2016•2017

FACULTY OF SCIENCES
Master of Statistics

Master's thesis

Variance-stabilizing transformations for image-based compound profiling features

Supervisors:
Prof. Dr. Ziv SHKEDY
Prof. Dr. Nolen Joy PERUALILA
Dr. Marjolein CRABBE
Dr. Steffen JAENSCH

Wafula LEonard
Thesis presented in fulfillment of the requirements for the degree of Master of Statistics

Abstract

Background: Image-based multiparametric compound profiling is an important biological method for characterizing compounds in drug research and discovery. Many highly correlated and non-normally distributed features are often extracted and used to differentiate these compounds. Yet multivariate compound discrimination techniques are not optimal when applied directly due to redundancy in features extracted and lack of homogeneity in variance. A range of glog transformations were applied to stabilize the variance and assess effects of these transformations on separation between treatment replicates and non-replicates. A treatment was defined as a compound at a specific concentration. Investigations on effects of transformations on the proportion of actively-called treatments and whether treatments separation was better when transformations were applied on single-cell or aggregated data were also done.

Methods: Data sets from colon and liver cancer cell lines were used in this study. Two approaches - the Hotelling's two-sample T^2 statistic and the Area Under the ROC curve - were used to assess the effects of glog transformations on separation between treatment replicates and non-replicates. Transformations that led to positive shifts in distributions of T^2 statistics and/or higher AUC values were concluded as having improved treatment separation compared to when data was untransformed.

Results: Here, effects of glog transformations were investigated on actively-called treatments present before and after every given transformation. For both cell lines, a few transformations led to marginal improvements in treatments separation. These observed improvements were largely insignificant. Most transformations resulted in treatments separation that was indifferent to results obtained before transformation. Effect of transformation on proportion of actively-called treatments was data-dependent. In the first data, the proportion was in general lower than proportion obtained pre-transformation. On the contrary, the proportion of actively-called treatments in the second data set was always higher than the proportion observed pre-transformation. Across both data sets, a lower(higher) proportion resulted from a higher(lower) number of selected features. Both approaches failed to identify preference between applying a transformation at cell- or well-level.

Conclusion: Both approaches showed that performing glog transformations generally failed to improve treatments separation beyond results obtained before transformation. Further, our results determined that the proportion of actively-called treatments was data-dependent. Finally, our study could not establish whether transformations improved separation between treatment replicates and non-replicates better when applied at cell- or well-level.

Keywords: High content imaging, Glog transformation, Hotelling's T^2, AUC

Contents

Dedication

This Thesis is written as a gift to my superb family and friends and, my academic and career mentors for your belief in my abilities and never-ending support both within and without the academic circles. Kindly receive my appreciation in one package since listing you wonderful people here is beyond the bounds of possibility.

Acknowledgements

To GOD and my family, *asanteni sana*.

I deeply appreciate my internal supervisors Prof. Dr. Ziv SHKEDY and Prof. Dr. Nolen Joy PERUALILA for effortlessly triggering ideas and providing alternatives in the course of the project. It is incredible how at peace you are with statistics and the wonders it can inform. I was blessed to have a patient and motivated team of Dr. Marjolein CRABBE and Dr. Steffen JAENSCH as my external supervisors. You were excellent and vastly knowledgeable in the subject matter. You always seemed to have all the time in the world to wade me through any storm I had. Your eye for detail is beyond reproach. Thank you all for not just listening to my words and reading my mails and drafts, but especially hearing and replying to the spaces in them! You have been a gale-force-wind at my back.

I also wish to acknowledge the academic support I got from all lecturers at UHasselt and the Flemish government for materially supporting my studies and stay through the VLIR-OUS program. Credit to my classmates for all the splendid fun we had and for providing real options whenever any of us was stuck.

LEonard

1 Background

1.1 Introduction

Image-based multi-parametric compound profiling is an important biological method for distinguishing compounds in the drug discovery chain using a range of features extracted from image-based assays by applying High Throughput Microscopy (HTM) [1–3]. In HTM, a cell is identified by partitioning an image into regions using a computer image analysis algorithm. These cellular regions then provide information on intensity, texture and spatial distribution of intracellular biomarkers as well as the shape, quantity and geometry of cells. There has been an increased interest in multi-parametrically characterizing compounds at an early stage due to a rising need to understand how compounds induce their desired properties and describe their mechanisms of action, how to preferentially identify highly specific compounds from a pool of compounds that have a considerable activity on isolated biological target molecules and how to identify as early as possible any likely undesired effects (for instance toxicity detection) of compounds on cells and cellular activity [3–5]. Multi-parametric compound profiling methods, unlike univariate methods, factor in possible dependencies among features and often lead to more precise characterization, comparison and prediction of compound effects on cells [6, 7].

Whereas a richer set of features is vital in discriminating between compounds and predicting modes of action of compounds by examining compound feature profiles relative to profiles of compounds with known mechanisms of action [3], there is need to limit feature sets to only features that discriminate as many compounds as possible without being highly correlated among themselves [8, 9]. The problem of dimensional reduction of features is a common difficulty since HTM outputs data on numerous features most of which are usually highly similar in describing respective compound modes of action. Most features extracted from images also exhibit non-normal distributions which renders application of multivariate methods dependent on normality less optimal. Multivariate techniques for assessing possible relationships amongst groups of compounds are highly dependent on data variability [10–12] and thus necessitate appropriate handling of the data's variance structure to maximize on power of results obtained. Briefly, variation between a pair of compound samples can be categorized as either 'technical' or 'biological' [12–14] with technical variability being associated with procedures performed during an experiment for example sample preparation, extraction, labeling and analytical measurements. Technical variability blurs discrimination of compound samples and samples from other compounds. Therefore to maximize on separation between samples of compounds, this component of variation needs to be minimized. Biological variability refers to differences in feature(s) measurements from a particular cell line due to differences in modes of action induced by different compounds. Biological variability rightly differentiates compounds and methods

aimed at classifying compounds target to maximize this component of variation.

To multi-parametrically discriminate compounds therefore, it is necessary to both select a set of features that is efficient in discriminating compounds, and particularly for this report, apply transformations aimed at maximizing separation between compounds samples. These transformations target at minimizing the dispensable effects of technical variability while concurrently optimizing compounds separation through their biological variability [12–15]. Such transformations (and data pre-processing techniques) are referred to as Variance Stabilizing Transformations (VST). VST techniques primarily transform data such that simple analysis methods like linear regression can be applied without violating underlying assumptions. Key amongst these normality assumptions is homoscedasticity which essentially implies independence between the mean and variance functions. High content imaging data, similar to other high throughput data, often present a lack of independence between their mean and standard deviation [16] and thus inherently require variance stabilization to maximize on respective compounds separation. VSTs are also applied to improve graphical exploration of data by rescaling it so that visualization is much clearer.

Variance stabilizing transformations come in different forms with the most commonly applied ones being auto-scaling, Pareto scaling [14], Yeo-Johnson power transformations [17, 18] and the generalized log transformation (glog) [19–23]. The ability of each of these transformations to improve compounds classification by stabilizing data variability and retaining the necessary biological variability varies. The glog transformation has been investigated and shown to lead to better variance stabilization compared to auto-scaled and Pareto-scaled data across a set of metabolomics data [12]. Noteworthy, improved classification and clustering resulting from effective variance stabilization was shown when a glog transformation was applied on metabolomics data [15]. As finely noted in [12], the major strength of the glog transformation is its ability to minimize technical variability among technical replicates, which is a strength that auto-scaling and Pareto-scaling lack. In fact, auto-scaling collapses all variability in the data to unity and therefore makes all biological variability in the data uniform.

Multiparametric high content imaging is a rich tool in research, discovery and characterization of compounds at an early stage in the drug discovery pipeline [1, 2, 4, 6] whose importance has been greatly documented [3, 5]. This report explored the possibility of applying variance stabilizing characteristics of transformations on high content data to promote better profiling of compounds using image-based multi-parametric phenotypes. We investigated effectiveness of transformations on two sets of disparate data categorized based on cancer cell lines from which they were extracted. The glog transformation was applied in this study since it has been shown to lead to better classification and clustering results when applied on other data sets [12, 15] and also compares favourably to other commonly applied variance stabilizing transformations [12]. We performed glog transfor-

mations on these data sets in combination with other data pre-processing techniques. We then assessed effects of a transformation on separation of compounds samples from non-samples using both Hotelling's T^2 statistic and area under the curve approaches. Both the Hotelling's T^2 statistic and area under the curve analysis methods were applied on data before and after applying a specific glog transformation.

1.2 Objectives

The primary objective of this study was to investigate effects of glog variance stabilizing transformations on the separation of treatment replicates from non-replicates on high content imaging data. A treatment was defined as a candidate compound at a specific concentration. Treatment replicates refer to individual samples of the same treatment, whereas non-replicates are individual samples of different treatments. Further, effects of variance stabilizing transformations on proportions of treatments determined as active or inactive was investigated. Investigations were also done to determine whether transformations resulted in better treatment discrimination when performed at single-cell level or after aggregating at well level.

1.3 Statistical analysis plan

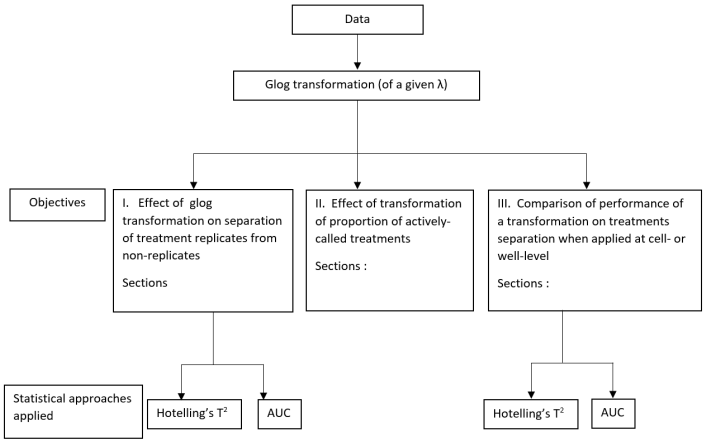


Figure 1: statistical analysis

Figure 1 gives a snapshot of the backbone statistical analyses applied in this report. For each data, a number of data pre-processing steps and a glog transformation at a specific λ value were initially applied (sections 2.1- 2.5). The first and third objectives were then assessed using both the Hotelling's T^2 statistic (section 3.1) and AUC (section 3.2) while assessment of transformation impact on proportion of active-calling was a direct product

of active-calling procedure described in section 2.5. Specific sections summarizing results and discussions for each objective are as listed in figure 1. Overall, the report is prepared in 5 sections. This section gives a brief introduction to high content imaging data and wide-ranging application and effectiveness of variance stabilizing transformations, with a particular inclination to glog transformation. Data description and, data pre-processing procedures are explained in section 2 while statistical analyses methods are introduced in section 3. Section 4 devotes to the study results whereas discussion of results from analyses and concluding remarks are offered in section 5.

2 Data and data pre-processing

Two high-content imaging data sets from two different cell lines were used to underline conclusions drawn from the analysis and assess the effectiveness of VSTs in improving discrimination between treatment replicates and non-replicates. The cell lines used were a colon cancer cell line fluorescently labelled for the cytoskeleton and endosomes and a liver cancer cell line labelled for the endoplasmic reticulum and mitochondria. Each data set consisted of three batches with each batch containing eighteen 384-well plates. Each data set had a total of 16236 wells. Every plate contained between 134909 and 281177 cells in the first data and between 152679 and 330117 cells in the second data. Each well comprised of between 73 and 1812, and between 95 and 2120 cells in the first and second data sets, respectively. We tested a total of 1253 treatments. Each treatment was a unique combination of a compound and a concentration level. In total, 311 compounds were tested including *Dimethyl SulfOxide* (DMSO) as a control (concentration level 0). Each compound had either 4 or 5 concentration levels. The concentration levels were 1 µM (microMolar), 3 µM, 3.34 µM, 9 µM and 11.1 µM. In total, 462 features on intensity, texture and spatial distribution of intracellular markers as well as cellular shape and geometry were extracted from each cell using a custom Acapella (PerkinElmer) image analysis script. A sketch of the data structure and dimensional reduction from cell to well level is shown using the first data set in figure 2. The data pre-processing procedures including transformation, aggregation, normalization, feature selection and active calling are described below.

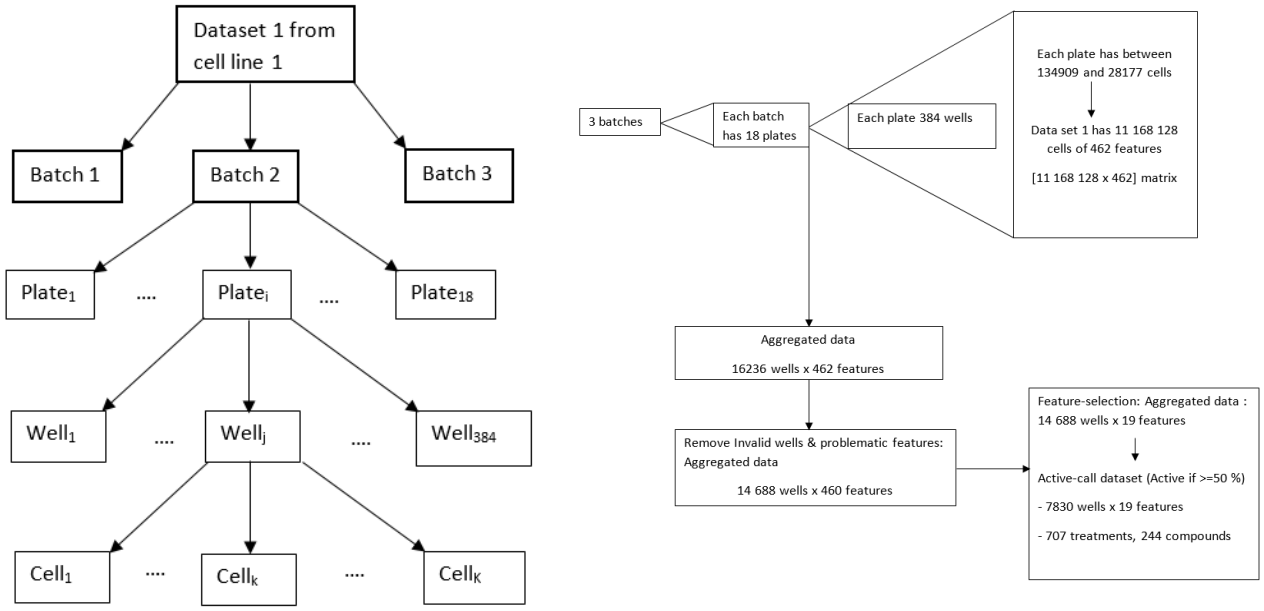
2.1 Generalized log transformation (glog)

Most of the multivariate methods, as has been pointed out in section 1.1, rely on the assumptions that data variability is independent of the mean and other underlying gaussian assumptions [15, 23, 24]. However, in high content imaging data, these assumptions rarely hold and skewness in the distributions is often observed. A variety of methods can be used to stabilize the variance [14,17,19–21]. The glog transformation, defined in equation 1, was used in this study.

$$z = \mathrm{Log}(y - \alpha + \sqrt{(y-\alpha)^2 + \lambda}) \tag{1}$$

where y is the untransformed data for a given feature, α is the feature mean for a feature across DMSO controls (the mean of low level measurements), λ is the transformation parameter while z is the glog-transformed data.

Estimation of λ can be done via a maximum likelihood method using a set of replicate measurements where the likelihood is maximized by minimizing the sum of squares of er-



Data structure

Data dimension

Figure 2: Data structure and dimension using data set 1

rors (SSE) of the transformed, Jacobian corrected data [15, 25]. In this study, however, the maximum likelihood estimate of λ was not estimated due to the demanding computations needed as a result of the rather voluminous data involved and also, there was need to examine a range of glog transformations without limiting to a single transformation calculated at the maximum likelihood estimate (MLE) of λ. Therefore, a range of λ values between 0.1 and 25 with an interval of 0.1 were simultaneously investigated. The α values were estimated from DMSO control measurements for each of the features per cell line as described in [23]. That is, the respective feature αs were the means of that feature for the measurements observed in the DMSO control treatment across all the wells and plates for that cell line.

Besides the glog transformation, many other variance stabilizing transformations exist. These include auto-scaling, Pareto scaling [14] and the Yeo-Johnson power transformations [17,18]. Autoscaling transforms a feature such that the mean is zero and has variance equal to one. While autoscaling results in a uniform variability such that there is no bias towards features with high variation, it implies that features are now treated as equally important and therefore limits the role of features that are more relevant in describing a treatment's mechanism of action. Further, as noted in [14], autoscaling is known to lead to inflation of the measurement error. Pareto scaling on the other hand transforms features such that the

relative impact of features with large variations is reduced while concurrently maintaining the initial data structure. Unlike in autoscaling where the difference in the feature value and the mean is divided by the standard deviation, this difference is divided by the square root of the standard deviation when performing a pareto transformation. As pareto-scaled data structure largely resembles an untransformed data and is also highly sensitive to large changes in feature values, it often results in a considerable number of features with comparatively large variances. Therefore, its variance stabilizing ability is quite limited. Yeo-Johnson power transformation is a single-parameter (λ) modification of the Box-Cox transformation [26] that includes transformations in cases of a negative dependent variable. Much like in Box-Cox transformation, the power parameter λ is estimated using maximum likelihood methods by selecting the λ value that maximizes the profile likelihood.

2.2 Aggregation

After the transformation process described in section 2.1, well-level data was obtained by calculating the mean for each feature of single-cell feature measurements per well. Further, the actual number of cells in a well after removing cells touching image borders were calculated. Before the succeeding phases of data processing, an aggregated data was then cleaned by dropping invalid wells containing fluorescent compounds. Also, two features whose measurements were dominated by zeros were excluded to avoid infinity results in later calculations. The resulting aggregated data for each cell line had 14688 wells with 460 features (figure 2).

2.3 Normalization

To account for plate-to-plate variability, all features of the sample treatments were normalized by re-expressing them as z-scores of the corresponding feature values observed in the DMSO control. The z-score transformation was calculated by subtracting the mean of DMSO control measurements of a feature in a plate and dividing by the pooled standard deviation of that feature across all plates.

2.4 Feature selection

To select a set of non-redundant and highly informative features, the Minimum Redundancy - Maximum Relevance (MRMR) feature selection [8,9] procedure was used. MRMR selects a set of 'optimal' features by simultaneously optimizing two conditions. The first condition quantifies the relevance of the feature to the classifying treatment variable - that is the mutual information between the feature and the treatment variable - using an F-statistic (calculated as the ratio of variance between treatments and the pooled variance within treatments). A feature with the strongest differentiating ability for the treatments has a

corresponding large F-statistic value, is ranked highest and therefore chosen for inclusion in the set of optimal features. The second condition is aimed at minimizing feature redundancy by selecting a feature that is maximally uncorrelated to the ones already selected. In this MRMR procedure, we identified a subset of features with low pairwise correlation and high reproducibility among replicates. In this regard, the approach aimed at identifying a feature subset that separated pairs of replicates (high correlation between pheno-signatures) from pairs of non-replicates (low correlation between pheno-signatures) as much as possible.

Figure 3 shows a schematic set of steps in identifying a set of features with optimal separation between treatments. For a given number of features selected, Pearson correlation coefficients [27] between treatment replicates and non-replicates are calculated. This results in two distributions, one for treatment replicates and another for non-replicates. An ROC curve quantifying the separation between the two distributions is then constructed and a corresponding AUC value calculated. This procedure is repeated for all sets of features between 2 and 75. An optimal number of features is then determined such that these features maximize the separation of treatment replicates from non-replicates within 1 standard error of the AUC .

2.5 Active calling

Each treatment sample was classified as active or inactive on the basis of the Euclidean distance to the center of the DMSO controls. The threshold for classification was calculated as the 95^{th} percentile of the distribution of the distance of the DMSO samples to the DMSO center. For every treatment, the percentage of active replicates was calculated and treatments with at least 50% active replicates were considered active, otherwise, the treatment was considered as inactive.

A pictorial representation of the analysis matrix after all data processing steps for an untransformed second cell line data is shown in figure 4. Only actively-called treatments were considered for analysis. All selected features were used in AUC analysis, while the top 10 ranked features from MRMR were used for calculating the Hotelling's T^2 statistic.

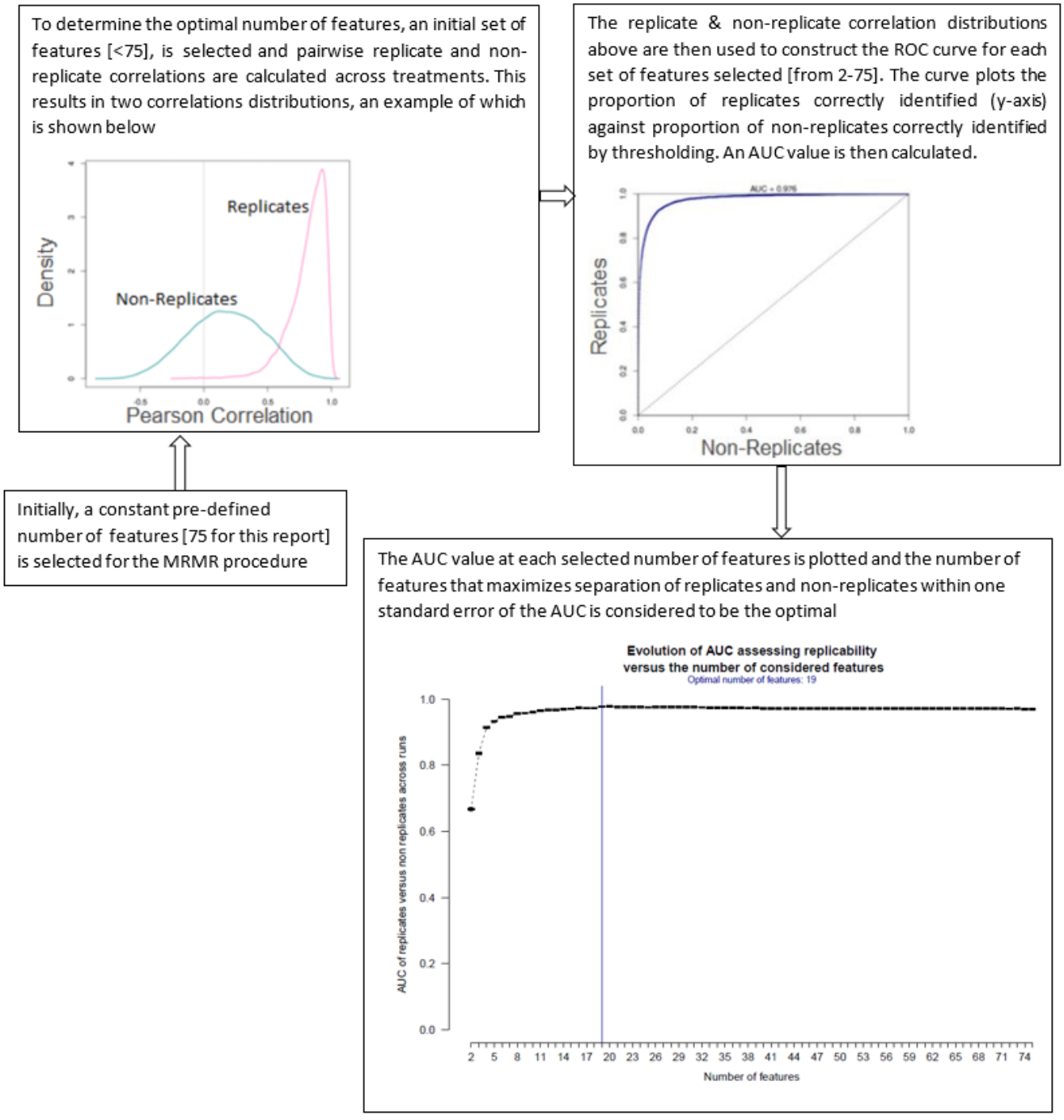


Figure 3: Feature selection schema

		Features			
Treatment	WELL_ID	Ratio_Nuc2Cyto_GFP_TotalInt	Cyto_NucCytoBFP_SERDark_0	.....	Cyto_RFP_TotalInt
cmpd101 @ 0.334 uM	11565590	0.931385212	1.404918838	.....	-3.299679747
cmpd101 @ 0.334 uM	11548694	-0.773279641	0.439804662	.....	-1.710712992
cmpd101 @ 0.334 uM	11549462	-0.024981534	-0.024499729	.....	-1.942772197
cmpd101 @ 0.334 uM	11565206	1.235073745	0.811573928	.....	-2.194526306
				.....	
cmpd230 @ 1 uM	11565957	0.824269512	1.169156578	.....	-2.222119054
cmpd230 @ 3 uM	11537092	-1.723239019	-0.34810657	.....	0.551858987
cmpd230 @ 3 uM	11549060	-0.260310018	0.608166677	.....	-1.299028855
cmpd230 @ 3 uM	11536324	-1.028556349	0.063148497	.....	0.496614392
cmpd230 @ 3 uM	11548676	0.108299414	2.03015273	.....	-0.65833225
				.....	
cmpd738 @ 9 uM	11565643	4.931418715	3.43754401	.....	-5.057760657
cmpd738 @ 9 uM	11565259	4.052273099	3.140591665	.....	-4.936178196
cmpd738 @ 9 uM	11537163	3.172814264	0.548239937	.....	-2.83115809
cmpd738 @ 9 uM	11536395	4.858004098	3.32898876	.....	-3.00825851
cmpd738 @ 9 uM	11566027	4.954179227	3.309243612	.....	-5.740644216
				.....	
DMSO @ 0 uM	11568871	-0.814475238	-1.424893332	.....	1.478200021
DMSO @ 0 uM	11563999	0.582510101	0.972662103	.....	-1.019464192
DMSO @ 0 uM	11548447	0.03602821	0.947501438	.....	0.685644676
DMSO @ 0 uM	11565487	0.080910187	-0.117160938	.....	0.302009268
DMSO @ 0 uM	11550895	-0.480077529	0.297963749	.....	-1.157415332

Figure 4: Data matrix for analysis

3 Methodology

To measure improvement in treatment separability before and after glog transformation, two approaches were used. The first method used the two-sample multivariate Hotelling's T^2 statistic, while the second method applied the area under the Receiver Operative Characteristic (ROC) curve approach, referred here as the AUC approach.

