Sparse reduced-rank regression for exploratory visualization of single cell patch-seq recordings

Dmitry Kobak[1], Marissa A. Weis[1,2], and Philipp Berens[1]

[1] *Institute for Ophthalmic Research, Center for Integrative Neuroscience and Bernstein Center for Computational Neuroscience, University of Tübingen, Tübingen, Germany*
[2] *Graduate Training Center for Neuroscience, University of Tübingen, Tübingen, Germany*

April 16, 2018

Abstract

High-throughput single cell transcriptomics is rapidly emerging as the technique of choice to establish a census of neurons in the nervous system. Integrating the resulting cell type census with a physiological and anatomical taxonomy has been difficult, as most techniques require the tissue to be dissociated before sequencing. The recently proposed patch-seq technique allows to acquire multi-modal single cell data, where RNA-seq data is collected together with physiological and morphological information from the same cells. The technique typically results in data sets which have many more dimensions (expression levels of genes and electrophysiological properties) than measurements (cells), making it computationally difficult to relate the two modalities. Here we present a framework based on sparse reduced-rank regression for obtaining an interpretable visualization of the relationship between high-dimensional transcriptomic data and electrophysiological information on the single-cell level.

Introduction

Since the days of Ramón y Cajal, neuroscientists have classified neurons into cell types, which are often considered the fundamental building blocks of neural circuits (Masland, 2004). Classically, these types have been defined based on their physiology or anatomy, but due to the recent rise of single cell transcriptomics, a definition of cell types based on genetics is becoming increasingly popular (Poulin et al., 2016). For example, high-throughput single cell transcriptomics approaches have been used to establish a census of neurons in the retina (Shekhar et al., 2016; Macosko et al., 2015) and the cortex (Tasic et al., 2016, 2017; Zeisel et al., 2015) of mice.

Despite this success, it has proven difficult to integrate the obtained cell type taxonomy based on the transcriptome with information about physiology and anatomy. *On the cell type level*, single genes have been shown to be correlated with physiological properties in a meta-study on a brain-wide database of cell types, characterized using microarrays and electrophysiology (Tripathy et al., 2017). To be able to relate gene expression patterns to physiological characteristics *on the single cell level*, we need to be able to obtain the transcriptomes as well as electrophysiological measurements from the same cells and then integrate the resulting data sets.

The experimental capability to achieve this was developed with patch-seq (Cadwell et al., 2016; Fuzik et al., 2016; Cadwell et al., 2017; Földy et al., 2016), a technique that allows obtaining the transcriptome of cells characterized electrophysiologically or morphologically (Figure 1a). In contrast to other single cell transcriptomics experiments, these expertiments are rather low throughput, resulting in a multi-modal dataset with particular statistical structure: for a few dozen of cells, we have expression data on several thousands of genes as well as dozens of electrophysiological measurements (Figure 1a). Integrating and properly visualizing genetic and physiological information in this $n \ll p$ regime requires specialized techniques, which allow extracting an interpretable subset of genes and exploit as much information about the relationship between genes and physiology as possible to increase statistical power.

Here we present a framework based on sparse reduced-rank regression for obtaining an interpretable visualization of the relationship between high-dimensional single cell transcriptomes and electrophysiological information obtained using techniques like patch-seq. The method yields an

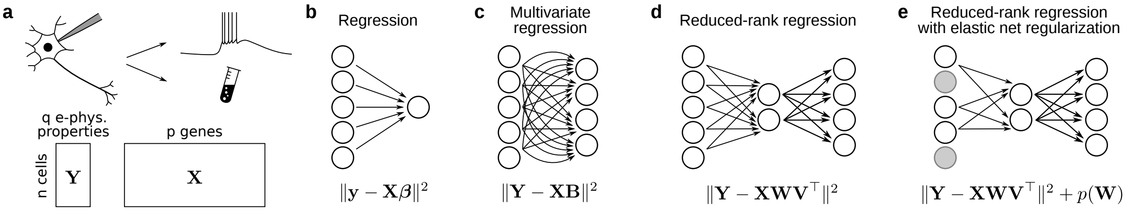
a
q e-phys. properties
p genes
n cells
Y
X
b
Regression
$\|\mathbf{y} - \mathbf{X}\beta\|^2$
c
Multivariate regression
$\|\mathbf{Y} - \mathbf{XB}\|^2$
d
Reduced-rank regression
$\|\mathbf{Y} - \mathbf{XWV}^\top\|^2$
e
Reduced-rank regression with elastic net regularization
$\|\mathbf{Y} - \mathbf{XWV}^\top\|^2 + p(\mathbf{W})$

Figure 1: a. Schematic illustration of a patch-seq experiment: electrophysiological activity is recorded by patch-clamping, followed by RNA extraction and sequencing. Below: data matrices after computational characterization of electrophysiological properties and estimation of gene counts. **b–e.** Schematic illustrations and loss functions for several regression methods. **b.** Simple regression. **c.** Multivariate regression. **d.** Reduced-rank regression. **e.** Regularized reduced-rank regression. Gray circles denote predictors that are left out of the sparse model.

intuitive low-dimensional representation of key features of the data, relating the dominant gene expression patterns that predict variation in the electrophysiological space.

Results

To relate gene expression patterns to electrophysiological properties, one can use the genetic data to predict any given electrophysiological property (Cadwell et al., 2016). This is a *regression* problem: each gene is a predictor and AP threshold is the response variable (Figure 1b). To predict multiple electrophysiological properties at the same time, one can combine individual regression problems into a *multivariate regression* problem where the response is a multivariate vector (Figure 1c).

However, different electrophysiological properties tend to be strongly correlated and so one could construct a more parsimonious model where gene expression is predicting several latent factors that in turn predict all the electrophysiological properties together (Figure 1d). These latent factors form a “bottleneck” in the linear mapping and allow exploiting correlations between the predicted elecrophysiological properties to increase statistical power. This is called *reduced-rank regression* (RRR) and can be solved by running principal component analysis (PCA) on the results of multivariate regression (see Methods). An attractive property of RRR is that it can be viewed not only as a prediction method, but also as a dimensionality reduction method, allowing visualization and exploration of the multi-modal dataset.

