9	HMGA1
SFRP1	H2AFZ
B2M	DEK
FBLN5	EZR
CLDN11	KRT18

Figure 2.15: Congruence between variance-driving genes and differentially expressed genes between sub-clusters in a fetal kidney data set. a g-phiclust for each cluster in the fetal kidney data set. b First two singular vectors of three clusters from the fetal kidney data set with high phiclust. The color indicates sub-clustering. Dashed grey lines indicate the 0 value on each of the axes. c First two singular vectors of clusters shown in panel a with color indicating the average log-transformed gene expression of genes with the 1% highest values in the first gene singular vector. d First two singular vectors of clusters shown in panel a with color indicating the average log-transformed gene expression of genes with the 1% lowest values in the first gene singular vector. e Genes driving the variance in the three clusters shown in b. These genes have the 20 highest/lowest values in the first gene singular vector respectively. In blue: top 20 upregulated genes based on differential expression (DE) test between the sub-clusters using findMarkers (from scran R package).

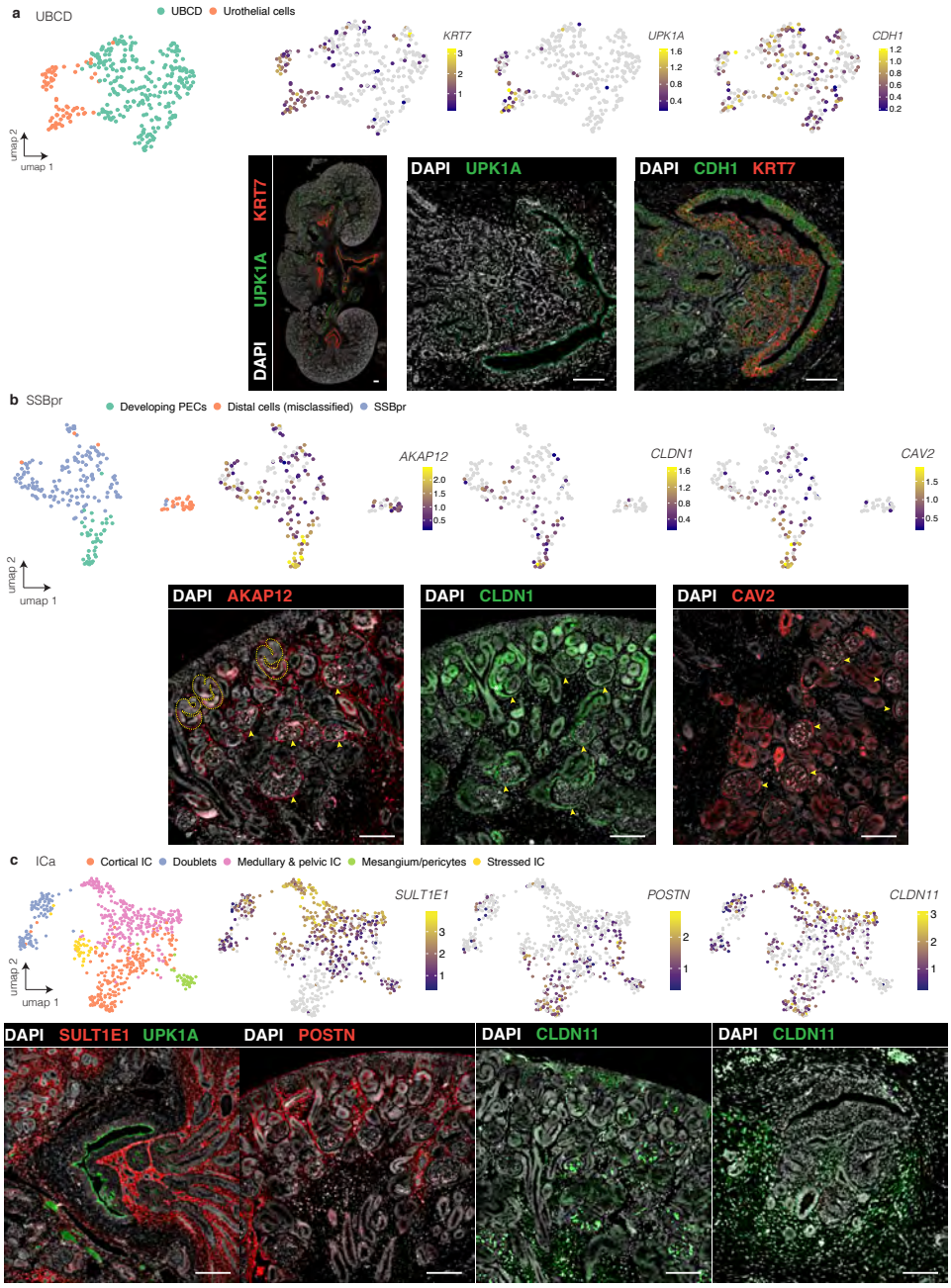


Figure 2.16: Immunostaining validates newly identified subclusters in fetal kidney data set. (Caption on the next page)

2.3 Discussion

Here, we presented phiclust, a clusterability measure that can help detect subtly different phenotypes in scRNA-seq data. Universal properties of the underlying theory make it possible to apply phiclust to arbitrary noise distributions, and the noise can be additive or multiplicative. Empirically, we find that the bulk of the singular value distribution of measured expression matrices is well-approximated by the MP distribution. This supports the assumption that the noise is generated by independent and identically distributed random processes.

While most of the technical and biological noise can likely be considered random, there are also known systematic errors and unwanted, confounding factors (such as the efficiency of RNA recovery, cell cycle phase etc.) Therefore, regressing out uninformative, deterministic factors, is an important part of the method.

The approach underlying phiclust also allows us to identify the genes that are most relevant for the biological interpretation of the observed variability. We found these genes to overlap strongly with differentially expressed genes identified after sub-clustering. The g-phiclust measure, a conjugate to phiclust, quantifies how distinguishable the expression profiles of different phenotypes are in the presence of noise.

The most important application of phiclust is the nomination of clusters for sub-clustering and subsequent experimental validation. All clusters that were nominated in the fetal kidney data set turned out to have subpopulations that could be validated by experiments: rare urothelial cells, which differ from nearby clusters in only a few genes; PECs and subtypes of interstitial cells, which had distinct spatial distributions.

There are several other methods that attempt to detect the presence of meaningful information in single-cell RNA-seq data. Below, we will compare phiclust to some of the most popular examples: the silhouette coefficient, ROGUE, robust PCA, the dip test and ZINB-WaVE.

The silhouette coefficient is a popular tool to assess clustering quality. In contrast to phiclust, this coefficient requires a (sub-)clustering and it cannot be used to decide, whether a cluster contains meaningful variability and should be sub-clustered further. As demonstrated, using the silhouette coefficient can lead to over-clustering of random noise as well as missing the presence of subtly different phenotypes. Likewise, phiclust appeared to be more sensitive than ROGUE, an entropy-based clusterability measure. Both ROGUE and the silhouette coefficient do not scale linearly with the tARI, which we introduced as an upper limit to the achievable agreement of an empirical clustering with the ground truth.

Figure 2.16: Immunostaining validates newly identified subclusters in fetal kidney data set. (Figure on the previous page) a-c Upper panels show UMAPs of the selected clusters in the fetal kidney data set. Log-normalized expression of selected genes is indicated by color. Lower panels show immunostainings of week 15 fetal kidney sections. a UBCD cluster. UPK1A, CDH1, and KRT7 expression is shown in a complete section (leftmost image) and in the urothelial epithelium. b SSBpr cluster. Expression of AKAP12, CLDN1 and CAV2 is shown. The dashed lines indicate S-shaped bodies, arrowheads indicate PECs in developing glomeruli c ICa cluster. Expression of SULT1E1 and UPK1A is visible around the ureter expression of POSTN is visible in cortical areas, CLDN11 is visible in the cortical area (CLDN11, left image) and around the ureter (CLDN11, right image). Scale bars: $100\mu m$.

Robust PCA [32, 33] decomposes a measured expression matrix into a sparse component and a low-rank component. Under the assumption that noise is sparse, the sparse component is identified with random noise. In our opinion, there is no reason to assume that the noise in scRNA-seq data is sparse, or sparser than the measured expression matrix itself. Likely, every non-zero gene expression measurement was corrupted by noise. Additionally, the remaining low-rank component cannot be identified as the noise-free signal. It is fundamentally impossible to reconstruct the noise-free signal from the measured expression because the noise is created by a random process. The low-rank component is therefore only a (noisy) approximation of the noise-free signal. Given the fundamental limit to signal reconstruction, the best thing we can do is quantify the closeness between signal and measured expression, as implemented by phiclust. In robust PCA, the low-rank matrix is often further subjected to dimensionality reduction, where it is difficult to determine the correct number of dimensions. phiclust does not require any dimensionality reduction and uses all available data.

The dip test [9], a method aimed at detecting the presence of clusters, tests whether there are multiple modes in the data. It can be applied directly to the distribution of distances between expression profiles or a low-dimensional representation of the data, such as principal component scores. The dip test will miss relevant variability, if it does not manifest itself as separate modes, which can easily occur, for example in the case of differentiation paths. It also just provides a binary result (modes present or not), whereas phiclust is a continuous measure and does not require the presence of modes.

