Using Control Genes to Correct for Unwanted Variation in Microarray Data

Johann A. Gagnon-Bartsch* Terence P. Speed*†

March 26, 2011

Abstract

Microarray expression studies suffer from the problem of batch effects and other unwanted variation. Many methods have been proposed to adjust microarray data to mitigate the problems of unwanted variation. Several of these methods rely on factor analysis to infer the unwanted variation from the data. A central problem with this approach is the difficulty in discerning the unwanted variation from the biological variation that is of interest to the researcher. We present a new method, intended for use in differential expression studies, that attempts to overcome this problem by restricting the factor analysis to negative control genes. Negative control genes are genes known *a priori* not to be differentially expressed with respect to the biological factor of interest. Variation in the expression levels of these genes can therefore be assumed to be unwanted variation. We name this method "Remove Unwanted Variation, 2-step" (RUV-2). We discuss various techniques for assessing the performance of an adjustment method, and compare the performance of RUV-2 with that of other commonly used adjustment methods such as Combat and SVA. We present several example studies, each concerning genes differentially expressed with respect to gender in the brain, and find that RUV-2 performs as well or better than other methods. Finally, we discuss the possibility of adapting RUV-2 for use in studies not concerned with differential expression, and conclude that there may be promise, but substantial challenges remain.

1 Introduction

Microarray expression studies are plagued by the problem of unwanted variation. In addition to the biological factor(s) that are of interest to the researcher, other factors, both technical and biological, influence observed gene expression levels. A typical example is a *batch effect*, which can occur when some samples are processed differently than others. For example, significant batch effects may arise when some samples are processed in a different laboratory, by a different technician, or even just on a different day (Leek et al., 2010; Scherer, 2009). Though infamous, batch effects are not the only source of unwanted variation. Other sources of unwanted technical variation can occur *within* batches and be just as problematic. Moreover, unwanted biological variation can be a problem as well.

Several methods have been proposed to adjust microarray data to mitigate the problems of unwanted variation. Despite substantial progress, there is still no "silver bullet" and perhaps never will be. As such, there remains a need for both improved methods and ways to evaluate the relative strengths of existing methods. Our primary goal in this paper is to contribute a new method based on *control genes*, and to encourage the use of control genes more generally. A secondary goal is to present some techniques we have found to be useful for comparing the performance of different adjustment methods. Finally, a less explicit though still important theme of this paper is that we believe that the most appropriate way to deal with unwanted variation depends critically on the final goal of the analysis — e.g. differential expression, classification, or clustering.

*Department of Statistics, University of California at Berkeley, Berkeley, CA 94720
†Bioinformatics Division, Walter and Eliza Hall Institute, Victoria 3050, Australia

In what remains of the introduction, we present a brief summary of existing methods to adjust for unwanted variation, followed by a brief summary of our own method. In Section 2 we present techniques to compare the performance of these methods, and in Section 3 provide examples. Details of our method are left for Section 4.

Methods to adjust for unwanted variation can be divided into two broad categories. In the first category are methods that can be used quite generally, and provide a *global adjustment*. An example would be quantile normalization, which is generally regarded as a self-contained step, and plays no role in the downstream analysis of the data. In the second category are *application specific* methods that incorporate the batch adjustment directly into the main analysis of interest. For example, in a differential expression study, batch effects may be handled by explicitly adding batch terms to a linear model. The method we present in this paper falls into this second category, where the application is differential expression.

Most of the progress that has been made with application-specific methods has been for differential expression studies, and has made use of linear models. Some methods presume the batches to be known; in this case the effects of the known batches can be directly modeled. Combat is one such successful and well-known method; in particular Combat has been shown to work well with small datasets (Johnson, Li, and Rabinovic, 2007). While Combat and other similar methods can be quite successful, they also have limitations. One limitation is that the batches must be known; in many situations this is not the case. Another limitation is that even when batch information is known, it may give only a partial hint of the underlying unwanted variation. This is because it is not "batch" itself that causes "batch effects," but rather some other physical variable that is correlated with batch. As a simple example, suppose changes in the operating temperature of the scanner leads to unwanted variation, and consider a study in which some samples are processed at lab A and the rest at lab B. If the scanner at lab A generally runs cooler than the scanner at lab B, this will lead to a "batch effect." However, if the operating temperature at lab A is itself quite variable, this will lead to additional within-batch unwanted variation that is not captured in the simple batch model. Section A in the supplementary material[1] (SM) provides a brief example from the MAQC study (Shi et al., 2006) of substantial within-batch unwanted variation.

Other linear-model based methods presume the sources of the unwanted variation to be unknown. These methods attempt to infer the unwanted variation from the data, and then adjust for it. Often this is accomplished via some form of factor analysis; several factors believed to capture the unwanted variation are computed and then incorporated into the model in just the same way known confounders are incorporated. In the simplest approach, factors are computed directly from the observed expression matrix by means of a singular value decomposition or some other factor analysis technique. This is often successful in practice, but can be dangerous — if the biological effect of interest is large, it too will be picked up by the factor analysis, and removed along with the unwanted variation. In other words, if one adjusts for unwanted variation by removing the first several principal components (PCs) from the data, one may very well throw out the baby with the bathwater. This problem has been acknowledged, and several attempts have been made to avoid it. One of the most well-known methods that directly address this problem is SVA (Leek and Storey, 2007, 2008). Other methods of potential interest include Kang et al. (2010), Kang, Ye, and Eskin (2008a), Kang et al. (2008b), Listgarten et al. (2010), Mecham, Nelson, and Storey (2010), Price et al. (2006), Stegle et al. (2008), Yu et al. (2005). Some of the first uses of factor analysis to adjust for unwanted variation can be found in Alter, Brown, and Botstein (2000) and Nielsen et al. (2002), although in these examples there is no explicit linear model.

Our strategy is to use control genes. Negative control genes are genes whose expression levels are known *a priori* to be truly unassociated with the biological factor of interest. Conversely, positive control genes are genes whose expression levels are known *a priori* to be truly associated with the factor of interest. For example, if the factor of interest is the presence or absence of ovarian cancer, CyclinE would be a positive control. Negative control genes are in general harder to identify with certainty. So-called housekeeping genes are often good candidates, but not always. Another example of negative controls are the spike-in controls found on many microarray platforms.[2]

[1] The supplementary material can be found at the end of this document.

[2] Two notes on terminology: 1) When we refer generally to "negative control genes" or simply "control genes" we often

Our method is to perform factor analysis on just the negative control genes, and incorporate the resulting factors into a linear regression model. The idea is that since the negative control genes are known to be unassociated with the factor of interest, there is no danger in picking up any of the relevant biology in the factor analysis step, and thus no danger of throwing out the baby with the bathwater. We name this method RUV-2, for "Remove Unwanted Variation, 2-step" — the two steps are the factor analysis and the regression.

2 Criteria for a Good Adjustment

Techniques to evaluate the quality of an adjustment are in many ways as important as the adjustment method itself. The statistical models on which adjustment methods are based are artificial. The models are most useful as sources of inspiration for improved methods; they are substantially less useful in proving the worth of a method. In the end, an adjustment method must prove its value by working in practice.

The question thus arises of how to know whether an adjustment is helping or hurting. This is not trivial. In many cases, evidence that seems to suggest an adjustment method is helping (or hurting) is actually ambiguous. As an example, consider a differential expression study, and consider assessing the quality of the study by counting the number of genes "discovered" at a certain FDR threshold. If the unwanted variation is roughly orthogonal to the factor of interest, the unwanted variation will manifest itself as additional "noise" that obscures any true association between the factor of interest and gene expression levels. An effective adjustment method would therefore increase the number of discovered genes. On the other hand, if the unwanted variation is correlated with the factor of interest, this will introduce spurious associations between the factor of interest and gene expression levels. An effective adjustment method would therefore decrease the number of discovered genes. As a second example, consider a classification study in which a researcher wants to classify tumor samples into one of several tumor sub-types. Suppose the researcher wants to test her classification algorithm on a set of tumor samples in which the sub-type is known, and she does her test once in combination with a method that adjusts for unwanted variation, and once without. Suppose the rate of misclassification is higher when the adjustment method is used. This would seem to suggest the adjustment method is hurting. However, it is equally possible that the adjustment method is working — if the tumor sub-types were processed in batches, the resulting batch effects could artificially help the classifier.

In the following few sections we present some techniques that we have found to be useful in evaluating the quality of an adjustment.[3] Only the first technique provides a (nearly) unambiguous assessment of the quality of an adjustment. However, its applicability is limited. The other two techniques are also very informative, if not entirely definitive, and can be used in a wider variety of situations.

2.1 Control Genes / Gene Rankings

Positive control genes can be used for quality assessment in differential expression studies. After computing p-values for each gene, we can rank the genes in order of increasing p-value. Positive controls should be towards the top of this list. We can therefore use the number of positive controls ranked in (for example) the top 50 genes as a quality metric. If an adjustment method substantially increases the number of top-ranked positive controls, we have reason to believe the method is effective.[4]

use the term to include the spike-in control probesets as well, despite the fact they are not genes. 2) In other contexts (e.g. Illumina) the term "negative controls" is used to denote probes that should never be expressed in any sample. This usage of the term "negative controls" is different than ours, and should not be confused.

[3] One fairly common technique that we do *not* discuss is clustering. In some circumstances, performing a cluster analysis both before and after adjustment and observing whether the adjustment causes samples to cluster more strongly by biology (or less strongly by batch) can be a highly effective way to assess the quality of the adjustment. Indeed, we use this technique ourselves in Section 3.4. However, there are also many circumstances in which clustering can be deceptive (the classification example we provide above applies just as well to clustering). We feel a full discussion is beyond the scope of this paper. In general, we feel that clustering is more helpful as an exploratory tool, and less helpful as a test of the quality of an adjustment.

[4] Even here, however, the evidence, strictly speaking, is not entirely conclusive. For example, consider a simple hypothetical adjustment method that simply shrinks gene-specific variances to a common average variance. This would have the effect of systematically increasing the magnitude of the t statistics for highly variable genes. If the positive controls all happened to be highly variable, the end result would be that the rankings of the positive controls improve, despite the fact that the

Note that we use the ranks of the p-values and not the p-values themselves. This is for reasons discussed above; a good adjustment may increase or decrease the positive controls' p-values depending on the nature of the unwanted variation. Ranking helps to resolve the ambiguity.

While ranking p-values is generally preferred, there remain some situations in which it may be better to look at the p-values themselves. An example might be when only a very small number of positive controls are available, and their rankings do not change substantially after the adjustment. In this case, one might wish to examine the p-values of both positive and negative controls. If the p-values of the positive controls substantially decrease — and the p-values of the negative controls do not — this would suggest the adjustment helps. On the other hand, if the p-values of the negative controls decrease as well, this may simply suggest an artifact of the adjustment method. Likewise, if the p-values of the positive and negative controls both increase, the result is ambiguous, but the technique decribed in 2.2 might help clarify matters.

Some caution is required when using negative controls to assess the quality of an adjustment. After all, if the method of adjustment is to fit and remove variation characteristic of a set of negative controls, then observing that the adjustment diminishes the association between the factor of interest and the negative controls is simply to be expected. A better strategy would be to use two different sets of negative controls — one to make the adjustment, and one to use in assessing the quality of the adjustment. Preferably, the two sets of negative controls will be different from each other in some important way. For example, we might use spike-in controls to make an adjustment, and housekeeping genes to assess the quality of the adjustment, or vice versa.

2.2 The p-value Distribution

Consider a differential expression study in which the factor of interest is assumed to be associated with the expression level of only a fraction of genes. The distribution of the p-values for the genes that are unassociated with the factor of interest would ideally be uniformly distributed over the unit interval, whereas the p-values for the genes that are associated with the factor of the interest will ideally be nearly 0. Thus, a histogram of the p-values will ideally be nearly uniform, with a spike near 0. In practice however, this is uncommon, as unwanted variation tends to introduce dependence across measured gene expression levels. Since adjusting for unwanted variation should remove this dependence, we might expect a good adjustment to result in p-value histograms closer to the "ideal."

2.3 RLE Plots

RLE (relative log expression) plots are boxplots that can be used to determine the overall quality of a dataset, and, in particular, identify bad chips. Consider a set of m chips, each with n genes, and denote the log expression level of the j^{th} gene on the i^{th} chip by y_{ij}. Denote the j^{th} column of the matrix (y_{ij}) by $y_{\star j}$. For each of the n genes we can calculate median$(y_{\star j})$, the median (over the m chips) log expression level. For each gene on each chip, we can then calculate $y_{ij} - \text{median}(y_{\star j})$, the deviation from the median gene expression level. For each chip, we can then produce a box plot of its n deviations. In most cases, if the chip is of good quality, this boxplot will be centered around zero and its width (IQR) will be around 0.2 or less. Examples of RLE plots can be found in Figure 1 (to be discussed more fully later). More information about RLE plots can be found in Bolstad et al. (2005) and Brettschneider et al. (2008).

3 Examples

We present four examples. In the first three we discover genes that are differentially expressed in the brain with respect to gender. The fourth example involves clustering tumors of known types. We chose these examples because "truth" is in some sense known.

We chose differential expression with respect to gender because it provides us with a clear set of potential positive controls — in this case, genes that are located on the X and Y chromosomes. Treating X and Y

"adjustment" is not really correcting for unwanted variation at all.

genes as positive controls and using our technique of counting the number of top-ranked positive controls (Section 2.1) allows us to compare the performance of various adjustment methods. We chose the brain because of its comparatively complex biology — a very large fraction of genes are expressed in the brain — and the availability of several interesting datasets. Indeed, our lead example is ideal, as the study was originally intended to discover differentially expressed genes in the brain,[5] and the data exhibit profound batch effects.

In all of our examples, a few practical issues must be considered. The first is what preprocessing should be done. Microarray data routinely go through three stages of preprocessing — background correction, normalization, and summarization. Several algorithms have been proposed for each of these steps. For simplicity, we limit ourselves to the "standard" sequence of algorithms used in RMA (Bolstad et al., 2003; Irizarry et al., 2003a,b). The preprocessing steps – particularly the quantile normalization – are nonlinear, and it is not clear how they might interact with the adjustment methods. Thus, we repeat many of our analyses omitting one or more stages of preprocessing in order to see what happens.

The second issue to consider is which negative controls to use. To effectively adjust for batch effects, our negative controls must both 1) be uninfluenced by the factor(s) of interest and 2) be influenced by the unwanted factors. In other words, they must actually be negative controls, and their expression levels must accurately reflect the unwanted variation. We focus on two classes of possible negative controls — housekeeping genes and spike-in controls. The housekeeping genes we use are those discovered in Eisenberg and Levanon (2003).[6] A good discussion of both spike-in controls and housekeeping genes can be found in Lippa et al. (2010).

3.1 Gender Study

Vawter et al. (2004) conducted a study to find genes differentially expressed in the brain with respect to gender. Samples were taken post-mortem from the brains of 10 individuals. Three samples were taken from each individual — one from the anterior cingulate cortex, one from the dorsolateral prefrontal cortex, and one from cortex of the cerebellar hemisphere. One aliquot of each sample was sent to each of three laboratories for analysis. The analyses were done using either Affymetrix HG-U95A or Affymetrix HG-U95Av2. We are unaware of how the decision was made to use which platform for which analysis. One of the laboratories used only HG-U95Av2. Note that there should have been $10 \times 3 \times 3 = 90$ chips total. However, six of the combinations were missing, leaving us with 84 chips.[7] Data are available on GEO (GSE2164).

The HG-U95A platform has 12626 probesets, and the HGu95Av2 platform has 12625 probesets. We identified 12600 probesets that were shared between the two platforms. We did not, however, attempt to map individual probes from one platform to the other. Since preprocessing (background correction, quantile normalization, summarization) requires probe level data, we did these preprocessing steps for each platform separately. As a result, large differences remained between the different platform types even after the standard preprocessing.[8] The raw HG-U95Av2 expression values are generally larger than their HG-U95A analogs by about a factor of 4, so the $\log_2$ expression values for the HG-U95Av2 expression values are generally greater than those of the HG-U95A by about 2. We therefore added an additional preprocessing step after summarization; we performed a location / scale adjustment in which we linearly re-scaled the data so that each chip had the same mean and standard deviation. See Figure 1 for RLE plots at different stages of preprocessing — no preprocessing; background correction and quantile normalization only (BG + QN); background correction, quantile normalization, and location / scale adjustment (BG + QN + LS). It

[5] In this paper we are primarily concerned with methodology, not biology, and we do not discuss the specific genes we find to be differentially expressed. Readers interested in the particular genes that are differentially expressed can find tables in the SM.

[6] Housekeeping genes are genes that are essential to basic cellular activities, such as metabolism. These genes are therefore expressed in all cells. The strategy used in Eisenberg and Levanon (2003) to discover housekeeping genes was to examine gene expression levels in many different tissues, and see which genes were expressed in all of the tissues.

[7] Additionally, 3 of the combinations were replicated, so in fact there are 87 chips available on GEO. We omitted the replicates in our analysis.

[8] In this respect, the situation is similar to the one found in Nielsen et al. (2002), one of the first uses of factor analysis to remove unwanted variation, where the primary source of unwanted variation was chip type.

is important to note that the vertical scale of the RLE plots in Figure 1 is substantially different than that of all other RLE plots in this paper.

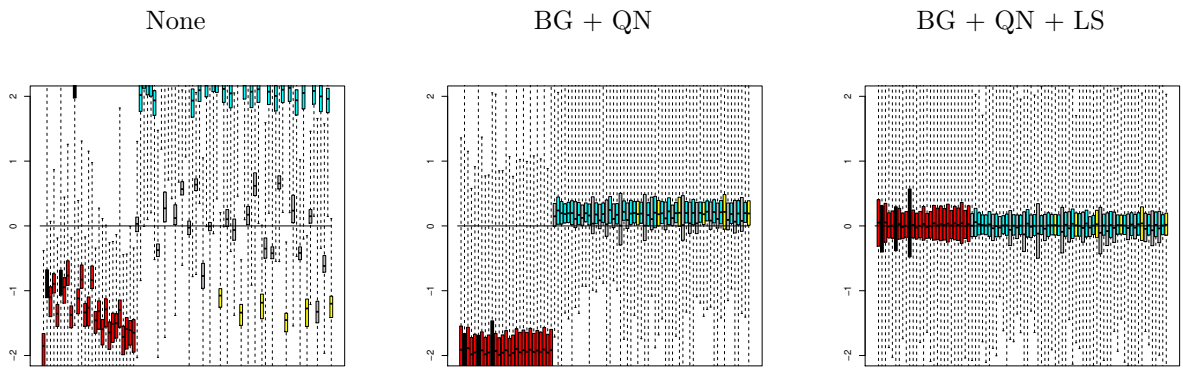


Figure 1: Gender study RLE plots at different stages of preprocessing. From left to right: No preprocessing; background correction / quantile normalization done separately for each platform type; background correction / quantile normalization followed by a final location/scale adjustment across all chips. Coloring: red – site A, HG-U95A; yellow – site A, HG-U95Av2; black – site B, HG-U95A; gray – site B, HG-U95Av2; cyan – site C, HG-U95Av2. NOTE: The scale on the y-axis is different for these RLE plots than for all other RLE plots in this paper.

The unwanted variation apparent in Figure 1 is striking. Even ignoring the differences between chip types, observed expression levels differ by up to an order of magnitude between laboratories. There is substantial within-laboratory unwanted variation as well. The average observed expression level varies from chip to chip by roughly a factor of two within laboratories. As a result, discovering genes that are differentially expressed with respect to gender is nearly impossible without adjusting for the unwanted variation. On un-preprocessed, unadjusted data, every gene has an FDR-adjusted p-value of approximately 1, and only 7 of the top 60 genes come from the X or Y chromosome. The preprocessing helps; after preprocessing (but no other adjustments), 15 of the top 60 genes are from the X or Y chromosome, and 8 of these have FDR-adjusted p-values that are significant at the 0.05 level. Even after preprocessing, however, substantial unwanted variation persists, as can be seen in RLE, p-value, and scree plots.

A critical step in RUV-2 is determining the number k of factors to remove. In general, this is difficult, and there is no clear way to determine k. We recommend pursuing several approaches and exercising judgment. We have found RLE plots and p-value histograms to be helpful.[9] In addition, if any positive controls are known, these should be used as well. To make use of RLE plots and p-value plots it is necessary to complete the analysis for several values of k and then examine the plots to evaluate the quality of the results. Several such plots can be found in the SM (Figures 11 and 12); based on these plots, we choose a k of 10 — see Figure 2.

Several questions remain: Which factor analysis method is best? To what extent does performance depend on k? Does RUV-2 work well in combination with Limma? Do we achieve a better adjustment using housekeeping genes or spike-in controls? How does RUV-2 compare to other adjustment methods such as standard linear regression, Combat, or SVA? How does preprocessing affect performance? We address each of these questions in turn.

Given the central role of factor analysis in RUV-2, one might expect that the choice of factor analysis method is quite important. In our examples this turns out not to be the case. We repeated our analysis with

[9]Many people find scree plots to be helpful as well (Venables and Ripley, 2002). In our examples, we found scree plots to be of questionable value. However, we provide them in the SM.

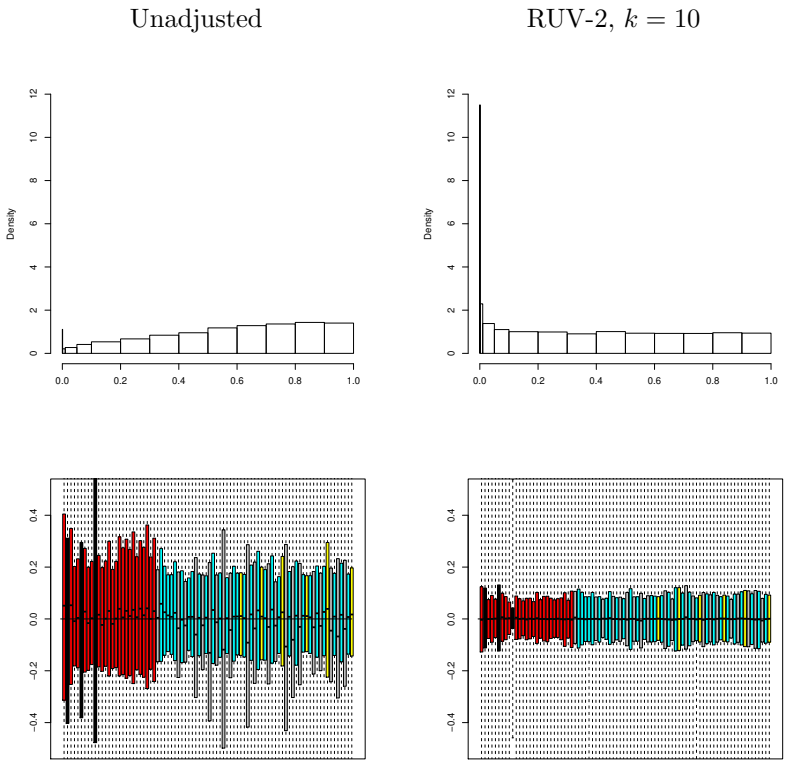


Figure 2: Gender study p-value histograms and RLE plots before and after adjustment. Histogram breakpoints are at 0.001, 0.01, 0.05, and 0.1, 0.2, 0.3, etc. Data were fully preprocessed (BG+QN+LS). The factors were computed by SVD on the housekeeping genes. P-values were computed using Limma.

three different factor analysis methods. The first is the singular value decomposition (SVD). The second is an EM algorithm based on a relatively simple probabilistic model that allows for gene-specific variances in the error term. The model and the algorithm are described in Section 12.2.4 of Bishop (2006). The implementation of the algorithm is our own. The third method is a "robust" method described in Hubert, Rousseeuw, and Vanden Branden (2005) and implemented in the `PcaHubert` fuction of the R package `rrcov`. RLE plots after adjustment looked nearly identical for all three factor analysis methods. P-value histograms were also remarkably similar. More convincingly, the number of top-ranked X/Y genes is roughly the same for all three methods, as we discuss below.

The choice of k turns out to be critical to the performance of RUV-2. This can be seen in RLE plots and p-value histograms (figures are in the SM), but can be seen most dramatically by counting the number of top-ranked X/Y genes at different values of k. We repeated the analysis for all possible values of k. At first, the number of top-ranked X/Y genes increases with increasing k, as additional unwanted variation is removed. After a certain point, however, increasing k only degrades performance, as adding additional factors to the model simply increases variance.[10] See Figure 3. Note from Figure 3 that although the performance of RUV-2 depends critically on the choice of k, the region over which RUV-2 performs well is fairly large. This is important because it implies, in this example at least, that while RLE plots and p-value histograms may not lead us to the single best choice of k, they can at least lead us to a good one. Finally,

[10]Note that as $k \to m$, we keep adjusting away additional dimensions, until we finally remove everything. At this point the OLS estimator becomes undefined. Thus, performance will certainly degrade for large k. The only question is when.

note also from Figure 3 that the choice of factor analysis method does not greatly impact the results.

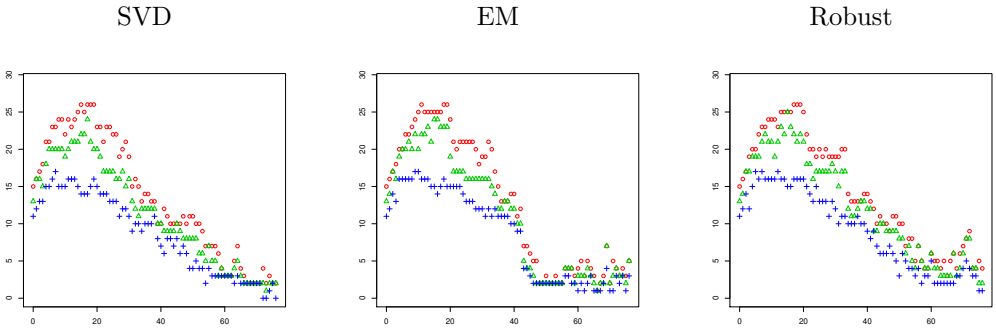


Figure 3: Comparison of performance of different factor analysis methods in the gender study. The number of X / Y genes discovered is plotted as a function of k. Genes were ranked by p-value; results are shown for the number of X / Y genes ranked in the top 20 (plus), top 40 (triangle), and top 60 (circle). $k = 0$ corresponds to no adjustment. Data were preprocessed (BG + NM + LS). PCs were computed using the housekeeping genes. P-values were computed using Limma.

It is common to analyze microarray data using the Limma package in Bioconductor (Smyth, 2004). Limma uses empirical Bayes methods to improve the estimates of the variances of individual genes. Since adjusting with RUV-2 or any other method can substantially affect the general structure of the residuals, we checked to ensure that RUV-2 and Limma work well together. To accomplish this we completed our analysis once using Limma and once using "ordinary" regression. We did not notice any substantial difference in performance (as assessed using RLE plots, p-value histograms, and counts of top-ranked X/Y genes) between the two methods. Details can be found in the SM.

The performance of RUV-2 depends greatly on the choice of negative controls. Adjustments based on both the housekeeping genes and the Affymetrix spike-in controls improved performance relative to unadjusted data, however the performance increase using housekeeping genes was substantially better. See Figure 4. We believe housekeeping genes may outperform the spike-ins in this example for three reasons. Firstly, housekeeping genes may be able to capture unwanted biological variation, whereas spike-in controls can only capture unwanted technical variation. Secondly, even with regards to technical variation, housekeeping genes may be more "representative" than spike-ins. For example, housekeeping genes may capture unwanted variation introduced during sample collection, whereas spike-ins would not. Conversely, spike-ins may exhibit some unique variation related to their own administration. Lastly, there are far more housekeeping genes than there are spike-in controls. There are 799 probesets that correspond to one of the housekeeping genes in Eisenberg and Levanon (2003). However, there are only 33 probesets corresponding to spike-in controls.

Finally, we wish to compare the performance of RUV-2 to that of other adjustment methods. Additionally we would like to investigate the effects of preprocessing. We use the number of top-ranked X/Y genes as our basis for comparison,[11] and present the results in Table 1. Several observations merit mention. RUV-2 outperforms Combat and ordinary regression in all cases, and SVA outperforms them in several cases as well. This is despite the fact that Combat and ordinary regression explicitly model known batches (lab / chip type), whereas RUV-2 and SVA infer the unwanted variation from the data. Moreover, it seems that even when we do explicitly model known batches with RUV-2 by including a "Z" term in the model (see Section 4) there is no substantial increase in performance.[12] Another observation is that the level of preprocessing does not seem to matter for the SVD and EM (but not robust) variants of RUV-2, matters slightly for Combat,

[11] Additionally, RLE plots and p-value histograms for Combat and SVA are given in the SM.

[12] It is true that more X / Y genes are found when we include a Z term in the "Top 40" and "Top 60" cases, but this should be interpreted with caution; including a Z term results in adjusting for three additional factors; similar increases in performance can be seen by dropping the Z term and increasing k to 13 — see Figure 3.

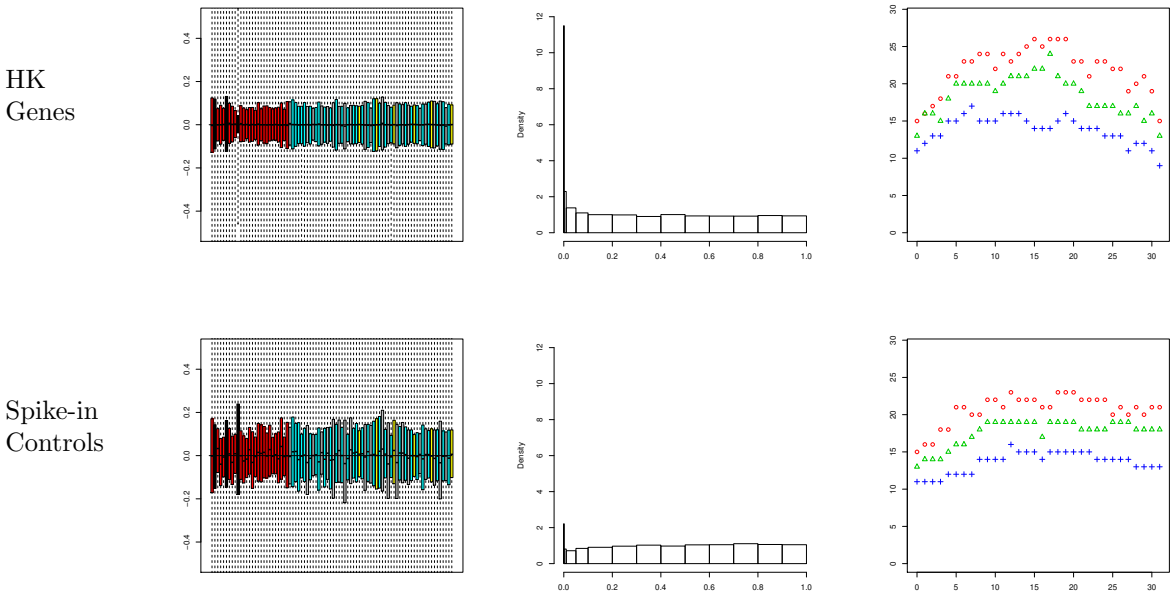


Figure 4: Comparison of results for housekeeping genes and Affymentrix spike-in controls in the gender study. For the RLE plots and p-value histograms, $k = 10$. Factors were computed by SVD. P-values were computed using Limma. Note that there are only 33 spike-in controls, so adjustments with $k > 33$ are undefined for the spike-in case. We truncate results in the housekeeping case as well for easy comparison.

and matters greatly for the other methods. This seems to suggest that, at least in some cases, RUV-2 can obviate the preprocessing. While this is of little immediate value in the current example, it may be useful in situations where there is concern that the nonlinearities introduced by preprocessing are problematic. For example, if a very large number of genes are differentially expressed with respect to the factor of interest, the nonlinearities in the quantile normalization may induce an artificial correlation between the negative control genes and the factor of interest. This would violate the assumptions of RUV-2, and create problems. In such a situation it may be best to skip the quantile normalization. In any case, we regard the fact that RUV-2 performs well even without preprocessing as quite encouraging; after all, the un-preprocessed data are extremely noisy. Finally, we observe that in general, the best performing methods are the SVD and EM variants of RUV-2.

