DUBStepR: correlation-based feature selection for clustering single-cell RNA sequencing data

Bobby Ranjan , Wenjie Sun , Jinyu Park , Ronald Xie , Fatemeh Alipour , Vipul Singhal , and Shyam Prabhakar ✉

[1]Laboratory of Systems Biology and Data Analytics, Genome Institute of Singapore, A*STAR, 60 Biopolis Street, Singapore - 138672

Feature selection (marker gene selection) is widely believed to improve clustering accuracy, and is thus a key component of single cell clustering pipelines. However, we found that the performance of existing feature selection methods was inconsistent across benchmark datasets, and occasionally even worse than without feature selection. Moreover, existing methods ignored information contained in gene-gene correlations. We therefore developed DUBStepR (Determining the Underlying Basis using Stepwise Regression), a feature selection algorithm that leverages gene-gene correlations with a novel measure of inhomogeneity in feature space, termed the Density Index (DI). Despite selecting a relatively small number of genes, DUBStepR substantially outperformed existing single-cell feature selection methods across diverse clustering benchmarks. In a published scRNA-seq dataset from sorted monocytes, DUBStepR sensitively detected a rare and previously invisible population of contaminating basophils. DUBStepR is scalable to large datasets, and can be straightforwardly applied to other data types such as single-cell ATAC-seq. We propose DUBStepR as a general-purpose feature selection solution for accurately clustering single-cell data.

single cell | feature selection | gene correlation | clustering
Correspondence: ***prabhakars@gis.a-star.edu.sg***

Introduction

Heterogeneity in single-cell RNA sequencing (scRNA-seq) datasets is frequently characterized by identifying cell clusters in gene expression space, wherein each cluster represents a distinct cell type or cell state. In particular, numerous studies have used unsupervised clustering to discover novel cell populations in heterogeneous samples (1). The steps involved in unsupervised clustering of scRNA-seq data have been well documented (2). i) Low-quality cells are first discarded in a quality control step (3). ii) Reads obtained from the remaining cells are then normalized to remove the influence of technical effects, while preserving true biological variation (4). iii) After normalization, feature selection is performed to select the subset of genes that are informative for clustering, iv) which are then typically reduced to a small number of dimensions using Principal Component Analysis (PCA) (5). v) In the reduced principal component (PC) space, cells are clustered based on their distance from one another (typically, Euclidean distance), and vi) the corresponding clusters are assigned a cell type or state label based on the known functions of their differentially expressed (DE) genes (6).
Although feature selection is a critical step in the canonical clustering workflow described above, only a few different approaches have been developed in this space. Moreover, there have been only a handful of systematic benchmarking studies of scRNA-seq feature selection methods (7, 8). A good feature selection algorithm is one that selects cell-type-specific (DE) genes as features, and rejects the remaining genes. More importantly, the algorithm should select features that optimize the separation between biologically distinct cell clusters. A comprehensive benchmarking study of feature selection methods would ideally use both of these metrics.

The most widely used approach for feature selection is mean-variance modeling: genes whose variation across cells exceeds a data-derived null model are selected as features (9, 10). Such genes are described as highly variable genes (HVGs) (11). Some earlier single cell studies instead selected genes with high loading on the top principal components of the gene expression matrix (high loading genes, or HLGs) as features (12). M3Drop, a more recent method, selects genes whose dropout rate (number of cells in which the gene is undetected) exceeds that of other genes with the same mean expression (8). As an alternative approach to detect rare cell types, GiniClust uses a modified Gini index to identify genes whose expression is concentrated in a relatively small number of cells (13). All of the above feature selection methods test genes individually, without considering expression relationships between genes. Another drawback is that existing methods for determining the size of the feature set do not bear direct relation to the separation of cells in the resulting space.

Here, we present Determining the Underlying Basis using Stepwise Regression (DUBStepR), an algorithm for feature selection based on gene-gene correlations. A key feature of DUBStepR is the use of a step-wise approach to identify an initial core set of genes that most strongly represent coherent expression variation in the dataset. Uniquely, DUBStepR defines a novel graph-based measure of cell aggregation in the feature space, and uses this measure to optimize the number of features. We benchmark DUBStepR against 6 commonly used feature selection algorithms on datasets from 4 different scRNA-seq protocols (10x Genomics, Drop-Seq, CEL-Seq2 and Smart-Seq2) and show that it substantially outperforms other methods. We used DUBStepR to detect a rare, contaminating basophil population in FACS-purified monocytes and dendritic cells (15). Finally, we show that DUBStepR could potentially be applied even to single-cell ATAC sequencing data.