ZINB-WaVE [25] performs dimensionality reduction based on a zero-inflated negative binomial distribution and is similar to principal component analysis, if no additional covariates are added to the model. ZINB-WaVE acknowledges the fact that principal components are prone to correlate with nuisance parameters, even after normalization. The problem is circumvented by adding such parameters as covariates to the model, which is similar to the confounder regression used for phiclust. However, the user has to decide the number of dimensions to use and currently there is no principled way to determine the optimal number. phiclust does not have any such adjustable parameters.

2.4 Conclusion

We hope that this manuscript will bring renewed awareness to random noise as a factor that imposes hard limits on clustering and identification of differentially expressed genes. We hope that quantitative measures of clusterability, such as phiclust, can play an important role in making single-cell RNA-seq analysis more reproducible and robust.

2.5 Methods

2.5.1 Preprocessing

Before applying the method to simulated or measured single-cell RNA-seq data sets, several preprocessing steps are necessary. The raw counts are first normalized and log-transformed. Next, the expression matrix is standardized, first gene-wise, then cell-wise. These steps assure the proper agreement of the bulk of the singular value distribution with the MP distribution (Additional file 1: Fig. S2). See also Supplementary Note, Section 3.1.

2.5.2 Phiclust

To derive phiclust, we assume that the expression matrix $\widetilde{X}$ measured by scRNA-seq, can be written as the sum of a random matrix X, which contains random biological variability and technical noise, and a signal matrix P, which contains the unobserved expression profiles of each cell:

$$\widetilde{X} = X + P$$

Note that in this decomposition, cells that belong to the same cell type (or phenotype) have identical expression profiles in the signal matrix P. Below we will show that multiplicative noise can be treated analogously.

We apply SVD to obtain the singular values, as well as the right and left singular vectors of $\widetilde{X}$. The left singular vectors span gene-space and the right singular vectors span cell-space. Hence, we call them gene-singular vectors and cell-singular vectors, respectively. If we use the term "singular vector" it is implied to mean cell-singular vector.

Considering the signal matrix P a perturbation to the random matrix X, we can apply results from both random matrix theory and low-rank perturbation theory. Random matrix theory [33, 34] predicts that the singular value distribution of X is a Marchenko-Pastur (MP) distribution [18, 19, 35], which coincides with the bulk of the singular value distribution [12–14] of $\widetilde{X}$. The singular values of $\widetilde{X}$ above the values predicted by the MP distribution characterize the signal matrix P. Since the agreement with the MP distribution holds strictly only for infinite matrices, we use two additional concepts to identify relevant singular values exceeding the range defined by the MP distribution. The Tracy-Widom [16, 36] (TW) distribution describes the probability of a singular value to exceed the MP distribution, if the matrix is finite. Additionally, since singular vectors of a random matrix are normally distributed, relevant singular vectors have to be significantly different from normal [14]. To test for normality we used the Shapiro-Wilk test.

We apply low-rank perturbation theory [17] to calculate the singular values (θ_i) of P from the outlying singular values (γ_i) of the measured expression matrix $\widetilde{X}$:

$$\theta_i(\gamma_i) = \sqrt{\frac{2c}{{\gamma_i}^2 - (c+1) - \sqrt{\left({\gamma_i}^2 - (c+1)\right)^2 - 4c}}}$$

where c is the cell-to-gene ratio, i.e. the total number of cells divided by the total number of genes.

The values of θ_i are then used to obtain the angles ϕ_i between the singular vectors of $\widetilde{X}$ and P. These angles are conveniently expressed in terms of their squared cosine as

$$\phi_i = cos(\alpha_i)^2 = 1 - \frac{c\left(1+\theta_i^2\right)}{\theta_i^2\left(\theta_i^2+c\right)}.$$

The leading singular vector of the measured expression matrix, which has the largest singular value, has the smallest angle to its unperturbed counterpart. The squared cosine of this smallest angle is then used as a measure of clusterability:

$$\phi_{clust} = \cos(min_i\phi_i)^2 = max_i \cos(\phi_i)^2, \quad phi_i \in [0, \frac{\pi}{2}]$$

For a detailed derivation of phiclust, see Additional File 2, Section 2.1-2.4.

Uncertainty of phiclust

The uncertainties for the values phiclust are estimated using a sampling approach. The basic idea is to approximate the signal matrix P and add new realizations of the noise matrix by sampling from a random distribution. The uncertainty is then obtained from the values phiclust calculated for this ensemble of sampled matrices. First, we decompose a simulated or measured expression matrix $\widetilde{X}$ into a noise matrix Xr and a matrix Xs that contains deterministic structure. Xs is constructed from the relevant singular vectors, which were identified as described in the previous section. Note that Xs contains noise and is thus different from the signal matrix P. To create an approximation Ps of the signal matrix P, we replace the singular values γ_i used in the construction of Xs with the singular values θ_i of P, calculated using low-rank perturbation theory as shown in the previous section. The entries of the noise matrix Xr have a mean of 0 and a standard deviation of 1, as a result of preprocessing. Since the results of RMT are valid irrespective of the particular noise distribution, we can create additional realizations of the noise matrix by sampling from a normal distribution with mean 0 and standard deviation 1. By adding sampled noise matrices to the approximated signal matrix Ps, we can create an ensemble of matrices with the same singular value spectrum as the original measured expression matrix but different realizations of the noise. The uncertainty for positive and negative deviations from the mean is then calculated as the standard deviation for at least 50 sampled matrices. See Supplementary Note, Section 2.4.3 for a detailed description.

Test for deviation from the MP distribution

To validate the use of the MP distribution, we test whether the bulk of the measured singular value distribution deviates significantly. Singular values are considered to be part of the bulk, if they are located below the MP upper bound and not associated with the transcriptome mode. We sample 1000 values from the MP distribution using the RMTstat R package (V 0.3) and subsequently test for similarity with the Kolmogorov-Smirnov test [35]. The resulting p-values are adjusted for multiple hypothesis testing with the Benjamini-Hochberg procedure [37].

Confounder regression

scRNA-seq data contains various confounding factors that drive uninformative variability. These either emerge from technical issues (such as the varying efficiency of transcript recovery, which cannot be fully eliminated by normalization) or biological factors (such as cell cycle phase, metabolic state, or stress). To account for these factors, a regression step, inspired by current gene expression normalization methods [23, 26], is included. We perform a linear regression by using each relevant singular vector as a response variable and the confounding factors as covariates. This is a valid approach because the singular vectors of the measured expression matrix contain normal distributed noise. The amount of variance explained by the nuisance parameters is then given by the value of the adjusted R2 (coefficient of determination) of this linear regression. To relate the regression result to the singular values, we consider the squared singular values (= eigenvalues) which correspond to the variance explained by the corresponding singular vectors / eigenvectors. Squared singular values are corrected by multiplication with (1 – adjusted R2) to retrieve the fraction of variance not explained by nuisance parameters. The square root of the result is the corrected singular vector. See also Supplementary Note, Section 3.2. For Additional file 1: Fig. S5a, each nuisance parameter was individually regressed on, to compare the influence of each factor.

Multiplicative noise

To model multiplicative noise, we use a rectangular random noise matrix X with the same dimensions as the measured expression matrix $\widetilde{X}$ and a square signal matrix P whose number of rows or columns is equal to the number of measured genes. The measured expression matrix $\widetilde{X}$ is then modeled as:

$$\widetilde{X} = (I + P)^{\frac{1}{2}} \; X,$$

Where I denotes the identity matrix. Importantly, the bulk of the singular vector distribution of the measured expression matrix $\widetilde{X}$ still follows the MP distribution in this model. The singular values of the signal matrix P are calculated from the outlying singular values of $\widetilde{X}$ by:

$$\theta_i = \frac{2c}{\lambda_i - c - 1 - \sqrt{(\lambda_i - a)(\lambda_i - b)}}$$

with $a, b = \left(1 \pm \sqrt{c}\right)^2$. The angles between the corresponding singular vectors of the measured expression matrix and the signal matrix are then calculated as: $\phi\sigma^i{}_{mult} = \frac{1}{\theta_i} \frac{{\theta_i}^2 - c}{\theta_{i(c+1)} + 2c}$
More information on multiplicative perturbation can be found in [38].

2.5.3 Clustering

Theoretically achievable clustering quality

To construct a Bayes classifier [22], which achieves the minimal error rate, we need to know the ground truth clustering. Hence, we used data simulated with Splatter [20], containing two clusters. For each ground truth cluster, we fit a multidimensional Gaussian to the corresponding entries of the singular vectors (see Additional file 1: Fig. S3a). We only

consider singular vectors with singular values larger than predicted by the MP distribution. For the fit, we use the mclust [39] R package (V 5.4.6). We then construct a classifier by assigning a cell to the cluster for which it has the highest value of the fitted Gaussian distribution. This classifier is thus approximately a Bayes classifier (for a true Bayes classifier, we would need to know the exact distributions of the singular vector entries). The ARI [21] calculated based on this classification is thus approximately the best theoretically achievable ARI (tARI). We also tested the silhouette coefficient [8] as a potential alternative to the ARI for quantifying our notion of clusterability. The silhouette coefficient was calculated on the first singular vector using Euclidean distances. In Additional file 1: Fig. S4 the silhouette coefficient averaged over all cells is reported. The theoretically achievable silhouette coefficient tSIL is defined as the silhouette coefficient of the Bayes classification described in the previous paragraph. The calculation of tARI and tSIL is described in more detail in Additional File 2, section 2.5.