3.2 Alzheimer's Study

Blalock et al. (2004) conducted a microarray study to investigate patterns of gene expression in Alzheimer's patients. We learned of this dataset from a colleague soon after we completed our initial analysis of the gender study, and decided to see if we could use it to replicate our findings. We continued to use gender as the factor of interest (instead of Alzheimer's disease state) so that we had clear positive controls. Note that our choice of this dataset was thus rather arbitrary — many other brain studies would have been just as suitable — but this one had our attention, was publicly available, and seemed to exhibit a fair degree of unwanted variation. Data are available on GEO (GSE1297).

Samples were taken post-mortem from the hippocampus of 35 individuals suffering from various stages of Alzheimer's disease. Four samples were discarded by the authors of the study because the severity of the patients' disease was not clear, leaving 31 samples available for analysis. These samples were assayed using Affymetrix HG-U133A microarrays. We were unable to obtain clinical gender data for the samples,

	Top 20						Top 40						Top 60					
	No Prepr.		BG,QN		BG,QN,LS		No Prepr.		BG,QN		BG,QN,LS		No Prepr.		BG,QN		BG,QN,LS	
	Std	Lim	Std	Lim	Std	Lim	Std	Lim	Std	Lim	Std	Lim	Std	Lim	Std	Lim	Std	Lim
No Adjustment	7	6	9	9	11	11	7	7	12	12	13	13	7	7	14	13	15	15
Regression (Z)	5	5	14	13	12	12	6	6	15	15	16	15	7	7	16	16	16	17
SVA (IRW)	5	6	12	11	14	14	6	7	13	14	17	16	7	8	15	14	19	19
SVA (Two-Step)	NA	NA	NA	NA	16	16	NA	NA	NA	NA	21	20	NA	NA	NA	NA	23	22
Combat	11	11	13	13	12	13	14	12	17	17	17	17	16	15	18	18	19	19
RUV-2 — SVD ($k = 10$)	15	15	15	15	15	15	22	22	21	19	20	19	24	23	22	21	22	22
RUV-2 — SVD ($k = 10$), w/Z	14	14	15	15	16	15	**23**	**23**	22	**22**	**22**	**22**	**25**	**25**	24	24	24	**25**
RUV-2 — EM ($k = 10$)	**17**	**17**	**17**	**17**	**17**	**17**	22	22	**23**	**22**	**22**	20	24	24	**25**	**25**	25	24
RUV-2 — EM ($k = 10$), w/Z	16	16	16	16	**17**	16	22	22	22	**22**	**22**	**22**	24	23	**25**	24	**26**	**25**
RUV-2 — Robust ($k = 10$)	8	7	**17**	16	**17**	16	11	10	21	21	**22**	21	12	11	**25**	**25**	24	24

Table 1: Summary of performance of different methods in the gender study. The number of X / Y genes found in the top 20 / 40 / 60 is shown for different levels of preprocessing (None; BG,QN; BG,QN,LS), different methods of computing p-values (standard and Limma), and different methods of adjustment (none, standard regression, SVA, Combat, and RUV-2). In the case of RUV-2, results are given for different methods of factor analysis (SVD, EM, robust). Additionally, we present results for models that include an explicit Z term, where Z is a matrix of dummy variables corresponding to site (A, B, or C) and chip type (HGU-95A or HGU-95Av2). For all RUV-2 methods, $k = 10$ and housekeeping genes were used as negative controls. Results for the two-step variant of SVA are not available in some cases, because the function exited with an error. The highest number in each column is shown in bold.

but found we could infer the gender using the expression levels of XIST and DDX3Y.[13]

Our analysis — and conclusions — closely parallel those of the gender study. Most of the details are provided in the SM; here we simply highlight key results. One important difference between this study and the gender study is that in this study there are no known batch effects. Therefore standard regression and Combat are inapplicable. Moreover, a final location / scale pre-processing step is unnecessary, since there is only one chip type. A summary of the performance of SVA and RUV-2 is provided in Table 2.

	Top 20				Top 40				Top 60			
	No Prepr.		BG,QN		No Prepr.		BG,QN		No Prepr.		BG,QN	
	Std	Lim	Std	Lim	Std	Lim	Std	Lim	Std	Lim	Std	Lim
No Adjustment	8	9	15	16	9	9	19	21	9	9	23	24
SVA (IRW)	8	9	16	16	10	13	17	19	12	13	18	19
SVA (Two-Step)	NA	NA	16	16	NA	NA	18	18	NA	NA	19	19
RUV-2 — SVD ($k = 10$)	**16**	**19**	**19**	**20**	**22**	**23**	**25**	**26**	**23**	**26**	26	26
RUV-2 — EM ($k = 10$)	15	17	**19**	**20**	18	22	24	25	**23**	25	25	**27**
RUV-2 — Robust ($k = 10$)	11	13	**19**	**20**	15	17	**25**	**26**	19	20	**27**	**27**

Table 2: Summary of performance of different methods in the Alzheimer's study. The number of X / Y genes found in the top 20 / 40 / 60 is shown for different levels of preprocessing (None, BG,QN), different methods of computing p-values (standard and Limma), and different methods of adjustment (none, SVA, and RUV-2). In the case of RUV-2, results are given for different methods of factor analysis (SVD, EM, robust). For all RUV-2 methods, $k = 10$ and housekeeping genes were used as negative controls. Results for the two-step variant of SVA are not available in some cases, because the function exited with an error. The highest number in each column is shown in bold.

In contrast with the gender study, we find that preprocessing improves performance in all cases. In addition, Limma provides a clear performance enhancement as well, particularly when preprocessing is omitted. This is as we might expect, since Limma is most helpful when the sample size is small, and here we have only 31 samples compared to the gender study's 84. Again, the best performance is attained with RUV-2. Note that in some cases, SVA actually decreases performance.

[13]Details are provided in the SM. Note in particular that we found 19 women and 12 men, and this contradicts the statement in Blalock et al. (2004) that, of the original 35 subjects, 16 were female and 19 were male.

Finally, we note that, unlike in the gender study example, RUV-2 does not perform well with spike-in controls. A comparison between housekeeping genes and spike-in controls of RLE plots, p-value histograms, and the number of top-ranked X / Y genes can be found in Figure 5. Note that on the HG-U133A platform there are 45 probesets for the spike-in controls, and 1112 corresponding to the housekeeping genes in Eisenberg and Levanon (2003). We cannot say with confidence why the spike-in controls fail in this case, but we note that two possible explanations would be that most of the unwanted variation is biological in nature, or due to sample quality.

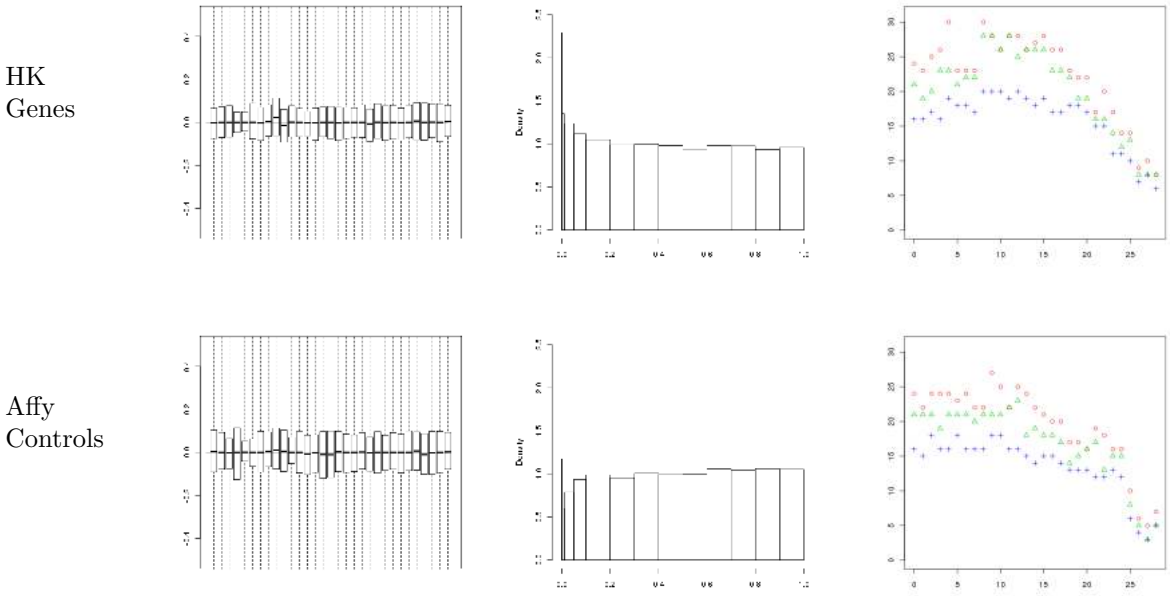


Figure 5: Comparison of results for housekeeping genes and Affymentrix spike-in controls in the Alzheimer's study. HK genes clearly perform better. For the RLE plots and p-value histograms, $k = 10$. Factors were computed by SVD. P-values were computed using Limma.

3.3 TCGA

The Cancer Genome Atlas (TCGA) is a large collaborative project established by the National Institutes of Health and managed by the National Cancer Institute and the National Human Genome Research Institute with the aim of collecting many types of data (e.g. expression, sequence, methylation) on a large number of samples from many different types of cancers. In particular, TCGA has already collected a few hundred glioblastoma multiforme (brain tumor) samples, and measured gene expression levels in these samples using (among others) the Affymetrix GeneChip Human Exon 1.0 ST array and the Affymetrix HT HG-U133A array. Because of the importance and size of this dataset, we decided it would be a good choice with which we could further replicate our findings from the gender study.

As with the Alzheimer's study, our analysis and conclusions closely parallel those of the gender study, and we limit our remarks here to highlight key results. Details can be found in the SM. We discuss the exon array data first, followed by the HT HG-U133A data.

We downloaded exon array data for 386 samples from TCGA. Clinical gender data were available for 316 of the samples. We processed the data using aroma.affymetrix (Bengtsson et al., 2008) using a custom

CDF[14] provided by colleagues at Lawrence Livermore National Laboratory. The custom CDF has 18632 probesets, corresponding to genes (not exons). The custom CDF does not include any probesets for spike-in controls, so we were only able to study adjustments using housekeeping genes. We identified 518 probesets as housekeeping genes.

Substantial unwanted variation is evident. The RLE plot of the un-preprocessed data (Figure 29 in the SM) suggests the presence of batch effects as well as additional within-batch variation. RLE plots and p-value histograms using various values of k (Figures 30 and 31 in the SM) do not suggest a clear choice for k. The RLE plots suggest that k should be at least 30. The p-value histograms suggest that the unwanted variation skews p-values downwards, but a k of 100 is sufficient to solve this problem. Since a wide variety of values of k may be appropriate, we include results for both $k = 50$ and $k = 100$ in Table 3. Indeed, it turns out that results do not vary much over this wide range in k (see Figure 34, SM).

	Top 20				Top 40				Top 60			
	No Prepr.		BG,QN		No Prepr.		BG,QN		No Prepr.		BG,QN	
	Std	Lim	Std	Lim	Std	Lim	Std	Lim	Std	Lim	Std	Lim
No Adjustment	20	20	20	20	25	25	30	30	28	28	33	33
SVA (IRW)	20	20	20	20	28	29	30	30	30	30	32	32
SVA (Two-Step)	NA	NA	20	20	NA	NA	34	34	NA	NA	36	36
RUV2 — SVD ($k = 50$)	20	20	20	20	34	34	35	35	36	36	36	36
RUV2 — EM ($k = 50$)	20	20	20	20	35	35	**36**	**36**	37	37	**37**	**37**
RUV2 — Robust ($k = 50$)	20	20	20	20	33	33	35	35	37	36	36	36
RUV2 — SVD ($k = 100$)	20	20	20	20	**36**	**36**	35	35	**39**	**39**	**37**	**37**
RUV2 — EM ($k = 100$)	20	20	20	20	**36**	**36**	**36**	**36**	**39**	**39**	**37**	**37**
RUV2 — Robust ($k = 100$)	20	20	20	20	**36**	**36**	**36**	**36**	37	37	**37**	**37**

Table 3: Summary of performance of different methods for TCGA exon array data. The number of X / Y genes found in the top 20 / 40 / 60 is shown for different levels of preprocessing (None, BG,QN), different methods of computing p-values (standard and Limma), and different methods of adjustment (none, SVA, and RUV-2). In the case of RUV-2, results are given for different methods of factor analysis (SVD, EM, robust), and different values of k (50 and 100). Housekeeping genes were used as negative controls. Results for the two-step variant of SVA are not available in some cases, because the function exited with an error. The highest number in each column is shown in bold.

One of the more striking features of Table 3 is the relatively small gain achieved by adjustment. Only an additional 5 or so genes are discovered, despite the substantial unwanted variation. This may be because the large sample size compensates for the unwanted variation, allowing most of the differentially expressed genes to be found even without adjustment. Nonetheless, adjustment does help. RUV-2 and the two-step variant of SVA perform the best. A final interesting observation is that in this example, unlike in the gender study and Alzheimer's study examples, the robust variant of RUV-2 performs well even on un-preprocessed data.

We now turn to the HT HG-U133A data. We downloaded data for 414 samples from TCGA. Clinical gender data were available for 319 of these. Unlike with the exon array data, spike-in control data are available, but the HT HG-U133A only has 9 spike-in probesets. We identified 1045 probesets as housekeeping genes.

As with the exon array data, substantial unwanted variation is evident. Indeed, the RLE plot of the un-preprocessed data (Figure 38 in the SM) suggests the presence of very substantial batch effects. RLE plots and p-value histograms (Figures 39 and 40 in the SM) suggest using a k of 30 is appropriate. Table 4 summarizes the results. Once again, RUV-2 and the two-step variant of SVA perform the best. Unlike with the exon array data, the robust variant of RUV-2 does not perform well on un-preprocessed data.

A comparison of the performance of spike-in controls and housekeeping genes is provided in Figure 6. As one might expect, considering there are only 9 spike-in controls, the housekeeping genes perform better.

[14] A CDF is a file that maps probes to probesets. A custom CDF is necessary because Affymetrix does not provide a CDF with their exon arrays. They provide a different file format, which is incompatible with aroma.affymetrix.

	Top 20				Top 40				Top 60			
	No Prepr.		BG,QN		No Prepr.		BG,QN		No Prepr.		BG,QN	
	Std	Lim	Std	Lim	Std	Lim	Std	Lim	Std	Lim	Std	Lim
No Adjustment	14	14	20	20	21	22	35	35	27	27	39	39
SVA (IRW)	**20**	**20**	20	20	36	36	37	37	38	38	44	44
SVA (Two-Step)	NA	NA	20	20	NA	NA	**39**	**39**	NA	NA	**53**	**53**
RUV-2 — SVD ($k = 30$)	**20**	**20**	20	20	**38**	**38**	38	38	**47**	**47**	50	50
RUV-2 — EM ($k = 30$)	**20**	**20**	20	20	**38**	**38**	**39**	**39**	46	46	51	50
RUV-2 — Robust ($k = 30$)	**20**	**20**	20	20	31	30	**39**	**39**	37	35	51	51

Table 4: Summary of performance of different methods for the TCGA HT HG-U133A data. The number of X / Y genes found in the top 20 / 40 / 60 is shown for different levels of preprocessing (None, BG,QN), different methods of computing p-values (standard and Limma), and different methods of adjustment (none, SVA, and RUV-2). In the case of RUV-2, results are given for different methods of factor analysis (SVD, EM, robust). For all RUV-2 methods, $k = 30$ and housekeeping genes were used as negative controls. Results for the two-step variant of SVA are not available in some cases, because the function exited with an error. The highest number in each column is shown in bold.

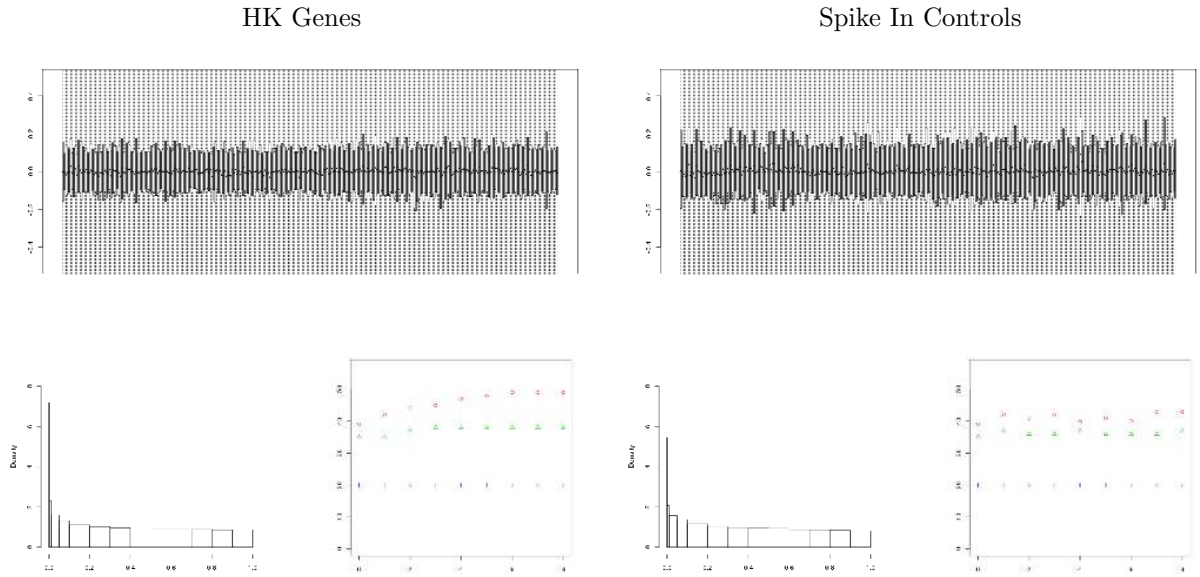


Figure 6: Comparison of results for housekeeping genes and Affymentrix spike-in controls in the TCGA HT HG-U133A data. Housekeeping genes clearly perform better. For the RLE plots and p-value histograms, $k = 30$. Factors were computed by SVD. P-values were computed using Limma. Note that there are only 9 spike-in controls, so adjustments with $k > 9$ are undefined for the spike-in case. We truncate results in the housekeeping case as well for easy comparison.

3.4 NCI-60

As discussed in Section 1, RUV-2 is an *application specific* adjustment method intended for use in differential expression studies. A natural question is whether the method can be adapted to provide a global adjustment, in which the entire dataset is adjusted to provide a modified dataset that can then be used just like the original dataset in any subsequent analysis. The advantages of such a method are obvious — not only would it allow one to use the method for applications other than differential expression (e.g. for classification or clustering), but the self-contained nature of the method would allow one to easily insert the algorithm into pre-existing code, and thus easily re-visit past analyses. Unfortunately, adapting RUV-2 to provide a global adjustment is not trivial. In this section we consider a naive adaptation of RUV-2 that provides a global adjustment and demonstrate that there may be some promise in the approach, but that there are many potential pitfalls.

Our naive method is simple — perform factor analysis on the control genes as before, regress the original dataset onto the factors, and use the residuals of the regression as the new dataset. In other words, subtract off (from the entire dataset) the components of variation that are seen to exist in the control genes. This method is "naive" because it implicitly assumes that all of the factors corresponding to the biology of interest are orthogonal to the unwanted factors. For applications such as clustering, in which the goal is to discover the biology of interest from the data, the assumption that the biology of interest is orthogonal to the unwanted variation is at best unverifiable. If the assumption is false, the problem of throwing out the baby with the bathwater returns, and the "adjustment" may very well hurt more than it helps.

Our example is the NCI-60 dataset. The National Cancer Institute maintains 60 cell lines derived from the tissues of 9 different types of human cancers (brain, blood, breast, colon, kidney, lung, ovary, prostate, skin). These cell lines have been well studied, and a great deal of public data are available. More information (including data) can be found at `http://discover.nci.nih.gov/`. We obtained expression data from a study that analyzed the NCI-60 cell lines using using both the Affymetrix HG-U95A and HG-U133A platforms (Shankavaram et al., 2007). We wanted to see whether the naive adjustment described above would result in a better clustering of the data into the 9 tissue types. To perform the clustering, we used the R functions `dist` and `hclust` with their default settings ("euclidian" and "complete linkage").

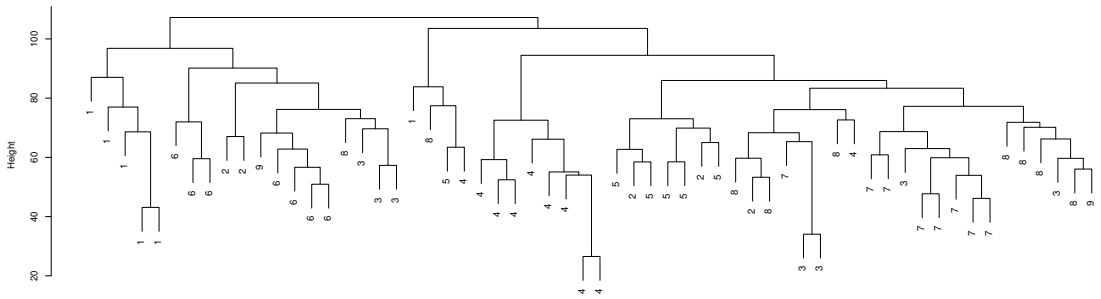
Figure 7 provides the clustering results for the HG-U95A data before and after an adjustment using the spike-in controls and $k = 1$. The adjustment helps. Before adjustment, the colorectal cancers were grouped into one clade of size 4 and another clade of size 3. After adjustment, they were all grouped into a single clade. This clade also included one lung cancer, however. Other clusterings improved as well. Before adjustment, 5 of the Leukemias grouped into a single clade, and 1 grouped by itself. After adjustment, all 6 Leukemias grouped into a single clade. Ovarian went from four clades of sizes 3, 2, 1 and 1 to three clades of size 4, 2, and 1. Renal went from 3 clades of size 5, 2 and 1 to two clades of size 7 and 1. Lung went from nine clades of size 1 to 6 clades — one of size 4 and five of size 1. The prostate cancers moved closer together, but still did not form their own clade. There was no substantial change in the quality of the clusterings of the other cancers (brain, breast, skin).

Despite the success in Figure 7, the naive global adjustment cannot be relied upon in general. Increasing k from 1 to 2 does not further improve the performance (some cancers cluster slightly better while others cluster slightly worse), and increasing k past 2 quickly leads to a *decrease* in the quality of the clustering. See Figures 46, 47, and 48 in the SM for dendrograms at various values of k. In this particular example, we only knew to set $k = 1$ because we knew the "correct answer." RLE plots provide poor guidance in choosing k, and p-value plots / positive controls are not even applicable since this is not a differential expression study. In a more realistic example, choosing an appropriate k would be very difficult.

Moreover, the method does not work with the HG-U133A data for any value of k. This may be because the unwanted variation is correlated with the biology of interest. We discuss this possibility in more detail in Section B in the SM. It is also possible that the spike-in controls are simply too noisy, or that the variation characteristic of the spike-in controls is not sufficiently representative of the unwanted variation affecting the majority of the probesets.[15] Indeed, recall from previous examples that spike-in controls generally did not

[15] A final logical possibility is that there is simply no unwanted variation to remove. In that case, we would see no improvement after adjustment. However, we do not believe this is the case, as we have evidence there is unwanted variation (although not

Unadjusted



$k = 1$, Affy Spike-in Controls

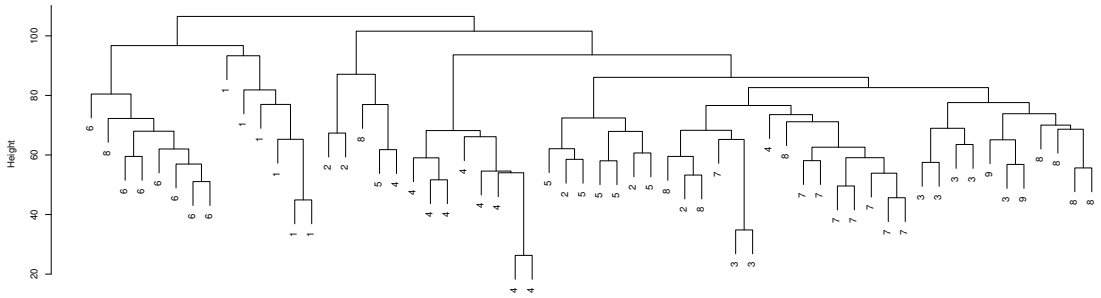


Figure 7: Dendrograms of NCI-60 HG-U95A dataset before and after adjustment. The data were preprocessed (BG + QN). The factors were computed by SVD. Key: 1 – Blood; 2 – Breast; 3 – Ovarian; 4 – Melanoma; 5 – Brain; 6 – Colorectal; 7 – Renal; 8 – Lung; 9 – Prostate.

perform as well as housekeeping genes.

Lastly, the method does not work if we use housekeeping genes instead of the spike-in controls. The housekeeping genes are highly correlated with cancer type. (Again, see Section B for a more complete discussion of this topic.) This is presumably due to the fact that the tissue types are sufficiently different that even the expression levels of the housekeeping genes actually do vary from one tissue type to the next, violating of the control gene assumption. See Figure 8 for a dendrogram of the HG-U133A data after an adjustment using housekeeping genes.

necessarily batch effects). Indeed, as we discuss in Section B in the SM, it seems there is unwanted variation that is correlated with biology.

$k = 1$, Housekeeping Genes

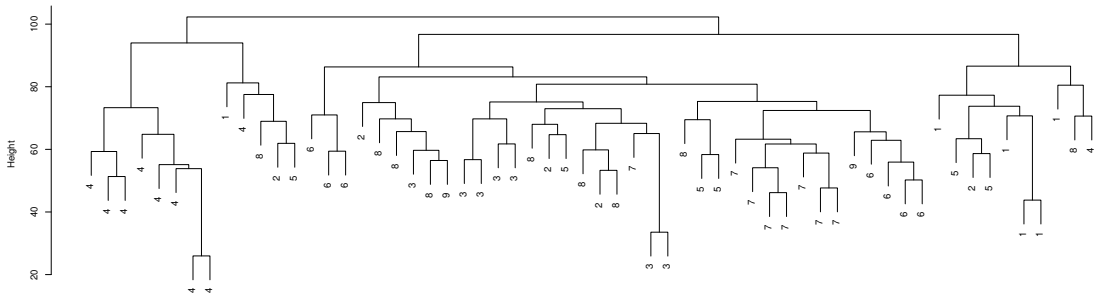


Figure 8: Dendrograms of NCI-60 HG-U95A dataset after adjustment using housekeeping genes. The data were preprocessed (BG + QN). The factors were computed by SVD. Key: 1 – Blood; 2 – Breast; 3 – Ovarian; 4 – Melanoma; 5 – Brain; 6 – Colorectal; 7 – Renal; 8 – Lung; 9 – Prostate.

4 Methods

Assume we have m arrays each with n genes (or probes or probesets). Let y_{ij} denote the observed log expression level of the j^{th} gene on the i^{th} array, and let Y denote the $m \times n$ matrix (y_{ij}). We model Y as:

$$Y_{m\times n} = X_{m\times p}\beta_{p\times n} + Z_{m\times q}\gamma_{q\times n} + W_{m\times k}\alpha_{k\times n} + \epsilon_{m\times n} \tag{1}$$

Here, X is a matrix whose columns are the factors of interest (e.g. disease state, treatment / control), Z is a matrix whose columns are observed covariates (e.g. batch, ethnicity), and W is a matrix whose columns are unobserved covariates (e.g. sample quality). Note that k is unobserved. The matrices β, γ, and α are all unobserved coefficients that determine the influence of a particular factor on a particular gene. The $Z\gamma$ term is optional; a researcher may not be aware of any observed covariates, or may wish to treat observed covariates as if they were unobserved. To complete our specification of the model, we make the following assumptions:

$$\text{Rank}\left[(X \mid Z \mid W)\right] = p + q + k < m \tag{2}$$

$$\mathbb{E}\left[\epsilon \mid X, Z, W\right] = 0 \tag{3}$$

$$\text{Var}\left[\epsilon_{ij} \mid X, Z, W\right] = \sigma_j^2 \tag{4}$$

$$\epsilon_{ij} \perp\!\!\!\perp \epsilon_{i'j'} \text{ if } (i,j) \neq (i',j') \tag{5}$$

In a differential expression study the goal is to estimate β. This is almost a standard linear regression problem. The important difference is that W is unobserved. To address this difficulty, we propose to estimate W from the data using negative control genes.[16] As stated previously, negative control genes are those genes whose expression levels are known *a priori* not to be "truly associated" with X. More formally, the j^{th} gene is a negative control if we can assume that $\beta_{\star j} = 0$, where $\beta_{\star j}$ denotes the j^{th} column of β.

Let Y_c, α_c, β_c, γ_c and ϵ_c be reduced versions of Y, α, β, γ and ϵ that only contain the columns of control genes. Then by (1)

$$Y_c = X\beta_c + Z\gamma_c + W\alpha_c + \epsilon_c. \tag{6}$$

The "control gene assumption" however is that $\beta_c = 0$, so (6) becomes

$$Y_c = Z\gamma_c + W\alpha_c + \epsilon_c. \tag{7}$$

For simplicity, first consider the case that there is no $Z\gamma$ term in the model. Then (7) becomes

$$Y_c = W\alpha_c + \epsilon_c. \tag{8}$$

This is a typical model in factor analysis, and many methods exist to estimate W (e.g. SVD). Note that, no matter what factor analysis method is used, W can only be estimated if $\text{rank}(\alpha_c) \geq k$. In practice, this means that the control genes must not only be unassociated with the factor of interest, but must indeed be associated with the unwanted factors — the negative controls must exhibit the unwanted variation so that the factor analysis can detect it!