3.1 Hotelling's T^2 statistic

The two sample Hotelling's T^2 statistic [11, 28] measures the difference in two multivariate means and was used here to assess the improvement in the differences between mean vectors for pairs of actively-called treatments before and after the glog transformation. Assuming that replicates from both treatment populations are independently sampled, that both treatment populations are normally distributed with associated population mean vectors for the respective treatment populations being μ_1 for the first treatment and μ_2 for the second treatment in a treatment pair, respectively estimated with sample mean vectors $\bar{X_1}$ and $\bar{X_2}$, and having respective population variance-covariance matrices Σ_1 and Σ_2 (more so that $\Sigma_1 = \Sigma_2 = \Sigma$) which are respectively estimated by sample variance-covariance matrices S_1 and S_2, the Hotelling's T^2 statistic is calculated using the formulation shown in equation 2.

$$T^2 = (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)\prime\{\boldsymbol{S}_p(\frac{1}{n_1} + \frac{1}{n_2})\}^{-1}(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2) \tag{2}$$

where;

$$\bar{\boldsymbol{X}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\boldsymbol{X}_{ij}, \quad n_i \text{ is the number of replicates for treatment } i,$$

$$\boldsymbol{S}_i = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(\boldsymbol{X}_{ij} - \bar{\boldsymbol{X}}_i)(\boldsymbol{X}_{ij} - \bar{\boldsymbol{X}}_i)\prime,$$

$$\text{and} \quad \mathbf{X_{ij}} = \begin{Bmatrix} X_{ij1} \\ X_{ij2} \\ . \\ . \\ . \\ X_{ijp} \end{Bmatrix} \quad \text{is the treatment-replicates data matrix}$$

with X_{ijk} representing the k^{th} feature of replicate j from actively-called treatment i, $i =$

$1, 2,\ \ j = 1, 2...$*total replicates for treatment* i & $k = 1, 2...$*total number of features selected*.

Since both S_1 and S_2 estimate Σ under the homogeneity assumption, they were pooled together to provide a more efficient estimator of Σ. This estimate is the pooled variance, denoted as S_p in equation 2 and calculated using the formula below:

$$S_p = \frac{(n_1 - 1)\boldsymbol{S}_1 + (n_2 - 1)\boldsymbol{S}_2}{n_1 + n_2 - 2}$$

The resulting T^2 statistic is distributed as $\frac{(n_1+n_2-2)p}{n_1+n_2-p-1}F_{p,n_1+n_2-p-1}$, where F_{p,n_1+n_2-p-1} denotes a random variable with p and $n_1 + n_2 - p - 1$ degrees of freedom (p being the dimension of the response matrix). If the observed statistical measure of distance between the two treatment mean vectors - T^2- is too large, that is, mean vector μ_1 is too different from μ_2, the null hypothesis $H_0 : \mu_1 = \mu_2$ is rejected and the alternative hypothesis $H_1 : \mu_1 \neq \mu_2$ is therefore concluded. We note here that conclusion of the alternative hypothesis implies that the treatment pairs are well separated from each other.

Hotelling's T^2 statistic and the research questions

To use the Hotelling's T^2 statistic to answer the research questions, further restrictions were imposed due to the limited number of replicates for most of the treatments in both sets of data. Since most of the actively-called treatment pairs have at most 9 replicates (and therefore 18 observations in total), the number of features selected for T^2 had to be strictly less than 18. Further, for every pair of treatments compared, 2 degrees of freedom were lost in estimating the mean vectors leaving 16 degrees of freedom as the rank of residuals. Simulation studies also show that to be able to maximize on the power of MANOVA based statistics, the number of response variables has to be sufficiently less than error degrees of freedom [29,30]. Therefore, the top 10 ranked features from the feature selection procedure described in 2.4 were used for calculating the T^2 statistic for untransformed and transformed datasets.

To measure improvement in reproducibility of replicates as well as separation amongst replicates of a treatment compared to non-replicates, T^2 statistics were calculated for both transformed and untransformed data sets and their distribution and distributional statistics investigated. Since treatments were mutually expected to be different regardless of whether they come from the same compound or not, pairwise T^2 statistics were calculated for all the treatment pairs qualified as active via active-calling. In total, for N treatments chosen as active, pairwise treatments combinations resulted in $\frac{N(N-1)}{2}$ T^2 statistics. Then a distribution of the T^2 statistics plotted for every transformation was compared against the distribution obtained for the untransformed dataset. A shift to the positive of the distri-

bution of T^2 statistic for a transformation compared to the untransformed result signified improvement in the reproducibility of the data for that transformation. Boxplots were also plotted, with effective transformations showing larger boxplot statistics.

In using T^2 results to determine whether a transformation performs better on single-cell or aggregated data, a comparison of T^2 statistics calculated from transforming single-cell data and transforming data after aggregation using the same transformation parameters was done. Similar guidelines to inferences were done as described in the preceding paragraph. In general, the following was the procedure followed to calculate and use the T^2 for assessing the effect of transformation(s) on the separation of treatment replicates from non-replicates.

Hotelling's T^2 procedure

i. Consider the N treatments actively-called both before and after transformation. Each treatment had a minimum of 9 (and a maximum of 45) replicates

ii. The 10 top-ranked features with strongest differentiating ability are chosen by feature selection for both the transformed and untransformed data. These two sets of top-ranked selected features were not necessarily the same

iii. The Hotelling's T^2 statistic for every possible pair of treatments in both data is calculated. This results in $\frac{N(N-1)}{2}$ T^2 statistics for untransformed and transformed data

iv. A distribution for T^2 statistic is plotted for the untransformed and all transformations to assess the transformational effect on the separation of treatment replicates & non-replicates. Boxplots of T^2 statistic for transformed and untransformed cases were also plotted

v. A transformation resulting in larger T^2 statistics implied improved separation of treatment replicates and non-replicates.

To account for multiplicity and control familywise error rates, the Bonferroni correction was used to determine a significance level for a difference between mean vectors of a pair of treatments. To obtain a 0.05 familywise significance level for N actively-called treatments, a Bonferroni corrected significance level was calculated as $\frac{0.05}{0.5*N(N-1)}$. Where $0.5*N(N-1)$ is the total number of tests under consideration.

3.2 Area Under the ROC Curve (AUC) approach

The area under the receiver operating characteristic curve (AUC) is a measure for classification accuracy that is extensively applied in various fields of research. The ROC - AUC

literature and merits have been broadly discussed [31–35] and is contextually described here as a measure for evaluating the effect of the glog transformations in relation to the separation of treatment replicates from non-replicates. Following the feature-selection and active calling for treatments in a given glog transformation (and untransformed case also) process as described in sections 2.4 and 2.5, the corresponding AUC value was calculated through a two-step procedure as described in subsections below.

Replicates and non-replicates correlation and distribution

Initially, Pearson correlation coefficients in terms of the selected features were computed between pairs of replicates and pairs of non-replicates of a set of actively-called treatments of a transformed (or untransformed) data set to provide a measure of similarity. This establishes two distributions of correlation values (one for pairs of replicates and a second for pairs of non-replicates). Correlation between pairs of replicates is expected to be high, whereas correlation between non-replicates is expected to be low. The separation between these two distributions therefore provides a measure of how well treatments can be discriminated.

Construction of ROC curve and the AUC

To quantify the separation between the two distributions, an ROC curve was constructed (as described in schematic diagram 3) using a series of correlation thresholds and the area under the curve (AUC) calculated, resulting in a single valued measure of separation. At a given threshold value t, the proportion of correctly identified replicates (y-axis) and the proportion of correctly identified non-replicates (x-axis) formed a point on the ROC curve.

AUC statistic and the research questions

To use the AUC approach in investigating the effect of the range of glog transformations on the separation between treatment replicates and non-replicates, a random sample of treatments was selected 10 times and corresponding AUC values computed for each transformed data set and its untransformed data determined by the dimension of common actively-called treatments. A transformation that leads to better separation between replicates and non-replicates would have in general higher values for the AUC across the random samples and thus numerically larger boxplots statistics than those observed in the untransformed case. A composite plot of all the AUC values across runs for transformed and untransformed scenarios was presented. Similarly, to compare the performance of a transformation on single-cell versus aggregated data, AUC values were calculated on data transformed at single-cell level and compared to AUC outputs for well-level transformed data. Similar intuition used to conclude effectiveness of a transformation compared to the untransformed case was applied in concluding effectiveness of transformation at single-cell or aggregated

data levels.

3.3 Statistical softwares

All the data management and analyses were done using R 3.4.0 software [36] and its in-built packages. In the MRMR procedure available in the MRMR package, the functions **run_mRMR** and **evaluationAUC** were used for feature selection and calculation of the AUC value for treatment separation respectively. The Hotelling's T^2 statistic was calculated using the **hotelling.test** function available in the Hotelling package in R.

4 Results

4.1 Exploratory data analysis

4.1.1 Treatment replicates

In each data set, 1192 (95.208%), 28 (2.236%) and 32 (2.556%) treatments had 9, 36 and 45 replicates respectively. DMSO control was replicated across 1512 wells. The large number of treatments with 9 replicates meant that for application of Hotelling's T^2 statistic as a measure of separation between treatment replicates and non-replicates, a limited number of selected features were used so as to maximize the power of the statistic. To safeguard against using a larger number of features as compared to the rank of residuals for any arbitrarily chosen pair of treatments therefore, 10 highest ranked selected features were used in calculating the T^2 statistic.

4.1.2 Cell and well level calculated mean α values

Figure 5 and table 1 (in Appendix) respectively show the ratio and actual calculated mean α values for all features in both data sets. Except for a few features where the calculated α values had relatively huge differences when calculated at well level compared to cell level, most of them were close. 2 (0.44%) of the features in the first data set and 17 (3.7%) for the second data set had α mean differences beyond 5% of each other in absolute values. The difference in mean α values is a result of reduced variability due to aggregation at well-level as compared to high variability observed at cell levels. Feature *Nuc_NucCytoBFP2RFP_PearsonCor* for the first data set had the highest ratio at 1.13 between α values calculated at cell and well levels. *Cell_NucCytoBFP2GFP_PearsonCor* resulted in the lowest ratio of 0.914. In the second data, *Cell_NucCytoBFP2GFP_PearsonCor* (*Cell_NucCytoBFP2RFP_PearsonCor*) had the highest (lowest) ratio of 1.217 (0.707). These features, which quantify correlation between cell and respective nuclear (or nuclear to nuclear for *Nuc_NucCytoBFP2RFP_PearsonCor* in first data) cytoplasms, were highly variable at cell level compared to when aggregated at well level. Features with high ratios between cell- and well-calculated α values were however few. Comparison between α values calculated at cell and well level was mandatory since very different αs would end up comparing different glog transformations even if applied using the same λ parameter.

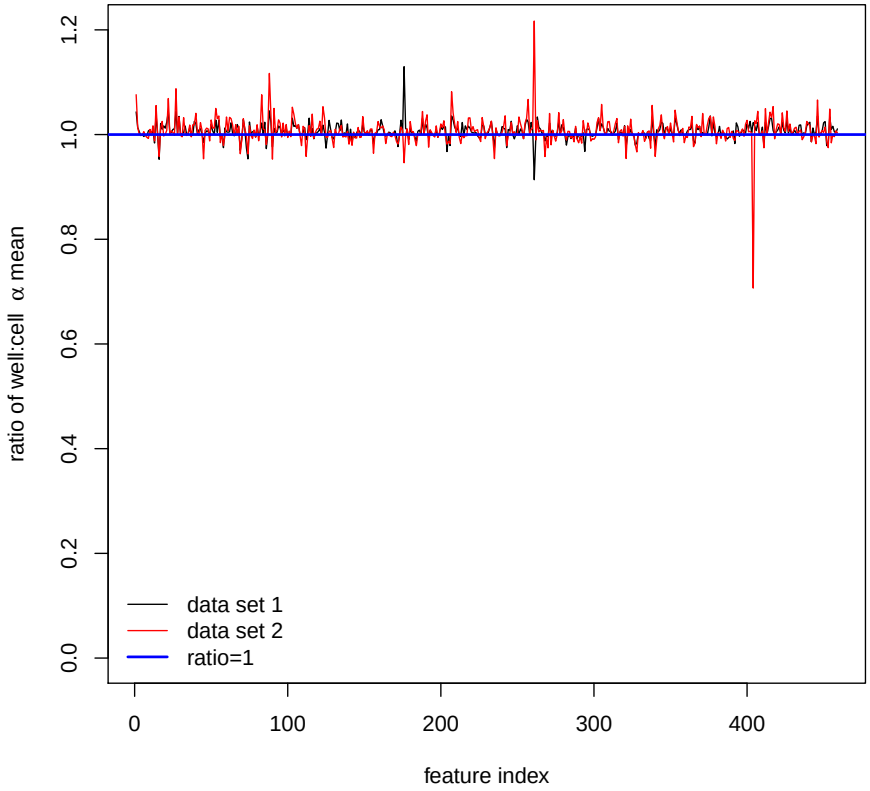
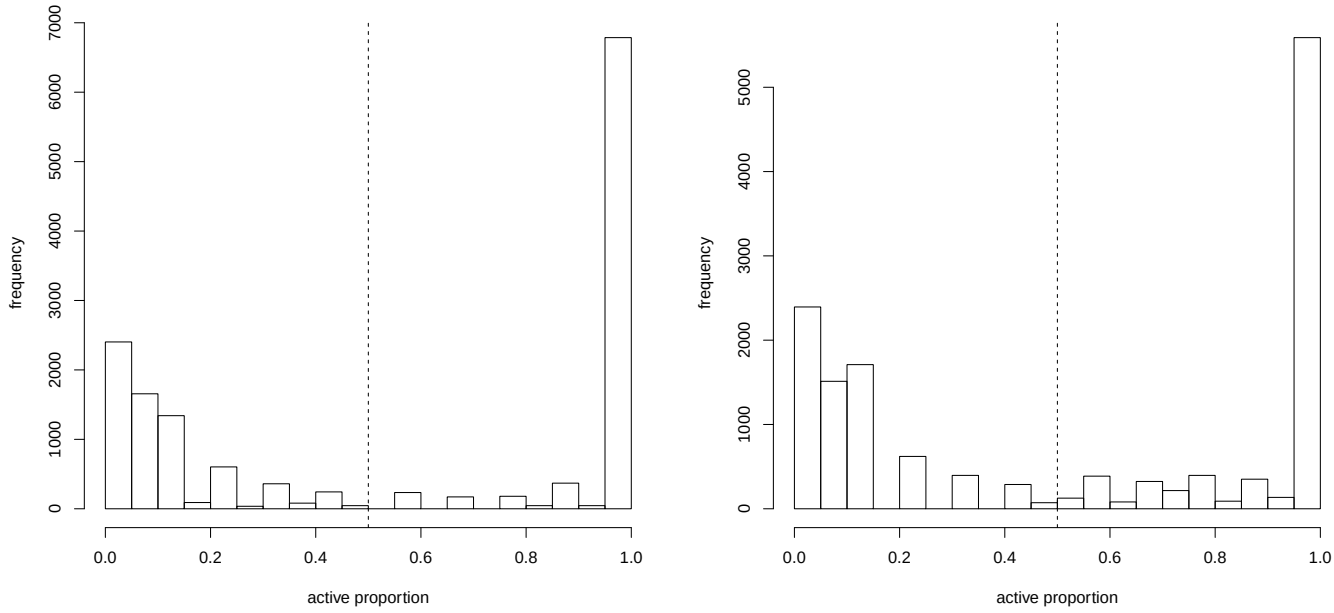


Figure 5: Ratio of calculated α values at well and cell level.

4.1.3 Actively-called treatments for untransformed data sets

Figure 6 shows graphic representations for categorization of active and inactive treatments. The dotted vertical line shows the cut-off point (0.5 mark). Treatments with more than 50% of replicates significantly different from DMSO controls - with significance from DMSO controls determined using the Euclidean distance to the center of DMSO - appear to the right of the cut-off point and are concluded as active. Treatments with less than 50% of replicates significantly different from the DMSO controls appear to the left of the cut-off point and were concluded as inactive. For instance, treatments with an active proportion of 1.0 imply that all replicates were different from DMSO controls. Treatments with an active proportion of 0.0 conversely had all replicates that were not different from DMSO controls. The number of replicates for all treatments at a given proportion is shown on the

y-axis. In total, 707 (56.42%) treatments in the first and 650 (51.88%) in the second data set were active.



First data set Second data

Figure 6: Proportion of actively-called treatments for untransformed data

4.2 Effect of transformations on separation of treatment replicates and non-replicates

4.2.1 Hotelling's T^2 results

To assess improvement in separation of treatments due to transformation, T^2 statistics were calculated for untransformed and transformed data sets using actively-called treatments present both pre- and post-transformation. Only glog transformations with λ values 0.1 and ranging between 0.5 to 25 at intervals of 0.5 were computed and plotted for due to time, computations involved and ease of presenting results obtained. The two sets of T^2 statistics (one for untransformed and another for transformed) were plotted in each combination of transformed and untransformed data sets. As a result, each transformed data had its own corresponding untransformed data set, which was determined by the number of actively-called treatments existing in both data sets. A transformation with its T^2 statis-

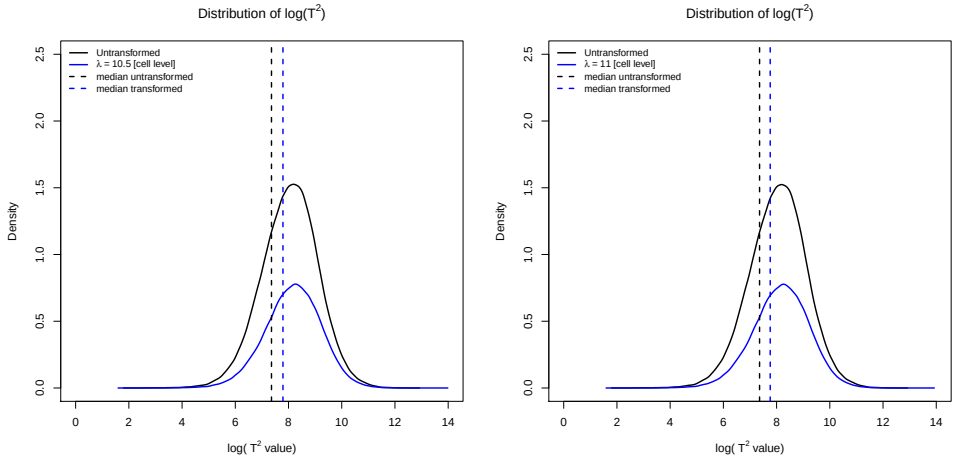
tics distribution positively shifted of its untransformed T^2 distribution implied improved separation for treatments compared to when data was untransformed (figures 19a to 19d in appendix). This is because large T^2 statistics point to greater distance between treatment pairs, have a higher chance of being classified as significant and thus better treatments separation. Figure 18 (in appendix) shows boxplots for $log(T^2)$ for each transformed data and its corresponding untransformed data.

Presence of outlying statistics both before and after transformation was evident, with differences between boxplots at every λ value often obscured as a result. However, as observed in figures 19a to 19d, there were some transformations that led to small improvements in treatments separation judging from shifts in their T^2 distribution medians. In particular, for transformations of $\lambda = 10.5$ and $\lambda = 11$ for the first and $\lambda = 0.1$ and $\lambda = 0.5$ for the second data set respectively, figure 7 shows distributions of their $log(T^2)$ and median locations compared to respective untransformed cases. For these transformations, the resulting medians of their $log(T^2)$ were slightly positively-shifted compared to untransformed cases. These transformations therefore led to marginally better separation of treatment replicates from non-replicates. Figure 20 on the contrary shows some transformations that either performed poorly or equally as in untransformed cases. All their transformed $log(T^2)$ medians were either equivalent or negatively shifted in comparison with untransformed distributions.

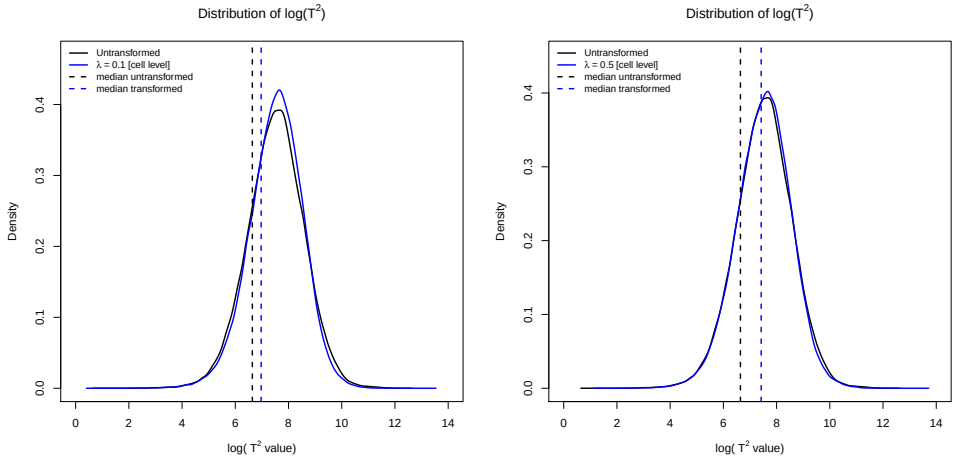
Figure 21 (in appendix) shows increase in the proportion of T^2 statistics concluded as significant when an effective transformation was applied. Proportion of T^2 statistics concluding that treatment pairs were not significantly different from each other dropped from *49.82 % to 45.93 %* when a transformation with $\lambda = 10.5$ was applied to the first data while for $\lambda = 0.5$, the proportion increased from *62.39 % to 62.79 %* when applied to the second data set. For $\lambda = 10.5$, the number of actively-called treatments was 703 compared to 707 for untransformed case in the first data set. Of these treatments, 683 were active both before and after transformation. Consequently, 45.93% of 232903 (i.e. 106968) treatment pairs were concluded to be non-significantly different compared to 116021 when untransformed. Therefore, a glog transformation of $\lambda = 10.5$ on the first data set led to 9053 more treatment combinations being classified as significantly different. Similarly, a $\lambda = 0.5$ transformation on the second data set resulted in 662 active treatments compared to 650 when untransformed. Of these, 642 treatments were actively-called both pre- and post-transformation. This was associated with 129198 non-significant T^2 statistics compared to 128371 non-significant T^2 when untransformed. This transformation was as a result credited with an decrease of 827 significant T^2 statistics.