Clustering methods

For the validation of the tARI and tSIL, several clustering methods were used on simulated data with two clusters. Seurat clustering [2] was performed with the Seurat R package with 10 principal components (PCs) and 20 nearest neighbors. Three different resolution parameters were used: 0.1, 0.6, and 1.6. Scanpy clustering [3] was performed with the scanpy python package with 10 PCs and 20 nearest neighbors. Three different resolution parameters were used: 0.1, 0.6, and 1.6. Hierarchical clustering [5] was performed on the first 10 PCs and Euclidean distances. The hierarchical tree was built with the Ward linkage and the tree was cut at a height where 2 clusters could be identified. K-means [4] was performed on the first 10 PCs using Euclidean distances and two centers. TSCAN [40] was calculated on the first 10 PCs. In Additional file 1: Fig. S7 k-means clustering was performed on the first 3 principal components and using Euclidean distances.

Clusterability measures

ROGUE [10] is an entropy-based clusterability measure. A null model is defined under the assumption of Gamma-Poisson distributed gene expression and its differential entropy is then compared to the actual differential entropy of the gene expression profile. For the RNA-mix data set ROGUE (V 1.0) was used with 1 sample (see Fig S6), "UMI" platform, and a span of 0.6. For the simulated data sets, ROGUE was used with k = 10 (Additional file 1: Fig. S4 d). The silhouette coefficient was calculated with the cluster R package (V 2.1.0) using euclidean distances in the space of the relevant singular vectors. The reported values for the silhouette coefficients are average values per cluster. The confidence intervals given in Additional file 1: Fig. S6 and S7 are standard deviations of its values per cluster.

2.5.4 Variance driving genes

Genes that drive the variance in the significant singular vectors can be used to explore the biological information in the sub-structures. Genes with large positive or negative entries in a gene-singular vector are localized in cells with high positive or negative entries in the corresponding cell-singular vector. It is also possible to assess the signal-to-noise ratio for the genes by calculating the angle between the gene singular vectors of the measured

expression matrix $\widetilde{X}$ and the gene singular vectors of the signal matrix P, given by15

$$\phi^g_{clust} = cos\left(\widetilde{\alpha\phi}\right)^2 = 1 - \frac{\left(c + \theta_i^2\right)}{\theta_i^2\left(\theta_i^2 + 1\right)},$$

2

where c is the cell-to-gene ratio. We call ϕ^g_{clust} the gene phiclust (g-phiclust). See Additional File 2, section 2.4 for a more detailed discussion.

2.5.5 Data sets

The simulated data sets in Additional file 1: Fig. S1 comprised 201 cells and 350 genes. The random noise matrix was sampled from a normal distribution with mean 0 and variance 1 in panels a and b, or from a Poisson distribution with parameter 1 in panels c and d. The rank 1 signal matrix was constructed from one cell-singular vector and one gene-singular vector. The cell-singular vector consisted of 67 entries equal to $1/\sqrt{N_{cell}}$ and all other entries equal to $-1/\sqrt{N_{cell}}$, where Ncell is the number of cells. The gene-singular vector consisted of 200 entries equal to $1/\sqrt{N_{gene}}$ and the rest equal to $-1/\sqrt{N_{gene}}$, where Ngene is the number of genes. The signal matrix was then created by matrix multiplication of the gene-singular vector and the transposed cell-singular vector times the singular value θ ($\theta = 2$ in a,c and $\theta = 5$ in b,d). In Additional file 1: Fig. S2f,g a rank 1 signal matrix was created similarly as described above. The cell-singular vector with a number of entries matching the number of cells in the cluster was constructed as before. The gene-singular vector was drawn from a normal distribution and subsequently normalized to unit length. The rank 1 signal matrix was then added to the preprocessed expression matrix of the indicated cluster. The remaining simulated data sets were produced with the splatter [20] R package (V 1.10.1). The parameters used for the simulation are shown in Table S1. For Fig. 1c,d, Additional file 1: Fig. S3b-e, Additional file 1: Fig. S4, and Additional file 1: Fig. S8a the simulations for each parameter were performed 50 times, each with a different seed. The results were averaged over the 50 runs. Confounder regression was performed for the total number of transcripts per cell. PBMC data [23] was downloaded from the 10x genomics website (). For the calculation of the tARI, clustering with Scanpy, TSCAN, k-means, and hierarchical clustering, preprocessing was performed with the scanpy python package (V 1.4.6) following the provided pipeline () for the filtering of cells and genes, normalization, and log-transformation as well as cluster annotation. For the clustering with Seurat, the provided Seurat pipeline was used () for preprocessing, such as cell and gene filtering, normalization, log-transformation and cluster annotation using the Seurat R package (V 3.1.5). CD8 T cells and B cells were extracted from the data and each cluster was standardized gene-wise and cell-wise before the calculation of the SVD. To remove any sub-structure in these clusters and before the reconstruction of the matrices from the SVD, singular values above the MP distribution were moved into the bulk, and the transcriptome mode (i.e. the singular vector that would have the largest singular value without normalization, see Supplementary Methods Note 1) was moved above the MP distribution. Then, two synthetic clusters containing 150 cells each were created from the cleaned-up original clusters. For cluster 1, a weighted average of a randomly picked B cell with expression profile c_B and a randomly picked CD8 T cell with expression profile $c_{CD8\ T}$ was calculated according to: $c_1 = \alpha \cdot c_B + (1 - \alpha) \cdot c_{CD8\ T}$. For cluster 2, the weights

were flipped: $c_2 = (1-\alpha) \cdot c_B + \alpha \cdot c_{CD8\ T}$. α was chosen in a range from 0 to 1. α close to 0.5 produced highly similar clusters, while α close to 0 or 1 produced maximally different clusters (see Fig S3e). For each value of α, the procedure was repeated 50 times, each with a different seed for selecting 300 cells per cell type, and the results were averaged. RNA-mix data [24] was downloaded from the provided GitHub page. The data were normalized with the R scran package (V 1.14.6) and then log-transformed. Confounder regression was performed for the total number of transcripts, average mitochondrial gene expression, and average ribosomal gene expression. Two different merged clusters were created from the provided RNA mixtures as shown in Additional file 1: Fig. S6. The bone marrow mononuclear cell data set (BMNC) [28] was downloaded from the R package SeuratData (bmcite, V 0.2.1). Normalization and calculation of the G2M score [41] were performed with the Seurat R package (V 3.1.5). Confounder regression was performed for the log-transformed total number of transcripts, cell cycle score, and average expression of: mitochondrial genes and ribosomal genes (list obtained from the HGNC website). For the fetal kidney data set [31], the same preprocessing and normalization was used as reported previously (scran R package [42]). The data was then log-transformed and the G2M score was calculated with the Seurat R package. Confounder regression was performed for the log-transformed total number of transcripts, G2M scores, and the average expression of: mitochondrial genes, ribosomal genes, and stress-related genes [43].

2.5.6 Single cell data analysis

Embedding

Uniform Manifold Approximation and Projections [44] (UMAPs) for individual clusters were calculated with the R package umap (V 0.2.7.0) on the first 10 PCs, 20 nearest neighbors, min_ dist = 0.3, and Euclidean distances. The umap for BMNC data was calculated with the Seurat R package using 2000 highly variable genes (hvg), d = 50, k = 50, min.dist = 0.6 and metric = cosine. For the fetal kidney data set a force-directed graph layout was calculated using the scanpy python package. The graph was constructed using 100 nearest neighbors, 50 PCs, and the ForceAtlas2 layout for visualization.

Differential expression test

Differentially expressed genes within the sub-clusters found in Additional file 1: Fig. S9 and Additional file 1: Fig. S10 were calculated with the function findMarkers of the scran R package on log-transformed normalized counts. Genes with a false discovery rate below 0.05 were selected and then sorted by log2 fold change. In Figures S9e and S10e, genes with the top 20 highest/lowest values in the gene singular vectors are listed and colored blue if they correspond to the top 20 DE genes.

2.5.7 Staining

A human fetal kidney (female) at week 15 of gestation was used for immunofluorescence using the same procedure as reported previously [31]. The following primary antibodies were used: rabbit anti-UPK1A (1:35, HPA049879, Atlas Antibodies), mouse anti-KRT7 (1:200, # MA5-11986, Thermo Fisher Scientific), rabbit anti-CDH1 (1:50, SC-7870, Santa Cruz), rabbit anti-CLDN1 (1:100, # 717800, Thermo Fisher Scientific), goat anti-CAV2 (1:100, AF5788-SP, R& D Systems), mouse anti-AKAP12 (1:50, sc-376740, Santa Cruz), rabbit anti-

CLDN11 (1:50, HPA013166, SIGMA Aldrich), mouse anti-POSTN (1:100, sc-398631, Santa Cruz) and goat anti-SULT1E1 (1:50, AF5545-SP, R& D Systems). The secondary antibodies were all purchased from Invitrogen and diluted to 1:500: Alexa Fluor 594 donkey anti-mouse (A21203), Alexa Fluor 594 donkey anti-rabbit (A21207), Alexa Fluor 647 donkey anti-mouse (A31571), Alexa Fluor 647 donkey anti-rabbit (A31573), Alexa Fluor 647 donkey anti-goat (A21447). The sections were imaged on a Nikon Ti-Eclipse epifluorescence microscope equipped with an Andor iXON Ultra 888 EMCCD camera (Nikon, Tokyo, Japan).