If there is a $Z\gamma$ term in the model, we can multiply both sides of (7) by the residual operator

$$R_Z \equiv I - Z\left(Z'Z\right)^{-1}Z'$$

to obtain

$$R_Z Y_c = R_Z W\alpha_c + \epsilon_c. \tag{9}$$

We can then use factor analysis to estimate $R_Z W$. Now, in general we cannot assume that $R_Z W = W$, but in practice we can safely treat our estimate of $R_Z W$ as if it were in fact an estimate of W. This is because the standard OLS estimator for β depends only on the column space of $(Z|W)$ — not on Z and W themselves — and the column space of $(Z|W)$ is equal to the column space of $(Z|R_Z W)$. Thus we see

[16] Recall that we use the term "negative control genes" loosely, and allow it to refer to spike-in controls as well.

that whether or not a $Z\gamma$ term is included in the model, a de-facto estimate $\hat{W}$ of W can be easily produced using factor analysis. We can then calculate $\hat{\beta}$ by OLS, substituting $\hat{W}$ for W. Note that in a trivial sense we include a Z term in all of our analyses, since we include a constant (intercept) term in our models.

We now consider the adaptation of this method introduced in Section 3.4. Now the goal is not an estimate $\hat{\beta}$ of β, but rather an adjusted expression matrix $Y^\star$. Additionally, we now regard X as unknown. Regarding X as unknown makes any effort to adapt RUV-2 highly non-trivial. For simplicity, consider the case in which there is no $Z\gamma$ term; extending to the case in which there is a $Z\gamma$ term is trivial. An ideal globally adjusted expression matrix would simply equal the original expression matrix Y minus the unwanted variation $W\alpha$; i.e. $Y^\star$ would equal $X\beta + \epsilon$. In the method of Section 3.4, we fall short of this ideal. We define:

$$Y^\star \equiv \left[I - \hat{W} \left(\hat{W}'\hat{W} \right)^{-1} \hat{W}' \right] Y \tag{10}$$

where $\hat{W}$ is found by factor analysis as in RUV-2. In words, we project onto the orthogonal complement of the column space of $\hat{W}$. This should effectively remove the unwanted variation. The problem is that it may remove some of the signal as well. Note that

$$Y^\star = R_{\hat{W}} \left(X\beta + W\alpha + \epsilon \right) \tag{11}$$

where $R_{\hat{W}}$ is the projector onto the orthogonal complement of the column space of $\hat{W}$. Thus if $\hat{W} \approx W$,

$$Y^\star \approx R_{\hat{W}} X\beta + R_{\hat{W}} \epsilon. \tag{12}$$

If $\hat{W}$ is orthogonal to X, then $R_{\hat{W}} X = X$ and $Y^\star$ is nearly our ideal globally adjusted expression matrix. If the unwanted factors are primarily technical factors (not biological), and samples are processed randomly, X and $\hat{W}$ may very well be nearly orthogonal. On the other hand, if $\hat{W}$ and X are not orthogonal, $Y^\star$ may be far from the ideal, and if X is regarded as unknown, it is not even possible to check whether $\hat{W}$ and X are orthogonal — let alone correct for the bias non-orthogonality would introduce. Thus, this naive adaptation of RUV-2 should be used with extreme caution, if at all.

5 Discussion

Methodologically, RUV-2 is extremely simple. Its two steps — factor analysis and regression — are well studied and well understood. Despite this simplicity, RUV-2 is highly effective. RUV-2 derives its strength not from any deep new statistical theory but from some powerful biological assumptions. RUV-2 is only as good as these assumptions on which it is based, and it is therefore worthwhile to reiterate these assumptions. The control genes must satisfy two key conditions — they must be 1) uninfluenced by the factor of interest, and they must be 2) influenced by the unwanted factors. Different situations will call for different sets of control genes. The choice of an appropriate set of control genes is central to RUV-2.

We have discussed two possible sets of control genes — housekeeping genes and spike-in controls. In the differential expression examples (gender, Alzheimer's, and TCGA) the housekeeping genes are the better choice, presumably because they are more "representative," and because there are more of them. In the NCI-60 example the spike-in controls are the better choice because the housekeeping genes are not negative controls with respect to tissue type. The NCI-60 example highlights the important fact that housekeeping genes are influenced by biology and cannot be casually assumed to be negative controls in every situation. Housekeeping genes are effective negative controls in the first several examples because they are unaffected by *gender*, not because they are unaffected by biology in general. In short, housekeeping genes are a good place to begin a search for negative controls, but cannot be relied upon in all cases — the factor of interest matters.

We need not restrict our search for control genes to housekeeping genes and spike-in controls. In other studies, still other sets of genes might be the best choice. For example, a researcher might wish to use genes known to be stably expressed within a particular tissue type (Stamova et al., 2009), or under certain

experimental conditions; choosing genes specially suited to the study at hand may improve performance. In other situations, a researcher might wish to include additional control genes with the intent of adjusting for specific types of unwanted variation. For example, if a researcher suspects that the cause of death is an important source of unwanted variation, it might be wise to include control genes that could possibly capture this information — e.g. genes associated with cellular stress, apoptosis, etc.

There may be a temptation to "discover" negative control genes. For example, a researcher may wish to find genes whose expression levels are not highly correlated with the factor of interest, label these genes as negative controls, and then adjust via RUV-2. The allure of this approach is clear — finding a set of negative controls would be much easier, and could in fact be automated. However, we feel this approach is misguided. If there are unwanted factors that are correlated with the factor of interest, then the expression levels of the true negative controls should in fact be correlated with the factor of interest. Excluding genes correlated with the factor of interest would bias our estimate of the unwanted factors.

Just as the researcher must exercise judgment when choosing a set of control genes, the researcher must also exercise judgment when choosing k. This can be difficult. We have seen that RLE plots and p-value histograms can be quite helpful. Positive controls, when available,[17] can be even more helpful. Some readers may question why we encourage choosing k based on these quality assessments when more "objective" and automated methods exist. For example, it is possible to choose k via a series of hypothesis tests, in which one keeps increasing k until no more "statistically significant" factors can be found. This is the approach taken in SVA. However, we feel there are problems with this approach. One reason is that including an additional term in a linear regression model may lead to a decrease in bias, but it can also lead to an increase in variance. Thus, it is possible that we might get a better estimate of β by leaving some of the unwanted factors out of the model. Using hypothesis testing to find k does not account for this bias-variance trade-off; instead the goal is simply to include all factors. A second reason is a bit more philosophical. Whenever a researcher calculates a small p-value, three logical possibilities are on the table — the null hypothesis is true and we have observed an unlikely event; the null hypothesis is false; the model is wrong. We know already the model is wrong. We don't know "how wrong," or in precisely which way. Nor do we know exactly how the model misspecification affects the results of any particular hypothesis test. Thus, we feel it is unwise to rely too heavily on a hypothesis test to give us a "good" answer, especially when the choice of k is so important.[18] We feel there is a role here for human judgment, and that quality assessments based on positive controls, p-value histograms, etc., are useful tools in guiding this judgment.

The simplicity of RUV-2 makes it relatively flexible, and an excellent starting point for new, more advanced methods. In addition, some of the basic ideas of RUV-2 can be useful in exploratory data analysis. For example, we have used methods similar to RUV-2 to identify the age of a formalin fixed, paraffin-embedded tissue sample as an important source of unwanted variation (data not shown). The extended NCI-60 discussion in the SM is another good example. In these final paragraphs, we discuss some ways in which RUV-2 might be improved, and suggest possibilities for future development.

One possible direction for improvement is in the regression step. While we considered three different methods of factor analysis (and found that a simple SVD seems to work as well as anything else), with the

[17] Some readers may question the availability of positive controls. Positive controls are genes known to be associated with the factor of interest. The goal of differential expression studies is to find genes associated with the factor of interest. Thus it may seem that if we have positive controls, we already have "the answer." To be sure, positive controls will not always be available, but the situation is not hopeless. In some cases, we might have a set of genes known to be highly enriched with differentially expressed genes, although we don't know exactly which of the genes are the differentially expressed ones. In this case, it might be possible to treat the entire set of genes as if they were all positive controls. For example, the method of ranking genes and counting the number of top-ranked positive controls can still be used. This is exactly the approach taken in our gender examples. Not all X and Y-linked genes are differentially expressed, but many of them are. Alternatively, we may begin with only a handful of known positive controls, when in fact many genes are differentially expressed. An example would have been if we had found any differentially expressed autosomal genes in our gender examples. (In actuality, we did not find any autosomal genes that appeared to be consistently differentially expressed with respect to gender across multiple datasets.)

[18] Our objections here are not "purely philosophical." We experimented with various "objective" methods for determining k, and found that many worked very well on simulated data, but very poorly on real data. Presumably, this was due to model misspecification. Some methods worked well on real data as well — but of course, we could only verify this on the relatively small number of datasets that we had available. Other datasets may suffer other forms of model misspecification, and the methods may no longer work.

exception of Limma, we did not consider any elaborations to the regression step of RUV-2. There may be room for improvement. Combat, for example, is essentially an advanced form of regression, and, as we saw in the gender example, it does in fact perform better than standard regression. Incorporating some of the techniques from Combat into RUV-2 may lead to a still better method.

A second avenue for future development concerns combining multiple datasets. It is an open question whether RUV-2 is effective enough to allow a researcher to combine multiple datasets from completely separate studies without introducing excessive unwanted variation. It is also an open question as to whether RUV-2 can be used when combining data from different platforms. We saw in the gender study example that we were able to effectively combine data from the HG-U95A and HG-U95Av2 platforms, but these platforms are relatively similar. It is not clear that such a procedure would also work when combining data from, for example, the HG-U95A and HG-U133A platforms, or when combining Affymetrix and Agilent chips.

Yet another open question is whether RUV-2 can be adapted to entirely different technologies. While microarrays are still used, high throughput sequencing technologies are becoming increasingly popular for use in gene expression studies. Adapting RUV-2 for use with these sequencing technologies could be very helpful. Other technologies, such as qrt-PCR, may benefit from an adaptation of RUV-2 as well.

Finally, we feel it may be possible to apply some of the ideas of RUV-2 to applications other than differential expression. As we have stated, we feel that unwanted variation is most effectively dealt with when it is considered in the context of the goal of the analysis at hand. RUV-2 deals effectively with unwanted variation in the context of differential expression studies. However, microarrays are also commonly used for classification and for clustering. We do not yet know how best to handle unwanted variation in these types of studies, but we believe control genes will play an important role.

6 Acknowledgements

We would like to thank Darya Chudova, Jun Li, Mark Vawter, Hui Shen, Matthew Ritchie, Pierre Neuvial, Juergen von Frese, Sue Wilson, Di Wu, Laurent Jacob, Tim Triche, Dongseok Choi and Prabhakara Choudary for their helpful comments on a draft version of this paper. We would also like to thank Francois Collin, Moshe Olshansky, Elizabeth Purdom, Pierre Neuvial, Sandrine Dudoit, Jun Li, Mark Vawter, Gordon Smyth, Mark Robinson, Keith Satterley, Leming Shi, Roel Verhaak, Victoria Wang, Ying Xu and Julia Brettschneider for helpful discussions and logistical support in the course of our research. Finally, we would like to thank Philip Musk and Gene Logic for providing us with their data.

References

O. Alter, P.O. Brown, and D. Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences of the United States of America*, 97(18):10101–10106, 2000.

H. Bengtsson, K. Simpson, J. Bullard, and K. Hansen. aroma.affymetrix: A generic framework in R for analyzing small to very large affymetrix data sets in bounded memory. Technical Report #745, Department of Statistics, University of California, Berkeley, February 2008.

C.M. Bishop. *Pattern recognition and machine learning*. Springer, 2006.

E.M. Blalock, J.W. Geddes, K.C. Chen, N.M. Porter, W.R. Markesbery, and P.W. Landfield. Incipient Alzheimer's disease: microarray correlation analyses reveal major transcriptional and tumor suppressor responses. *Proceedings of the National Academy of Sciences of the United States of America*, 101(7):2173, 2004.

B. Bolstad, F. Collin, J. Brettschneider, K. Simpson, L. Cope, R. Irizarry, and T.P. Speed. Quality assessment of Affymetrix GeneChip data. *Bioinformatics and computational biology solutions using R and bioconductor*, pages 33–47, 2005.

B.M. Bolstad, R.A. Irizarry, M. Astrand, and T.P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185, 2003. ISSN 1367-4803.

J. Brettschneider, F. Collin, B.M. Bolstad, and T.P. Speed. Quality assessment for short oligonucleotide microarray data. *Technometrics*, 50(3):241–264, 2008. ISSN 0040-1706.

E. Eisenberg and E.Y. Levanon. Human housekeeping genes are compact. *TRENDS in Genetics*, 19(7): 362–365, 2003. ISSN 0168-9525.

M. Hubert, P.J. Rousseeuw, and K. Vanden Branden. ROBPCA: a new approach to robust principal component analysis. *Technometrics*, 47(1):64–79, 2005. ISSN 0040-1706.

R.A. Irizarry, B.M. Bolstad, F. Collin, L.M. Cope, B. Hobbs, and T.P. Speed. Summaries of Affymetrix GeneChip probe level data. *Nucleic acids research*, 31(4):e15, 2003a. ISSN 0305-1048.

R.A. Irizarry, B. Hobbs, F. Collin, Y.D. Beazer-Barclay, K.J. Antonellis, U. Scherf, and T.P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249, 2003b. ISSN 1465-4644.

WE Johnson, C. Li, and A. Rabinovic. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1):118–127, 2007.

H.M. Kang, C. Ye, and E. Eskin. Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. *Genetics*, 180(4):1909–1925, 2008a.

H.M. Kang, N.A. Zaitlen, C.M. Wade, A. Kirby, D. Heckerman, M.J. Daly, and E. Eskin. Efficient control of population structure in model organism association mapping. *Genetics*, 178(3):1709, 2008b.

H.M. Kang, J.H. Sul, S.K. Service, N.A. Zaitlen, S.Y. Kong, N.B. Freimer, C. Sabatti, and E. Eskin. Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, 42(4):348–354, 2010. ISSN 1061-4036.

J.T. Leek and J.D. Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*, 3:e161, 2007.

J.T. Leek and J.D. Storey. A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences*, 105(48):18718–18723, 2008. ISSN 0027-8424.

J.T. Leek, R.B. Scharpf, H.C. Bravo, D. Simcha, B. Langmead, W.E. Johnson, D. Geman, K. Baggerly, and R.A. Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11:733–739, 2010.

K.A. Lippa, D.L. Duewer, M.L. Salit, L. Game, and H.C. Causton. Exploring the use of internal and external controls for assessing microarray technical performance. *BMC Research Notes*, 3(349), 2010. ISSN 1756-0500.

J. Listgarten, C. Kadie, E.E. Schadt, and D. Heckerman. Correction for hidden confounders in the genetic analysis of gene expression. *Proceedings of the National Academy of Sciences*, 107(38):16465, 2010. ISSN 0027-8424.

B.H. Mecham, P.S. Nelson, and J.D. Storey. Supervised normalization of microarrays. *Bioinformatics*, 26 (10):1308–1315, 2010. ISSN 1367-4803.

T.O. Nielsen, R.B. West, S.C. Linn, O. Alter, M.A. Knowling, J.X. O'Connell, S. Zhu, M. Fero, G. Sherlock, J.R. Pollack, P.O. Brown, D. Botstein, and M. van de Rijn. Molecular characterisation of soft tissue tumours: a gene expression study. *The Lancet*, 359(9314):1301–1307, 2002. ISSN 0140-6736.

A.L. Price, N.J. Patterson, R.M. Plenge, M.E. Weinblatt, N.A. Shadick, and D. Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909, 2006.

A. Scherer. *Batch Effects and Noise in Microarray Experiments: Sources and Solutions*. Wiley, 2009. ISBN 0470741384.

U.T. Shankavaram, W.C. Reinhold, S. Nishizuka, S. Major, D. Morita, K.K. Chary, M.A. Reimers, U. Scherf, A. Kahn, D. Dolginow, J. Cossman, Kaldjian E.P., D.A. Scudiero, E. Petricoin, L. Liotta, J.K. Lee, and J.N. Weinstein. Transcript and protein expression profiles of the NCI-60 cancer cell panel: an integromic microarray study. *Molecular Cancer Therapeutics*, 6(3):820–832, 2007. ISSN 1535-7163.

L. Shi, L.H. Reid, W.D. Jones, R. Shippy, J.A. Warrington, S.C. Baker, P.J. Collins, F. de Longueville, E.S. Kawasaki, K.Y. Lee, et al. The MicroArray Quality Control (MAQC) project shows inter-and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology*, 24(9):1151–1161, 2006. ISSN 1087-0156.

G.K. Smyth. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, 3(1), 2004.

B.S. Stamova, M. Apperson, W.L. Walker, Y. Tian, H. Xu, P. Adamczy, X. Zhan, D.Z. Liu, B.P. Ander, I.H. Liao, J.P. Gregg, R.J. Turner, G. Jickling, L. Lit, and F.R. Sharp. Identification and validation of suitable endogenous reference genes for gene expression studies in human peripheral blood. *BMC Medical Genomics*, 2(49), 2009. ISSN 1755-8794.

O. Stegle, A. Kannan, R. Durbin, and J. Winn. Accounting for non-genetic factors improves the power of eQTL studies. In *Proceedings of the 12th annual international conference on Research in computational molecular biology*, pages 411–422. Springer-Verlag, 2008. ISBN 3540788387.

M.P. Vawter, S. Evans, P. Choudary, H. Tomita, J. Meador-Woodruff, M. Molnar, J. Li, J.F. Lopez, R. Myers, D. Cox, S.J. Watson, H. Akil, E.G. Jones, and W.E. Bunney. Gender-specific gene expression in post-mortem human brain: localization to sex chromosomes. *Neuropsychopharmacology*, 29(2):373–384, 2004.

W.N. Venables and B.D. Ripley. *Modern applied statistics with S*. Springer-Verlag, 2002. ISBN 0387954570.

J. Yu, G. Pressoir, W.H. Briggs, I.V. Bi, M. Yamasaki, J.F. Doebley, M.D. McMullen, B.S. Gaut, D.M. Nielsen, J.B. Holland, S. Kresovich, and E.S. Buckler. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, 38(2):203–208, 2005.

Supplementary Material to **Using Control Genes to Correct for Unwanted Variation in Microarray Data**

Johann A. Gagnon-Bartsch* Terence P. Speed*†

A MAQC Example

The Microarray Quality Control (MAQC) project (Shi et al., 2006) sought to investigate the replicability of microarray results by processing multiple technical replicates of a few biological samples at multiple laboratories on multiple microarray platforms. In particular, they processed 5 technical replicates of Stratagene Universal Human Reference RNA ("Sample A") and 5 technical replicates of Ambion Human Brain Reference RNA ("Sample B") at each of 6 laboratories (for a total of 60 chips) using Affymetrix HG-U133 Plus 2 arrays. To investigate unwanted variation in this dataset, we first subtracted off sample means gene by gene (to remove the biological signal), and then computed the SVD of the resulting matrix. Substantial variation between laboratories is evident, but so is substantial within-laboratory variation.

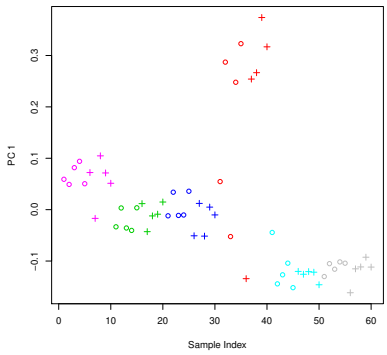


Figure 9: The first PC of the MAQC HG-U133 Plus 2 data after removal of the biological signal. Different colors represent different laboratories. Circles represent Sample A; pluses represent Sample B. Unwanted variation occurs both between and within batches. The data was not preprocessed.

We contacted MAQC to enquire if there was any known explanation for the relatively large (though still minor in absolute terms) unwanted variation at the fourth laboratory (the E.P.A., shown in red). We were informed that at this laboratory, the person who ran the experiment did not get enough training and hands-on experience.

*Department of Statistics, University of California at Berkeley, Berkeley, CA 94720
†Bioinformatics Division, Walter and Eliza Hall Institute, Victoria 3050, Australia

B Extended NCI-60 Discussion

In our discussion of the NCI-60 data in Section 3.4 we raise the possibility that, in the case of the HG-U133A data, unwanted variation is correlated with biology. As discussed in Section 4, the "naive" global adjustment method can fail when the unwanted variation is correlated with biology. Thus, this correlation may explain why the adjustment fails in the case of the HG-U133A data. We also asserted that the (true, wanted) biological variation in the housekeeping genes is highly correlated with biology, and therefore the housekeeping genes cannot be regarded as negative controls in this example. In this Section, we present the methods and results by which we arrived at these (tentative) conclusions.

In Section 4 we observe that the "naive" global adjustment method will fail if X and $\hat{W}$ are substantially correlated. Recall that this is because the method will only work when $R_{\hat{W}}X = X$, and $R_{\hat{W}}X = X$ only when X and $\hat{W}$ are orthogonal. Now, since $\hat{W}$ is calculated from data, X and $\hat{W}$ will never be perfectly orthogonal in practice. We would therefore like some measure of just how correlated X and $\hat{W}$ actually are. We choose to use canonical correlations as a measure. Since canonical correlations may be unfamiliar to some readers, we present here a brief review, beginning with some motivation. More information can be found in Venables and Ripley (2002).

Consider two full-rank matrices $U_{n\times p}$ and $V_{n\times q}$. Assume that each column of U and each column of V has a mean of 0. Note that U and V have the same number of rows. If we choose one column of U and one column of V it is possible to compute the correlation coefficient between these two columns. We could do this for all pq pairs of columns, and this would, in some sense, tell us about the correlation between U and V. This is called the "cross correlation" of U and V. For our purposes, however, the cross correlation presents some difficulties. Firstly, it consists of pq individual numbers, and it is not clear how to interpret all of these numbers simultaneously to get a clear sense of the extent to which U and V are correlated. Perhaps more importantly, if we reparameterize either U or V, so that its individual columns are changed but its column space is not, the cross correlation of U and V will, in general, change. This makes the cross correlation undesriable as a metric for the "correlation" of X and $\hat{W}$, since the quantity $R_{\hat{W}}X$ is unaffected by reparameterizations of $\hat{W}$; we would like our metric of "correlation" between X and $\hat{W}$ to be likewise unaffected by reparameterizations of $\hat{W}$. Canonical correlations achieve this.

Let u be an element of the column space of U, and let v be an element of the column space of V. Let $r(u, v)$ be the correlation of u and v. We can define the first canonical correlation as

$$r_1 \equiv \max_{u,v} r(u, v).$$

Note that this number is unaffected by reparameterizations of U and V. It also has the advantage of being a single number. We consider it to be a good metric by which to measure the extent to which X and $\hat{W}$ are correlated, and the extent to which $R_{\hat{W}}X \neq X$.

(Note that we could also define the second canonical correlation — and the third, fourth, etc., up to $\min(p, q)$ — but these are not needed for our purposes. For sake of completeness, however, we remark that the second canonical correlation can be defined similarly to how we defined the first canonical correlation, but constraining the maximization over u and v to include only u and v that are orthogonal to u_1 and v_1, where $r_1 = r(u_1, v_1)$.)

It can be shown that the first canonical correlation is equal to the square root of the largest eigenvalue of

$$(V'V)^{-\frac{1}{2}}V'U(U'U)^{-1}U'V(V'V)^{-\frac{1}{2}}. \tag{13}$$

Note that in the case that $q = 1$ (i.e. the case that V is a single column),

$$(V'V)^{-\frac{1}{2}}V'U(U'U)^{-1}U'V(V'V)^{-\frac{1}{2}} = \frac{V'U(U'U)^{-1}U'V}{V'V} \quad (14)$$

$$= \frac{V'U(U'U)^{-1}U'U(U'U)^{-1}U'V}{V'V} \quad (15)$$

$$= \frac{\left[U(U'U)^{-1}U'V\right]'\left[U(U'U)^{-1}U'V\right]}{V'V} \quad (16)$$

$$= \frac{\hat{V}'\hat{V}}{V'V} \quad (17)$$

$$= R^2 \quad (18)$$

where $\hat{V} \equiv U(U'U)^{-1}U'V$ and R^2 is the coefficient of determination in a regression of V on U. Thus, in the case that $q = 1$, the first canonical correlation between U and V is equal to the familiar quantity $\sqrt{R^2}$ in a regression of V on U. This completes our review of canonical correlations.

We now return to the NCI-60 example. The biology of interest in our example is the cancer tissue type (blood, brain, etc.). Since tissue type is known (despite the fact we treat it as unknown in the example), we can construct X as a 9-column matrix of dummy variables. We can then calculate the first canonical correlation between X and $\hat{W}$.

If we are interested in whether or not the variation characteristic of the control genes is "systematically" correlated with biology, we might want to know whether the observed canonical correlation is "statistically significant." Strictly speaking, this question does not make sense unless the samples were processed at random. Nonetheless, we can generate an interesting "null distribution" by randomly permuting tissue type labels (i.e. randomly permuting the rows of X) and recalculating the canonical correlation each time. We can then calculate the fraction of these re-calculated canonical correlations that are greater than the observed canonical correlation to get an approximate "p-value." If this p-value is small, we might conclude that the unwanted variation is systematically correlated with biology (tissue type). Alternatively, we might conclude that the "control genes" are not control genes after all, and their expression levels are in fact influenced by the tissue type.

We calculated the canonical correlation between X and $\hat{W}$ and an associated p-value for both the HG-U95A and HG-U133A datasets, both with preprocessing and without preprocessing, for both the spike-in controls and the housekeeping genes, with a k of 1 and a k of 5. We present the results in Table 5.

		Spike-in Controls				Housekeeping Genes			
		No Prepr.		BG,QN		No Prepr.		BG,QN	
		$k=1$	$k=5$	$k=1$	$k=5$	$k=1$	$k=5$	$k=1$	$k=5$
HG-U95A	1st Can. Corr.	0.33	0.55	0.29	0.56	0.37	**0.93**	**0.82**	**0.92**
	p-value	0.6	0.4	0.8	0.4	0.5	**0**	**0**	**0**
HG-U133A	1st Can. Corr.	**0.55**	0.65	0.44	0.56	**0.56**	**0.93**	**0.83**	**0.93**
	p-value	**0.02**	0.06	0.2	0.4	**0.009**	**0**	**0**	**0**

Table 5: First canonical correlation between X and $\hat{W}$ (and an associated p-value) in various cases of the NCI-60 data. Statistically significant entries are shown in bold.

Table 5 is a bit complicated, and difficult to interpret. We begin with the HG-U95A dataset; it is a bit simpler. The spike-in controls are not significantly correlated with the biology in any of the cases shown. This is as we might expect. The housekeeping genes, however, are significantly correlated with biology in some cases. This seems to suggest either that unwanted variation is correlated with biology, or that the housekeeping genes are not actually negative controls. Given that the source of the biological signal is quite strong (completely different tissue types) and that the spike-in data suggest that unwanted variation

is not correlated with biology, we conclude that the housekeeping genes are not actually negative controls. This mostly explains the HG-U95A results, but one puzzle remains — why are the housekeeping genes not significantly correlated with biology in the case of no preprocessing and $k = 1$? Our interpretation of this result is that without preprocessing, the unwanted variation dominates the biological signal, and thus the first PC is unwanted variation, not biology. This is supported by the fact that the correlation of the first PC of the housekeeping genes (without preprocessing) and the first PC of the spike-in controls (again, without preprocessing) is 0.77. (By contrast, after preprocessing, this correlation is -0.01.)

We now turn to the HG-U133A results. Here we see some confirmation that the housekeeping genes are not negative controls. Indeed, it is interesting to note that the canonical correlations in the last three columns of Table 5 are nearly identical between the HG-U95A case and the HG-U133A case. However, the spike-in controls, without preprocessing, are also significantly correlated with biology. This seems to suggest that the unwanted variation in this example is in fact significantly correlated with biology. The puzzle is now why the spike-in controls are *not* significantly correlated with biology after preprocessing. Our interpretation is that the preprocessing — in particular, the quantile normalization — removes much of the component of the unwanted variation that is correlated with biology. As an aside, we note that once again the correlation between the first PC of the housekeeping genes (without preprocessing) and the first PC of the spike-in controls (without preprocessing) is quite strong — 0.88. Thus it seems that the significant correlation between the housekeeping genes and biology in the case of no preprocessing and $k = 1$ is actually driven by unwanted variation, not the biological signal. In other words, it seems that in this particular case, we get the "right answer for the wrong reason."

To summarize, Table 5 suggests that housekeeping genes are not negative controls, that the HG-U133A samples were processed in such a manner as to partially confound tissue type with technical artifacts, and that preprocessing at least partially removes these artifacts.

An Unsolved Mystery

We were interested to learn the source of the unwanted variation that is partially confounded with tissue type in the HG-U133A data. The CEL files indicated that all samples were scanned on a single day. We arranged the samples in order of scan time and made box plots of the log perfect-match probe intensities. No temporal patterns were evident. There were no obvious batch effects / clusters. We therefore contacted Gene Logic, where the samples were assayed. Gene Logic was able to provide us with data on the scanner, hybridization station / position, chip lot, and fluidics station used for each of the samples.[1] There were 12 scanners, 8 hybridization stations, 4 fluidics stations, and 2 chip lots employed in the processing of the arrays. Each of the 8 hybridization stations had 4 positions, for a total of 32 station / position combinations. Only 27 of these combinations were actually used. For each of the four main factors (scanner, hybridization station, fluidics station, chip lot), we created a matrix of dummy variables for the factor. In other words, we created a 12-column matrix of dummy variables for the scanners, a 2-column matrix of dummy variables for the chip lots, etc. We also created a 27-column matrix of dummy variables for each of the hybridization station / position combinations. We then took each of these five matrices and calculated the first canonical correlation between it and a matrix of dummy variables representing tissue types. We also computed a p-value for this canonical correlation using a permutation test. Results are presented in Table 6.

Chip lot stands out; its correlation with tumor type is highly significant. Forty-six chips came from the first chip lot and 11 came from the second. Of the 11 chips from the second chip lot, 8 were used to assay melanoma samples, 2 were used to assay prostate samples, and one was used to assay an ovarian sample.

[1]There was missing / mislabeled / mismatched data for two of the samples. All of the CEL files we downloaded from `http://discover.nci.nih.gov/` were timestamped on April 30, 2002. In the data provided by Gene Logic, all samples were reported as having been processed on April 30, 2002, with the exception of two that were reported to have been processed on May 16. These two samples did not stand out in any obvious way. For example, their log perfect-match box plots seemed normal when compared to the rest of the samples. We re-calculated the canonical correlation between the first PC of the spike-in controls and a dummy matrix representing tumor tissue type, this time omitting the two suspicious samples. The results were essentially unchanged. We therefore determined that it was not these two samples that were driving the correlation between tissue type and the first PC of the spike-in controls. In the discussion / analysis that follows, we omit these two samples.

	Scanner	Hyb. Station	Hyb. Stat. / Pos.	Fluidics Stat.	Chip Lot
Correlation w/ Tissue Type	0.69	0.66	0.86	0.48	0.85
p-value	0.5	0.2	0.7	0.5	0

Table 6: First canonical correlation between known potential confounders (scanner, hybridization station, hybridization station / position combination, fluidics station, chip lot) and tumor tissue type. P-values are approximated using a permuation test (10,000 iterations).

Conversely, only 2 of the melanoma samples were assayed on chips from the first lot, none of the prostate samples, five of the ovarian samples, and all of the other remaining samples.

