Title
Stepwise Normalization of Two-Channel Spotted Microarrays

Authors
Xiao, Yuanyuan
Segal, Mark R
Yang, Yee Hwa

Publication Date
2004-11-05

Stepwise Normalization of Two-Channel Spotted Microarrays

Yuanyuan Xiao[1] Mark R. Segal[2] Yee Hwa Yang[3]

November 5, 2004

Departments of [1]Biopharmaceutical Sciences, [2]Epidemiology and Biostatistics, [3]Medicine,
[2,3]Center for Bioinformatics and Molecular Biostatistics,
University of California, San Francisco, CA 94143, USA.

[3]To whom correspondence should be addressed.

Abstract

Intensities measurements of spotted microarrays embody many undesirable systematic variations. Very commonly, varying amounts and types of such variations are observed in different arrays. Although various normalization methods have been proposed to remove such systematic effects, it has not been well studied how to assess or select the most appropriate method for different arrays and data sets. To address this issue, we present a novel normalization technique, STEPNORM, for data-dependent and adaptive normalization of two-channel spotted microarrays. STEPNORM performs a stepwise interrogation of a range of different normalization models and selects the appropriate method based on formal model selection criteria. In addition, we evaluate the effectiveness of STEPNORM and other commonly used normalization methods utilizing a set of specially constructed splicing arrays.

1 Introduction

DNA microarray technology, fast emerging as one of the most widely used and powerful tools for a suite of genomic applications, can profile gene expression of any organism on a whole genome scale (Lockhart et al. (1996); Cho et al. (2001); DeRisi et al. (1996); Schulze and Downward (2001);Alizadeh et al. (2000)). Two-channel microarrays, which are our focus here, employ a two-color (usually red and green) labeling scheme to measure the *relative* abundance of gene expression in two mRNA populations via competitive hybridization. Like other measuring technologies, two-channel microarray data contain inherent systematic errors arising from variation in labeling, hybridization,

spotting or other non-biological sources (Schena (1999)). Normalization procedures, which adjust microarray data to remove such systematic effects, are therefore crucial for subsequent analysis of either differential expression or gene expression profiling. Throughout, we define normalization as a procedure applied *after* appropriate background adjustment has been made. Discussions on effects of various background correction approaches are described in Yang et al. (2002a).

Several approaches to normalization have been previously proposed. Our emphasis here is on within-array normalization; note, the nature of within-array normalization allows us to normalize between-array locations. What is commonly referred to as between-array normalization actually refers to between-array scale normalization and the reader is referred to Yang and Thorne (2003) and Quackenbush (2002) for further review of more normalization methods. One of the most pronounced biases embodied in relative intensity (fluorescence) measurements results from imbalance in green and red dye incorporation. This imbalance is manifested as the dependence of relative expression ratios on primarily two factors, the fluorescent intensity and spatial heterogeneity. The intensity and spatial biases can be best illustrated using MA-plots, boxplots and spatial plots developed by Yang et al. (2002b). In general, the systematic variation between log-ratios M ($M = \log_2(R/G)$, where R and G are the fluorescent intensity measurements of the red and green channels) and log-intensities A ($A = log_2\sqrt{RG}$) can be removed by linear or nonlinear methods; see Yang et al. (2002b), Finkelstein et al. (2000) and Kepler et al. (2000)) for example. In addition, Fan et al. (2004) develop a procedure based on within-array replications via a semi-linear model. Wang et al. (2002) propose an iterative procedure for estimating normalized coefficients. To remove systematic errors dependent on the spatial layout of spots (S), Yang et al. (2002b) apply the scatter plot smoother `loess` (Cleveland (1979); Cleveland and Devlin (1988)) within each print-tip. Sellers et al. (2003) employ an ANOVA model to remove effects localized within array rows and columns. Wilson et al. (2003) propose using a moving median filter, consisting of a 3×3 block of spots, for the correction of streaky spatial artifacts. Other normalization efforts include the variance-stabilization transformation "vsn" described by Huber et al. (2002) and the quantile approach described in Yang and Thorne (2003). These last two methods are known as single-channel normalization and can be used for between-array normalization.

Despite this multitude of normalization methods there has been very little research in two critical and interrelated areas: (a) the development of formal criteria to assess the performance of a given normalization procedure, and (b) the comparison of competing normalization methods. Here, in addition to tackling these problems, we devise a novel normalization technique, STEPNORM, for two-channel spotted microarrays. This technique aims at applying appropriately calibrated, data-adaptive normalization corrections by stepwise interrogation a range of adjustment models and then invoking formal model selection criteria.