2.5.8 Data availability

All sequencing data sets were obtained from publicly available resources. The BMNC data can be downloaded with the R package SeuratData, named "bmcite." The fetal kidney data is available in the GEO database under the accession number GSE114530. The PBMC data can be downloaded at https://cf.10xgenomics.com/samples/cell/pbmc3k/pbmc3k_filtered_gene_bc_matrices.tar.gz and the RNA-mix data is available at https://github.com/LuyiTian/sc_mixology, named "mRNAmix_qc". Supplementary tables are available in the online version at https://doi.org/10.1186/s13059-021-02590-x.

2.6 Supplementary Note 1

Introduction

Our aim is to develop a clusterability measure for scRNA-seq data. As we define more precisely in section (2.6), we consider clusterability to be the clustering quality that is optimally achievable, given a certain amount of noise in the data. Clustering quality can only be assessed quantitatively if the ground truth is known, which is strictly only the case for simulated data. A clusterability measure must thus be able to reflect clustering quality without knowledge of the ground truth. Such a measure would be highly useful, since it would allow us to detect the presence of meaningful (non-random) variability, and thus determine the necessity to sub-cluster measured data. For the development of this clusterability measure we will use concepts from random matrix theory and perturbation theory. In short, we decompose the single-cell gene expression matrix $\tilde{X}$ into a random matrix X, which contains technical and biological noise, and a signal matrix P, which contains the expression profiles of different cell types or states. Then, we apply perturbation theory, treating the signal matrix P as a low-rank perturbation of the noise matrix X. Perturbation theory then allows us to calculate the angle between the singular vectors of the measured single cell expression matrix $\tilde{X}$ and the corresponding singular vectors of the unobserved signal matrix P. The cosine of this angle constitutes a useful clusterability measure because a large value (small angle) indicates a high signal-to-noise ratio (and thus high clusterability) and a small value (large angle) indicates a low signal-to-noise-ratio (and thus low clusterability). We show empirically that this clusterability measure is a proxy for the theoretically achievable adjusted rand index [Fig. 1d].

In what follows, we first present our model of gene expression data (2.6) and introduce matrix decomposition (2.6). Subsequently, we introduce the Marchenko-Pastur (MP) distribution (2.6), which describes the eigenvalue spectrum of a random matrix and apply perturbation theory to link the (unobserved) signal matrix to the spectrum of the measured expression matrix (2.6). In section (2.6), we establish our notion of clusterability. In section (2.6), we describe the preprocessing steps necessary for the application of the theory to single-cell RNAseq data. Then, in section (2.6), we develop a method to remove the effect of nuisance variables (i.e. sources of systematic, non-random variability that should not drive clustering.) The complete algorithm can be found in section (2.6).

Phiclust

Model

Let $\tilde{X} \in \mathbb{R}^{M \times N}$ be the measured single-cell expression matrix with M the number of genes (rows) and N the number of cells (columns). We model the measurement $\tilde{X}$ as the sum of a random noise matrix $X \in \mathbb{R}^{M \times N}$ and a "signal" matrix $P \in \mathbb{R}^{M \times N}$.

$$\tilde{X} = X + P \tag{2.1}$$

In our model, X contains both technical and biological noise. For example, if there was only one cell type or cell state present in a data set, P would consist of identical columns. Note that we only observe the matrix $\tilde{X}$ experimentally. We will show below, that we can make a statement about the influence of the noise X on the signal P, without knowing X or P. To achieve that we invert the logic of conventional models: instead of modeling

the influence of random noise on the signal, we consider the influence of a deterministic perturbation on a random matrix. All results rely on matrix decomposition, which will be introduced next.

MATRIX DECOMPOSITION

1. Eigendecomposition

We first define the cell-cell correlation matrix. To that end, we assume that $\tilde{X}$ has been standardized cell-wise (i.e. column-wise) to mean 0 and standard deviation 1. The cell-cell correlation matrix $C \in [-1,1]^{N\times N}$ is then defined as:

$$C = \frac{1}{M-1}\tilde{X}^T\tilde{X} \tag{2.2}$$

The correlation matrix is a square and symmetric matrix which can hence, by the spectral theorem, undergo eigendecomposition into the form

$$C = V\Sigma V^T = \sum_{i=1}^{N} \lambda_i v_i v_i^T. \tag{2.3}$$

$V \in \mathbb{R}^{N\times N}$ contains the eigenvectors v_i of C in the columns and $\Sigma \in \mathbb{R}^{N\times N}$ is a diagonal matrix containing the eigenvalues λ_i of C. If $M < N$, then C is a singular matrix and will contain at least $N - M$ eigenvalues equal to 0, which is an important consideration for the definition of the Marchenko-Pastur distribution (see below).

In full analogy to the cell-cell correlation matrix we can define a gene-gene correlation matrix $\hat{C}$, now assuming that the expression matrix $\tilde{X}$ has been standardized gene-wise (row-wise) to mean 0 and standard deviation 1:

$$\hat{C} = \frac{1}{N-1}\tilde{X}\tilde{X}^T. \tag{2.4}$$

If $M > N$, then $\hat{C}$ is a singular matrix and will contain at least $M - N$ eigenvalues equal to 0. Therefore either C (if $M < N$) or $\hat{C}$ (if $M > N$) is a singular matrix (unless $M = N$) with at least $|N - M|$ eigenvalues equal to 0.

2. Singular value decomposition

To decompose the (rectangular) expression matrix $\tilde{X}$ into noise and signal, we use singular value decomposition:

$$\tilde{X} = \sum_{i=1}^{N} \gamma_i u_i v_i^T.$$

The v_i's are the right singular vectors of $\tilde{X}$ and correspond to the eigenvectors of the cell-cell correlation matrix. We will call them cell singular vectors or singular vectors in the following. The u_i's are the left singular vectors of $\tilde{X}$ and correspond to the eigenvectors of the gene-gene correlation matrix, which we will call gene singular vectors. The singular values are denoted by γ_i. The singular values of $\tilde{X}$ and the eigenvalues of the corresponding correlation matrix have a known connection given by:

$$\lambda_i = \gamma_i^2.$$

Random Matrix Theory

The Marchenko-Pastur (MP) distribution is widely used to reveal nonrandom properties of empirical correlation matrices in physics and finance [12, 13]. The MP distribution describes the distribution of eigenvalues of a random correlation matrix in the asymptotic limit [18, 19, 36] (for $N \longrightarrow \infty$ and $M \longrightarrow \infty$, $\frac{N}{M} < 1$). The entries of the random matrix are arbitrary as long as they are distributed identically and independently. scRNA-seq data are typically modeled by a Poisson, a negative binomial or a zero-inflated negative binomial distribution, which are in principle admissible in random matrix theory.

Theorem 1 (Marchenko-Pastur) *([18, 19, 36]) Let Y be a $M \times N$ matrix with entries that are independent identically distributed (i.i.d.), mean 0 and variance $v^2 < \infty$. The corresponding Wishart matrix is defined as $W = \frac{1}{M} Y^T Y$. For $N \longrightarrow \infty$, $M \longrightarrow \infty$ and $0 < c < 1$, where c is defined as $\frac{N}{M}$. The distribution of the eigenvalues λ of W is given by*

$$\mu(\lambda) = \frac{\sqrt{(b-\lambda)(\lambda-a)}}{2\pi c \lambda v^2} \, d\lambda \quad \text{if} \quad a \leq \lambda \leq b$$

For $c > 1$ the distribution has an additional number of 0 eigenvalues:

$$\mu(\lambda) = \frac{\sqrt{(b-\lambda)(\lambda-a)}}{2\pi c \lambda v^2} \mathbb{1}_{[a,b]} + (1 - \frac{1}{c}) \delta_0(\lambda)$$

with

$$a, b = v^2 \left[1 \pm \sqrt{c}\right]^2.$$

$\delta_0(\lambda)$ is the Dirac delta function, which is 1 if $\lambda = 0$ and 0 otherwise. For the correlation matrix we obtain $v = 1$ because the mean of all eigenvalues is 1.
This theorem places the eigenvalues of a random correlation matrix into a compact interval between $[a, b]$. All eigenvalues of an empirical correlation matrix that fall within this interval can be considered to be due to random noise. The presence of eigenvalues above this distribution indicates the existence of non-random structure in the data. An empirical (measured) correlation matrix can therefore be decomposed into a random part C^r and a signal part C^s [19]:

$$C = \sum_{\lambda \leq b} \lambda_i v_i v_i^T + \sum_{\lambda > b} \lambda_i v_i v_i^T = C^r + C^s$$

C^s contains the non-random and therefore biologically relevant correlations.

For the application of the MP distribution to an empirical correlation matrix we need to consider that the eigenvalues of a correlation matrix always sum up to 1. Thus, if there are eigenvalues above the MP distribution the bulk of the distribution (which is described by MP) will shift to the left. To approximately account for this shift, we introduce a modified MP-distribution as follows:

$$\mu^*(\lambda) = \frac{\mu(\lambda)}{\alpha},$$
$$a^* = \alpha a, \; b^* = \alpha b.$$

where $\alpha = 1 - \frac{\lambda_{max}}{N}$ and a^* and b^* replace a and b respectively.