Boxplots for both sets of data in figure 18 showed that differences between transformed and untransformed medians of $log(T^2)$ statistics observed in figure 7 were mainly due to outlying values and that these differences were very small in themselves. Further, effects of transformations on proportion of treatment pairs concluded as significantly different were

data-dependent. In the first data, a transformation with $\lambda = 10.5$ which was associated with a marginal improvement in treatments separation increased the proportion of significantly different pairs by 9053. In contrast, transformation with $\lambda = 0.5$ which also seemed to slightly improve treatments separation in the second data decreased the proportion of significant treatment pairs by 827. The Hotelling's T^2 statistics method consequently suggests that the transformations failed to substantially and consistently improve separation between treatment replicates and non-replicates beyond separation observed before applying transformations.



Transformation on first data set ($\lambda = 10.5$) Transformation on first data set ($\lambda = 11$)



Second data set ($\lambda = 0.1$) Second data set ($\lambda = 0.5$)

Figure 7: Distribution of $\log(\mathrm{T}^2)$ statistics after transformation compared to untransformed cases

Proof of treatment separation concept

To illustrate that glog transformations associated with marginal improvements in separation between treatment replicates and non-replicates did indeed lead to some enhanced separation between treatment profiles, pairs of treatments that resulted in large differences between the Hotelling's T^2 before and after transformations were plotted for both data sets. In the first data set, pairs of treatments whose difference in pre- and post-transformation T^2 increased by more than 250000 units were plotted under a glog transformation with $\lambda = 10.5$. Four uniquely different treatment pair combinations- *cmpd193* @ 9 µM and *cmpd306* @ 1 µM, *cmpd275* @ 3 µM and *cmpd469* @ 0.334 µM, *cmpd370* @ 0.334 µM and *cmpd694* @ 9 µM, and, *cmpd807* @ 9 µM and *cmpd815* @ 3 µM - had differences in T^2 higher than 250000. 19 features were selected before transformation while 21 were selected after transformation. Four features were commonly selected both before- and after-transformation. The respective paired treatment profiles are shown in figure 8.

In line with earlier results obtained when discussing T^2 statistics of assessing improvement in treatments separation before and after transformation, treatment profiles where there were huge increases in the value of the T^2 statistics remained largely similar. For comparison of *cmpd193* @ 9 µM and *cmpd306* @ 1 µM treatment pairs in figures 8a and 8b for instance, the four features present before and after transformations showed mixed degrees of changes in profiling the treatments. For *Cell_NucCytoBFP_SERHole_1* and *Nuc_RFP_SERValley_1*, before transformation, treatment values for these features were very similar. However, after transformation and with a different set of optimal features, they appeared slightly better separated. *Nuc_GFP2RFP_PearsonCor* and *Nuc_NucCytoBFP_SERHole_2* features showed minor improvements in separation of treatments profiles even after combining them with a different set of features. Even though some of these specific features present before and after transformation individually profiled differently, the general treatments profiles were clearly not different. These shows that transformational effects on treatment separation was at bare minimum and most cases non-existent.

Three treatment combinations from the second data set - *cmpd279* @ 9 µM and *cmpd897* @ 1 µM, *cmpd735* @ 0.334 µM and *cmpd94* @ 9 µM, and, *cmpd749* @ 1 µM and *cmpd94* @ 9 µM- had differences above 200000 and are plotted in figure 9. Figures 9a and 9b through to figures 9e and 9f show that transformation was helpful in rescaling the data leading to better visualization and treatments separation. Before transforming, treatment profiles appeared to overlap for many features. These overlapping was reduced and in some cases not present after transformation. Analogous to results obtained where large changes in T^2 statistics were observed, figures 9g and 9h (also 22a and 22b in appendix) show that treatment pairs that were poorly separated pre-transformation likewise remained hardly changed after transformation and with different sets of selected features.

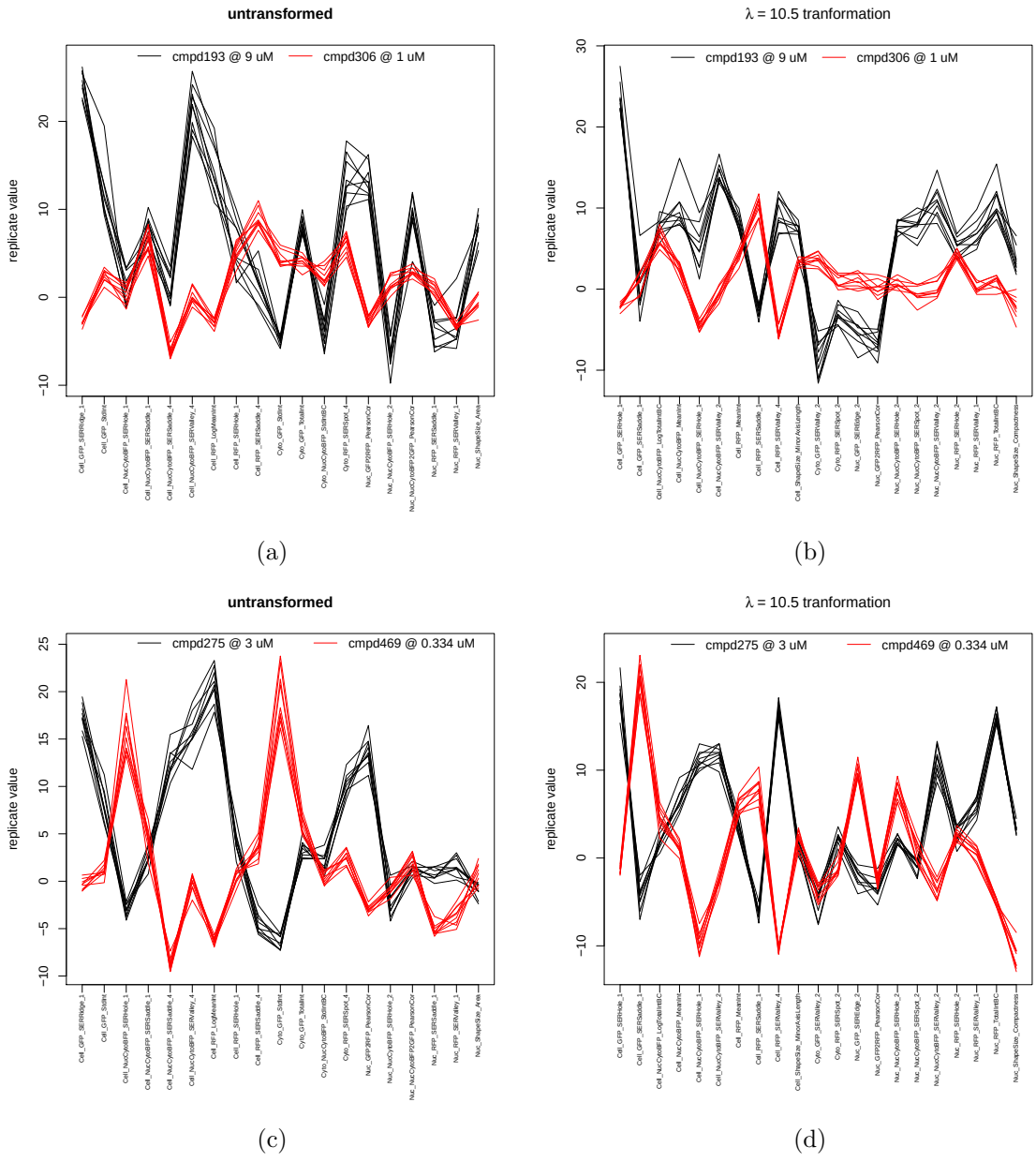
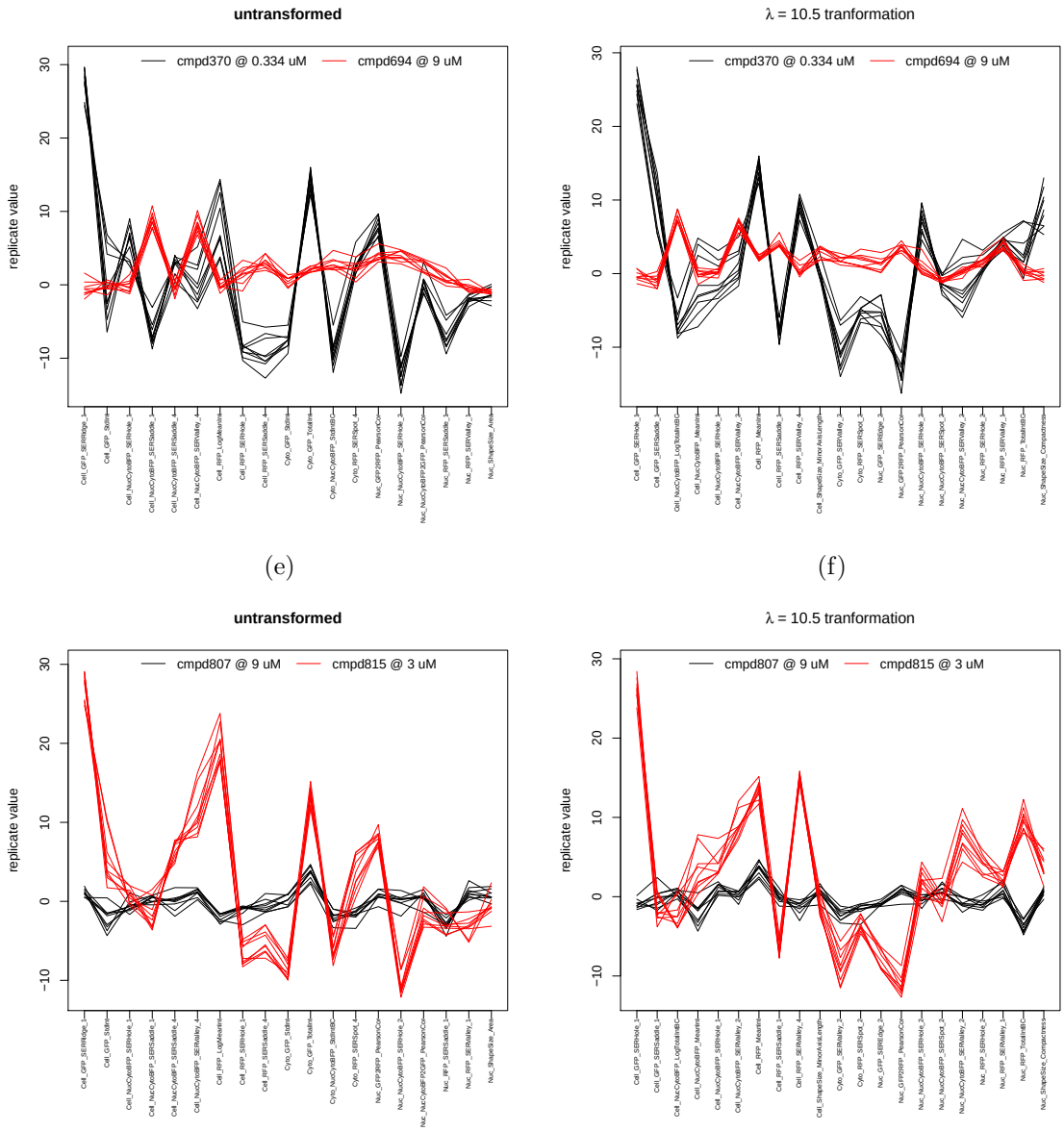


Figure 8: Treatment profiles before and after transformation (first data set, $\lambda = 10.5$)



(e) (f)

(g) (h)

Figure 8: Treatment profiles before and after transformation (first data set, $\lambda = 10.5$)

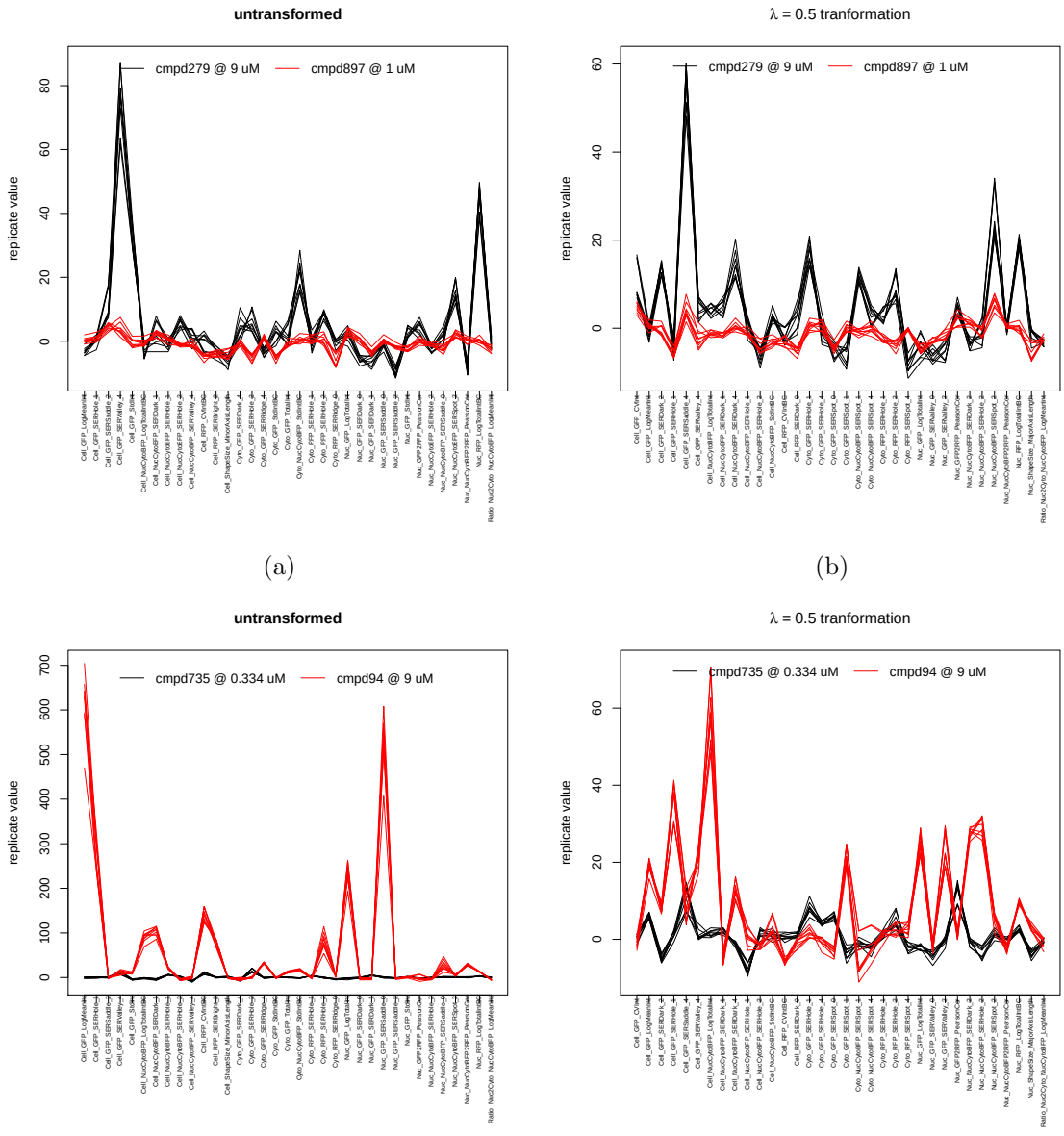
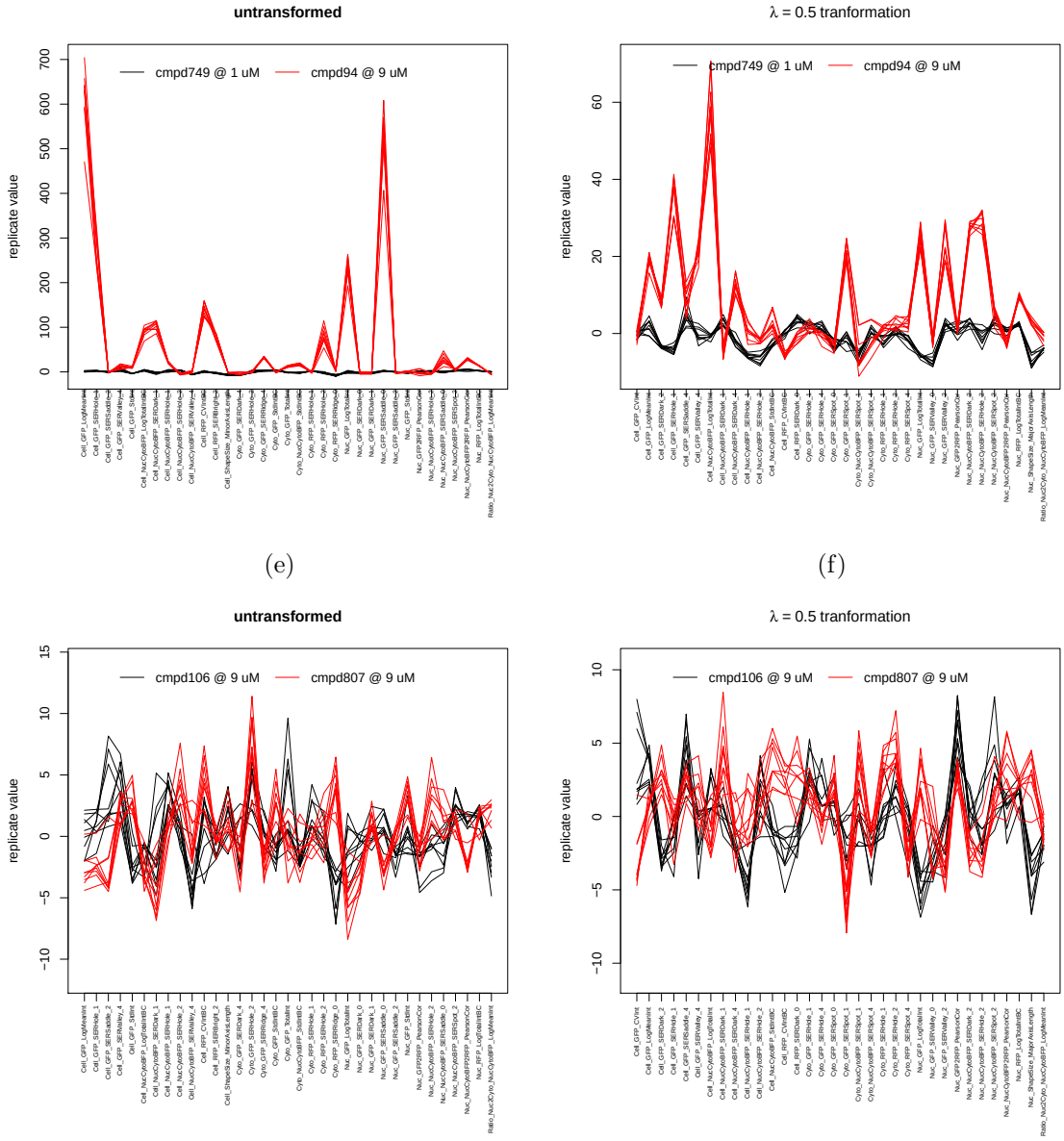


Figure 9: Treatment profiles before and after transformation (second data set, $\lambda = 0.5$)



(e) (f)

(g) (h)

Figure 9: Treatment profiles before and after transformation (second data set, $\lambda = 0.5$)

4.2.2 AUC results

AUC values were calculated using all features selected across all actively-called treatments common for both untransformed and transformed data sets as an alternative method of assessing the effect of transformations on separation of treatment replicates from non-replicates. The results for both cell lines were summarized using boxplots and are displayed in figure 10. Only λs equal to 0.1 and 0.5 to 25 at intervals of 0.5 were plotted for similar reasons as listed in section 4.2.1. Each transformation was compared to an untransformed subset data-set with the same actively-called treatments.

For all transformed and untransformed cases in the first data set, AUC values ranged between 0.9801 and 0.9848 with a median of 0.9828. Most boxplots for transformed data had slightly higher AUC values than corresponding untransformed cases. In a few transformations where untransformed cases showed higher AUC values compared to when transformed, the difference was not large as evidenced by values after transformation falling well within upper and lower quantiles of AUC values realized when untransformed. The respective differences were in general minimal. The minimal improvement in AUC values after transformation were due to effective treatments separation pre-transformation, all of which had AUC values above 0.98. Transformations of λ values 0.1, 6, 11 and 22 had distinctly elevated boxplot values. Although the differences were small, these transformations led to marginal improvements in treatments separation. Conversely, transformations with λ values 0.5, 10.5 and 22.5 showed very similar AUC values with their untransformed cases. They therefore had minimal effect on treatments separation beyond what was observed pre-transformation. It is important to note that improvement (or lack of) in treatments separation was highly sensitive to small changes in λ values.

AUC values for the second data set ranged between 0.9498 and 0.9633 with a median of 0.9562 across all transformed and untransformed cases. Similar to results for the first data set, treatments separation pre-transformation was highly effective. Therefore, improvement in separation of treatments as a result of transformation was limited. Unlike in the first data set however, where transformations either performed marginally better or equivalent to untransformed cases, in the second data set, all possible comparison results were observed. Some transformations led to slight improvements in treatments separation, others were just as effective as when untransformed while a few led to slightly lower separation. Still, these differences were small. Transformations with λ values at 3.5, 4.5 and 5 resulted in AUC values distinctly lower than in respective untransformed cases. These transformations were associated with lowered treatments separation compared with when data was untransformed. Transformations with λ values equal to 0.1, 15.5 and 17.5 led to insignificantly higher AUC values. They therefore resulted in slight improvement in treatments separation.