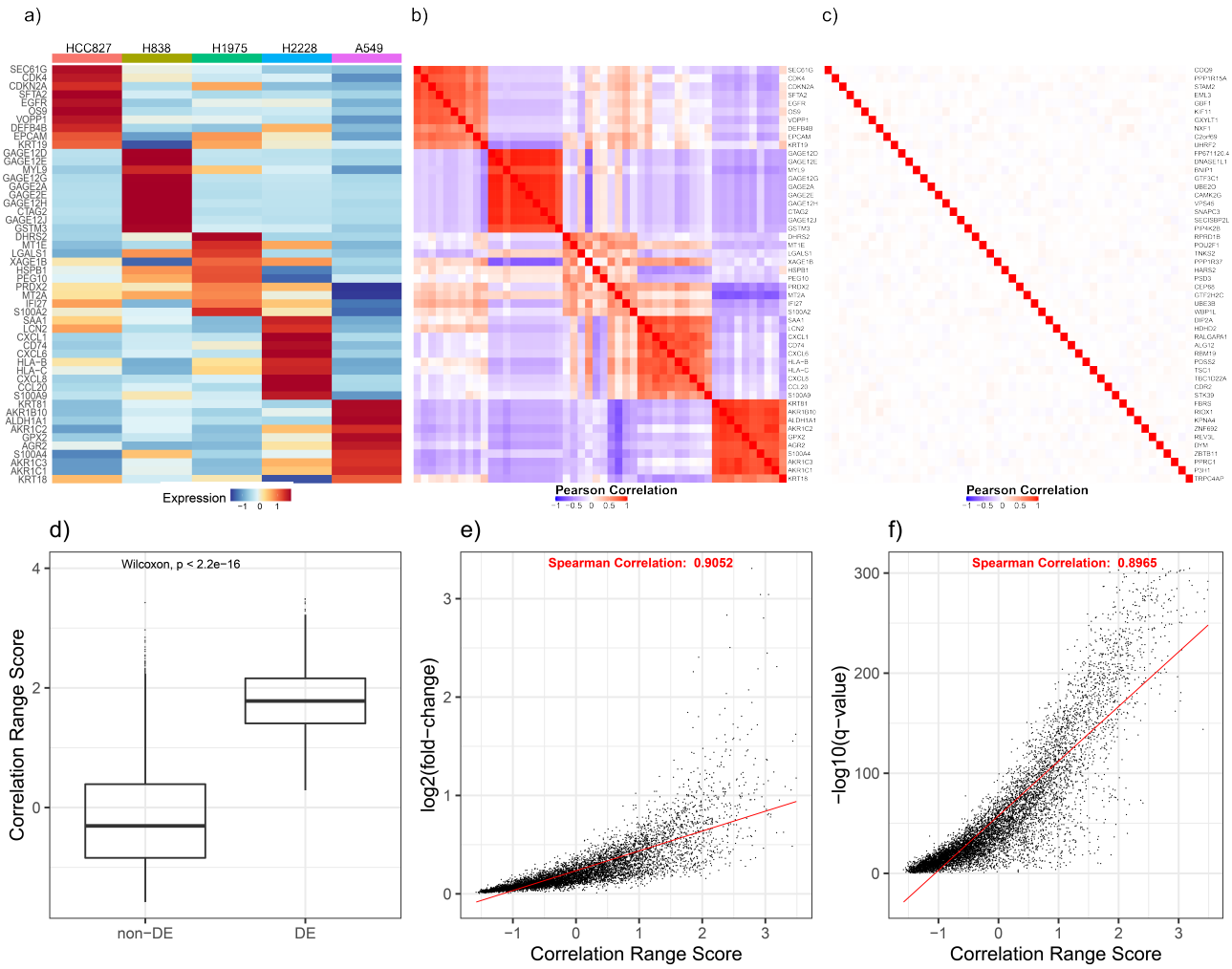

Fig. 1. Expression correlations of DE genes: scRNA-seq data from 5 lung adenocarcinoma cell lines (14). a) Average expression of top 10 DE genes for each cell type. b) Gene-gene correlations of the same genes. c) Gene-gene correlations for non-DE genes. d) Boxplot showing correlation range scores for non-DE and DE genes. e-f) Scatter plot of genes showing correlation between d) fold change of cell-type-specific expression and e) q-value of cell-type-specific expression with correlation range score.

Results

Gene-gene correlations predict cell-type-specific DE genes. The first step in DUBStepR is to select an initial set of candidate features based on known properties of cell-type-specific DE genes (marker genes). DE genes specific to the same cell types would tend to be highly correlated with each other, whereas those specific to distinct cell types are likely to be anti-correlated (Figure 1a-b; Methods). In contrast, non-DE genes are likely to be only weakly correlated (Figure 1c). We therefore hypothesized that a correlation range score derived from the difference between the strongest positive and strongest negative correlation coefficients of a gene (Methods), would be substantially elevated among DE genes. Indeed, we found that the correlation range score was significantly higher for DE genes relative to non-DE genes (Figure 1d). Moreover, the correlation range score of a gene was highly predictive of its greatest fold-change between cell types, and also its most significant differential expression q-value (Figure 1e-f). Due to the strong association between correlation range and marker gene status, DUBStepR selects genes with high correlation range score as the initial set of candidate feature genes (Methods).

Stepwise regression identifies a minimally redundant feature subset. We observed that candidate feature genes formed correlated blocks of varying size in the gene-gene correlation (GGC) matrix (Figure 2a), with each block presumably representing a distinct pattern of expression variation across the cells. To ensure more even representation of the diverse expression signatures within the candidate feature set, we sought to identify a representative minimally-redundant subset, which we termed "seed" genes. For this purpose, DUBStepR performs stepwise regression on the GGC matrix, regressing out, at each step, the gene explaining the largest amount of variance in the residual from the previous step (Figure 2b-d). We devised an efficient implementation of this procedure that requires only a single matrix multiplication at each step (Methods).

This approach selects seed genes with diverse patterns of cell-type-specificity (Figure 2e-h). DUBStepR then uses the elbow point of the stepwise regression scree plot to determine the optimal number of steps (Methods), i.e. the size of the seed gene set (Figure 2i,j).

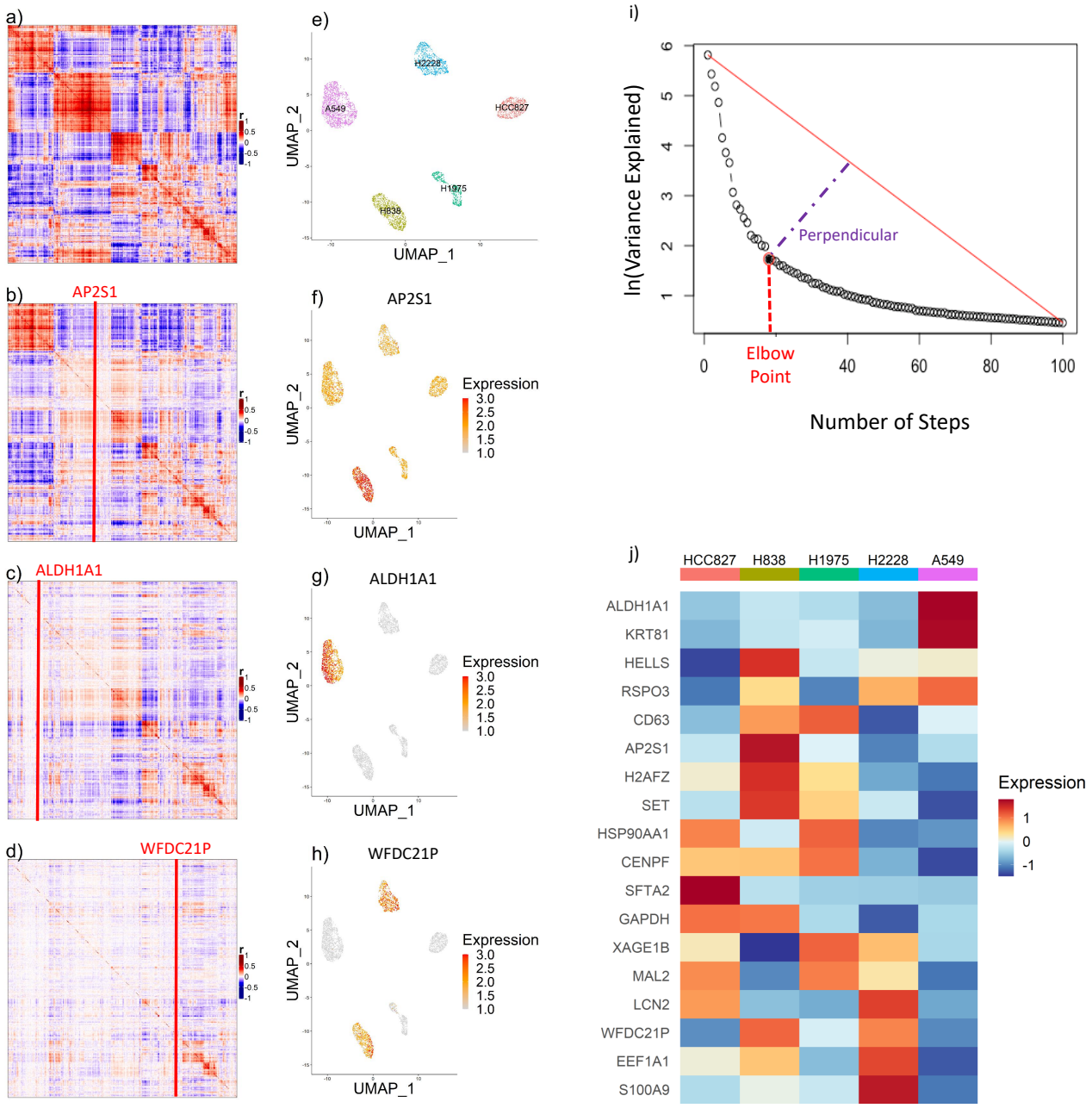

Fig. 2. a) Gene-gene correlation matrix of candidate feature genes (high correlation range score). b-d) Residuals from stepwise regression on the gene-gene correlation matrix. e) UMAP visualization of cells in gene expression space, colored by cell line. f-h) Same UMAP, colored by expression of genes regressed out in the first 3 steps. i) Scree plot: variance in GGC matrix explained by the gene regressed out at each step. j) Standardized average expression of the final seed gene set in each of the 5 cell lines.