We can formulate the MP distribution also for singular values, via a variable transform, and obtain the following density:

$$d\rho(\gamma) = \frac{\sqrt{(b-\gamma^2)(\gamma^2-a)}}{\pi\gamma c}\, d\gamma \quad \text{if} \quad \sqrt{a} \leq \gamma \leq \sqrt{b} \tag{2.5}$$

In this case, all singular values that lie within the compact interval of $[\sqrt{a}, \sqrt{b}]$ can be considered to arise from random noise and singular values above this threshold indicate deterministic biological relevant signal. Thus, we can decompose the matrix $\tilde{X}$ into two parts:

$$\tilde{X} = \sum_{\gamma \leq \sqrt{b}} \gamma_i u_i v_i^T + \sum_{\gamma > \sqrt{b}} \gamma_i u_i v_i^T = \tilde{X}^r + \tilde{X}^s \tag{2.6}$$

The first part $\tilde{X}^r$ is random noise, the second part $\tilde{X}^s$ contains relevant signal.

The MP theorem holds strictly only in the asymptotic limit, but provides a very good approximation for big enough N and M. For finite dimensions, there is however a non-zero probability that a random i.i.d matrix has eigenvalues above the MP distribution. That probability is described by the Tracy-Widom (TW) distribution.

Theorem 2 (Tracy-Widom) *([36]) For empirical correlation matrices of size $N \times N$ of i.i.d. random variables with a finite fourth moment, the distance between the upper edge of the spectrum of the MP distribution b and the largest eigenvalue λ_{max} converges towards the Tracy-Widom distribution*

$$\text{Prob}\left(\lambda_{\max} \leq b + \gamma N^{-2/3} u\right) = F_1(u),$$

where γ in this case is given by $\gamma = \sqrt{c}\, b^{2/3}$.

$F_1(u)$ is the TW distribution, the probability distribution of the re-scaled eigenvalues of a random Hermitian matrix. We are interested in the type-1 distribution which holds for Gaussian orthogonal ensembles [15]. The distribution function can not be explicitly stated but relies on numerical approximations.
The TW distribution can be formulated, as well, for the singular values via the variable transform:

$$\text{Prob}\left(\gamma_{max} \leq \sqrt{b + \gamma N^{-2/3} u}\right) = F_1(u), \tag{2.7}$$

Since we always work with finite matrices in practice, we use the TW distribution to discriminate between singular values that belong to noise and signal, respectively. Specifically, we use $u = 1$ as a cutoff, so that $F_1(1) \approx 0.95$. In other words, there is a probability of 0.05 that a singular value bigger than $\sqrt{b + \gamma N^{-2/3}}$ is observed, if the matrix is entirely random. If N is very low, the MP distribution is not a good approximation anymore. For $N < 50$, we create an empirical distribution of noise-related singular values, by permuting the entries

of the measured expression matrix $\tilde{X}$. For each permutation we calculate the singular values and note the largest singular value. The 95th quantile of the distribution of the largest singular values across permutations is then taken to be the cutoff between singular values stemming from noise and signal respectively.

To discriminate random from non-random matrix components we can also look at the singular vectors [14]. Singular vectors that correspond to random components are "de-localized" and their elements have the following distribution:

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

If N is large, this distribution can be estimated by a Gaussian distribution with mean zero and variance $\frac{1}{N}$.

$$f(\psi) \sim \frac{N}{\sqrt{2\pi}} e^{\frac{-N\psi^2}{2}} \tag{2.8}$$

In order to distinguish localized from de-localized singular vectors, we can therefore assess the normality of the singular vectors. In our implementation we use a Shapiro-Wilk test. We assign singular vectors that obtain a p-value < 0.01 or are associated to singular values far from the bulk (the highest 50% of signal singular values) to real variability above the MP distribution.

Perturbation theory

As explained above, we model the observed expression matrix $\tilde{X}$ as a random matrix X perturbed by a deterministic signal matrix P. There is an important difference between the perturbation matrix P in equation 2.1 and the matrix $\tilde{X}^s$ in equation 2.6. $\tilde{X}^s$ does contain biologically relevant information, but is still influenced by the effects of random noise, whereas the matrix P consists of the pure signal without any added noise. The only case where these two matrices are identical is when the singular vectors of the noise matrix X and the perturbation matrix P are linearly independent, which is rarely the case. It is thus not possible to recover the unobserved, noise-free signal matrix by using those singular vectors that are associated with the highest singular values.

While it is not possible to reconstruct the signal matrix from measured data, perturbation theory [17] establishes a simple relationship between the singular value of the observed expression matrix $\tilde{X}$ and those of the signal matrix P. P is assumed to have finite rank r. Its singular value decomposition is thus:

$$P = \sum_{i=1}^{r} \theta_i u_i v_i^T, \text{ where } r \ll N, M$$

For scRNA-seq data, we only have to consider singular values $\theta_i > 0$, which means that $\tilde{X}$ potentially has singular values above the MP distribution. Thus, we only need to consider the largest singular values of $\tilde{X}$.

Theorem 3 (Largest Singular Value for MP) *([17]) The r largest singular values $\gamma_i(\tilde{X})$ of the $M \times N$ perturbed matrix $\tilde{X}$ exhibit the following behaviour as $M, N \longrightarrow \infty$ and $\frac{N}{M} \longrightarrow c$: For each fixed $1 \leq i \leq r$,*

$$\gamma_i(\tilde{X}) \xrightarrow{a.s.} \begin{cases} \sqrt{\frac{\left(1+\theta_i^2\right)\left(c+\theta_i^2\right)}{\theta_i^2}} & \text{if } i \leq r \text{ and } \theta_i > c^{1/4}, \\ b & \text{otherwise} \end{cases} \tag{2.9}$$

Moreover, for each fixed $i > r$, we have that $\gamma_i\left(\tilde{X}_n\right) \xrightarrow{a.s.} b$.

This theorem establishes a functional relationship between the largest singular values γ_i of the measured expression matrix and the singular values θ_i of the signal matrix P. Note that if θ_i is smaller than or equal to $c^{\frac{1}{4}}$, the corresponding γ_i will be equal to b, which is the upper limit of the MP distribution. In other words, if the perturbation (signal) is too small, the singular value spectrum of the observed expression matrix $\tilde{X}$ will be just the MP distribution and hence, no meaningful signal can be extracted.

From the above formula we are able to calculate the singular values of the perturbation matrix P. These are the values that describe the actual variances of the signal matrix without any contribution of the noise. This is achieved by calculating the inverse function

$$\theta_i(\gamma_i) \xrightarrow{\text{a.s.}} \begin{cases} \sqrt{\frac{2c}{\gamma_i^2-(c+1)-\sqrt{(\gamma_i^2-(c+1))^2-4c}}} & \text{if } \gamma_i > b, \\ c^{\frac{1}{4}} & \text{otherwise} \end{cases} \tag{2.10}$$

Phiclust

Next, we want to establish how the singular vectors of $\tilde{X}$ depend on the perturbation P. In section 2.6 it is described that the elements of the singular vectors will follow a Gaussian distribution for a random matrix and large N. The elements of the singular vectors of the perturbation P are deterministic and correspond to biological variance. The following theorem describes the scalar product between the singular vector of the perturbation P and the perturbed matrix $\tilde{X}$.

Theorem 4 (Norm of Projection of Largest Singular Vectors for MP) *([17]) Let $\tilde{v}$ the right unit singular vectors of $\tilde{X}$. Then, the norm of projection of the right singular vector is given by*

$$\left|\langle \tilde{v}_i, v_i \rangle\right|^2 \xrightarrow{a.s.} \begin{cases} 1 - \frac{c\left(1+\theta_i^2\right)}{\theta_i^2\left(\theta_i^2+c\right)} & \text{if } \theta_i \geq c^{1/4} \\ 0 & \text{otherwise} \end{cases} \tag{2.11}$$

This theorem shows the same qualitative behavior as equation 2.9. If the singular value θ_i of the perturbation matrix is below the threshold of $c^{\frac{1}{4}}$, the scalar product is zero, indicating that the perturbed matrix $\tilde{X}$ has no relationship to the perturbation P. In other words, no relevant signal can be extracted. In the other limit, when the scalar product goes to 1, the

singular vectors of the perturbation P are perfectly aligned with the singular vectors of the perturbed matrix $\tilde{X}$. Thus, random noise has a negligible influence on the signal.

The scalar product given by $|\langle \tilde{v}_i, v_i \rangle|^2$ is identical to the squared cosine of the angle between the vectors:

$$\phi_{\text{clust}} = cos(\alpha)^2 = \left(\frac{\tilde{v} \cdot v}{\|v\| \, \|\tilde{v}\|} \right)^2 = (\tilde{v} \cdot v)^2 = |\langle \tilde{v}_i, v_i \rangle|^2.$$

This holds because the singular vectors are assumed to have norm 1.