Despite the very strong confounding of chip lot and tissue type, it does not seem that this confounding explains the correlation between the first PC of the spike-in controls and tissue type. We calculated the canonical correlation between chip lot and the first PC of the spike in controls (un-preprocessed data). It is only 0.03 ($p \approx 0.8$). Chip lot is not a major source of unwanted variation.

Interestingly, it seems that in fact *none* of these factors are major sources of unwanted variation. In Table 7 we present canonical correlations (and p-values) between each of the factors and the first PC of the spike-in controls.

	Scanner	Hyb. Station	Hyb. Stat. / Pos.	Fluidics Stat.	Chip Lot
Correlation w/ first PC	0.47	0.40	0.69	0.33	0.03
p-value	0.3	0.2	0.4	0.1	0.8

Table 7: First canonical correlation between known potential confounders (scanner, hybridization station, hybridization station / position combination, fluidics station, chip lot) and the first PC of the spike-in controls (unpreprocessed data). P-values are approximated using a permuation test (10,000 iterations).

We also regressed the first PC of the spike-in controls onto scanner, hybridization station, fluidics station and chip lot simultaneously. The design matrix included a column of 1s for the intercept, 11 columns for the scanners, 7 columns for the hybridization stations, 3 columns for the fluidics stations, and one for chip lot. The value of R^2 is only 0.44, with a permutation test p-value of 0.31. (Equivalently, the canonical correlation between the design matrix and the first PC is 0.66.) Again, the conclusion seems to be that these four factors are not major sources of unwanted variation.

Finally, we computed the residuals of this regression. The canonical correlation between the residuals and tissue type is 0.5, with a p-value of 0.06. This seems to support the assertion that there is some source of unwanted variation partially confounded with tissue type, but the source of this unwanted variation is associated neither with scanner, hybrid station, fluidics station, nor chip lot. Since this unwanted variation is evident in the spike-in controls, it seems it must be technical variation that enters somewhere in the final stages of the assay. What it is, however, remains a mystery.

C Example Code

RUV-2 is very simple to implement and does not warrant an R pacakge. Instead, we include here some sample code implementing RUV-2 using SVD and Limma. Note that the expression matrix in this example is $n \times m$ instead of $m \times n$, since this is how expression data is often stored in practice. The variable `ctl` must index the genes to be used as control genes.

```
RUV2 = function(Y, X, ctl, k, Z=matrix(rep(1, ncol(Y))))
{
  library(limma)

  # Project onto the orthogonal complement of Z
  RZY = Y - Y%*%Z%*%solve(t(Z)%*%Z)%*%t(Z)

  # Perform SVD
  W = svd(RZY[ctl,])$v

  # Keep the first k factors
  W = W[,1:k]

  # Fit using Limma and return
  fit = lmFit(Y,cbind(X,Z,W))
  fit = eBayes(fit)
  return(fit)
}
```

D Additional Gender Study Figures and Tables

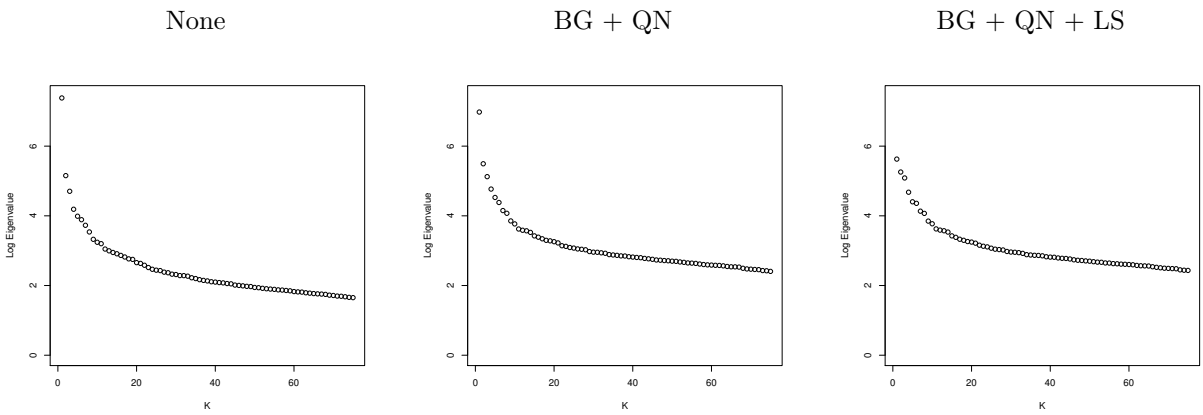


Figure 10: Gender study scree plots on a log scale at different stages of preprocessing. From left to right: No preprocessing; background correction / quantile normalization done separately for each platform type; background correction / quantile normalization followed by a final location/scale adjustment across all chips. All genes were included in the eigenanalysis.

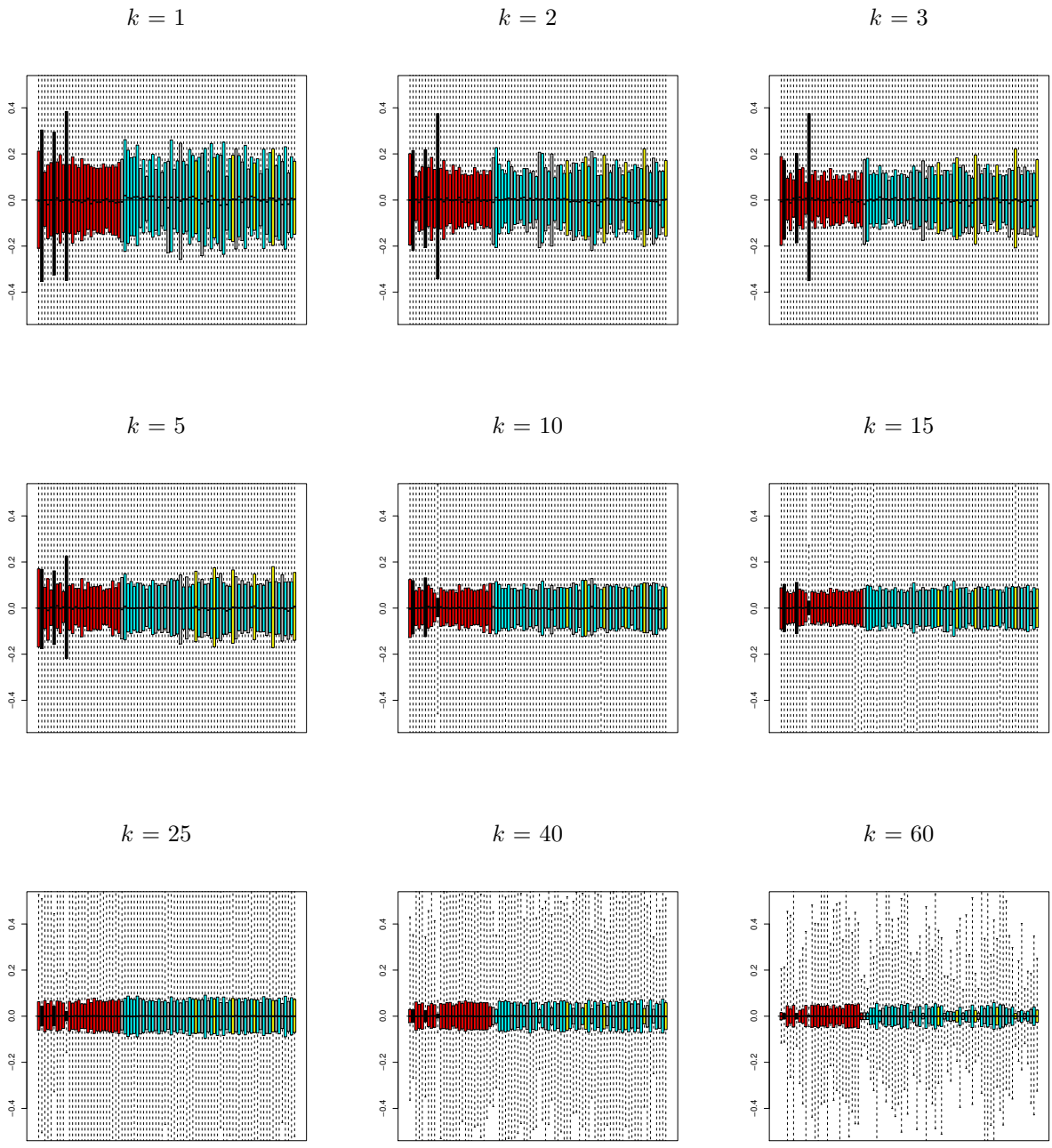


Figure 11: Gender study RLE plots after adjustment, using various values of k. Quality improves with increasing k at first, then decreases as too many factors are removed. Data was fully preprocessed (BG+QN+LS). The factors were computed by SVD on the housekeeping genes.

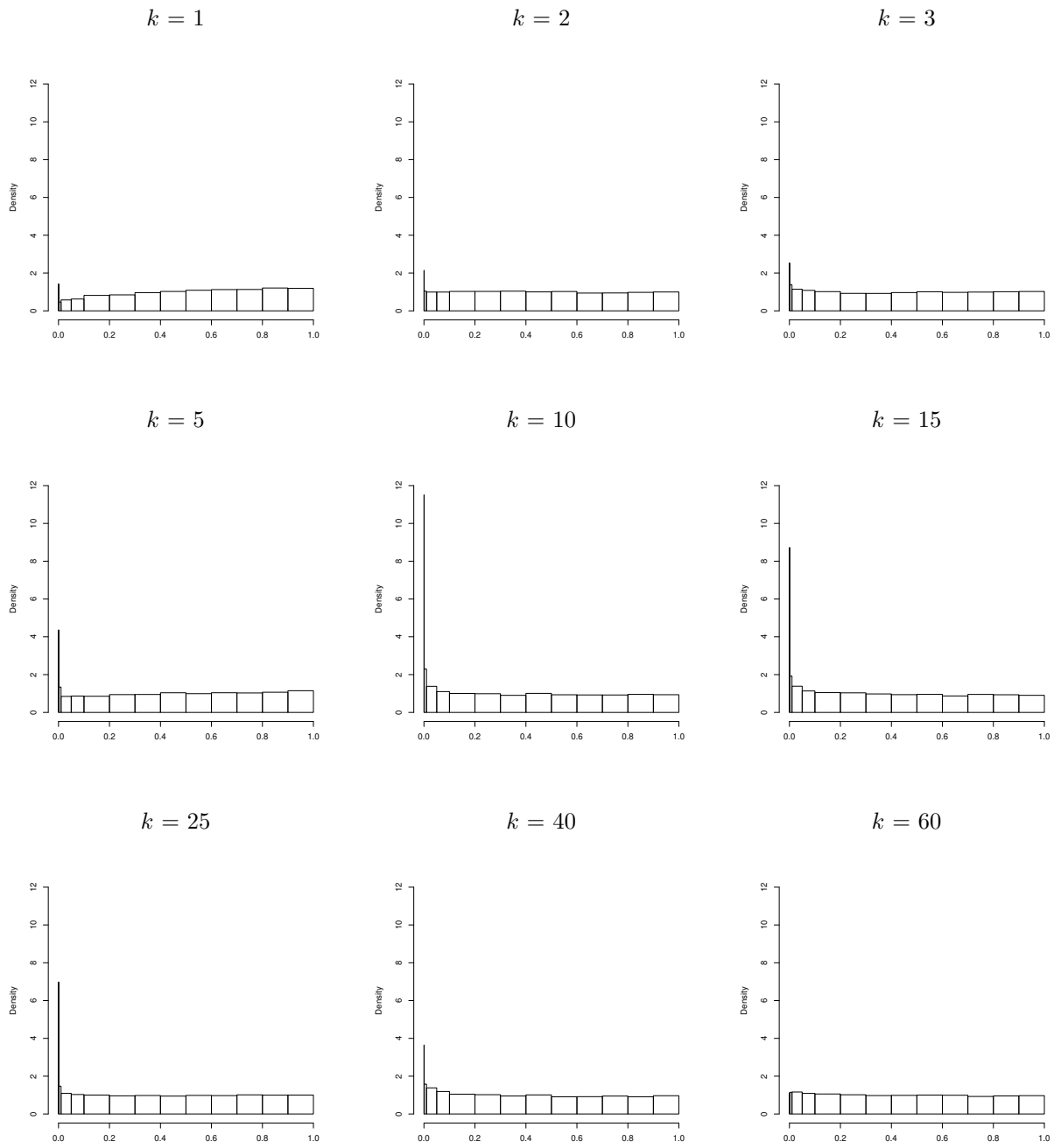


Figure 12: Gender study p-value histograms after adjustment, using various values of k. Histogram breakpoints are at 0.001, 0.01, 0.05, and 0.1, 0.2, 0.3, etc. Data was fully preprocessed (BG+QN+LS). The factors were computed by SVD on the housekeeping genes. P-values were computed using Limma.

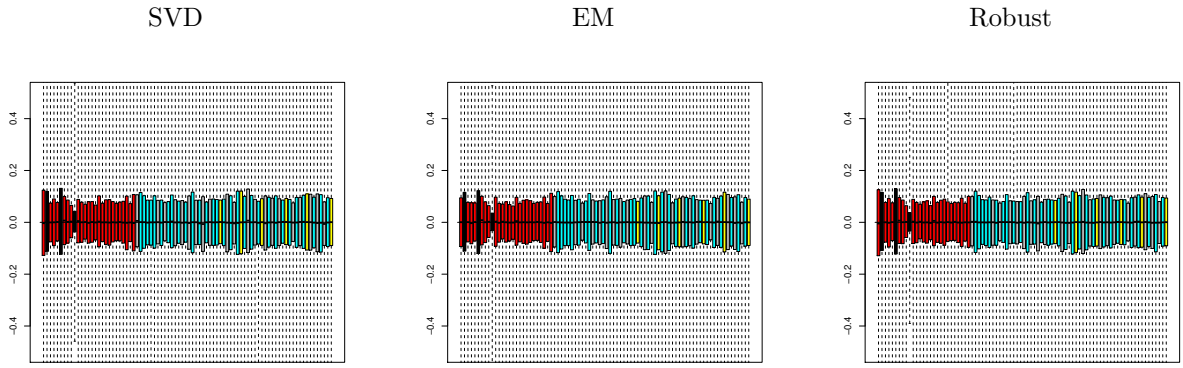


Figure 13: Gender study RLE plots after adjustment ($k = 10$), using different methods of factor analysis. The results are remarkably similar. The data was preprocessed (BG + QN + LS). The factors were computed using the housekeeping genes.

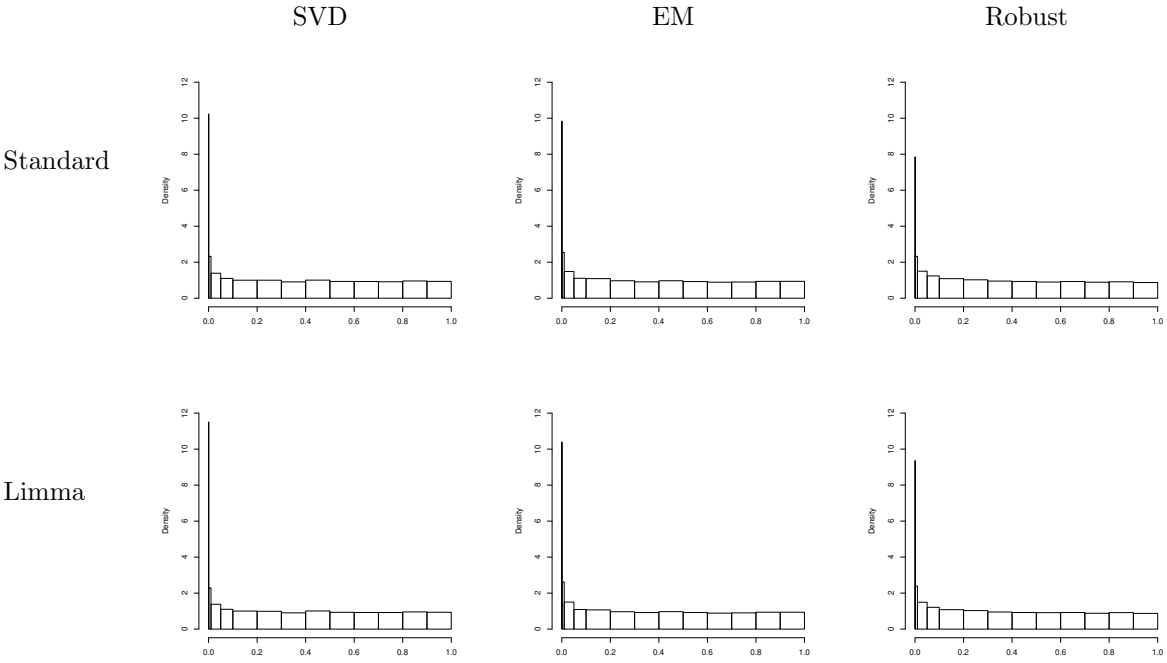


Figure 14: Gender study p-value histograms after adjustment ($k = 10$), using different methods of factor analysis, and different methods of computing p-values. The data was preprocessed (BG + QN + LS). The factors were computed using the housekeeping genes.

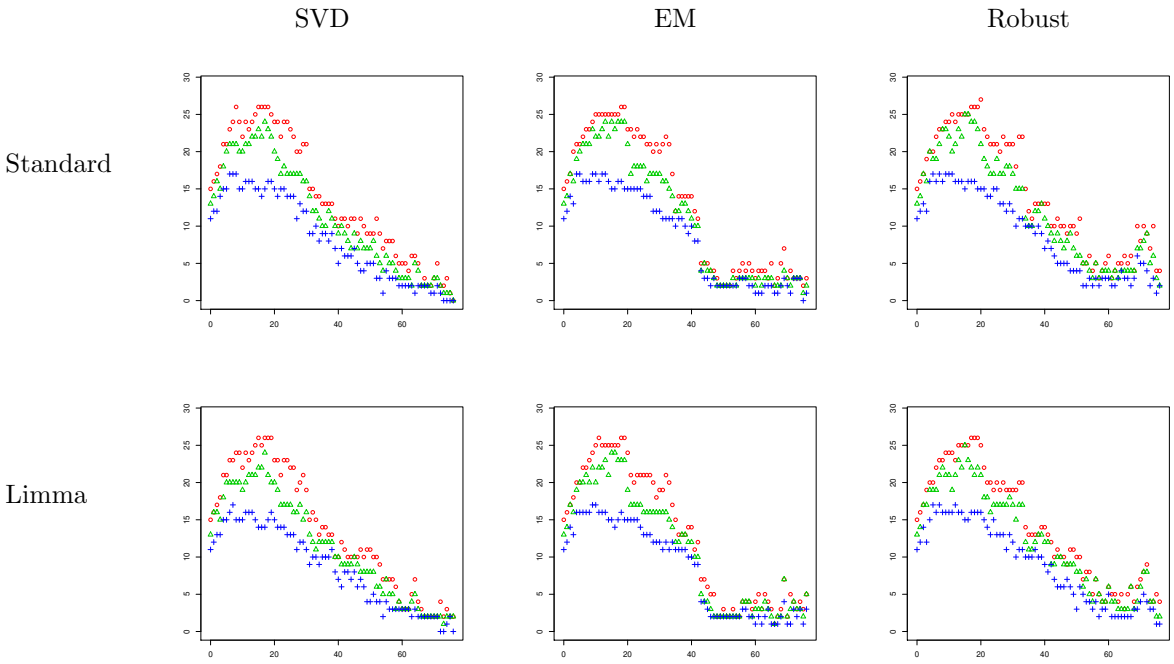


Figure 15: Comparison of the performance of variants of RUV-2 in the gender study. The number of X / Y genes discovered is plotted as a function of k. Genes were ranked by p-value; results are shown for the number of X / Y genes ranked in the top 20 (plus), top 40 (triangle), and top 60 (circle). $k = 0$ corresponds to no adjustment. Data was preprocessed (BG + NM + LS). PCs were computed using the housekeeping genes.

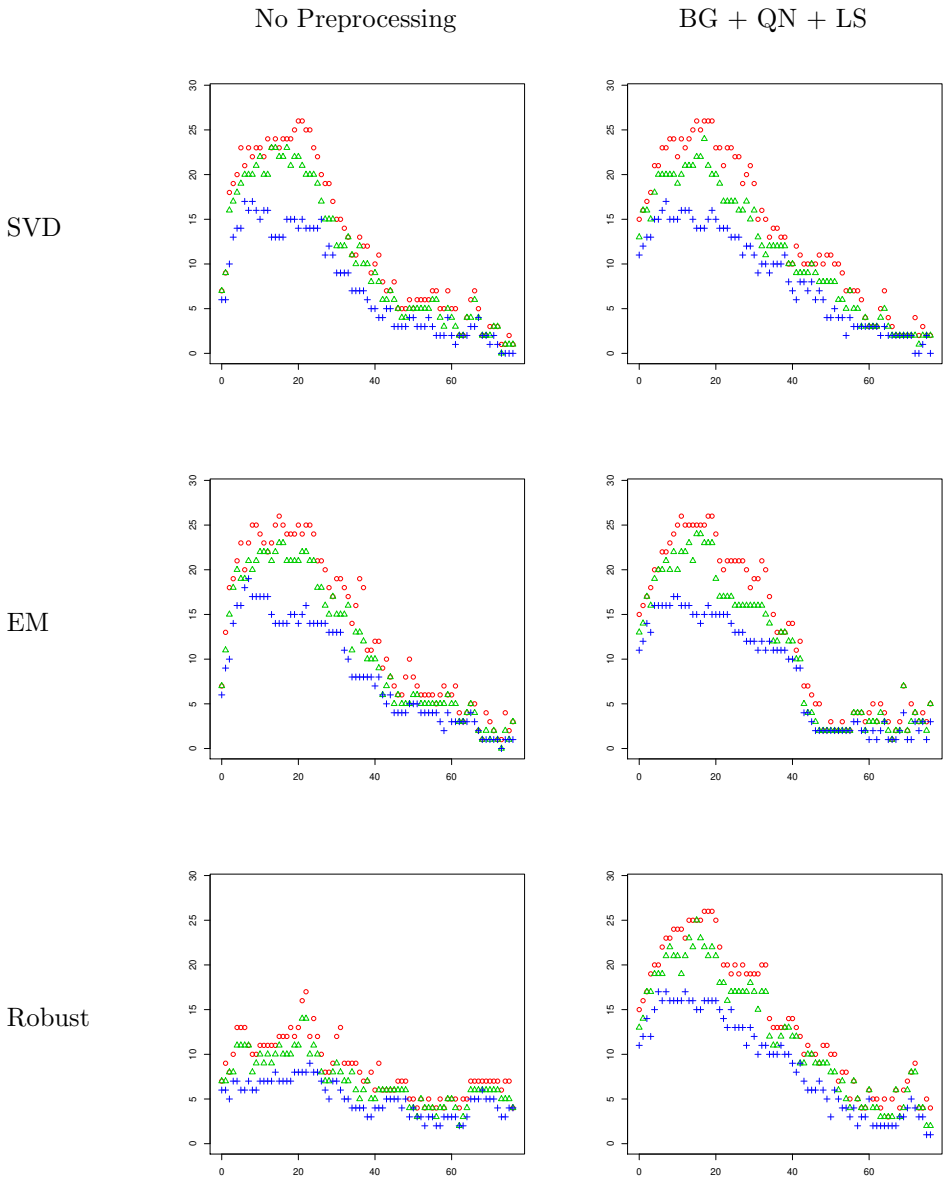


Figure 16: Comparison of the performance of different factor analysis methods with and without preprocessing in the gender study. The number of X / Y genes discovered is plotted as a function of k. Genes were ranked by p-value; results are shown for the number of X / Y genes ranked in the top 20 (plus), top 40 (triangle), and top 60 (circle). $k = 0$ corresponds to no adjustment. PCs were computed using the housekeeping genes. P-values were computed using Limma.

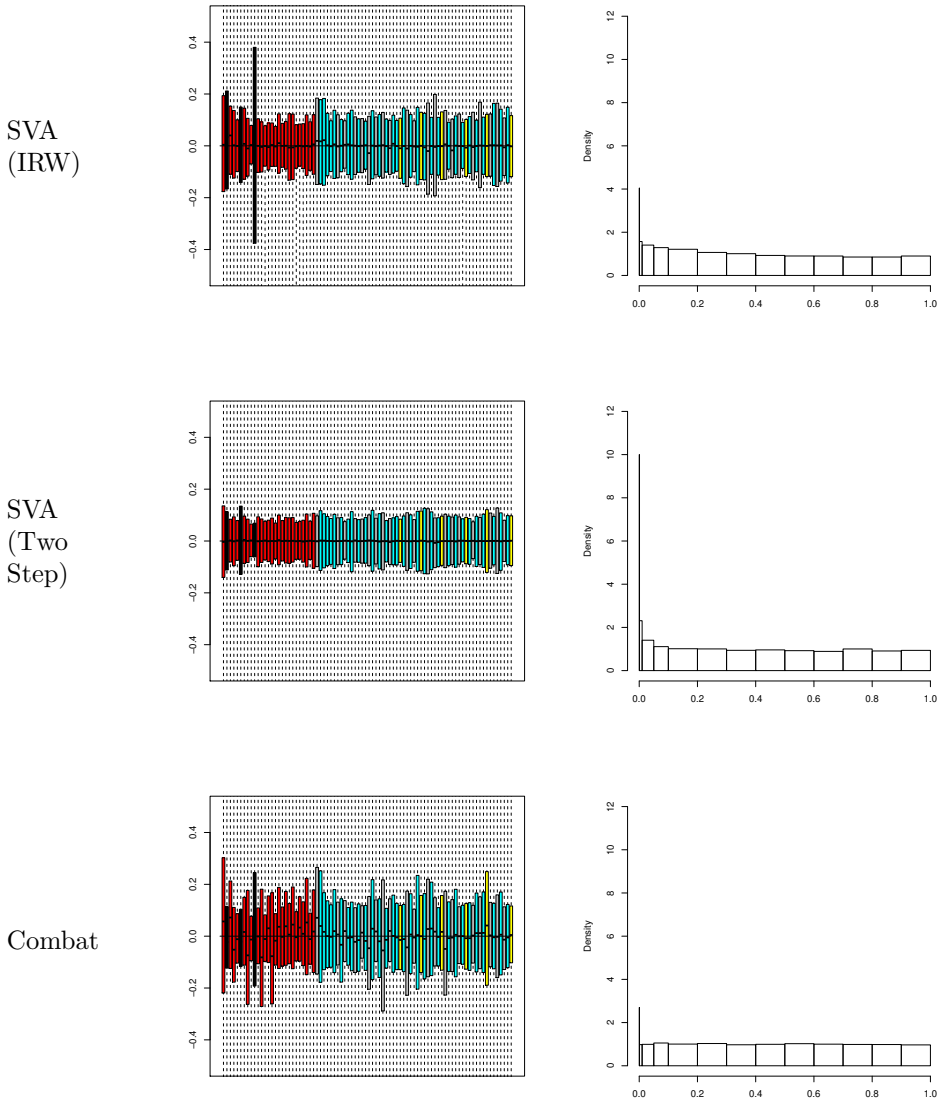


Figure 17: Gender study RLE plots and p-value histograms after adjustments by SVA / Combat. The "two-step" variant of SVA appears to do fairly well. P-values were computed using Limma.

Unadjusted

Rank	Gene	Chrom	Adj. P	Rank	Gene	Chrom	Adj. P	Rank	Gene	Chrom	Adj. P
1	RPS4Y1	Y	6×10^{-39}	11	IL2RA	10	0.4	21	PTPN20B	10	1
2	KDM5D	Y	2×10^{-19}	12	ASMTL	X Y	0.5	22	SLC25A6	X Y	1
3	DDX3Y	Y	3×10^{-19}	13	IFITM1	11	0.7	23	CDC42	1	1
4	XIST	X	2×10^{-13}	14	USP9X	X	0.9	24	COASY	17	1
5	USP9Y	Y	4×10^{-11}	15	CD24	6	1	25	VWF	12	1
6	TTTY15	Y	2×10^{-08}	16	PECAM1	17	1	26	DBC1	9	1
7	UTY	Y	2×10^{-07}	17	HBA1	16	1	27	CD59	11	1
8	EIF1AY	Y	3×10^{-05}	18	PCDH11X	X	1	28	IL10RB	21	1
9	HBB	11	5×10^{-02}	19	HBG1	11	1	29	DNAJB1	19	1
10	CIRBP	19	0.3	20	HBB	11	1	30	ANKRD26	10	1

RUV-2 (SVD), $k = 10$, housekeeping genes

Rank	Gene	Chrom	Adj. P	Rank	Gene	Chrom	Adj. P	Rank	Gene	Chrom	Adj. P
1	RPS4Y1	Y	1×10^{-43}	11	DDX3X	X	5×10^{-06}	21	NOMO1	16	3×10^{-03}
2	KDM5D	Y	2×10^{-26}	12	LITAF	16	4×10^{-05}	22	TF	3	3×10^{-03}
3	DDX3Y	Y	6×10^{-24}	13	PCDH11X	X	6×10^{-05}	23	HBA1	16	4×10^{-03}
4	XIST	X	1×10^{-19}	14	HBG1	11	3×10^{-04}	24	GTPBP6	X Y	4×10^{-03}
5	USP9Y	Y	2×10^{-16}	15	USP9X	X	6×10^{-04}	25	ENPP2	8	4×10^{-03}
6	UTY	Y	7×10^{-14}	16	SLC25A6	X Y	6×10^{-04}	26	TUBA1B	12	5×10^{-03}
7	CD99	X Y	7×10^{-12}	17	FAM153A	5	1×10^{-03}	27	KLK6	19	5×10^{-03}
8	CYorf15B	Y	5×10^{-10}	18	RPS4X	X	1×10^{-03}	28	IFITM1	11	5×10^{-03}
9	TTTY15	Y	2×10^{-08}	19	PPP2CB	8	2×10^{-03}	29	HBB	11	5×10^{-03}
10	EIF1AY	Y	4×10^{-07}	20	H2AFY	5	2×10^{-03}	30	PRKY	Y	7×10^{-03}

Table 8: Comparison of gender study gene rankings before and after adjustment. The data has been fully preprocessed (BG + NM + LS). The p-values were calculated using Limma.