Algorithm	Description	Software
GiniClust	Gini index-based feature selection	GiniClust (13)
HLG	Features ranked by sum of magnitude of PC loadings	irlba (8)
HVGDisp	Highly Variable Genes by dispersion value	Seurat (9)
HVGVST	Highly Variable Genes after variance-stabilized transformation	Seurat (9)
M3Drop/DANB	Dropout-based feature selection: M3Drop for read counts, DANB for UMIs	M3Drop (8)
trendVar	Biological and technical components of the gene-specific variance	scran (10)

Table 1. Feature selection methods used for the benchmarking comparison.

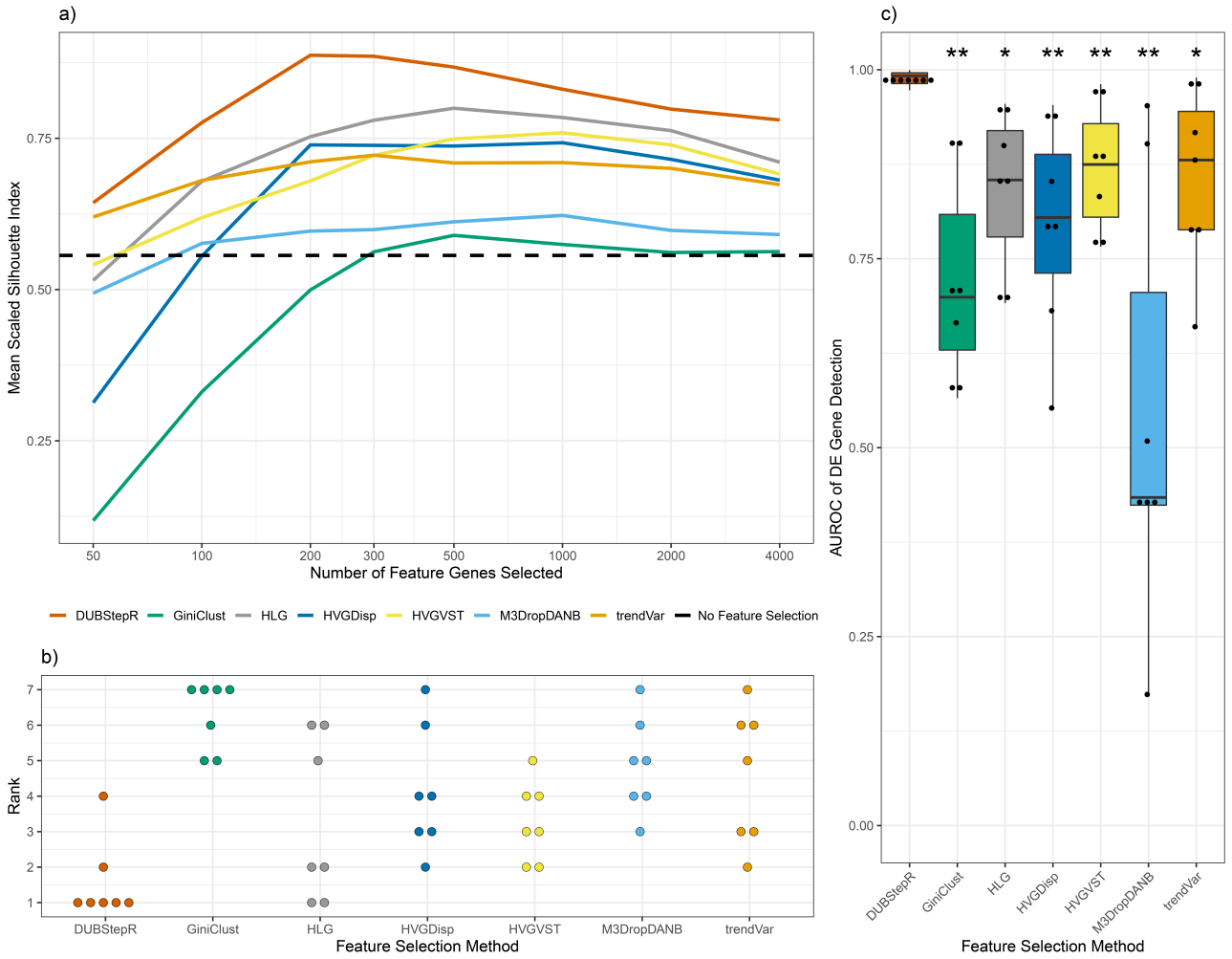

Fig. 3. Benchmarking feature selection methods. a) Mean scaled Silhouette Index of feature sets ranging from 50 to $4,000$ features. b) Rank distribution of feature selection methods. For each dataset, the 7 methods are ranked from 1 to 7 by their best SI across all feature set sizes. c) AUROC of DE gene detection. *: p <= 0.05, **: p <= 0.01.

Guilt-by-association expands the feature set. Although the seed genes in principle span the major expression signatures in the dataset, each individual signature is now represented by only a handful of genes (2-5 genes, in most cases). Given the high level of noise in scRNA-seq data, it is likely that this is insufficient to fully capture coherent variation across cells. DUBStepR therefore expands the seed gene set by iteratively adding correlated genes from the candidate feature set (Supp. Fig. S4; Methods). This process is continued until DUBStepR reaches the optimal number of feature genes (see below).

Benchmarking. To benchmark the performance of DUBStepR, we compared it against 5 other algorithms for feature selection in scRNA-seq data: three variants of the HVG approach (HVGDisp, HVGVST, trendVar), HLG and M3Drop/DANB (Table 1). For completeness we also benchmarked GiniClust, though it was designed only for identifying markers of rare cell types. Each algorithm was benchmarked on 7 datasets spanning 4 scRNA-seq protocols: 10X Genomics, Drop-Seq, CEL-Seq2, and Smart-Seq on the Fluidigm C1 (Supplementary Note 2A). These datasets were selected because the true cell type could be independently ascertained based on cell line identity or FACS gate. Our benchmarking approach thus avoids the circularity of using algorithmically defined cell type labels as ground truth.