We propose ϕ_{clust} (phiclust) as a measure of clusterability in scRNA-seq data. If, for a given cluster, there are no values above the MP distribution the signal of the perturbation matrix P can not be recognized any more and phiclust will be zero. If there are singular values above the MP distribution, phiclust evaluates how closely related the singular vectors of the expression matrix $\tilde{X}$ are to those of the perturbation matrix P.
We obtain a value of phiclust for each singular value that can be found above the MP distribution. Each of them indicates the signal-to-noise ratio for the variance that the corresponding singular vector explains. Thus, the more singular values are above the MP distribution, the more variances can be found in the data and it can be interpreted as proportional to the number of clusters. In the definition of phiclust, we have decided to use the maximum of all angles, thus indicating the maximal clusterability that can be achieved from clustering.

G-phiclust
In accordance with the above definition of phiclust (2.6), we can also define the clusterability, or signal-to-noise ratio, for the gene space. The following theorem describes the equation.

Theorem 5 (Norm of Projection of Largest Singular Vectors for MP) *([17]) Let $\tilde{u}$ be the left unit singular vectors of $\tilde{X}$. Then, the norm of projection of the left singular vector by*

$$|\langle \tilde{u}_i, u_i \rangle|^2 \xrightarrow{a.s.} \begin{cases} 1 - \frac{\left(c+\theta_i^2\right)}{\theta_i^2\left(\theta_i^2+1\right)} & \text{if } \theta_i \geq c^{1/4} \\ 0 & \text{otherwise} \end{cases} \tag{2.12}$$

For the gene singular vector, ϕ^g_{clust} (g-phiclust) indicates how closely the variance among genes is related to the original variance in the perturbation matrix P. For each singular vector, the variance-driving genes correspond to those with the highest absolute loading in the corresponding gene singular vector. Cells with high positive or negative entries in the singular vector have high expression of genes with large positive or negative entries in the corresponding gene singular vector, respectively. This relationship is not a replacement for the calculation of differentially expressed genes, but merely indicates the genes that drive the variance across cells for each singular vector. Based on the value of g-phiclust, it is possible to evaluate how accurate the determination of differentially expressed genes will be. With a low signal-to-noise ratio, it is more likely to obtain genes differentially expressed that can be attributed to noise. As well as for phiclust, we obtain several angles,

one for each singular value above the MP distribution. Thus, genes driving the variances in gene singular vectors with a higher g-phiclust are more accurate. We decided, to be consistent, to define g-phiclust as the highest squared cosine of the angle.

Uncertainty of phiclust

The theory presented above holds as the expected value in the infinite limit, however we do not know about the variations within the finite limit. To address this, we constructed a confidence interval for the values of phiclust using the following sampling approach. The basic idea is to approximate the signal matrix P and add new realizations of the noise matrix by sampling from a random distribution. The standard deviation is then constructed from the values of phiclust calculated for this ensemble of sampled matrices.

First, the matrix $\tilde{X}$ is pre-processed as described in section 2.6. By applying the MP distribution, we then determine the singular values associated with signal and noise. We decompose the simulated or measured expression matrix $\tilde{X}$ into a noise matrix X^r and a matrix X^s that contains deterministic structure (see equation 2.6).

Then, we estimate the first two moments of X^r, which due to the pre-processing of the measured expression matrix are equal to a mean of 0 and a standard deviation of 1. It is thus possible, given the universality property of the MP distribution, to sample a new noise matrix X with the same two first moments (mean = 0 and variance = 1) from a normal distribution.

To approximate the perturbation matrix, we use the singular values λ_i of X^s to calculate the expected singular values θ_i of the perturbation matrix based on equation 2.10. We replace the singular values λ_i of the matrix X^s with those of the perturbation matrix θ_i and call it P^s. In this way we have created a perturbation matrix with the expected singular values θ_i and unit singular vectors. Note that P^s contains noise and is thus different from the signal matrix P. Luckily, low rank-perturbation theory is independent of the exact distribution of the signal singular vectors.

Together, we obtain a sample measurement matrix (Step 1):

$$\tilde{X}^* = X + P^s.$$

We next calculate the values phiclust of $\tilde{X}^*$ (Step 2). By sampling new values for the noise matrix X several times (~ 50), and repeating step 1 and 2, we are now able to estimate the influence of random variations, in finite limits, on the additive perturbation and thus on phiclust.

We can subsequently calculate the upper ϕ^{up}_{clust} and lower $\phi^{down}_{\text{clust}}$ standard deviation as follows. Let k be the number of values above the original value ϕ^*_{clust} and N the total number of sampled values then

$$\phi^{up}_{\text{clust}} = \left(\frac{1}{k-1} \sum_{\phi^*_{\text{clust}} \geq \phi_{\text{clust}}} (\phi^*_{\text{clust}} - \phi_{\text{clust}})^2 \right)^{1/2} \tag{2.13}$$

$$\phi^{down}_{\text{clust}} = \left(\frac{1}{N-k-1} \sum_{\phi^*_{\text{clust}} < \phi_{\text{clust}}} (\phi^*_{\text{clust}} - \phi_{\text{clust}})^2 \right)^{1/2} \tag{2.14}$$

are the upper and lower boundaries of the interval.

Clusterability

Assessing clustering quality

We use two different methods to assess clustering quality, the adjusted rand index (ARI) and the silhouette coefficient.

Assuming two partitions, A and B, of a set of N cells, the rand index is defined as[21]:

$$RI(A,B) = \frac{N_{11} + N_{00}}{\binom{N}{2}},$$

where N_{11} is the number of pairs of elements that are in the same cluster in A and in the same cluster in B. N_{00} is the number of pairs of elements that are in a different cluster in A and in a different cluster in B. The rand index takes values between 0 and 1, where 0 indicates the complete lack of agreement between the partitions and 1 would indicate identical partitions. Even a random clustering of elements produces a non-zero rand index. The ARI is defined in such a way, that its value is on average 0 for a pair of partitions with randomly permuted cluster labels. A positive ARI thus indicates that partitions agree more than expected to happen by random chance. Let partition A have K_A clusters of sizes a_i and partition B have K_B clusters of sizes b_j, then the adjusted rand index is defined as:

$$ARI(A,B) = \frac{RI(A,B) - \mathbb{E}[RI(A,B)]}{1.0 - \mathbb{E}[RI(A,B)]} = \frac{\binom{N}{2}\sum_{k,m=1}^{K_A K_B}\binom{n_{km}}{2} - \sum_{m=1}^{K_A}\binom{a_k}{2}\sum_{m=1}^{K_B}\binom{b_m}{2}}{\frac{1}{2}\binom{N}{2}\left[\sum_{k=1}^{K_A}\binom{a_k}{2}\sum_{m=1}^{K_B}\binom{b_m}{2}\right] - \sum_{k=1}^{K_A}\binom{a_k}{2}\sum_{m=1}^{K_B}\binom{b_m}{2}}$$

For synthetic data, we take a high ARI between a clustering and the ground truth partition to indicate a clustering of high quality.

Another useful measure for clustering quality is the silhouette coefficient. Let $a(i)$ be the mean distance from point i to all other data points in the same cluster and $b(i)$ be the mean distance from point i to all other points from different clusters, then the silhouette coefficient is defined as [8]:

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

For the calculation of the distance, we consider the euclidean distance metric in the space spanned by the singular vectors that are associated with singular values above the MP distribution of the expression matrix $\tilde{X}$ (see 2.6). The final silhouette coefficient is taken as the mean value over all data points. For the calculation of the silhouette coefficient we use the *cluster R package (V 2.1.0)*.

Theoretically achievable clustering quality

A perfect clustering would coincide with the ground truth and obtain an ARI of 1. Here we argue that such a perfect clustering is in general not achievable, if there is noise in the data. In other words there is always a finite Bayes error rate (also called irreducible error) for assigning cells to the appropriate cluster. To construct a Bayes classifier, which achieves the minimal error rate, we need to know the ground truth partition. Hence, we use simulated data. For each ground truth cluster, we fit a multidimensional Gaussian to

the elements of the singular vectors of the expression matrix $\tilde{X}$ that correspond to the cells in the respective cluster (see Additional file 1: Fig. S3a). We only consider singular vectors with singular values above the MP distribution. For the fit we use the *mclust R package (V 5.4.6)*. We then construct a classifier by assigning a cell to the cluster for which it has the highest value of the fitted Gaussian distribution. This corresponds to the best clustering one can achieve if the ground truth partition is known. We define the theoretically achievable adjusted rand index (tARI) as the ARI between this best achievable clustering and the ground truth partition. Similarly, we define the theoretically achievable silhouette coefficient (tSIL) as the silhouette coefficient of the best achievable clustering. Since we use the fitted Gaussian distributions instead of the actual (unknown) distribution of singular vector elements, the constructed classifier only approximates the Bayes classifier. However, we confirmed empirically, that the tARI defined above is an upper bound for all tested clustering methods, which comprises the currently most popular tools used for single-cell RNA-seq data [Additional file 1: Fig. S3 b, c].
The tARI embodies our notion of clusterability. We define high clusterability as a low Bayes error rate for cluster assignments, which corresponds to a high tARI. We show empirically that our clusterability measure is a proxy of the tARI and thus a way to assess clusterability without knowing the ground truth [Fig. 1d].

Application to single-cell RNA-seq data

Preprocessing of scRNA-seq data

In the following the necessary preprocessing steps for the application of the clusterability measure for scRNA-seq data are described.