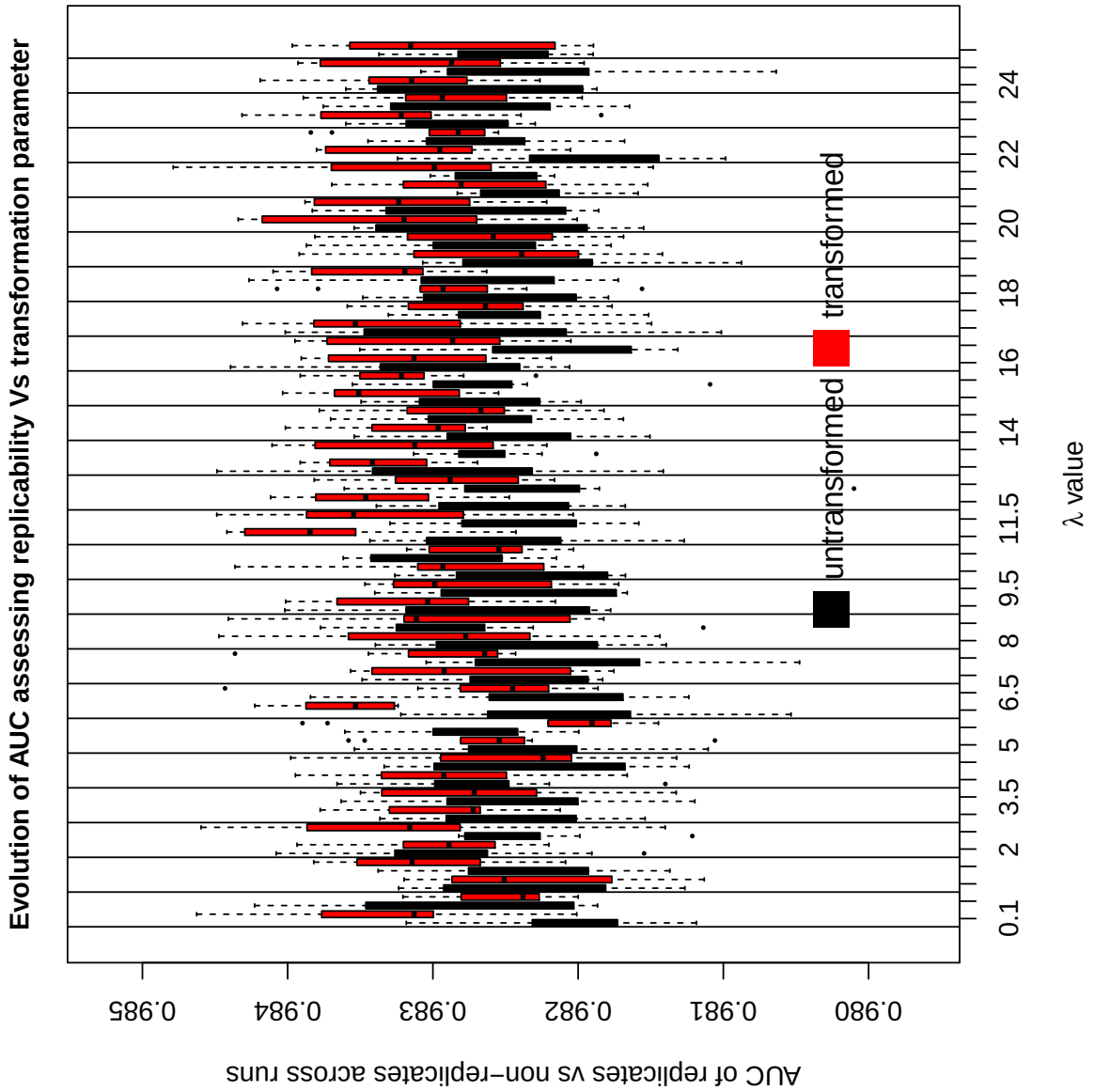
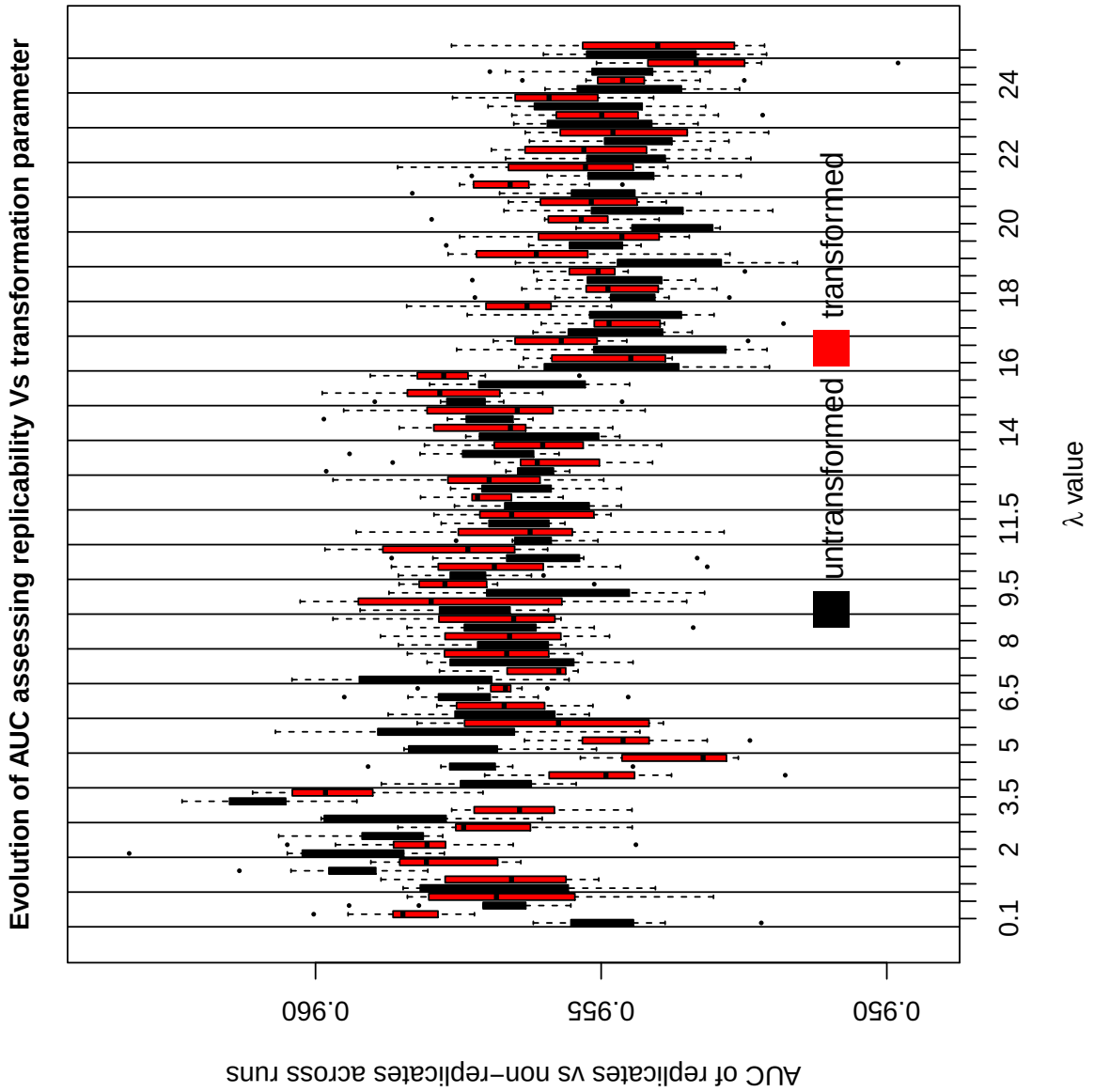


Figure 10: Boxplots of AUC values across runs for transformed and untransformed data sets



Second data set

Figure 10: Boxplots of AUC values across runs for transformed and untransformed data sets

In all cases, whether a transformation led to a small improvement, no improvement or in cases where treatments separation due to a transformation was poorer compared to when untransformed, it was evident from AUC boxplots and distributions of T^2 that transformations generally failed to markedly improve treatments separation beyond what was observed pre-transformation. Two factors were observed to have most likely contributed to a general lack of effectiveness of transformations in improving treatments separation. Firstly, transformations had virtually no effect on features distributions. Secondly, treatments separation before transformation was highly effective. These factors are delved into in sections below.

Effect of transformations on distribution of selected features

Before transformation, all features selected in both data sets showed lack of normality (figures 11 and 23 (in appendix)). Both positive and negative skewness were present. The non-normality in distributions intrinsically points to a lack of independence between the mean and standard deviation of each of the features. To investigate whether glog transformations improved features distributions, a pair of transformations that showed a slight improvement in treatments separation were applied in both sets of data. A transformation of $\lambda = 10.5$ was applied across all features in the first data set, while for the second data, a transformation of $\lambda = 0.5$ was similarly applied.

Results for distributions of some features that were not selected in the optimal set before but were selected after transformation are shown in figures 11 and 23 (in appendix). For the first data, 17 features not initially selected in untransformed case were selected after a glog transformation with $\lambda = 10.5$. In the second data, 18 features not selected pre-transformation were selected after a transformation of $\lambda = 0.5$. While a few features distributions clearly improved, distributions for most selected features showed little improvement towards normality after transformation. This was not surprising, however, since transformations were not tailored to address feature-specific lack of normality. Normality in response distributions is a cornerstone assumption for applying multivariate methods (particularly for T^2) and since after transformation there was selective improvement in normality, both pre- and post-transformation data remained sub-optimal for application of the Hotelling's T^2 methodology and resulted in similar statistics.

Differentiating ability of features selected

Figures 12a and 12b for the first data set, and 13a and 13b for the second dataset, show the discriminating ability of features selected using the MRMR procedure for highly unlike treatments and treatments with similar mechanisms of action. The features are selected on the untransformed data sets. In the first data set, 19 features were selected while 34

features were selected in the second data set. For both data sets, feature profiles for highly similar treatments are indifferent, while less similar treatment pairs show very different profiles. The features selected on these untransformed data show a high ability to differentiate treatments replicates from non-replicates based on treatment mechanisms of action. As a result, applying a transformation was unlikely to show more impressive treatments separation since excellent separation was already achieved on untransformed data by selecting a very effective set of features.

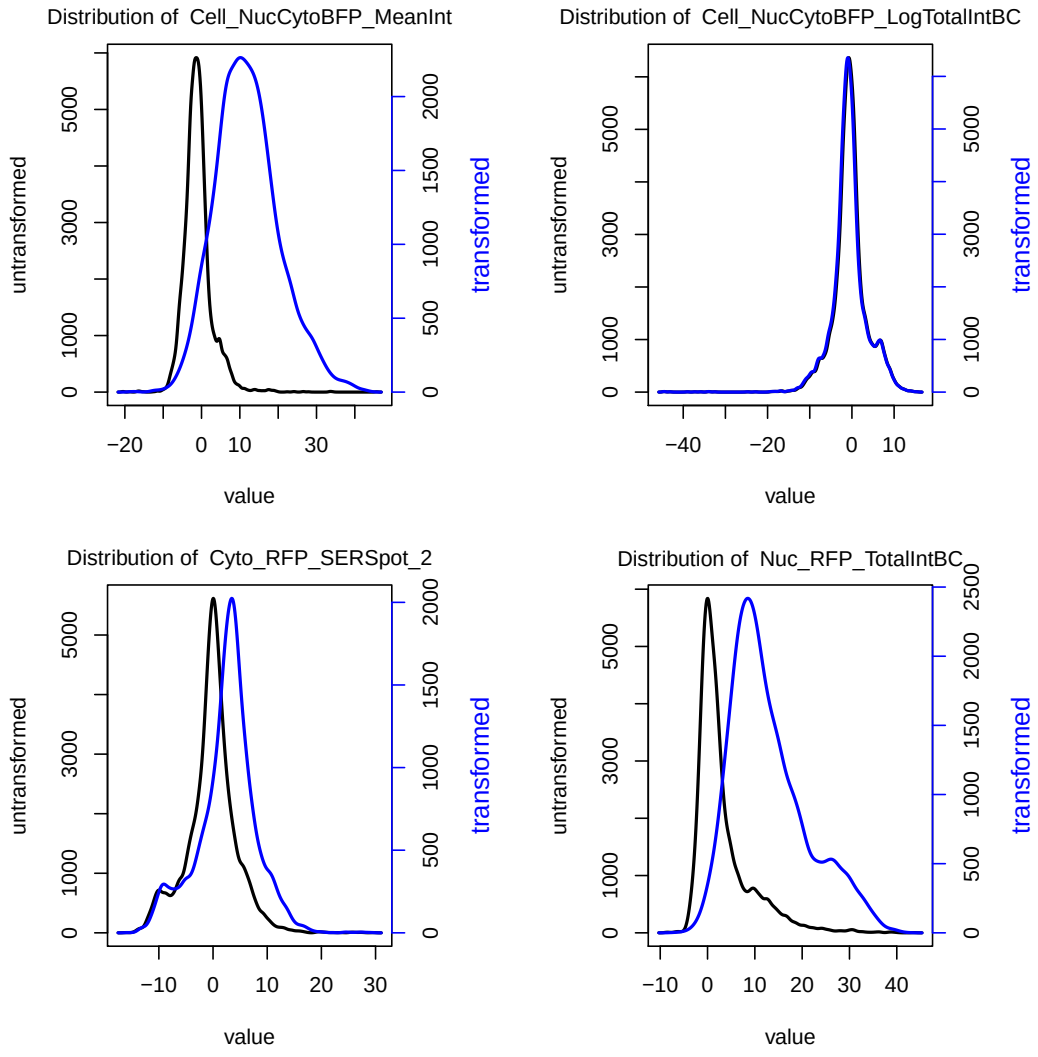
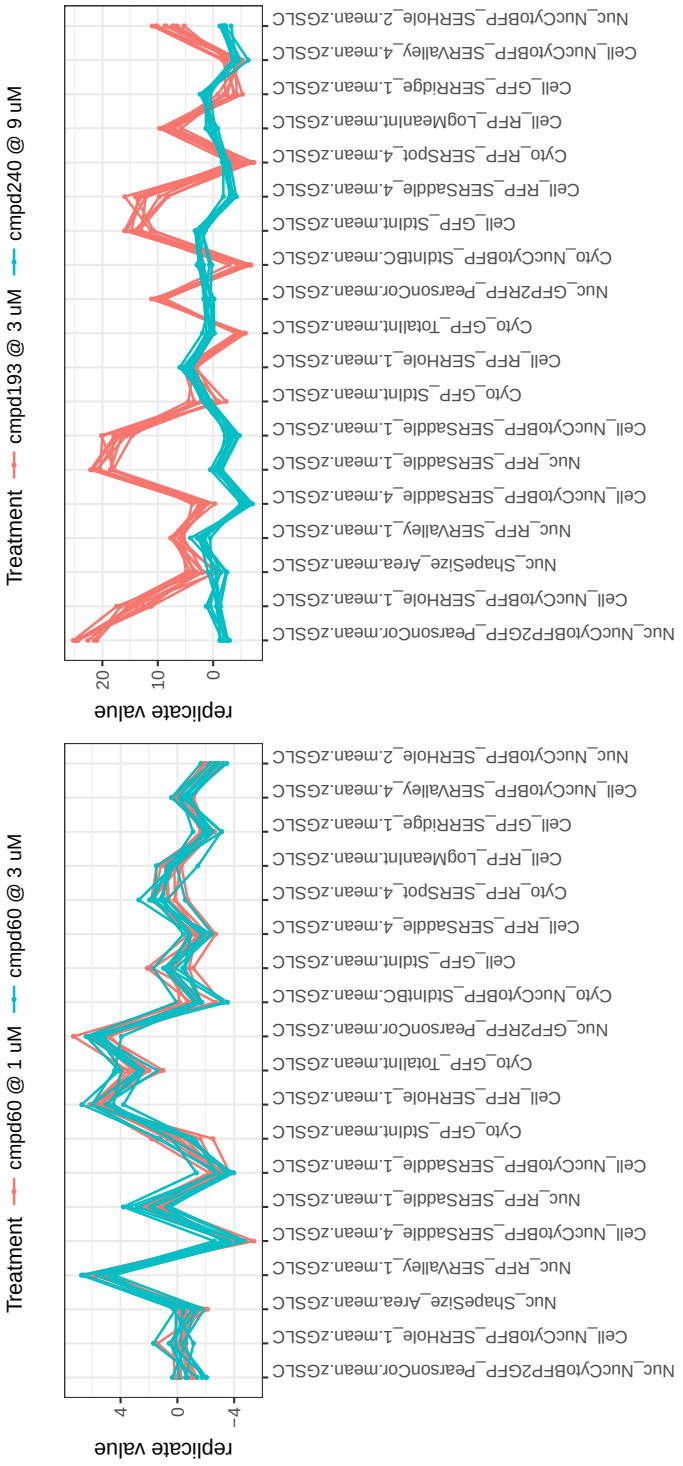


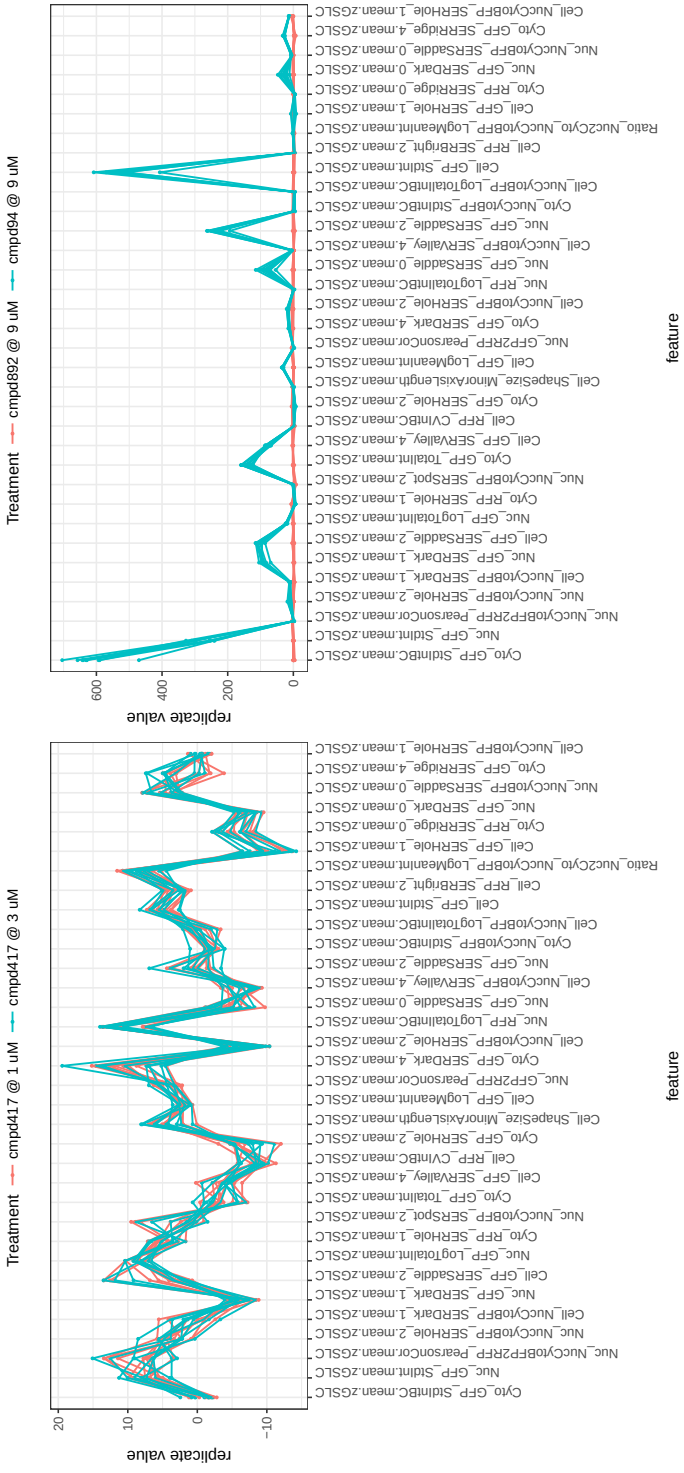
Figure 11: Data set 1 - feature distributions before (black line) and after (blue line) transformation ($\lambda = 10.5$)



(a) Similar treatments

(b) Less similar treatments

Figure 12: Differential ability of features selected- first data set

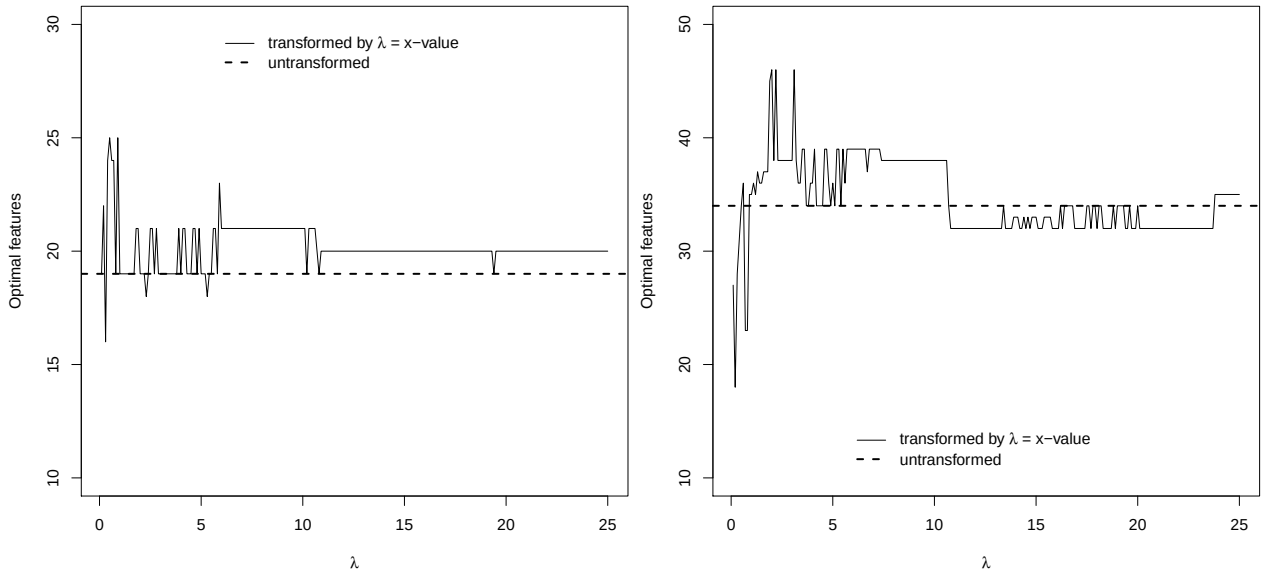


(a) Similar treatments (b) Less similar treatments

Figure 13: Differential ability of features selected- second data set

4.3 Effect of transformation on number of features selected and proportion of active treatments

Figures 14a and 14b show the effect of transformation on the number of features selected for the first and second data set respectively for all λ parameters investigated. For untransformed data, the optimal number of features was 19 for the first data set and 34 for the second data. The number of features selected for the first data set after transformation was generally higher compared to when untransformed. The minimum number of features selected was 16 for $\lambda = 0.3$ and maximum at 25 features for λ values 0.5 and 0.9.



(a) First data set

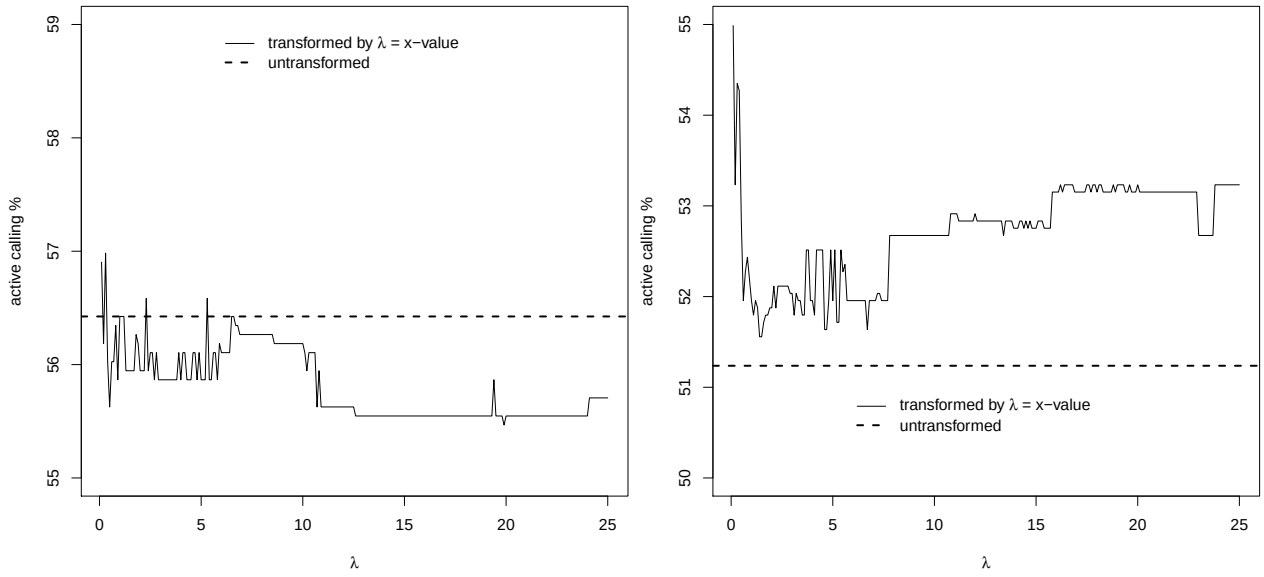
(b) Second data set

Figure 14: Effect of transformation on features selected

Unlike in the first data set where features selected after transformation were generally higher than features selected when untransformed, the number of features selected in the second data set showed patches of values highly sensitive to the range in which the transforming parameter λ falls. The minimum number of features of 18 was selected at $\lambda = 0.2$ while the maximum number was 46 selected for $\lambda = 2.2$ and $\lambda = 3.1$. Only λ values between 0.9 and 10.7, and greater than 23.8 consistently selected more than the 34 features selected for untransformed data.