The paper is organized as follows. Section 2 presents details on some pertinent testbed datasets and then describes our proposed approach. The compendium of normalization methods included in a comprehensive comparison study is also provided. Results obtained from use of STEPNORM along with the findings of the comparison study are presented in Section 3. Finally, Section 4 discusses issues, extensions and open questions.

2 Material and Methods

2.1 Data

Experiment A: Quality control arrays (QC)

The UCSF NHLBI Shared Microarray Facility (`http://arrays.ucsf.edu/`) conducted a series of experiments designed to assess print-run quality control. Hybridizations were performed to measure differential gene expression between two RNA samples after passing the BioAnalyzer RNA quality check: K562 erythroleukemia RNA and Stratagene Universal Human Reference (SFUHR), a pool of 10 different cell lines. the BioAnalyzer mRNA check. After hybridization, arrays were scanned using an Axon 4000B laser scanner and images were processed using GenePix 5.0 software. The arrays themselves are fabricated with 70-mer oligonucleotide probes from the Operon Human Genome Oligo version 2, supplemented with some custom-designed 70-mer oligonucleotides. A total of 21,357 probes for 10,801 gene clusters were printed by an arrayer with 4 by 12 print-tips. Each print-tip group consists of 21×23 spots. Similar data can be found in Barczak et al. (2003). Our novel stepwise normalization algorithms were developed with data from this series of experiments and illustrative results are provided in Section 3.1.

Experiment B: Apolipoprotein AI (Apo AI) experiment

In this experiment, gene expression in tissue samples from eight *apo AI* knock-out and eight wild-type mice was studied. The reference sample used in all hybridizations was prepared by pooling cDNA from the eight wildtype mice and was labeled with Cy3. For each of the sixteen mice, cDNA was labeled with Cy5 and co-hybridized with the reference sample to microarrays containing 6384 cDNA probes. Probes were spotted onto the glass slides using a 4×4 print head. For further details the reader is referred to Callow et al. (2000) and Yang et al. (2002b).

Experiment C: Splice arrays (Spt)

Clark et al. (2002) have recently designed a DNA microarray for the analysis of splicing in yeast. Yeast, the simplest eukaryotic organism, has only 250 intron-containing genes and only a handful of these possess multiple introns or are alternatively spliced (Barrass and Beggs (2003)). To discriminate between spliced and unspliced transcripts for intron-containing yeast genes, oligonucleotide probes were designed for splice junctions (SJ), introns (Int) and second exons (Ex) of all intron-containing genes (Figure 1). Splice junctions are found in spliced transcripts whereas introns exist only in unspliced transcripts. The second exon is present in both spliced and unspliced transcripts, and provides a good measure of total transcript level. Clark et al. (2002) quantify the loss of splicing in various mutants by normalizing the change (relative to wildtype) of the splice junction probe signal by the change of the related exon probe signal:

$$\text{SJ index} = M^{SJ} - M^{Ex},$$

where M^{SJ} and M^{Ex} are the log-ratios of the splice junction probe and the corresponding exon

probe [1]. An analogous index is used to quantify intron accumulation:

$$\text{IA index} = M^{Int} - M^{Ex},$$

where M^{Int} is the log-ratio of the intron probe. Using these splice indices, Xiao et al. (2004b) investigate the roles of the chromatin elongation factors, Spt4 and Spt5, in splicing. Spt4 and Spt5 form a complex that regulates the transcription elongation of cellular genes. A collection of 22 hybridizations were performed comparing 5 different splice mutants, *ceg1-250*, *spt4*Δ, *spt5-194*, *spt5-242* and *spt5-4*, to wildtype. Details of this experiment can be found in Xiao et al. (2004b). The unusual design of the splicing arrays, whereby every splice junction probe is paired with a constitutively expressed exon probe, enables comparison of competing normalization methods, including our STEPNORM procedure. The assumptions underpinning these comparisons are described in Section 2.3 and results given in Section 3.2.