Unadjusted

Rank	Gene	Chrom	Adj. P	Rank	Gene	Chrom	Adj. P	Rank	Gene	Chrom	Adj. P
1	RPS4Y1	Y	1	11	CRYM	16	1	21	GAP43	3	1
2	DDX3Y	Y	1	12	XIST	X	1	22	GNAS	20	1
3	HBB	11	1	13	CALM1	14	1	23	SPP1	4	1
4	HBA1	16	1	14	ACTB	7	1	24	SNAP25	20	1
5	HBB	11	1	15	PFN2	3	1	25	UTY	Y	1
6	CIRBP	19	1	16	PRKAR1A	17	1	26	SLC17A7	19	1
7	KDM5D	Y	1	17	ACTB	7	1	27	RPS23	5	1
8	GAPDH	12	1	18	SLC25A6	X Y	1	28	RPS21	20	1
9	GAD1	2	1	19	GAPDH	12	1	29	HSP90AB1	6	1
10	USP9Y	Y	1	20	GNAS	20	1	30	ATP5A1	18	1

RUV-2 (SVD), $k = 10$, housekeeping genes

Rank	Gene	Chrom	Adj. P	Rank	Gene	Chrom	Adj. P	Rank	Gene	Chrom	Adj. P
1	RPS4Y1	Y	8×10^{-30}	11	HBB	11	2×10^{-05}	21	PDE4DIP	1	1×10^{-03}
2	KDM5D	Y	4×10^{-21}	12	HBA1	16	4×10^{-05}	22	DDX3X	X	1×10^{-03}
3	DDX3Y	Y	4×10^{-14}	13	SLC25A6	X Y	1×10^{-04}	23	EIF1AY	Y	3×10^{-03}
4	XIST	X	1×10^{-11}	14	PCDH11X	X	1×10^{-04}	24	NCL	2	3×10^{-03}
5	CD99	X Y	4×10^{-11}	15	HBB	11	2×10^{-04}	25	HDHD1A	X	3×10^{-03}
6	UTY	Y	5×10^{-10}	16	SLC25A6	X Y	5×10^{-04}	26	GTPBP6	X Y	3×10^{-03}
7	USP9Y	Y	3×10^{-09}	17	USP9X	X	6×10^{-04}	27	ASMTL	X Y	3×10^{-03}
8	RPS4X	X	5×10^{-07}	18	FDPS	1	6×10^{-04}	28	IFITM1	11	4×10^{-03}
9	TTTY15	Y	7×10^{-07}	19	HBG1	11	6×10^{-04}	29	PIN1	19	5×10^{-03}
10	CYorf15B	Y	1×10^{-05}	20	NLGN4Y	Y	6×10^{-04}	30	POLD2	7	6×10^{-03}

Table 9: Comparison of gender gene rankings before and after adjustment. The data has not been preprocessed. The p-values were calculated using Limma.

E Additional Alzheimer's Study Figures and Tables

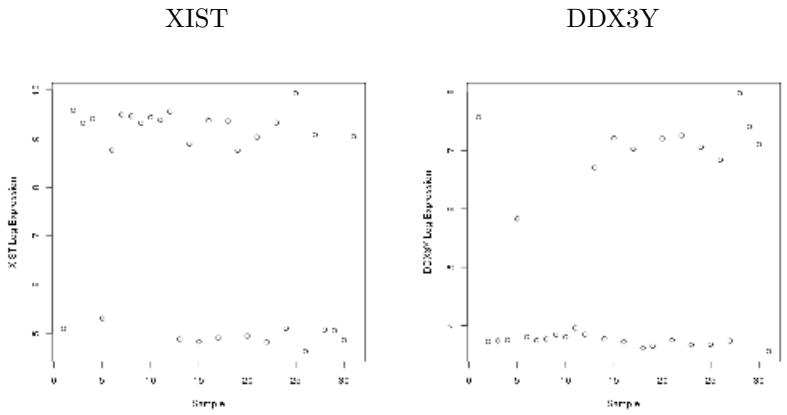


Figure 18: Plots of XIST and DDX3Y expression levels in the Alzheimer's study. The horizontal axis is just sample index. It is clear which samples are male and which are female. Data was preprocessed.

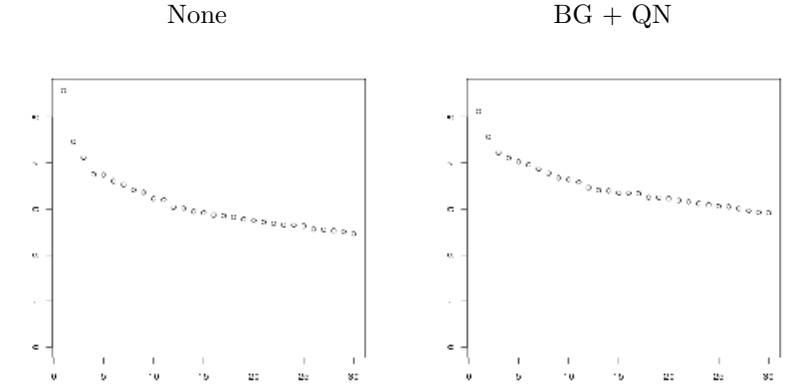


Figure 19: Alzheimer's study scree plots with and without preprocessing. Left: No preprocessing; Right: Background correction / quantile normalization. All genes were included in the eigenanalysis.

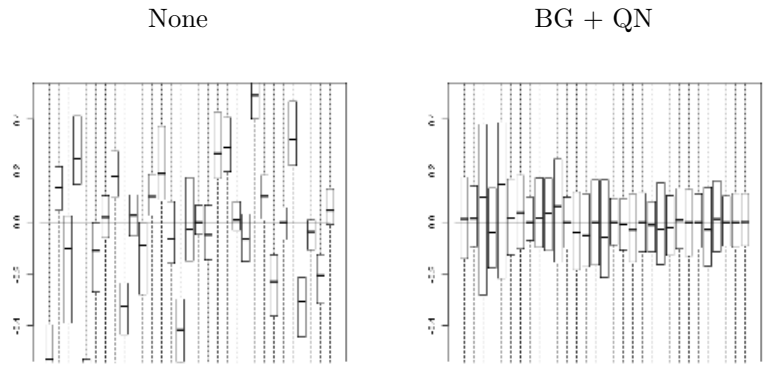


Figure 20: Alzheimer's study RLE plots with and without preprocessing. Left: No preprocessing; Right: Background correction / quantile normalization.

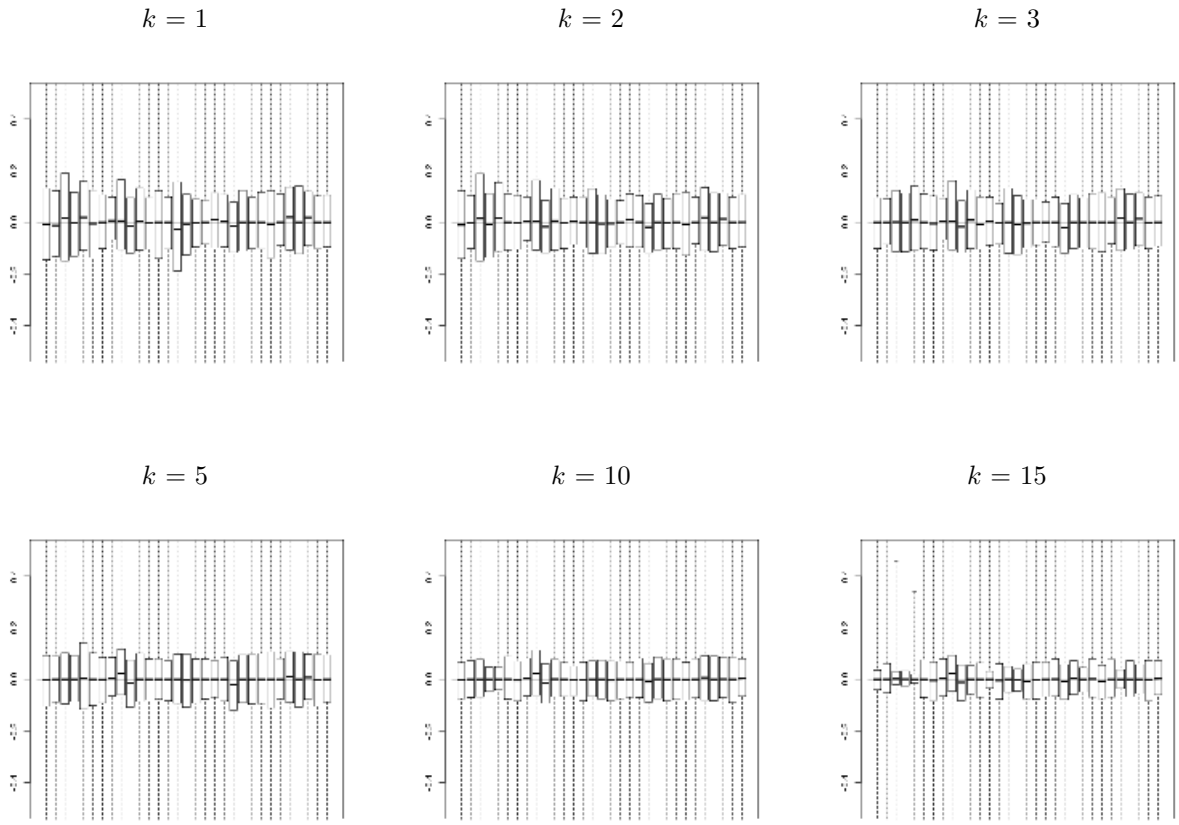


Figure 21: Alzheimer's study RLE plots after adjustment, using various values of k. Data was preprocessed (BG+QN). The factors were computed by SVD on the housekeeping genes.

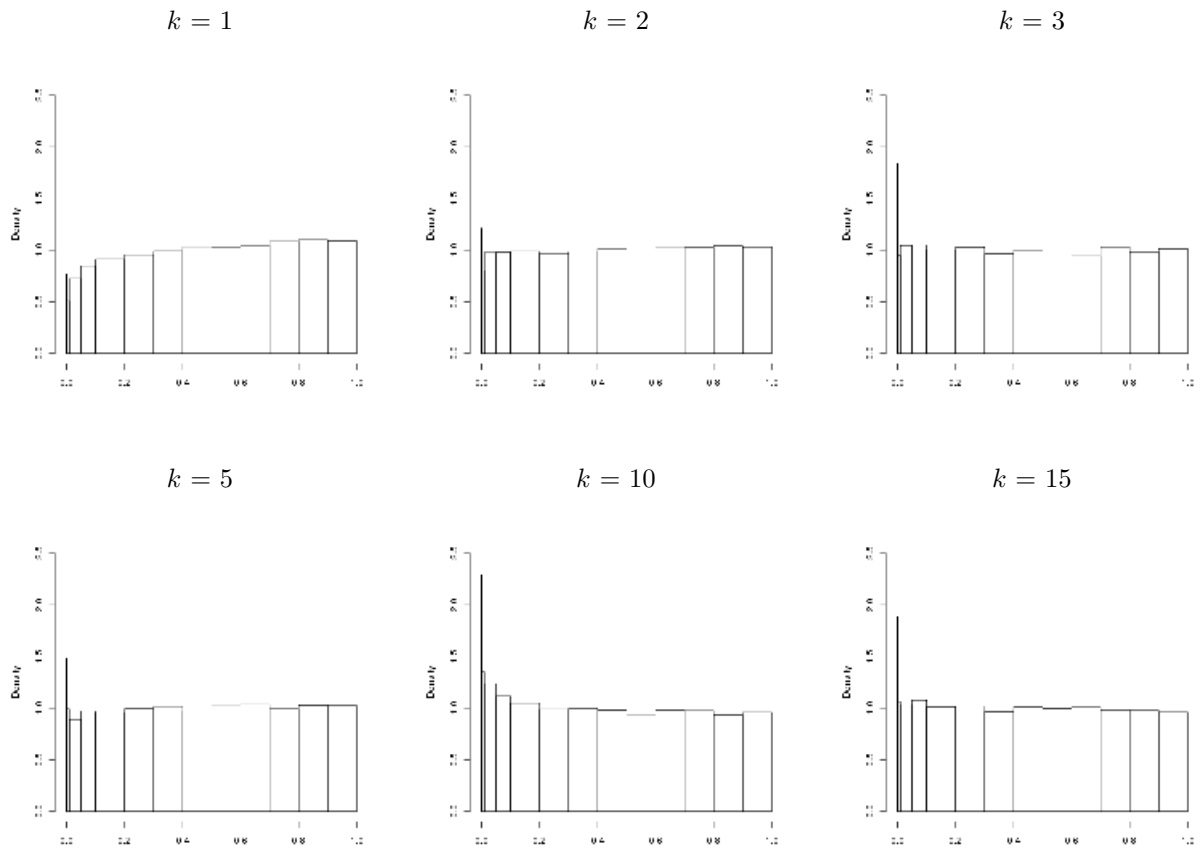


Figure 22: Alzheimer's study P-value histograms after adjustment, using various values of k. Histogram breakpoints are at 0.001, 0.01, 0.05, and 0.1, 0.2, 0.3, etc. Data was preprocessed (BG+QN). The factors were computed by SVD on the housekeeping genes. P-values were computed using Limma.

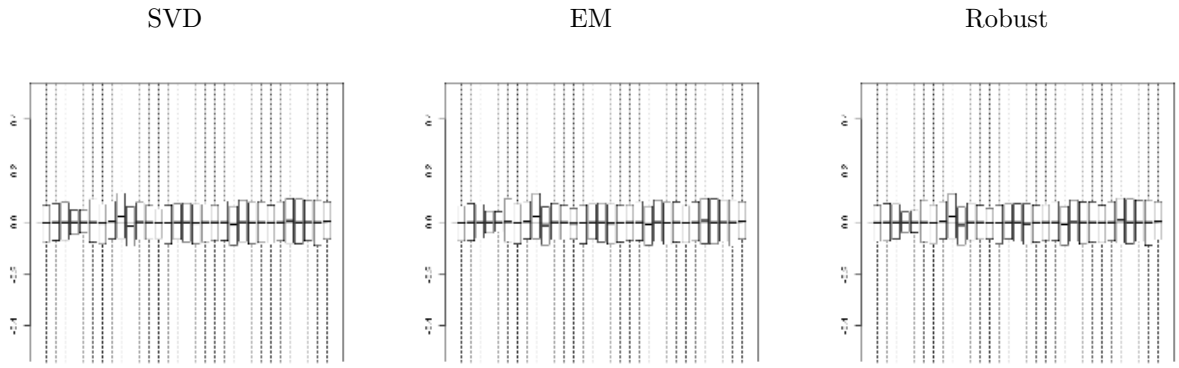


Figure 23: Alzheimer's study RLE plots after adjustment ($k = 10$), using different methods of factor analysis. The data was preprocessed (BG + QN). Factors were computed using the housekeeping genes.

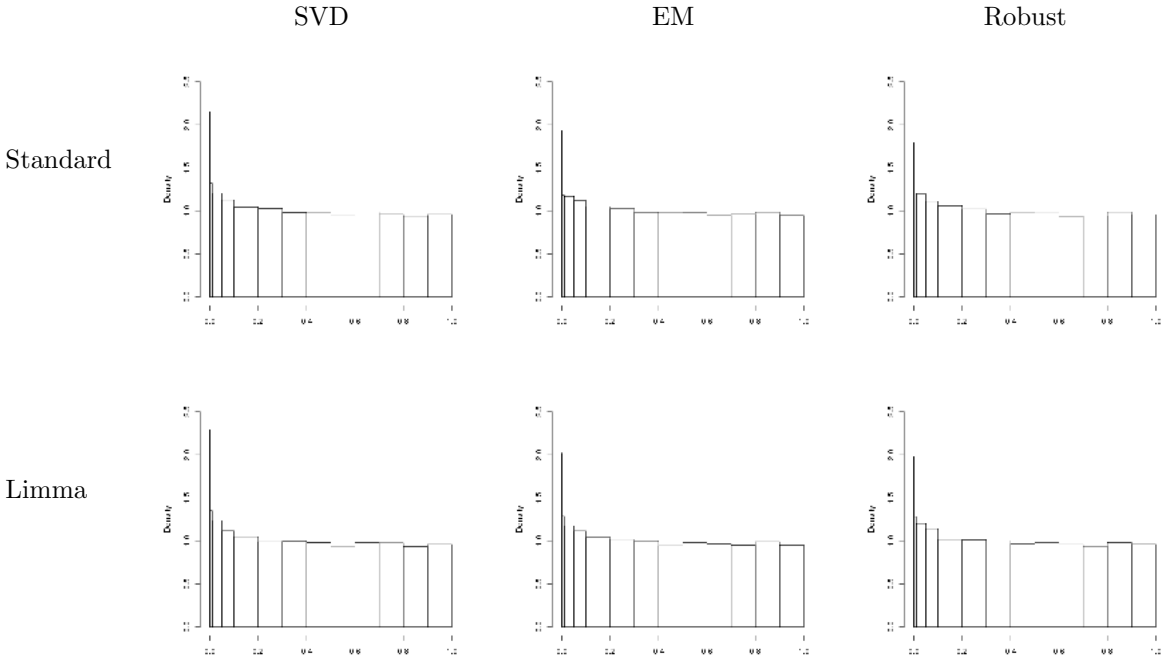


Figure 24: Alzheimer's study p-value histograms after adjustment ($k = 10$), using different methods of factor analysis, and different methods of computing p-values. The data was preprocessed (BG + QN). Factors were computed using the housekeeping genes.

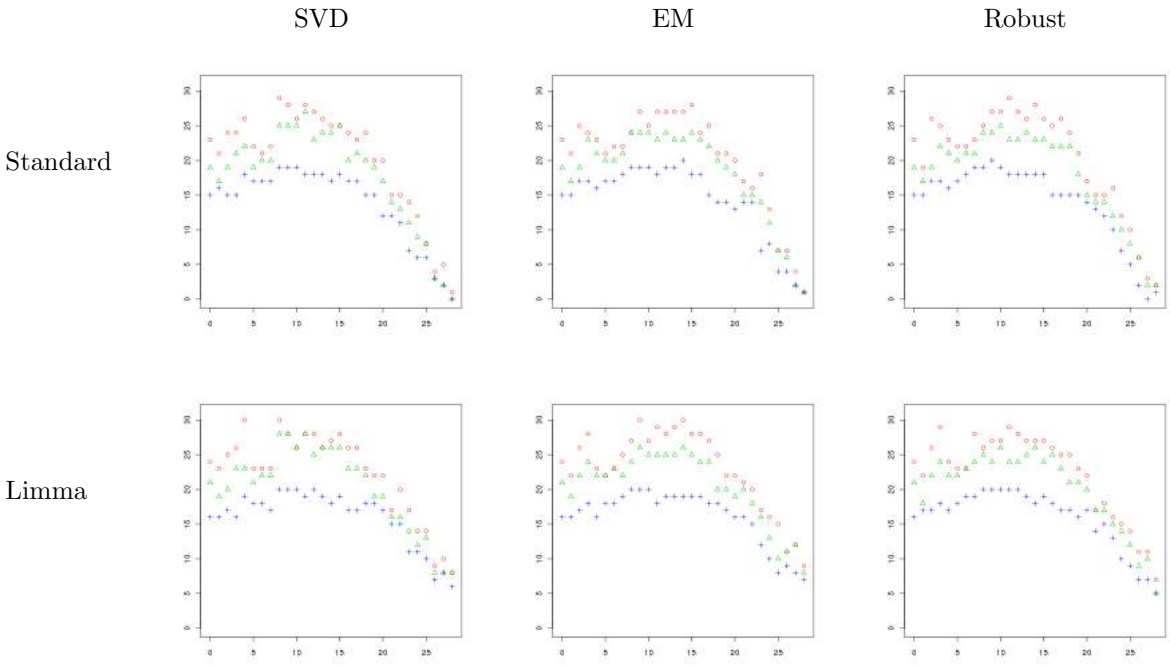


Figure 25: Comparison of the performance of variants of RUV-2 in the Alzheimer's study. The number of X / Y genes discovered is plotted as a function of k. Genes were ranked by p-value; results are shown for the number of X / Y genes ranked in the top 20 (plus), top 40 (triangle), and top 60 (circle). $k = 0$ corresponds to no adjustment. Data was preprocessed (BG + QN). PCs were computed using the housekeeping genes.

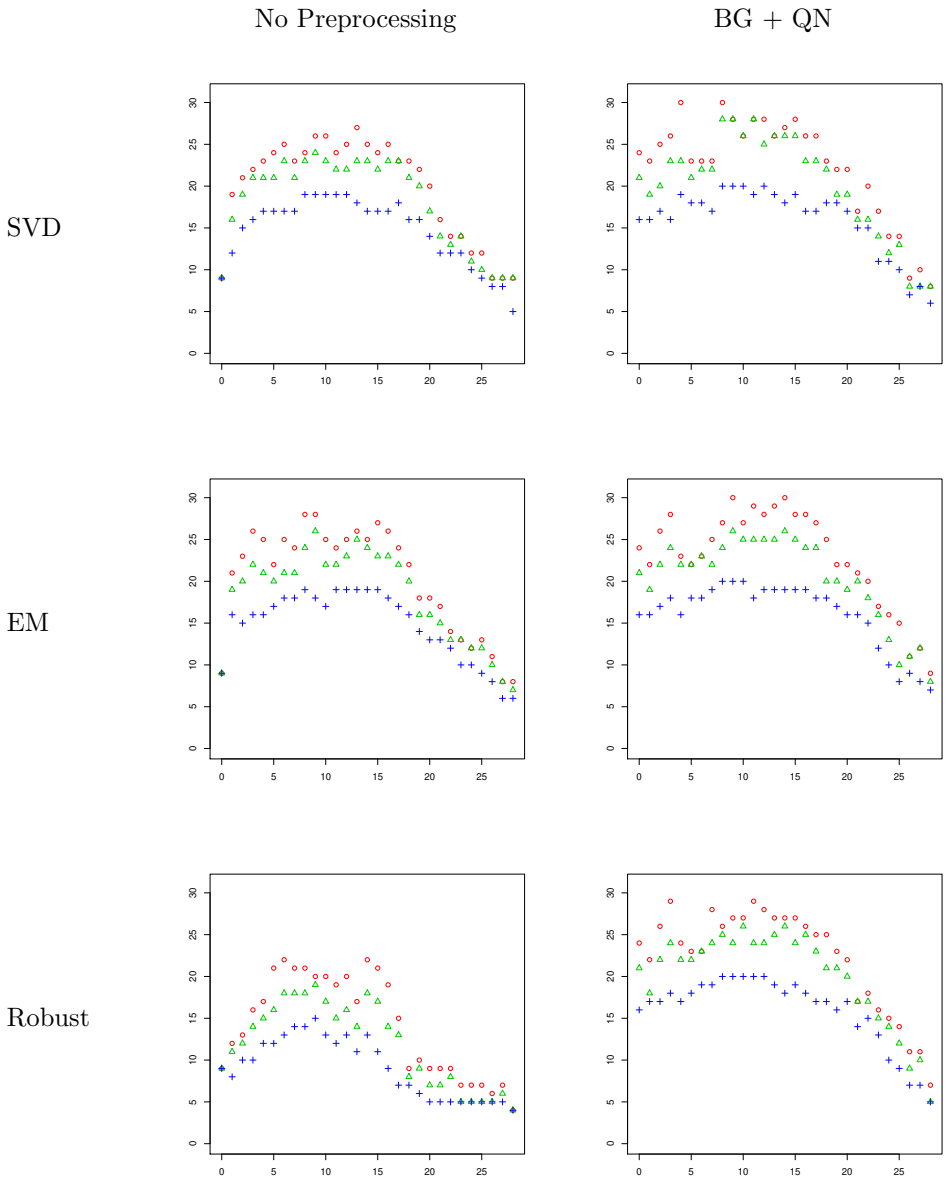
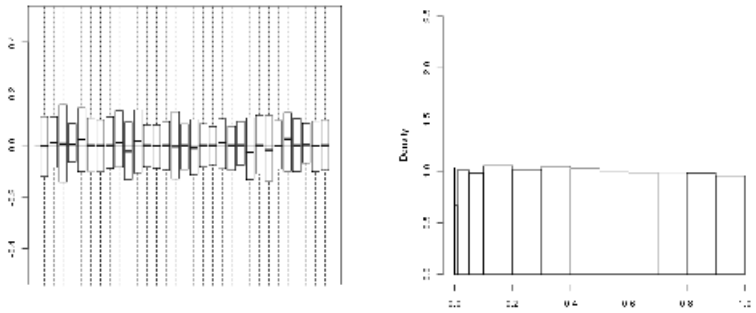


Figure 26: Comparison of the performance of different factor analysis methods with and without preprocessing in the Alzheimer's study. The number of X / Y genes discovered is plotted as a function of k. Genes were ranked by p-value; results are shown for the number of X / Y genes ranked in the top 20 (plus), top 40 (triangle), and top 60 (circle). $k = 0$ corresponds to no adjustment. PCs were computed using the housekeeping genes. P-values were computed using Limma.

SVA IRW



SVA Two Step

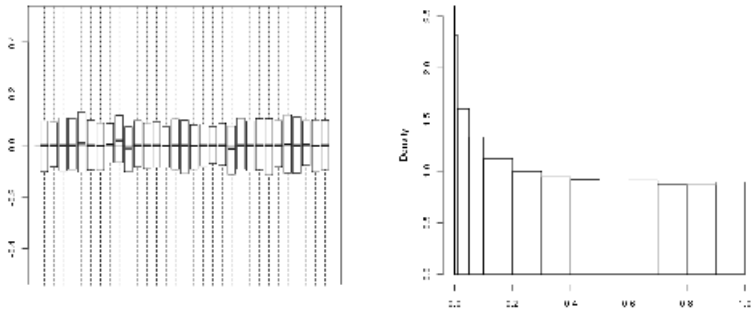


Figure 27: Alzheimer's study RLE plots and p-value histograms after adjustments by SVA. P-values were computed using Limma. Data was preprocessed (BG + QN).

Unadjusted

Rank	Gene	Chrom	Adj. P	Rank	Gene	Chrom	Adj. P	Rank	Gene	Chrom	Adj. P
1	XIST	X	3×10^{-28}	11	TTTY15	Y	1×10^{-03}	21	KLHL22	22	1
2	XIST	X	8×10^{-27}	12	UTY	Y	1×10^{-03}	22	TAPBPL	12	1
3	DDX3Y	Y	2×10^{-20}	13	CYorf15B	Y	2×10^{-03}	23	LPAR6	13	1
4	RPS4Y1	Y	3×10^{-15}	14	DDX3Y	Y	8×10^{-02}	24	PRKAB1	12	1
5	KDM5D	Y	2×10^{-11}	15	ZBED1	X Y	1	25	GPX3	5	1
6	EIF1AY	Y	1×10^{-07}	16	PLCXD1	X Y	1	26	HDHD1A	X	1
7	USP9Y	Y	8×10^{-07}	17	CDK10	16	1	27	KDM6A	X	1
8	EIF1AY	Y	1×10^{-06}	18	GPX3	5	1	28	CD99	X Y	1
9	NLGN4Y	Y	5×10^{-06}	19	NCKAP1	2	1	29	PRKAR1B	7	1
10	NCRNA00185	Y	1×10^{-05}	20	UBAP2L	1	1	30	MARCH1	4	1

RUV-2 (SVD), $k = 10$, housekeeping genes

Rank	Gene	Chrom	Adj. P	Rank	Gene	Chrom	Adj. P	Rank	Gene	Chrom	Adj. P
1	XIST	X	1×10^{-21}	11	TTTY15	Y	1×10^{-05}	21	COL4A5	X	7×10^{-02}
2	XIST	X	1×10^{-21}	12	CYorf15B	Y	1×10^{-05}	22	FBXO9	6	7×10^{-02}
3	DDX3Y	Y	4×10^{-16}	13	CD99	X Y	5×10^{-04}	23	PLCL1	2	7×10^{-02}
4	RPS4Y1	Y	1×10^{-15}	14	NA	X	3×10^{-03}	24	DDX3X	X	0.1
5	KDM5D	Y	1×10^{-11}	15	CD99	X Y	3×10^{-03}	25	KDM6A	X	0.1
6	USP9Y	Y	8×10^{-09}	16	UTY	Y	1×10^{-02}	26	DOPEY1	6	0.1
7	EIF1AY	Y	4×10^{-08}	17	DDX3Y	Y	3×10^{-02}	27	DDX3X	X	0.1
8	EIF1AY	Y	2×10^{-07}	18	RPS4X	X	3×10^{-02}	28	ZBED1	X Y	0.1
9	NLGN4Y	Y	4×10^{-07}	19	KDM6A	X	4×10^{-02}	29	RARS	5	0.2
10	NCRNA00185	Y	1×10^{-06}	20	PLCXD1	X Y	6×10^{-02}	30	WNT8B	10	0.2

Table 10: Comparison of gene rankings before and after adjustment (SVD, $k = 10$) in the Alzheimer's study. The data has been preprocessed (BG + NM). The p-values were computed using Limma.

F Additional TCGA Figures and Tables

F.1 Exon Array Data

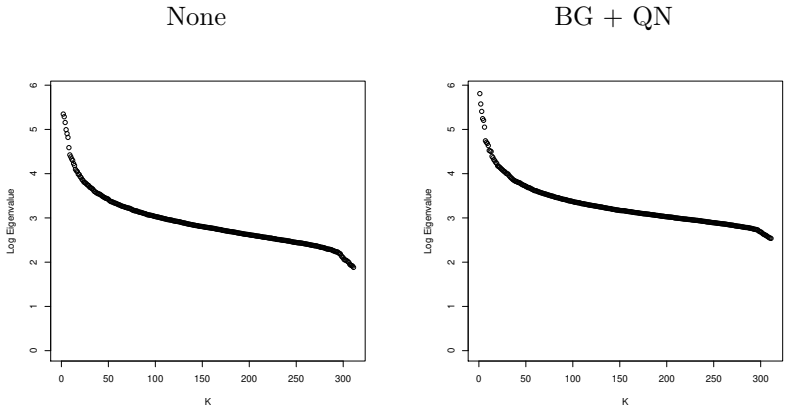


Figure 28: TCGA exon array scree plots with and without preprocessing. Left: No preprocessing; Right: Background correction / quantile normalization. All genes were included in the eigenanalysis.

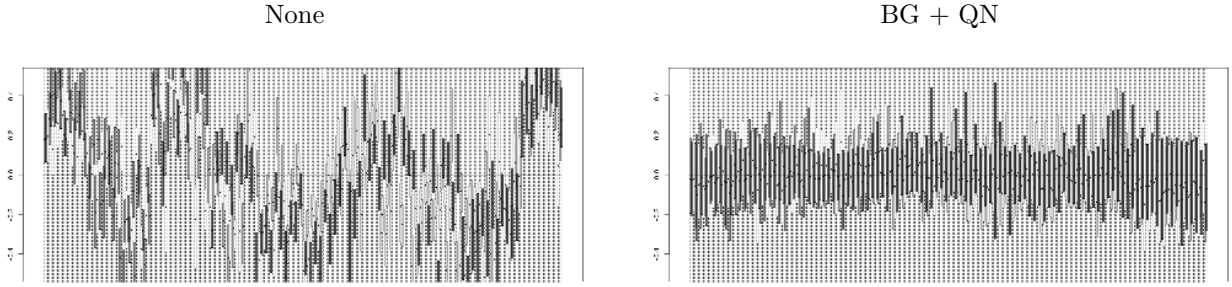


Figure 29: TCGA exon array RLE plots with and without preprocessing. Left: No preprocessing; Right: Background correction / quantile normalization.