To evaluate the quality of the selected features, we used the well-established Silhouette Index (SI), which quantifies cluster separation, i.e. closeness between cells belonging to the same cluster, relative to distance to cells from other clusters (16). In addition to being a well established measure of single cell cluster separation (17–19), the SI has the advantage of being independent of any downstream clustering algorithm. We evaluated the SI of each algorithm across a range of feature set sizes (50-$4,000$), scaled the SI values to a maximum of 1 for each dataset and then averaged the scaled SIs across the 7 datasets (Figure 3a; Supplementary Figure S2). Remarkably, HLG, an elementary PCA-based method that predates scRNA-seq technology, achieved greater average cell type separation than existing single cell algorithms at most feature set sizes. M3Drop, HVGDisp and trendVAR remained close to their respective performance peaks over a broad range from 200 to $2,000$ features, and dropped off on either side of this range. HVGVST, GiniClust and HLG showed a more peaked performance characteristic, with maxima at 500-$1,000$ features. DUBStepR substantially outper-

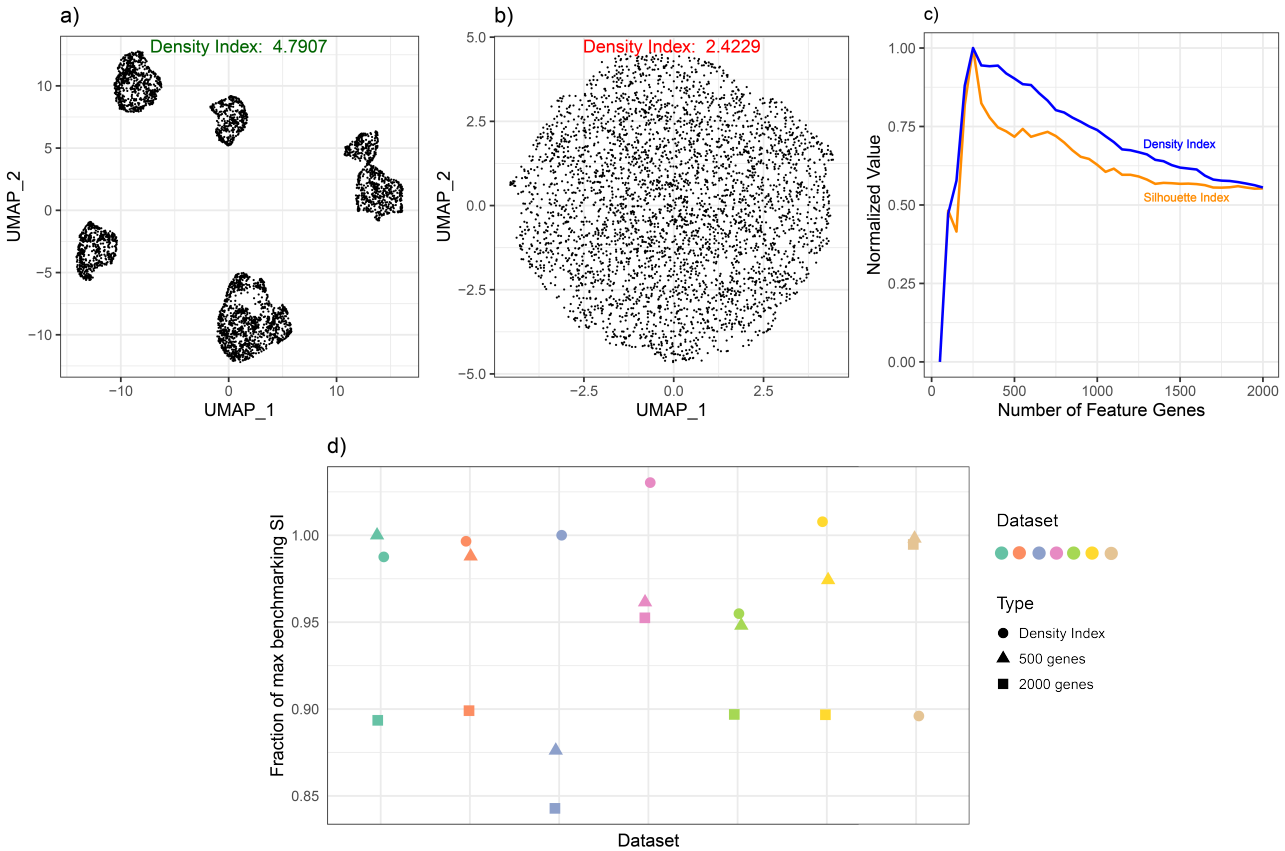

Fig. 4. a-b) UMAP visualizations of the 5 cell lines comparing the local neighbourhood density of feature spaces in (a) good feature selection versus (b) poor feature selection. c) Comparison of optimal Density Index and Silhouette Index over feature set size range. d) Fraction of the maximum benchmarking SI achieved using feature set sizes determined by the Density Index as compared to fixing feature set sizes at 500 genes and 2000 genes.

formed all other methods across the entire range of feature set size. Moreover, DUBStepR was the top-ranked algorithm on 5 of the 7 datasets (Figure 3b).

Ideally, a feature selection algorithm should select only cell-type-specific genes (DE genes) as features. As an independent benchmark, we therefore quantified the ability of feature selection algorithms to discriminate between DE and non-DE genes. To minimize the effect of ambiguously classified genes, we designated the top 500 most differentially expressed genes in each dataset as DE, and the bottom 500 as non-DE (Methods), and then quantified performance using the area under the receiver operating characteristic (AUROC). Remarkably, DUBStepR achieved an AUROC in excess of 0.97 on all 7 datasets, indicating near-perfect separation of DE and non-DE genes (Figure 3c). In contrast, none of the other methods exceeded a median AUROC of 0.88. Thus, DUBStepR vastly improves our ability to select cell-type-specific marker genes (DE genes) for clustering scRNA-seq data.

Density Index predicts the optimal feature set. As shown above, selecting too few or too many feature genes can result in sub-optimal clustering (Figure 3a). Ideally, we would want to select the feature set size that maximized cell type separation (i.e. the SI) in the feature space. However, since the feature selection algorithm by definition does not know the true cell type labels, it is not possible to calculate the SI for any given feature set size. We therefore endeavoured to define a proxy metric that would approximately model the SI without requiring knowledge of cell type labels. To this end, we defined a measure of the inhomogeneity or "clumpiness" of the distribution of cells in feature space. If each cell clump represented a distinct cell type, then this measure would tend to correlate with the SI. The measure, which we termed Density Index (DI), equals the root mean squared distance between all cell pairs, divided by the mean distance between a cell and its k nearest neighbours (Methods). Intuitively, when cells are well clustered and therefore inhomogeneously distributed in feature space, the distance to nearest neighbours should be minimal relative to the distance between random pairs of cells, and thus DI should be maximal (Figure 4a,b). Empirically, we found that DI and SI were indeed positively correlated and tended to reached their maxima at approximately the same feature set size (Figure 4c). Further, for 5 out of the 7 benchmarking datasets, the feature set with the highest DI also maximized SI (Figure 4d). Since our earlier analysis only tested a discrete number of feature set sizes (Figure 3a; Supplementary Note 3), the DI-guided approach even improved on the maximum SI in 2 cases (Figure 4d). One additional advantage of the DI is that it is relatively straightforward to compute, since the numerator is proportional to the square of the Frobenius norm of the gene expression matrix (Methods). By default, DUBStepR

