Taylor & Francis
Taylor & Francis Group

Journal of Computational and Graphical Statistics

ISSN: 1061-8600 (Print) 1537-2715 (Online) Journal homepage: https://amstat.tandfonline.com/loi/ucgs20

Automatic Transformation and Integration to Improve Visualization and Discovery of Latent Effects in Imaging Data

Gregory J. Hunt, Mark A. Dane, James E. Korkola, Laura M. Heiser & Johann A. Gagnon-Bartsch

To cite this article: Gregory J. Hunt, Mark A. Dane, James E. Korkola, Laura M. Heiser & Johann A. Gagnon-Bartsch (2020) Automatic Transformation and Integration to Improve Visualization and Discovery of Latent Effects in Imaging Data, Journal of Computational and Graphical Statistics, 29:4, 929-941, DOI: 10.1080/10618600.2020.1741379

To link to this article: https://doi.org/10.1080/10618600.2020.1741379

JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2020, VOL. 29, NO. 4, 929–941
https://doi.org/10.1080/10618600.2020.1741379

Taylor & Francis
Taylor & Francis Group

Automatic Transformation and Integration to Improve Visualization and Discovery of Latent Effects in Imaging Data

Gregory J. Hunt[a], Mark A. Dane[b], James E. Korkola[b], Laura M. Heiser[b], and Johann A. Gagnon-Bartsch[c]

[a]Department of Mathematics, College of William & Mary, Williamsburg, VA; [b]Department of Biomedical Engineering, Knight Cancer Institute, OHSU Center for Spatial Systems Biomedicine, Oregon Health and Science University, Portland, OR; [c]Department of Statistics, University of Michigan, Ann Arbor, MI

ABSTRACT

Proper data transformation is an essential part of analysis. Choosing appropriate transformations for variables can enhance visualization, improve efficacy of analytical methods, and increase data interpretability. However, determining appropriate transformations of variables from high-content imaging data poses new challenges. Imaging data produce hundreds of covariates from each of thousands of images in a corpus. Each of these covariates will have a different distribution and needs a potentially different transformation. As such imaging data produce hundreds of covariates, determining an appropriate transformation for each of them is infeasible by hand. In this article, we explore simple, robust, and automatic transformations of high-content image data. A central application of our work is to microenvironment microarray bio-imaging data from the NIH LINCS program. We show that our robust transformations enhance visualization and improve the discovery of substantively relevant latent effects. These transformations enhance analysis of image features individually and also improve data integration approaches when combining together multiple features. We anticipate that the advantages of this work will likely also be realized in the analysis of data from other high-content and highly multiplexed technologies like Cell Painting or Cyclic Immunofluorescence. Software and further analysis can be found at *gjhunt.github.io/rr*. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received May 2019
Revised October 2019

KEYWORDS
Automatic transformation; Data integration; Imaging; Latent variables; PCA; Visualization

1. Introduction

Transformation of data is an essential component in many areas of analysis. Consider principal components analysis (PCA), one of the primary statistical techniques for visualization and recovery of latent variables. PCA is well-known to be sensitive to skewed distributions and outliers (Hubert, Rousseeuw, and Verdonck 2009; Maadooliat, Huang, and Hu 2015). Using PCA in such cases can lead to results that are unduly influenced by arbitrary data scales and often describe only a few particular outlying points. Thus, an active area of research is methods for making PCA robust to such problems (Locantore et al. 1999; Hubert, Rousseeuw, and Verboven 2002; Higuchi and Jp 2004; Croux and Ruiz-Gazen 2005; Maronna 2005). One approach to dealing with data scale and outliers is data transformation (Hu, Wright, and Zou 2006; Huang, Shen, and Buja 2008; Zimmerman and Nunez-Anton 2010; Maadooliat, Huang, and Hu 2015). However, choosing the correct transformation is highly application specific and typically entails substantial domain-specific knowledge (Maadooliat, Huang, and Hu 2015). While recommended transformations are established in some fields, determining an appropriate transformation is often itself a substantial question to answer. In this article, we tackle this problem by exploring automatic transformation and integration of features as part of a PCA data analysis pipeline. One important area in which automatic transformation of data is necessary is high-content imaging data (Caicedo et al. 2017). Such data consist of hundreds of features extracted from thousands of images using automatic feature detection software. While classically there is only one data matrix to transform, for high-content image data there will be hundreds of matrices to consider: one for each image feature. An appropriate transformation will need to be chosen for each of these feature matrices. In this article, we consider simple and robust ways of adaptively doing this. We show that transforming data can improve visualization and discovery of substantive latent effects in features individually and when integrating together multiple features. Consequently, we explore the interaction of transformation with integrating multiple features to extract a common set of latent variables. As part of all these analyses we employ methods to estimate principal components (PCs) from data with missing values.

2. Motivating Application: MEMAs

The motivating application of this work is data from the high-content bio-technology called the microenvironment microarray (MEMA; Labarge, Parvin, and Lorens 2014; Lin et al. 2017; Watson et al. 2018; Smith et al. 2019). MEMAs aid exploration of cellular microenvironment: the cells' immediate physical and bio-chemical surroundings. This microenvironment is important as it is implicated in many cell and tissue level processes, diseases, and dysfunctions (Lin, Lee, and LaBarge 2012; LaBarge

CONTACT Gregory J. Hunt *ghunt@wm.edu* Department of Mathematics, College of William & Mary, P.O. Box 8795, Williamsburg, VA 23187-8795.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JCGS*.

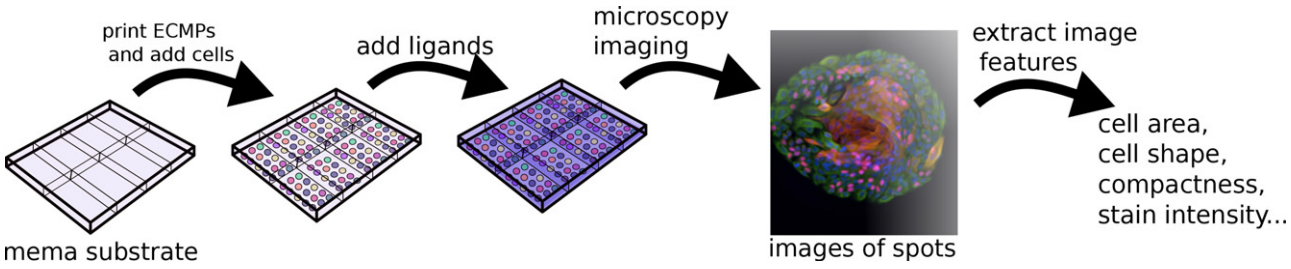


Figure 1. MEMAs: a plastic substrate partitioned into wells. Each well is an array of hundreds of ∼400 μm printed "spots." Added cells randomly bind to spots and interact with: (1) an extra-cellular matrix protein (ECMp) specific to the spot and (2) a ligand specific to a well. After growing, cells are immunofluorescently stained, imaged, and quantified cellular features extracted.

2013; Januschke and Näthke 2014; Pelissier et al. 2014; Maman and Witz 2018). Better understanding of the microenvironment benefits basic research and furthers an understanding of the interaction between therapeutic agents and regulatory behavior (Teti 1992; Bissell and Labarge 2005; LaBarge, Petersen, and Bissell 2007). Understanding cellular microenvironments has been a long-term research aim of the NIH and consequently is a major component of the Library of Integrated Network-Based Cellular Signatures (LINCS) NIH Common Fund program.

MEMAs facilitate the study of combinations of microenvironmental factors on molecular and biological endpoints via high-throughput image-based profiling of cells. After growing the cells on the MEMA substrate under various conditions, the cells are stained and imaged with microscopy (Figure 1). The images are then analyzed with software to extract quantitative cellular features. There is one feature matrix for each cellular feature. For example, there will be a feature matrix recording cell size for several hundred MEMA plates (the observations, rows) across several hundred microenvironmental conditions (the variables, columns). The features (and feature matrices) cover a wide range of cellular aspects. There are morphological features like cell area, compactness, eccentricity, perimeter, or solidity. The MEMAs also produce stain intensity features and features capturing the cell cycle state, cell lineage, cell count, texture, and many others. Typically, this amounts to several hundred features however the features are not only diverse but may vary from one experiment to the next depending on research interests, software utilized, and underlying biology.

While this plethora of feature data presents new opportunities for discovery, it also necessitates an adaptive approach that can handle a disparate and changing landscape of features. Determining a good transformation for MEMA data is more complicated than traditional -omics experiments due to the number, flexibility, and disparate nature of the features. Consider a nuclei orientation feature (approximately normally distributed) and a DAPI intensity feature (highly right-skewed and always positive, see Supplementary Figures 1 and 2). The appropriate transformation for these two features will likely not be the same since they have very different distributions. We might want to log-transform the intensity feature while leaving orientation feature alone. With this in mind, this article studies methods for automatically and adaptively choosing appropriate data transformations for any image feature produced by a high-throughput image-analysis platform like a MEMA. Our successful application on MEMAs points to promising applications of this methodology to other image-based cell-profiling technologies like Cyclic Immunofluorescence and Cell Painting (Bray et al. 2016; Lin et al. 2016; Tsujikawa et al. 2017).

3. Methods

In this section, we discuss our transformation approach (Section 3.1) and how it helps visualization and discovery of substantive latent effects (Section 3.2). We will also outline calculating the SVD in the presence of missing data (Section 3.3), and how we recover latent effects through data integration (Section 3.4).

3.1. Robust Rescaling

To process the image-feature matrices we follow three sequential transformation steps:

Procedure 1 Three-step robust rescaling (RR)

Step 1: (G) robustly "Gaussianize" the data,
Step 2: (Z) convert the data to robust z-scores,
Step 3: (O) remove outliers.

These three steps are applied to each of the feature matrices individually.

3.1.1. The Gaussianizing Step (G)

The (G), or "Gaussianizing," step transforms the data using a robust Box–Cox-like procedure (Box and Cox 1964). The traditional Box–Cox procedure for finding an optimal transformation is not well suited for this data because it is not robust. Indeed, central to the Box–Cox estimator is a rescaled sample variance. As the breakdown point of the sample variance is exactly zero, we expect the breakdown point of the Box–Cox estimator to also be near zero (Cook and Wang 1983; Atkinson 1986; He, Simpson, and Portnoy 1990; Marazzi and Yohai 2004). A breakdown point near zero indicates that even one outlier can lead to a drastically over-fit Box–Cox transformation. To overcome this, the (G) step uses an interpretable and robust "divide-and-summarize" approach to improve stability and avoid over-fitting. The procedure first divides the data column-wise and estimates a Gaussianizing (Box–Cox) transformation for each column of the feature matrix. To avoid over-fitting, the procedure then robustly summarizes the transformation parameters, choosing the median among the column-wise estimates. This median transformation is then used to transform the origi-

nal feature matrix. This model-averaging approach is similar to attribute bagging seen in supervised learning applications (Bryll, Gutierrez-Osuna, and Quek 2003). However, because all attributes (columns) are used in (G), as opposed to a random subspace procedure, the estimates we calculate are interpretable. If we consider a model like that of Carroll (1980) where a certain percentage of the data is arbitrarily corrupted, then by using the consensus median estimate of the transformation parameter, up to 50% of the columns of the data can be arbitrarily corrupted without arbitrarily corrupting the estimate of the transformation.

Consider MEMA data where each column is a different microenvironment being studied. Our final transformation is a transformation that works well for the typical (median) microenvironment and ignores any aberrant microenvironments. This allows (G) to avoid undue influence of technical problems. We will see in Section 4 that this avoids technical problems of the NID and ELN microenvironments. Furthermore, the extreme transformations proposed by these two microenvironments allows us to flag them for further quality-control scrutiny.

Let $Y \in \mathbb{R}^{M\times N}$ be a specific feature matrix and let $\mathcal{T}_1, \ldots, \mathcal{T}_Q$ be a collection of Q parameterized transformation families so that for any $q = 1, \ldots, Q$ the family $\mathcal{T}_q = \{T_\lambda^{(q)} \mid \lambda \in \Lambda^{(q)}\}$ consists of differentiable, monotonic, transformations $T_\lambda^{(q)} : S \to \mathbb{R}$ on some $S \subseteq \mathbb{R}$. The goal is to optimize over the union of these families and choose the transformation that makes the data close to being normal (without over-fitting).

In our application (Section 4), we will choose $Q = 2$ families over which to search: a power family $\mathcal{T}_1 = \mathcal{T}_{\text{power}} = \{(\text{sign}(y)|y|^\lambda - 1)/\lambda, \lambda \in \mathbb{R}\}$ and an arc-hyperbolic-sine family $\mathcal{T}_2 = \mathcal{T}_{\text{ahs}} = \{\text{asinh}(\lambda y)/\lambda, \lambda \geq 0\}$. We choose these two families because they cover a range of power and sigmoidal shapes. While the family of power transformation is versatile, it cannot deal elegantly with negative values in the data. Thus, we have included the arc-hyperbolic-since family because it is a well-studied family of transformations that can deal with negative values (Nowicka et al. 2019). Many reasonable choices of parameterized families can be made and nothing in our discussion depends on the specific choices. We include other options of families in our software.

Before we describe the procedure for optimizing over many families, we will first consider the simpler case when $Q = 1$ and discuss how to choose an optimal transformation over a single family generically denoted $\mathcal{T} = \{T_\lambda \mid \lambda \in \Lambda\}$. Define Y_{*j} be the jth column of Y, and for any $\lambda \in \Lambda$ let $Y_{*j}(\lambda) = T_\lambda\left(Y_{*j}\right)$ be Y_{*j} under the transform T_λ. The goal is to choose a $\widehat{\lambda}$ so that $Y_{*j}\left(\widehat{\lambda}\right)$ is approximately normally distributed for each $j = 1, \ldots, N$. The (G) approach follows two steps:

Procedure 2 (G) Transformation estimation for one family ($Q = 1$)

Step 1: (divide, estimate) determine the optimal $\widehat{\lambda}_j$ for each column Y_{*j} so that $Y_{*j}\left(\widehat{\lambda}_j\right)$ is as normal as possible,
Step 2: (summarize) set $\widehat{\lambda} = \text{median}_j \widehat{\lambda}_j$

For the first step, estimate $\widehat{\lambda}_j$ using the traditional Box–Cox approach on Y_{*j}. Assume there is some λ_j so that $Y_{ij}(\lambda_j) \overset{\text{iid}}{\sim} N(\mu_j, \sigma_j^2)$ for $\mu_j \in \mathbb{R}$ and $\sigma_j \geq 0$. Then let $\widehat{\lambda}_j$ be the MLE of λ_j. This is obtained by profiling the likelihood over μ_j and σ_j^2 and then maximizing the profile likelihood over λ_j. If L_j is the profile likelihood of λ_j profiling over μ_j and σ_j^2 then $\widehat{\lambda}_j \overset{\text{def}}{=} \arg\max_{\lambda_j \in \Lambda} L_j(\lambda_j)$.

After estimating $\widehat{\lambda}_j$ for each column, the second step is to summarize the collection of $\widehat{\lambda}_j$'s into a single $\widehat{\lambda}$. This is done with the median. Define $\widehat{\lambda}$ as the element-wise median $\widehat{\lambda} = \text{median}_j \widehat{\lambda}_j$.

Concisely, the procedure when $Q = 1$ is to first divide and optimize within each column and then median-summarize across the column-wise estimates. When $Q > 1$ we add an additional step to first determine which family among $\mathcal{T}_1, \ldots, \mathcal{T}_Q$ is best. The procedure is described in Procedure 3.

Procedure 3 (G) Transformation estimation for multiple families ($Q > 1$)

Step 1: determine which family is the best over-all, call it $\widehat{q}$
Step 2: estimate λ using just the optimal family $\mathcal{T}_{\widehat{q}}$ (following the previous procedure of Procedure 2 for $Q = 1$).

This procedure first determines the best family individually for each column and then uses the family that is best among a plurality of columns. More specifically, let $L_j(q_j, \lambda_j)$ be the likelihood of the jth column after transformation using the q_jth family and transformation parameter λ_j. Optimize $L_j(q_j, \lambda_j)$ jointly over λ_j and q_j and let $(\widehat{q}_j, \widetilde{\lambda}_j) = \arg\max_{\substack{q_j=1,\ldots,Q \\ \lambda_j \in \Lambda^{(q)}}} L_j(q_j, \lambda_j)$ and $\widehat{q} = \text{mode}_j\, \widehat{q}_j$ so that $\widehat{q}$ is the family that is the best among a plurality of the columns. Once we have determined this optimal family $\widehat{q}$ we then estimate $\widehat{\lambda}$ following the procedure when $Q = 1$ using the family $\mathcal{T}_{\widehat{q}}$. Finally, define the Gaussianized version of Y as $G(Y) \overset{\text{def}}{=} T_{\widehat{\lambda}}^{(\widehat{q})}(Y)$.

3.1.2. The z-Score Step (Z)

The second step in the (RR) procedure is the (Z) step, a robust z-score transformation. Let $\widetilde{Y}$ be a vectorized version of Y and $\widetilde{Y}_{(q)}$ be the q-winsorized version of $\widetilde{Y}$. In this article, we will use $q = 0.001$ replacing everything below the qth quantile of $\widetilde{Y}$ by the qth quantile and replacing everything above the $(1-q)$th quantile of $\widetilde{Y}$ by the $(1-q)$th quantile. Given this, the robust z-score version of Y is defined as $Z(Y) \overset{\text{def}}{=} (Y - \widehat{\mu})/\widehat{\sigma}$ where $\widehat{\mu}$ and $\widehat{\sigma}$ are mean and SD estimates of $\widetilde{Y}_{(q)}$. Notice that the final values $Z(Y)$ have not been winsorized themselves, the winsorization has only been used in the calculation of $\hat{\mu}$ and $\hat{\sigma}$.

3.1.3. The Outlier Removal Step (O)

The final of the three (RR) steps is outlier removal. The outlier removal procedure simply thresholds z-scores and marks as missing anything beyond four standard deviations. First let $Z(Y)$ be the robust z-scored version of Y. We then define $O(Y)_{ij} = Y_{ij}$ if $|Z(Y)_{ij}| \leq z$, and $O(Y)_{ij} =$ "NA," otherwise. Here, "NA" denotes a missing value. To be conservative in this article, we use $z = 4$ although this is somewhat arbitrary. With

$z = 4$ if the data are truly normal this removes only about 3e−3 percent of the data from each tail.

3.1.4. The Three-Step (RR) Procedure

Given these definitions, the three step (RR) transformation is to apply the (G), (Z), and then (O) transformations. If Y is a feature matrix then we define $RR(Y)$ as $RR(Y) = O(Z(G(Y)))$.

3.2. Transformations and Latent Effects

A central component in data analysis is the identification of important latent effects both visually and quantitatively. For MEMAs, we divide latent effects into two categories: (1) biological effects and (2) technical effects. Biological effects include, for example, differences in biological endpoints due to ECMps or ligands. Technical effects are unwanted and we are interested in identifying them so that we may remove them. Examples include batch across plates or spatial effects within wells. Discovery of latent effects is typically done through visual inspection of plots or quantitative analysis like PCA. Unfortunately, methods like PCA are often misled by prominent aspects of the data are unrelated to substantive latent effects. As an example, consider how PCA can be misled when used to identify groups in skewed data. Let $u^{(1)}, u^{(2)}, v^{(1)}, v^{(2)} \in \mathbb{R}^N$ have elements that are iid from a standard log-normal distribution. For a small $\delta \in \mathbb{R}$ and noise $\epsilon \in \mathbb{R}^{2N \times N}$ define a block data matrix Y as $Y = \begin{bmatrix} A \\ B \end{bmatrix} + \epsilon$ where $A = u^{(1)} v^{(1)\prime}$ and $B = \delta + u^{(2)} v^{(2)\prime}$ so that the first N rows of the data matrix and the last N rows of the data matrix constitute two groups with a mean difference of δ. The left side of Figure 2 displays a histogram of the elements of Y for a simulation using $\delta = 1/2$ and ϵ distributed iid standard normal. Visually, it is difficult to distinguish between the two groups in the left-hand panel of Figure 2 because the group difference is over-shadowed by the data's long tails. Consequently, PCA identifies the variance due to tail skewness, not the group difference, as the most prominent variation in the data. While the first two PCs capture more than 99% of the total variance in this example, they only capture about 50% of the group difference (Supplementary Figure 3). The right side of Figure 2 shows that a log transformation makes the groups more prominent through visual inspection. The transformation also helps quantitative analyses. The transformation un-skews the data thereby attenuating the effect of the tails on PCA. In this case, while the first two PCs only capture about 80% of the total variation, they capture about 94% of the variation due to group difference (Supplementary Figure 3).

Motivated by the previous example, we want to attenuate the influence of prominent, yet uninformative, variation when visualizing data and when applying quantitative methods. The (RR) transformation does this by ameliorating the effects of two commonly encountered, and potentially misleading, aspects of data: (1) skewness in measurement scales and (2) anomalous outliers. By anomalous outliers we mean extremely unusual data points that are not informative of much beyond their own uniqueness. For example, cells may have difficulty growing on a spot, or software might produce image-analysis artifacts like segmentation anomalies.

To guard against nonsubstantive variation (RR) applies three robust rescaling steps. (G) robustly prevents a feature's naturally long-tailed measurement scale from dominating analysis by de-skewing each feature's distribution. While the traditional Box–Cox procedure is likely to over-fit in the presence of genuine outliers, (G) will not. Indeed, traditional Box–Cox will propose an extreme transformation to rectify even a single outlier. (G) only transforms the data to un-skew fundamentally skewed data, not simply to reign in a few points. Instead, outliers are handled more parsimoniously by a procedure that specifically targets them. To remove outliers (RR) first converts the data to robust z-scores using (Z) and then removes any entry of the feature matrix bigger in magnitude than four with (O).

3.3. Complete Singular Vectors

A common statistical tool used to recover latent effects is the singular value decomposition (SVD). For a data matrix Y with a singular value decomposition (SVD) of $Y = U\Sigma V'$ we call the columns of the U left singular vectors and the columns of V the right singular vectors. If one were to mean-center the columns of Y, these singular vectors define the PCs. We avoid this term because we do not mean-center.