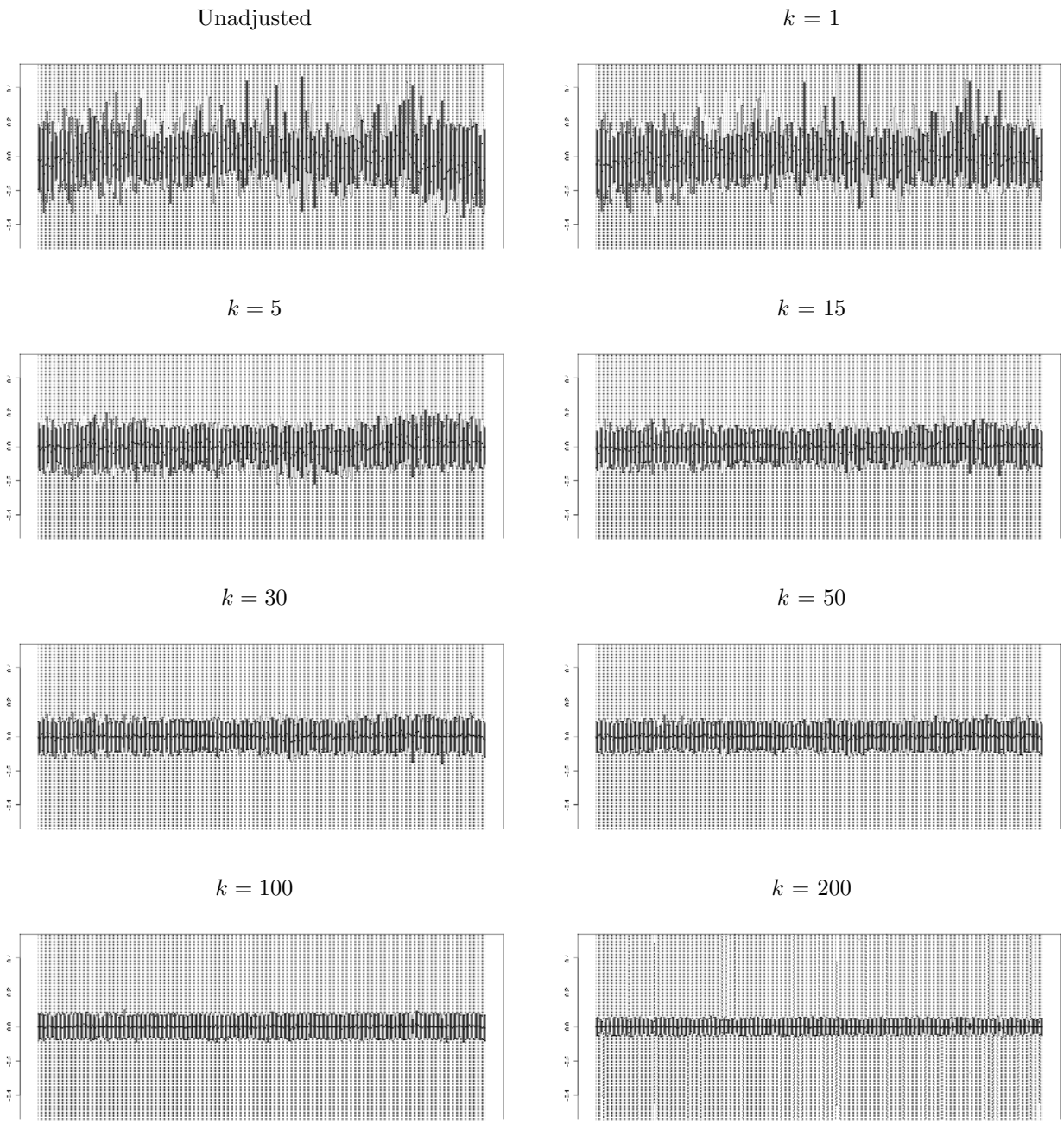


Figure 30: TCGA exon array RLE plots after adjustment, using various values of k. Data was preprocessed (BG+QN). The factors were computed by SVD on the housekeeping genes.

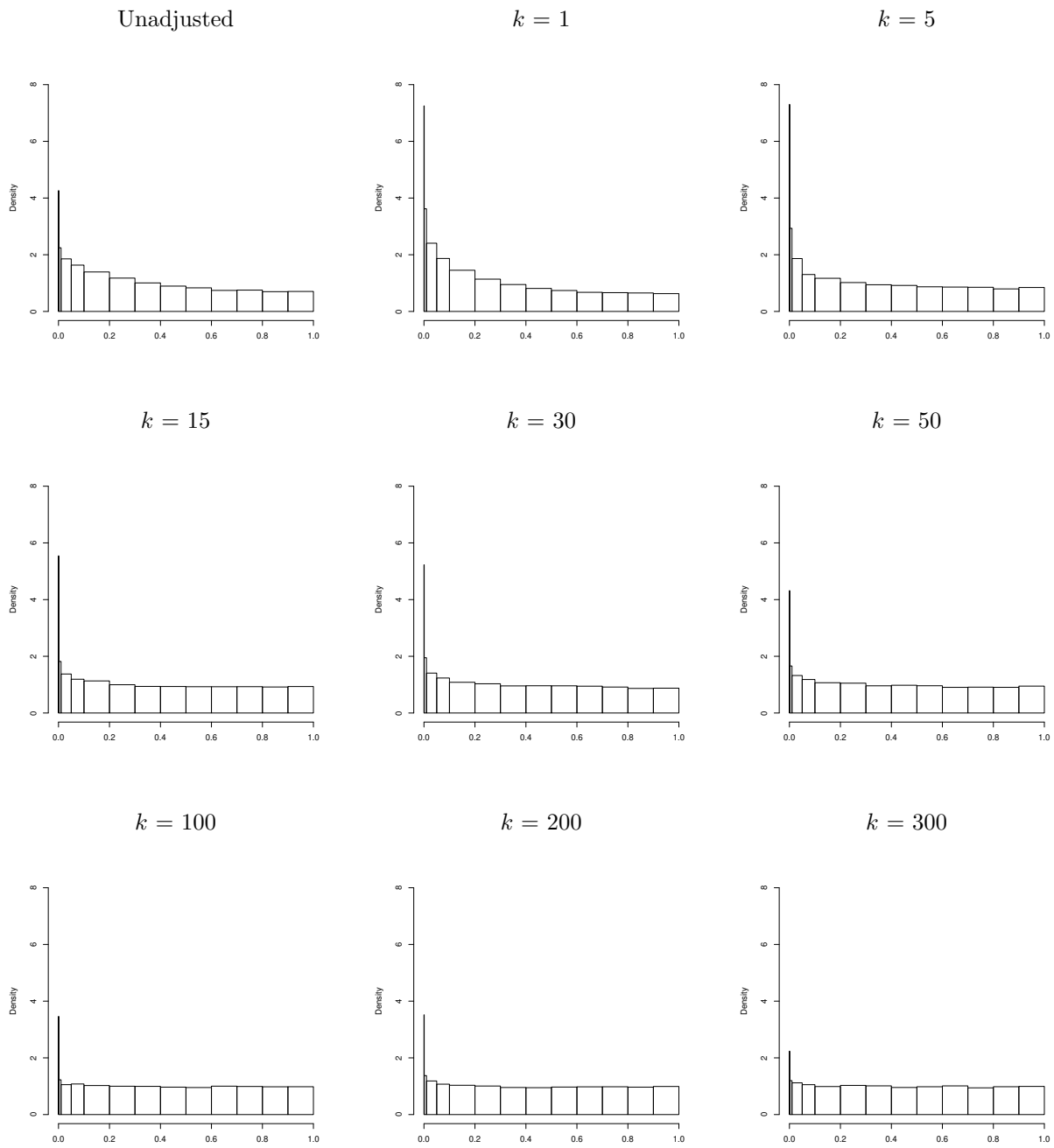
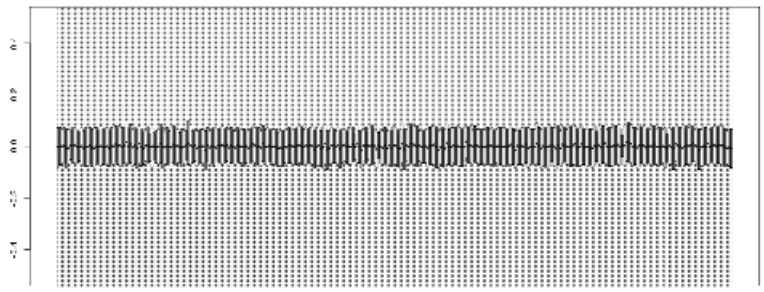
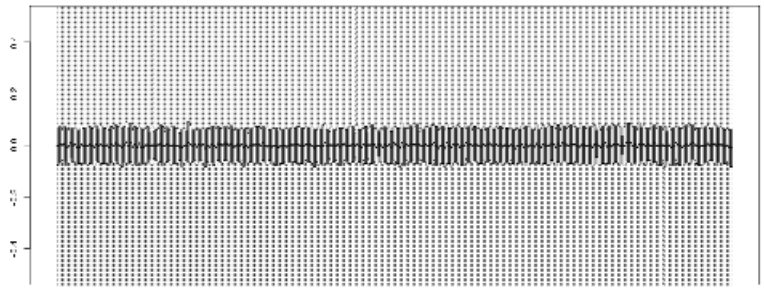


Figure 31: TCGA exon array p-value histograms after adjustment, using various values of k. Histogram breakpoints are at 0.001, 0.01, 0.05, and 0.1, 0.2, 0.3, etc. Data was preprocessed (BG+QN). The factors were computed by SVD on the housekeeping genes. P-values were computed using Limma.

SVD



EM



Robust

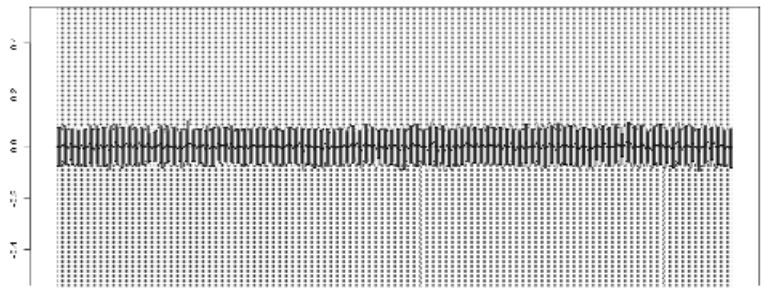


Figure 32: TCGA exon array RLE plots after adjustment ($k = 100$), using different methods of factor analysis. The data was preprocessed (BG + QN). The factors were computed using housekeeping genes.

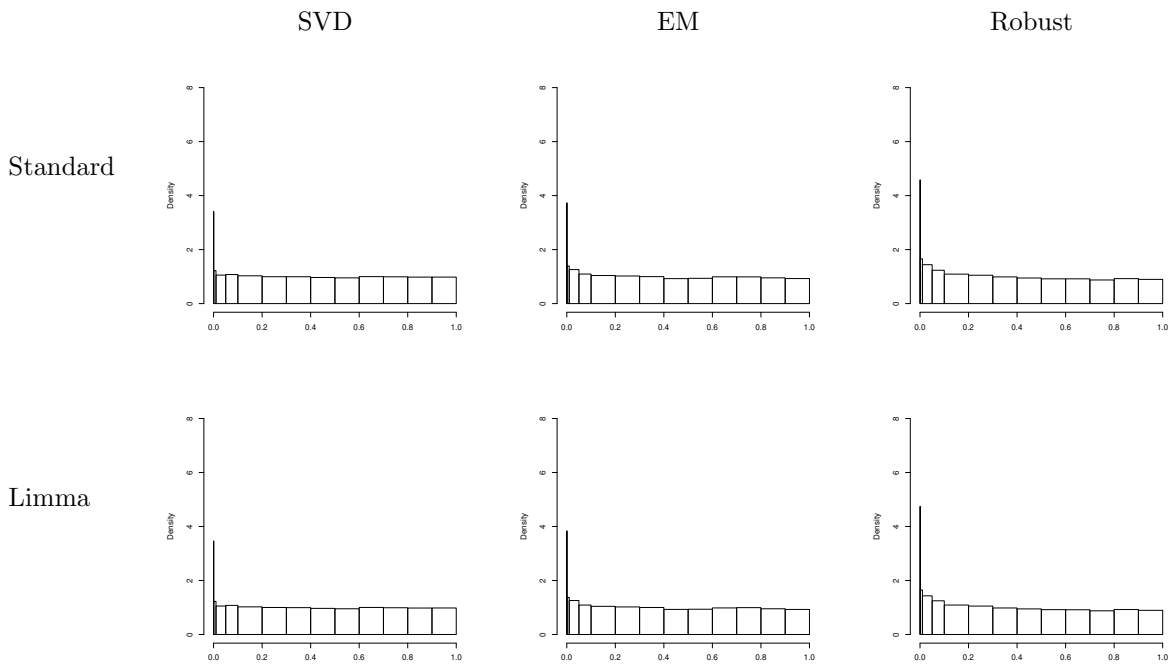


Figure 33: TCGA exon array p-value histograms after adjustment ($k = 100$), using different methods of factor analysis, and different methods of computing p-values. The data was preprocessed (BG + QN). The factors were computed using housekeeping genes.

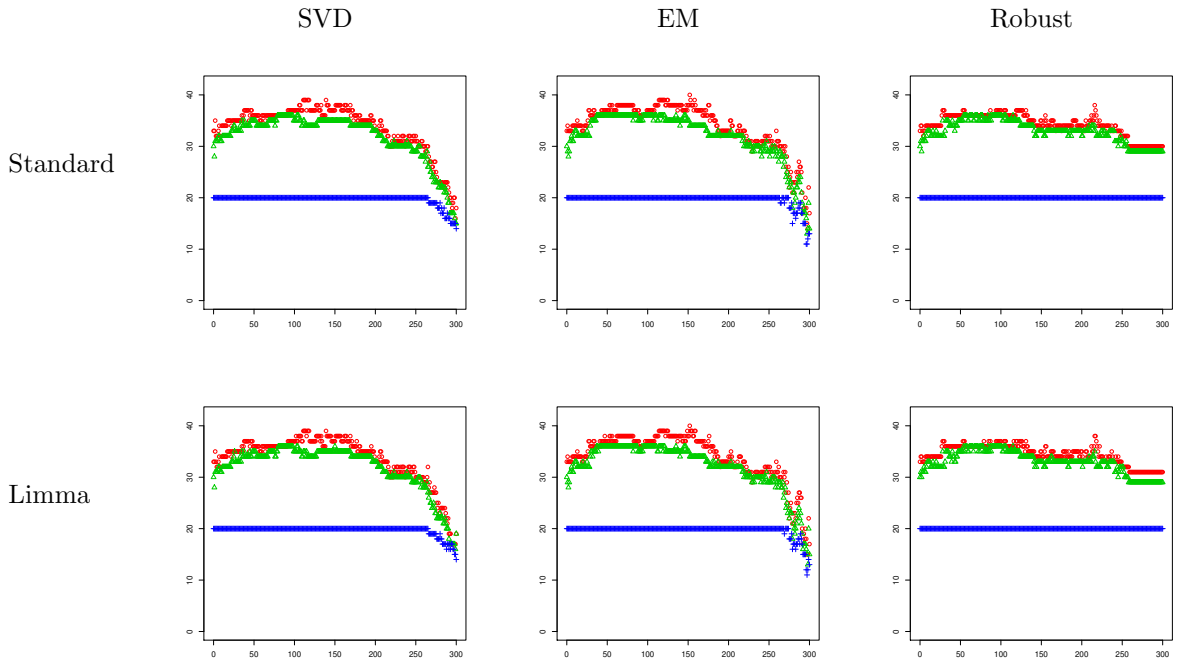


Figure 34: Comparison of the performance of variants of RUV-2 in the TCGA exon array data. The number of X / Y genes discovered is plotted as a function of k. Genes were ranked by p-value; results are shown for the number of X / Y genes ranked in the top 20 (plus), top 40 (triangle), and top 60 (circle). $k = 0$ corresponds to no adjustment. Data was preprocessed (BG + NM). PCs were computed using the housekeeping genes.

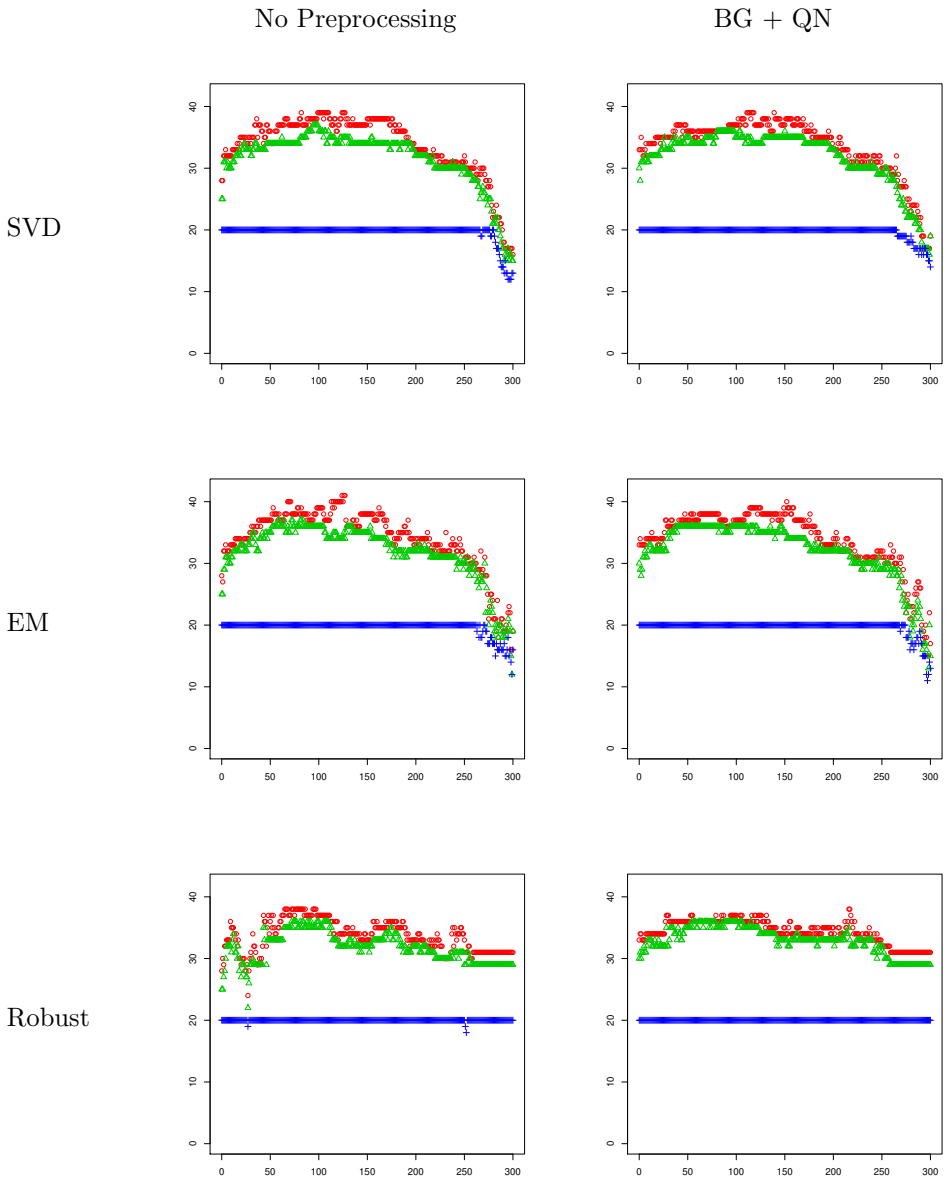
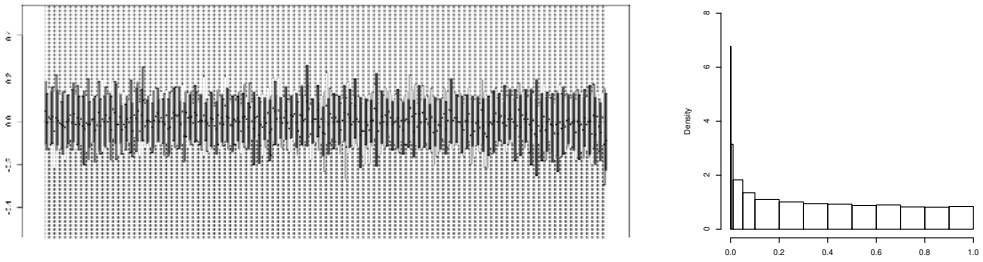


Figure 35: Comparison of the performance of different factor analysis methods with and without preprocessing in the TCGA exon array data. The number of X / Y genes discovered is plotted as a function of k. Genes were ranked by p-value; results are shown for the number of X / Y genes ranked in the top 20 (plus), top 40 (triangle), and top 60 (circle). $k = 0$ corresponds to no adjustment. PCs were computed using the housekeeping genes. P-values were computed using Limma.

SVA IRW



SVA Two Step

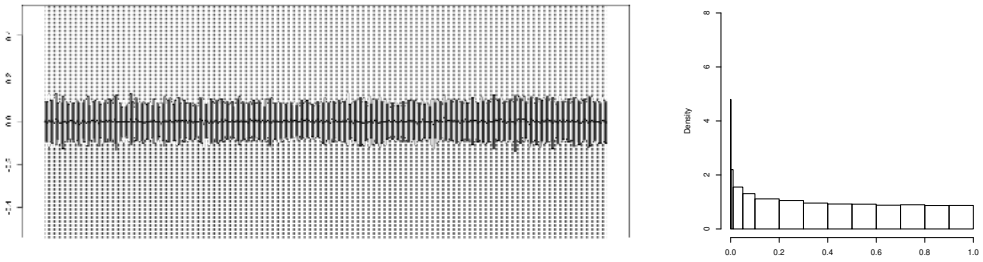


Figure 36: TCGA exon array RLE plots and p-value histograms after adjustments by SVA. The data was preprocessed. The p-values were computed using Limma.

Unadjusted

Rank	Gene	Chrom	Adj. P	Rank	Gene	Chrom	Adj. P	Rank	Gene	Chrom	Adj. P
1	RPS4Y1	Y	1×10^{-159}	16	JARID1C	X	7×10^{-21}	31	LRRC47	1	3×10^{-2}
2	DDX3Y	Y	6×10^{-137}	17	ZFX	X	3×10^{-18}	32	SYAP1	X	3×10^{-2}
3	EIF1AY	Y	2×10^{-131}	18	UTX	X	1×10^{-13}	33	MYLK	3	3×10^{-2}
4	UTY	Y	3×10^{-125}	19	LOC554203	X	1×10^{-12}	34	IHPK1	3	5×10^{-2}
5	USP9Y	Y	3×10^{-120}	20	HDHD1A	X	1×10^{-12}	35	HTATIP2	11	5×10^{-2}
6	CYorf15A	Y	3×10^{-117}	21	CXorf15	X	5×10^{-8}	36	NLGN4X	X	6×10^{-2}
7	ZFY	Y	4×10^{-113}	22	DDX3X	X	9×10^{-7}	37	ZRSR2	X	6×10^{-2}
8	JARID1D	Y	2×10^{-112}	23	EIF1AX	X	2×10^{-6}	38	PRKCH	14	6×10^{-2}
9	RPS4Y2	Y	3×10^{-94}	24	RPS4X	X	1×10^{-5}	39	EDG3	na	7×10^{-2}
10	NLGN4Y	Y	1×10^{-92}	25	SRY	Y	2×10^{-4}	40	NUPL2	7	7×10^{-2}
11	CYorf15B	Y	4×10^{-81}	26	STIM2	4	4×10^{-3}	41	SH3PXD2A	10	7×10^{-2}
12	TMSB4Y	Y	7×10^{-48}	27	EIF2S3	X	5×10^{-3}	42	STS	X	7×10^{-2}
13	TTTY10	Y	1×10^{-47}	28	ZFP2	5	5×10^{-3}	43	PAPSS2	10	7×10^{-2}
14	PRKY	Y	5×10^{-31}	29	GEMIN8	X	7×10^{-3}	44	ABCA11	4	7×10^{-2}
15	TTTY14	Y	3×10^{-26}	30	MICAL2	11	1×10^{-2}	45	GGTLA1	22	7×10^{-2}

RUV-2 (SVD), $k = 100$, housekeeping genes

Rank	Gene	Chrom	Adj. P	Rank	Gene	Chrom	Adj. P	Rank	Gene	Chrom	Adj. P
1	RPS4Y1	Y	1×10^{-104}	16	TTTY14	Y	3×10^{-30}	31	CA5B	X	4×10^{-5}
2	DDX3Y	Y	2×10^{-88}	17	ZFX	X	3×10^{-28}	32	GEMIN8	X	4×10^{-5}
3	UTY	Y	2×10^{-83}	18	PRKY	Y	9×10^{-21}	33	ORC4L	2	2×10^{-3}
4	EIF1AY	Y	5×10^{-81}	19	HDHD1A	X	1×10^{-19}	34	CHM	X	3×10^{-3}
5	USP9Y	Y	5×10^{-81}	20	DDX3X	X	4×10^{-19}	35	GPR88	1	4×10^{-3}
6	CYorf15A	Y	2×10^{-73}	21	RPS4X	X	4×10^{-17}	36	FUNDC1	X	5×10^{-3}
7	ZFY	Y	3×10^{-73}	22	EIF2S3	X	9×10^{-17}	37	SRY	Y	1×10^{-2}
8	JARID1D	Y	5×10^{-71}	23	EIF1AX	X	1×10^{-16}	38	CENPA	2	2×10^{-2}
9	RPS4Y2	Y	8×10^{-71}	24	SYAP1	X	6×10^{-16}	39	INSR	19	3×10^{-2}
10	NLGN4Y	Y	3×10^{-60}	25	LOC554203	X	1×10^{-15}	40	MYL3	3	8×10^{-2}
11	CYorf15B	Y	2×10^{-57}	26	CXorf15	X	3×10^{-15}	41	FMNL2	2	8×10^{-2}
12	JARID1C	X	2×10^{-38}	27	SMC1A	X	4×10^{-15}	42	RNF213	17	8×10^{-2}
13	TMSB4Y	Y	1×10^{-34}	28	STS	X	6×10^{-13}	43	GIP	17	9×10^{-2}
14	UTX	X	1×10^{-31}	29	USP9X	X	1×10^{-5}	44	UBE1	X	9×10^{-2}
15	TTTY10	Y	3×10^{-30}	30	ZRSR2	X	2×10^{-5}	45	TRIM23	5	0.1

Table 11: Comparison of gene rankings before and after adjustment (SVD, $k = 100$) using the TCGA exon array data. The data has been preprocessed (BG + NM). The p-values were computed using Limma.

F.2 HG-U133a Data

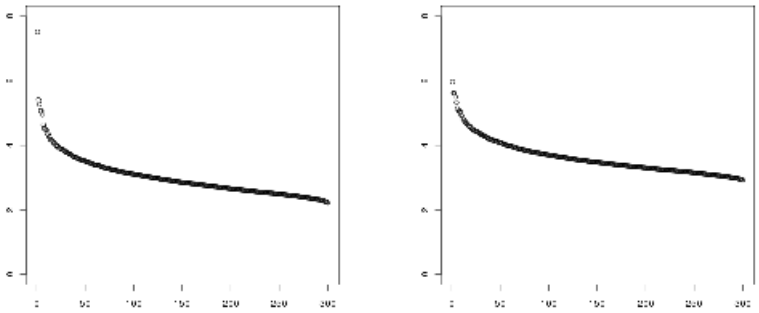


Figure 37: TCGA HT HG-133A scree plots with and without preprocessing. Left: No preprocessing; Right: Background correction / quantile normalization. All genes were included in the eigenanalysis.

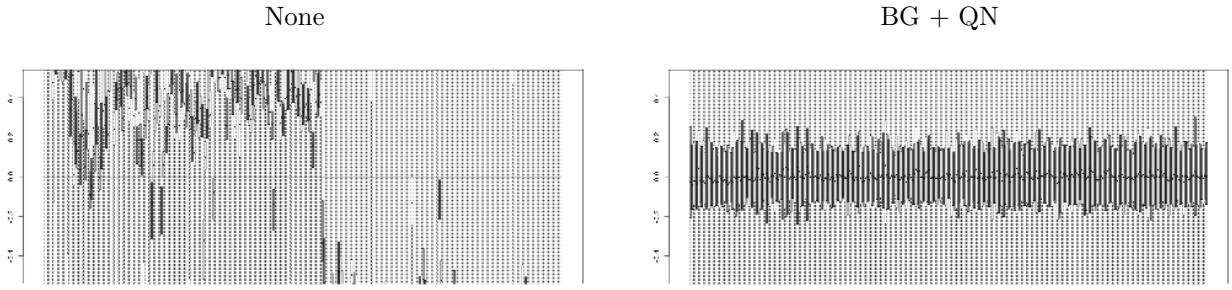


Figure 38: TCGA HT HG-133A RLE plots with and without preprocessing. Left: No preprocessing; Right: Background correction / quantile normalization.

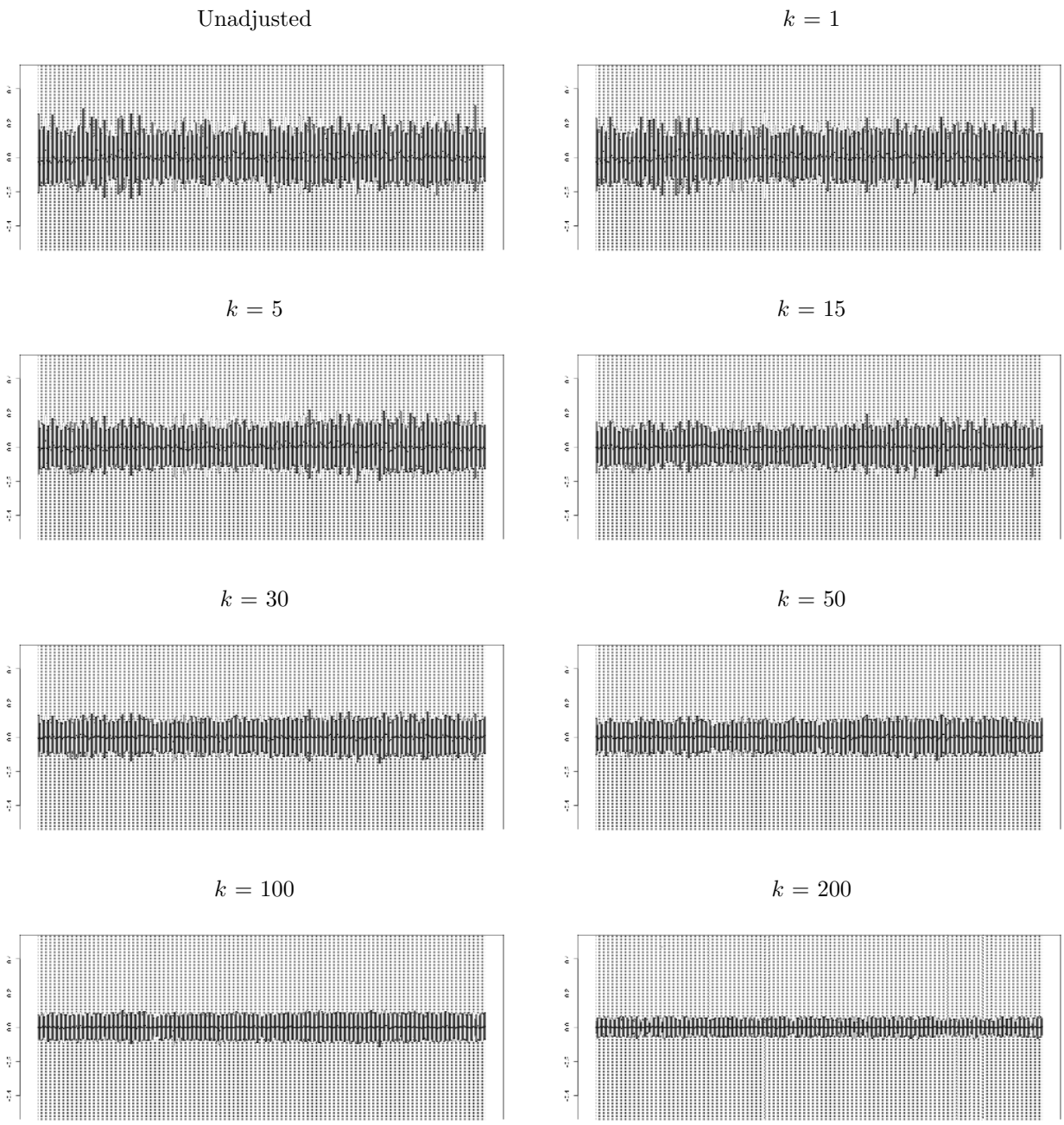


Figure 39: TCGA HT HG-133A RLE plots after adjustment, using various values of k. Data was preprocessed (BG+QN). The factors were computed by SVD on the housekeeping genes.