As there are over 20 thousand genes in a mouse genome and the typical sample size of a patch-seq data set is on the order of $n \approx 100$, all of these regression problems are in the $n \ll p$ regime and need to be regularized. Here we use elastic net regularization, which combines ℓ_1 (lasso) and ℓ_2 (ridge) penalties. This enforces sparsity and performs feature selection: only a small subset of genes are selected into the model while all other genes get zero regression coefficients (Figure 1e).

Mathematically, our method minimizes the following loss function (see Methods for details):

$$\mathcal{L} = \|\mathbf{Y} - \mathbf{XWV}^\top\|^2 + \lambda_1 \sum_{i=1}^{p} \|\mathbf{W}_{i\cdot}\|_2 + \lambda_2 \|\mathbf{W}\|^2 \quad \text{s.t. } \mathbf{V}^\top\mathbf{V} = \mathbf{I}.$$

Our elastic net RRR extends a recently suggested sparse RRR (Chen and Huang, 2012) and can be implemented using the `glmnet`-package (Friedman et al., 2010), a popular library for elastic net regression (see Methods). This model yields latent factors **XW** that can be interpreted as a low-dimensional genetic variability that is predictive of electrophysiological variability. Similarly, it allows to interpret **YV** as a low-dimensional electrophysiological variability that can be predicted from the genetic variability.

We applied our RRR approach to the patch-seq data set from Cadwell et al. (Cadwell et al., 2016) (see Methods for preprocessing). This data set encompasses two classes of interneurons, single bouqet cells (SBC) and elongated neurogliaform cells (eNGC), and contains $n = 44$ neurons. We used a permutation approach to estimate the rank r (dimensionality of the bottleneck) given the available data. For this data set, we obtained $r = 2$ (see Methods). We then used cross-validation (see Methods) to select the optimal values of the regularization parameters λ_1 and λ_2 (Figure 2a,b). At the maximum, the model achieved test-set $R^2 \approx 0.22$, a test-set correlation

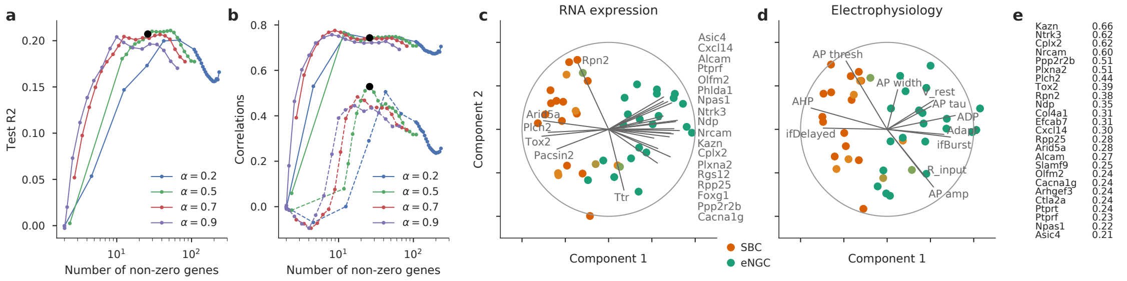

Figure 2: Reduced-rank regression (RRR) of the Cadwell et al. data set. **a.** Cross-validation performance: test-set R^2 depending on regularization parameters α and λ. Horizontal axis shows the mean number of non-zero genes selected on the training sets. Black dot shows the parameter values used for panels c–e: $\alpha = .5, \lambda = 1.4$. **b.** Test-set correlations between the first pair of RRR components (solid lines) and between the second pair of RRR components (dashed lines). **c.** Biplot in the transcriptomic space. Dots show single cells (color denotes cell type; intermediate colors correspond to cells that were not categorized unambiguously), lines show genes that were selected into the model. Each line shows correlations of a gene with the first two RRR components. The circle shows maximal possible correlation. **d.** Biplot in the space of electrophysiological properties. **e.** Probabilities for each gene to be selected into the model, estimated with bootstrapping. Top 25 genes shown.

between the first pair of components $\rho_1 \approx 0.75$, and test-set correlation between the second pair of components $\rho_2 \approx 0.5$.

We can use the bottleneck representation in the RRR model ($\mathbf{XW}$) to visualize and explore the genetic data (Figure 2c). The first component is clearly associated with the cell type. In contrast, the second one is uncorrelated to cell type and corresponds to within-type variation. Only 24 genes are selected as contributing to the model, shown as lines on Figure 2c: each line represents a gene's correlations to the RRR components 1 and 2 (the circle shows the maximum possible correlation). In the PCA literature, such visualization is called a "biplot" and we will adopt this terminology here. 22 out of the 24 genes are strongly correlated with the first component, with 18 of them having higher expression in the eNGC cells and four having higher expression in the SBC cells. Many of these genes were identified as differentially expressed in the original publication (Cadwell et al., 2016). The two remaining genes are strongly correlated with the second component.

We can visualize the electrophysiological space in a similar manner (using $\mathbf{YV}$) as a biplot with all 11 available electrophysiological properties (Figure 2d). Comparing the directions of variables on both biplots can suggest which electrophysiological variables are associated with which genes (e.g. AP threshold is positively correlated with Rpn2 expression level). We suggest to call the pair of RRR biplots a "bibiplot".

One important caveat is that the list of selected genes (Figure 2c) should not be interpreted as definite. This is for two reasons. First, the model performance (Figure 2a,b) was unaffected in some range of parameters corresponding to selecting from ∼10 to ∼50 genes, meaning that the choice of regularization strength in this interval remains an analyst's call. As an example, we show gene-space biplots with 10 and 40 genes in Figure S1. Second, even for fixed regularization parameters, a somewhat different set of genes may be selected every time when bootstrapping the model (this is often true for lasso-regularized models, especially when $n \ll p$). We show the frequencies with which some genes are selected during bootstrapping in Figure 2e. There is an interplay between these two factors. Stronger ℓ_1 regularization leads to a sparser model with less bootstrap reliability. Weaker ℓ_1 regularization leads to a less sparse model with more bootstrap reliability.