When calculating the SVD for image-based data we often need to account for missing data. Missing values arise for domain-specific reasons (e.g., cells failed to grow on a MEMA), image-analysis reasons (e.g., the software could not detect any features), and because (RR) introduces missing values as part of (O). As the SVD is undefined for matrices with missing values,

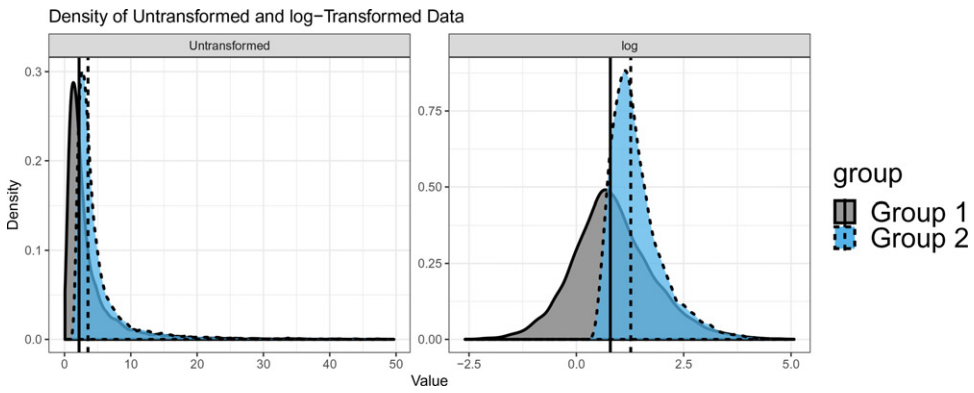


Figure 2. Skewed data (left) before a transformation and (right) after a log-transformation. The transformation makes the group difference more prominent.

we use "complete" singular vectors calculated from rescaled pairwise-complete gram matrices. This is similar to pairwise-complete covariance matrices (e.g., `cor` in R).

Let $Y \in \mathbb{R}^{M\times N}$ be a feature matrix with missing values. Define Y_0 as Y with missing values replaced by zeros and $Y_{\mathbb{1}}$ so that $(Y_{\mathbb{1}})_{ij} = \mathbb{1}\{Y_{ij} \text{ is not missing}\}$. We define the rescaled pairwise-complete left Gram matrix $Y \cdot Y'$ so that

$$(Y \cdot Y')_{ij} = \frac{N}{n_{ij}} \sum_{k=1}^{N} (Y_0)_{ik}(Y_0)_{jk},$$

where $n_{ij} = (Y_{\mathbb{1}} Y'_{\mathbb{1}})_{ij}$ is the number of pairwise-complete entries between row i and j of Y.

$Y \cdot Y'$ is a matrix of rescaled inner products of the rows of Y accounting for the number of nonmissing pairs between rows. We similarly define the right gram matrix $Y' \cdot Y$ replacing Y for Y' above. We call the eigenvectors of $Y \cdot Y'$ and $Y' \cdot Y$ the complete left/right singular vectors. WLOG they are ordered decreasing by eigenvalue and we keep only those with positive eigenvalues. For brevity we henceforth omit the adjective "complete," referring to these simply as the "singular vectors." If there are no missing values they are identical. While other methods exist for calculating the SVD on matrices with missing values (e.g., Hastie et al. 2015), because the total number of missing values is small for MEMA data, our simple procedure will allow quick and intuitive calculations of singular vectors that are highly commensurate existing approaches.

3.4. Average Singular Vectors

In addition to recovering important latent effects in individual features, we are interested in latent effects common to multiple features. To extract a common set of latent effects from a collection of P features $Y^{(1)}, \ldots, Y^{(P)}$ we use the eigenvectors from the average of their rescaled pairwise-complete left and right gram matrices, respectively,

$$\frac{1}{P}\sum_{p=1}^{P} Y^{(p)} \cdot Y^{(p)\prime} \quad \text{and} \quad \frac{1}{P}\sum_{p=1}^{P} Y^{(p)\prime} \cdot Y^{(p)}.$$

We call these eigenvectors the left and right average singular vectors (ASVs).

4. Application to MEMA Data

4.1. Structure of MEMA Data

We work with MEMA data from the MEP-LINCS Center at the Oregon Health and Science University. The data are accessible through Synapse with identifiers syn10155286, syn10155289, and syn10155292 (Gray, Heiser, and Korkola 2014) available at *http://www.synapse.org/MEP_LINCS*. In total, we analyze 24 MEMAs of human epithelial mammary tissue (MCF10A). The 24 MEMAs come in three batches of eight plates. Each MEMA plate is divided evenly into eight wells. Each well contains 700 spots in a 20 by 35 grid. Cells are added to the wells and bind to the spots. Subsequently, a buffer solution containing a specific ligand is added to each well. Thus, the cells can grow out in the presence of different ECM proteins and ligands. The pattern of ECMps is identical across all wells (see Supplementary Figure 4) however a (potentially) different ligand is added to each well (see Supplementary Figure 5). After incubating the cells for 72 hr they are fluorescently stained, imaged, and cell-level features are extracted with image analysis software. For the analysis in this article, we work with spot-level features (median summarized cell-level features). For each image feature, we have a data matrix of 192 wells (3 batches × 8 plates × 8 wells) by 694 spots (we remove 6 alignment spots with no cells from the 700). In total, we will work with 103 image features (Supplementary Table 1).

4.2. Features and Transformations Considered

The MEMA plates we analyze are grown, stained, and imaged in three separate processing batches. A different set of stains is used in each batch. Those sets are (1) "SS1" (containing stains DAPI, Actin, CellMask and MitoTracker), (2) "SS2noH3" (containing stains DAPI, Fibrillarin and EdU), and (3) "SS3" (containing stains DAPI, KRT5, KRT19, and CellMask). Because each of these batches use a different staining set we refer to them as the "staining batches." While these batches are separate experiments, aside from the staining set the experimental conditions were made as identical as possible.

In total, there are 103 different image features extracted from the MEMAs. A different set of features is extracted in each staining batch with some being common across multiple batches. There are 50 features extracted in at least two of the staining batches and 18 features that are extracted from all three. We focus on four features in this article: (1) cell area (notated on synapse as "Cells_CP_AreaShape_Area"), (2) cell compactness ("Cells_CP_AreaShape_Compactness"), (3) spot cell count ("Spot_PA_SpotCellCount"), and (4) total cytoplasm DAPI intensity ("Cytoplasm_CP_Intensity_IntegratedIntensity_Dapi"). We choose these features because they represent several different feature types. The first two are morphological traits of cells, the third is the cell count, and the last is an intensity. We have deposited the full results for all features at *gjhunt.github.io/rr*. To explore the effects of (G), (Z), and (O), we consider five transformations of the features: (1) no transformation (NT), (2) the (G) step only, (3) the (Z) step only, (4) the (O) step only, and (5) the three-step (RR) transformation.

4.3. Visualization

4.3.1. Feature Distributions

A typical first step in exploratory analysis is data visualization. Simple data visualizations can succinctly summarize the major features of the data and inform qualitative analyses. In Figure 3, we plot the distribution of cell area for the five transformations. The densities correspond to staining batches. The bold line is the density of all data combined. Notice in Figure 3 that the density of (NT) largely reflects the data's long tail. The same can be said for (Z). Conversely, the other transformations reveal the staining batches. Both (O) and (G) de-emphasize the data's long tail in favor of the group difference. Furthermore, in (RR) these groups are approximately Gaussian. Supplementary Figures 6–9 show the similar plots for other features and effects.

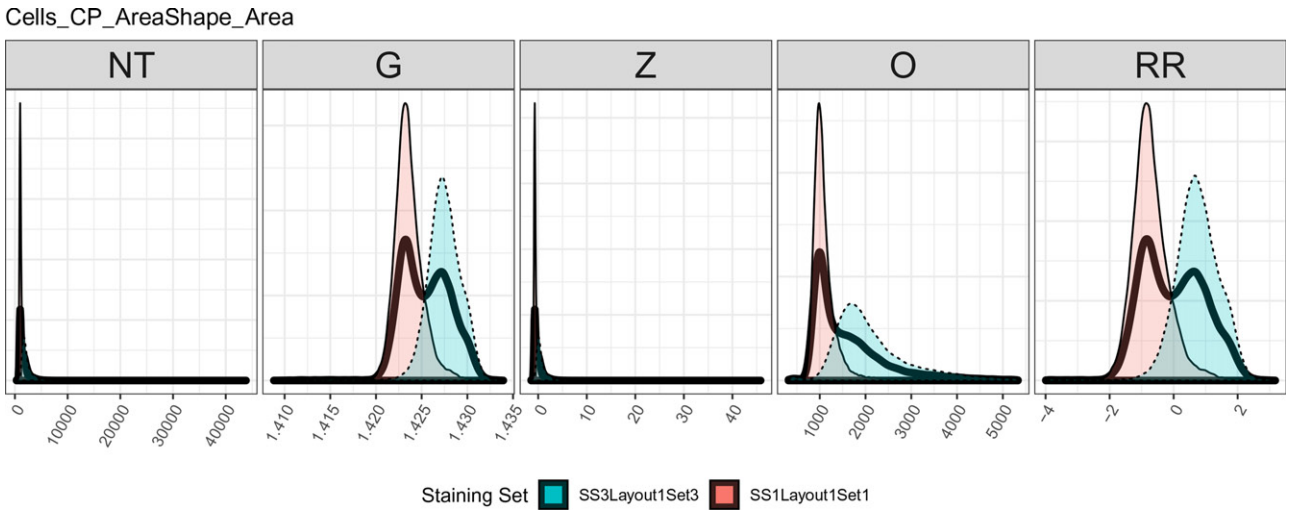


Figure 3. Density of elements of cell area feature matrix. Bold density is all elements combined. Other densities are the densities for the two staining batches. Subplots are for five processing transformations.

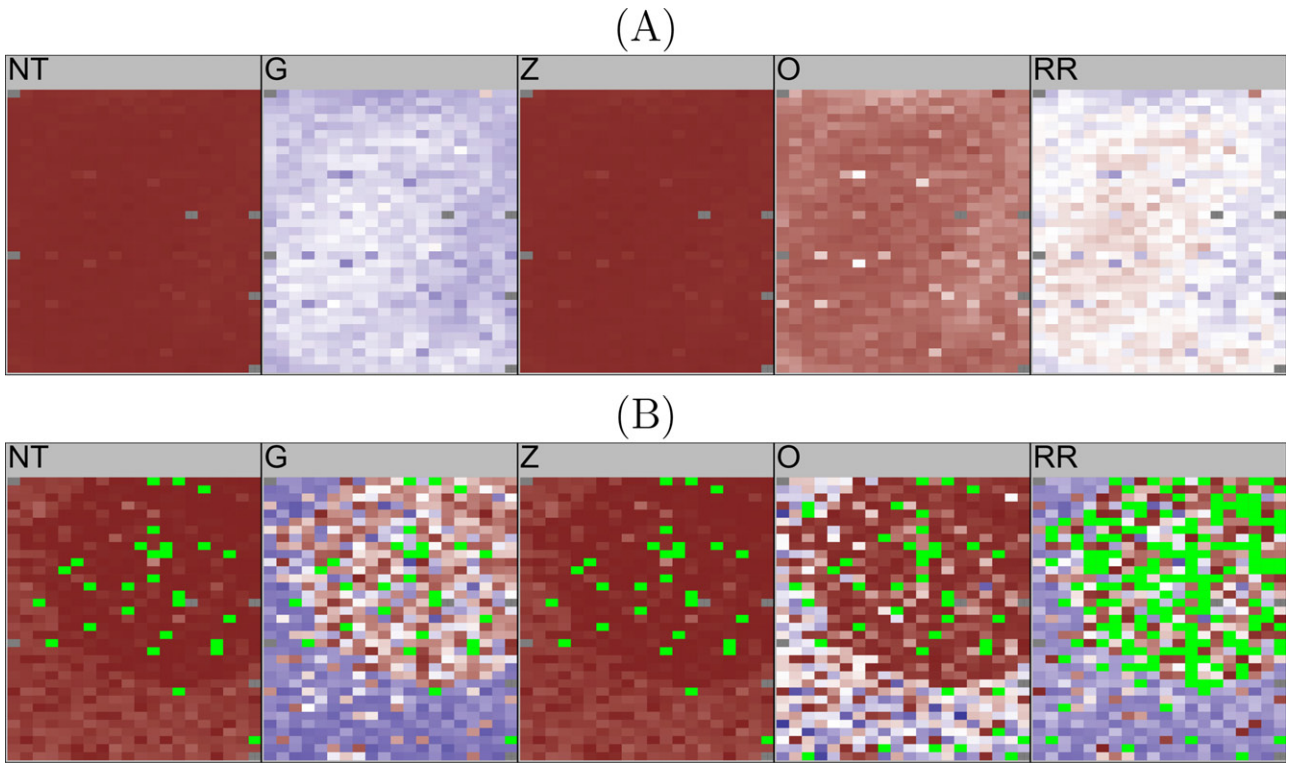


Figure 4. Heat map of cell area for two wells (a) and (b) across the five transformations (NT), (G), (Z), (O), (RR). Subplots of Supplementary Figure 10. Color scaled is determined globally over all spots, wells, and plates in the dataset to reflect the fact that the transformation is similarly calculated over this data. Thus, we see no blue in this (NT) subplot as we see almost no blue in Supplementary Figure 10. This plot is a representative microcosm of the larger plot. Green indicates missing data.

4.3.2. Heat-Maps

Another way to visualize the MEMA data is through heat-map pseudo-images. These pseudo-images are heat-maps of the value of a feature for each spot plotted following the same physical layout as the MEMA. These pseudo-images can be useful for discovering spatial effects and assessing the quality of data. As an example, we visualize cell area this way in Supplementary Figure 10. In Figure 4(a), we display a single well from Supplementary Figure 10 across the five transformations. The colors are more blue if they are close to the minimum cell area, red if they are close to the maximum, and white if they are half-way between. Dark gray spots are omitted according to the MEMA design. Note that the color scale is determined globally over all wells and all plates in the dataset (Supplementary Figure 10).

This figure is not very informative for (NT) or (Z). The skewness and outliers assign the bulk of data to a tiny range of colors meaning the plots are essentially a single color. Conversely, for (G) and (O) we see a spatial effect between the right and edges and the rest of the well. We also see a nonspatial effect where certain spots are much different than their surroundings. We circle these spots in orange in Supplementary

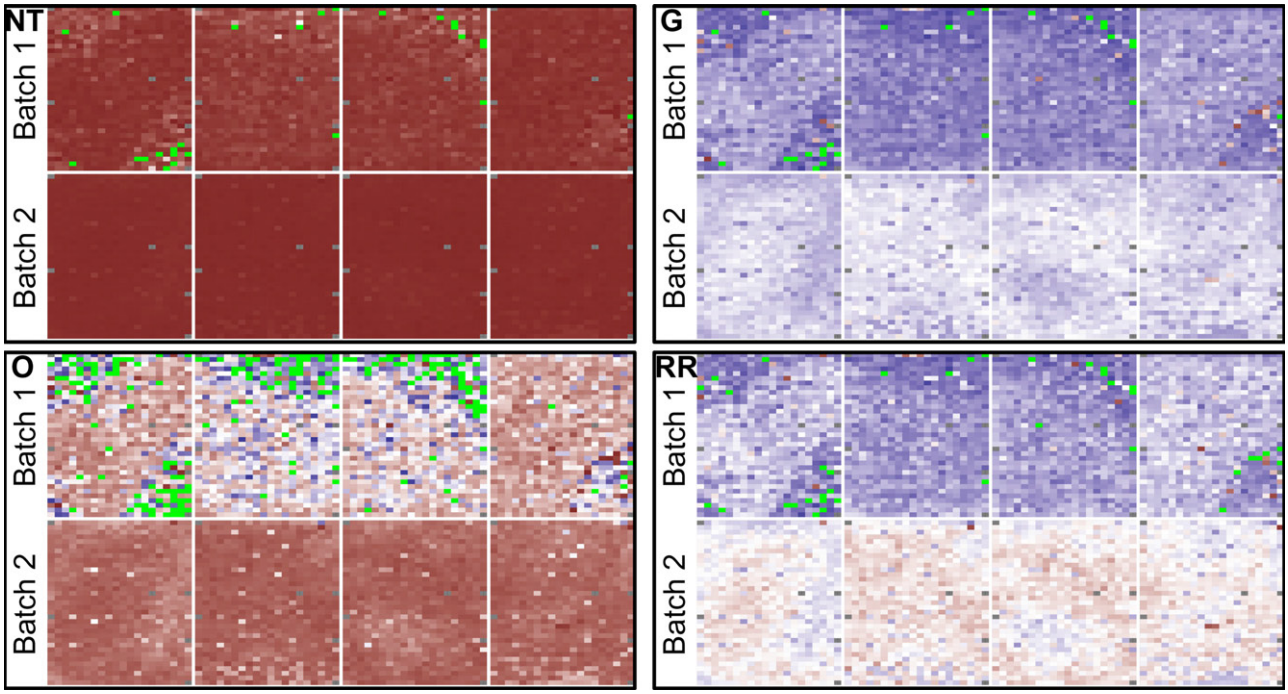


Figure 5. Heat map of eight wells across the five transformations (NT), (G), (Z), (O), (RR). Top row of each subplot is from first staining batch. Bottom row is from second staining batch. Colors are more blue if they are close to the minimum, red if they are close to the maximum, and white if they are close to half-way between. Green spots are missing. Dark gray spots are omitted according to the MEMA design.

Figure 11. In Section 4.6, we show that this is an effect of the ECMps NID1 and ELN. We can see from these plots that (RR) strongly highlights the spatial effects as well as the NID/ELN effect.

In Figure 4(b), we focus on a different well of Supplementary Figure 10. Here, the green spots indicate missing data. These spots are missing either due to experimental error or because they have been removed as part of analysis. We see similar behavior where (G) and (O) reveal spatial differences between the upper right and the rest of the well. This spatial effect is also seen in (RR) however the number of points removed is much different in (RR) compared to (O). This highlights the difference between (O) thresholding outliers without transformation and (RR) thresholding outliers after (G). We believe thresholding based on a z-score makes most sense on a Gaussianized scale (as in (RR)). Note also that outliers are defined in a global context of the entire data, so while many values are marked as outliers by (RR) in this particular well it is a small percentage of the entire data.

In addition to highlighting spatial effects, these transformations also reveal batch effects between plates, wells, and staining batches. In Figure 5, we display the heat-map pseudo-image of cell area for eight wells across (NT), (G), (O) and (RR). ((Z) is identical to (NT).) The top four wells in each subplot are from the first staining batch, the bottom four wells are from the second. Nonetheless, we see little indication of batch in (NT). However, batch is visible in (G), (O), and (RR). The bottom of (G) is lighter blue than the top, and the top of (O) is lighter red than the bottom. In (RR), we have solid-blue in the top and mostly red in the bottom. Being better able to identify batch effects hopefully will aid down-stream procedures to account for such effects.

4.4. Recovering Technical Effects Across Wells

Batch effects are a common and well-studied problem in high-throughput biological experiments like MEMAs (Leek et al. 2010). Often, such batch effects obscure biological variation of interest. To deal with this problem, unwanted variation, like batch, is typically identified using the SVD and projected out of the data. In this section, we explore how (RR) helps identify unwanted variation like batch using the SVD. We focus on the large staining batch effect as it was visible by eye in Figure 5.

We assess the transformations by measuring the percentage of the batch captured by the first k singular vectors of the transformed feature matrix. Let $U = [u_1, \ldots, u_N] \in \mathbb{R}^{M\times N}$ be the (complete) left singular vectors of a feature matrix and $B \in \mathbb{R}^{M\times D}$ be the batch indicator matrix so that $B_{ij} = 1$ if well i is in batch j for $j = 1, \ldots, D$. Here we have $D = 3$ for the three staining batches. For $k = 1, \ldots, N$ and $t = 1 \ldots \min(k, D)$ define $C_k^{(t)}$ to be the tth canonical correlation between the first k left singular vectors $U_k = [u_1, \ldots, u_k]$ and the batch B. Then let $C_k^2 = \frac{1}{D}\sum_{t=1}^{\min(k,D)} \left(C_k^{(t)}\right)^2$ to be the average of these squared canonical correlations. We can interpret C_k^2 as the percentage of the batch B that is captured by these first k singular vectors. In Figure 6(a), we plot C_k^2 on the y-axis and vary k across the x-axis from $k = 1$ to 192. From this figure, we see that the transformations enhance identification of the staining batch. Consider the cell area and total DAPI intensity features. As compared with no transformation (NT), these plots show that (G), (O), and (RR) increase how much of the staining batch is captured by the first several singular vectors. These transformations attenuate the non-informative tails of the distributions and focus the singular vectors on the differences across the staining batches.