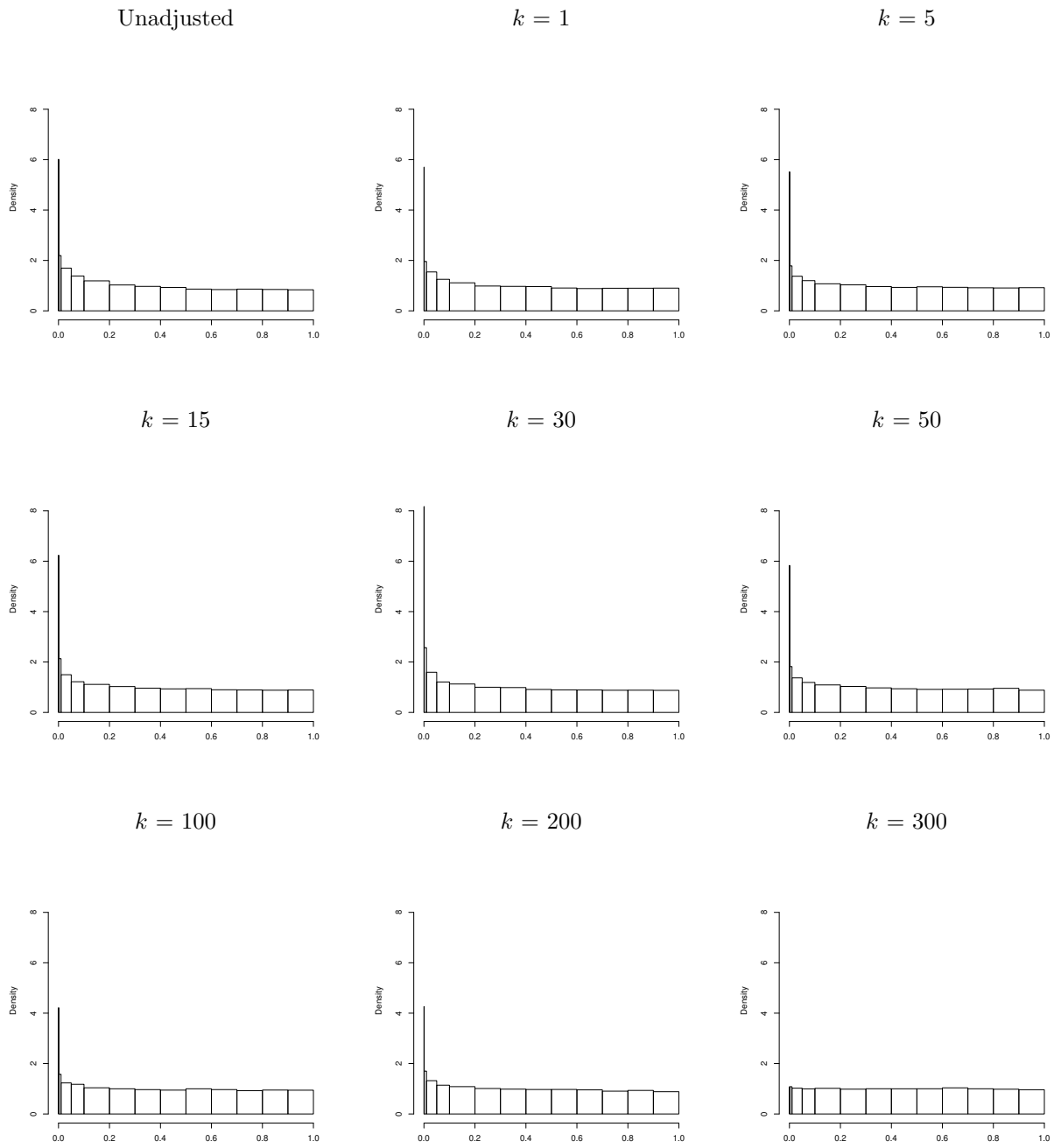
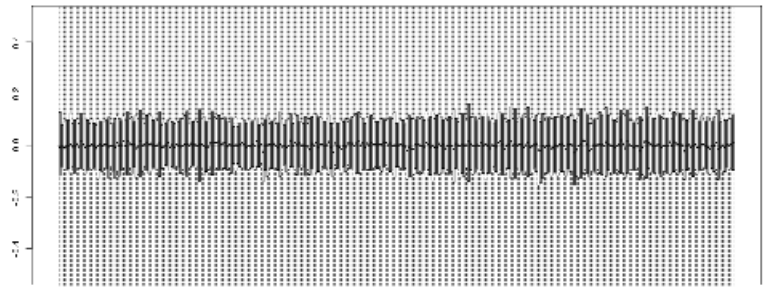
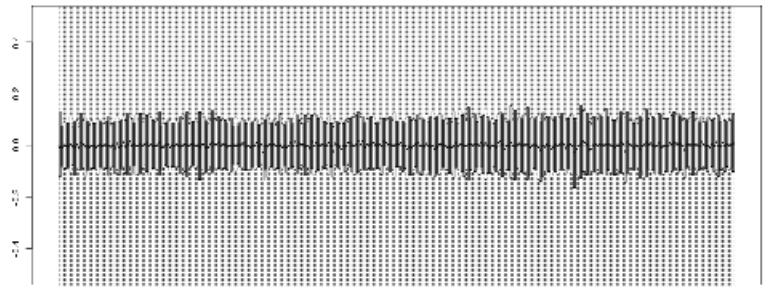


Figure 40: TCGA HT HG-133A p-value histograms after adjustment, using various values of k. Histogram breakpoints are at 0.001, 0.01, 0.05, and 0.1, 0.2, 0.3, etc. Data was preprocessed (BG+QN). The factors were computed by SVD on the housekeeping genes. P-values were computed using Limma.

SVD



EM



Robust

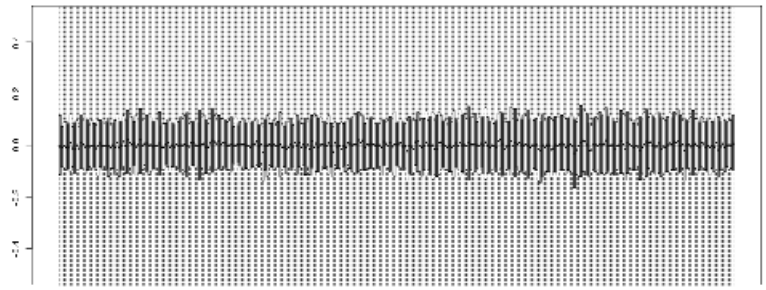


Figure 41: TCGA HT HG-133A RLE plots after adjustment ($k = 30$), using different methods of factor analysis. The data was preprocessed (BG + QN). The factors were computed using housekeeping genes.

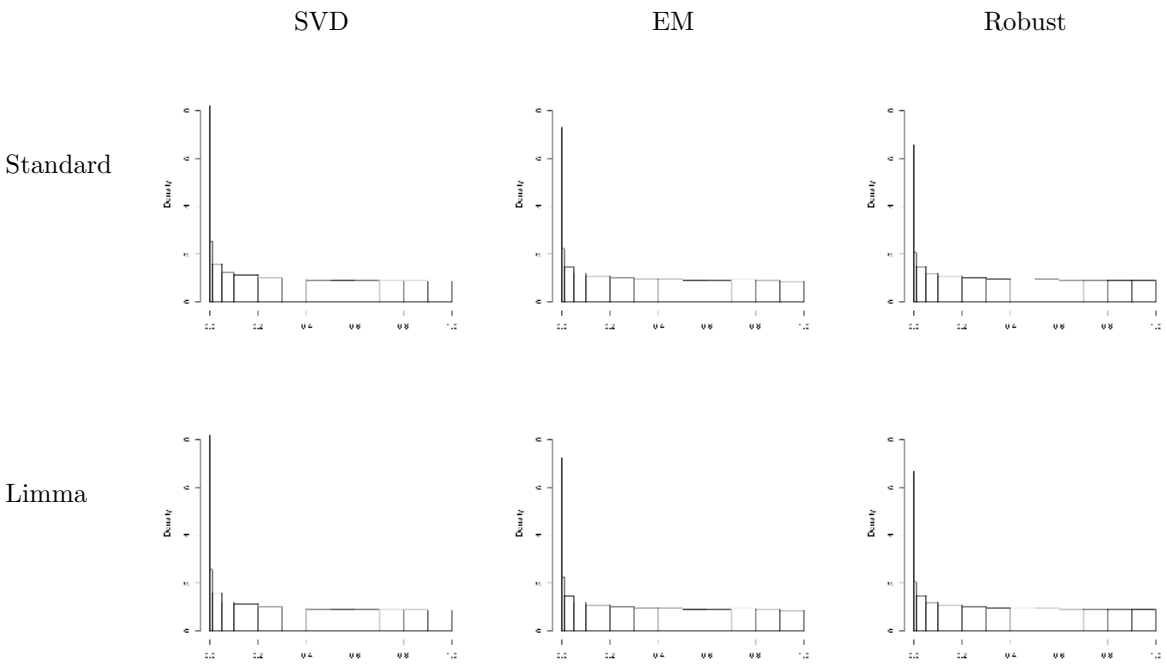


Figure 42: TCGA HT HG-133A p-value histograms after adjustment ($k = 30$), using different methods of factor analysis, and different methods of computing p-values. The data was preprocessed (BG + QN). The factors were computed using housekeeping genes.

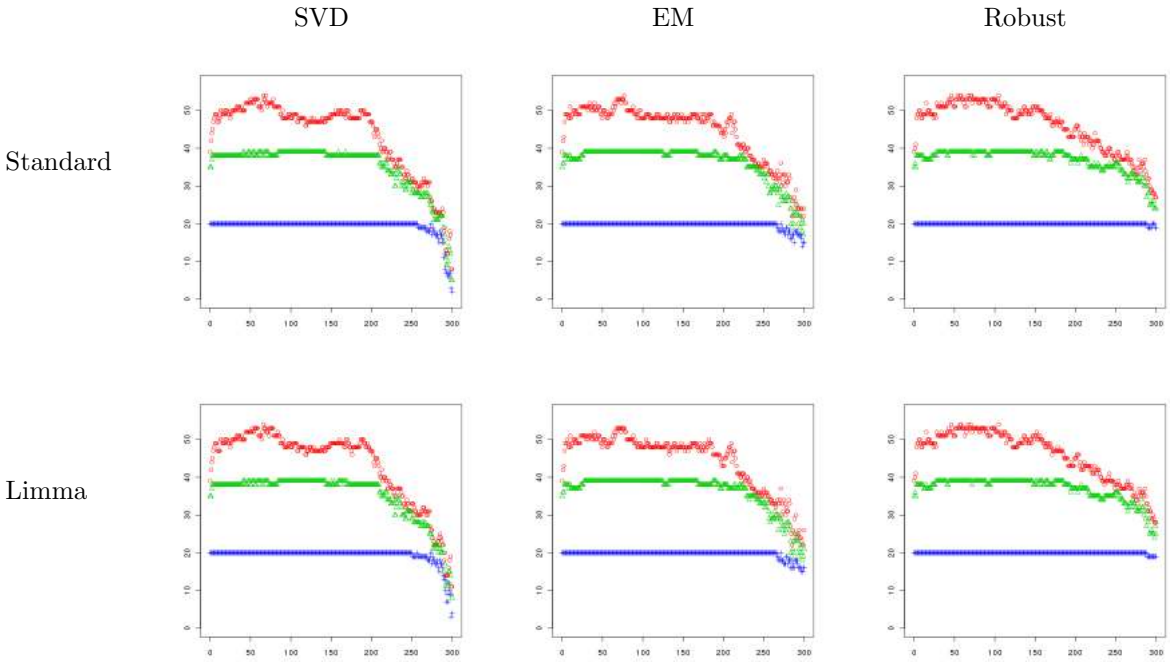


Figure 43: Comparison of the performance of variants of RUV-2 in the TCGA HT HG-U133A data. The number of X / Y genes discovered is plotted as a function of k. Genes were ranked by p-value; results are shown for the number of X / Y genes ranked in the top 20 (plus), top 40 (triangle), and top 60 (circle). $k = 0$ corresponds to no adjustment. Data was preprocessed (BG + NM). PCs were computed using the housekeeping genes.

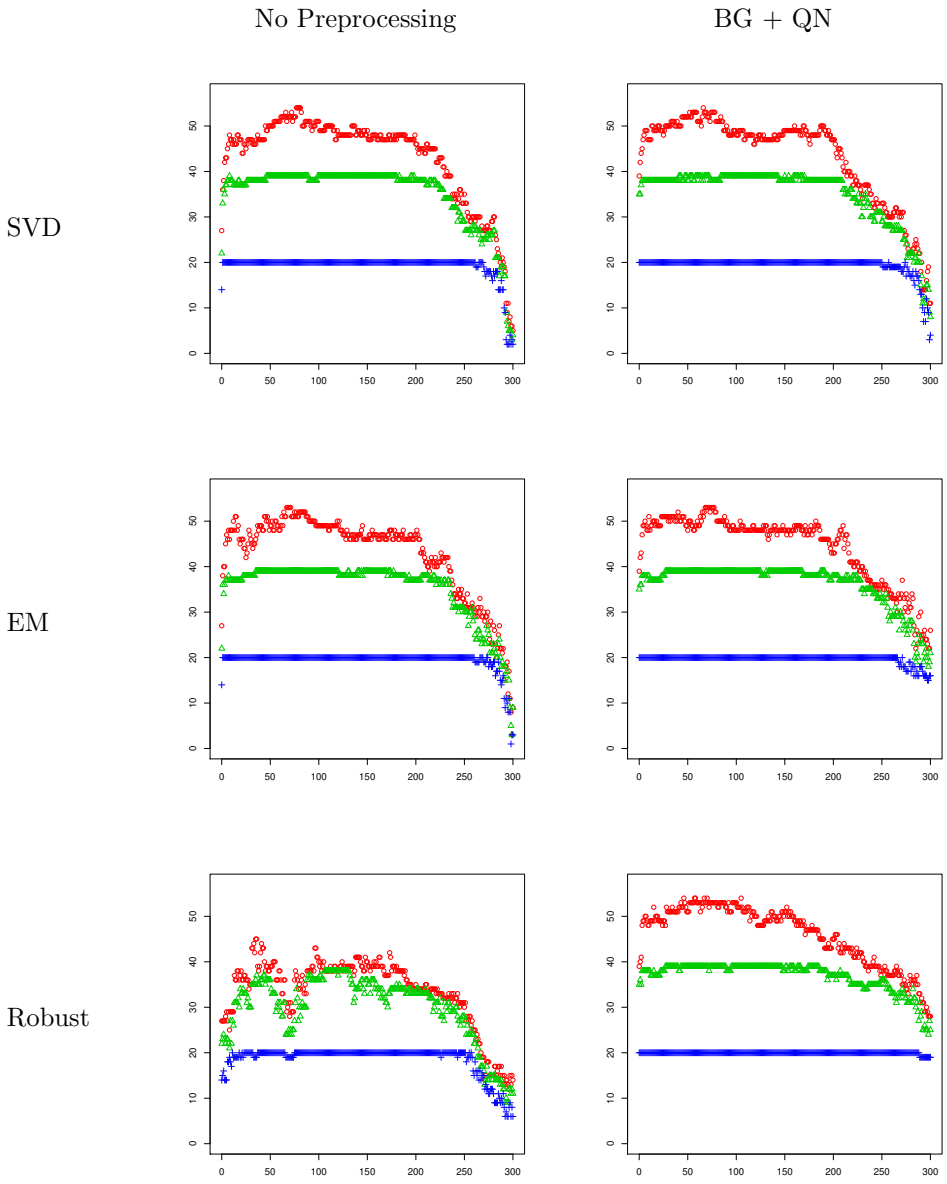
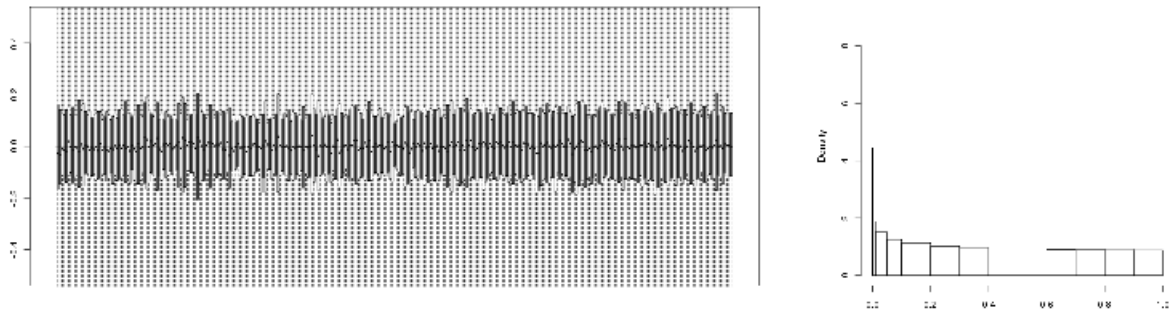


Figure 44: Comparison of the performance of different factor analysis methods with and without preprocessing in the TCGA HT HG-U133A data. The number of X / Y genes discovered is plotted as a function of k. Genes were ranked by p-value; results are shown for the number of X / Y genes ranked in the top 20 (plus), top 40 (triangle), and top 60 (circle). $k = 0$ corresponds to no adjustment. PCs were computed using the housekeeping genes. P-values were computed using Limma.

SVA IRW



SVA Two Step

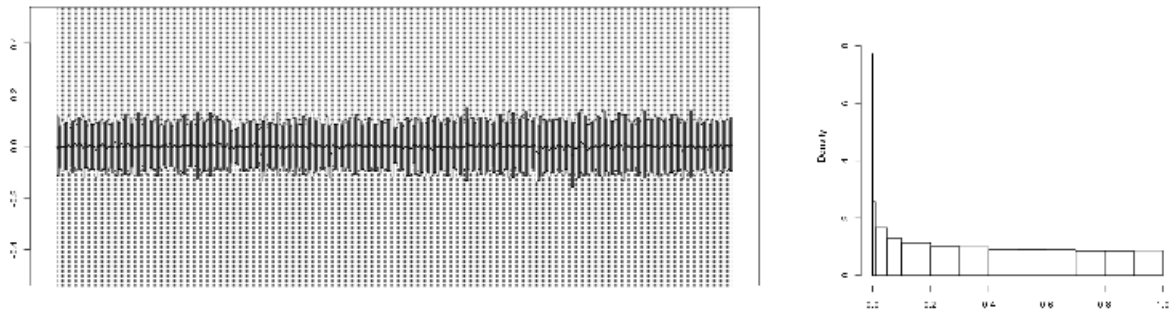


Figure 45: TCGA HT HG-U133A RLE plots and p-value histograms after adjustments by SVA. The data was preprocessed (BG + QN). P-values were computed using Limma.

Unadjusted

Rank	Gene	Chrom	Adj. P	Rank	Gene	Chrom	Adj. P	Rank	Gene	Chrom	Adj. P
1	RPS4Y1	Y	2×10^{-157}	21	KDM6A	X	4×10^{-12}	41	MICAL2	11	3×10^{-03}
2	DDX3Y	Y	5×10^{-139}	22	TMSB4Y	Y	2×10^{-11}	42	GPR88	1	4×10^{-03}
3	XIST	X	3×10^{-113}	23	KDM6A	X	2×10^{-09}	43	ZRSR2	X	4×10^{-03}
4	EIF1AY	Y	5×10^{-111}	24	KDM5C	X	9×10^{-09}	44	CCDC71	3	5×10^{-03}
5	XIST	X	1×10^{-109}	25	RPS4X	X	1×10^{-08}	45	FILIP1L	3	6×10^{-03}
6	KDM5D	Y	5×10^{-94}	26	CD99	X Y	5×10^{-08}	46	GEMIN8	X	7×10^{-03}
7	EIF1AY	Y	7×10^{-93}	27	KDM6A	X	1×10^{-07}	47	CADM4	19	1×10^{-02}
8	DDX3Y	Y	6×10^{-89}	28	CD99	X Y	2×10^{-07}	48	LASS4	19	2×10^{-02}
9	UTY	Y	1×10^{-77}	29	SRY	Y	2×10^{-07}	49	CA5BP	X	2×10^{-02}
10	TTTY15	Y	4×10^{-77}	30	RPS4X	X	1×10^{-06}	50	NA	NA	2×10^{-02}
11	NLGN4Y	Y	1×10^{-64}	31	PRKY	Y	3×10^{-06}	51	TMEM147	19	3×10^{-02}
12	NCRNA00185	Y	4×10^{-64}	32	NA	NA	4×10^{-06}	52	DDX3X	X	3×10^{-02}
13	CYorf15B	Y	3×10^{-41}	33	EIF1AX	X	4×10^{-06}	53	ZFYVE9	1	4×10^{-02}
14	USP9Y	Y	2×10^{-39}	34	TRIM31	6	1×10^{-04}	54	HTATIP2	11	4×10^{-02}
15	HDHD1A	X	2×10^{-23}	35	RPS4X	X	2×10^{-04}	55	NA	NA	4×10^{-02}
16	UTY	Y	1×10^{-22}	36	ZRSR2	X	4×10^{-04}	56	RTCD1	1	5×10^{-02}
17	ZFY	Y	5×10^{-21}	37	USP19	3	1×10^{-03}	57	THY1	11	5×10^{-02}
18	UTY	Y	9×10^{-18}	38	USF2	19	1×10^{-03}	58	HTATIP2	11	5×10^{-02}
19	DDX3X	X	1×10^{-14}	39	USF2	19	2×10^{-03}	59	DYRK1B	19	5×10^{-02}
20	ZFX	X	2×10^{-13}	40	EIF1AX	X	2×10^{-03}	60	TUSC2	3	5×10^{-02}

RUV-2 (SVD), $k = 30$, housekeeping genes

Rank	Gene	Chrom	Adj. P	Rank	Gene	Chrom	Adj. P	Rank	Gene	Chrom	Adj. P
1	RPS4Y1	Y	2×10^{-141}	21	TMSB4Y	Y	5×10^{-19}	41	ZRSR2	X	5×10^{-06}
2	DDX3Y	Y	8×10^{-126}	22	ZFX	X	3×10^{-17}	42	EIF1AX	X	7×10^{-06}
3	XIST	X	8×10^{-105}	23	KDM6A	X	3×10^{-16}	43	USP9X	X	3×10^{-05}
4	XIST	X	1×10^{-101}	24	DDX3X	X	2×10^{-15}	44	EIF1AX	X	5×10^{-05}
5	EIF1AY	Y	7×10^{-99}	25	KDM6A	X	2×10^{-14}	45	TMEM147	19	3×10^{-04}
6	EIF1AY	Y	4×10^{-93}	26	RPS4X	X	2×10^{-13}	46	EIF2S3	X	3×10^{-04}
7	KDM5D	Y	5×10^{-91}	27	KDM5C	X	5×10^{-12}	47	DDX3X	X	3×10^{-04}
8	DDX3Y	Y	1×10^{-84}	28	NA	NA	1×10^{-11}	48	ATP10A	15	3×10^{-04}
9	TTTY15	Y	9×10^{-81}	29	RPS4X	X	5×10^{-11}	49	STIM1	11	5×10^{-04}
10	UTY	Y	6×10^{-79}	30	STS	X	7×10^{-11}	50	USP9X	X	7×10^{-04}
11	NLGN4Y	Y	4×10^{-71}	31	CD99	X Y	1×10^{-10}	51	TMEM204	16	8×10^{-04}
12	NCRNA00185	Y	6×10^{-67}	32	CD99	X Y	4×10^{-10}	52	NA	NA	1×10^{-03}
13	USP9Y	Y	6×10^{-49}	33	PRKY	Y	1×10^{-09}	53	ZFX	X	2×10^{-03}
14	CYorf15B	Y	3×10^{-46}	34	RPS4X	X	1×10^{-09}	54	GPR88	1	2×10^{-03}
15	UTY	Y	2×10^{-33}	35	EIF1AX	X	5×10^{-09}	55	DOCK9	13	3×10^{-03}
16	KDM6A	X	2×10^{-28}	36	ZRSR2	X	3×10^{-07}	56	UBA1	X	3×10^{-03}
17	HDHD1A	X	6×10^{-26}	37	DDX3X	X	4×10^{-07}	57	CA5BP	X	5×10^{-03}
18	DDX3X	X	2×10^{-25}	38	TRIM31	6	7×10^{-07}	58	STS	X	6×10^{-03}
19	UTY	Y	4×10^{-23}	39	SRY	Y	8×10^{-07}	59	GEMIN8	X	6×10^{-03}
20	ZFY	Y	6×10^{-23}	40	STS	X	2×10^{-06}	60	PDE4DIP	1	7×10^{-03}

Table 12: Comparison of gene rankings before and after adjustment (SVD, $k = 30$) using the TCGA HT HG-U133A data. The data has been preprocessed (BG + NM). P-values were calculated using Limma.

G Additional NCI-60 Figures

The following figures are dendrograms of the NCI-60 datasets (HG-U95A and HG-U133A) before and after preprocessing, before and after adjustment using various values of k, and using different sets of negative controls (housekeeping genes and spike-in controls). There are a lot of dendrograms, and it would take a long time to examine all of them. We therefore verbally summarize the dendrograms below. For most readers, the summary should suffice; the dendrograms themselves are included only for the especially curious.

Figures 46, 47 and 48: HG-U95A dataset, preprocessed, adjusted by spike-in controls, $k = 0$ to 8.

The quality improves going from unadjusted to $k = 1$. Increasing k does not lead to further improvement, and the quality soon decreases.

Figure 49: HG-U95A dataset, preprocessed, adjusted by housekeeping genes, $k = 0$ to 2.

Quality decreases going from unadjusted to $k = 1$, and decreases further when k is increased to 2.

Figure 50: HG-U95A dataset, not preprocessed, adjusted by spike-in controls, $k = 0$ to 2.

Quality increases somewhat going from unadjusted to $k = 1$, then decreases somewhat going from $k = 1$ to $k = 2$.

Figure 51: HG-U95A dataset, not preprocessed, adjusted by housekeeping genes, $k = 0$ to 2.

Quality increases going from unadjusted to $k = 1$, and does not change much going from $k = 1$ to $k = 2$.

Figure 52: HG-U133A dataset, preprocessed, adjusted by spike-in controls, $k = 0$ to 2.

Quality stays about the same going from unadjusted to $k = 1$, and decreases somewhat going from $k = 1$ to $k = 2$.

Figure 53: HG-U133A dataset, preprocessed, adjusted by housekeeping genes, $k = 0$ to 2.

Quality decreases going from unadjusted to $k = 1$, and decreases further going from $k = 1$ to $k = 2$.

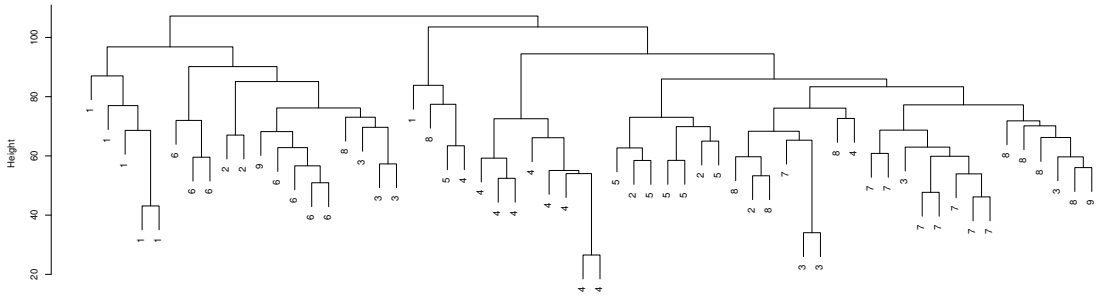
Figure 54: HG-U133A dataset, not preprocessed, adjusted by spike-in controls, $k = 0$ to 2.

Quality increases going from unadjusted to $k = 1$, and increases slightly more going from $k = 1$ to $k = 2$.

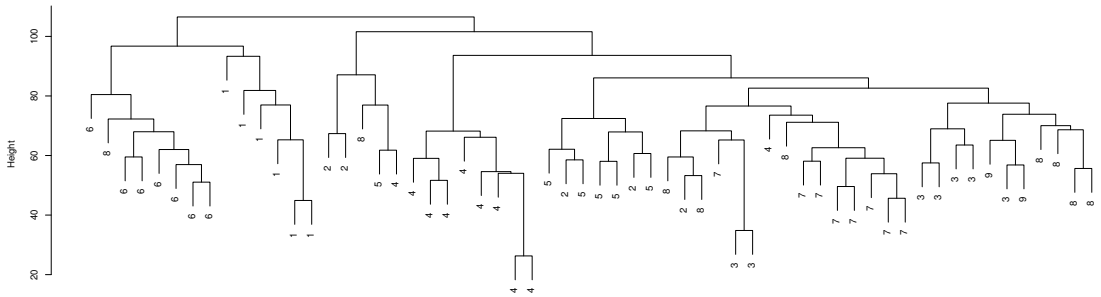
Figure 55: HG-U133A dataset, not preprocessed, adjusted by housekeeping genes, $k = 0$ to 2.

Quality increases going from unadjusted to $k = 1$, and stays about the same going from $k = 1$ to $k = 2$.