The RRR biplots can be compared to PCA biplots, made in the gene space and in the electrophysiological space independently from each other (Figure S2; a similar analysis on another data set with $n = 11$ neurons was done in (Harris et al., 2017)). In the electrophysiological space, the RRR biplot is almost identical to the PCA biplot, meaning that our RRR model explains the dominant modes of variation among the dependent variables. In the gene space, the situation is different. The first PCA component also separates the two cell types (albeit less clearly), but the second component is only weakly related to the PC2 in the electrophysiological space (correlations between the first/second pair: 0.74 and 0.22). Apart from that, the PCA in the gene space is not sparse, making the biplot practically impossible to display and interpret as it would have to show

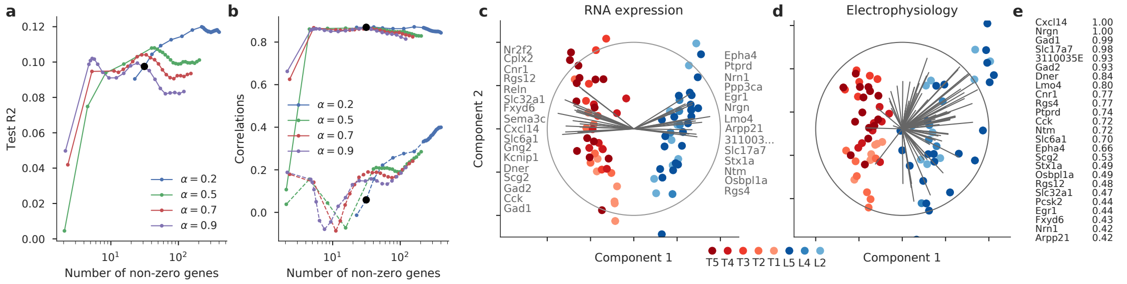

Figure 3: Reduced-rank regression (RRR) of the Fuzik et al. data set. **a.** Cross-validation performance: test-set R^2 depending on regularization parameters α and λ. Horizontal axis shows the mean number of non-zero genes selected on the training sets. Black arrow shows the parameter values used for panels c–e: $\alpha = .2, \lambda = 8$. **b.** Test-set correlations between the first pair of RRR components (solid lines) and between the second pair of RRR components (dashed lines). **c.** Biplot in the transcriptomic space. Dots show single cells (color denotes cell type), lines show genes that were selected into the model. Each line shows correlations of a gene with the first two RRR components. The circle shows maximal possible correlation. **d.** Biplot in the space of electrophysiological properties. Individual properties are not labeled because there were too many of them. **e.** Probabilities for each gene to be selected into the model, estimated with bootstrapping. Top 25 genes shown.

thousands of genes.

In addition, we applied our framework to the patch-seq dataset from Fuzik et al. (Fuzik et al., 2016) which encompasses $n = 80$ inhibitory and excitatory neurons from layers $1/2$ of mouse somatosensory cortex (Figure 3). The first RRR component strongly separated excitatory from inhibitory neurons, which is not surprising given the large differences in gene expression and in firing patterns between these two classes of neurons. At the same time, applying PCA separately to each modality results in pronounced class difference along multiple PCs (PC1, PC2, and some others; Figure S3), whereas RRR is capable of identifying and isolating this co-variation between modalities as the first component.

However, the second RRR component did not seem to carry a lot of signal in this dataset. Correlation was very weak (Figure 3b), in particular when the regularization was strong enough to select < 100 genes. Cross-validation indicated that when using several hundreds of genes or more the second RRR component became more pronounced (Figure 3a,b), but given that this could not be attributed to a smaller set of genes we suspect that it represents some unidentified experimental bias. Also, when running RRR separately for the inhibitory and the excitatory classes, we were unable to identify meaningful RRR components. For that reason here we chose the values of regularization parameters that did not yield any genes associated with the second component.

Discussion

We suggested regularized cross-validated sparse reduced-rank regression as a tool for interpretable data exploration and visualization of patch-seq datasets. It allows to visualize the variability across cells in transcriptomic and electrophysiological modalities in a consistent way, and to find a sparse set of genes explaining electrophysiological variability. Cross-validation allows to estimate the out-of-sample validity of the model. We expect that our method will also be relevant beyond the scope of patch-seq data: Spatial transcriptomics (Lein et al., 2017) combined with two-photon imaging may allow characterizing the transcriptome and physiology of individual cells in the intact tissue, yielding large multi-modal data sets. Similarly, other types of "multi-omics" data where single-cell or bulk transcriptomic data are combined with some other type of measurements (e.g. chemical, medical, or even behavioural), may benefit from interpretable visualization techniques.

Reduced rank-regression is closely related to the two other classical dimensionality reduction methods analyzing two data matrices ("two views") together: canonical correlation analysis (CCA) and partial least squares (PLS). These can be understood as looking for projections with maximal correlation (CCA) or maximal covariance (PLS), whereas RRR looks for projections with maximal explained variance in $\mathbf{Y}$. In recent years, multiple approaches to sparse CCA and sparse PLS have been suggested (Witten et al., 2009; Wilms and Croux, 2015; Lê Cao et al., 2008; Chun and Keleş, 2010) (among others). Here, we chose sparse RRR at the core of our framework, because it seemed more meaningful to predict electrophyiological properties from transcriptomic data instead of treating them symmetrically. Also, sparse RRR allows a mathematically simple formulation for rank $r > 1$ (using group lasso, see Methods) and can be conveniently implemented using existing implementations of elastic net regression.

To regularize the RRR model and to achieve sparse solutions, we used the elastic net penalty. It has two parameters, α and λ, and cross-validation will often indicate that α can be varied in some range without affecting the validation performance (see e.g. Figure 2a). This allows the researcher to control the trade-off between a sparser solution and a more comprehensive gene selection. If there is a set of genes that are highly correlated between each other, then large α will tend to select only one of them, whereas small α will tend to assign similar weights to all of them. Using $\alpha = 0$ corresponds to RRR with pure lasso regularization as suggested in (Chen and Huang, 2012). In the datasets analyzed here, we found that values $\alpha \approx .5$ yielded a good compromise.

In principle, it would be possible to generalize this regression framework to nonlinear mappings, using e.g. a neural network with a bottleneck instead of the low-rank linear mapping shown in Figure 1e. This can be an interesting direction for future research, but fitting such models would require much larger sample sizes than currently available for patch-seq data.

Python code for this manuscript is available at `https://github.com/berenslab/patch-seq-rrr`.

Methods

Data

For the data by Cadwell et al. (Cadwell et al., 2016) we used the RPKM values, for the Fuzik et al. data set (Fuzik et al., 2016) the UMI counts as gene expression data. In the Cadwell et al. data set there are $n = 51$ interneurons (from 53 sequenced interneurons, 2 were excluded in the original publication as "contaminated"), $p = 15074$ genes identified by the authors as "detected", and $q = 11$ electrophysiological properties. In Fuzik et al. data set there are $n = 83$ cells, $p = 24378$ genes after excluding ERCC spike ins, and $q = 89$ electrophysiological properties. Out of 83 sequenced cells, we were only able to match $n = 80$ to the electrophysiological data. We used only $q = 80$ electrophysiological properties for which the data were available for all these cells (the fact that $n = q = 80$ is coincidental).