2.2 Stepwise Normalization

It is commonly observed that different arrays exhibit varying amounts and types of bias. This is especially evident from the spectrum of spatial trends seen across array studies. Applying the same normalization method to all arrays may not adequately address the complexities of array biases. Seemingly, a more appropriate scheme would be to adaptively capture the bias characteristics of each array by quantitatively assessing the adequacy and extent of corresponding corrections/models. A helpful perspective is provided by considering the variance-bias trade-off as a function of model complexity; see Hastie et al. (2001). Typically, the more complex a normalization model the lower the bias, but the higher the variance. The goal of STEPNORM is to achieve a good balance between variance and bias. We believe it is the first microarray normalization procedure to adaptively pursue this objective and, simultaneously, objectively measure performance.

Figure 2 illustrates the STEPNORM methodology using a typical QC array in Experiment A. Here, it consists of four steps – within each step a particular bias is targeted for correction. A prescribed range of competing models, of varying complexity, is compared for effecting these within step corrections. More generally, both the number of steps and the bias correction models can be flexibly chosen, depending on experimental specifics and objectives. Algorithmic details of STEPNORM are provded in Appendix.

The intensity bias, A, is usually the major source of bias (Yang et al. (2002b); Sellers et al. (2003)) and is therefore examined first. Correction models to be tested in this step include: (i) the "null" model, which doesn't fit any parameters and represents the scenario that the A bias is not sufficient to warrant any correction; (ii) the global median shift model; (iii) `rlm`, the robust linear regression model (Finkelstein et al. (2000)); and (iv) `loess`, the locally weighted scatter plot smoother (Yang et al. (2002b)). These four models are of increasing complexity as quantified by their respective degrees of freedom (df). Note that the approximate df for `loess` are obtained from the trace of the attendant smoother matrix linking observed and fitted values (Hastie and Tibshirani (1990)). The

[1] $M^{SJ} = \log(mut_{SJ}/wt_{SJ})$, where mut_{SJ} and wt_{SJ} represent fluorescent intensities of the splice junction probe in a mutant and wildtype respectively. Note here the simplification of similar notations from Xiao et al. (2004b) and Clark et al. (2002)

application of the above correction models can be generalized by the following form:

$$M_i = f(A) + \epsilon_i, \tag{1}$$

where M_i is the observed log-ratio and ϵ_i is the error term that is typically referred to as the normalized log-ratio. The term $f(A)$ is the normalization factor, which is a function of log-intensities (A) and the formalism of which distinguishes the differing models.

After correction of the A bias, normalized log-ratios are subject to further adjustments to address spatial biases. We decompose spatial bias into three components: print-tip (PT), plate (PL), and residual two-dimensional spatial effects. The reason for adjusting for PT first is that the PT effect is usually the dominant spatial bias (Sellers et al. (2003)). Furthermore, the number of print-tip-groups on an array is usually smaller than the number of (384-well) plates used, and so PT can be effected more parsimoniously. The correction models employed in the PT and PL steps are similar to those in the A step, except that they are fitted within each print-tip group or well-plate group.

We base (within step) selection of a correction model on the Bayesian Information Criterion (BIC; Schwartz (1979)). Although, again, alternate approaches can be entertained. BIC is (asymptotically) consistent as a selection criterion, whereas the familiar Akaike Information Criterion (AIC; Akaike (1983)) is not. However, it is difficult to relate this to finite sample performance. Because of the size of their respective penalty terms ($\log(N) \cdot p$ (BIC) vs $2 \cdot p$ (AIC) for models with p df) BIC will penalize complex models more heavily. A multitude of other information criteria exist (e.g. RIC, Foster and George (1994); CIC, Tibshirani and Knight (1999)), but perhaps the simplest and most widely used approach for model selection is cross validation (CV). We compare BIC and CV in the Discussion and note similar selections. However, in the present microarray setting BIC enjoys a substantial computational advantage.

Under a Gaussian model for the errors, BIC is simply calculated as:

$$\begin{aligned} BIC &= -2 \cdot log(\hat{L}) + p \cdot log(N), & (2) \\ &= N \cdot log(\sum_{i=1}^{N} \epsilon_i^2/N) + p \cdot log(N), & (3) \end{aligned}$$

where $\hat{L}$ is the maximum likelihood of the model and is a function of the residual sum of squares obtained from Equation 1 (for the correction of the A bias). In the STEPNORM framework, for each step of the normalization, the correction model with the lowest BIC value is the preferred model. We illustrate the utility of STEPNORM using BIC as a mode selection criterion on a QC array in Section 3.1