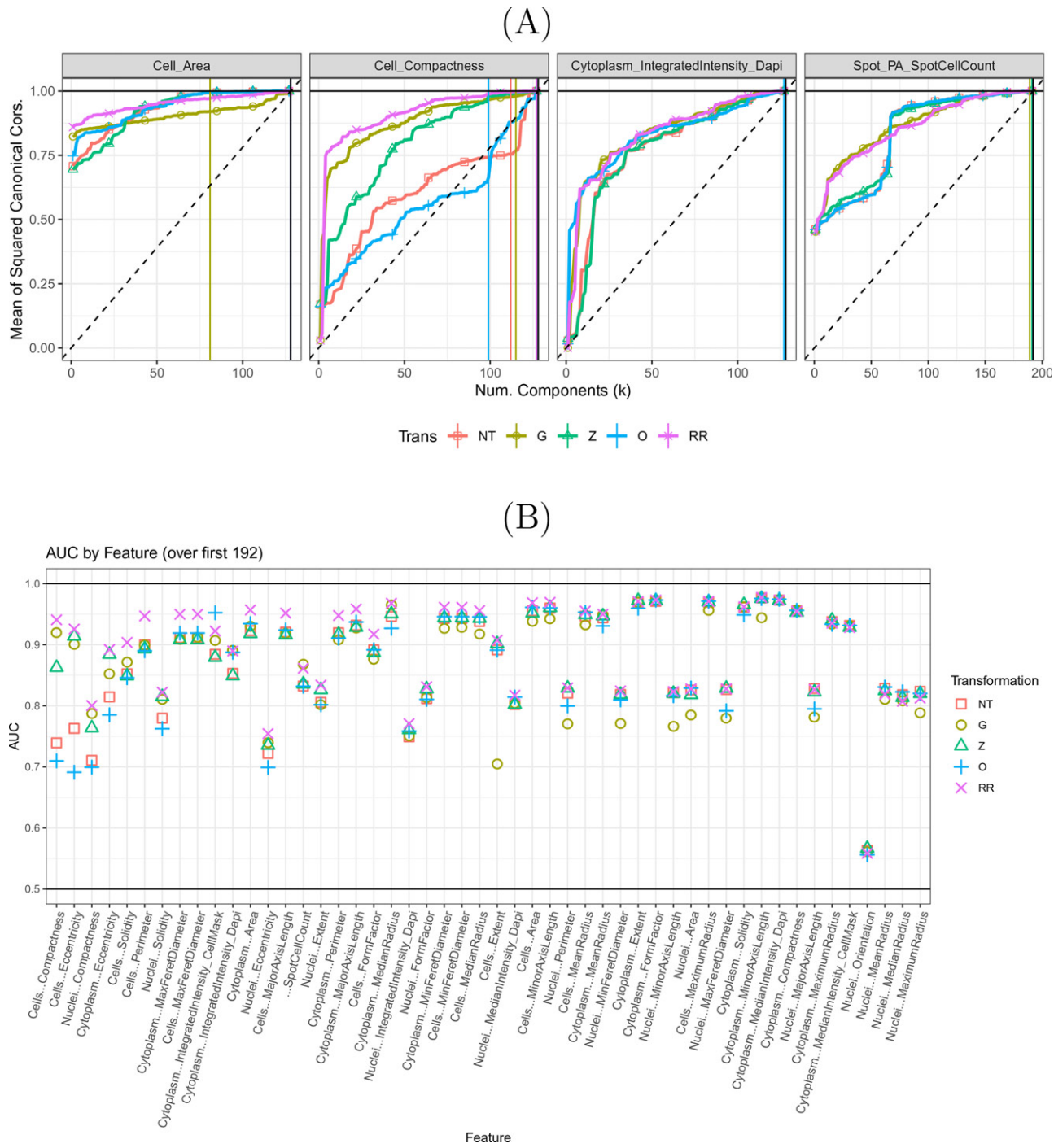


Figure 6. (a) Mean of the squared canonical correlations between the first k left singular vectors and the staining batch dummy variables. (b) Grand mean of the squared canonical correlations across number of components. Features ordered left to right decreasing by the difference in the AUC between (RR) and (NT). Thus, those on the left (RR) perform relatively better than (NT).

We summarize batch recovery for all features in Figure 6(b). Here, we calculate the area under the CC curves (AUC) for each feature as AUC $= \sum_{k=1}^{192} C_k^2$. Broadly, we see the same behavior in Figure 6(b) as displayed in Figure 6(a). (RR) seems to generally improve recovery of the staining batch. Sometimes we see a substantial improvement (e.g., Cell Compactness) and rarely do we see that (RR) is detrimental. In Supplementary Figures 15–20, we display similar plots for the recovery of plate, well, and ligand effects.

4.5. Data Integration for Discovering Between-Well Effects

Given the close relationship among many of the MEMA image features, latent effects that appear in one feature may show up in other features. Shared effects give insight into biological and technical aspects that are important across many features. To extract these common effects we integrate information across MEMA features using the left average singular vectors (ASVs) as described in Section 2.

In the left panel of Figure 7, we plot the mean squared canonical correlations (C_k^2) between the first k left ASVs and the stain-

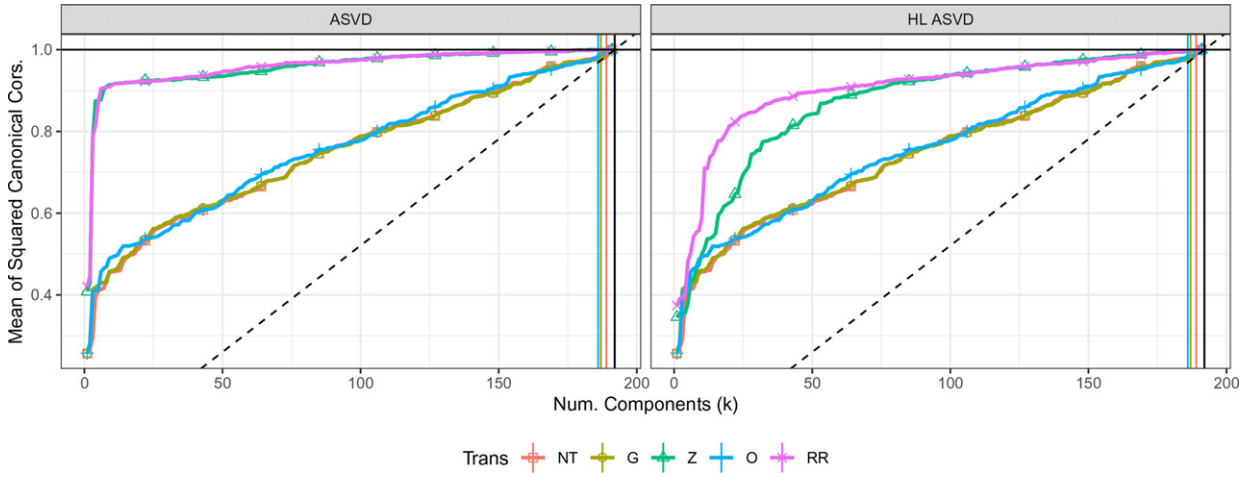
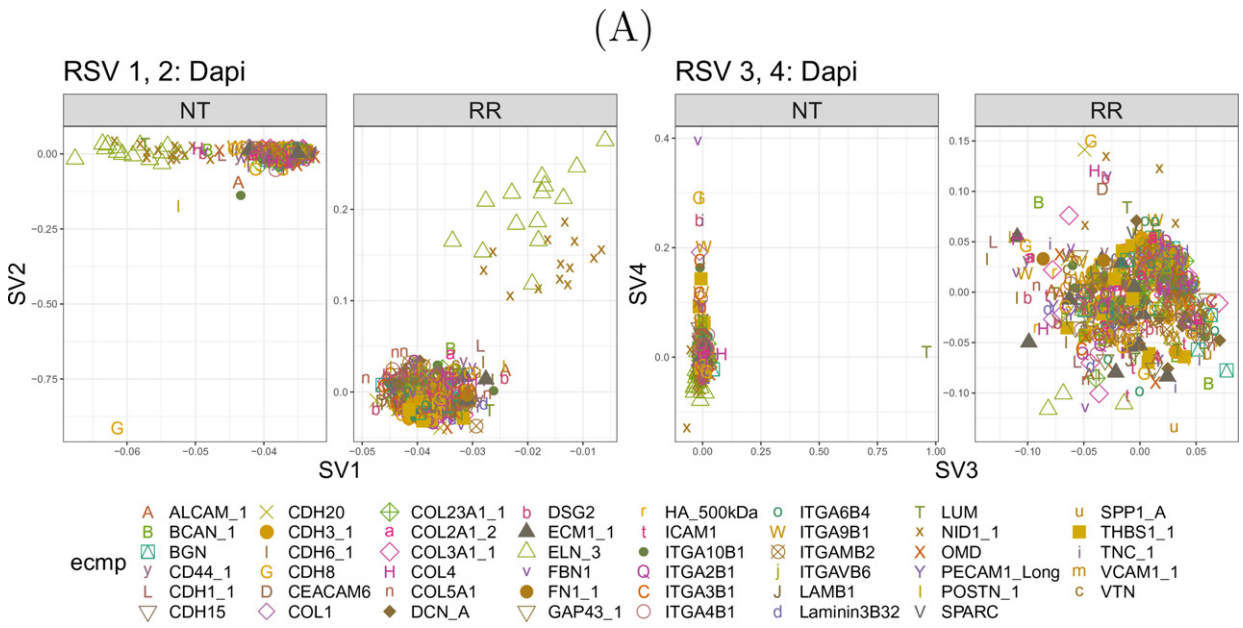


Figure 7. Mean of the squared canonical correlations between the first k average left singular vectors and the staining batch dummy variables. The average left singular vectors come from integration of (left) the 18 features that are measured across all MEMAs, and, (right) the five with the highest leverage points (among those 18).



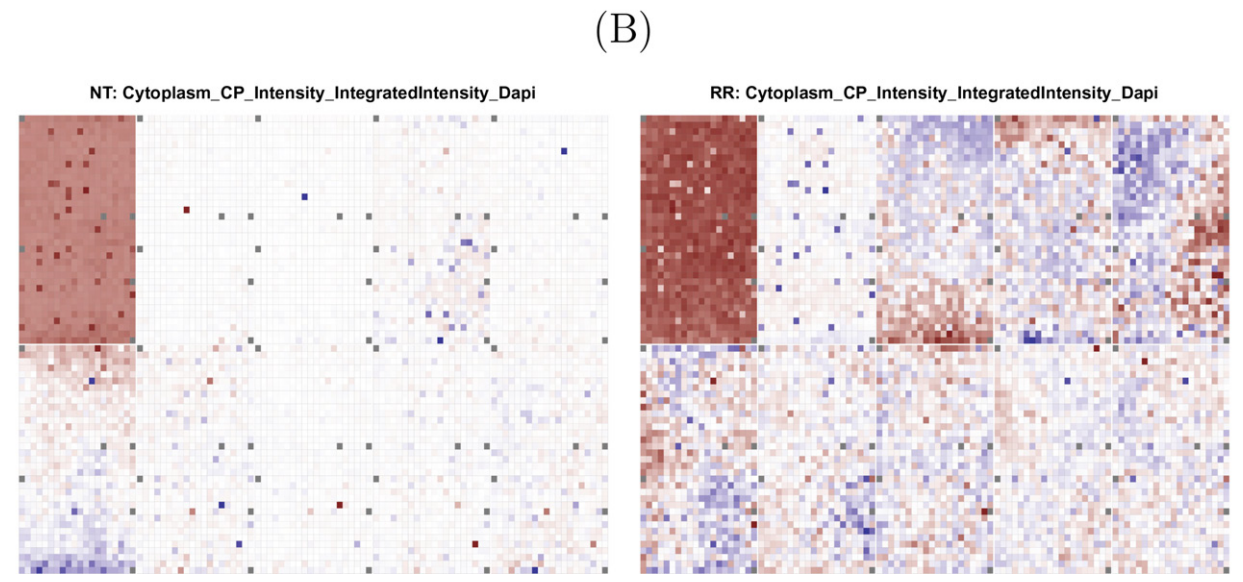


Figure 8. (a) Scatterplot of elements of first four right singular vectors against each other for the total cytoplasmic DAPI intensity feature. Shape and color indicate ECMp of the spot corresponding to the elements of the singular vector. (b) Heat map of elements of top 10 right singular vectors for the total cytoplasmic DAPI intensity feature.

ing batch. We calculate the ASVs using the 18 features measured in every plate. From this figure, we can see that (Z) and (RR) quickly and strongly recover the staining batch. The AUC for these curves is in excess of 0.95, meaning it recovers batch better than the majority of individual features. The ASVs "average-out" feature-specific effects and amplify common effects like staining batch.

It is notable that the (Z) and (RR) recover batch significantly better than (O), (G), and (NT). This happens because the ASVs element-wise average Gram matrices across features. If these Gram matrices are on vastly different scales their average is biased toward the largest features. This arbitrarily weights features' by their scales. To equitably integrate information all features should have values in a similar range as in (Z) and (RR). Thus, they recover batch better.

Finally, it is notable how well (Z) does alone. This happens because the averaging used to compute the ASVs conveys some of the same benefits as (G) and (O). This is true so long as we do not have a small number of features or systematic skewness or outliers across features. In the right panel of Figure 7, we calculate the ASVs using only five features with several high-leverage points. Here, we see a separation between (Z) and (RR) since the average is over a small number of highly skewed features. In any case, including (G) and (O) steps does not seem to hurt the analysis and thus we still recommend the full three-step (RR) transformation for integrating features in this manner. Similar, but attenuated results for plate, well and ligand are shown in Supplementary Figures 21–23.

4.6. Discovering Biological and Spatial Effects Within Wells

The left singular vectors of the feature matrices reveal latent effects across the wells, plates, and staining batches. Similarly, the right singular vectors (RSVs) reveal effects across the spots. In Figure 8(a), we display a scatterplot of the first four RSVs of total cytoplasmic DAPI intensity for (NT) and (RR).

A prominent feature of Figure 8(a) is the separation between the ECMps ELN, NID1, and the rest. Upon further investigation of the underlying MEMA images we find that this effect manifests because the cells have difficulty adhering to the ELN and NID1 ECMp substrates. Notice the cell count heat-map in Supplementary Figure 13 shows that the cell count in the ELN and NID1 spots are significantly lower than other spots. While this ELN-NID1 effect is present in the untransformed data, it is more prominent in (RR). The first RSV from the untransformed data does capture the effect; however, the second through fourth RSVs are focused on explaining several outliers. Moreover, (RR) separates NID1 and ELN from the other ECMps and from each other.

In Figure 8(b), we plot pseudo-image heat-maps of the first ten RSVs for cytoplasmic DAPI intensity arranging elements of the RSVs according to the MEMA plate spatial layout. In addition to the ELN/NID effect, these plots reveal common spatial patterns across wells. These patterns are more visible for (RR) than (NT) as the RSVs of (NT) mostly capture outliers. It is important to identify such unwanted effects so that we can properly account for them downstream. In Supplementary Figures 25–29, we display similar scatterplots and heat-maps for the other example features. They tell similar stories.

To see what biological effects can be found if we a priori remove the dominating ELN-NID effect, we reanalyze the MEMA data after removing these spots. Now we find an effect separating THBS from the other ECMps. This is particularly prominent in morphological features. As an example, in Figure 9 we plot a scatterplot of the top four RSVs for cell compactness. (RR) reveals a difference between THBS and the other ECMps. This effect shows up in many of the morphological features but not cell count. Thus, this THBS

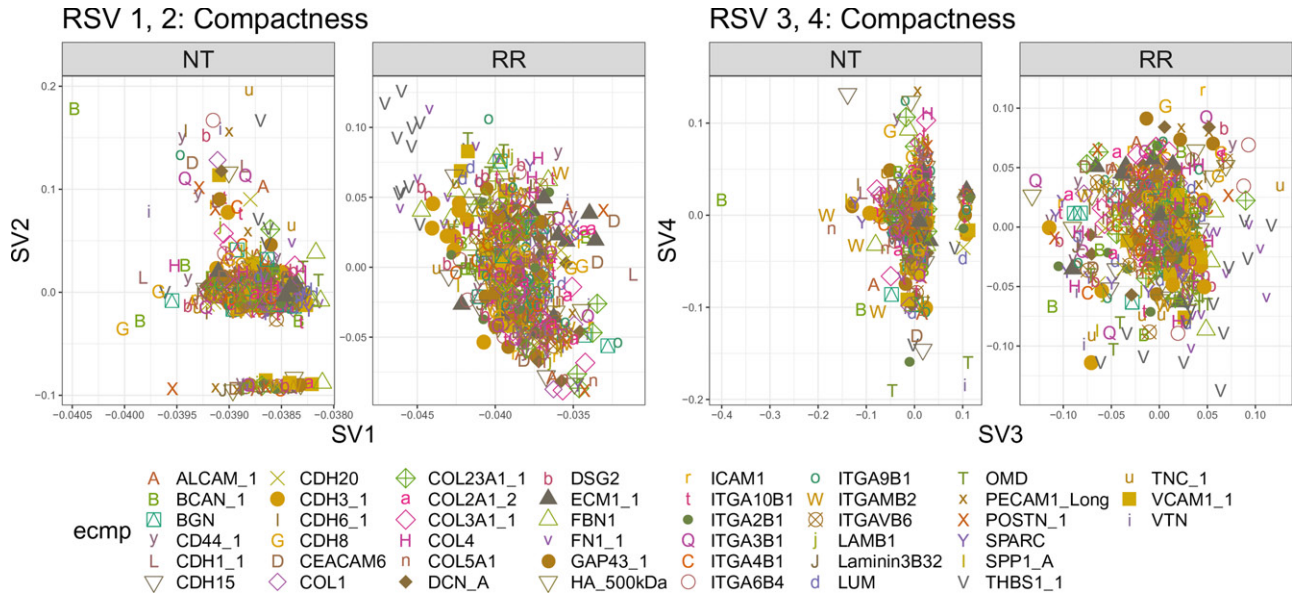


Figure 9. Scatterplot (after removing ELN and NID from analysis) of elements of top four right singular vectors against each other for the cell compactness feature. Shape and color indicate ECMp of the spot corresponding to the elements of the singular vector. The clusters seen in the NT panels are from missing spots on an outlier plate (see Supplementary Figure 31).

effect does not appear to be of similar origin to the ELN-NID effect. Instead, it appears to be a biological effect on cell morphology.

4.7. Data Integration for Discovering Within-Well Effects

In Section 4.5, we saw that data integration helped make salient important between-well effects. In a similar fashion, the average right singular vectors (ASVs) help bring out within-well effects. In Figure 10(a), we plot the first two right ASVs against each other. Again, (RR) equitably integrates information from all features and helps highlight the NID/ELN effect. Finally, we display heat-maps pseudo-images for the first ten right ASVs in Figure 10(b). The right ASVs for (NT) seem to be mostly picking up a couple of outliers. On the other hand, (RR) strongly picks up several interesting spatial effects.

5. Discussion

In this article, we have explored the effects of several transformations as part of a preprocessing pipeline of high-content image data, in particular, data from MEMAs. The goal of these transformations is to emphasize important latent effects in the data and attenuate common and misleading aspects.

Untransformed feature data is often encumbered by skewed measurement scales, outliers, or both. These aspects can hinder discovery of substantive latent effects. To de-emphasize such misleading aspects of the data (O), (G), and their combination

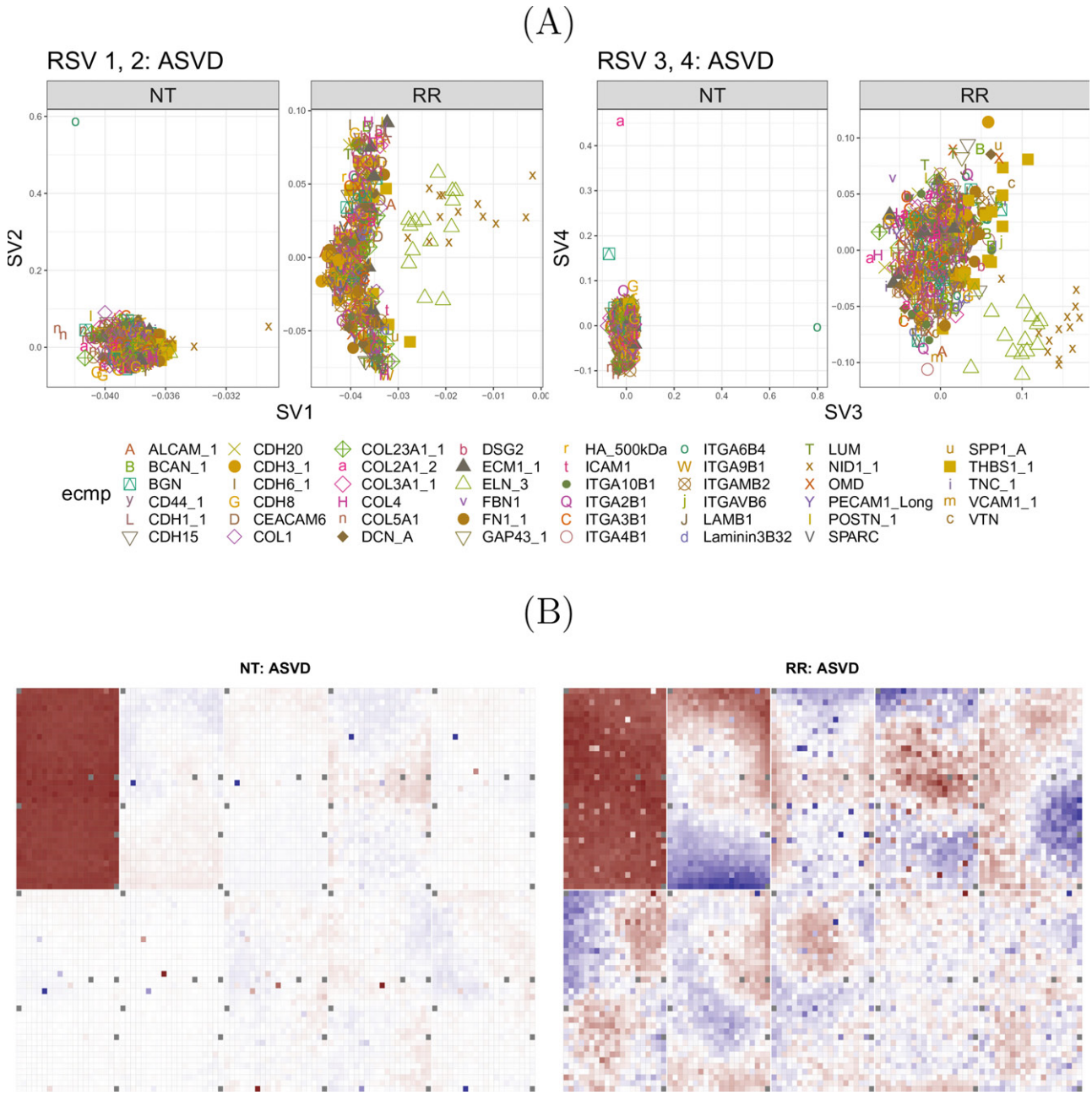


Figure 10. (a) Scatterplot of elements of top four right ASVs calculated over 18 features measured on all MEMAs. Shape and color indicate ECMp of the spot corresponding to the elements of the singular vector. (b) Heat-map of top 10 right ASVs calculated over 18 features measured on all MEMAs.

in (RR) were helpful. (O) removed outliers using a conservative threshold and (G) reduced skewness by Gaussianizing the data. Additionally, (RR) included a (Z) step that converted values to robust z-scores. (Z) and (RR) allowed features to be straightforwardly integrated with a simple arithmetic average of Gram matrices. In the analysis of MEMA data, we showed that a combination of a Gaussianizing transformation (G), z-score transformation (Z), and removal of outliers (O) can improve visualization and discovery of biological and technical latent effects in both features individually and when combining features together. Finally, as (RR) automatically chose transformations for each feature this allowed adaptive application of (RR) to a data containing many different features. This adaptive ability makes (RR) a promising candidate for data generated by other image-based technologies like Cyclic Immunofluorescence (CycIF) or Cell Painting (Bray et al. 2016; Lin et al. 2016; Tsujikawa et al. 2017). For example, CycIF, through a series of imaging and washing steps, allows up to 30-channel immunofluorescent imaging and thus extraction of potentially several hundred features. Like MEMAs, each of these features will need a transformation to be adaptively chosen. Exploring the application of (RR) to other high-content and highly multiplexed technologies is a direction we hope to explore in future work.