We performed library size normalization by dividing the values for each cell i by the cell sum over all genes ("library size") and multiplying the result by the median library size across all cells:

$$X_{ij} = \frac{X_{ij}}{\sum_g X_{ig}} \cdot \operatorname*{Median}_c \Big[\sum_g X_{cg}\Big].$$

We then log-transformed the data using $\log_2(x+1)$ transformation. We excluded all cells for which at least one electrophysiological property was not estimated. Further, we excluded all genes that had exactly zero expression for all remaining cells. Finally, we standardized all gene expression values and all electrophysiological properties (to zero mean and unit variance).

For the Cadwell et al. data set, this yielded the final sample size of $n = 44$, $p = 15054$ genes, and $q = 11$ electrophysiological properties. We restricted the gene pool to the $p = 3000$ most variable genes (the same ones identified in the original publication) for all our analyses. We used the expert classification of cells into two classes performed in the original publication for annotating cell types. Out of $n = 44$ cells, only 35 cells were classified unambiguously (score 1 or score 5 on the scale from 1 to 5); the remaining 9 cells received intermediate scores.

Using the Cadwell et al. data, we tried two modifications of the above preprocessing pipeline: first, we left out the standardization of the transcriptomic data; second, we left out feature selection and used all available genes instead of 3000 most variables one. In both cases the cross-validated R^2 was somewhat lower than with our default pipeline, but when using $\alpha \approx .5$ and λ necessary to get $\sim$20 genes selected, we obtained very similar RRR projections.

For the Fuzik et al. data set, the same preprocessing pipeline yielded $n = 80$, $p = 13089$, and $q = 80$. We selected $p = 1384$ genes with average expression above .5 (before standardization) for the RRR analysis. The $n = 80$ cells have been classified in the original publication into L2, L4, and L5 excitatory neurons and into five classes of interneurons (labeled T1 to T5 in Figure 3).

All data sets were provided by the authors.

Algorithm

We consider two data matrices, $\mathbf{X}$ of $n \times p$ size and $\mathbf{Y}$ of $n \times q$ size that contain two sets of measurements on the same n samples. We assume that both matrices are *centered*, i.e. column means have been subtracted.

For simplicity, we first consider the special case of rank $r = 1$. The loss function of reduced-rank regression (RRR) in this case can be written as

$$\mathcal{L}_{\mathrm{RRR}} = \|\mathbf{Y} - \mathbf{X}\mathbf{w}\mathbf{v}^\top\|^2, \tag{1}$$

where without loss of generality it is convenient to require that $\|\mathbf{v}\| = 1$. Here and below all matrix norms are Frobenius norms. The product $\mathbf{w}\mathbf{v}^\top$ forms the matrix of regression coefficients that has rank $r = 1$. This decomposition allows to interpret $\mathbf{w}$ as a mapping that transforms $\mathbf{X}$ into latent variables and $\mathbf{v}$ as a mapping that transforms latent variables into $\mathbf{Y}$ (Figure 1e).

RRR can be directly solved using singular vector decomposition (SVD). Indeed, the loss can be decomposed into the ordinary least squares (OLS) loss and the rank approximation loss:

$$\mathcal{L}_{\mathrm{RRR}} = \|\mathbf{Y} - \mathbf{X}\mathbf{w}\mathbf{v}^\top\|^2 = \|\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}_{\mathrm{OLS}}\|^2 + \|\mathbf{X}\hat{\mathbf{B}}_{\mathrm{OLS}} - \mathbf{X}\mathbf{w}\mathbf{v}^\top\|^2, \tag{2}$$

where $\hat{\mathbf{B}}_{\mathrm{OLS}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$ is the solution to the un-penalized OLS regression. The first term corresponds to the variance of $\mathbf{Y}$ that is un-explainable by any linear model. The minimum of the second term can be obtained using SVD of $\mathbf{X}\hat{\mathbf{B}}_{\mathrm{OLS}}$. The right singular vector corresponding to the largest singular value gives $\hat{\mathbf{v}}$, and $\hat{\mathbf{u}} = \hat{\mathbf{B}}_{\mathrm{OLS}}\hat{\mathbf{v}}^\top$.

We now add the elastic net penalty to the loss function that linearly combines the lasso (ℓ_1-norm) and the ridge (ℓ_2-norm) penalties:

$$\mathcal{L}_{\mathrm{enRRR}} = \|\mathbf{Y} - \mathbf{X}\mathbf{w}\mathbf{v}^\top\|^2 + \lambda_1\|\mathbf{w}\|_1 + \lambda_2\|\mathbf{w}\|_2^2. \tag{3}$$

The penalties are only applied to the vector $\mathbf{w}$ because the vector $\mathbf{v}$ has a fixed ℓ_2 norm anyway, and ℓ_1 penalty would be inappropriate because we do not wish to make it sparse. This optimization problem is biconvex and can be solved with an iterative "alternating" approach: in turn, we fix $\mathbf{v}$ and find the optimal $\mathbf{w}_{\mathrm{opt}}$ and then fix $\mathbf{w}$ and find the optimal $\mathbf{v}_{\mathrm{opt}}$ until convergence.