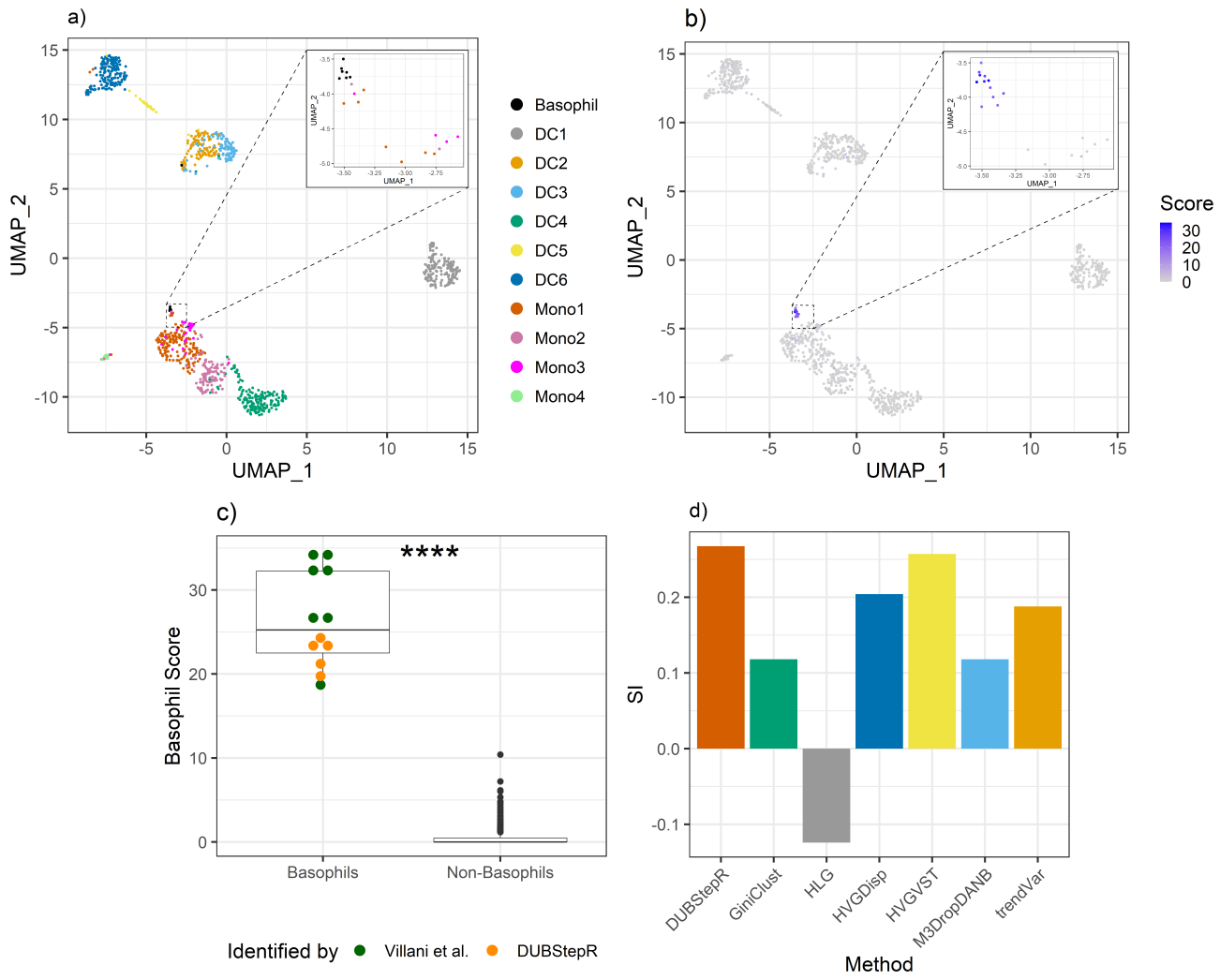

Fig. 5. a) UMAP visualization of scRNA-seq data from Villani et al. after feature selection using DUBStepR, with cells colored by cell type labels as defined by Villani et al. Inset: zoomed-in UMAP visualization of the basophils identified by DUBStepR. b) Cumulative score of canonical basophil markers (*CPA3*, *HDC*, *CLC*, *MS4A2*). c) Comparison of basophil marker expression between basophils and non-basophils in the dataset. d) SI of basophils identified by DUBStepR compared to existing methods. ****: p-value <= 0.0001

therefore selects the feature set size that maximizes DI.

DUBStepR robustly clusters rare contaminating basophils in sorted PBMCs. The above benchmarking analysis was largely based on detection of relatively common cell types ($> 10\%$ of all cells). To examine the performance of feature selection methods in detecting a rare cell type, we analyzed a published scRNA-seq dataset generated from human dendritic cells (DC) and monocytes purified from PBMCs (15). In this study, the authors manually discarded 7 contaminating basophil cells based on expression of the marker genes *CLC* and *MS4A2*. To evaluate the possibility of automatically detecting this contaminating population, we applied DUBStepR to the entire dataset of $1,085$ cells. We found that DUBStepR clearly separated the 7 basophils into a distinct group (Figure 5a), which intriguingly also included 5 additional cells. We scored each cell by its average expression of the canonical basophil markers (*CPA3*, *HDC*, *CLC*, *MS4A2*) and found that the 5 additional cells showed high basophil marker expression, indicating that they were indeed basophils (Figure 5b,c). In contrast to DUBStepR, some of the other feature selection methods did not clearly separate this rare population of cells (Figure 5d). This result suggests that DUBStepR could also be used to detect rare cell types in scRNA-seq data.

DUBStepR generalizes to scATAC-seq data. Feature selection is typically not performed on scATAC-seq data, since their mostly binary nature renders them refractory to conventional single cell feature selection based on highly variable features (21). However, since the logic of feature correlations applies even to binary values, we hypothesized that DUBStepR could also be applied to this data type. To test this hypothesis, we applied DUBStepR to scATAC-seq data from 8 FACS-purified subpopulations of human bone marrow cells (20). In contrast to the common approach of using all scATAC-seq peaks, we found that peaks selected by DUBStepR more clearly revealed the emergence of the three major lineages from the haematopoietic stem cells: lymphoid, myeloid and megakarocyte/erythroid (Figure 6a, b). Pseu-