Effect of transformation on proportion of treatment active calling

Figures 15a and 15b show the effect of transformation on the proportion of treatments concluded as active for the first and second data set respectively. Figure 15a shows that the proportion of actively-called treatments was largely below 56.42% achieved for the first untransformed data, except for λ values of 0.1, 0.3, 2.3 and 5.3 that respectively had proportions of 56.90% , 56.98% , 56.58% and 56.58% . Transformations that showed marginal improvement in separation of treatment replicates from non-replicates were in general associated with the highest decrease in proportions of actively-called treatments compared to untransformed data. For instance, transformation with $\lambda = 10.5$ resulted in a proportion of 56.11% while $\lambda = 11$ resulted in 55.63%. $\lambda = 19.9$ gave the lowest proportion at 55.467.



(a) First data set

(b) Second data set

Figure 15: Effect of transformation proportion of active calling

All transformations in the second on the other hand were associated with increased proportions of actively-called treatments as shown in figure 15b. In this case, transformations that appeared to perform relatively better in treatment separation compared to untransformed data generally had higher proportions of actively-called treatments. A case in point is $\lambda = 0.1$ that was associated with a proportion of 54.99% which was also the highest increase in proportion of active-called treatments observed. The minimum proportion of 51.56% was achieved with $\lambda = 1.4$. Observing the respective data-specific graphs for features

selected and proportion of active treatments also reveals a suspect functional relationship. For the first data set, the high number of features selected was in general followed by lower proportions of actively-called treatments compared to the untransformed case. The second data set on the hand, save for the rather weird observations for low λ values, splits the graphs into two phases. The phase with higher number of features selected compared to untransformed case clearly associates with lowered proportions of active treatments, albeit still above the proportion under no transformation. The second phase with less features selected was associated with a higher proportion of treatments qualified as active.

4.4 Performance of transformation when applied at cell and well levels

To investigate whether the glog transformation had improved treatments separation when applied at single cell or well level, both data sets were transformed using a select set of transformations that showed small improvements in separating treatments in comparison with untransformed cases when applied on single cell data. For the first data set, a corresponding transformation with $\lambda = 10.5$ and α mean values calculated on aggregated data was applied to well level data. For the second data set similarly, a transformation with $\lambda = 0.5$ was applied to well-aggregated data. Only treatments that were classified as active for untransformed, and cell- and well-level transformations for each data set were analyzed. The Hotelling's T^2 and AUC methods were then used to compare performance for untransformed and when transformations were applied at cell- and well levels. Distributions for the T^2 and boxplots for both the AUC and T^2 statistics were used and inferences done in a similar way as has been discussed in earlier sections. To assess reproducibility of results, further investigations were done for an extra set of transformations that performed marginally better than in untransformed cases at single cell level, and also a pair of transformations that performed indifferently to untransformed cases at single cell level. In that regard, to confirm that results using $\lambda = 10.5$ in the first data set were reproducible, transformation results for $\lambda = 11$ were also investigated. In the second data set, to check reproducibility of results for $\lambda = 0.5$, results for $\lambda = 0.1$ were also investigated. For the transformations that were indifferent at single cell level, transformations for $\lambda = 8.5$ and $\lambda = 10$ were investigated for the first data set. Similarly, transformations for $\lambda = 8.5$ and $\lambda = 10.5$ were investigated for the second data set as a pair of transformations that showed little improvement in the separation of treatments when applied at single cell level.

4.4.1 Hotelling's T^2 results

Figures 16a and 16b show distribution of logarithm of the Hotelling's T^2 statistic for a pair of transformations ($\lambda = 10.5$ and $\lambda = 11$) that minimally improved treatments separation on single cell first cell line data. These transformations showed small but non-significant improvements in treatments separation when applied at cell level compared to when applied

at well level. Box-plots comparing performance at cell- and well-level did not reveal any clear-cut improvement to results obtained in the untransformed case. For transformations that were indifferent to the untransformed cases at single cell level ($\lambda = 8.5$ and $\lambda = 10$) in figures 24a and 24b, transforming data at well level was associated with non-significantly reduced effectiveness in treatments separation compared to when untransformed or transformed at single-cell level.

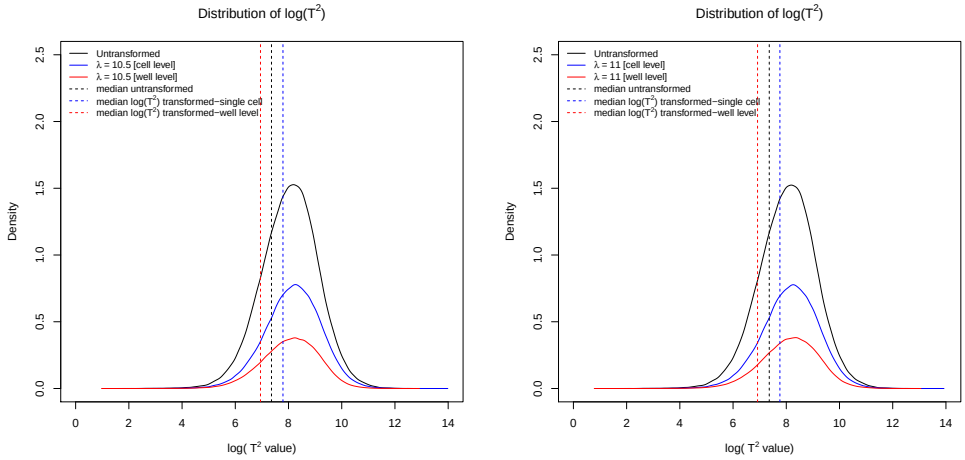
In the second data, similar conclusions were drawn for transformations that led to slight improvements in treatment separation at single cell level. Figures 16c, 16d show distributions for transformations with $\lambda = 0.1$ and $\lambda = 0.5$. For these pair of transformations, applying the glog transformation at single-cell level was associated with a negligible improvement in treatments separation. For transformations that led to no improvement in treatment separation on single cell data, transformations investigated were $\lambda = 8.5$ and $\lambda = 10.5$ and results are as graphically represented in figure 24c and 24d. They both showed minimal improvement in separation when transformations were done at well level compared to cell level.

In summary, when using the Hotelling's T^2 statistic, transformations that led to minimal improvements in treatment separation at single cell level were consistent in their output for both data sets. They all pointed to small improvements that were non-significant in treatments separation when applied on single cell data. However, applying transformations that failed to improve treatment separation over the untransformed data gave non-significant but data-dependent results. The first data set showed that for $\lambda = 8.5$ and $\lambda = 10$, transformations non-significantly improved separation between treatments when applied on single-cell data compared to when applied at well level. For the second data, however, transformations with $\lambda = 8.5$ and $\lambda = 10.5$ suggested non-significantly improved treatments separation when applied at well level.

4.4.2 AUC results

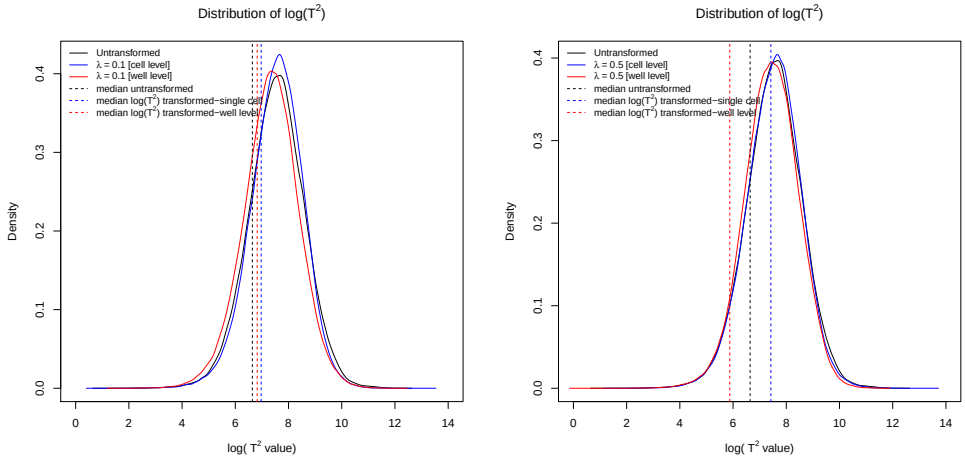
Transformations that showed improved treatments separation for the first data set (λ values 10.5 and 11) performed similarly when applied at cell or well level (figures 17a and 17b). Treatments separation when applied at $\lambda = 10.5$ lost all their marginal effectiveness observed at single cell level when considering only actively-called treatments across the untransformed, cell and well-level transformed data sets. At $\lambda = 11$, both cell- and well-level transformed data sets marginally improved treatments separation compared to when data was untransformed. However, neither of cell- nor well-level transformations was clearly better. Similarly, for transformations that failed to give any improvement in treatment separation at cell level (λ values 8.5 and 10), no clear improvement in treatments separation was visible between cell- and well-level transformations (figures 25a and 25b in appendix).

All transformations applied at single cell level separated treatments marginally better than when applied at well-level for the second data set (figures 17c and 17d, 25c and 25d in appendix). Effectiveness of transformations with λ values 0.5 and 8.5 however were indifferent to when untransformed for actively-called treatments across the three sets of data. Just as we have noted throughout this report, however, any slight improvements in treatments separation were very small and largely insignificant.



(a) *first data set* ($\lambda = 10.5$)

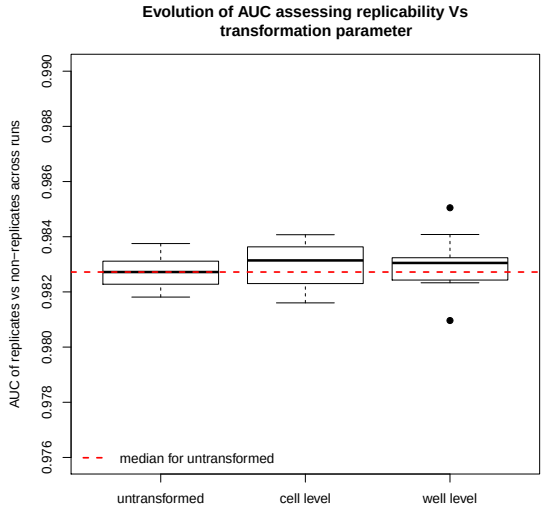
(b) *first data set* ($\lambda = 11$)



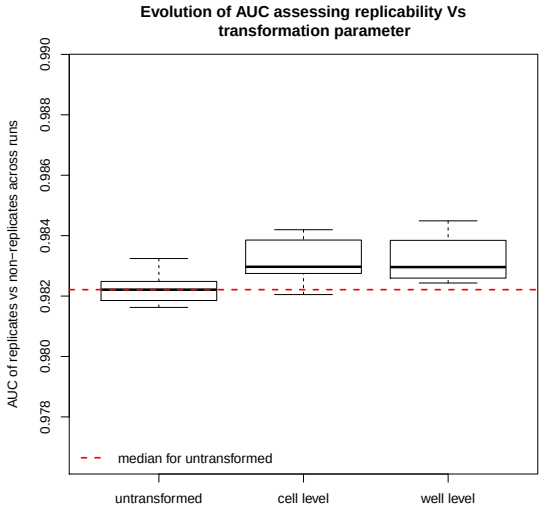
(c) *second data set* ($\lambda = 0.1$)

(d) *second data set* ($\lambda = 0.5$)

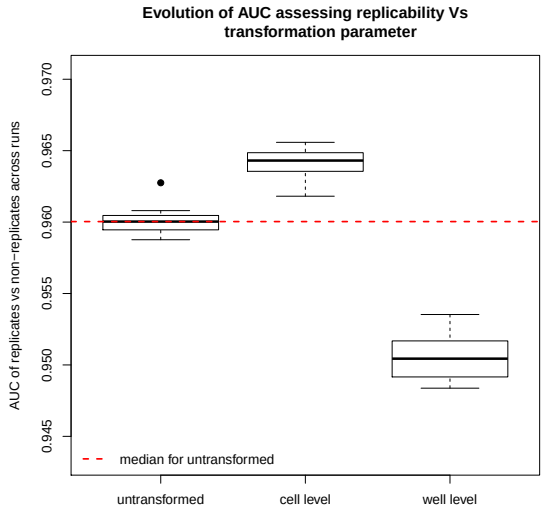
Figure 16: Distribution of T^2 statistics comparing untransformed, cell and well level performance



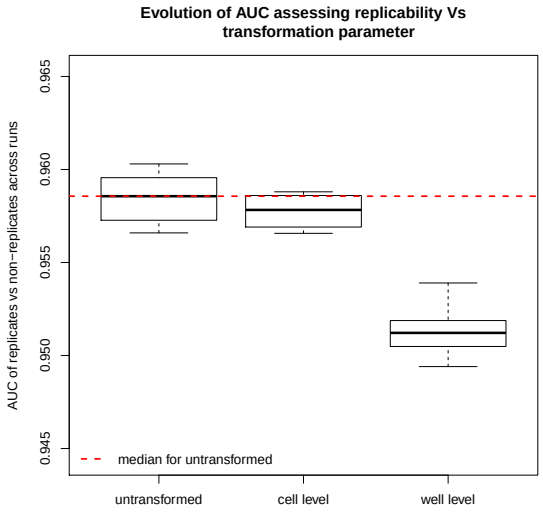
(a) *First data set* ($\lambda = 10.5$)



(b) *First data set* ($\lambda = 11$)



(c) *Second data set* ($\lambda = 0.1$)



(d) *Second data set* ($\lambda = 0.5$)

Figure 17: Boxplots of AUC values across runs for untransformed, cell and well level transformed data

5 Discussion and Conclusion

The primary objective of this study was to investigate effects of glog transformations on separation of treatment replicates from non-replicates. Two approaches were explored- using the Hotelling's T^2 statistic and assessment by AUC approach. In both data sets using both approaches, a few transformations led to slight improvement in treatments separation while some other transformations showed no improvement beyond what was observed before transformation. In cases where improvement was observed, the improvement was often very small that respective boxplots were indistinguishable. Our findings suggest that transformations in general failed to authoritatively improve separation of treatment replicates from non-replicates beyond results observed when data was untransformed. This findings compare unfavourably with findings in other settings that showed improved classification results after applying glog transformations [12, 15].

The inability of transformations to improve treatments separation was possibly influenced by multiple factors. Firstly, applying the same transformation across all features was clearly not optimal in regards to improving respective lack of normality in their distributions. Feature-specific distributions showed no improvement in distributions after transformations and therefore similar problems that were undermining treatments separation using multivariate classification methods were still present even after transformations. We hypothesize that for glog transformations to improve treatments separation, feature-specific transformations should be applied to address lack of normality seen at feature level. Secondly, improvement in separation, even though insignificant, was observed at specific ranges of λ values. While we did not explicitly determine the maximum likelihood λ estimate, results done using MLE λ values [12] led to improved classification results. We therefore encourage future studies to be transformed at optimal λ values for respective features. Thirdly, the small improvements in treatment separation for transformations compared to untransformed cases was largely due to efficient feature selection algorithms such that even when untransformed, treatments separation was already impressive. This means that transformations were left with a very small window of improving treatments separation over what was achieved in untransformed cases. Only an extremely efficient transformation would have separated treatments better.

Our findings suggest that effect of transformations on the proportion of actively-called treatments is cell-line specific. While in the first data set transformations led to a slight decrease in the proportion of actively-called treatments, similar transformations were associated with increased proportions of actively-called treatments in the second data set. Further, the number of features selected highly influenced the proportion of actively-called treatments. In the first data set where the number of features selected was often higher than number of features selected pre-transformation, the associated proportion of actively-

called treatments was in general lower than the proportion pre-transformation. For the second data set, where number of features selected was higher than in untransformed case, the proportion of actively-called treatments was higher than in untransformed case, but comparatively lower than for the set of transformations that selected less features in comparison with the untransformed case. This relationship between number of features selected and proportion of active treatments exists because an optimal features set containing more features leads to a higher precision in treatments discrimination and therefore only undoubtedly better treatments than the DMSO control are selected as active.

This study found no preference in applying transformations on single cell data to aggregated data using both the Hotelling's T^2 and AUC approaches. While some transformations showed small improvements (and in others, small declines) in treatments separation when transformations were applied at single cell level compared to well-level, the differences were consistently minimal and not-significant. The lack of inclination towards single-cell or well level transformation was likely influenced by failure of transformations to address feature-specific lack of normality and the high treatments separation achieved due to selection of highly effective features at all levels of transformation. As was observed for untransformed cases, selection of a highly effective set of features discriminates treatment replicates from non-replicates admirably well regardless of whether a transformation was applied or not.

Lastly, but certainly not the least, settling on top 10 ranked features from MRMR for calculation of Hotelling's T^2 statistic was a convenient and mutually agreed choice across all λ-transformations. We recognize that firstly, it was practically possible to use up to 16 of the selected features while still guarding against treatment pairs with 9 replicates apiece as motivated for in section 3.1. And secondly, at every λ transformation, it was possible to determine the number of replicates needed that would have maximized treatment pairs separation and avoid subjecting all transformations to same feature dimension restrictions. A future study could endeavor to address these.

References

[1] J. H. Price, A. Goodacre, K. Hahn, L. Hodgson, E. A. Hunter, S. Krajewski, R. F. Murphy, A. Rabinovich, J. C. Reed, and S. Heynen, "Advances in molecular labeling, high throughput imaging and machine intelligence portend powerful functional cellular biochemistry tools," *Journal of Cellular Biochemistry*, vol. 87, no. S39, pp. 194–210, 2002.

[2] S. M. Lin, P. Du, W. Huber, and W. A. Kibbe, "Model-based variance-stabilizing transformation for Illumina microarray data," *Nucleic Acids Research*, vol. 36, no. 2, pp. e11–e11, nov 2007.

[3] F. Reisen, A. Sauty de Chalon, M. Pfeifer, X. Zhang, D. Gabriel, and P. Selzer, "Linking Phenotypes and Modes of Action Through High-Content Screen Fingerprints." *Assay and drug development technologies*, vol. 13, no. 7, pp. 415–27, 2015.

[4] P. Lang, K. Yeow, A. Nichols, and A. Scheer, "Cellular imaging in drug discovery," pp. 343–356, 2006.

[5] Z. E. Perlman, M. D. Slack, Y. Feng, T. J. Mitchison, L. F. Wu, and S. J. Altschuler, "Multidimensional drug profiling by automated microscopy." *Science (New York, N.Y.)*, vol. 306, no. 5699, pp. 1194–8, 2004.

[6] L.-H. Loo, L. F. Wu, and S. J. Altschuler, "Image-based multivariate profiling of drug responses from single cells," *Nature Methods*, 2007.

[7] C. Laufer, B. Fischer, M. Billmann, W. Huber, and M. Boutros, "Mapping genetic interactions in human cancer cells with RNAi and multiparametric phenotyping." *Nature methods*, vol. 10, no. 5, pp. 427–31, 2013.

[8] N. De Jay, S. Papillon-Cavanagh, C. Olsen, N. El-Hachem, G. Bontempi, and B. Haibe-Kains, "MRMRe: An R package for parallelized mRMR ensemble feature selection," *Bioinformatics*, vol. 29, no. 18, pp. 2365–2368, 2013.

[9] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data." *Journal of bioinformatics and computational biology*, vol. 3, no. 2, pp. 185–205, 2005.

[10] J. K. Patel, *Handbook of the normal distribution*, ser. Statistics, textbooks and monographs 40. M. Dekker, 1982.

[11] D. W. W. Richard A. Johnson, *Applied Multivariate Statistical Analysis (6th Edition)*, 6th ed. Prentice Hall, 2007.

[12] H. M. Parsons, C. Ludwig, U. L. Günther, and M. R. Viant, "Improved classification accuracy in 1- and 2-dimensional NMR metabolomics data using the variance stabilising generalised logarithm transformation," *BMC bioinformatics*, vol. 8, no. 1, p. 234, 2007.

[13] G. A. Churchill, "Fundamentals of experimental design for cDNA microarrays," *Nat Genet*, vol. 32 Suppl, pp. 490–495, 2002.

[14] R. A. van den Berg, H. C. J. H. Hoefsloot, and Westerhuis, "Centering, scaling, and transformations: improving the biological information content of metabolomics data." *BMC genomics*, vol. 7, no. 1, p. 142, 2006.

[15] P. V. Purohit, D. M. Rocke, M. R. Viant, and D. L. Woodruff, "Discrimination models using variance-stabilizing transformation of metabolomic NMR data." *Omics : a journal of integrative biology*, vol. 8, no. 2, pp. 118–130, 2004.

[16] S. A. Haney, "Rapid assessment and visualization of normality in high-content and other cell-level data and its impact on the interpretation of experimental results," *Journal of biomolecular screening*, vol. 19, no. 5, pp. 672–684, 2014.

[17] I. Yeo and R. A. Johnson, "A new family of power transformations to improve normality or symmetry," *Biometrika*, vol. 87, no. 4, p. 954, 2000.

[18] S. Weisberg, "Yeo-Johnson Power Transformations," *Department of Applied Statistics, University of Minnesota*, no. 2, pp. 1–4, 2001.

[19] W. Huber, A. von Heydebreck, H. Sültmann, A. Poustka, and M. Vingron, "Variance stabilization applied to microarray data calibration and to the quantification of differential expression." *Bioinformatics (Oxford, England)*, vol. 18 Suppl 1, no. 1997, pp. S96–104, 2002.

[20] B. P. Durbin, J. S. Hardin, D. M. Hawkins, and D. M. Rocke, "A variance-stabilizing transformation for gene-expression microarray data," *Bioinformatics*, vol. 18, no. suppl 1, pp. S105–S110, 2002.

[21] D. Rocke and B. Durbin, "Approximate variance-stabilizing transformations for gene-expression microarray data," *Bioinformatics*, vol. 19, no. 8, pp. 966–972, 5 2003.

[22] P. Munson, "A consistency test for determining the significance of gene expression changes on replicate samples and two convenient variance-stabilizing transformations," *Gene Logic Workshop of Low Level Analysis of Affymetrix geneChip Data*, 2001.

[23] D. M. Rocke and B. Durbin, "A model for measurement error for gene expression arrays." *Journal of Computational Biology*, vol. 8, no. 6, pp. 557–69, 2001.

[24] D. Johnson, "Signal-to-noise ratio," *Scholarpedia*, vol. 1, no. 12, p. 2088, 2006.

[25] B. Durbin and D. M. Rocke, "Estimation of transformation parameters for microarray data," *Bioinformatics*, vol. 19, no. 11, pp. 1360–1367, 2003.

[26] G. E. P. Box and D. R. Cox, "An analysis of transformations," *Journal of the Royal Statistical Society. Series B (Methodological*, vol. 26, no. 2, pp. 211–252, 1964.

[27] C. J. Nachtsheim, J. Neter, W. Li, and M. H. Kutner, "Applied linear statistical models," *Journal Of The Royal Statistical Society Series A General*, vol. Fifth, pp. 83–84, 2004.

[28] H. Hotelling, "The generalization of student's ratio," *Ann. Math. Statist.*, vol. 2, no. 3, pp. 360–378, 08 1931.

[29] K. J. Preacher and R. C. MacCallum, "Exploratory factor analysis in behavior genetics research: Factor recovery with small sample sizes," *Behavior Genetics*, vol. 32, no. 2, pp. 153–161, 2002.

[30] Y.-Y. Chi and K. E. Muller, "Two-Step Hypothesis Testing When the Number of Variables Exceeds the Sample Size," *Communications in Statistics - Simulation and Computation*, vol. 42, no. 5, pp. 1113–1125, 2013.

[31] C. E. Metz, "Basic principles of ROC analysis," *Seminars in Nuclear Medicine*, vol. 8, no. 4, pp. 283–298, 1978.

[32] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve." *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

[33] J. A. Hanley, "Receiver operating characteristic (ROC) methodology: the state of the art." pp. 307–335, 1989.