Supplementary Materials

The supplementary materials contain figures for a wider range of features and conditions.

Funding

The authors gratefully acknowledge support from the National Science Foundation (grant no. DMS-1646108) and the National Institutes of Health (NIH grant nos. U54HG008100 and 1U54CA209988).

References

Atkinson, A. C. (1986), "Diagnostic Tests for Transformations," *Technometrics*, 28, 29–37. [930]

Bissell, M. J., and Labarge, M. A. (2005), "Context, Tissue Plasticity, and Cancer: Are Tumor Stem Cells Also Regulated by the Microenvironment?," *Cancer Cell*, 7, 17–23. [930]

Box, G. E. P., and Cox, D. R. (1964), "An Analysis of Transformations," *Journal of the Royal Statistical Society*, Series B, 26, 211–252. [930]

Bray, M.-A., Singh, S., Han, H., Davis, C. T., Borgeson, B., Hartland, C., Kost-Alimova, M., Gustafsdottir, S. M., Gibson, C. C., and Carpenter, A. E. (2016), "Cell Painting, a High-Content Image-Based Assay for Morphological Profiling Using Multiplexed Fluorescent Dyes," *Nature Protocols*, 11, 1757–1774. [930,940]

Bryll, R., Gutierrez-Osuna, R., and Quek, F. (2003), "Attribute Bagging: Improving Accuracy of Classifier Ensembles by Using Random Feature Subsets," *Pattern Recognition*, 36, 1291–1302. [931]

Caicedo, J. C., Cooper, S., Heigwer, F., Warchal, S., Qiu, P., Molnar, C., Vasilevich, A. S., Barry, J. D., Bansal, H. S., Kraus, O., Wawer, M., Paavolainen, L., Herrmann, M. D., Rohban, M., Hung, J., Hennig, H., Concannon, J., Smith, I., Clemons, P. A., Singh, S., Rees, P., Horvath, P., Linington, R. G., and Carpenter, A. E. (2017), "Data-Analysis Strategies for Image-Based Cell Profiling," *Nature Methods*, 14, 849–863. [929]

Carroll, R. (1980), "A Robust Method for Testing Transformations to Achieve Approximate Normality," *Journal of the Royal Statistical Society*, Series B, 42, 71–78. [931]

Cook, R. D., and Wang, P. C. (1983), "Transformations and Influential Cases in Regression," *Technometrics*, 25, 337–343. [930]

Croux, C., and Ruiz-Gazen, A. (2005), "High Breakdown Estimators for Principal Components: The Projection-Pursuit Approach Revisited," *Journal of Multivariate Analysis*, 95, 206–226. [929]

Gray, J., Heiser, L., and Korkola, J. (2014), "Microenvironment Perturbagen (MEP) LINCS, Sage Bionetworks." www.synapse.org/#! [933]

Hastie, T., Mazumder, R., Lee, J. D., and Zadeh, R. (2015), "Matrix Completion and Low-Rank SVD via Fast Alternating Least Squares," *Journal of Machine Learning Research*, 16, 3367–3402. [933]

He, X., Simpson, D. G., and Portnoy, S. L. (1990), "Breakdown Robustness of Tests," *Journal of the American Statistical Association*, 85, 446. [930]

Higuchi, I., and Jp, E. A. (2004), "Robust Principal Component Analysis With Adaptive Selection for Tuning Parameters Shinto Eguchi," Technical Report. [929]

Hu, J., Wright, F. A., and Zou, F. (2006), "Estimation of Expression Indexes for Oligonucleotide Arrays Using the Singular Value Decomposition," *Journal of the American Statistical Association*, 101, 41–50. [929]

Huang, J. Z., Shen, H., and Buja, A. (2008), "Functional Principal Components Analysis via Penalized Rank One Approximation," *Electronic Journal of Statistics*, 2, 678–695. [929]

Hubert, M., Rousseeuw, P. J., and Verboven, S. (2002), "A Fast Method for Robust Principal Components With Applications to Chemometrics," *Chemometrics and Intelligent Laboratory Systems*, 60, 101–111. [929]

Hubert, M., Rousseeuw, P., and Verdonck, T. (2009), "Robust PCA for Skewed Data and its Outlier Map," Technical Report. [929]

Januschke, J., and Näthke, I. (2014), "Stem Cell Decisions: A Twist of Fate or a Niche Market?," *Seminars in Cell & Developmental Biology*, 34, 116–123. [930]

LaBarge, M. (2013), "Breaking the Canon: Indirect Regulation of Wnt Signaling in Mammary Stem Cells by MMP3," *Cell Stem Cell*, 13, 259–260. [930]

LaBarge, M. A., Parvin, B., and Lorens, J. B. (2014), "Molecular Deconstruction, Detection, and Computational Prediction of Microenvironment-Modulated Cellular Responses to Cancer Therapeutics," *Advanced Drug Delivery Reviews*, 69–70, 123–131. [929]

LaBarge, M. A., Petersen, O. W., and Bissell, M. J. (2007), "Of Microenvironments and Mammary Stem Cells," *Stem Cell Reviews*, 3, 137–146. [930]

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K., and Irizarry, R. A. (2010), "Tackling the Widespread and Critical Impact of Batch Effects in High-Throughput Data," *Nature Reviews Genetics*, 11, 733–739. [935]

Lin, C.-H., Jokela, T., Gray, J., and LaBarge, M. A. (2017), "Combinatorial Microenvironments Impose a Continuum of Cellular Responses to a Single Pathway-Targeted Anti-Cancer Compound," *Cell Reports*, 21, 533–545. [929]

Lin, C.-H., Lee, J. K., and LaBarge, M. A. (2012), "Fabrication and Use of MicroEnvironment microArrays (MEArrays)," *Journal of Visualized Experiments*, 68, 1–7. [929]

Lin, J.-R., Fallahi-Sichani, M., Chen, J.-Y., and Sorger, P. K. (2016), "Cyclic Immunofluorescence (CycIF), a Highly Multiplexed Method for Single-cell Imaging," *Current Protocols in Chemical Biology*, 8, 251–264. [930,940]

Locantore, N., Marron, J. S., Simpson, D. G., Tripoli, N., Zhang, J. T., Cohen, K. L., Boente, G., Fraiman, R., Brumback, B., Croux, C., Fan, J., Kneip, A., Marden, J. I., Peña, D., Prieto, J., Ramsay, J. O., Valderrama, M. J., Aguilera, A. M., Locantore, N., Marron, J. S., Simpson, D. G., Tripoli, N., Zhang, J. T., and Cohen, K. L. (1999), "Robust Principal Component Analysis for Functional Data," *Test*, 8, 1–73. [929]

Maadooliat, M., Huang, J. Z., and Hu, J. (2015), "Integrating Data Transformation in Principal Components Analysis," *Journal of Computational and Graphical Statistics*, 24, 84–103. [929]

Maman, S., and Witz, I. P. (2018), "A History of Exploring Cancer in Context," *Nature Reviews Cancer*, 18, 359–376. [930]

Marazzi, A., and Yohai, V. J. (2004), "Robust Box-Cox Transformations for Simple Regression," in *Theory and Applications of Recent Robust Methods*, eds. A. Struyf, S. Van Aelst, & M. Hubert, Basel: Birkhäuser Basel, pp. 173–182. [930]

Maronna, R. (2005), "Principal Components and Orthogonal Regression Based on Robust Scales," *Technometrics*, 47, 264–273. [929]

Nowicka, M., Krieg, C., Crowell, H. L., Weber, L. M., Hartmann, F. J., Guglietta, S., Becher, B., Levesque, M. P., and Robinson, M. D. (2019),

"CyTOF Workflow: Differential Discovery in High-Throughput High-Dimensional Cytometry Datasets," *F1000Research*, 6, 748. [931]

Pelissier, F. A., Garbe, J. C., Ananthanarayanan, B., Miyano, M., Lin, C. H., Jokela, T., Kumar, S., Stampfer, M. R., Lorens, J. B., and LaBarge, M. A. (2014), "Age-Related Dysfunction in Mechanotransduction Impairs Differentiation of Human Mammary Epithelial Progenitors," *Cell Reports*, 7, 1926–1939. [930]

Smith, R., Devlin, K., Kilburn, D., Gross, S., Sudar, D., Bucher, E., Nederlof, M., Dane, M., Gray, J. W., Heiser, L., and Korkola, J. E. (2019), "Using Microarrays to Interrogate Microenvironmental Impact on Cellular Phenotypes in Cancer," *Journal of Visualized Experiments*, 147, e58957. [929]

Teti, A. (1992), "Regulation of Cellular Functions by Extracellular Matrix," Technical Report. [930]

Tsujikawa, T., Kumar, S., Borkar, R. N., Azimi, V., Thibault, G., Chang, Y. H., Balter, A., Kawashima, R., Choe, G., Sauer, D., El Rassi, E., Clayburgh, D. R., Kulesz-Martin, M. F., Lutz, E. R., Zheng, L., Jaffee, E. M., Leyshock, P., Margolin, A. A., Mori, M., Gray, J. W., Flint, P. W., and Coussens, L. M. (2017), "Quantitative Multiplex Immunohistochemistry Reveals Myeloid-Inflamed Tumor-Immune Complexity Associated With Poor Prognosis," *Cell Reports*, 19, 203–217. [930,940]

Watson, S. S., Dane, M., Chin, K., Jonas, O., Gray, J. W., and Korkola, J. E. (2018), "Microenvironment-Mediated Mechanisms of Resistance to HER2 Inhibitors Differ Between HER2+ Breast Cancer Subtypes," *Cell Systems*, 6, 329–342.e6. [929]

Zimmerman, D. L., and Nunez-Anton, V. A. (2010), *Antedependence Models for Longitudinal Data*, Boca Raton, FL: Chapman & Hall/CRC. [929]

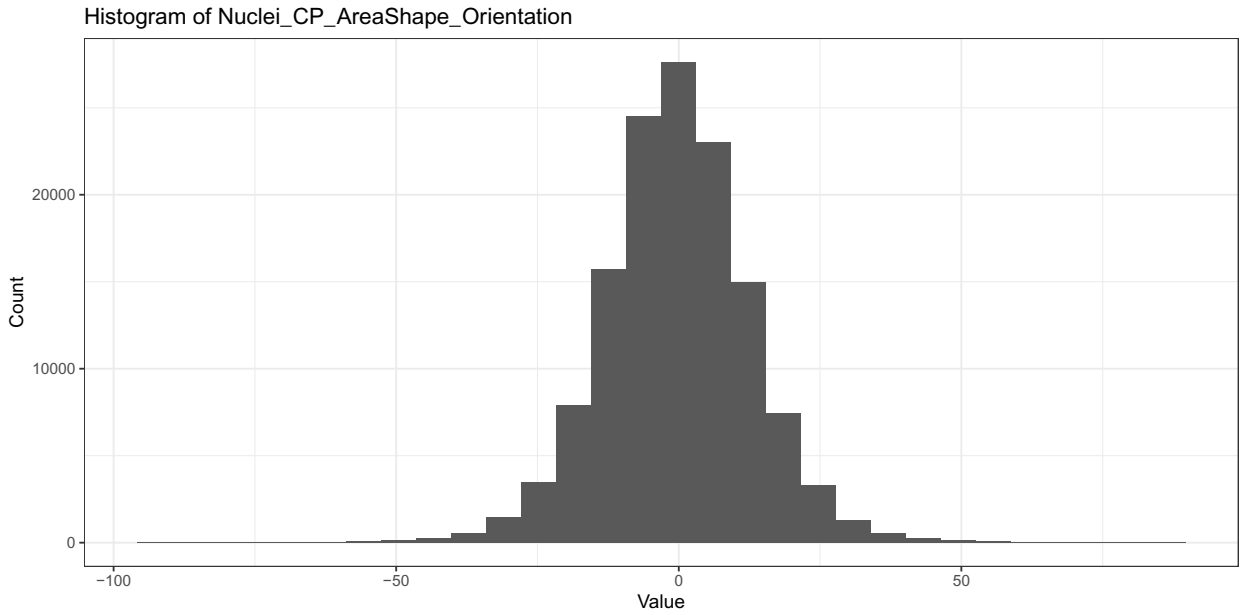


Figure 1: Histogram of nuclei orientation across all wells and spots.

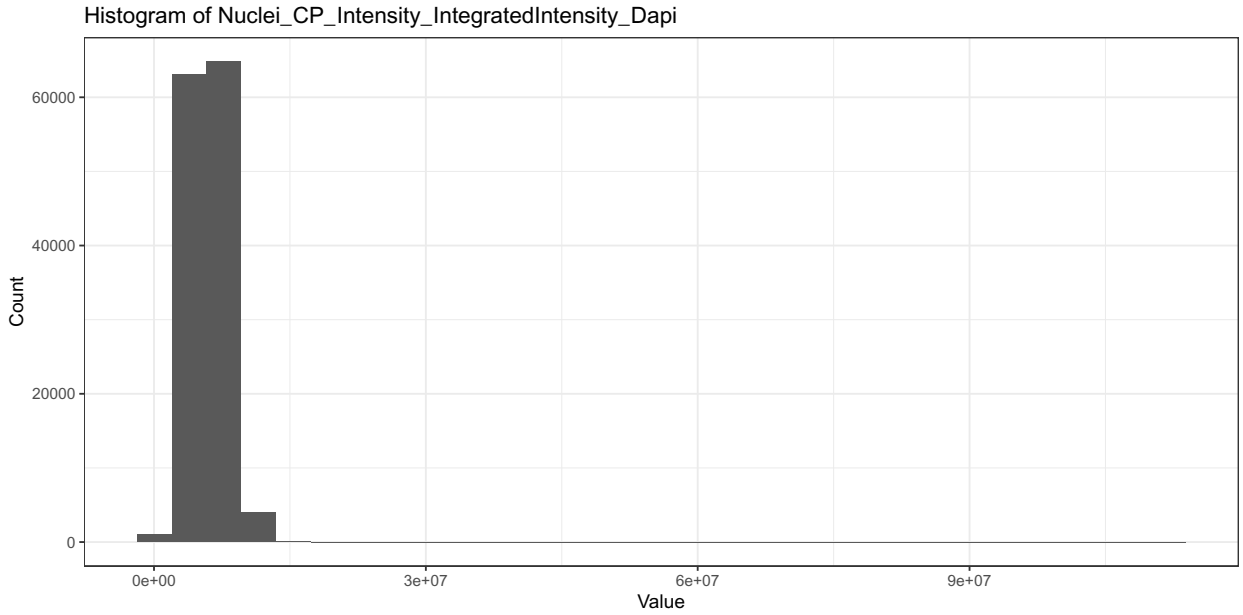


Figure 2: Histogram of total nuclei DAPI intensity across all wells and spots.

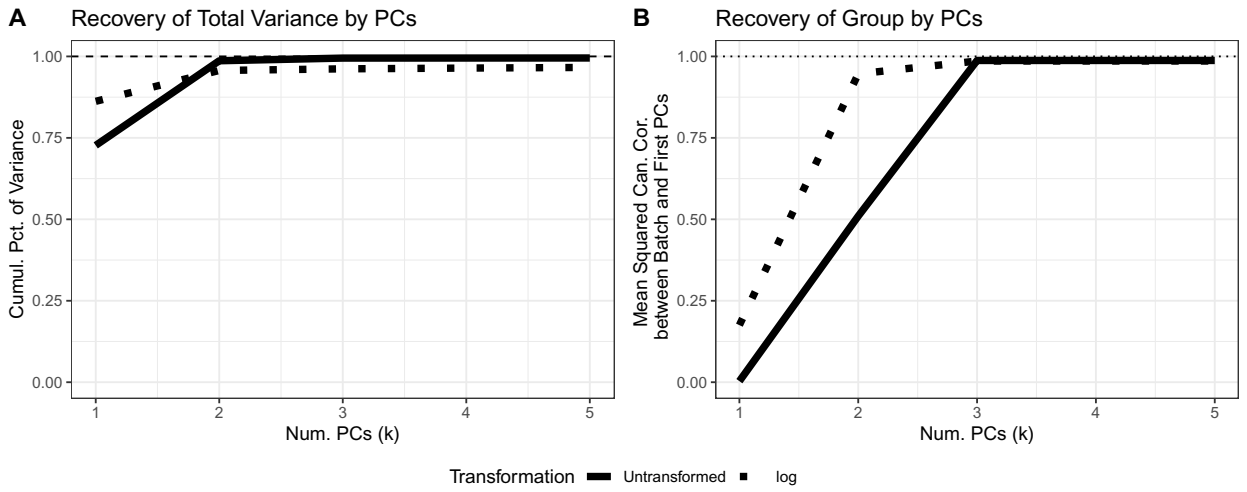


Figure 3: (A) The percentage of cumulative variance captured by first k principal components for both un-transformed data and log-transformed data. (B) The mean squared canonical correlations between the grouping factor and the first k principal components.

For clarity, we reproduce the explanation of (B) from main text.
We assess the transformations by measuring the percentage of the batch (group difference) captured by the first k singular vectors of the transformed feature matrix. Let $U = [u_1, \ldots, u_N] \in \mathbb{R}^{M \times N}$ be the (complete) left singular vectors of a feature matrix and $B \in \mathbb{R}^{M \times D}$ be the batch indicator matrix so that $B_{ij} = 1$ if well i is in batch j for $j = 1, \ldots, D$. Here we have $D = 2$ for the two groups. For $k = 1, \ldots, N$ and $t = 1 \ldots \min(k, D)$ define $C_k^{(t)}$ to be the t^{th} canonical correlation between the first k left singular vectors $U_k = [u_1, \ldots, u_k]$ and the batch B. Then let

$$C_k^2 = \frac{1}{D} \sum_{t=1}^{\min(k,D)} \left(C_k^{(t)} \right)^2$$

to be the average of these squared canonical correlations. We can interpret C_k^2 as the percentage of the batch B that is captured by these first k singular vectors.

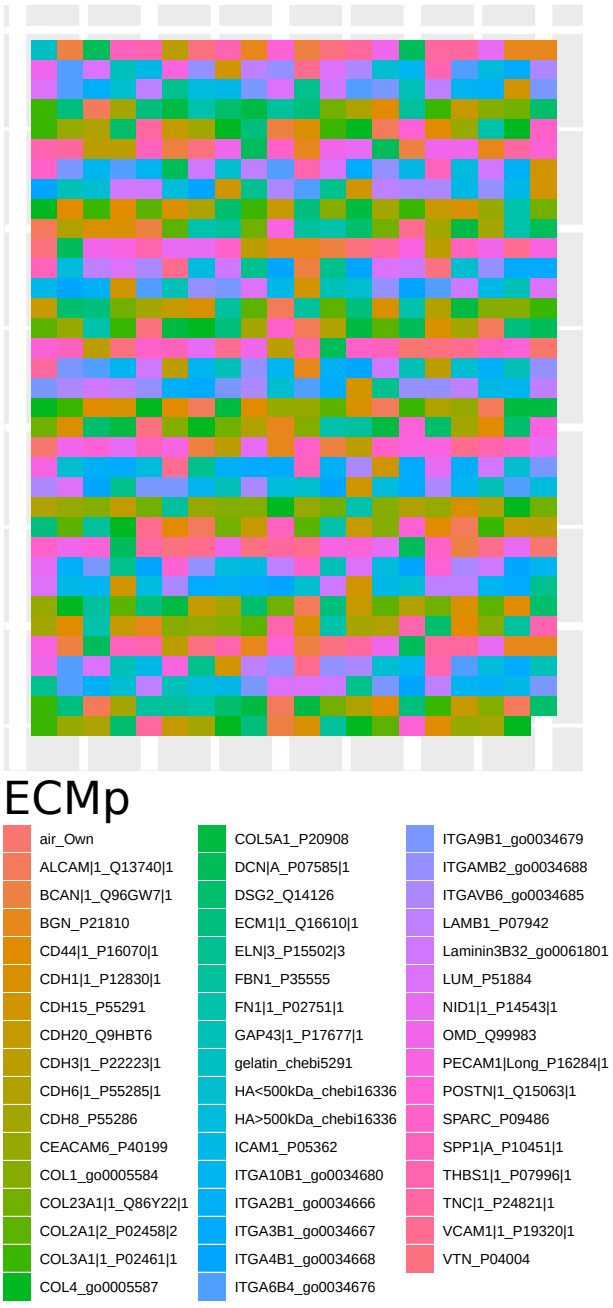


Figure 4: MEMA layout design of ECMps for each well. ECMp layout is identical across all wells.