Transcriptome Mode

The largest eigenvalue λ_1 of an expression matrix is typically much larger than all the other singular values and its corresponding singular vector has entries of equal sign, which often have similar magnitude (of order $\frac{1}{\sqrt{N}}$, which is the ideal value in the perfectly homogeneous case). This singular vector reflects a general, global trend in the data. This structure has been observed for many empirical data matrices. (In time series analysis of the stock market, this singular vector is called the "market mode" since it corresponds to a trend that is common across many stocks [19]). Here, we refer to this singular vector as "transcriptome mode" since it reflects a trend that is shared across the whole transcriptome (see Additional file 1: Fig. S2 a-d). In order to reduce the influence of this singular value on the calculation of the MP fit, we center the expression matrix $\tilde{X}$ gene-wise. As a result, the singular value of the transcriptome mode will be reduced to a value close to 0.

Normalization

The efficiency of the capture of transcripts and their conversion to cDNA is known to be highly variable between cells. Hence, single-cell gene expression data is usually normalized cell-wise. We have tested several normalization methods but none of them seemed sufficient to remove all technical variability in the data. Thus, in section 2.6 we describe a method to reduce these effects for our clusterability measure phiclust. Nevertheless, we normalize the expression to the total counts per cell and subsequently log-transform to stabilize the variance.

Gene distribution
Gene expression is typically modelled by a Poisson, negative binomial or zero inflated negative binomial distribution. However, the parameters of these distributions differ between genes, this violates the assumptions of the MP theorem, where all values are sampled from the same distribution. In practice, gene-wise standardization to a mean of 0 and standard deviation of 1 mostly circumvents this problem. Additionally, we have observed that there is a bias resulting from variations in cells. These biases are as well reduced by standardising the cells to a mean of 0 and standard deviation of 1 (see Additional file 1: Fig. S2 c,d). This is equivalent to calculating the eigenvalues and vectors of a correlation matrix instead of a covariance matrix.

Zero inflation
Another factor to be considered is the large amount of zero values in scRNA-seq data. These zeros might be on the one hand due to technical artefacts (low efficiency, dropout) or simply due to low, stochastic gene expression. After performing the above mentioned preprocessing steps we mostly do not observe deviations from the MP distribution. However, this is a known problem discussed within the framework of sparsity induced singular values. For single cell RNA-seq data an extensive analysis has been performed in [14], where the authors observe deviations from the MP distribution caused by sparsity. The authors suggest the exclusion of outlier genes that can be identified through the fit of the MP distribution. For phiclust we do not use this preprocessing step, however we do exclude genes that have a high expression in only a few number of cells.

Regressing out unwanted sources of variability (Confounder Regression)
scRNA-seq data suffers from several sources of technical variability that can obscure or even be mistaken for relevant biological signal. One of the most important of these is the variable efficiency of mRNA capture and cDNA conversion. The total number of detected transcripts per cell is typically taken as a proxy of this efficiency. There are also biological processes that can cause unwanted signal. Most cells are stressed due to the tissue dissociation necessary for single-cell library preparation. The percentage of expression coming from mitochondrial genes or the expression of marker genes for stress can be used to estimate the level of stress. Different metabolic states of cells might be reflected in the level of ribosomal gene expression and many genes fluctuate with the cell cycle. Here, we seek to establish a method to remove any effect of these nuisance variables on the clusterability measure.
We model the signal matrix P as a sum of relevant signal B and unwanted signal due to nuisance variables Y. Inspired by published approaches to expression data normalization [23, 26], we model the influence of Y by linear regression. This is a valid approach because the regression is performed on the singular vectors of $\tilde{X}$, which contain Gaussian distributed noise. Given the singular value decomposition of $\tilde{X}$ and singular vectors $\tilde{v}_i$,

$$\tilde{v}_i = \beta Z, \quad \text{with } \beta \in \mathbb{R}^k \tag{2.15}$$

where $Z \in \mathbb{R}^{N \times k}$ is a matrix of covariates, such as the total counts per cell, with k the number of covariates and N the number of cells. Each covariate is normalized to a length

of 1 such that the range agrees with the range of the singular vectors. The amount of variance explained by the nuisance parameters is then given by the value of the adjusted R squared (R^2_{adj}) of this linear regression. Since the eigenvalues of the cell-cell correlation matrix can be interpreted as the amount of variance explained, we reduce the eigenvalues λ_i by $\tilde{\lambda}_i = (1 - R^2_{adj})\lambda_i$. In the next step, we calculate adjusted singular values by $\tilde{\gamma}_i = \sqrt{\tilde{\lambda}_i}$ and use these adjusted singular values $\tilde{\gamma}_i$ for the consecutive steps in the calculation of the clusterability measure.

Algorithm

The procedure to obtain the clusterability measure involves the following steps:

1. Preprocess the single cell expression matrix as described in section 2.6:
   (a) Normalization
   (b) Log-transformation
   (c) Standardization gene-wise
   (d) Standardization cell-wise
2. Calculate the singular value decomposition of the gene expression matrix $\tilde{X}$.
3. Fit the MP distribution to the singular values (equation 2.5).
4. Determine singular values/vectors that correspond to non-random variability using the Tracy-Widom distribution (equation 2.7) or the Shapiro-Wilk test (equation 2.8), respectively.
5. Adjust the singular values for effects of nuisance variables by linear regression (equation 2.15).
6. Calculate the singular values θ_i of the signal matrix P using the inverse of equation 2.9, given by 2.10.
7. Calculate the projections of the singular vectors of the expression matrix $\tilde{X}$ on the corresponding singular vector of the signal matrix P with equations 2.11 for the singular vectors and 2.12 for the gene singular vectors.
8. The clusterability measure is the largest of the projections for the singular vectors obtained in the previous step.

2.7 Supplementary Note 2

Application of phiclust to our previously published single-cell RNA-sequencing study of the human fetal kidney [31] revealed two distinct groups of clusters (Fig. 3a). Connecting tubule (CnT), nephron progenitor cells-a (NPCa), nephron progenitor cells-b (NPCb), and mesangial cells (Mes) all obtained a phiclust of 0, which signified that these clusters consist of pure populations with homogeneous gene-expression profiles. The rest of the clusters obtained higher values of phiclust, indicating that they contained subpopulations that were previously overlooked. For further analysis, we explored all clusters with highest phiclust and chose to further investigate clusters in which new cell populations were identified: the ureteric bud/collecting duct (UBCD), the S-Shaped Body proximal precursor cells (SSBpr), and the Interstitial cells a (ICa), with phiclust of 0.97, 0.95, and 0.93, respectively.

UBCD The analysis of this cluster yielded two clearly separate subpopulations (Fig. 3b, Additional file 1: Fig. S10b). The bigger subpopulation contained developing collecting duct cells and their precursors (ureteric bud), indicated by the expression of genes such as WFDC2, AQP2, CLDN3, MMP7, and CALB1. In contrast, the smaller sub-cluster showed little or no expression of the aforementioned genes and was characterized by UPK1A and UPK1B, well-known markers of the urothelial epithelium, which constitutes the inner lining of the ureter. The presence of such cells in our data is plausible given that the whole fetal kidney was used in our sequencing experiment. Both DE analysis (Table S4) and inspection of the top variance-driving genes (Additional file 1: Fig. S10e) revealed SPINK1, UPK2, S100A6, KRT7, and KRT19 as additional markers. Staining of week 15 fetal kidney sections with UPK1A and KRT7 antibodies confirmed our interpretation (Fig. 3c, Additional file 1: Fig. S11a). UPK1A was restricted to the superficial urothelial cells in major and minor calyces as well as the developing ureter. KRT7 was expressed more broadly, across the superficial, intermediate, and basal urothelium. Both KRT7 and UPK1A were completely absent from the whole collecting system and the branching ureteric bud, marked by CDH1 (Additional file 1: Fig. S11a).

SSBpr Sub-clustering the SSBpr population showed the presence of 3 subpopulations (Fig. 3b, Additional file 1: Fig. S10b). One subpopulation contained markers of proximal cell precursors (GPC3, LHX1, CADM2) together with low expression of AMN and APOE (see Table S4), which is consistent with the original annotation of the cluster. A second subpopulation, contiguous to the previous one, showed the expression of CLDN1, which is expressed in the proximal epithelium, together with CITED2, expressed in developing podocytes. This suggested parietal epithelial cells (PECs) as the most likely cell type, as these cells were reported to share several markers with both proximal epithelium and podocytes [45]. To confirm this interpretation, we performed Immunostaining of CLDN1, as well as CAV2 and AKAP12 which were found by DE analysis (Fig. 3d, Additional file 1: Fig. S11b). Interestingly, CLDN1 was found in all segments of the S-shaped body except in the precursors of the PECs, which are the thin layer of cells at the lateral side of the proximal segment of the SSB. CLDN1 appeared in the parietal epithelium only at the capillary loop stage and continued to be expressed in all PECs in more mature glomeruli. CAV2 was present in the parietal epithelium in developing glomeruli, but also in the endothelial cells of both the glomerular capillaries and the surrounding vasculature. Intriguingly, CAV2

overlapped with CLDN1 only in a subpopulation of PECs in individual glomeruli, which might indicate previously unobserved heterogeneity within these cells in the developing kidney. Only AKAP12 marked the precursors of the PECs in S-shaped bodies and continued to be abundantly expressed. However, AKAP12, was not specific to PECs, as it was also expressed in interstitial cells in the cortex. Finally, a third, small and distinct subpopulation in the SSBpr cluster expressed distal tubule markers (SPP1, ODC1, IRX3, and S100A10), suggesting that these cells were misclassified during the original clustering. This shows that phiclust can pinpoint clustering errors, making it a useful tool for clustering quality control.