[34] M. H. Zweig and G. Campbell, "Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine," pp. 561–577, 1993.

[35] D. J. Hand, "Measuring classifier performance: A coherent alternative to the area under the ROC curve," *Machine Learning*, vol. 77, no. 1, pp. 103–123, 2009.

[36] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2017.

6 Appendix

Table 1: mean α values calculated at cell & well levels for both data sets

	Data set 1		**Data set 2**	
Feature	cell	well	cell	well
Ratio_Nuc2Cyto_GFP_TotalInt	1.87	1.95	1.58	1.7
Cyto_NucCytoBFP_SERDark_0	179.85	182.29	159.88	162.08
Nuc_NucCytoBFP_SERHole_1	1.2	1.21	1.28	1.28
Nuc_RFP_SERSaddle_2	9.03	9.01	26.48	26.48
Nuc_NucCytoBFP_LogTotalIntBC	5.56	5.56	5.34	5.34
Nuc_GFP_SERSaddle_2	11.76	11.71	11.46	11.52
Nuc_NucCytoBFP_LogMeanInt	3.06	3.06	2.88	2.88
Cell_GFP_MeanIntBC	3932.4	3922.84	390.1	388.05
Cell_RFP_StdIntBC	155.95	157.05	141.03	139.99
Nuc_NucCytoBFP_SERDark_1	7.27	7.34	6	6.06
Cell_RFP_SERBright_0	153.97	154.23	313.11	313.09
Ratio_Nuc2Cell_Area	0.56	0.57	0.49	0.5
Nuc_GFP_SERDark_1	42.72	42.03	40.39	39.9
Ratio_Nuc2Cyto_NucCytoBFP_TotalIntBC	4.49	4.66	2.72	2.87
Ratio_Nuc2Cell_NucCytoBFP_LogTotalInt	0.98	0.98	0.96	0.96
Cell_RFP_SERHole_4	0.3	0.29	2.4	2.3
Cyto_GFP_SERSpot_2	7.53	7.69	6.36	6.44
Cyto_GFP_CVInt	0.41	0.42	0.52	0.53
Cell_RFP_SERRidge_4	3.94	4.01	6.79	6.77
Cyto_RFP_SERHole_1	8.63	8.74	16.42	16.43
Nuc_GFP_SERBright_0	99.54	101.95	164.64	165.98
Nuc_GFP_SERSpot_2	4.99	5.21	4.28	4.57
Ratio_Nuc2Cyto_NucCytoBFP_LogMeanInt	1.13	1.13	1.1	1.11
Cell_NucCytoBFP_SERRidge_0	52.45	52.68	56.95	57.36
Nuc_GFP_SERSaddle_0	44.41	44.85	62.82	63.07
Cell_RFP_SEREdge_1	97.56	98.62	209.17	208.86
Ratio_Nuc2Cyto_GFP_TotalIntBC	1.9	1.98	1.81	1.97
Cell_RFP_SERSaddle_1	21.65	21.73	62.3	62.33
Cyto_GFP_SERValley_0	82.06	84.9	120.28	124.34
Cell_RFP_LogMeanIntBC	2.45	2.45	2	2
Cell_RFP_SERDark_1	22.83	22.87	77.08	76.74
Ratio_Nuc2Cell_RFP_MeanIntBC	1.15	1.16	0.76	0.78
Cyto_GFP_SERRidge_0	83.3	84.66	99.19	99.47

Cell_NucCytoBFP_LogMeanIntBC	2.89	2.88	2.67	2.66
Nuc_RFP_SERRidge_0	78.23	78.43	171.11	171.64
Cyto_NucCytoBFP_SERBright_2	3.55	3.6	2.45	2.5
Cyto_NucCytoBFP_LogMeanIntBC	2.6	2.6	2.47	2.45
Nuc_ShapeSize_Symmetry03	0.08	0.08	0.07	0.07
Cyto_GFP_SERSaddle_0	61.35	62.9	75.18	76.59
Cyto_NucCytoBFP_SERRidge_4	0.75	0.77	0.46	0.48
Ratio_Nuc2Cell_RFP_LogTotalIntBC	0.96	0.96	0.9	0.91
Cell_RFP_SERSpot_1	12.24	12.34	33.22	33.19
Cell_NucCytoBFP_SEREdge_1	115.57	116.86	82.22	84.09
Ratio_Nuc2Cell_RFP_LogMeanIntBC	1.02	1.03	0.92	0.93
Cyto_NucCytoBFP_TotalIntBC	139131.63	137047.71	136886.18	130551.33
Cell_GFP_StdIntBC	1548.73	1552.36	196.15	197.7
Ratio_Nuc2Cell_NucCytoBFP_TotalIntBC	0.75	0.76	0.67	0.68
Nuc_NucCytoBFP_SERBright_2	20.41	20.54	15.74	15.91
Nuc_RFP_SERDark_0	136.47	136.06	367.68	363.23
Cell_GFP_SERBright_2	24.53	25	19.57	20.12
Ratio_Nuc2Cell_NucCytoBFP_TotalInt	0.73	0.74	0.63	0.64
Cell_GFP_CVInt	0.44	0.44	0.44	0.45
Cyto_GFP_SEREdge_2	139.06	142.95	118.24	124.15
Nuc_ShapeSize_Symmetry04	0.07	0.07	0.06	0.06
Cell_GFP_SERSpot_2	6.79	6.98	5.79	6
Cyto_RFP_SERBright_1	26.57	26.46	81.76	80.5
Ratio_Nuc2Cell_NucCytoBFP_LogTotalIntBC	0.98	0.98	0.97	0.97
Cell_RFP_SERDark_4	1.98	1.93	8.04	7.88
Cell_RFP_SEREdge_4	51.77	52.43	72.69	72.55
Nuc_ShapeSize_Symmetry15	0.07	0.07	0.06	0.06
Nuc_ShapeSize_Perimeter	66.79	67.1	63.76	64.2
Nuc_ShapeSize_Symmetry05	0.04	0.04	0.04	0.04
Nuc_RFP_SERBright_2	14.28	14.61	21.21	21.85
Ratio_Nuc2Cell_RFP_MeanInt	1.1	1.11	0.85	0.87
Cyto_RFP_SERHole_2	1.78	1.78	3.93	3.92
Nuc_ShapeSize_Symmetry13	0.12	0.12	0.11	0.11
Cyto_GFP_SERBright_0	126.95	129.39	171.45	171.79
Cyto_GFP2RFP_PearsonCor	0.52	0.53	0.57	0.58
Cell_GFP_SERHole_4	1.77	1.72	1.35	1.3
Cell_ShapeSize_Symmetry15	0.1	0.1	0.1	0.1
Cyto_RFP_SERValley_2	13.34	13.72	21.86	22.52
Nuc_GFP_StdInt	1255.11	1266.65	149.4	151.02
Cyto_RFP_SERHole_4	0.49	0.48	0.89	0.89

Cell_NucCytoBFP_SERHole_4	0.39	0.37	0.48	0.46
Cell_GFP_SERSpot_1	13.33	13.65	14.29	14.52
Cell_RFP_SERSaddle_0	52.41	52.51	140.72	140.48
Nuc_NucCytoBFP_TotalIntBC	434327.92	436627.2	255088.81	253135.07
Nuc_NucCytoBFP_SERDark_2	1.37	1.38	0.95	0.96
Cell_ShapeSize_Symmetry12	0.3	0.3	0.31	0.31
Cyto_NucCytoBFP_SERRidge_1	12.16	12.35	8.77	8.93
Nuc_RFP_SERValley_0	74.31	74.15	231.69	228.95
Nuc_NucCytoBFP_SERSaddle_0	28.86	29.02	29.09	29.35
Ratio_Nuc2Cyto_RFP_MeanIntBC	1.76	1.8	0.99	1.07
Cyto_NucCytoBFP_StdIntBC	217.33	217.37	126.1	127.45
Cyto_RFP_SEREdge_2	106.48	108.98	127.04	128.22
Nuc_GFP_SERValley_2	16.32	15.88	12.65	12.41
Ratio_Nuc2Cyto_RFP_LogMeanInt	1.05	1.05	0.95	0.96
Ratio_Nuc2Cyto_RFP_TotalIntBC	3.1	3.24	1.74	1.94
Cyto_NucCytoBFP_SEREdge_4	122.96	125.81	70.44	73.18
Cyto_GFP_TotalIntBC	1246167.21	1217248.5	153694.96	146484.77
Ratio_Nuc2Cell_RFP_TotalIntBC	0.65	0.66	0.39	0.41
Ratio_Nuc2Cyto_NucCytoBFP_LogTotalIntBC	1.11	1.11	1.07	1.07
Nuc_ShapeSize_MajorAxisLength	13.75	13.81	13.1	13.2
Ratio_Nuc2Cell_GFP_TotalIntBC	0.54	0.55	0.52	0.53
Cell_NucCytoBFP_SERSpot_2	8.96	9.08	6.27	6.4
Cell_RFP_SERValley_2	7.38	7.38	25.21	25.09
Cell_NucCytoBFP_SERSaddle_4	6.63	6.68	4.78	4.89
Nuc_RFP_CVInt	0.24	0.24	0.66	0.66
Cyto_NucCytoBFP_SERDark_2	35.09	35.53	21.91	22.33
Cell_GFP_MeanInt	3967.38	3956.27	446.62	444.08
Nuc_RFP_LogMeanIntBC	2.51	2.51	1.85	1.86
Cyto_RFP_LogMeanInt	2.53	2.52	2.26	2.25
Cell_GFP_SERSpot_4	3.77	3.88	2.64	2.78
Cell_NucCytoBFP_SERRidge_4	4.6	4.68	2.91	3.02
Cyto_NucCytoBFP_SERDark_1	66.49	67.63	40.5	41.35
Cyto_GFP_SERBright_4	10.24	10.37	6.91	6.99
Cell_GFP_SERBright_1	46.23	47.11	44.14	44.98
Cell_RFP_LogMeanInt	2.61	2.61	2.22	2.22
Cell_NucCytoBFP_TotalIntBC	573459.54	573674.91	391974.99	383686.4
Ratio_Nuc2Cyto_RFP_LogMeanIntBC	1.09	1.1	0.92	0.94
Nuc_GFP_TotalInt	1484125.46	1503997.72	168438.26	170452.46
Nuc_NucCytoBFP2GFP_PearsonCor	-0.51	-0.51	-0.43	-0.41
Nuc_GFP_SERDark_0	103.85	103.96	162.32	161.56

Cyto_GFP_SERHole_0	46.82	48.29	99.07	100.77
Nuc_ShapeSize_Roundness	0.89	0.89	0.91	0.91
Nuc_RFP_TotalIntBC	136057.7	137639.83	31607.08	32843.4
Cell_ShapeSize_Symmetry03	0.1	0.1	0.11	0.11
Ratio_Nuc2Cyto_NucCytoBFP_LogTotalInt	1.09	1.1	1.05	1.06
Cell_GFP_SERDark_1	31.37	31.47	39.97	40.39
Nuc_NucCytoBFP_SERRidge_2	11.9	11.97	8.94	9.05
Cyto_RFP_SERValley_4	3.95	4	5.38	5.52
Cell_NucCytoBFP_LogMeanInt	2.94	2.94	2.76	2.76
Cyto_GFP_SERSaddle_4	7.88	8.02	6.69	7.04
Nuc_ShapeSize_Symmetry14	0.1	0.1	0.09	0.09
Nuc_GFP2RFP_PearsonCor	0.25	0.25	0.51	0.51
Cell_NucCytoBFP_SERHole_1	3.56	3.55	3.75	3.73
Cyto_GFP_SERHole_1	7.68	7.89	11.71	11.77
Cell_GFP_SERSaddle_1	24.42	24.67	28.01	28.23
Cyto_RFP_SERSpot_2	4.8	4.77	13.17	13.07
Nuc_RFP_SERDark_1	19.77	19.6	104.74	102.15
Ratio_Nuc2Cyto_RFP_LogTotalIntBC	1.07	1.08	0.96	0.97
Cyto_RFP_SEREdge_1	136.06	139.05	199.4	199.7
Cyto_GFP_SERSpot_1	15.98	16.32	15.28	15.28
Nuc_NucCytoBFP_SERValley_2	1.47	1.48	1.04	1.05
Cyto_RFP_SERValley_1	30.96	31.83	60.65	61.72
Cyto_RFP_LogTotalInt	4.95	4.94	4.8	4.78
Cell_RFP_SERValley_1	18.95	19.01	66.11	65.91
Cell_RFP_MeanInt	422.58	422.66	174.33	173.51
Cell_GFP_SERBright_0	112.42	114.71	169.2	170.18
Cell_RFP_SERHole_2	1.63	1.61	7.48	7.34
Cell_RFP_SERRidge_1	22.21	22.44	55.19	55.02
Cyto_NucCytoBFP_MeanIntBC	441.37	435.06	321.63	315.04
Cyto_GFP_StdIntBC	1395.53	1399.42	205.16	206.12
Cell_NucCytoBFP_SERValley_2	9.15	9.09	7.69	7.64
Nuc_GFP_SERHole_0	54.75	54.91	95.55	94.92
Cell_NucCytoBFP_SERRidge_1	18.67	18.84	14.18	14.38
Nuc_RFP_SEREdge_0	123.91	124.6	346.45	344.55
Cyto_GFP_LogMeanIntBC	3.48	3.47	2.5	2.48
Cyto_NucCytoBFP_SERBright_4	0.73	0.74	0.47	0.49
Nuc_NucCytoBFP_SERRidge_1	20.87	20.96	16.98	17.08
Nuc_RFP_LogTotalInt	5.19	5.2	4.67	4.68
Nuc_NucCytoBFP_SERSpot_1	16.83	16.91	13.27	13.37
Cell_GFP_LogMeanInt	3.51	3.51	2.62	2.62

Ratio_Nuc2Cell_GFP_MeanIntBC	0.94	0.95	1.04	1.05
Cell_NucCytoBFP_CVIntBC	0.53	0.54	0.48	0.48
Nuc_RFP_SERDark_2	6.76	6.64	46.77	45.1
Nuc_RFP_LogMeanInt	2.65	2.65	2.14	2.14
Cell_NucCytoBFP_SERDark_0	91.51	91.64	108.63	109.12
Nuc_NucCytoBFP_SERSaddle_2	7.49	7.56	5.25	5.38
Nuc_NucCytoBFP_SERSaddle_1	12.83	12.95	10.13	10.29
Cyto_RFP_SERSaddle_4	7.53	7.74	8.41	8.52
Cyto_NucCytoBFP_SERHole_1	11.32	11.5	8.63	8.78
Nuc_NucCytoBFP_SERHole_0	37.24	37.31	51.3	51.78
Cell_GFP_TotalIntBC	2716768.37	2708212.06	301393.56	296205.84
Cell_NucCytoBFP_SERValley_1	18.96	18.96	15.26	15.24
Nuc_GFP_MeanInt	3747.03	3764.73	458.99	460.5
Cyto_RFP_SERRidge_0	85.16	85.43	175.61	175.08
Nuc_RFP_SERSaddle_0	49.36	49.39	147.05	146.46
Cyto_GFP_SERHole_2	1.9	1.9	3.4	3.42
Cell_RFP_SEREdge_0	139.56	140.58	324.32	323.73
Cyto_RFP_StdIntBC	154.9	154.44	153.26	151.26
Cell_GFP_SERDark_4	5.39	5.27	4.16	4.08
Cell_RFP_LogTotalInt	5.41	5.41	5.08	5.07
Cell_RFP_SERSpot_4	3.01	3.1	3.75	3.8
Nuc_GFP_SEREdge_0	168.53	168.93	175.58	177.03
Nuc_NucCytoBFP2RFP_PearsonCor	0.1	0.11	-0.32	-0.3
Cell_RFP_SEREdge_2	75.25	76.13	136.28	136.12
Cyto_NucCytoBFP_SERRidge_0	61.9	62.47	64.28	64.98
Cyto_RFP_MeanInt	356.11	352.79	195.61	191.89
Nuc_RFP_TotalInt	182408.62	184519.34	54209.52	55555.13
Cell_NucCytoBFP_SERValley_0	55.44	55.6	58.12	58.41
Nuc_GFP_SERSaddle_1	21.1	21.19	25.51	25.65
Cell_RFP_SERHole_1	7.04	7.01	21.25	21.07
Cyto_GFP_MeanInt	4034.8	3985.16	422.99	414.02
Cell_NucCytoBFP_StdInt	427.1	429.75	235.01	235.32
Cyto_RFP_SERSaddle_0	60.6	61.11	138.59	138.46
Nuc_GFP_CVInt	0.37	0.37	0.32	0.33
Cyto_GFP_SEREdge_1	186.3	191.99	163.01	170.2
Nuc_RFP_SERSaddle_1	19.69	19.7	63.55	63.54
Nuc_RFP_SERRidge_2	10.84	11.05	16.68	17.11
Ratio_Nuc2Cyto_RFP_MeanInt	1.39	1.4	0.86	0.89
Nuc_RFP_SERValley_1	15.7	15.59	87.31	85.21
Cyto_NucCytoBFP_SERBright_0	109.62	110.48	121.93	123.05

Cell_NucCytoBFP_SERHole_0	49.8	49.77	66.88	67.15
Cell_NucCytoBFP_LogTotalIntBC	5.69	5.68	5.53	5.52
Cyto_GFP_LogTotalInt	5.93	5.91	5.13	5.1
Cell_NucCytoBFP_SEREdge_0	136.49	137.81	107.71	109.5
Cell_ShapeSize_Symmetry02	0.21	0.21	0.22	0.22
Ratio_Nuc2Cell_GFP_LogTotalInt	0.95	0.95	0.94	0.94
Nuc_RFP_SERBright_1	30.09	30.54	61.19	62.43
Nuc_RFP_StdIntBC	112.43	113.52	97.89	99.22
Cell_GFP_SERRidge_4	7.59	7.71	5.17	5.3
Cell_RFP_SERRidge_2	10.37	10.52	22.37	22.3
Cyto_NucCytoBFP_SERHole_4	1.37	1.33	1.32	1.29
Cell_ShapeSize_Area	695.44	694.98	795.4	785.41
Cell_NucCytoBFP_SERHole_2	1.02	1	0.97	0.95
Ratio_Nuc2Cyto_RFP_TotalInt	2.28	2.37	1.15	1.25
Nuc_RFP_SERSpot_2	5.44	5.59	7.56	7.87
Cyto_NucCytoBFP_SERValley_2	36.02	36.54	22	22.46
Ratio_Nuc2Cyto_NucCytoBFP_MeanInt	2.32	2.34	1.91	1.91
Cell_NucCytoBFP_SEREdge_4	71.49	72.37	51.26	52.54
Ratio_Nuc2Cell_NucCytoBFP_MeanIntBC	1.36	1.36	1.4	1.39
Cyto_RFP_SERRidge_2	8.12	8.09	24.23	23.81
Cyto_GFP_SERSpot_0	62.26	63.7	98.23	98.34
Cyto_RFP_LogTotalIntBC	4.73	4.72	4.58	4.55
Cyto_RFP_SEREdge_0	178.93	181.91	314.8	314.33
Cell_RFP_SERBright_1	30.11	30.4	77.07	76.88
Cyto_NucCytoBFP2GFP_PearsonCor	0.59	0.6	0.55	0.56
Cyto_NucCytoBFP_CVIntBC	0.52	0.53	0.42	0.44
Nuc_NucCytoBFP_SERBright_1	32.75	32.9	26.28	26.45
Cell_GFP_SERSaddle_2	12.77	12.81	12.61	12.8
Ratio_Nuc2Cyto_GFP_LogTotalInt	1.02	1.02	1.01	1.02
Nuc_NucCytoBFP_SERRidge_0	49.33	49.52	52.65	53.07
Nuc_GFP_SEREdge_1	136.38	136.04	122.07	123.21
Cell_ShapeSize_Symmetry14	0.13	0.13	0.14	0.14
Cyto_RFP_SERSpot_1	11.1	11.04	35.61	35.14
Cell_NucCytoBFP_SERBright_4	8.17	8.32	5.4	5.57
Cyto_NucCytoBFP_SERBright_1	15.84	16.1	11.28	11.45
Cell_NucCytoBFP_MeanInt	911.79	912.09	610.08	605.32
Nuc_RFP_SERBright_0	149.23	149.54	308.76	309.5
Cell_NucCytoBFP_SERBright_0	99.97	100.36	113.24	114.05
Nuc_GFP_TotalIntBC	1470601.16	1490963.55	147698.6	149721.07
Cyto_RFP_SERDark_2	13.36	13.69	22.84	23.43

Cell_ShapeSize_MinorAxisLength	12.24	12.24	13.06	12.98
Cyto_RFP_TotalIntBC	74001.37	72177.81	56791.37	54180.59
Cyto_RFP_CVIntBC	0.7	0.71	1.26	1.28
Ratio_Nuc2Cell_NucCytoBFP_LogMeanInt	1.04	1.04	1.04	1.04
Nuc_GFP_SERValley_0	64.59	64.55	91.3	91.05
Cell_NucCytoBFP_LogTotalInt	5.74	5.74	5.62	5.61
Cell_GFP_SERDark_0	94.51	95.97	165.7	166.4
Nuc_NucCytoBFP_StdInt	312.53	316.28	157.95	160.44
Ratio_Nuc2Cell_RFP_TotalInt	0.62	0.63	0.43	0.44
Nuc_GFP_SERDark_2	24.12	23.52	17.72	17.34
Cyto_GFP_SERRidge_1	42.17	42.57	34.54	34.49
Cyto_GFP_SERRidge_2	22.14	22.31	15.98	15.98
Cyto_GFP_SERSaddle_1	33.84	34.7	33.13	33.91
Cell_RFP_SERHole_0	87.9	87.86	158.05	158.29
Cell_GFP_SERDark_2	14.54	14.42	15.56	15.64
Cell_RFP_SERValley_0	80.51	80.57	202.25	201.88
Cell_GFP_SERHole_0	46.01	46.79	92.89	92.97
Cyto_GFP_SERDark_4	4.09	4.09	4.62	4.77
Nuc_NucCytoBFP_SEREdge_2	83.35	84.23	62.33	63.61
Nuc_NucCytoBFP_SERSpot_2	11.67	11.75	9.25	9.34
Cell_NucCytoBFP_SERDark_2	8.98	8.91	7.67	7.61
Cyto_NucCytoBFP_SERValley_1	62.89	63.99	36.83	37.64
Nuc_GFP_SERRidge_1	32.68	33.63	33.75	34.8
Cyto_GFP_SERValley_2	12.58	12.88	18.15	19.36
Ratio_Nuc2Cell_GFP_LogTotalIntBC	0.95	0.95	0.94	0.94
Nuc_GFP_SERRidge_0	66.56	68.11	96.6	97.55
Nuc_NucCytoBFP_SERValley_1	7.08	7.15	5.56	5.62
Cell_NucCytoBFP2GFP_PearsonCor	-0.15	-0.14	0.11	0.13
Cell_GFP_SERHole_2	5.67	5.63	5.1	5.04
Cyto_GFP_SERDark_0	113.02	116.83	191	196.07
Cyto_NucCytoBFP_SERSpot_2	1.05	1.07	0.73	0.73
Cyto_RFP_SERDark_0	177.55	179.42	297.32	299.9
Ratio_Nuc2Cyto_RFP_LogTotalInt	1.05	1.06	0.98	0.98
Cell_GFP_StdInt	1548.86	1552.5	196.31	197.86
Cyto_NucCytoBFP_TotalInt	168493.69	166435.45	183198.09	175507.17
Ratio_Nuc2Cell_NucCytoBFP_MeanInt	1.32	1.32	1.32	1.32
Nuc_NucCytoBFP_SERHole_2	0.12	0.11	0.07	0.06
Ratio_Nuc2Cyto_GFP_MeanIntBC	1.07	1.09	1.31	1.36
Cell_NucCytoBFP_TotalInt	640159.82	641247.79	477738.8	468384.18
Cyto_GFP_SERRidge_4	8.04	8.11	5.2	5.24