Feature
Cells_CP_AreaShape_Area
Cells_CP_AreaShape_Compactness
Cells_CP_AreaShape_Eccentricity
Cells_CP_AreaShape_Extent
Cells_CP_AreaShape_FormFactor
Cells_CP_AreaShape_MajorAxisLength
Cells_CP_AreaShape_MaxFeretDiameter
Cells_CP_AreaShape_MaximumRadius
Cells_CP_AreaShape_MeanRadius
Cells_CP_AreaShape_MedianRadius
Cells_CP_AreaShape_MinFeretDiameter
Cells_CP_AreaShape_MinorAxisLength
Cells_CP_AreaShape_Perimeter
Cells_CP_AreaShape_Solidity
Cytoplasm_CP_AreaShape_Area
Cytoplasm_CP_AreaShape_Compactness
Cytoplasm_CP_AreaShape_Eccentricity
Cytoplasm_CP_AreaShape_Extent
Cytoplasm_CP_AreaShape_FormFactor
Cytoplasm_CP_AreaShape_MajorAxisLength
Cytoplasm_CP_AreaShape_MaxFeretDiameter
Cytoplasm_CP_AreaShape_MaximumRadius
Cytoplasm_CP_AreaShape_MeanRadius
Cytoplasm_CP_AreaShape_MedianRadius
Cytoplasm_CP_AreaShape_MinFeretDiameter
Cytoplasm_CP_AreaShape_MinorAxisLength
Cytoplasm_CP_AreaShape_Perimeter
Cytoplasm_CP_AreaShape_Solidity
Nuclei_CP_AreaShape_Area
Nuclei_CP_AreaShape_Compactness
Nuclei_CP_AreaShape_Eccentricity
Nuclei_CP_AreaShape_Extent
Nuclei_CP_AreaShape_FormFactor
Nuclei_CP_AreaShape_MajorAxisLength
Nuclei_CP_AreaShape_MaxFeretDiameter
Nuclei_CP_AreaShape_MaximumRadius
Nuclei_CP_AreaShape_MeanRadius
Nuclei_CP_AreaShape_MedianRadius
Nuclei_CP_AreaShape_MinFeretDiameter
Nuclei_CP_AreaShape_MinorAxisLength
Nuclei_CP_AreaShape_Orientation
Nuclei_CP_AreaShape_Perimeter
Nuclei_CP_AreaShape_Solidity
Cells_CP_Intensity_IntegratedIntensity_CellMask
Cells_CP_Intensity_IntegratedIntensity_KRT19
Cells_CP_Intensity_IntegratedIntensity_KRT5
Cells_CP_Intensity_MedianIntensity_CellMask
Cells_CP_Intensity_MedianIntensity_KRT19
Cells_CP_Intensity_MedianIntensity_KRT5
Cytoplasm_CP_Intensity_IntegratedIntensity_CellMask
Cytoplasm_CP_Intensity_IntegratedIntensity_Dapi
Cytoplasm_CP_Intensity_IntegratedIntensity_KRT19
Cytoplasm_CP_Intensity_IntegratedIntensity_KRT5
Cytoplasm_CP_Intensity_MedianIntensity_CellMask
Cytoplasm_CP_Intensity_MedianIntensity_Dapi
Cytoplasm_CP_Intensity_MedianIntensity_KRT19
Cytoplasm_CP_Intensity_MedianIntensity_KRT5
Nuclei_CP_Intensity_IntegratedIntensity_Dapi
Nuclei_CP_Intensity_IntegratedIntensity_KRT19
Nuclei_CP_Intensity_IntegratedIntensity_KRT5
Nuclei_CP_Intensity_MedianIntensity_Dapi
Nuclei_CP_Intensity_MedianIntensity_KRT19
Nuclei_CP_Intensity_MedianIntensity_KRT5
Cytoplasm_PA_Intensity_LineageRatio
Spot_PA_SpotCellCount
Cells_CP_Intensity_IntegratedIntensity_Actin
Cells_CP_Intensity_IntegratedIntensity_MitoTracker
Cells_CP_Intensity_MedianIntensity_Actin
Cells_CP_Intensity_MedianIntensity_MitoTracker
Cytoplasm_CP_Intensity_IntegratedIntensity_Actin
Cytoplasm_CP_Intensity_IntegratedIntensity_MitoTracker
Cytoplasm_CP_Intensity_MedianIntensity_Actin
Cytoplasm_CP_Intensity_MedianIntensity_MitoTracker
Nuclei_CP_Texture_AngularSecondMoment_Fibrillarin_3_0
Nuclei_CP_Texture_AngularSecondMoment_Fibrillarin_3_90
Nuclei_CP_Texture_Contrast_Fibrillarin_3_0
Nuclei_CP_Texture_Contrast_Fibrillarin_3_90
Nuclei_CP_Texture_Correlation_Fibrillarin_3_0
Nuclei_CP_Texture_Correlation_Fibrillarin_3_90
Nuclei_CP_Texture_DifferenceEntropy_Fibrillarin_3_0
Nuclei_CP_Texture_DifferenceEntropy_Fibrillarin_3_90
Nuclei_CP_Texture_DifferenceVariance_Fibrillarin_3_0
Nuclei_CP_Texture_DifferenceVariance_Fibrillarin_3_90
Nuclei_CP_Texture_Entropy_Fibrillarin_3_0
Nuclei_CP_Texture_Entropy_Fibrillarin_3_90
Nuclei_CP_Texture_InfoMeas1_Fibrillarin_3_0
Nuclei_CP_Texture_InfoMeas1_Fibrillarin_3_90
Nuclei_CP_Texture_InfoMeas2_Fibrillarin_3_0
Nuclei_CP_Texture_InfoMeas2_Fibrillarin_3_90
Nuclei_CP_Texture_InverseDifferenceMoment_Fibrillarin_3_0
Nuclei_CP_Texture_InverseDifferenceMoment_Fibrillarin_3_90
Nuclei_CP_Texture_SumAverage_Fibrillarin_3_0
Nuclei_CP_Texture_SumAverage_Fibrillarin_3_90
Nuclei_CP_Texture_SumEntropy_Fibrillarin_3_0
Nuclei_CP_Texture_SumEntropy_Fibrillarin_3_90
Nuclei_CP_Texture_SumVariance_Fibrillarin_3_0
Nuclei_CP_Texture_SumVariance_Fibrillarin_3_90
Nuclei_CP_Texture_Variance_Fibrillarin_3_0
Nuclei_CP_Texture_Variance_Fibrillarin_3_90
Nuclei_CP_Intensity_IntegratedIntensity_EdU
Nuclei_CP_Intensity_IntegratedIntensity_Fibrillarin
Nuclei_CP_Intensity_MedianIntensity_EdU
Nuclei_CP_Intensity_MedianIntensity_Fibrillarin

Table 1: List of all features extracted from at least one MEMA plate.

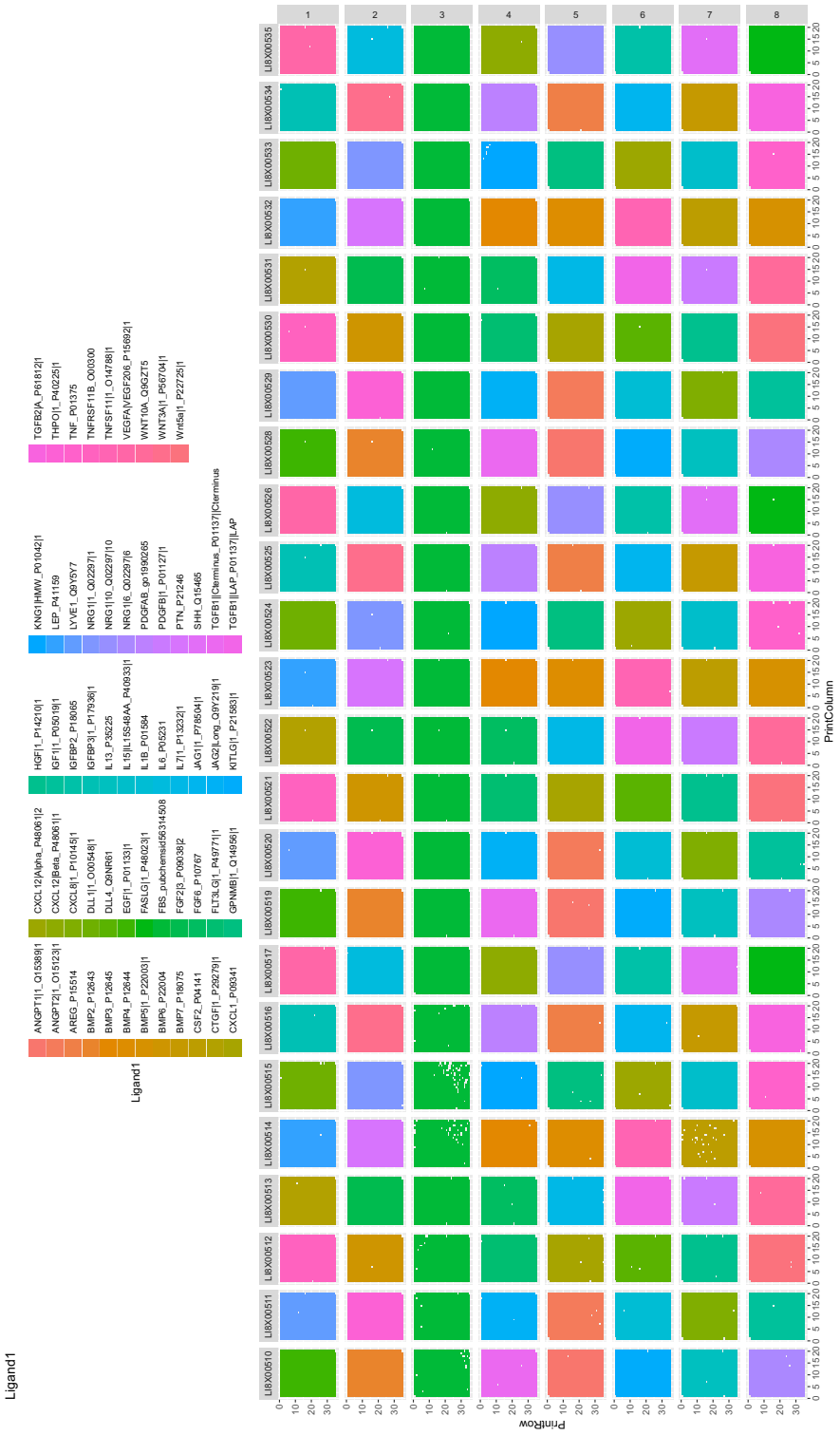


Figure 5: MEMA layout of ligands across plates and wells. Each row is a plate consisting of eight wells (columns). Color indicates ligand added to buffer solution of the well.

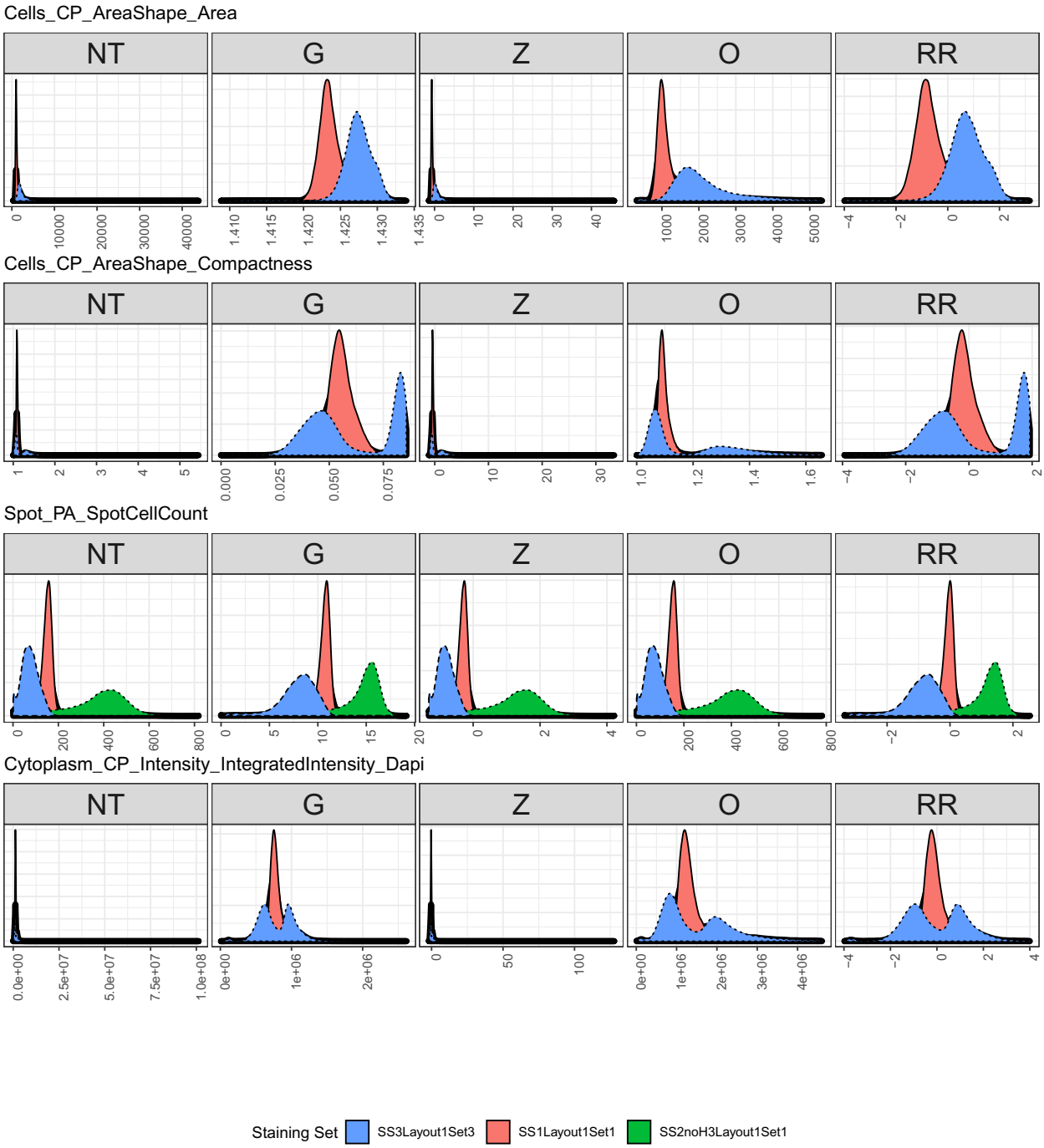


Figure 6: Density of elements of feature matrices. Black density is all elements combined. Colored densities are the densities denote staining batch. Subplots are for five processing transformations of this matrix: (NT) no transformation, (G) Gaussianization, (Z) z-score, (O) outlier removal, (RR) the three-step (G), (Z), and (O), robust re-scaling.

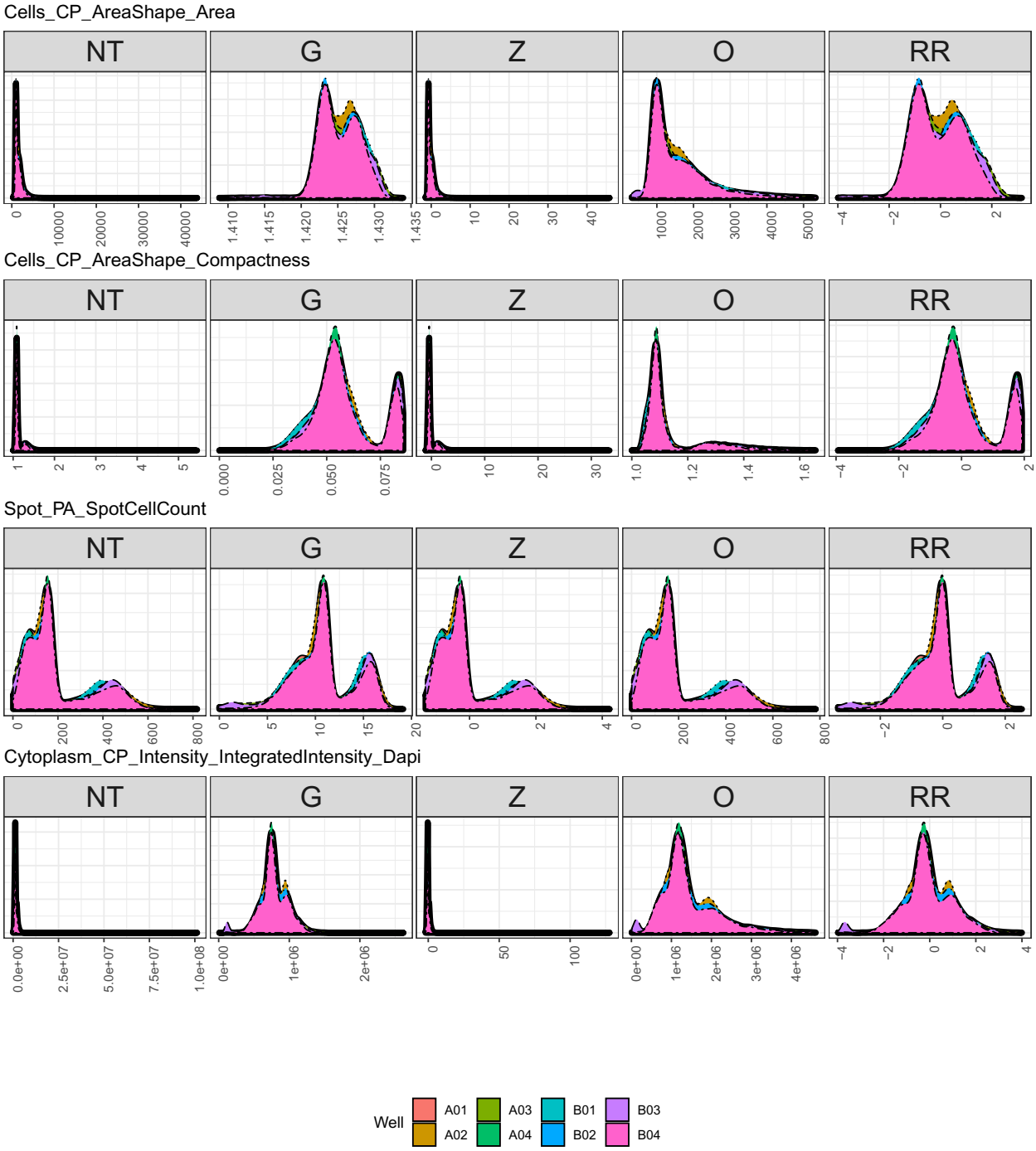


Figure 7: Similar to Figure 6 except colors indicate well.

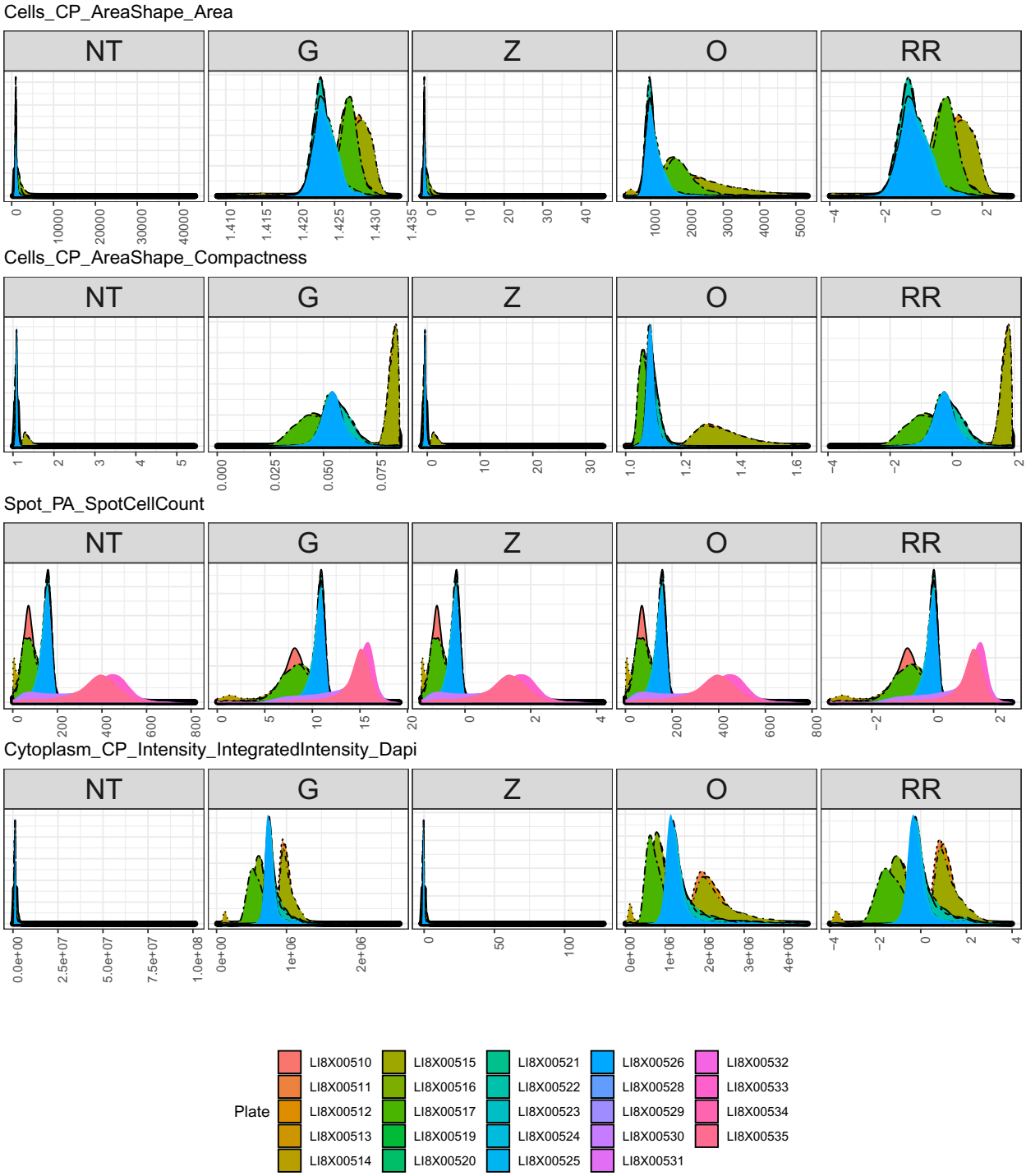


Figure 8: Similar to Figure 6 except colors indicate plate.