For fixed $\mathbf{w}$, the loss does not depend on the penalty terms and the least-squares term can be written as

$$\begin{aligned}\|\mathbf{Y} - \mathbf{X}\mathbf{w}\mathbf{v}^\top\|^2 &= \|\mathbf{Y}\|^2 - \mathrm{tr}(\mathbf{v}\mathbf{w}^\top\mathbf{X}^\top\mathbf{X}\mathbf{w}\mathbf{v}^\top) - 2\,\mathrm{tr}(\mathbf{Y}^\top\mathbf{X}\mathbf{w}\mathbf{v}^\top)\\ &= \mathrm{const} - 2\,\mathrm{tr}(\mathbf{Y}^\top\mathbf{X}\mathbf{w}\mathbf{v}^\top),\end{aligned} \tag{4}$$

which is minimized when $\mathbf{v}$ is aligned with $\mathbf{Y}^\top\mathbf{X}\mathbf{w}$, i.e.

$$\mathbf{v}_{\mathrm{opt}} = \frac{\mathbf{Y}^\top\mathbf{X}\mathbf{w}}{\|\mathbf{Y}^\top\mathbf{X}\mathbf{w}\|}. \tag{5}$$

For fixed $\mathbf{v}$, the least-squares term can be re-written as

$$\begin{aligned}\|\mathbf{Y} - \mathbf{X}\mathbf{w}\mathbf{v}^\top\|^2 &= \mathrm{tr}(\mathbf{Y}^\top\mathbf{Y}) + \mathrm{tr}(\mathbf{v}\mathbf{w}^\top\mathbf{X}^\top\mathbf{X}\mathbf{w}\mathbf{v}^\top) - 2\,\mathrm{tr}(\mathbf{v}\mathbf{w}^\top\mathbf{X}^\top\mathbf{Y})\\ &= \mathrm{const} + \mathrm{tr}(\mathbf{v}^\top\mathbf{Y}^\top\mathbf{Y}\mathbf{v}) + \mathrm{tr}(\mathbf{w}^\top\mathbf{X}^\top\mathbf{X}\mathbf{w}) - 2\,\mathrm{tr}(\mathbf{w}^\top\mathbf{X}^\top\mathbf{Y}\mathbf{v})\\ &= \mathrm{const} + \|\mathbf{Y}\mathbf{v} - \mathbf{X}\mathbf{w}\|^2,\end{aligned} \tag{6}$$

meaning that the loss is equivalent to

$$\|\mathbf{Y}\mathbf{v} - \mathbf{X}\mathbf{w}\|^2 + \lambda_1\|\mathbf{w}\|_1 + \lambda_2\|\mathbf{w}\|_2^2. \tag{7}$$

This is the loss of elastic net regression of $\mathbf{Y}\mathbf{v}$ on $\mathbf{X}$, and so the optimal $\mathbf{w}_{\mathrm{opt}}$ can be obtained using any of the many available elastic net libraries. We used `glmnet` (Friedman et al., 2010) which is readily available for Matlab, Python, and R. It uses the following parameterization of the loss which we also adopt here:

$$\mathcal{L}_{\mathrm{enRRR}} = \frac{1}{2n}\|\mathbf{Y}\mathbf{v} - \mathbf{X}\mathbf{w}\|^2 + \lambda\Big(\alpha\|\mathbf{w}\|_1 + (1-\alpha)\|\mathbf{w}\|_2^2/2\Big). \tag{8}$$

Here α controls the trade-off between the lasso and the ridge and λ controls the overall regularization strength.

Now we can consider the general case of any rank r. Instead of $\mathbf{w}$ and $\mathbf{v}$ being vectors we now have $\mathbf{W}$ and $\mathbf{V}$ being matrices of $p \times r$ and $q \times r$ shapes, respectively. Without loss of generality, it is convenient to constrain the encoder matrix $\mathbf{V}$ to have orthonormal columns: $\mathbf{V}^\top \mathbf{V} = \mathbf{I}$. The RRR loss term and the ridge term have the same form as before, but the lasso term needs to be modified. We want the matrix $\mathbf{W}$ to be sparse in the sense that some of the genes are left out of the model entirely. This means that the entire rows of $\mathbf{W}$, and not just its individual elements, should be zeroed out. This can be achieved by a group lasso penalty term $\sum_{i=1}^p \|\mathbf{W}_{i\cdot}\|_2 = \sum_{i=1}^p \sqrt{\sum_{j=1}^r W_{ij}^2}$ that computes the sum of ℓ_2 norms of each row of $\mathbf{W}$. This is lasso in disguise because it can be seen as the ℓ_1 norm of the vector of row norms. Conveniently, `glmnet` allows to fit such models using the `family="mgaussian"` option.

Using `glmnet`-like parameterization, the loss is

$$\mathcal{L}_{\mathrm{enRRR}} = \frac{1}{2n}\|\mathbf{Y} - \mathbf{X}\mathbf{W}\mathbf{V}^\top\|^2 + \lambda\Big(\alpha \sum_{i=1}^{p} \|\mathbf{W}_{i\cdot}\|_2 + (1-\alpha)\|\mathbf{W}\|^2/2\Big). \tag{9}$$

For fixed $\mathbf{V}$, the optimal $\mathbf{W}_{\mathrm{opt}}$ can be obtained by `glmnet`. For fixed $\mathbf{W}$, this is an example of the orthogonal Procrustes problem (Gower and Dijksterhuis, 2004). Using the same argument as in Equation 4, we need to maximize $\mathrm{tr}(\mathbf{Y}^\top \mathbf{X}\mathbf{W}\mathbf{V}^\top)$. This can be achieved by the "thin" SVD of $\mathbf{Y}^\top\mathbf{X}\mathbf{W}$. If the left and right singular vectors are stacked in columns of $\mathbf{L}$ and $\mathbf{R}$ respectively, then $\mathbf{V}_{\mathrm{opt}} = \mathbf{L}\mathbf{R}^\top$. We provide a short proof below.

Given that the loss function is biconvex but possibly not jointly convex in $\mathbf{V}$ and $\mathbf{W}$, it is important to choose a reasonable initialization. We initialized $\mathbf{V}$ by the r leading right singular vectors of $\mathbf{X}^\top\mathbf{Y}$ and found this strategy to work well.

Relaxed elastic net

It is well-known that elastic net or even the lasso penalty on its own can lead to an over-shrinkage effect when the non-zero coefficients are shrunk too much. There have been several suggestions in the literature on how to address this problem (Efron et al., 2004; Zou and Hastie, 2005; Meinshausen, 2007). For example, *relaxed lasso* (Meinshausen, 2007) does lasso regression with a penalty λ_1 and then, using only the terms with non-zero coefficients, does another lasso regression with a different penalty λ_2. If $\lambda_2 = 0$ this is also called "LARS-OLS hybrid" (Efron et al., 2004). Similar procedures for elastic net are not as established. We found that we obtain an improvement if after RRR with elastic net penalty with coefficients λ and α, we take only the genes with non-zero coefficients and run RRR with pure ridge penalty ($\alpha = 0$) with the same value of λ. This procedure does not introduce any additional tuning parameters but substantially outperformed pure elastic net RRR on our data.