ICa This cluster consisted of 5 subpopulations (Fig. 3b, Additional file 1: Fig. S10b). All subpopulations expressed markers of the renal interstitium. One also expressed genes found uniquely expressed in other cell types (EPCAM, CD24, BST2, NNAT, DAPL1) and thus likely contains doublets. Another small subset was characterized by markers of mesangial cells (MGP, ACTA2, PDGFRB), suggesting that it contains mesangial cells erroneously grouped with the ICa or renal pericytes, which share a similar gene expression profile [46]. Another subpopulation showed high expression of non-specific genes related to components and regulators of microtubules together with metabolic, mitochondrial and stress-related genes (H2AFZ, TUBA1B, TYMS, STMN1, DUT, MT-CO3, MT-ND5). The two remaining subpopulations were clearly interstitial but their gene expression profiles could not be linked to known interstitial populations, likely due to the dearth of knowledge about the renal stroma. We hypothesized that these two subpopulations were localized in different regions of the kidney. To test this idea, we stained fetal kidney sections with POSTN, CLDN11, and SULT1E1 (Fig. 3e, Additional file 1: Fig. S11c), which were identified by DE analysis and inspection of the top variance-driving genes. SULT1E1 was highly expressed in the pelvic area in the immediate vicinity of the developing ureter, as well as the inner and outer medulla, preferentially surrounding tubules. This marker might thus indicate the medullary interstitium as well as pelvic smooth muscle cells. Staining with CLDN11 showed a higher signal in the medulla and papilla, similar to SULT1E1, but with a wider spatial distribution. In contrast to SULT1E1, CLDN11 was also expressed in groups of cortical interstitial cells, situated directly underneath the renal capsule, in the nephrogenic zone. CLDN11 might thus also be expressed by the interstitial progenitor cells or their immediate progeny. Lastly, POSTN was mainly found in the renal cortex surrounding tubules and glomerular microvasculature. POSTN was also expressed in cortical blood vessels with larger diameters together with their arborizations. POSTN is a secreted extracellular matrix protein known to be expressed in cardiac smooth muscle cells, as well as connective tissues. Here, POSTN might mark smooth muscle cells of the cortical vasculature.

In conclusion, a reanalysis of our previously published data showed the ability of phiclust to reveal overlooked subpopulations. Interestingly, phiclust identified sub-clusters with only a few cells (41 developing PECs, 68 urothelial cells, 29 distal cells), highlighting its sensitivity to relevant substructure hidden within a bigger cluster. Finally, phiclust was also useful to pinpoint clustering errors and the presence of doublets, which makes it useful for quality control prior to DE analysis.

References

[1] M. Mircea et al. Phiclust: a clusterability measure for single-cell transcriptomics reveals phenotypic subpopulations. *Genome Biology*, 23(1):1–24, dec 2022.

[2] R. Satija et al. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*, 33(5):495–502, may 2015.

[3] F. A. Wolf, P. Angerer, and F. J. Theis. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1):15, feb 2018.

[4] J. A. Hartigan and M. A. Wong. Algorithm AS 136: A K-Means Clustering Algorithm. *Applied Statistics*, 28(1):100, 1979.

[5] F. Murtagh and P. Legendre. Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *Journal of Classification*, 31(3):274–295, oct 2014.

[6] V. Y. Kiselev, T. S. Andrews, and M. Hemberg. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nature Reviews Genetics*, 20(May):273–282, 2019.

[7] S. Ackerman, Margareta; Ben-David. Clusterability : A Theoretical Study. In M. van Dyk, David; Welling, editor, *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, volume 5, pages 1–8. PMLR, 2009.

[8] P. J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

[9] A. Adolfsson, M. Ackerman, and N. C. Brownstein. To cluster, or not to cluster: An analysis of clusterability methods. *Pattern Recognition*, 88:13–26, apr 2019.

[10] B. Liu et al. An entropy-based metric for assessing the purity of single cell populations. *Nature Communications*, 11(1):1–13, dec 2020.

[11] D. Paul and A. Aue. Random matrix theory in statistics: A review. *Journal of Statistical Planning and Inference*, 150:1–29, jul 2014.

[12] M. Potters, J.-P. Bouchaud, and L. Laloux. Financial Applications of Random Matrix Theory: Old Laces and New Pieces. *Acta Physica Polonica*, 35(9):2767–2784, 2005.

[13] O. Bohigas, M. J. Giannoni, and C. Schmit. Characterization of Chaotic Quantum Syectra and Universality of Level Fluctuation Laws. *Physical Review Letters*, 52(1):1–4, 1984.

[14] L. Aparicio, M. Bordyuh, A. J. Blumberg, and R. Rabadan. A Random Matrix Theory Approach to Denoise Single-Cell Data. *Patterns*, 1(3), 2020.

[15] G. Livan, M. Novaes, and P. Vivo. *Introduction to Random Matrices Theory and Practice*. Springer, Switzerland, 2018.

[16] C. A. Tracy and H. Widom. Level-spacing distributions and the Airy kernel. *Communications in Mathematical Physics*, 159(1):151–174, jan 1994.

[17] F. Benaych-Georges and R. R. Nadakuditi. The singular values and vectors of low rank perturbations of large rectangular random matrices. *Journal of Multivariate Analysis*, 111:120–135, 2012.

[18] E. P. Wigner. Characteristic Vectors of Bordered Matrices With Infinite Dimensions. Technical Report 3, 1955.

[19] M. Macmahon and D. Garlaschelli. Community Detection for Correlation Matrices. *Physical Review X*, 021006(5):1–34, 2015.

[20] L. Zappia, B. Phipson, and A. Oshlack. Splatter: Simulation of single-cell RNA sequencing data. *Genome Biology*, 18(1):174, sep 2017.

[21] A. J. Gates and Y.-Y. Ahn. The Impact of Random Models on Clustering Similarity. Technical report, 2017.

[22] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Elsevier, San Diego, 2 edition, 1990.

[23] C. Hafemeister and R. Satija. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biology*, 20(1):296, dec 2019.

[24] L. Tian et al. experiments. *Nature Methods*, 16(June), 2019.

[25] D. Risso et al. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nature Communications 2018 9:1*, 9(1):1–17, jan 2018.

[26] D. Grün. Revealing dynamics of gene expression variability in cell state space. *Nature Methods*, 17(1):45–49, jan 2020.

[27] W. Saelens, R. Cannoodt, H. Todorov, and Y. Saeys. A comparison of single-cell trajectory inference methods. *Nature Biotechnology*, 37(5):547–554, may 2019.

[28] T. Stuart and R. Satija. Integrative single-cell analysis. *Nature Reviews Genetics*, 20(5):257–272, jan 2019.

[29] F. V. Mello et al. Maturation-associated gene expression profiles during normal human bone marrow erythropoiesis. *Cell Death Discovery*, 5(1):69, dec 2019.

[30] A. C. Villani et al. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science*, 356(6335), apr 2017.

[31] M. Hochane et al. Single-cell transcriptomics reveals gene expression dynamics of human fetal kidney development. *PLOS Biology*, 17(2):e3000152, feb 2019.

[32] B. Adamson et al. A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. *Cell*, 167(7):1867–1882.e21, dec 2016.

[33] S. Bullett, T. Fearn, F. Smith, and I. E. Smolyarenko. An Introduction to Random Matrix Theory. *Advanced Techniques in Applied Mathematics*, pages 139–171, 2016.

[34] R. Mingo, James A; Speicher. *Free Probability and Random Matrices*. Springer New York LLC, 1 edition, 2017.

[35] K. Kendall and M. George. Kolmogorov–Smirnov Test. *The Concise Encyclopedia of Statistics*, pages 283–287, feb 2008.

[36] J. Bun, J.-p. Bouchaud, and M. Potters. Cleaning large correlation matrices : Tools from Random Matrix Theory. *Physics Reports*, 666:1–109, 2017.

[37] W. Haynes. Benjamini–Hochberg Method. *Encyclopedia of Systems Biology*, pages 78–78, 2013.

[38] F. Benaych-Georges and R. R. Nadakuditi. The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Advances in Mathematics*, 227(1):494–521, 2011.

[39] L. Scrucca, M. Fop, T. B. Murphy, and A. E. Raftery. Mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *R Journal*, 8(1):289–317, 2016.

[40] Z. Ji and H. Ji. TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Research*, 44(13):e117, jul 2016.

[41] I. Tirosh et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science*, 352(6282):189–196, apr 2016.

[42] A. T. Lun, K. Bach, and J. C. Marioni. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biology*, 17(1):1–14, apr 2016.

[43] S. C. van den Brink et al. Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. *Nature Methods*, 14(10):935–936, 2017.

[44] L. McInnes, J. Healy, and J. Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *Journal of Open Source Software*, 3(29):861, feb 2018.

[45] S. S. Guhr et al. The expression of podocyte-specific proteins in parietal epithelial cells is regulated by protein degradation. *Kidney International*, 84(3):532–544, sep 2013.

[46] X. Wang et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science*, 361(6400), jul 2018.