2.3 Comparison study of normalization methods

To compare the effectiveness of competing normalization procedures we need to address the issues of bias and variance simultaneously. In practice, it is relatively easy to show whether a new method decreases the variance of the normalized log-ratios (residuals of the normalized model). For example, in comparison of the effects of various within-array location and scale normalization methods,

Yang et al. (2002b) show that both the intensity dependent and within-print-tip-group location normalization methods reduce the spread of the log-ratios compared to a global normalization. The more challenging task is to establish whether this reduction in variance comes at a cost of attenuating absolute and relative intensity values. To properly address these concerns it is essential to have datasets with known levels of absolute and relative differential gene expression, as well as a reasonable amount of replication. Examples of such datasets using spiked-in genes are available for Affymetrix technology (`http://www.affymetrix.com/index.affx`) and some initial analyses are available at `http://www.stat.berkeley.edu/users/terry/zarry/affy/affy_index.html`. The ApoAI dataset in Experiment B contains RT-PCR verifications for the differentially expressed (DE) genes, which we use for a quick evaluation of various normalization methods in Section 3.2. However, the ApoAI dataset and, more generally, data sets for which only selective verification is performed, are limited with respect to evaluating normalizing methods since they do not provide genome-wide measures of accuracy. As explained next, it is the unusual design of splice arrays that provides a unique opportunity to assess both variance and bias on an individual gene basis which, in turn, enables us to formally compare normalization methods.

As shown in Figure 1, for each intron-containing gene there are three oligonucleotide probes targeting the corresponding splice junction, intron and exon. These are printed on the glass slide in quadruplicates. Clark et al. (2002) employ a probe-specific adjustment method to remove non-biological biases for these arrays. This probe-specific normalization is established based on a few reasonable assumptions and observations: (i) exons, and their corresponding splice junctions, are spotted closely on the array and have very similar expression levels, and (ii) exon log-ratios are not expected to exhibit any biological variation. Therefore, any systematic deviations from zero in the exon log-ratios are indicative of artefactual variations present in both the exons and their corresponding splice junction ratios. Thus, we can treat the exon probes as probe-specific normalization factors for their corresponding splice junction probes.

It is this existence of target "true normalization values" that enables assessment of both variance and bias and thereby allows comparison of competing normalization methods in the context of a real and complex dataset. To put it simply, we treat the exon log-ratio as the "ground truth" measurement of the systematic bias in the corresponding splice junction log-ratio. The six commonly used normalization models in Table 1, as well as our STEPNORM procedure, attempt to estimate this "ground truth" . Since these values are known the splice array platform provides a testbed for evaluating both variance (precision) and bias (accuracy). This contrasts with, for example, replicated studies which only furnish estimates of precision. It is important to note that these methods implicitly assume that relatively few genes are differentially expressed, and that there is no systematic relationship between differential gene expression and intensity or location of the spots.

A simple and widely used comparison criterion is mean square errors (MSE). To begin, we denote the expression log-ratio for the splice junction probe of the kth replicate ($k = 1, \ldots, 4$) of the jth gene ($j = 1, \ldots, 254$) as M_{jk}^{SJ}. Let c_{jk}^{SJ} be the normalization factor estimating the systematic variation in M_{jk}^{SJ}, obtained by subtracting the estimated (for each of the seven models) the normalized log-ratio (denoted M_{jk}^{*SJ}) from the corresponding unnormalized log-ratio: $c_{jk}^{SJ} = M_{jk}^{SJ} - M_{jk}^{*SJ}$. Then

the MSE of c_j^{SJ} $(= 1/4 \sum_{k=1}^{4} c_{jk}^{SJ})$ and associated bias - variance components are given by

$$\hat{MSE}(c_j^{SJ}) = \frac{1}{4}\sum_{k=1}^{4}(c_{jk}^{SJ} - M_j^{Ex})^2 \quad (4)$$

$$\hat{Var}(c_j^{SJ}) = \frac{1}{4}\sum_{k=1}^{4}(c_{jk}^{SJ} - c_j^{SJ})^2 \quad (5)$$

$$\hat{Bias}^2(c_j^{SJ}) = (c_j^{SJ} - M_j^{Ex})^2, \quad (6)$$

where M_j^{Ex}, the average of the four exon log-ratios for the jth gene, is the true normalization factor as explained above. In view of the familiar decomposition $\hat{MSE}(c_j^{SJ}) = \hat{Bias}^2(c_j^{SJ}) + \hat{Var}(c_j^{SJ})$, we simply compute bias by subtraction. Having computed probe-wise estimates of MSE, variance and bias, we summarize by pooling across genes and arrays using medians to derive estimates for each normalization model. Normalization models with smaller MSE are desirable.