Cross-validation

We used repeated k-fold cross-validation (CV) to select the values of λ and α. We used two measures of performance: (i) the test-set reconstruction error $\|\mathbf{Y}_{\mathrm{test}} - \mathbf{X}_{\mathrm{test}}\hat{\mathbf{W}}\hat{\mathbf{V}}^\top\|^2$ and (ii) the test-set correlation coefficients $\mathrm{corr}(\mathbf{Y}_{\mathrm{test}}\hat{\mathbf{v}}, \mathbf{X}_{\mathrm{test}}\hat{\mathbf{w}})$ for all columns of $\hat{\mathbf{W}}$ and $\hat{\mathbf{V}}$. The correlation is not directly optimized by RRR but it is arguably a more intuitive metric and is what we are mostly looking for when we look at how well the two biplots in a bibiplot match each other.

We found it convenient to work with $k \approx 10$. Leave-one-out CV is less applicable in this case because it does not allow to look at the test-set correlations. We found that cross-validation curves were quite sensitive to the random splitting into k folds. For that reason we used repeated cross-validation, averaging all the error estimates across 100 random splits into folds.

Biplots

For any linear dimensionality reduction method that reduces dataset $\mathbf{X}$ with n samples and p columns to $\mathbf{X}\mathbf{W}$ with 2 columns, we construct the corresponding biplot as follows.

The scatter plot shows n points with x-coordinates given by $\mathbf{X}\mathbf{W}_{\cdot 1}$ and y-coordinates given by $\mathbf{X}\mathbf{W}_{\cdot 2}$, both standardized to have unit variance. The p lines show correlations between the original variables in $\mathbf{X}$ and the projections $\mathbf{X}\mathbf{W}$ such that the i-th variable is represented as a vector with coordinates $\big(\mathrm{corr}(\mathbf{X}_{\cdot i}, \mathbf{X}\mathbf{W}_{\cdot 1}), \mathrm{corr}(\mathbf{X}_{\cdot i}, \mathbf{X}\mathbf{W}_{\cdot 2})\big)$. It is convenient to scale these vectors with a constant factor γ for better visibility. We used $\gamma = 2$. We also display a circle with radius γ that shows the maximal possible extent of the vectors (assuming that the columns of $\mathbf{XW}$ are uncorrelated).

If $\mathbf{W}$ is sparse, then we only show the variables corresponding to non-zero rows of $\mathbf{W}$ (even though other variables can also have non-zero correlations with $\mathbf{XW}$). In case of reduced-rank regression, we use $\mathbf{XW}$ as the dimensionality reduction mapping for $\mathbf{X}$ and $\mathbf{YV}$ as the dimensionality reduction mapping for $\mathbf{Y}$.

Permutation-based rank estimation

We used a permutation-based procedure to estimate the rank of the linear mapping between $\mathbf{X}$ and $\mathbf{Y}$ (Figure S4). First, we estimate the dimensionality of $\mathbf{X}$ as follows. PCA on $\mathbf{X}$ yields a sequence of eigenvalues sorted in the decreasing order. Randomly permuting (shuffling) the rows of $\mathbf{X}$ for each of the columns separately, we can obtain a dataset $\tilde{\mathbf{X}}$ that preserves marginal variances (and so the sum of the PCA eigenvalues as well) but sets all the population correlations to zero. Consequently, PCA on $\tilde{\mathbf{X}}$ yields a sequence of decreasing eigenvalues under the null hypothesis that all variables are uncorrelated. We repeat this $n_\mathrm{rep} = 100$ times and estimate dimensionality d of $\mathbf{X}$ as the number of its eigenvalues that are above the 95th percentile of the shuffled eigenvalues. The procedure yields $d = 13$ and $d = 16$ for the Cadwell et al. and the Fuzik et al. datasets respectively.

Now we apply PCA to reduce the dimensionality of $\mathbf{X}$ to d. Let us call this reduced dataset $\mathbf{Z}$. We perform unregularized RRR of $\mathbf{Y}$ on $\mathbf{Z}$ by doing SVD of $\mathbf{Z}\hat{\mathbf{B}}_\mathrm{OLS}$, as described above. This yields a sequence of d singular values sorted in decreasing order. Randomly permuting the rows of $\mathbf{Z}$ yields a dataset $\tilde{\mathbf{Z}}$ that has exactly the same covariance matrix as $\mathbf{Z}$ but is statistically independent of $\mathbf{Y}$. RRR gives a sequence of decreasing singular values under the null hypothesis that $\mathbf{Z}$ and $\mathbf{Y}$ are unrelated. We repeat this $n_\mathrm{rep} = 100$ times and estimate the rank r of the linear mapping as the number of singular values that are above the 95th percentile of the shuffled singular values. This procedure yields $r = 2$ and $r = 3$ for the Cadwell et al. and the Fuzik et al. datasets, respectively. For both datasets we used RRR with rank $r = 2$.

Bootstrapping

We used bootstrapping to estimate the reliability of the gene selection. The following procedure was repeated $n_\mathrm{rep} = 100$ times: we selected n out of n cells with repetition and applied regularized RRR to the resulting data. This allows to estimate frequency with which each gene gets selected into the model. We used the same values of regularization parameters for all bootstrap repetitions.

Procrustes problem

Given $\mathbf{A}$, the problem is to maximize $\mathrm{tr}(\mathbf{AV}^\top)$ subject to $\mathbf{V}^\top\mathbf{V} = \mathbf{I}$. Let us denote by $\mathbf{A} = \mathbf{LQR}^\top = \tilde{\mathbf{L}}\tilde{\mathbf{Q}}\mathbf{R}^\top$ the "thin" and the "full" SVD of $\mathbf{A}$. Now we have:

$$\mathrm{tr}(\mathbf{AV}^\top) = \mathrm{tr}(\tilde{\mathbf{L}}\tilde{\mathbf{Q}}\mathbf{R}^\top\mathbf{V}^\top) = \mathrm{tr}(\tilde{\mathbf{Q}}\mathbf{R}^\top\mathbf{V}^\top\tilde{\mathbf{L}}) = \mathrm{tr}(\tilde{\mathbf{Q}}\mathbf{H}) = \sum q_i H_{ii} \le \sum q_i = \mathrm{tr}(\mathbf{Q}).$$

Here $\mathbf{H} = \mathbf{R}^\top\mathbf{V}^\top\tilde{\mathbf{L}}$ is a matrix with orthonormal rows as can be verified directly, and so it must have all its elements not larger than one. It follows that the whole trace is not larger than the sum of singular values of $\mathbf{A}$. Using $\mathbf{V} = \mathbf{LR}^\top$ yields exactly this value of the trace, hence it is the optimum.