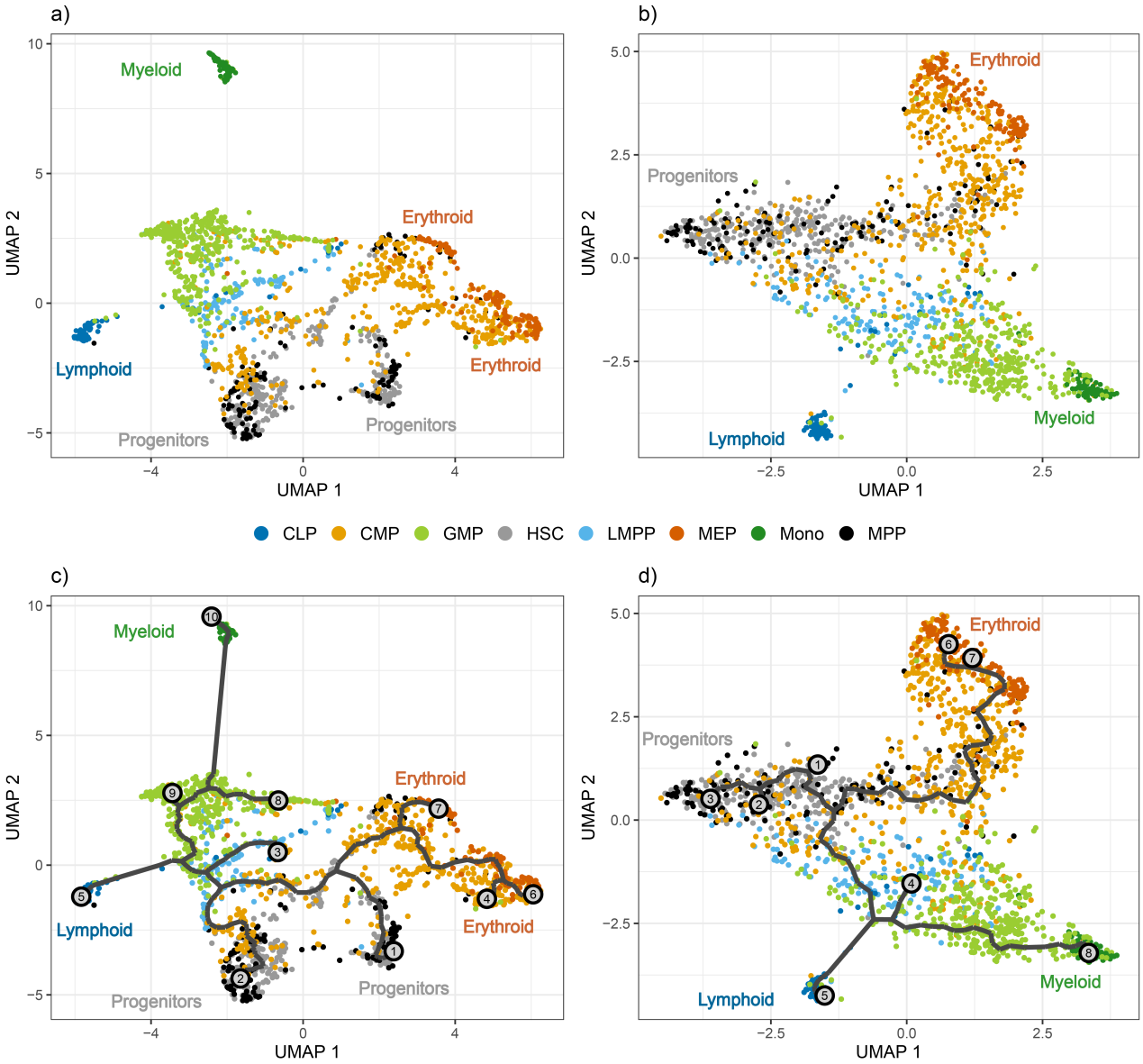

Fig. 6. UMAP visualization of scATAC-seq data from FACS-purified hematopoietic cell populations from human bone marrow (20). a) All scATAC-seq peaks, without any feature selection. b) Peaks selected by DUBStepR. c-d) Trajectories generated by pseudotemporal ordering using Monocle 3. c) All scATAC-seq peaks, and d) Peaks selected by DUBStepR. CLP: Common Lymphoid Progenitor, CMP: Common Myeloid Progenitor, GMP: Granulocyte/Macrophage Progenitor, HSC: Haematopoietic Stem Cell, LMPP: Lymphoid-primed Multi-Potential Progenitor, MEP: Megakaryocyte/Erythroid Progenitor, Mono: Monocyte, MPP: Multi-Potent Progenitor.

dotemporal trajectories generated using Monocle 3 (22) further corroborated this result (Figure 6c, d).

Discussion

DUBStepR is based on the intuition that cell-type-specific marker genes tend to be well correlated with each other, i.e. they typically have strong positive and negative correlations with other marker genes. Based on this intuition, the algorithm exploits structure in the gene-gene correlation matrix to prioritize genes as features for clustering. To benchmark this feature selection strategy, we used a stringently defined collection of single cell datasets for which cell type annotations could be independently ascertained (14). Note that this avoids the circularity of defining the ground truth based on the output of one of the algorithms being tested. Results from our benchmarking analyses indicate that, regardless of feature set size, DUBStepR separates cell types more clearly other methods (Figure 3a, b). This observation is further corroborated by the fact that DUBStepR predicts cell-type-specific marker genes substantially more accurately than other methods (Figure 3c). Thus, our results demonstrate that gene-gene correlations, which are ignored by conventional feature selection algorithms, provide a powerful basis for feature selection.

Intriguingly, DUBStepR approaches its maximum Silhouette Index value at 200-500 feature genes (Supplementary Figure S2), which is well below the default feature set size of 2,000 used in most single cell studies (9, 11). Thus, our results suggest that, if feature section is optimized, it may not be necessary to select a larger number of feature genes. Note

however that the optimum feature set size can vary across datasets (Supplementary Figure S2). Selecting a fixed number of feature genes for all datasets could therefore result in sub-optimal clustering (Figure 4d).

From the perspective of cell clustering, the optimal feature set size is that which maximizes cell type separation in feature space, which can be quantified using the SI. As an indirect correlate of cell type separation, we have defined a measure of the inhomogeneity or "clumpiness" of cells in feature space, which we termed the Density Index (DI). To our knowledge, DI is the only metric for scoring feature gene sets based on the distribution of cells in feature space. Our results suggest that the DI correlates with the SI, and that cluster separation is improved in most cases when the feature set is chosen to maximize DI. Another important advantage of the DI is that it is computationally straightforward to calculate from the Frobenius norm of the data matrix. It is possible that the DI measure could also be applied to other stages of the clustering pipeline, including dimensionality reduction (selecting the optimal number of PCs) and evaluation of normalization strategies.

Interestingly, although DUBStepR was not specifically designed to detect rare cell types, it nevertheless showed the greatest efficacy in distinguishing the 12 contaminating basophils present in the Villani et al. dataset (Figure 5d). Notably, the original study identified only the subset of 7 basophils contained within the dendritic cell pool - our analysis demonstrates that basophil contamination was also present in the sorted monocyte population (Figure 5a).