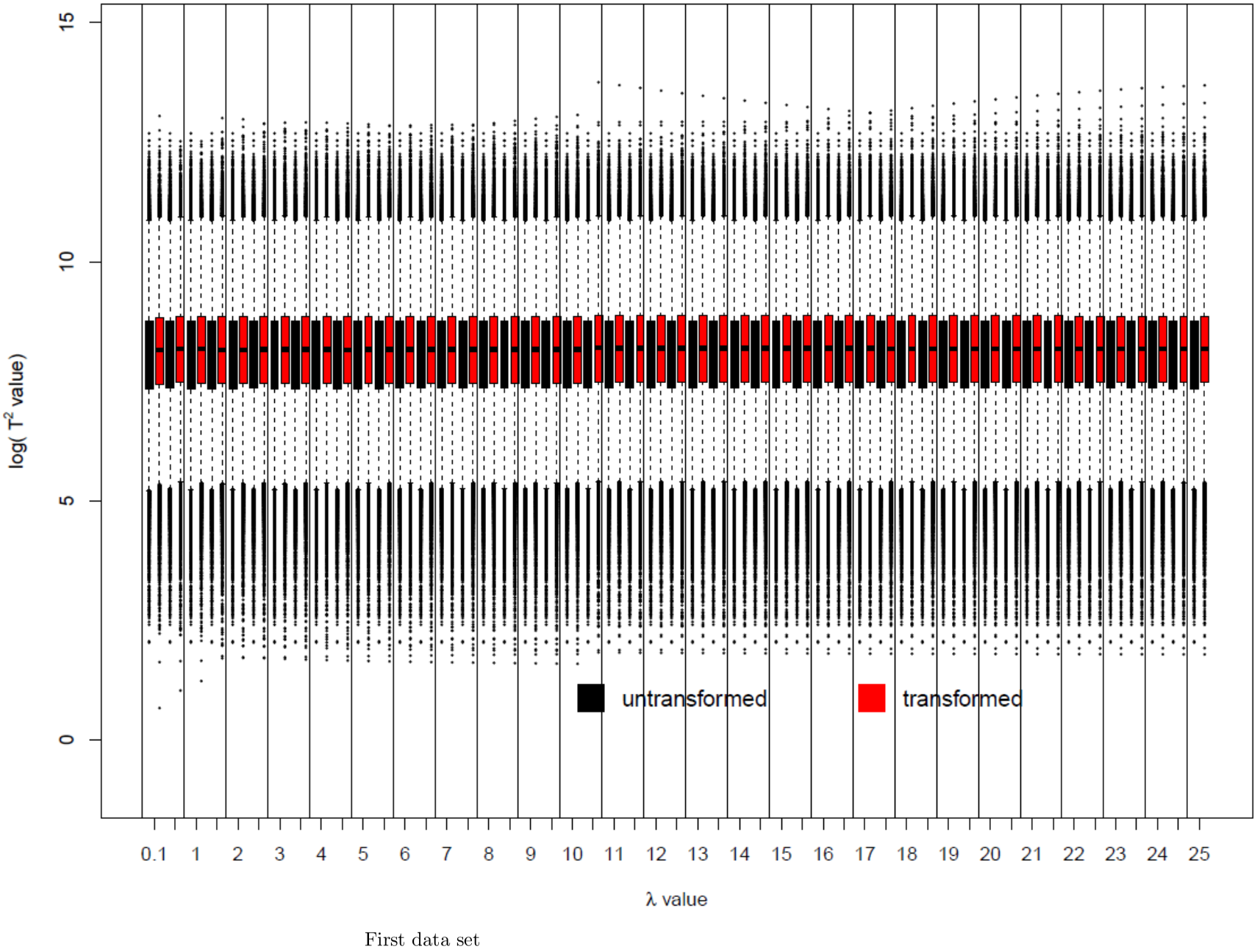
Cell_ShapeSize_Eccentricity	0.7	0.7	0.71	0.71
Cyto_RFP_StdInt	155.18	154.73	154.98	153.02
Nuc_ShapeSize_MinorAxisLength	9.06	9.1	9.05	9.05
Cyto_NucCytoBFP_SEREdge_1	190.26	194.93	106.73	111.21
Nuc_RFP_SERSpot_0	93.43	93.6	183.95	184.29
Cell_GFP_SERRidge_0	74.62	76.01	98.62	99.35
Nuc_NucCytoBFP_CVInt	0.26	0.26	0.2	0.2
Nuc_GFP_LogMeanIntBC	3.44	3.44	2.55	2.56
Cell_GFP_SERValley_4	4.35	4.26	3.4	3.37
Cyto_NucCytoBFP_SERHole_2	3.67	3.66	2.7	2.71
Nuc_ShapeSize_Eccentricity	0.7	0.7	0.67	0.67
Cell_ShapeSize_Symmetry13	0.15	0.15	0.16	0.16
Nuc_NucCytoBFP_LogMeanIntBC	3.02	3.02	2.81	2.8
Cyto_RFP_SERBright_2	11.26	11.21	32.66	32.2
Nuc_NucCytoBFP_StdIntBC	312.53	316.28	157.95	160.44
Cell_ShapeSize_Symmetry05	0.06	0.06	0.06	0.06
Cell_RFP_TotalInt	292930.7	292855.31	137484.67	135317.01
Cyto_GFP_LogMeanInt	3.51	3.5	2.58	2.57
Cyto_RFP_SERRidge_1	19.36	19.3	58.2	57.25
Nuc_RFP_MeanIntBC	345.85	346.95	84.74	87
Cyto_GFP_SERHole_4	0.49	0.47	0.89	0.89
Cyto_NucCytoBFP_SERSpot_4	0.09	0.09	0.08	0.08
Cell_NucCytoBFP_SERRidge_2	9.69	9.8	6.61	6.76
Cyto_NucCytoBFP_SERHole_0	90.48	91.3	96.76	97.9
Nuc_NucCytoBFP_MeanInt	1199.58	1197.62	806.65	796.65
Cell_RFP_SERDark_2	7.94	7.91	28.84	28.62
Cell_NucCytoBFP_MeanIntBC	815.79	814.77	502.54	497.91
Cell_RFP_MeanIntBC	303.39	303.13	112.7	112.16
Cyto_NucCytoBFP_SERRidge_2	3.01	3.05	2.08	2.12
Nuc_GFP_SERBright_1	37.95	39.11	40.23	41.53
Nuc_ShapeSize_Area	388.48	391.75	366.7	369.41
Cyto_GFP_SERDark_2	12.83	13.09	19.14	20.25
Cell_GFP_SEREdge_0	176.63	179.17	186.77	189.61
Cell_RFP_SERSaddle_2	10.55	10.59	25.47	25.53
Nuc_GFP_LogTotalIntBC	5.98	5.98	5.09	5.09
Cell_NucCytoBFP_SEREdge_2	98.54	99.69	69.82	71.54
Cyto_NucCytoBFP_CVInt	0.41	0.41	0.3	0.31
Cell_NucCytoBFP_StdIntBC	427.1	429.75	235	235.31
Cell_GFP_LogTotalIntBC	6.28	6.28	5.4	5.39
Nuc_NucCytoBFP_SERBright_0	96.47	96.76	107.27	108.11

Cyto_RFP_SERDark_4	3.87	3.89	5.52	5.64
Ratio_Nuc2Cyto_GFP_MeanInt	1.06	1.08	1.21	1.25
Cyto_NucCytoBFP_MeanInt	537.37	532.38	429.17	422.45
Cell_NucCytoBFP_SERSaddle_0	35.83	36	35.24	35.46
Nuc_NucCytoBFP_SERDark_0	66.56	66.79	82.99	83.76
Cyto_NucCytoBFP_LogMeanInt	2.7	2.7	2.61	2.6
Cyto_NucCytoBFP_SERValley_0	116.07	118.04	87.3	88.68
Cyto_GFP_TotalInt	1256990.75	1227477.22	177979.17	169843.15
Nuc_ShapeSize_Symmetry12	0.29	0.29	0.26	0.27
Cell_GFP_LogMeanIntBC	3.48	3.48	2.54	2.54
Cyto_GFP_CVIntBC	0.47	0.47	0.64	0.66
Cyto_RFP_SERBright_0	161.62	161.95	314.76	314.22
Cell_GFP_LogTotalInt	6.31	6.31	5.48	5.47
Cell_NucCytoBFP_SERDark_4	3	2.94	2.95	2.89
Nuc_RFP_SERValley_2	5.92	5.82	38.17	36.91
Cell_ShapeSize_Roundness	0.85	0.85	0.82	0.82
Ratio_Nuc2Cell_NucCytoBFP_LogMeanIntBC	1.05	1.05	1.05	1.05
Ratio_Nuc2Cell_GFP_LogMeanInt	0.99	0.99	1	1.01
Nuc_NucCytoBFP_SERSpot_0	61.68	61.83	70.67	71.21
Cyto_RFP_SERRidge_4	2.94	2.95	7.39	7.3
Cell_ShapeSize_Symmetry04	0.08	0.08	0.09	0.09
Nuc_GFP_MeanIntBC	3712.05	3731.29	402.4	404.4
Nuc_GFP_SEREdge_2	106.92	106.14	83.81	84.41
Nuc_RFP_SERHole_1	6.89	6.81	31.39	30.57
Cyto_GFP_SERDark_1	35.04	36.21	52.09	54.99
Cyto_NucCytoBFP_StdInt	217.33	217.38	126.12	127.47
Nuc_RFP_SERHole_2	1.77	1.73	14.92	14.29
Nuc_RFP_SEREdge_1	81.8	82.56	234.97	232.8
Cyto_NucCytoBFP_SERValley_4	12.29	12.35	8.3	8.39
Nuc_NucCytoBFP_SEREdge_1	98.8	99.72	74.42	75.7
Cyto_NucCytoBFP_SERSaddle_2	26.6	27.24	14.59	15.14
Cell_RFP_SERBright_4	6.06	6.19	9.19	9.22
Cell_ShapeSize_MajorAxisLength	18.76	18.73	20.32	20.16
Ratio_Nuc2Cyto_NucCytoBFP_MeanIntBC	2.71	2.75	2.3	2.32
Cell_NucCytoBFP_SERSpot_1	14.04	14.19	10.14	10.3
Nuc_GFP_LogTotalInt	6.01	6.02	5.16	5.16
Cell_NucCytoBFP_SERBright_2	16.17	16.36	11.15	11.4
Cell_ShapeSize_Compactness	0.98	0.98	0.95	0.96
Nuc_NucCytoBFP_MeanIntBC	1103.59	1100.3	699.1	689.22
Ratio_Nuc2Cyto_NucCytoBFP_TotalInt	3.74	3.86	2.2	2.3

Cyto_NucCytoBFP_SERSaddle_1	35.69	36.46	22.15	22.74
Ratio_Nuc2Cell_GFP_MeanInt	0.94	0.95	1.03	1.04
Cell_RFP_SERSaddle_4	4.6	4.61	8.46	8.44
Cell_GFP2RFP_PearsonCor	0.34	0.34	0.55	0.56
Cell_GFP_CVIntBC	0.5	0.5	0.53	0.54
Cell_RFP_TotalIntBC	210059.07	209817.64	88398.45	87023.99
Nuc_ShapeSize_Compactness	1.04	1.04	1.06	1.06
Nuc_NucCytoBFP_TotalInt	471666.13	474812.34	294540.71	292877
Cell_NucCytoBFP_SERBright_1	28.45	28.72	21.15	21.46
Nuc_RFP_StdInt	112.45	113.54	99.41	100.68
Cell_GFP_SERBright_4	9.97	10.17	6.84	7.08
Cyto_GFP_MeanIntBC	3999.83	3951.73	366.56	358.09
Cell_NucCytoBFP_SERValley_4	3.05	3	2.89	2.84
Cell_GFP_SERRidge_1	38.68	39.36	35.77	36.47
Cyto_RFP_SERHole_0	104.84	105.81	147.94	149.88
Cyto_NucCytoBFP_SERSaddle_0	61.31	62.29	48.03	48.8
Cyto_NucCytoBFP_LogTotalInt	5.12	5.11	5.16	5.13
Cyto_GFP_SERValley_4	4.12	4.14	4.38	4.56
Nuc_GFP_LogMeanInt	3.47	3.47	2.63	2.63
Cell_RFP_SERSpot_0	96.44	96.56	185.63	185.68
Nuc_ShapeSize_Symmetry02	0.21	0.21	0.18	0.19
Nuc_NucCytoBFP_CVIntBC	0.29	0.29	0.23	0.24
Nuc_GFP_SERSpot_1	9.87	10.22	11.72	12.14
Nuc_GFP_CVIntBC	0.43	0.43	0.39	0.39
Cyto_NucCytoBFP_SEREdge_0	221.99	227.11	138.64	143.32
Cell_GFP_SERValley_1	24.21	24.29	31.93	32.42
Cell_GFP_TotalInt	2741116.21	2731474.94	346417.42	340295.62
Ratio_Nuc2Cell_RFP_LogMeanInt	1.02	1.02	0.96	0.97
Cyto_RFP_SERSaddle_1	27.29	27.72	62.36	62.44
Cyto_RFP_LogMeanIntBC	2.32	2.31	2.04	2.03
Cell_NucCytoBFP_CVInt	0.47	0.47	0.38	0.39
Ratio_Nuc2Cyto_NucCytoBFP_LogMeanIntBC	1.16	1.16	1.14	1.15
Nuc_RFP_SERHole_0	82.7	82.4	190.6	188.23
Cell_GFP_SERValley_2	10.88	10.79	12.59	12.74
Cell_GFP_SERSaddle_4	5.79	5.76	4.84	4.91
Cyto_GFP_LogTotalIntBC	5.89	5.88	5.04	5.01
Cyto_NucCytoBFP_SERDark_4	11.86	11.87	8.41	8.46
Cell_RFP_StdInt	156.1	157.2	142.6	141.55
Nuc_GFP_SERValley_1	30.21	29.69	29.8	29.46
Cyto_RFP_SERSaddle_2	15.18	15.52	25.97	26.17

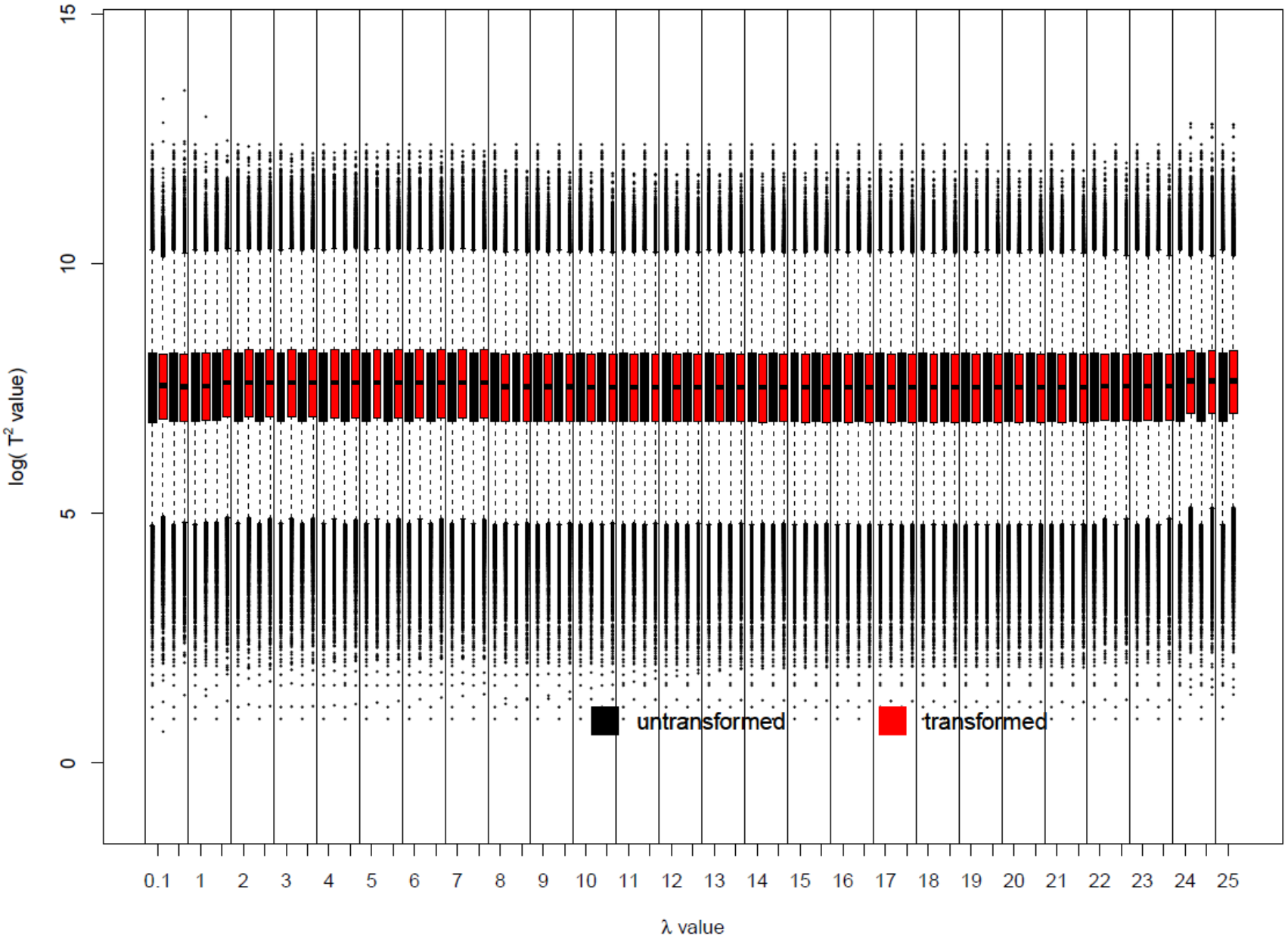
Cell_GFP_SERSaddle_0	48.64	49.35	66.59	67.03
Cyto_RFP_SERSpot_4	2.05	2.04	4.32	4.32
Cyto_GFP_StdInt	1395.68	1399.56	205.38	206.37
Nuc_GFP_StdIntBC	1254.97	1266.51	149.35	150.97
Cell_RFP_CVIntBC	0.53	0.54	1.3	1.3
Cell_GFP_SEREdge_4	66.83	67.01	52.33	53.36
Cell_NucCytoBFP_SERSpot_4	4.81	4.91	3.31	3.4
Cyto_RFP_SERValley_0	99.53	100.71	193.34	194.36
Nuc_GFP_SERSpot_0	47.26	48.46	92.93	93.51
Cyto_RFP_CVInt	0.43	0.43	0.78	0.78
Cell_NucCytoBFP2RFP_PearsonCor	0.47	0.48	-0.07	-0.05
Ratio_Nuc2Cell_GFP_TotalInt	0.54	0.55	0.51	0.53
Cell_GFP_SEREdge_2	106.32	107.13	91.87	93.82
Cyto_GFP_SERSaddle_2	18.16	18.58	16.35	17.08
Nuc_NucCytoBFP_LogTotalInt	5.6	5.6	5.41	5.41
Cell_RFP_LogTotalIntBC	5.25	5.25	4.85	4.84
Cell_GFP_SERValley_0	62.33	63.22	97.73	98.43
Nuc_GFP_SERHole_2	11.17	10.93	7.65	7.46
Cyto_GFP_SEREdge_4	85.47	87.45	68.98	72.4
Cyto_RFP_SERSpot_0	100.68	100.79	186.59	186.49
Nuc_RFP_SERSpot_1	11.83	12.01	25.56	26.14
Nuc_GFP_SERRidge_2	17.39	17.93	14.84	15.47
Cyto_GFP_SEREdge_0	232.71	239.97	218.47	225.59
Cyto_NucCytoBFP2RFP_PearsonCor	0.71	0.71	0.35	0.37
Cyto_RFP_SERBright_4	4.34	4.35	10.21	10.13
Nuc_NucCytoBFP_SERValley_0	39.32	39.5	44.15	44.56
Cell_NucCytoBFP_SERSaddle_2	11.46	11.56	8.06	8.23
Nuc_RFP_SERRidge_1	22.56	22.89	44.59	45.45
Cell_RFP_SERRidge_0	80.65	80.83	174.62	174.55
Cyto_NucCytoBFP_SEREdge_2	163.27	167.2	91.49	95.29
Nuc_NucCytoBFP_SEREdge_0	116.7	117.63	96.4	97.64
Ratio_Nuc2Cyto_GFP_LogTotalIntBC	1.02	1.02	1.01	1.02
Cyto_NucCytoBFP_SERSaddle_4	17.95	18.44	9.52	9.95
Ratio_Nuc2Cyto_GFP_LogMeanIntBC	0.99	0.99	1.03	1.04
Cell_GFP_SEREdge_1	140.27	141.86	131.94	134.6
Ratio_Nuc2Cell_RFP_LogTotalInt	0.96	0.96	0.92	0.92
Ratio_Nuc2Cyto_GFP_LogMeanInt	0.99	0.99	1.02	1.03
Cell_NucCytoBFP_SERSaddle_1	17.63	17.77	13.83	14
Nuc_RFP_MeanInt	465.12	466.56	146.39	148.46
Cell_RFP_SERDark_0	146.36	146.37	313.79	313.68

Cell_RFP_SERSpot_2	5.7	5.8	11.63	11.7
Cyto_NucCytoBFP_SERSpot_1	5.95	6.06	4.14	4.18
Cell_ShapeSize_Perimeter	94.41	94.2	104.38	103.35
Cell_GFP_SERHole_1	11.51	11.54	13.41	13.35
Nuc_RFP_LogTotalIntBC	5.05	5.05	4.37	4.38
Cell_GFP_SERRidge_2	20.82	21.17	16.33	16.74
Cyto_GFP_SERSpot_4	3.64	3.71	2.71	2.77
Cyto_RFP_SEREdge_4	69.16	70.79	65.6	66.71
Nuc_GFP_SERHole_1	18.6	18.35	16.39	16.17
Cyto_GFP_SERBright_1	51.48	52.13	43.78	43.73
Cell_NucCytoBFP_SERDark_1	20.05	20.04	16.86	16.83
Nuc_RFP_CVIntBC	0.33	0.33	1.29	1.27
Cyto_GFP_SERValley_1	31.77	32.88	46.44	49.51
Cyto_NucCytoBFP_LogTotalIntBC	5.02	5.01	5.01	4.99
Ratio_Nuc2Cell_GFP_LogMeanIntBC	0.99	0.99	1	1.01
Cyto_NucCytoBFP_SERSpot_0	64.18	64.59	75.88	76.48
Cell_GFP_SERSpot_0	54.1	55.31	96.19	96.59
Cyto_RFP_SERDark_1	35.03	35.9	68.14	69.11
Cell_RFP_SERValley_4	1.97	1.93	6.78	6.68
Cyto_RFP_MeanIntBC	237.06	233.41	134.17	130.81
Nuc_GFP_SERBright_2	19.94	20.61	16.97	17.79
Nuc_RFP_SEREdge_2	62.35	62.96	165.99	163.37
Cell_RFP_SERBright_2	14.07	14.29	29.7	29.7
Cell_RFP_CVInt	0.37	0.37	0.8	0.79
Cell_NucCytoBFP_SERSpot_0	62.58	62.8	73.23	73.76
Cyto_GFP_SERBright_2	26.28	26.57	19.7	19.77