Acknowledgements

- This work was funded by the German Ministry of Education and Research (FKZ 01GQ1601), the German Research Foundation (EXC307, BE5601/4-1) and the National Institutes of Health BRAIN Initiative (1U19MH114830-01).
- We thank Andreas Tolias, Rickard Sandberg, Cathryn Cadwell, Jiaolong Xiang and Frederico Scala for discussion, Cathryn Cadwell and Janos Fuzik and their co-authors for making their data available and Shreejoy Tripathy for help with data processing.

- PB and DK conceptualized the project, DK and MW developed statistical methods and wrote the software, PB supervised the project, DK and PB wrote the paper.
- The authors declare that they have no competing financial interests.
- Correspondence and requests for materials should be addressed to P.B. (email: philipp.berens@uni-tuebingen.de).

References

Cathryn R Cadwell, Athanasia Palasantza, Xiaolong Jiang, Philipp Berens, Qiaolin Deng, Marlene Yilmaz, Jacob Reimer, Shan Shen, Matthias Bethge, Kimberley F Tolias, et al. Electrophysiological, transcriptomic and morphologic profiling of single neurons using patch-seq. *Nature Biotechnology*, 34(2):199, 2016.

Cathryn R Cadwell, Federico Scala, Shuang Li, Giulia Livrizzi, Shan Shen, Rickard Sandberg, Xiaolong Jiang, and Andreas S Tolias. Multimodal profiling of single-cell morphology, electrophysiology, and gene expression using patch-seq. *Nature Protocols*, 12(12):2531, 2017.

Lisha Chen and Jianhua Z Huang. Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *Journal of the American Statistical Association*, 107(500): 1533–1545, 2012.

Hyonho Chun and Sündüz Keleş. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1):3–25, 2010.

Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004.

Csaba Földy, Spyros Darmanis, Jason Aoto, Robert C Malenka, Stephen R Quake, and Thomas C Südhof. Single-cell rnaseq reveals cell adhesion molecule profiles in electrophysiologically defined neurons. *Proceedings of the National Academy of Sciences*, 113(35):E5222–E5231, 2016.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1, 2010.

János Fuzik, Amit Zeisel, Zoltán Máté, Daniela Calvigioni, Yuchio Yanagawa, Gábor Szabó, Sten Linnarsson, and Tibor Harkany. Integration of electrophysiological recordings with single-cell rna-seq data identifies neuronal subtypes. *Nature Biotechnology*, 34(2):175, 2016.

John C Gower and Garmt B Dijksterhuis. *Procrustes problems*, volume 30. Oxford University Press on Demand, 2004.

K Harris, C Bengtsson Gonzales, Hannah Hochgerner, N Skene, Lorenza Magno, Linda Katona, Peter Somogyi, Nicoletta Kessaris, Sten Linnarsson, and Jens Hjerling-Leffler. Classes and continua of hippocampal ca1 inhibitory neurons revealed by single-cell transcriptomics. *bioRxiv*, 2017.

Kim-Anh Lê Cao, Debra Rossouw, Christele Robert-Granié, and Philippe Besse. A sparse pls for variable selection when integrating omics data. *Statistical Applications in Genetics and Molecular Biology*, 7(1), 2008.

Ed Lein, Lars E Borm, and Sten Linnarsson. The promise of spatial transcriptomics for neuroscience in the era of molecular cell typing. *Science*, 358(6359):64–69, 2017.

Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.

Richard H Masland. Neuronal cell types. *Current Biology*, 14(13):R497–R500, 2004.

Nicolai Meinshausen. Relaxed lasso. *Computational Statistics & Data Analysis*, 52(1):374–393, 2007.

Jean-Francois Poulin, Bosiljka Tasic, Jens Hjerling-Leffler, Jeffrey M Trimarchi, and Rajeshwar Awatramani. Disentangling neural cell diversity using single-cell transcriptomics. *Nature Neuroscience*, 19(9):1131, 2016.

Karthik Shekhar, Sylvain W Lapan, Irene E Whitney, Nicholas M Tran, Evan Z Macosko, Monika Kowalczyk, Xian Adiconis, Joshua Z Levin, James Nemesh, Melissa Goldman, et al. Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. *Cell*, 166(5): 1308–1323, 2016.

Bosiljka Tasic, Vilas Menon, Thuc Nghi Nguyen, Tae Kyung Kim, Tim Jarsky, Zizhen Yao, Boaz Levi, Lucas T Gray, Staci A Sorensen, Tim Dolbeare, et al. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nature Neuroscience*, 19(2):335, 2016.

Bosiljka Tasic, Zizhen Yao, Kimberly A Smith, Lucas Graybuck, Thuc Nghi Nguyen, Darren Bertagnolli, Jeff Goldy, Emma Garren, Michael N Economo, Sarada Viswanathan, et al. Shared and distinct transcriptomic cell types across neocortical areas. *bioRxiv*, page 229542, 2017.

Shreejoy J Tripathy, Lilah Toker, Brenna Li, Cindy-Lee Crichlow, Dmitry Tebaykin, B Ogan Mancarci, and Paul Pavlidis. Transcriptomic correlates of neuron electrophysiological diversity. *PLoS Computational Biology*, 13(10):e1005814, 2017.

Ines Wilms and Christophe Croux. Sparse canonical correlation analysis from a predictive point of view. *Biometrical Journal*, 57(5):834–851, 2015.

Daniela M Witten, Robert Tibshirani, and Trevor Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10 (3):515–534, 2009.