Unadjusted



$k = 1$, Affy Spike-in Controls



$k = 2$, Affy Spike-in Controls

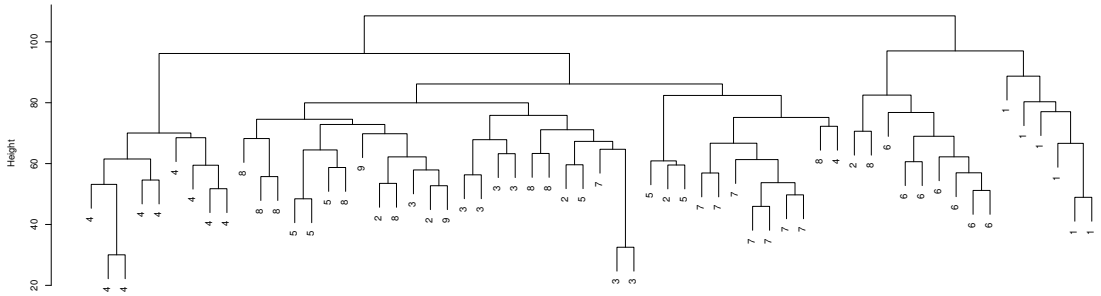
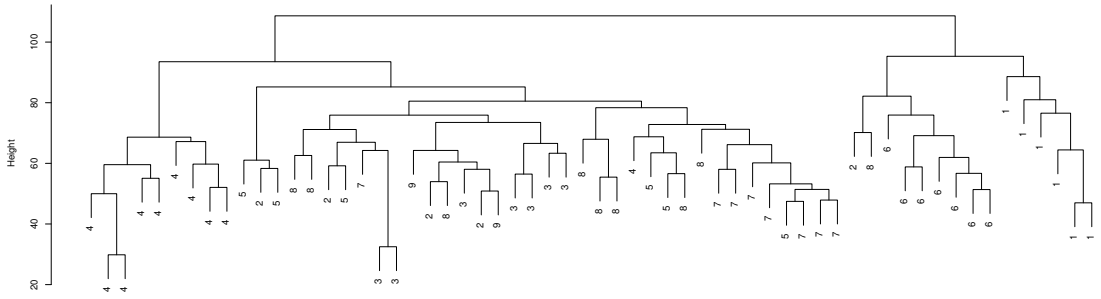


Figure 46: Dendrograms of NCI-60 HG-U95A dataset before and after adjustment. The data was preprocessed (BG + QN). The factors were computed by SVD. Key: 1 – Blood; 2 – Breast; 3 – Ovarian; 4 – Melanoma; 5 – Brain; 6 – Colorectal; 7 – Renal; 8 – Lung; 9 – Prostate.

$k = 3$, Affy Spike-in Controls

$k = 4$, Affy Spike-in Controls



$k = 5$, Affy Spike-in Controls

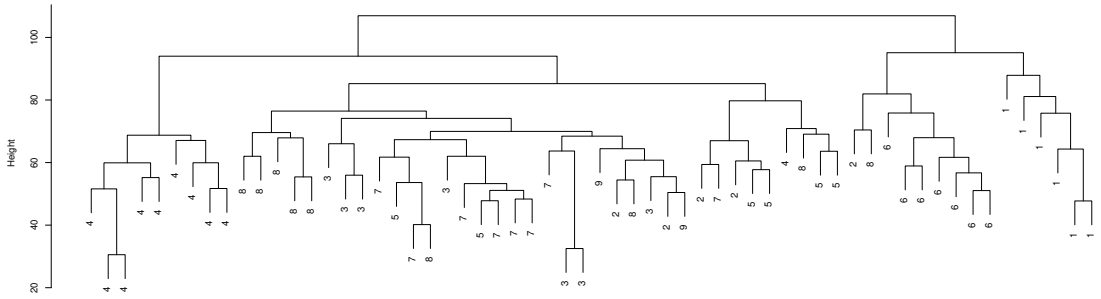
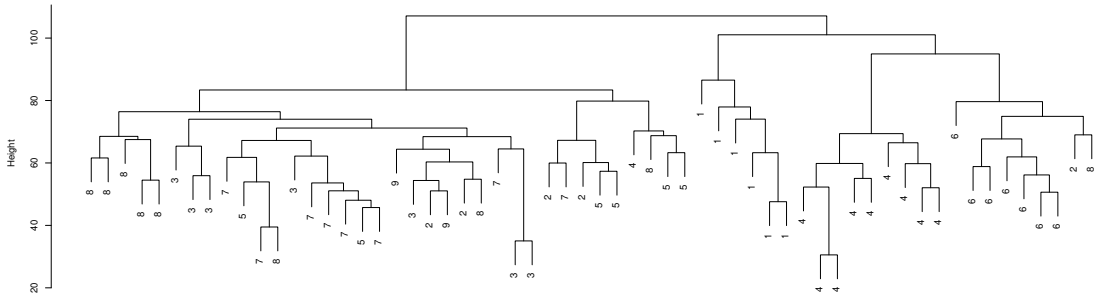
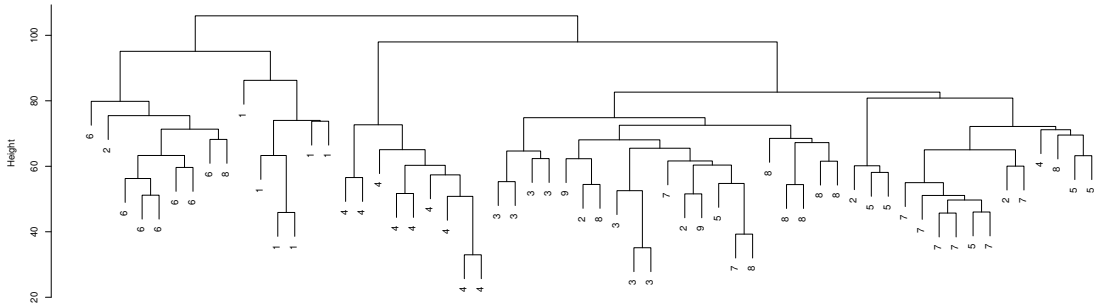


Figure 47: This is a continuation of Figure 46

$k = 6$, Affy Spike-in Controls



$k = 7$, Affy Spike-in Controls



$k = 8$, Affy Spike-in Controls

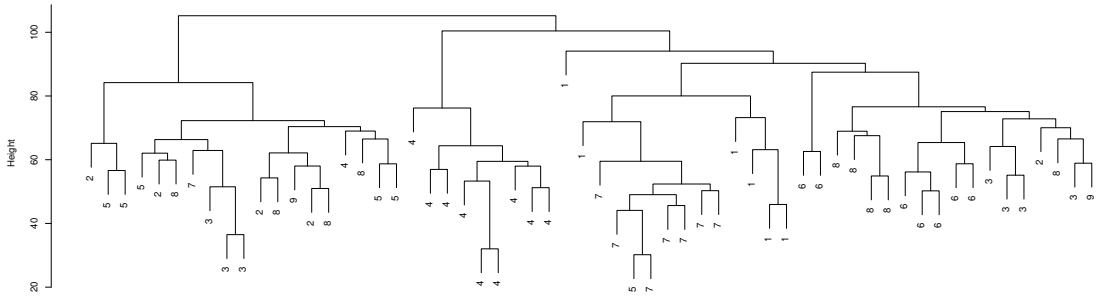
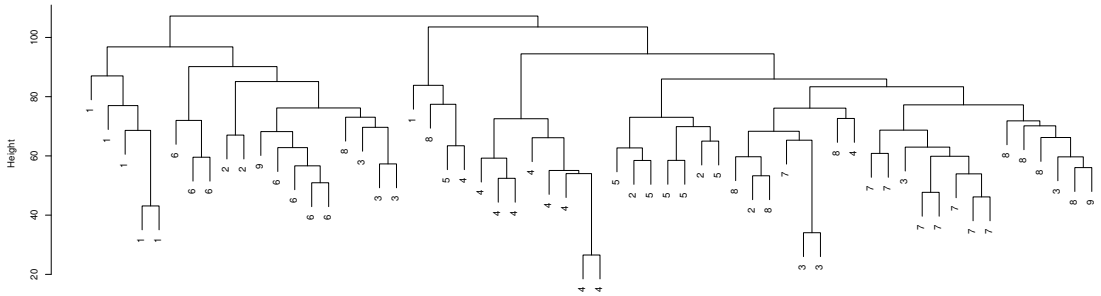
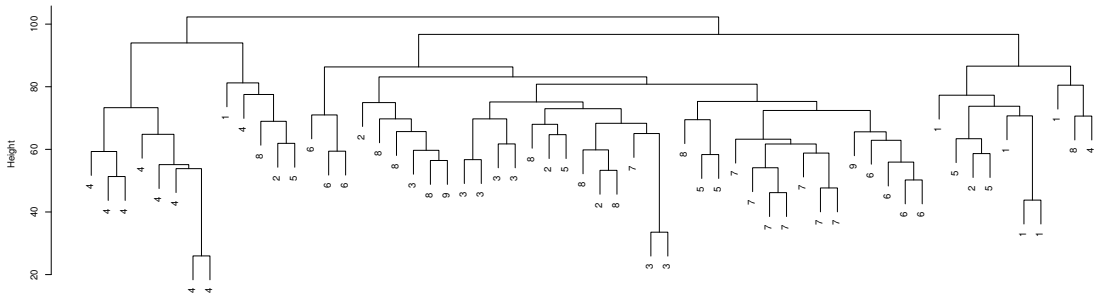


Figure 48: This is a continuation of Figure 46

Unadjusted



$k = 1$, Housekeeping Genes



$k = 2$, Housekeeping Genes

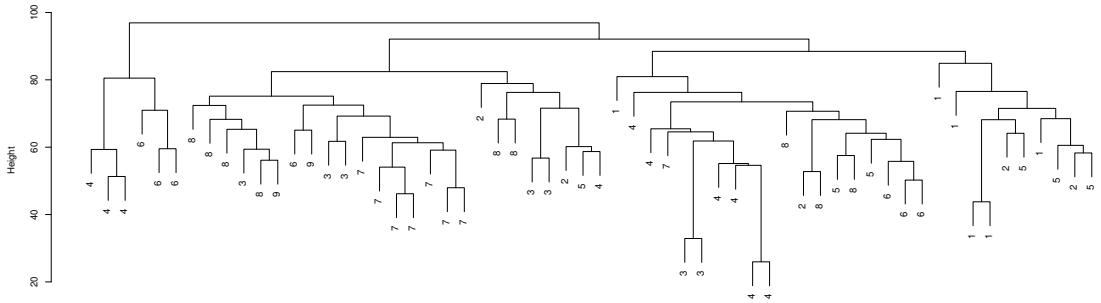
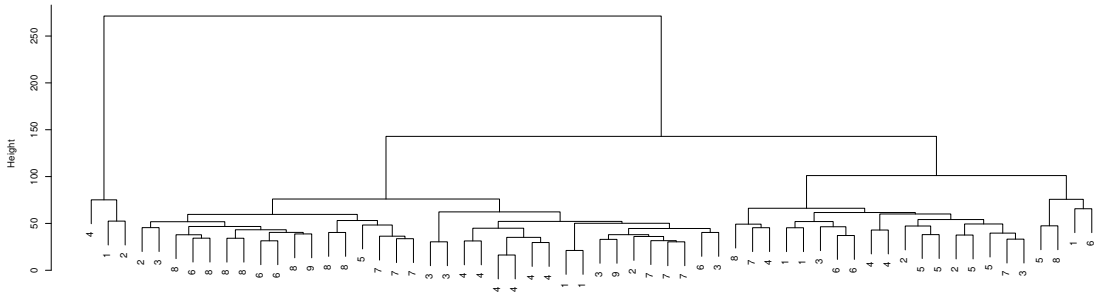
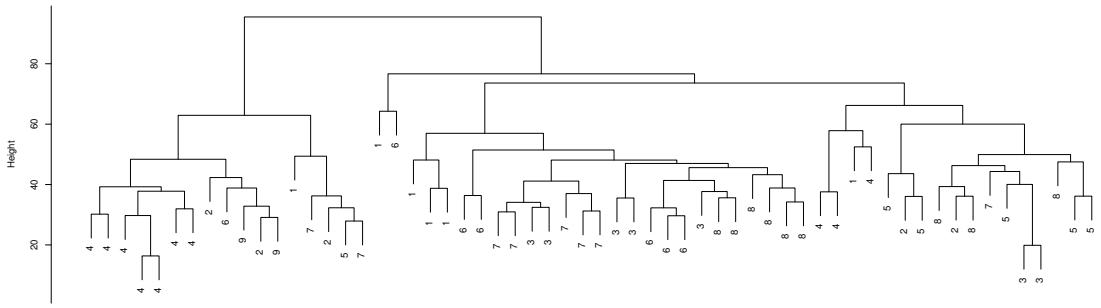


Figure 49: Dendrograms of NCI-60 HG-U95A dataset before and after adjustment. The data was preprocessed (BG + QN). The factors were computed by SVD. Key: 1 – Blood; 2 – Breast; 3 – Ovarian; 4 – Melanoma; 5 – Brain; 6 – Colorectal; 7 – Renal; 8 – Lung; 9 – Prostate.

Unadjusted



$k = 1$, Affy Spike-in Controls



$k = 2$, Affy Spike-in Controls

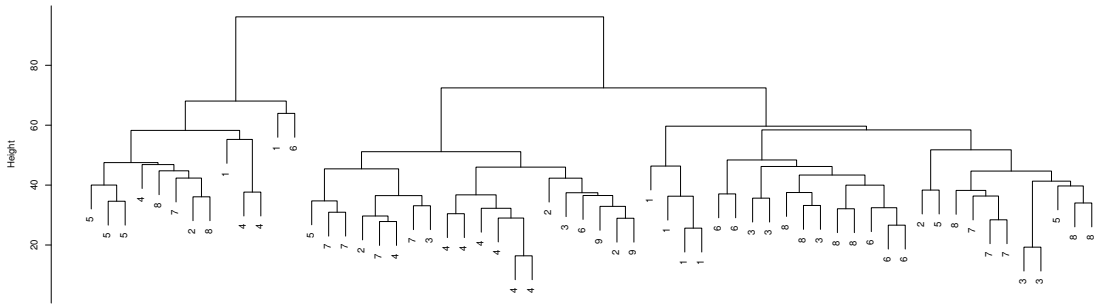
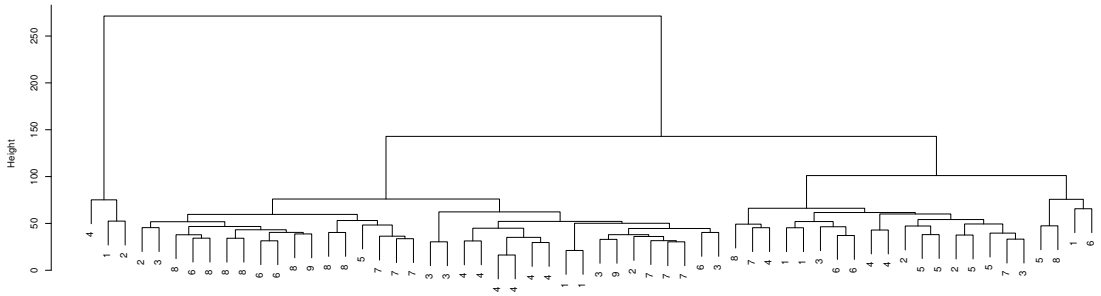
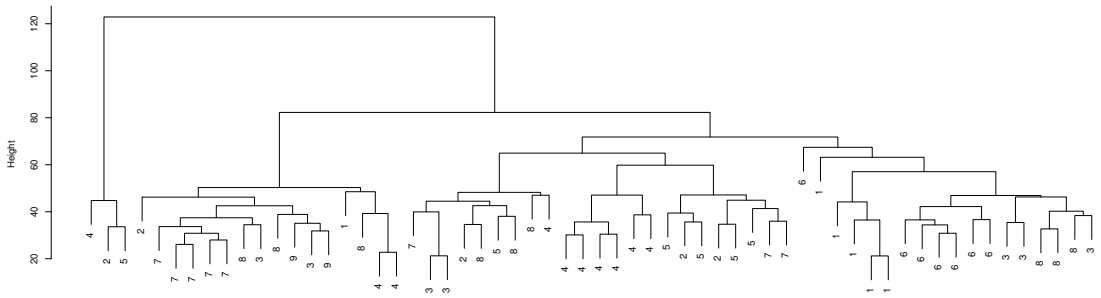


Figure 50: Dendrograms of NCI-60 HG-U95A dataset before and after adjustment. The data was not preprocessed. The factors were computed by SVD. Key: 1 – Blood; 2 – Breast; 3 – Ovarian; 4 – Melanoma; 5 – Brain; 6 – Colorectal; 7 – Renal; 8 – Lung; 9 – Prostate.

Unadjusted



$k = 1$, Housekeeping Genes



$k = 2$, Housekeeping Genes

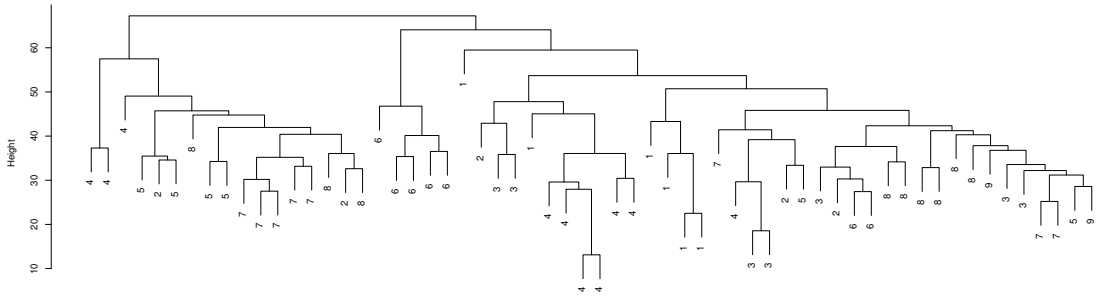
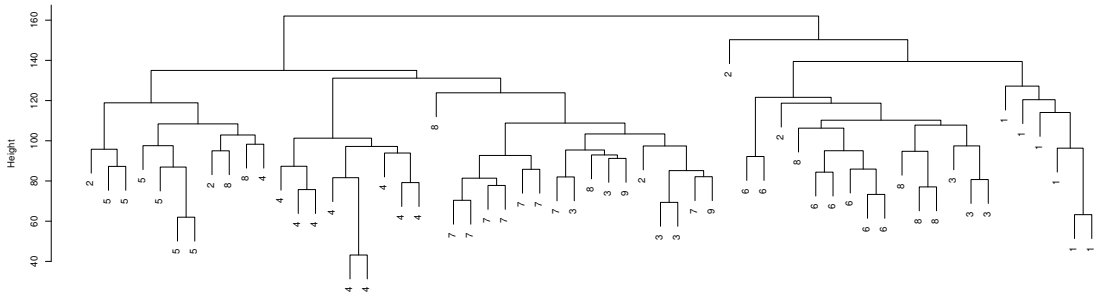
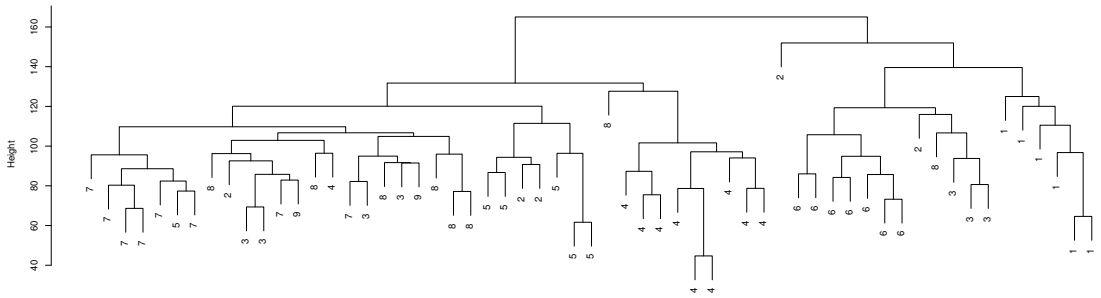


Figure 51: Dendrograms of NCI-60 HG-U95A dataset before and after adjustment. The data was not preprocessed. The factors were computed by SVD. Key: 1 – Blood; 2 – Breast; 3 – Ovarian; 4 – Melanoma; 5 – Brain; 6 – Colorectal; 7 – Renal; 8 – Lung; 9 – Prostate.

Unadjusted



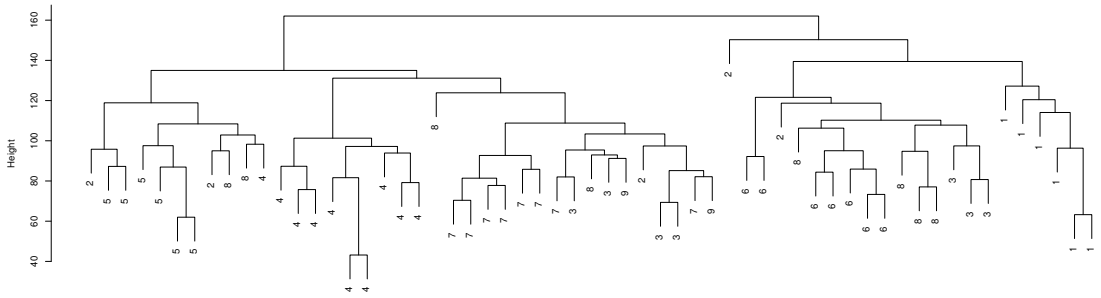
$k = 1$, Affy Spike-in Controls



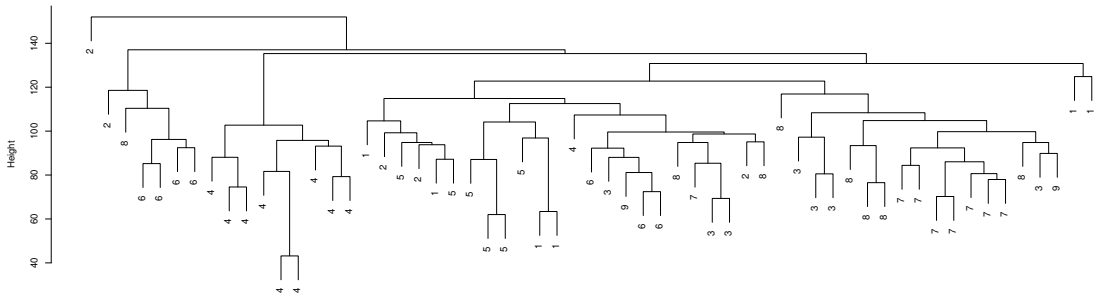
$k = 2$, Affy Spike-in Controls

Figure 52: Dendrograms of NCI-60 HG-U133A dataset before and after adjustment. The data was preprocessed (BG + QN). The factors were computed by SVD. Key: 1 – Blood; 2 – Breast; 3 – Ovarian; 4 – Melanoma; 5 – Brain; 6 – Colorectal; 7 – Renal; 8 – Lung; 9 – Prostate.

Unadjusted



$k = 1$, Housekeeping Genes



$k = 2$, Housekeeping Genes

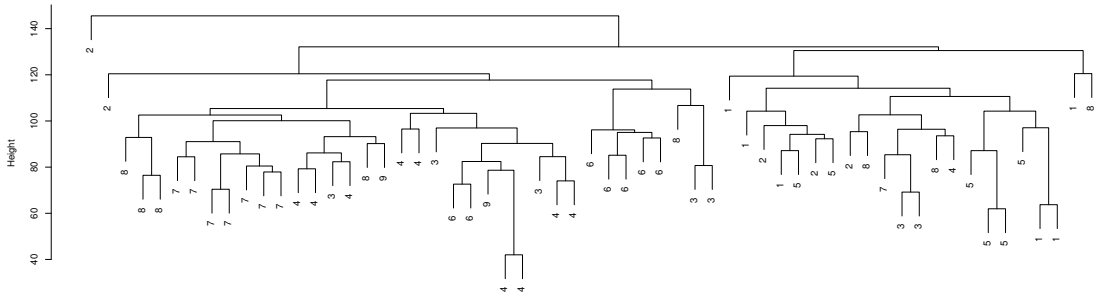
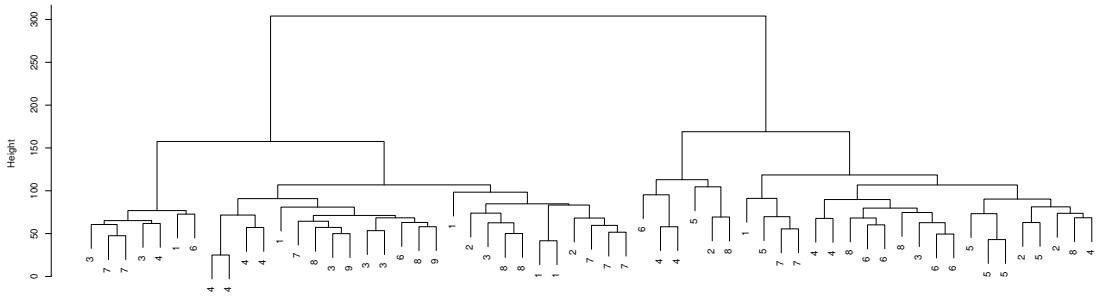
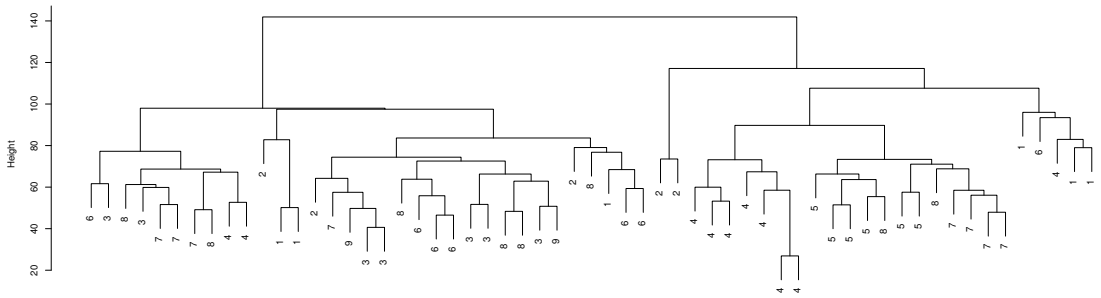


Figure 53: Dendrograms of NCI-60 HG-U133A dataset before and after adjustment. The data was preprocessed (BG + QN). The factors were computed by SVD. Key: 1 – Blood; 2 – Breast; 3 – Ovarian; 4 – Melanoma; 5 – Brain; 6 – Colorectal; 7 – Renal; 8 – Lung; 9 – Prostate.

Unadjusted



$k = 1$, Affy Spike-in Controls



$k = 2$, Affy Spike-in Controls

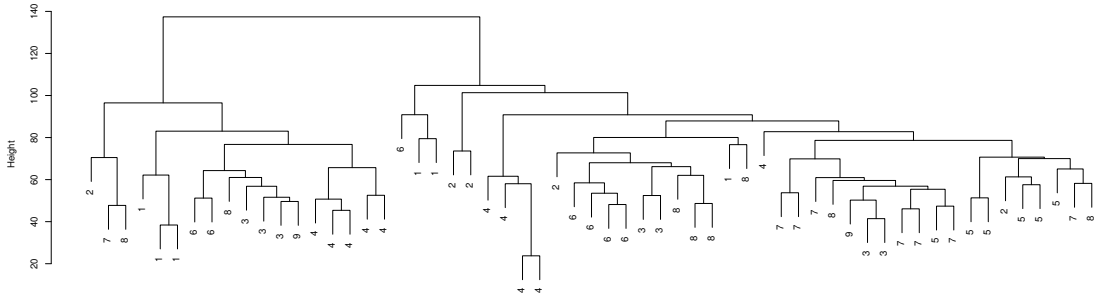
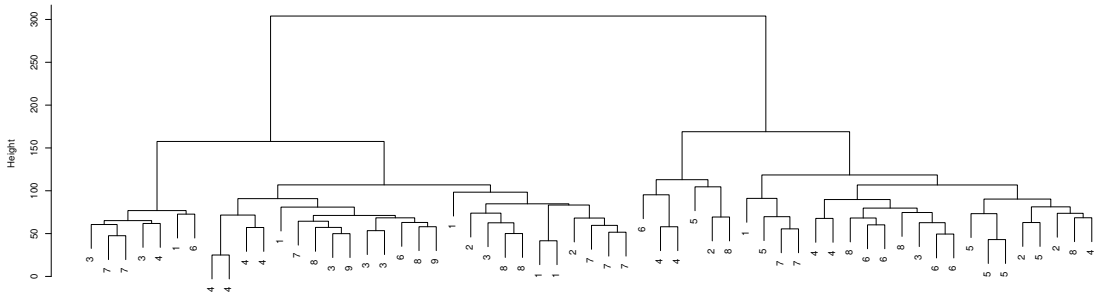
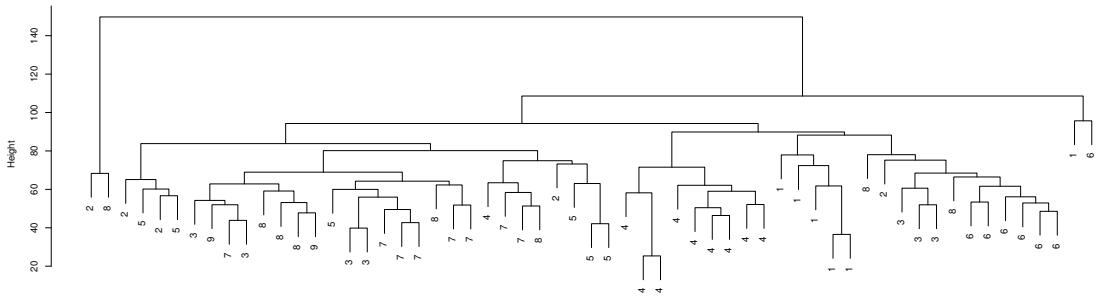


Figure 54: Dendrograms of NCI-60 HG-U133A dataset before and after adjustment. The data was not preprocessed. The factors were computed by SVD. Key: 1 – Blood; 2 – Breast; 3 – Ovarian; 4 – Melanoma; 5 – Brain; 6 – Colorectal; 7 – Renal; 8 – Lung; 9 – Prostate.

Unadjusted



$k = 1$, Housekeeping Genes



$k = 2$, Housekeeping Genes

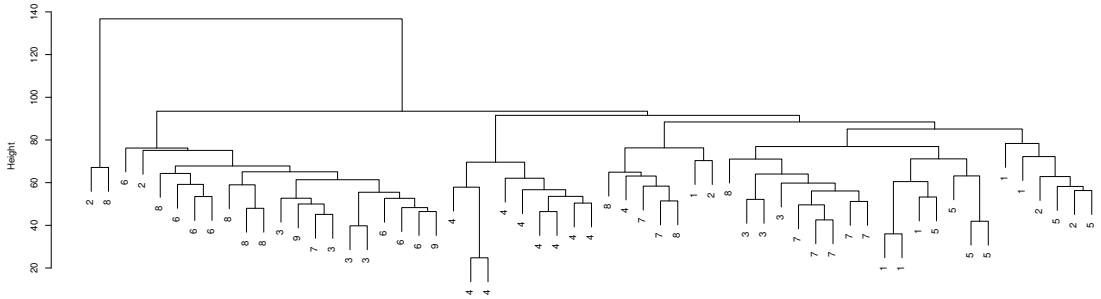


Figure 55: Dendrograms of NCI-60 HG-U133A dataset before and after adjustment. The data was not preprocessed. The factors were computed by SVD. Key: 1 – Blood; 2 – Breast; 3 – Ovarian; 4 – Melanoma; 5 – Brain; 6 – Colorectal; 7 – Renal; 8 – Lung; 9 – Prostate.