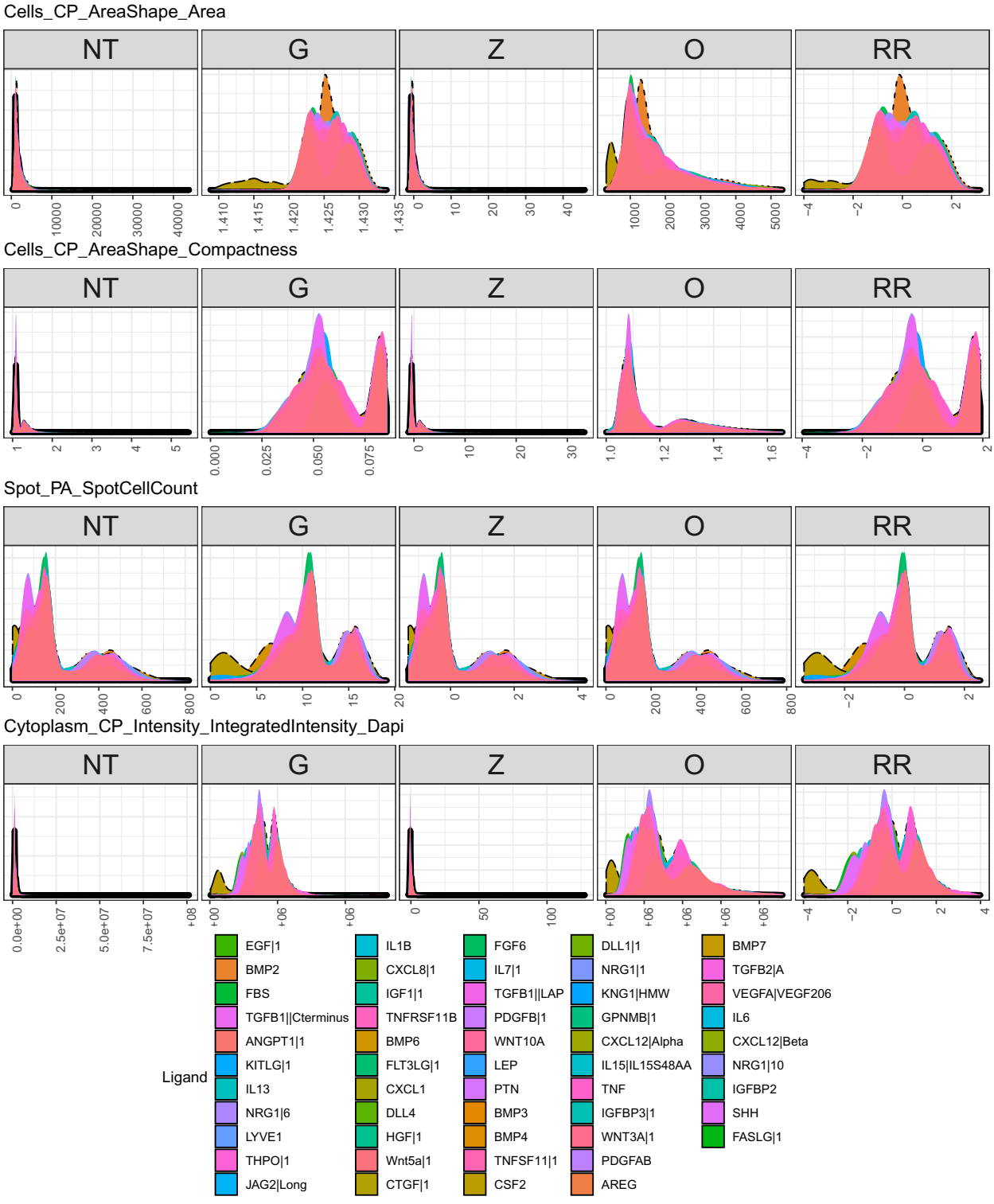


Figure 9: Similar to Figure 6 except colors indicate ligand.

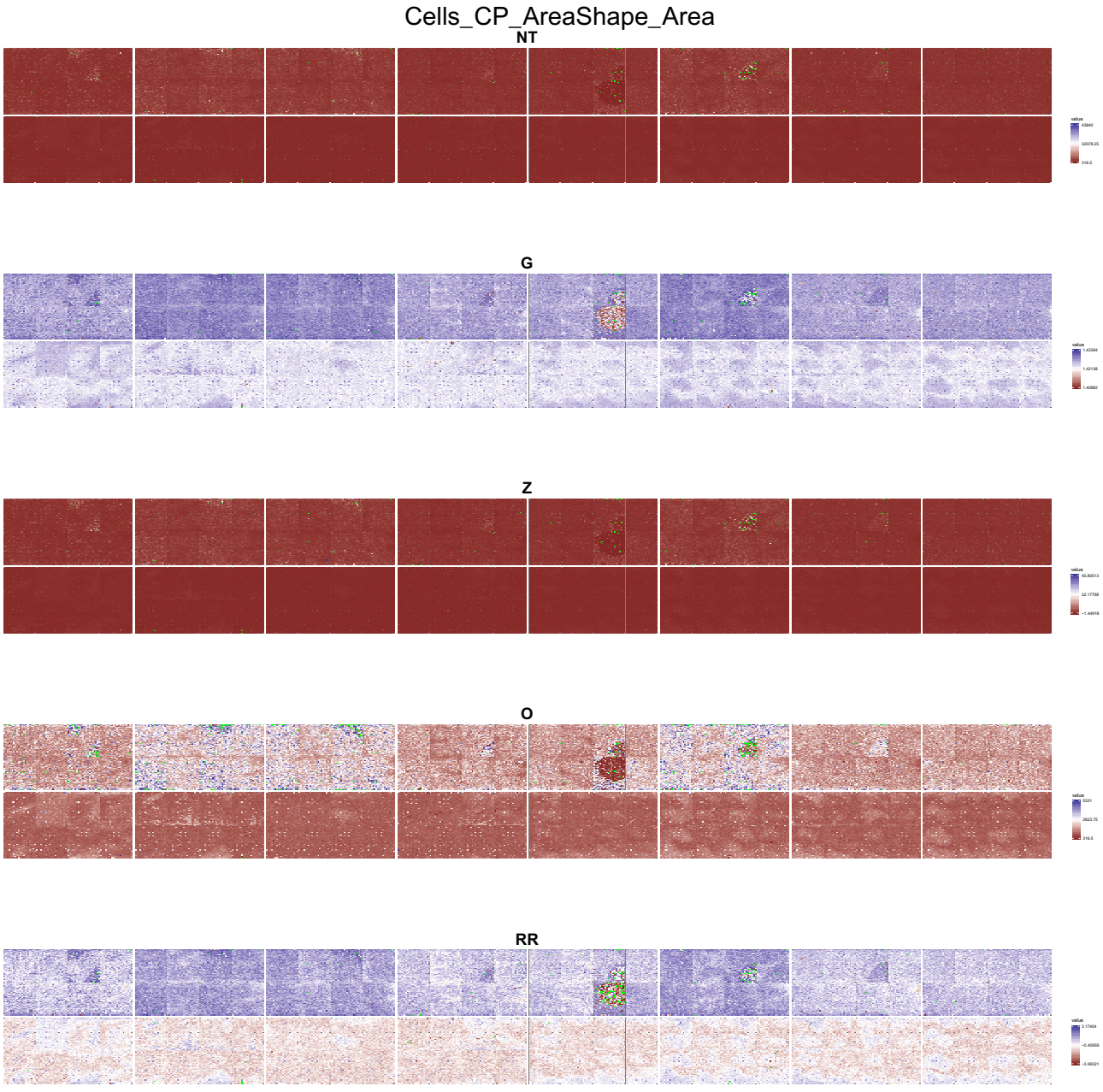


Figure 10: The next series of plots are heat-maps of MEMA plates across the five transformations (NT), (G), (Z), (O), (RR). Rows of each plot are the staining three batches. Colors are more blue if they are close to the minimum, red if they are close to the maximum, and white if they are close to half-way between. Green spots are missing. Dark grey spots are omitted according to the MEMA design.

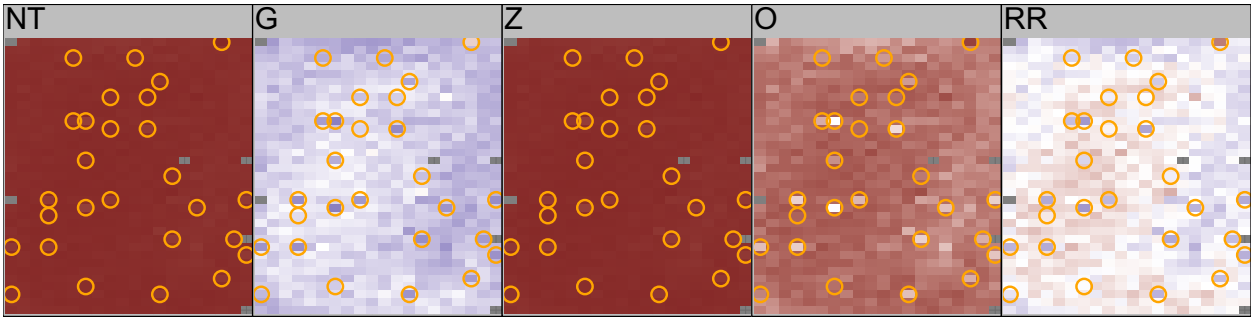


Figure 11: Heat map of a single well across the five transformations (NT), (G), (Z), (O), (RR). This is a sub-plot of Supplementary Figure 10. Color scaled is determined globally over all spots, wells, and plates in the dataset to reflect the fact that the transformation is similarly calculated over this data. Thus we see no blue in this (NT) sub-plot as we see almost no blue in Supplementary Figure 10. This plot is a representative microcosm of the larger plot. Orange circles highlight the ELN and NID1 spots.

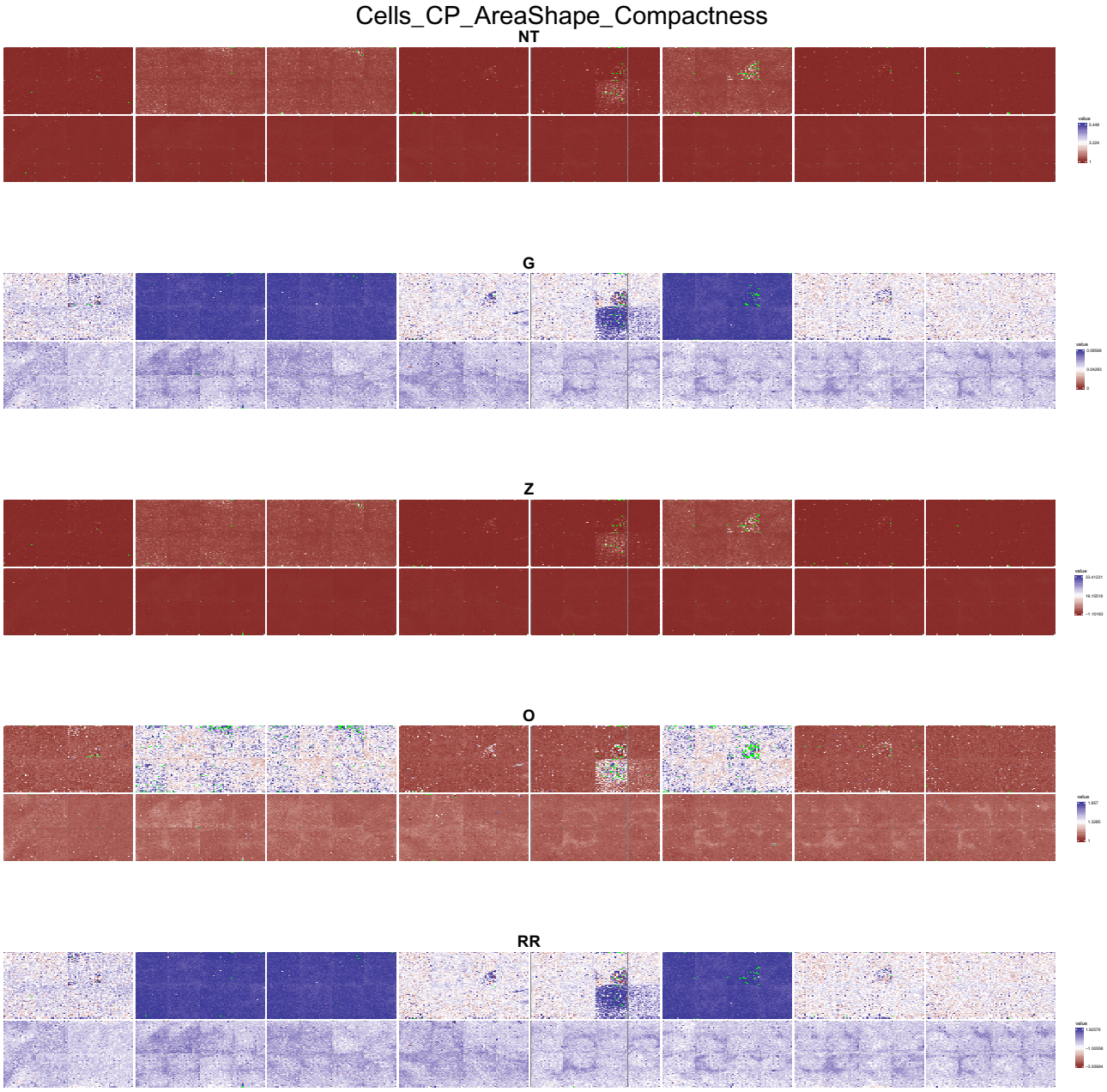


Figure 12: Similar to Figure 10 but for compactness.

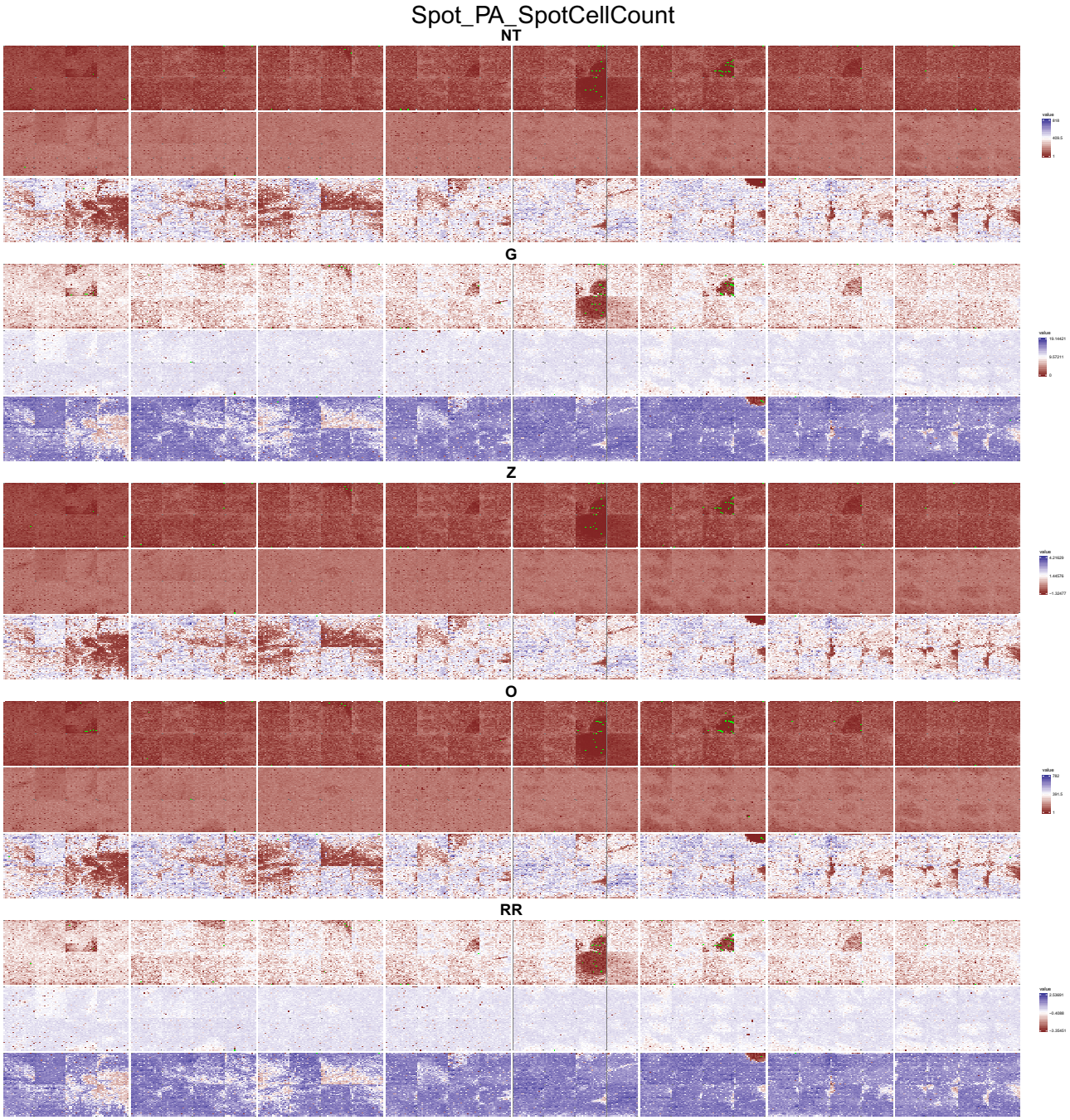


Figure 13: Similar to Figure 10 but for cell count.

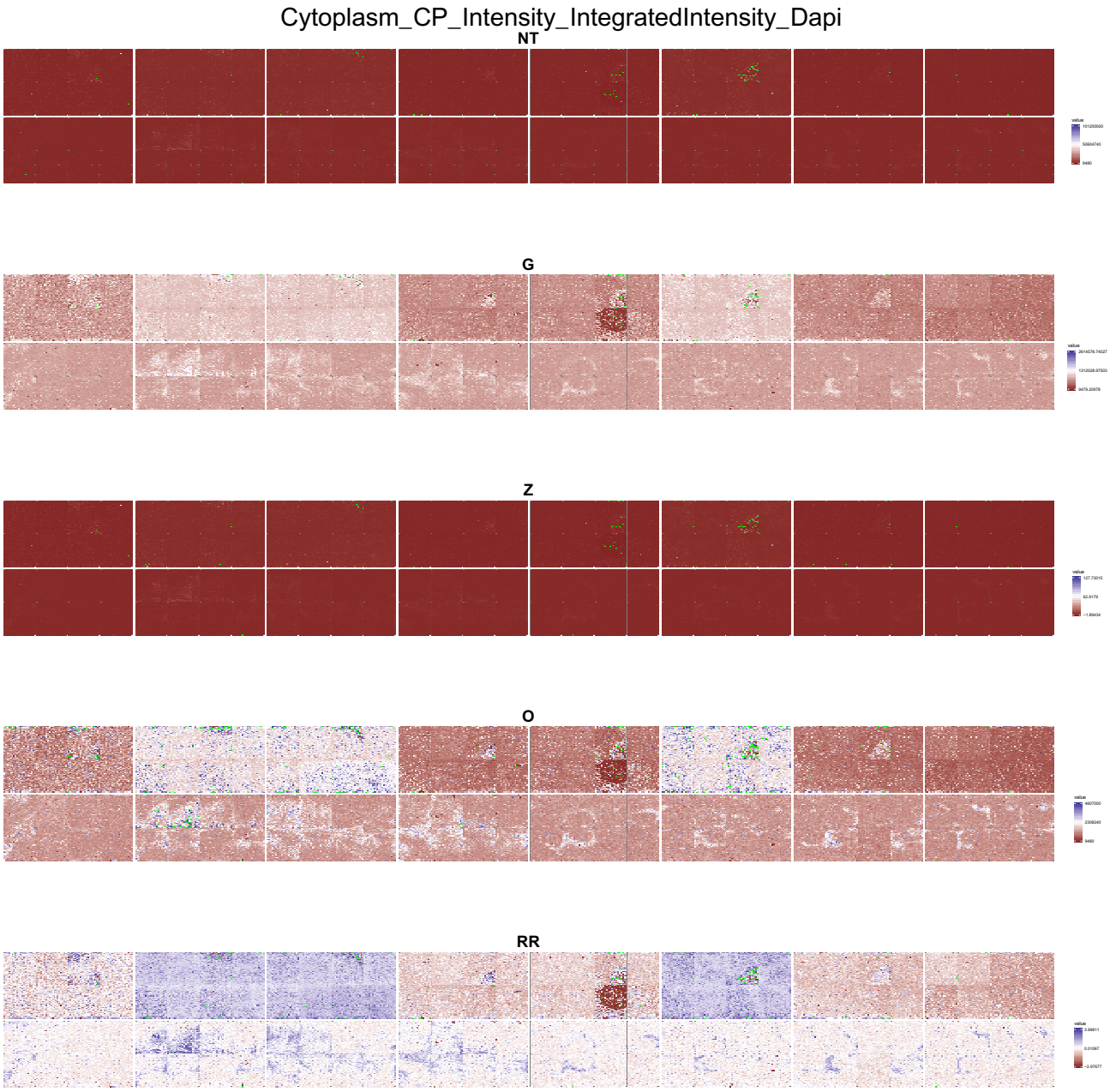


Figure 14: Similar to Figure 10 for for DAPI intensity.

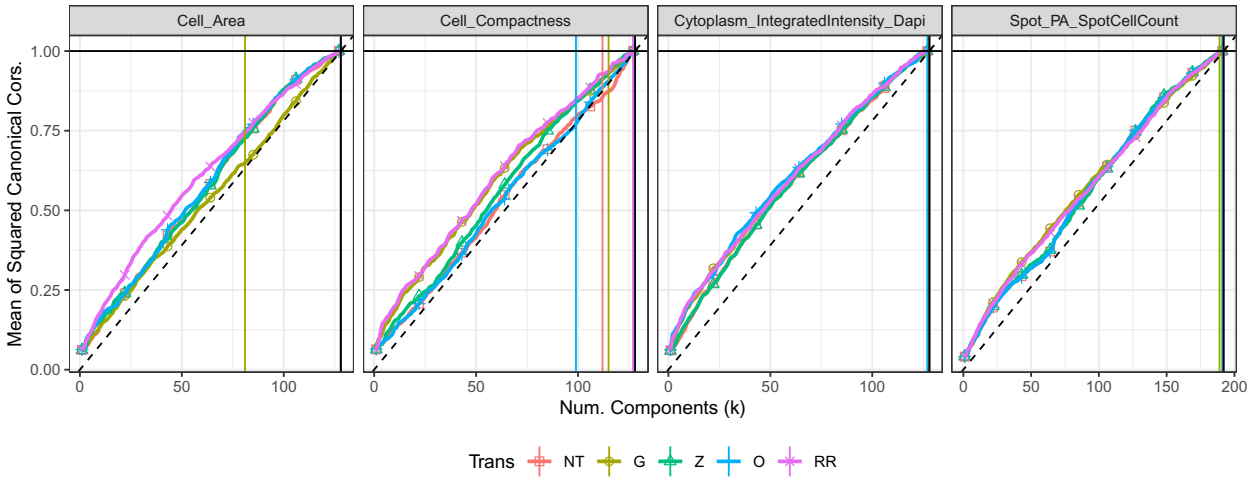


Figure 15: Mean of the squared canonical correlations between the first k principal components and the plate batch indicator variables.