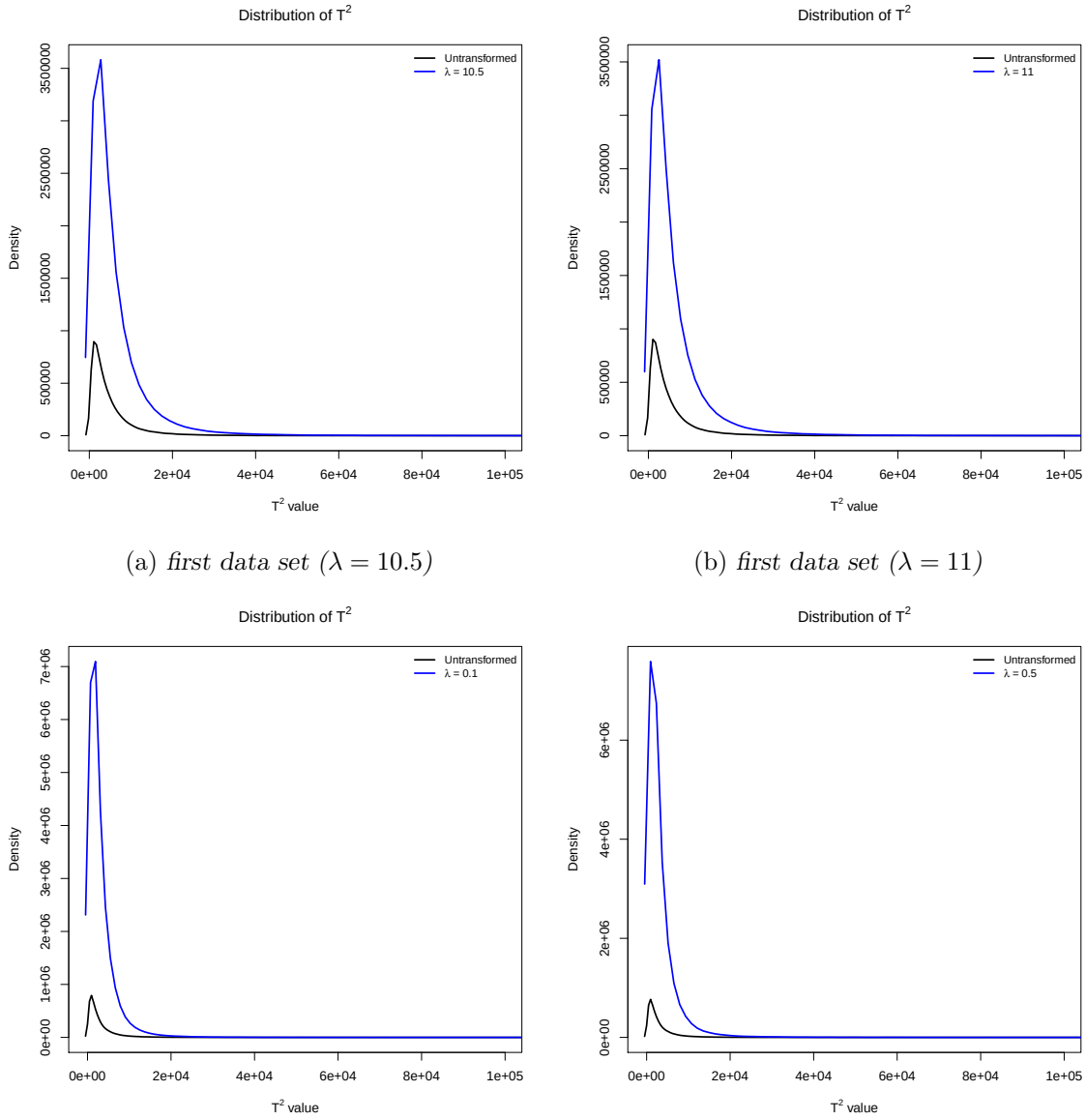
First data set

Figure 18: Boxplots of $\log(T^2)$ statistics after transformation compared to untransformed cases



Second data set

Figure 18: Boxplots of $\log(T^2)$ statistics after transformation compared to untransformed cases



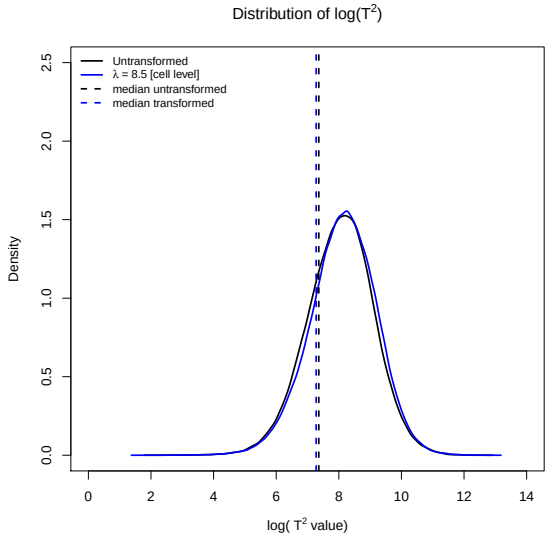
(a) *first data set* ($\lambda = 10.5$)

(b) *first data set* ($\lambda = 11$)

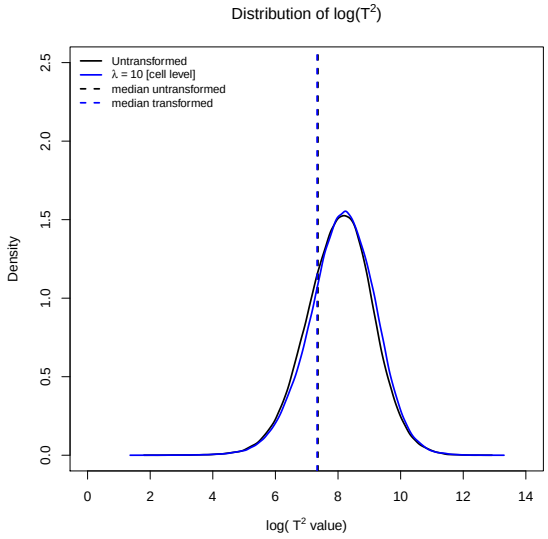
(c) *second data set* ($\lambda = 0.1$)

(d) *first data set* ($\lambda = 0.5$)

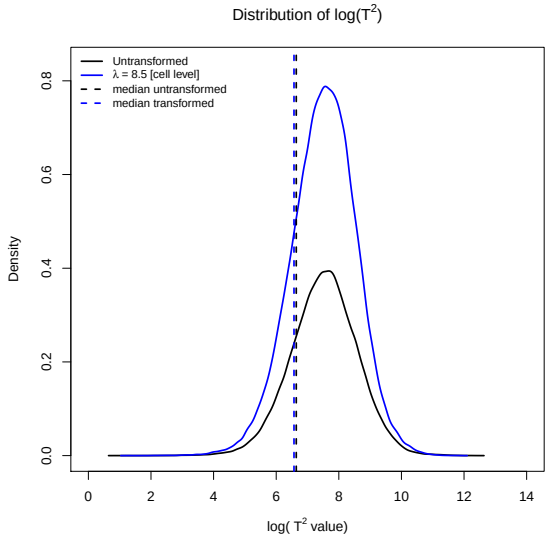
Figure 19: Distribution of T^2 statistics after transformation compared to untransformed cases



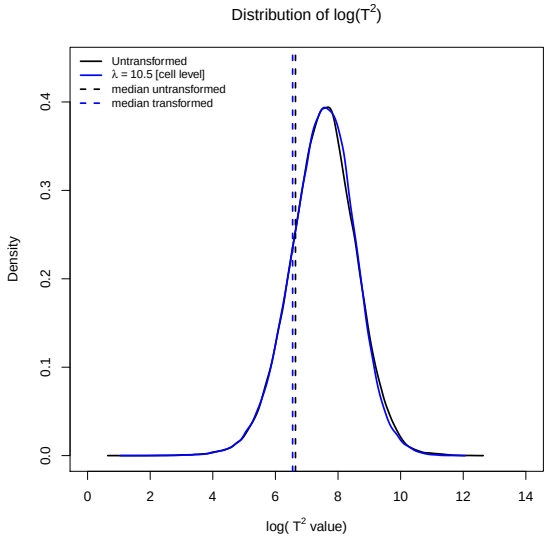
Transformation on first data set ($\lambda = 8.5$)



Transformation on first data set ($\lambda = 10$)

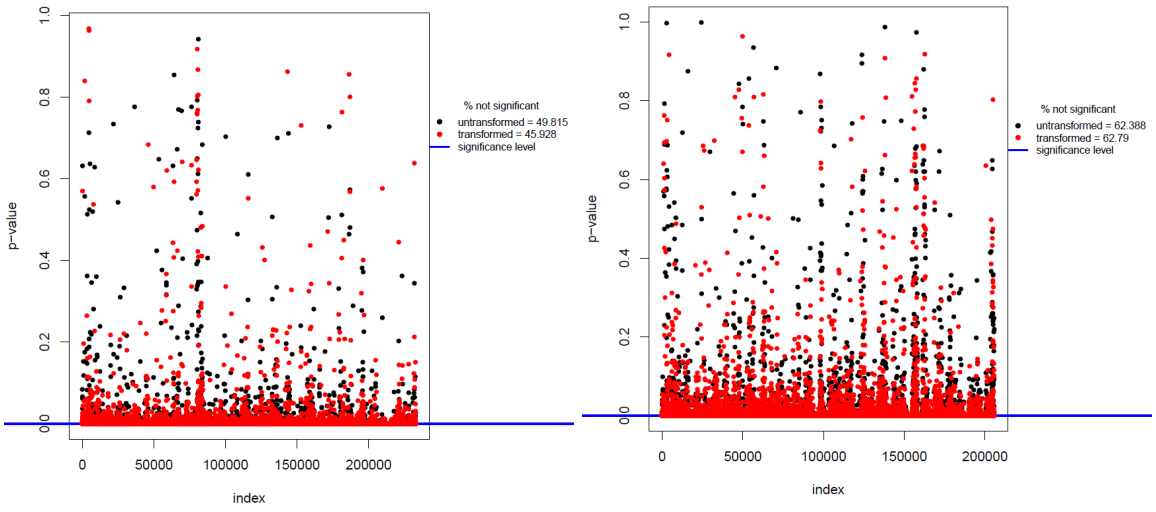


Second data set ($\lambda = 8.5$)



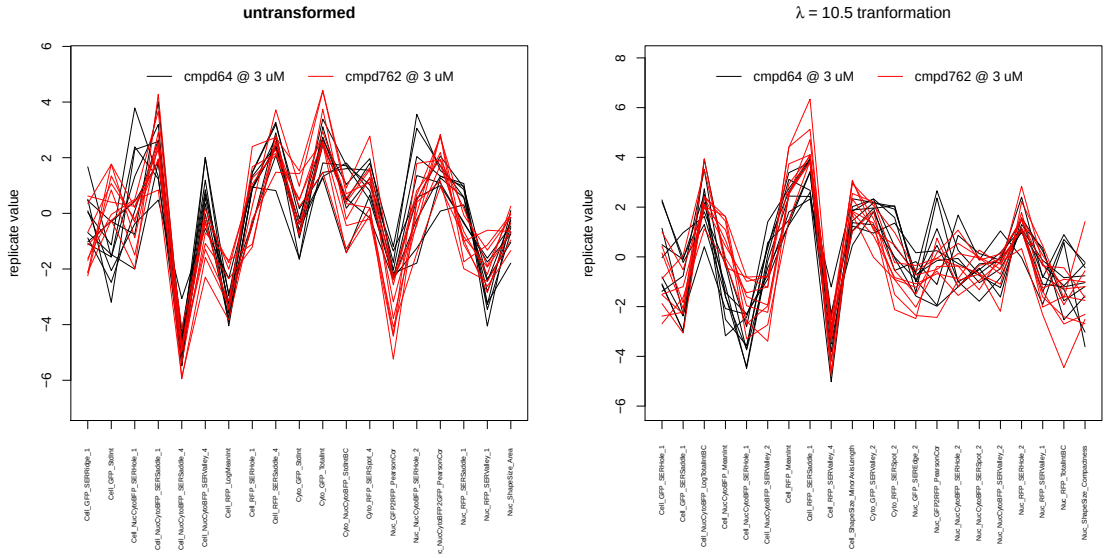
Second data set ($\lambda = 10.5$)

Figure 20: Distribution of $\log(\mathrm{T}^2)$ statistics after transformation compared to untransformed cases- less effective transformations



First data set ($\lambda = 10.5$) Second data ($\lambda = 0.5$)

Figure 21: Effect of transformations on proportion of significant T^2 statistics



(a) (b)

Figure 22: Treatment profiles before and after transformation (first data set, $\lambda = 10.5$)

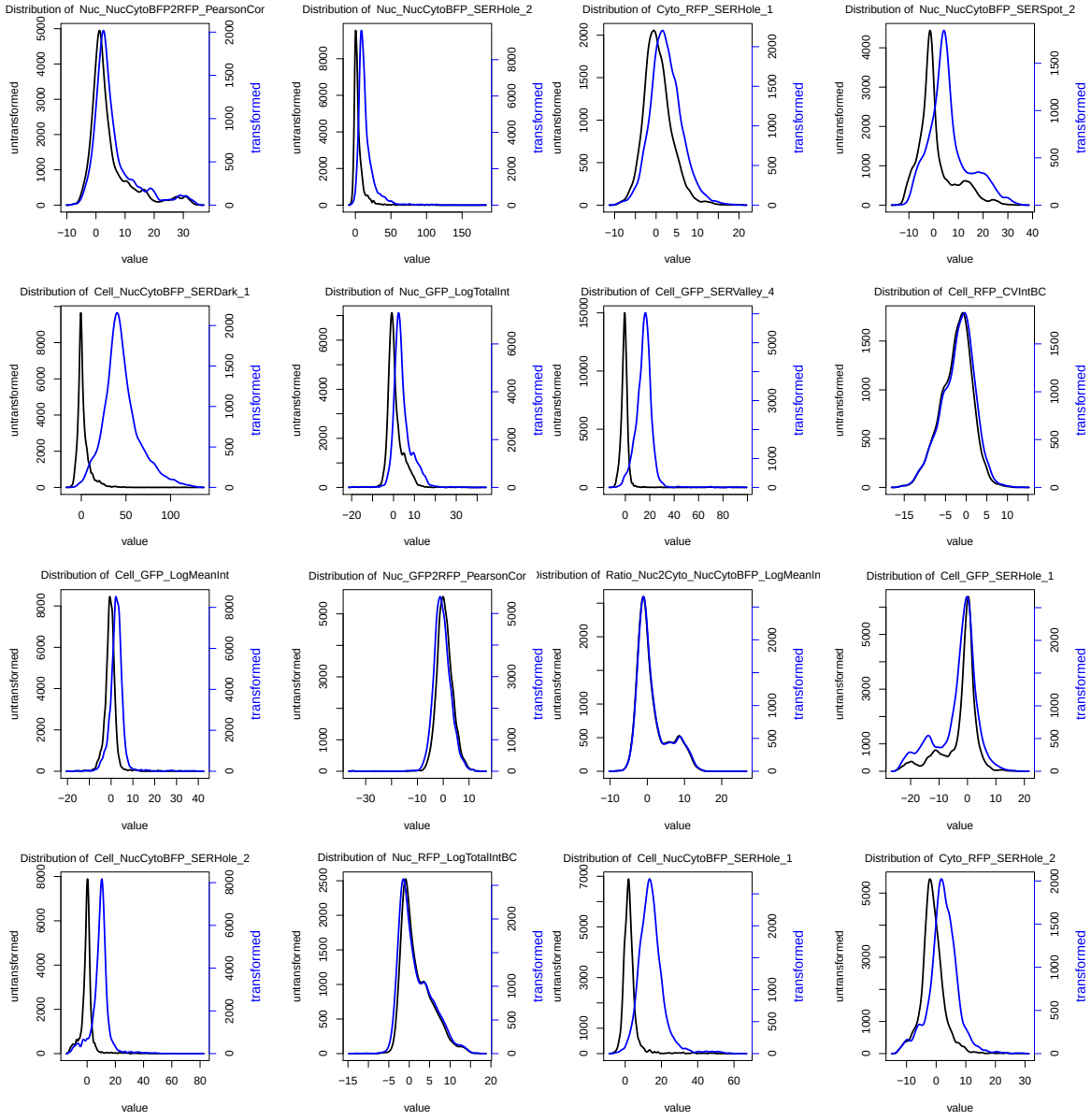
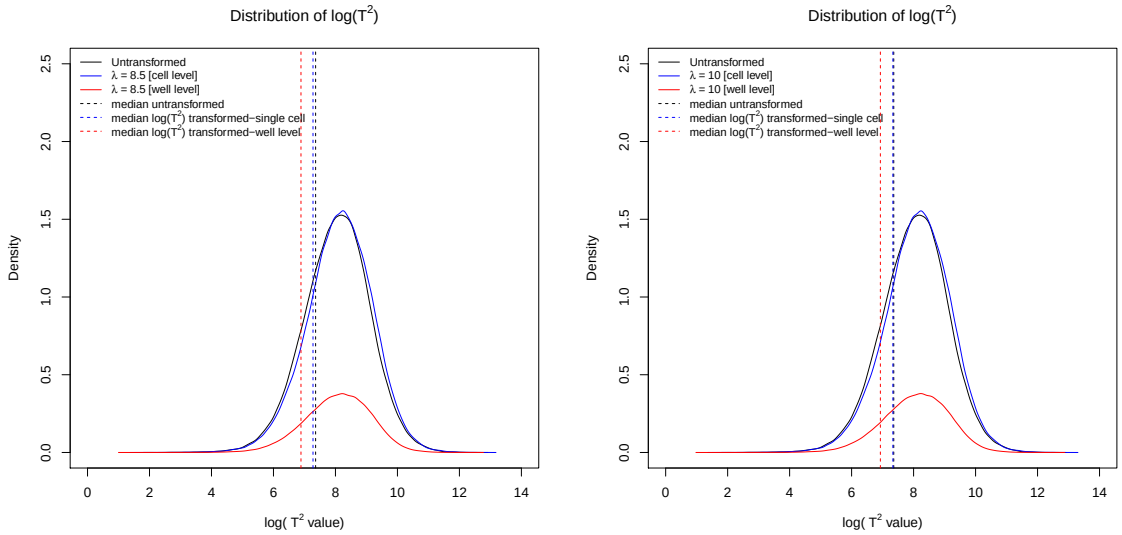
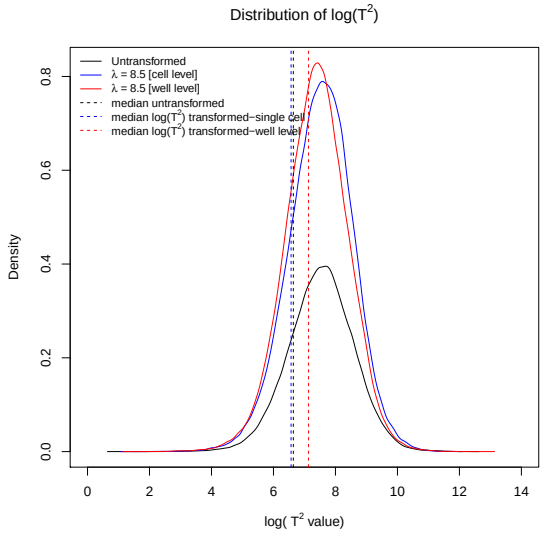


Figure 23: Data set 2 - feature distributions before (black line) and after (blue) transformation ($\lambda = 0.5$)

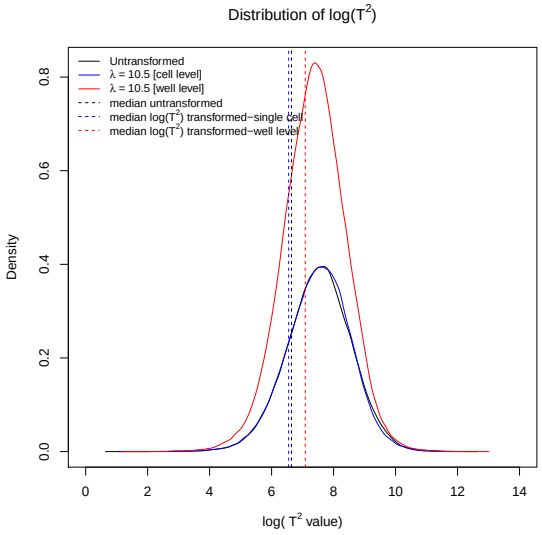


(a) *first data set* ($\lambda = 8.5$)

(b) *first data set* ($\lambda = 10$)

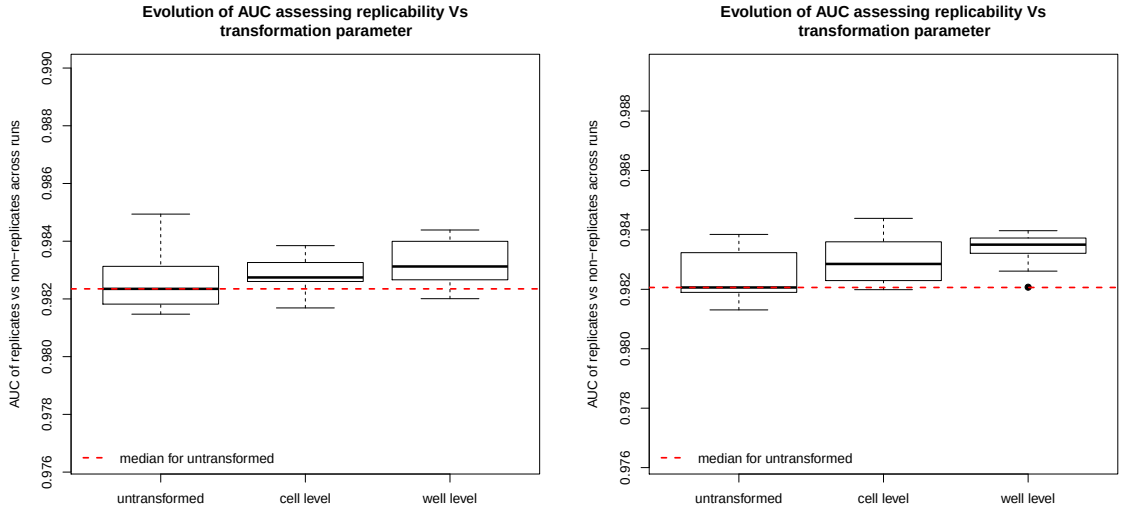


(c) *second data set* ($\lambda = 8.5$)



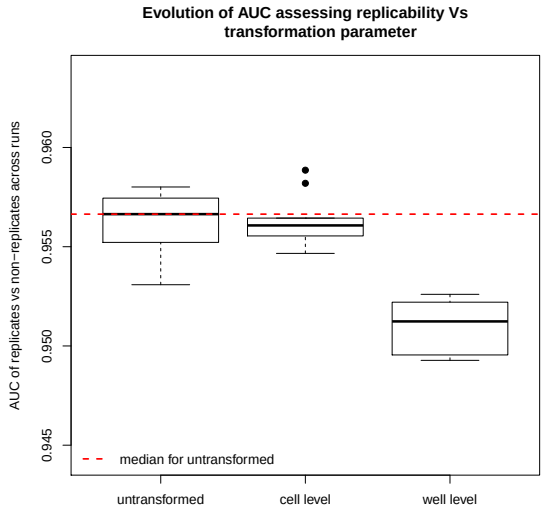
(d) *second data set* ($\lambda = 10.5$)

Figure 24: Distribution of T^2 statistics comparing untransformed, cell and well level performance

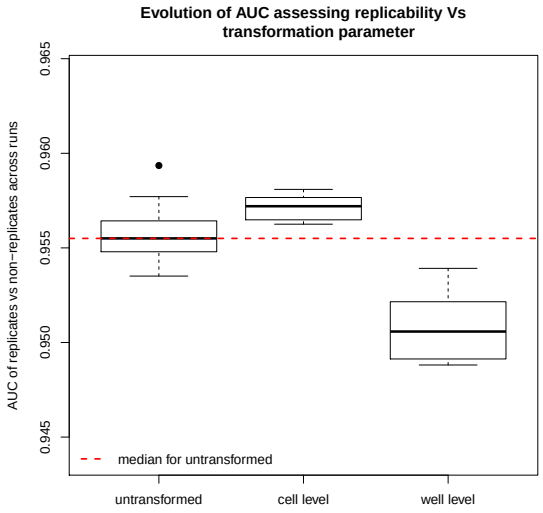


(a) *First data set* ($\lambda = 8.5$)

(b) *First data set* ($\lambda = 10$)



(c) *Second data set* ($\lambda = 8.5$)



(d) *Second data set* ($\lambda = 10.5$)

Figure 25: Boxplots of AUC values across runs for transformed, cell and well level transformed data

CODES

Codes used in this analysis can be accessed via `https://github.com/lwafula/MSc.-Thesis`