Amit Zeisel, Ana B Muñoz-Manchado, Simone Codeluppi, Peter Lönnerberg, Gioele La Manno, Anna Juréus, Sueli Marques, Hermany Munguba, Liqun He, Christer Betsholtz, et al. Cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq. *Science*, 347(6226): 1138–1142, 2015.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

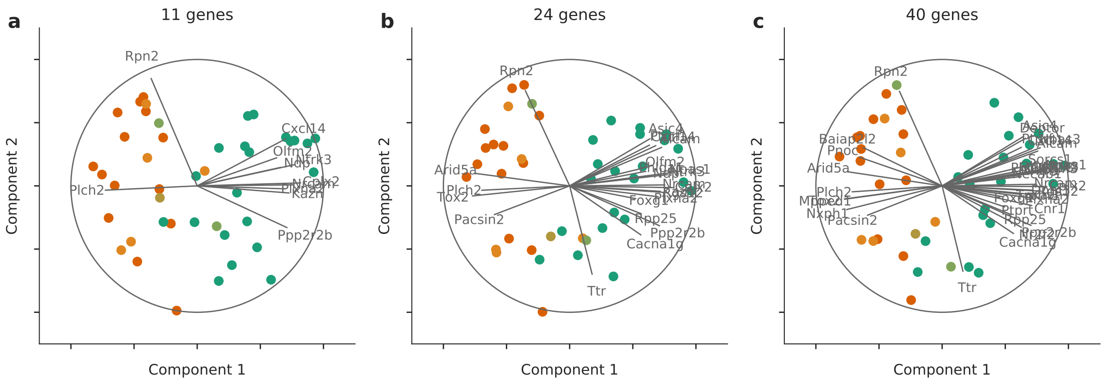

Figure S1: Biplots in the transcriptomic space for the Cadwell et al. data set using three sets of values of regularization parameters. The corresponding biplots in the electrophysiological space were all very similar to the one in Figure 2d and hence not shown. **a.** $\alpha = 9, \lambda = .9$. **b.** $\alpha = .5, \lambda = 1.4$ (identical to the one shown in Figure 2c). **c.** $\alpha = .2, \lambda = 4$.

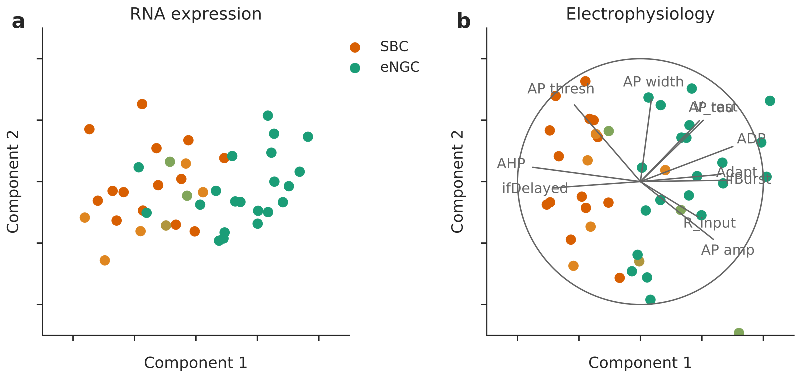

Figure S2: Principal component analysis (PCA) of the Cadwell et al. dataset. **a.** PCA in the transcriptomic space. Dots show single cells (color denotes cell type). Intermediate colors correspond to cells that were not categorized unambiguously. The scatter plot looks squeezed vertically because both PC1 and PC2 are standardized but there is one outlier cell (with respect to PC2) not visible on this plot. **b.** PCA biplot in the electrophysiological space. Each line shows correlations of an electrophysiological property with the first two PCA components. The circle shows maximal possible correlation.

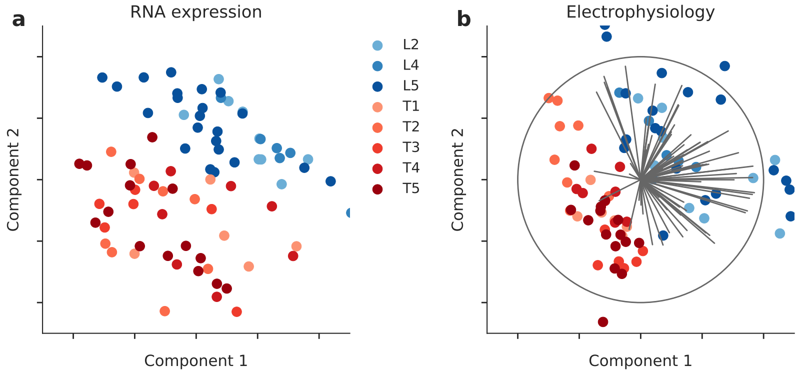

Figure S3: Principal component analysis (PCA) of the Fuzik et al. dataset. **a.** PCA in the transcriptomic space. Dots show single cells (color denotes cell type). **b.** PCA biplot in the electrophysiological space. Each line shows correlations of an electrophysiological property with the first two PCA components. The circle shows maximal possible correlation.

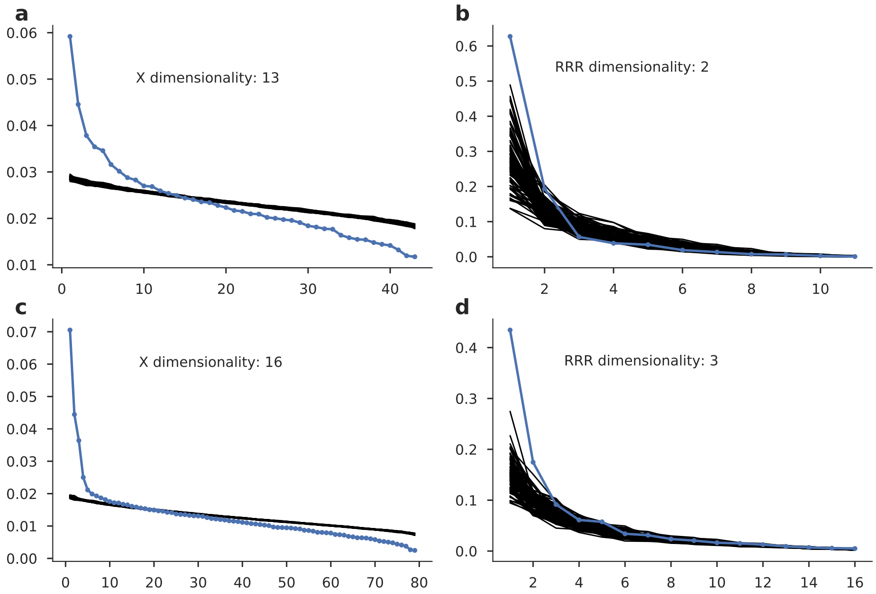

Figure S4: Permutation-based estimation of the transcriptomic dimensionality (left) and RRR rank (right) for the Cadwell et al. (top) and Fuzik et al. (bottom) data sets. Blue lines show actual eigenvalues. Black lines show eigenvalues after 100 shuffles.