Algorithmic pipelines for single cell epigenomic data, for example scATAC-seq, typically do not incorporate a formal feature selection step (21, 23). In most cases, such pipelines merely discard genomic bins at the extremes of high and low sequence coverage. This is because the sparsity and near-binary nature of single-cell epigenomic reduces the efficacy of conventional feature selection based on mean-variance analysis. Since DUBStepR uses an orthogonal strategy based on correlations between features, it is less vulnerable to the limitations of single cell epigenomics data (Figure 6). Thus, DUBStepR opens up the possibility of incorporating a feature selection step in single cell epigenomic pipelines, including scATAC-seq, scChIP-seq and single-cell methylome sequencing.

Code Availability. DUBStepR is freely available as an R package on GitHub at https://github.com/prabhakarlab/DUBStepR, and is well documented for easy integration into the Seurat pipeline. Code for generating all the figures in this paper is provided in the Supplementary Files.

Data Availability. All datasets used in this paper are publicly available, as described in Supplementary Note 2. Processed data used for generating the figures in this paper are available on Zenodo at https://doi.org/10.5281/zenodo.4072261.

Acknowledgements. The authors would like to acknowledge Muzlifah Haniffa and Gary Reynolds for assistance with analysis of the Villani et al. dataset, and members of the Prabhakar lab for critical feedback and discussions.

Author Contributions. BR, WS, JP, FA, VS and SP developed DUBStepR. BR, WS, JP and RX performed benchmarking analyses. BR wrote the manuscript. BR and SP edited and reviewed the manuscript.

Funding. The work was supported by the Call for Data Analytics Proposal (CDAP) grant no. 1727600056.

Methods

Gene Filtering. By default, DUBStepR filters out genes that are not expressed in at least 5% of cells. We allow the user to adjust this parameter if they are interested in genes that are more sparsely expressed in their data set. Additionally, mitochondrial, ribosomal and pseudogenes (using the Ensembl pseudogene reference from GRCh38 (24)) are removed.

Correlation Range. Correlation range c_i for gene i can be defined for a gene-gene correlation matrix G as

$$c_i = max3(G_i)) - 0.75 \cdot min(G_i)).$$

Correlation range uses the second-largest non-self correlation value (denoted here as $max3$) to calculate the range, to protect against genes with overlapping 5' or 3' exons (25).

The minimum correlation value has been down-weighted to 0.75 to give greater importance to stronger positive correlations over negative correlations.

We first binned genes based on their mean expression level, as mean expression tends to correlate with technical noise (26). In each bin, we compute a z-score of correlation range of gene i as

$$z_i = \frac{c_i - \mu_c}{\sigma_c},$$

where μ_c is the mean correlation range of the gene and σ_c refers to variance in the correlation range scores of a gene. Genes with a z-score ≤ 0.7 are filtered out at this step.

Step-wise Regression. We define the step-wise regression equation as

$$G = gw^T + \epsilon,$$

where G is the column-wise zero-centered gene-gene correlation matrix, g is the column of the matrix G to be regressed out, ϵ is the matrix of residuals and w is a vector of regression coefficients. The squared error ($\epsilon^T\epsilon$) is minimized when

$$w^T = \frac{g^T G}{g^T g}$$

Thus,

$$G = \frac{gg^T G}{g^T g} + \epsilon.$$

We calculate variance explained by the regression step as $V = \|G - \epsilon\|_F^2$, where F indicates the Frobenius norm. To efficiently compute V for all genes, we define $X = G^T G$ and

x as the row of X corresponding to gene g. Thus, $x = g^T G$. We can simplify V as

$$\begin{aligned} V &= \|G - \epsilon\|_F^2 \\ &= \left\| \frac{gg^T G}{(g^T g)} \right\|_F^2 \\ &= \frac{\|gx\|_F^2}{(g^T g)^2} \\ &= \frac{\mathrm{Tr}((gx)^T (gx))}{(g^T g)^2} \\ &= \frac{\mathrm{Tr}(x^T x)}{g^T g} \\ &= \frac{xx^T}{g^T g}. \end{aligned}$$

Thus, we can use a single matrix multiplication $G^T G$ to efficiently calculate variance explained by each gene in the gene-gene correlation matrix, and then regress out the gene explaining the greatest variance. The residual from each step k is then used as the gene-gene correlation matrix for the next step. In other words,

$$\begin{aligned} G_k &= g_k w_k^T + \epsilon_k \\ G_{k+1} &= \epsilon_k. \end{aligned}$$

For computational efficiency, we repeat this for 100 steps.

To select the genes contributing to the major directions in G, we use the elbow point on a scree plot. The elbow point is the point on the scree plot that has the largest perpendicular distance to the line connecting the first and last points of the plot (27).

Once the genes representing the major directions of variation in gene correlations are identified using the elbow plot, they are used as seed genes to add correlated genes through guilt-by-association (28) in a single-linkage manner.

Density Index. For a given feature set, Principal Component Analysis (PCA) (5) is performed on the gene expression matrix and the top D principal components (PCs) are selected, where D is a user-specified parameter with a default value of 20. Let M be the matrix of embeddings of the gene expression vectors of N cells in D principal components. The root-mean-squared distance d_{rms} between pairs of cells i and j can be calculated as

$$d_{rms} = \sqrt{< d_{i,j}^2 >} = \sqrt{< \Sigma_{p=1}^{D} (M_{i,p} - M_{j,p})^2 >},$$

where $<>$ denotes the average over all pairs $\{(i,j) | i \in [1,N], j \in [1,N]\}$. Note that, for simplicity of the final result, we include pairs in which $i = j$. This can be further simplified as follows:

$$\begin{aligned} d_{rms} &= \sqrt{< \Sigma_{p=1}^{D} (M_{i,p}^2 + M_{j,p}^2 - 2M_{i,p}M_{j,p}) >} \\ &= \sqrt{\Sigma_{p=1}^{D} (< M_{i,p}^2 > + < M_{j,p}^2 > - 2 < M_{i,p}M_{j,p} >)} \\ &= \sqrt{\Sigma_{p=1}^{D} (< M_{i,p}^2 > + < M_{j,p}^2 >)} \\ &= \sqrt{2\Sigma_{p=1}^{D} < M_{i,p}^2 >} \\ &= \sqrt{\frac{2}{N}} \|M\|_F. \end{aligned}$$