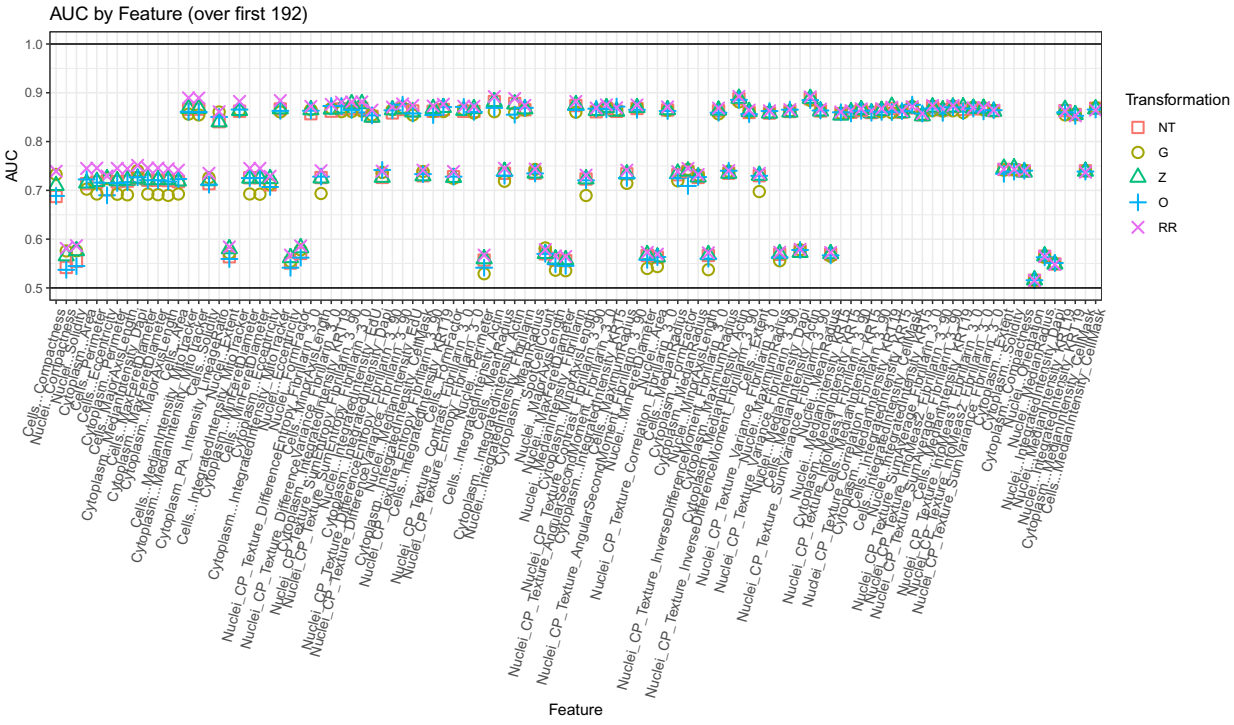


Figure 16: Grand mean of the squared canonical correlations across number of components (k). Canonical correlation is calculated between the first k principal components and the plate indicator variables.

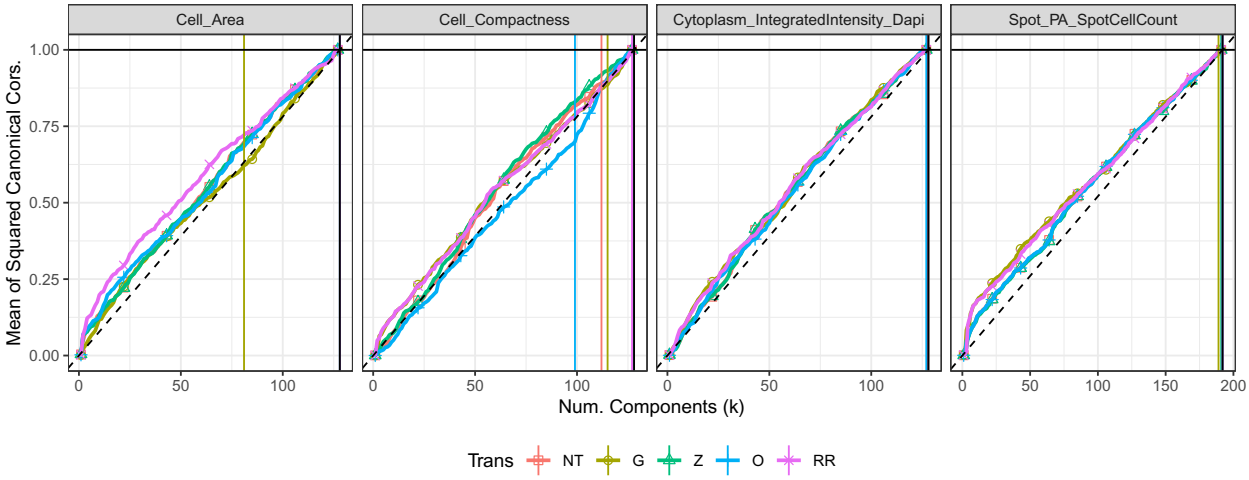


Figure 17: Similar to Figure 15 except correlation with well batch indicators.

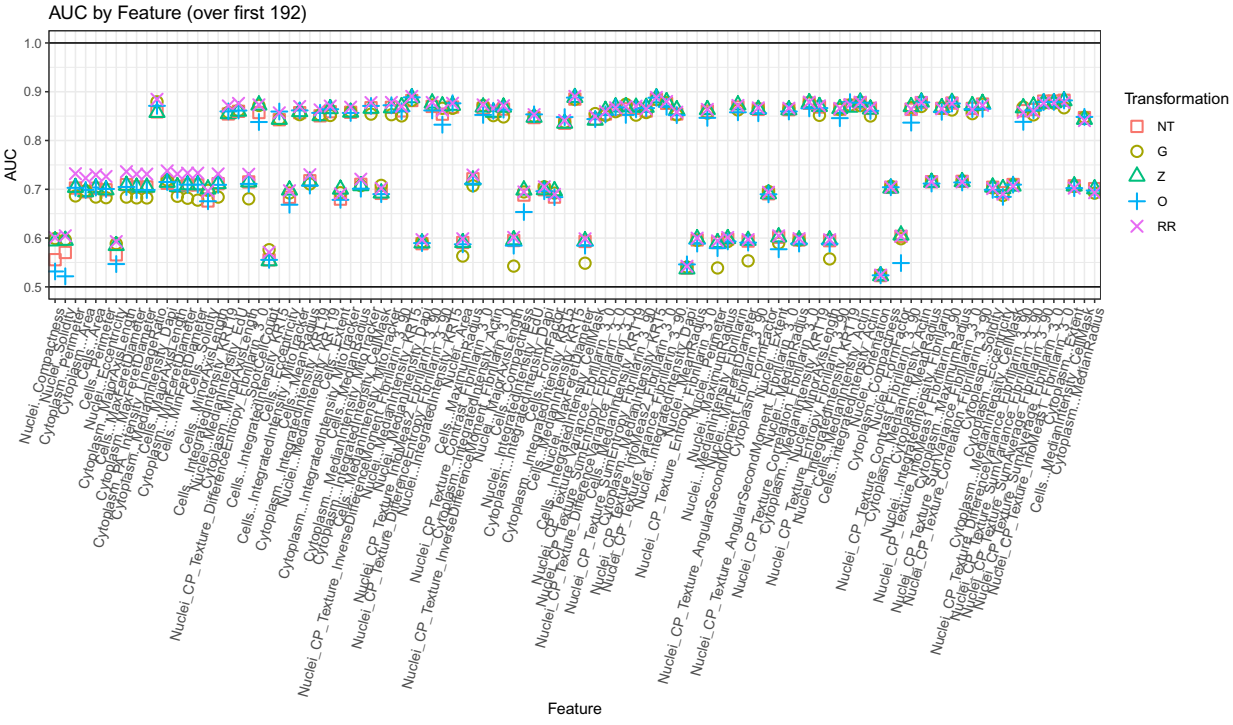


Figure 18: Similar to Figure 16 except correlation with well batch indicators.

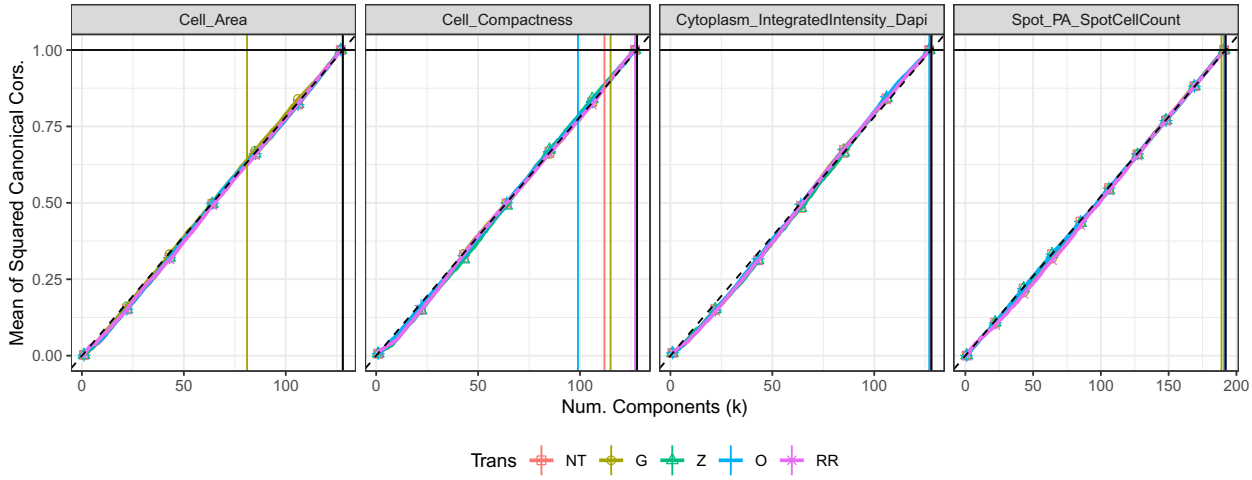


Figure 19: Similar to Figure 15 except correlation with ligand batch indicators.

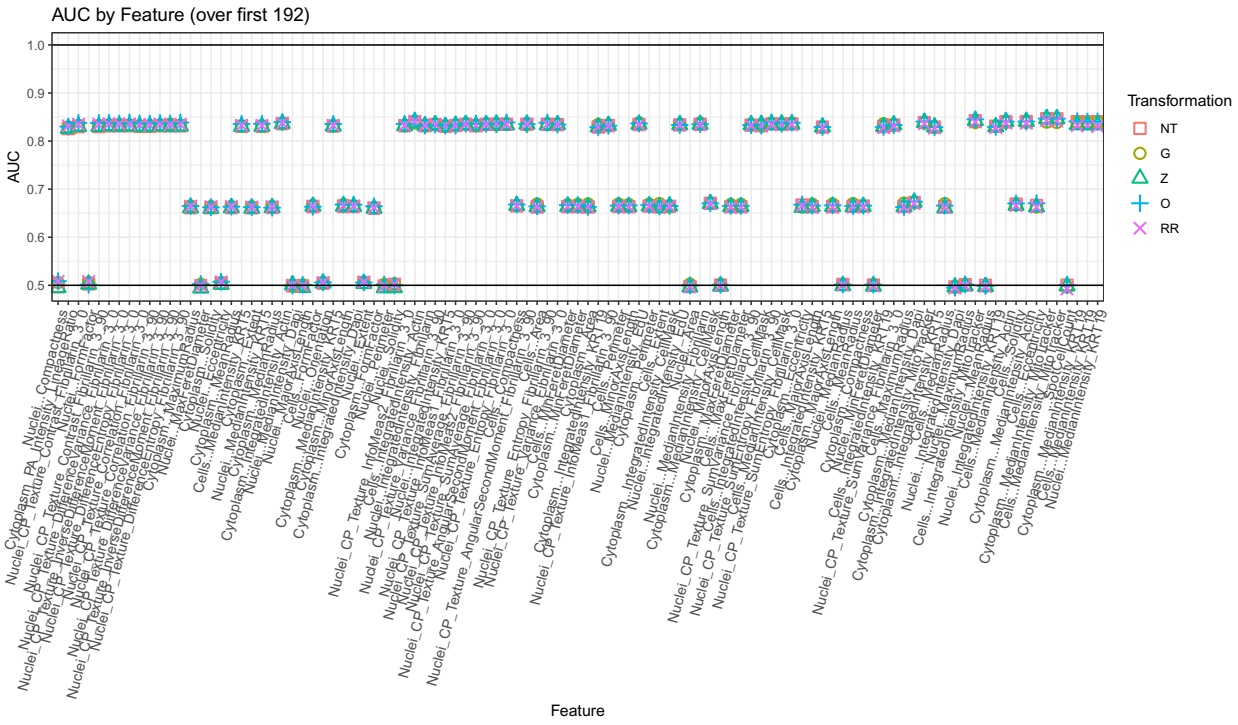


Figure 20: Similar to Figure 16 except correlation with well batch indicators.

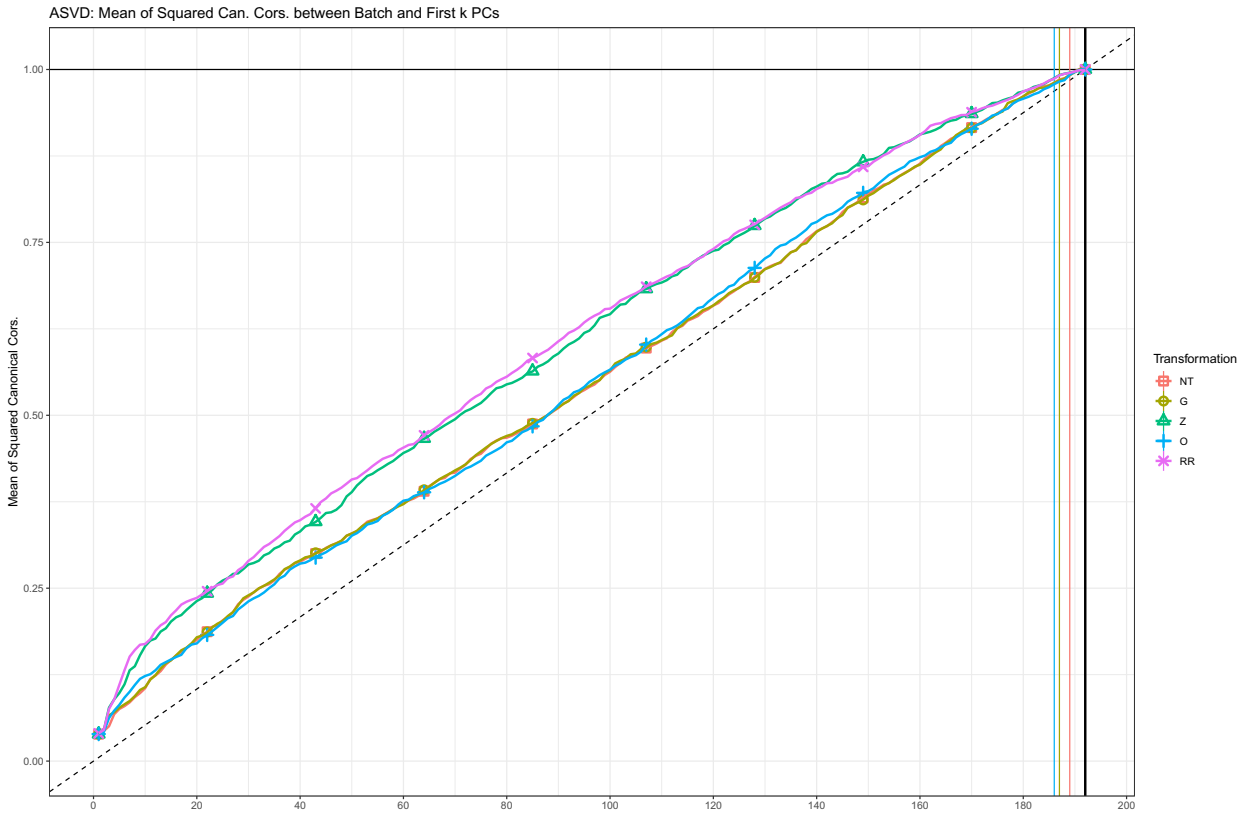


Figure 21: Mean of the squared canonical correlations between the first k principal components and the plate indicator variables. Principal components come from integration of the 21 features that are measured across all MEMAs.

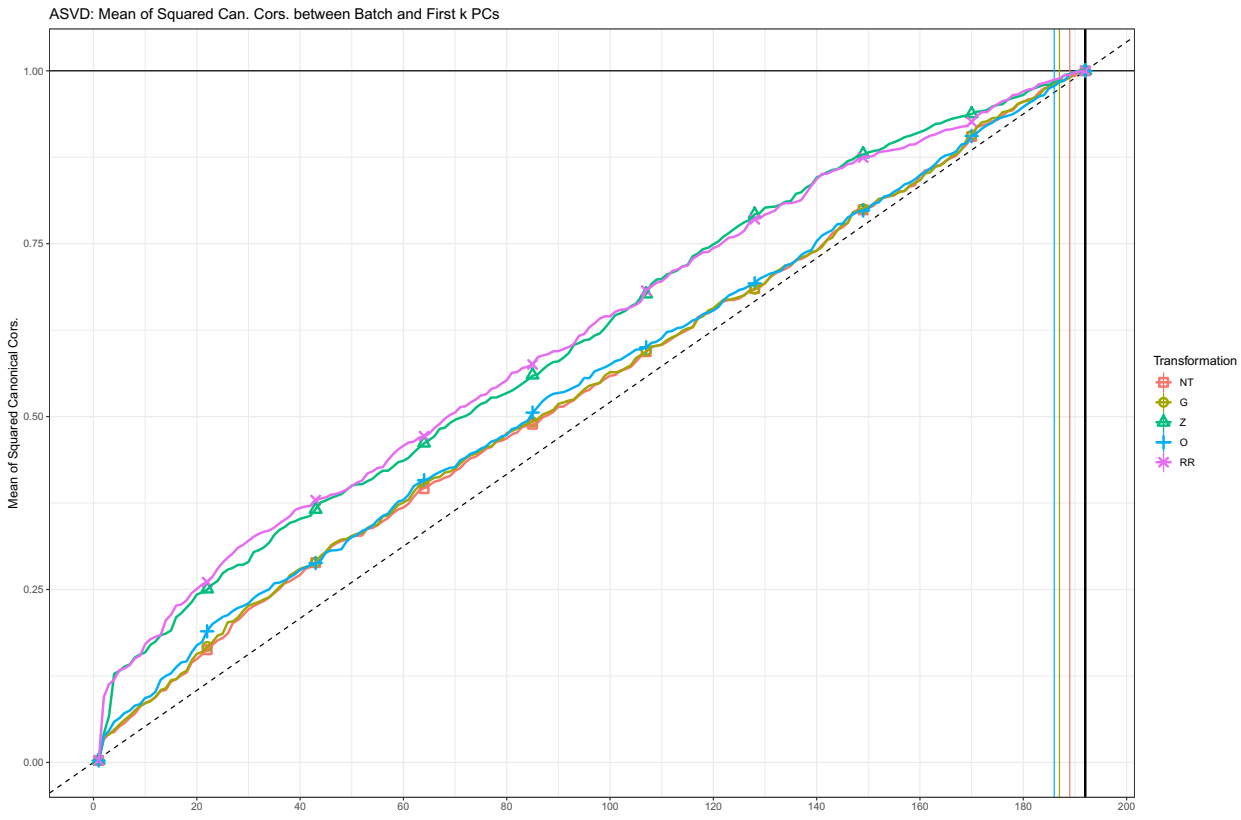


Figure 22: Similar to Figure 21 but calculating correlation with well indicators.

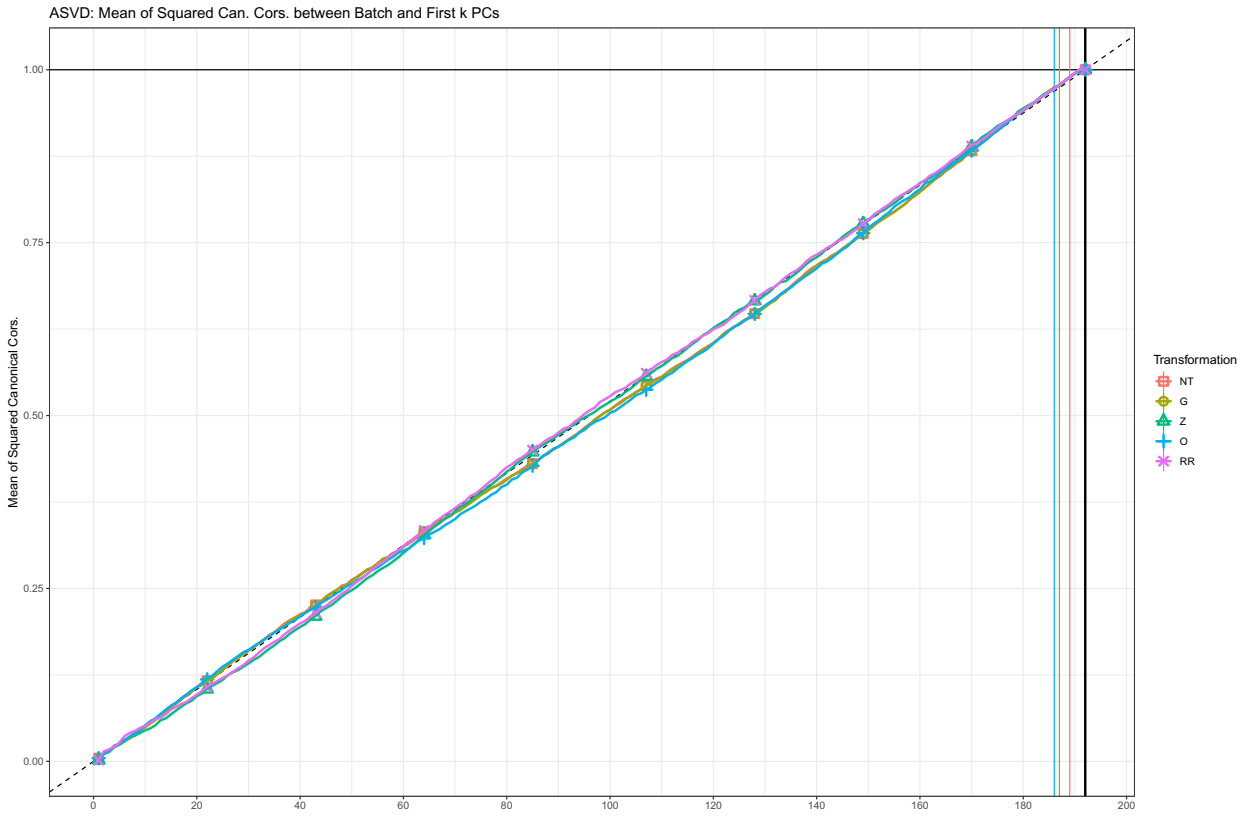


Figure 23: Similar to Figure 21 but calculating correlation with ligand indicators.