In the above derivation, the mean product term $< M_{i,p}M_{j,p} >$ is zero because $M_{i,p}$ and $M_{j,p}$ have zero mean across i and j respectively. Let k_i denote the average distance of cell i from its k nearest neighbours, and k_m denote the mean of k_i across all cells. We define the Density Index (DI) as

$$DI = \frac{d_{rms}}{k_m} = \sqrt{\frac{2}{N}} \frac{\|M\|_F}{k_m}.$$

Bibliography

1. Tallulah S Andrews and Martin Hemberg. Identifying cell populations with scrnaseq. *Molecular aspects of medicine*, 59:114–122, 2018.
2. Vladimir Yu Kiselev, Tallulah S Andrews, and Martin Hemberg. Challenges in unsupervised clustering of single-cell rna-seq data. *Nature Reviews Genetics*, 20(5):273–282, 2019.
3. Peng Jiang. *Quality Control of Single-Cell RNA-seq*, pages 1–9. Springer New York, New York, NY, 2019. ISBN 978-1-4939-9057-3. doi: 10.1007/978-1-4939-9057-3_1.
4. Catalina A Vallejos, Davide Risso, Antonio Scialdone, Sandrine Dudoit, and John C Marioni. Normalizing single-cell rna sequencing data: challenges and opportunities. *Nature methods*, 14(6):565, 2017.
5. Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.
6. Tianyu Wang, Boyang Li, Craig E Nelson, and Sheida Nabavi. Comparative analysis of differential gene expression analysis tools for single-cell rna sequencing data. *BMC bioinformatics*, 20(1):40, 2019.
7. Shun H Yip, Pak Chung Sham, and Junwen Wang. Evaluation of tools for highly variable gene discovery from single-cell rna-seq data. *Briefings in bioinformatics*, 20(4):1583–1589, 2019.
8. Tallulah S Andrews and Martin Hemberg. M3drop: Dropout-based feature selection for scrnaseq. *Bioinformatics*, 2018.
9. Rahul Satija, Jeffrey A Farrell, David Gennert, Alexander F Schier, and Aviv Regev. Spatial reconstruction of single-cell gene expression data. *Nature biotechnology*, 33(5):495, 2015.
10. Aaron TL Lun, Davis J McCarthy, and John C Marioni. A step-by-step workflow for low-level analysis of single-cell rna-seq data with bioconductor. *F1000Research*, 5, 2016.
11. Malte D Luecken and Fabian J Theis. Current best practices in single-cell rna-seq analysis: a tutorial. *Molecular systems biology*, 15(6), 2019.
12. Barbara Treutlein, Doug G Brownfield, Angela R Wu, Norma F Neff, Gary L Mantalas, F Hernan Espinoza, Tushar J Desai, Mark A Krasnow, and Stephen R Quake. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell rna-seq. *Nature*, 509(7500): 371–375, 2014.
13. Lan Jiang, Huidong Chen, Luca Pinello, and Guo-Cheng Yuan. Giniclust: detecting rare cell types from single-cell gene expression data with gini index. *Genome biology*, 17(1):144, 2016.
14. Luyi Tian, Xueyi Dong, Saskia Freytag, Kim-Anh Lê Cao, Shian Su, Abolfazl JalalAbadi, Daniela Amann-Zalcenstein, Tom S Weber, Azadeh Seidi, Jafar S Jabbari, et al. Benchmarking single cell rna-sequencing analysis pipelines using mixture control experiments. *Nature methods*, 16(6):479–487, 2019.
15. Alexandra-Chloé Villani, Rahul Satija, Gary Reynolds, Siranush Sarkizova, Karthik Shekhar, James Fletcher, Morgane Griesbeck, Andrew Butler, Shiwei Zheng, Suzan Lazo, et al. Single-cell rna-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science*, 356(6335):eaah4573, 2017.
16. Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.
17. Xianwen Ren, Liangtao Zheng, and Zemin Zhang. Sscc: a novel computational framework for rapid and accurate clustering large-scale single cell rna-seq data. *Genomics, proteomics & bioinformatics*, 17(2):201–210, 2019.
18. Changde Cheng, John Easton, Celeste Rosencrance, Yan Li, Bensheng Ju, Justin Williams, Heather L Mulder, Yakun Pang, Wenan Chen, and Xiang Chen. Latent cellular analysis robustly reveals subtle diversity in large-scale single-cell rna-seq data. *Nucleic Acids Research*, 47(22):e143–e143, 2019.
19. Fan Zhang, Kevin Wei, Kamil Slowikowski, Chamith Y Fonseka, Deepak A Rao, Stephen Kelly, Susan M Goodman, Darren Tabechian, Laura B Hughes, Karen Salomon-Escoto,

et al. Defining inflammatory cell states in rheumatoid arthritis joint synovial tissues by integrating single-cell transcriptomics and mass cytometry. *Nature immunology*, 20(7):928–942, 2019.

20. Jason D Buenrostro, M Ryan Corces, Caleb A Lareau, Beijing Wu, Alicia N Schep, Martin J Aryee, Ravindra Majeti, Howard Y Chang, and William J Greenleaf. Integrated single-cell analysis maps the continuous regulatory landscape of human hematopoietic differentiation. *Cell*, 173(6):1535–1548, 2018.
21. Signac: A framework for the analysis and exploration of single-cell chromatin data., 2020.
22. Junyue Cao, Malte Spielmann, Xiaojie Qiu, Xingfan Huang, Daniel M Ibrahim, Andrew J Hill, Fan Zhang, Stefan Mundlos, Lena Christiansen, Frank J Steemers, et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, 566(7745):496–502, 2019.
23. Huidong Chen, Caleb Lareau, Tommaso Andreani, Michael E Vinyard, Sara P Garcia, Kendell Clement, Miguel A Andrade-Navarro, Jason D Buenrostro, and Luca Pinello. Assessment of computational methods for the analysis of single-cell atac-seq data. *Genome biology*, 20(1):1–25, 2019.
24. Fiona Cunningham, Premanand Achuthan, Wasiu Akanni, James Allen, M Ridwan Amode, Irina M Armean, Ruth Bennett, Jyothish Bhai, Konstantinos Billis, Sanjay Boddu, et al. Ensembl 2019. *Nucleic acids research*, 47(D1):D745–D751, 2019.
25. Tomohiro Nakayama, Satoshi Asai, Yasuo Takahashi, Oto Maekawa, and Yasuji Kasama. Overlapping of genes in the human genome. *International journal of biomedical science: IJBS*, 3(1):14, 2007.
26. Anna Mantsoki, Guillaume Devailly, and Anagha Joshi. Gene expression variability in mammalian embryonic stem cells using single cell rna-seq data. *Computational biology and chemistry*, 63:52–61, 2016.
27. Find the elbow in a curve: http://paulbourke.net/geometry/pointlineplane/pointline.r, 1996.
28. Michael G Walker, Wayne Volkmuth, Einat Sprinzak, David Hodgson, and Tod Klingler. Prediction of gene function by genome-scale expression analysis: prostate cancer-associated genes. *Genome research*, 9(12):1198–1203, 1999.
29. Bobby Ranjan, Florian Schmidt, Wenjie Sun, Jinyu Park, Mohammad Amin Honardoost, Joanna Tan, Nirmala Arul Rayan, and Shyam Prabhakar. scconsensus: combining supervised and unsupervised clustering for cell type identification in single-cell rna sequencing data. *bioRxiv*, 2020.
30. Itay Tirosh, Benjamin Izar, Sanjay M Prakadan, Marc H Wadsworth, Daniel Treacy, John J Trombetta, Asaf Rotem, Christopher Rodman, Christine Lian, George Murphy, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell rna-seq. *Science*, 352(6282):189–196, 2016.
31. Huipeng Li, Elise T Courtois, Debarka Sengupta, Yuliana Tan, Kok Hao Chen, Jolene Jie Lin Goh, Say Li Kong, Clarinda Chua, Lim Kiat Hon, Wah Siew Tan, et al. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nature genetics*, 49(5):708, 2017.