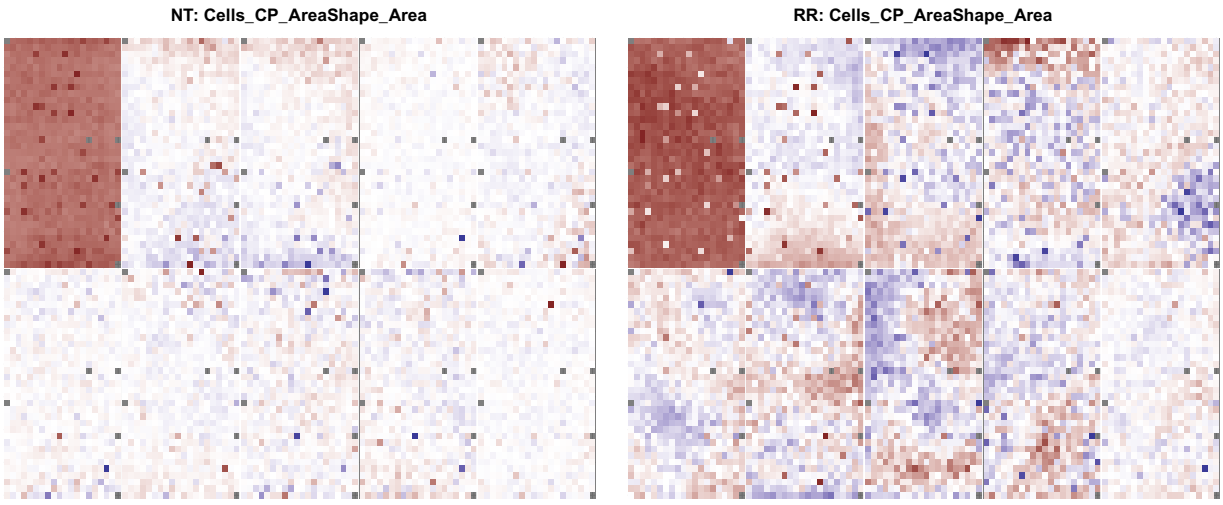


Figure 24: Heat map of elements of top 3 right singular vectors for the cell area feature.

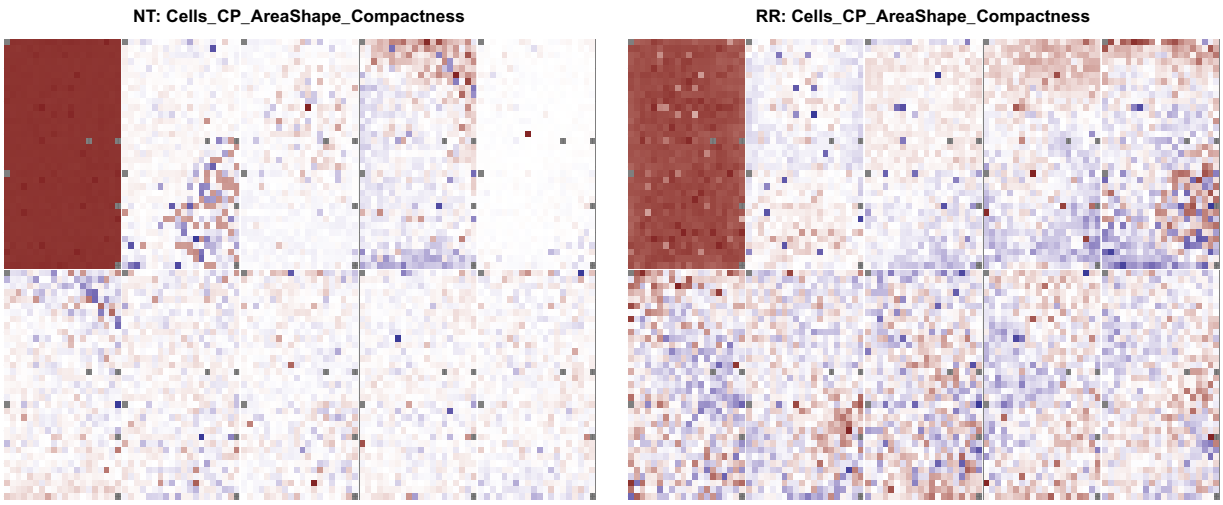


Figure 25: Similar to Figure 24 but for cell compactness feature.

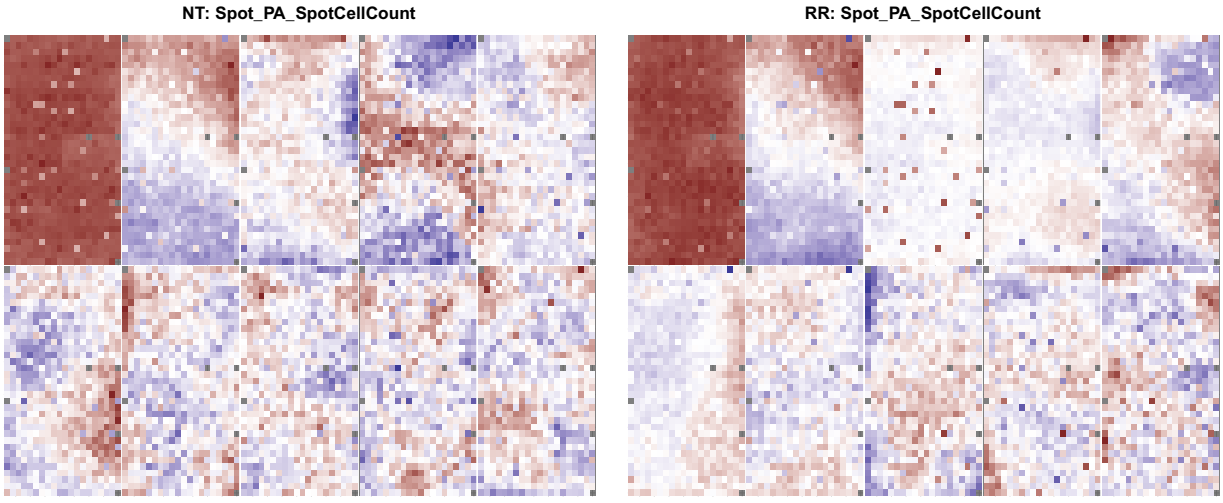


Figure 26: Similar to Figure 24 but for cell count feature.

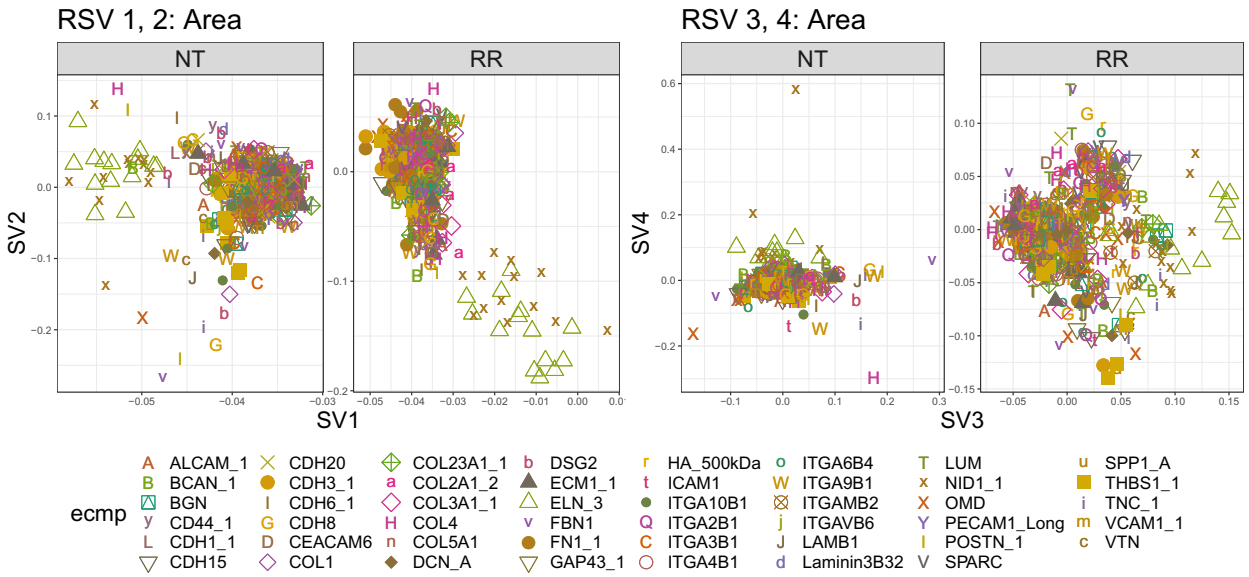


Figure 27: Scatter plot of elements of top two right singular vectors against each other for the cell area feature. Shape and color indicate ECMp of the spot corresponding to the elements of the singular vector.

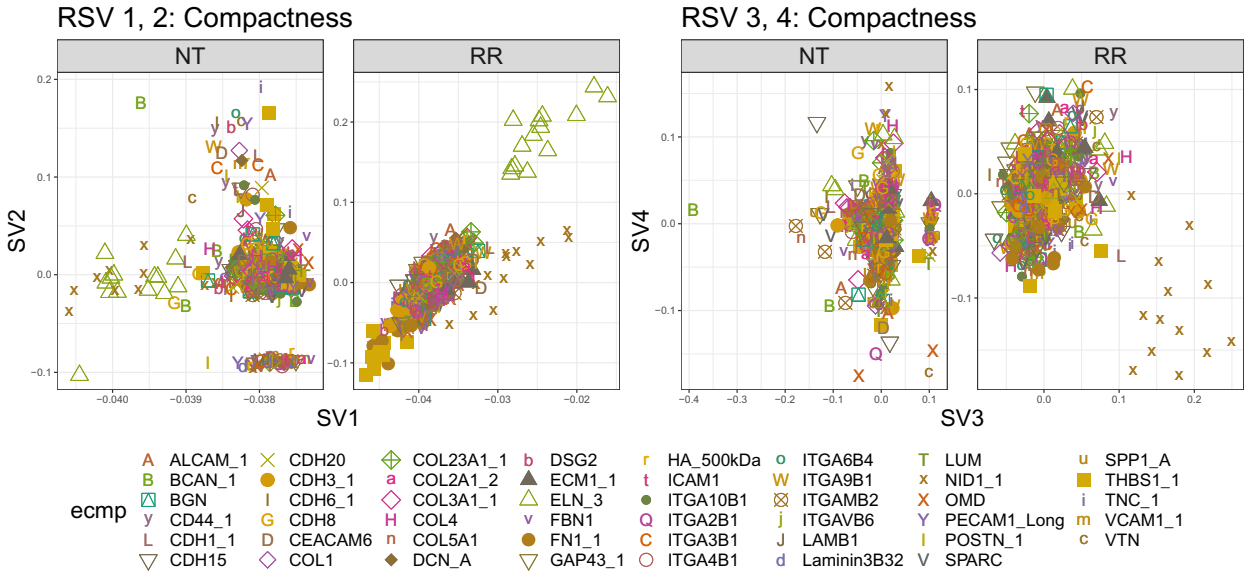


Figure 28: Similar to Figure 27 but for cell compactness feature.

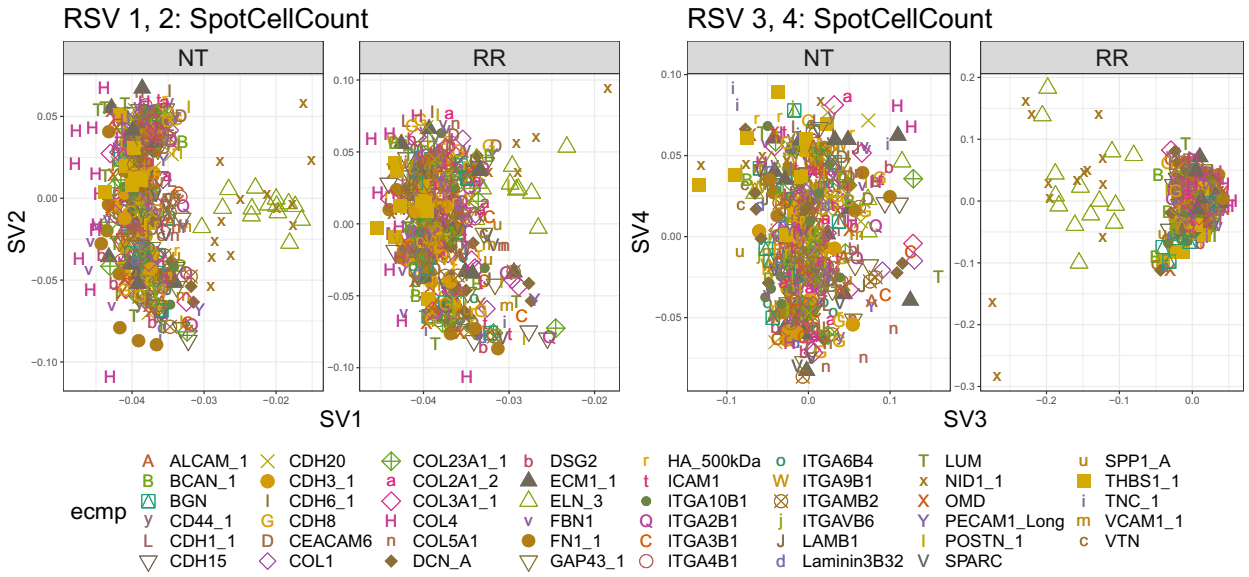


Figure 29: Similar to Figure 27 but for cell count feature.

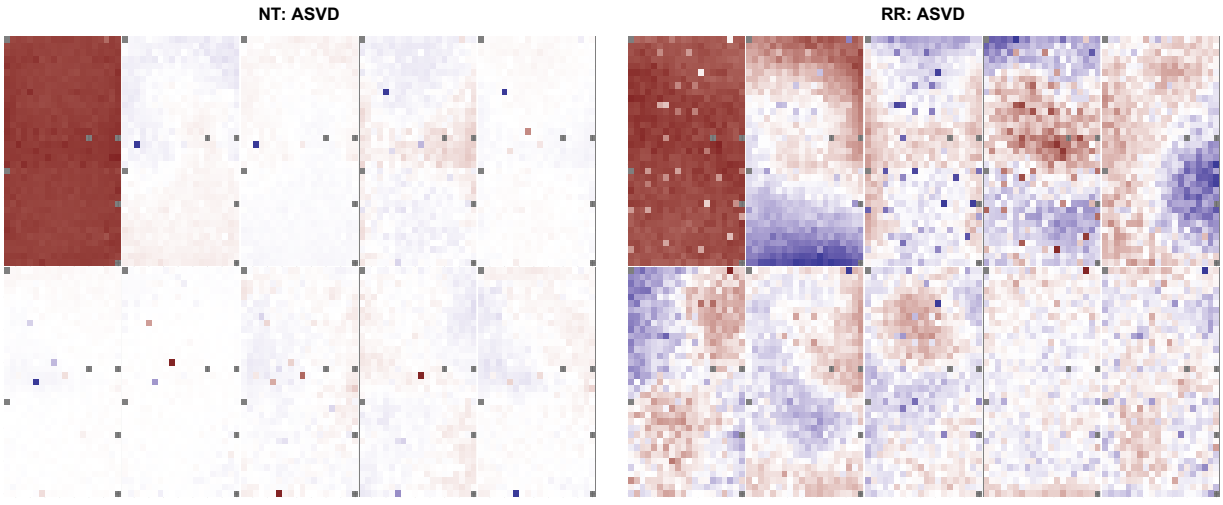


Figure 30: Heat-map of top 3 right ASVs calculated over 21 features measured on all MEMAs.

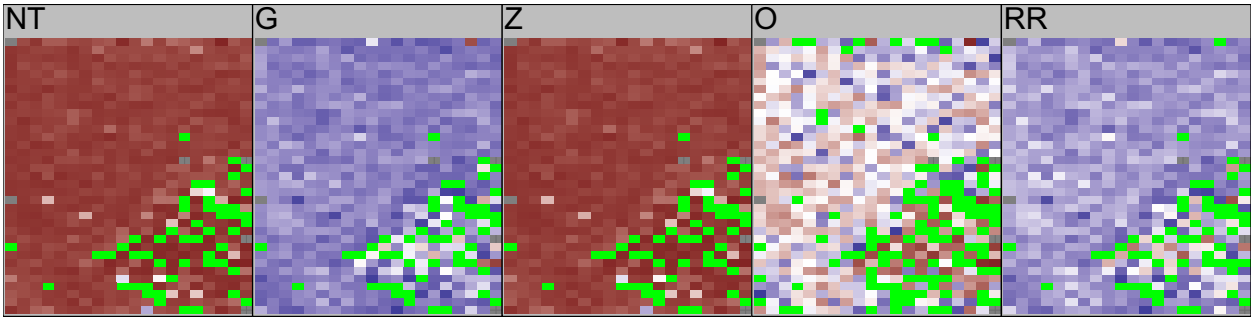


Figure 31: Missing values for well A03 on outlier plate LI8X00515. This plate is an outlier because it was processed using a different version of imaging processing software. Missing spots are indicated in green. Other colors indicate cell compactness feature. Notice that the missing values for (NT) are nearly identical to the the dark red spots in the second right singular vector in Supplementary Figure 25. This is what forms the group structure in Figure 11 as these missing spots are picked up on the outlying plate.

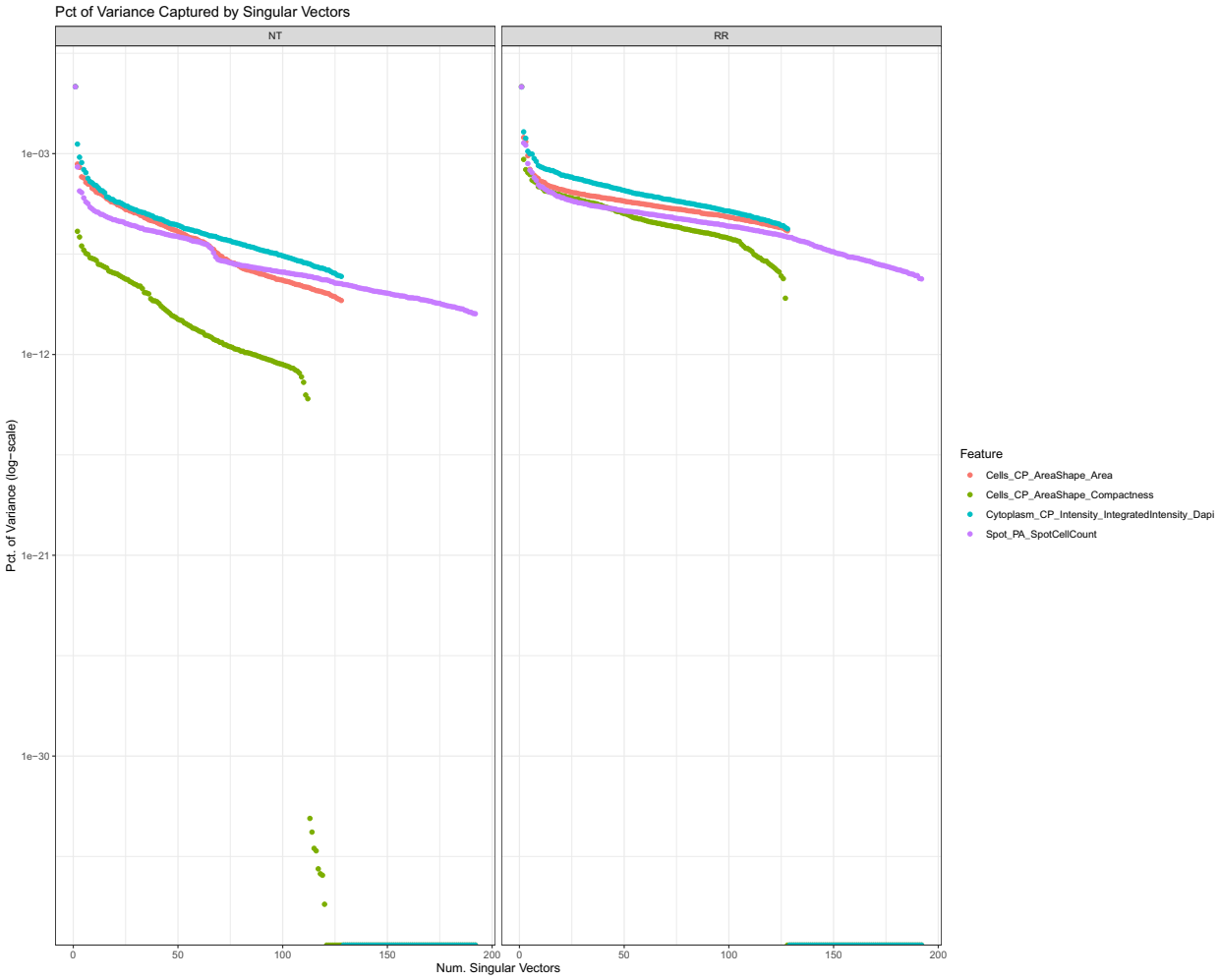


Figure 32: Pct. of variance captured by successive singular vectors for our four example features.