3 Results

In any microarray experiment it is important to adjust for inherent biases, recognizing assumptions and limitations of the adjustment procedures. Additionally, it is important to check that such normalization reduces systematic errors. Diagnostic plots, such as $MA-$plots, spatial plots, density and boxplots can qualitatively inform as to the level of adjustment needed and similarly provide qualitative guidance as to whether biases have been successfully removed. For example, investigators may decide whether to perform within-array scale normalization for a dataset by examining boxplots of log-ratios stratified by different print-tip-groups. However, as noted previously, there is a need to balance variance and bias and to perform correspondingly appropriate degrees of adjustment. Ensuring this requires a more quantitative, algorithmic approach. Our proposed stepwise normalization method provides such an approach. Next, we illustrate its utility using an array from the QC experiment with the aid of various diagnostic plots and subsequently contrast its performance with existing correction methods.

3.1 Application of stepwise normalization on Experiment A

We choose a typical array from the QC experiment to demonstrate the performance of STEPNORM. For the adjustment of the intensity A bias, we and others (Yang et al. (2002b); Wilson et al. (2003); Kepler et al. (2000)) have observed that almost all microarray data exhibit varying degrees of trend between M and A. The intensity-dependent bias is noticeable in the $MA-$plot of the QC array and spots exhibit apparent curvature (Figure 3(a)). To find the most appropriate model for the QC array, we compare three models, median shift, `rlm` and `loess`. Table 2 indicates that, among the three candidate models, `loess` has the lowest BIC value and therefore is chosen as the best model for correction by STEPNORM. The number of degrees of freedom spent in the `loess` fitting is approximately 5.5.

We proceed next to the removal of the PT bias. Figure 3(a) reveals that before any normalization is carried out intensity trends within print-tips show nonlinear tendencies. Yet, such nonlinear trends largely disappear after the first step A-bias correction and `loess` fits within each print-tip (Figure 3(b)) reveal essentially linear departures from the zero line. In addition, the spreads of log-ratios within each print-tip and well plate are greatly decreased by the first step correction (not shown). Appropriately then, STEPNORM chooses the `rlm` model for removal of PT bias. For both of the next two steps, PL and Spatial 2D biases, STEPNORM selects the null model, reflecting absence of systematic variation. These selections are affirmed by Figure 3 (d) which displays boxplots of log-ratios stratified by the well plates after the correction of PT bias and the spatial plot comparing the 2D spatial trend before and after stepwise normalization (result not shown).

Table 2 lists the accuracy of model fits ($-2log(\hat{L})$), degrees of freedom, and corresponding BIC values for the tested models along the normalization steps. As more and more normalizations are applied, we see the familiar improvement in (resubstitution-based measures of) fit, here reflected by decreasing values of $-2log(\hat{L})$. However, this decrease is only modest after the first two normalization steps (A and PT bias corrections). Correspondingly, BIC values increase for the last two steps (PL and Spatial 2D), confirming the slightly improved fit of these latter models is not sufficient to offset the increase in model complexity.

3.2 Comparison of normalization methods

Experiment B: Apo AI experiment

We applied the various normalization methods listed in Table 1 to the ApoAI dataset provided in Callow et al. (2000), where a selection of genes were verified by RT-PCR. A rough way to evaluate the existence of bias is to examine the pair-wise correlation of log-ratios between replicate arrays. Given that little difference is expected between wildtype and reference samples (the latter being a pool of the former), two independent replicates should have zero correlation between their log-ratios. Large positive correlation indicates undesirable systematic experimental effects between replicates. We see that the positive correlation observed before normalization (Figure 4(a)) is reduced after STEPNORM normalization (Figure 4(b)); see also Table 3. To assess whether this reduction in systematic bias comes at the cost of attenuating DE signals, we further compared the pair-wise correlation of log-ratios from the eight DE genes between control and knockout mice among replicate arrays. These eight genes had been verified by RT-PCR. The correlation coefficients were adjusted by subtracting background (excluding the eight DE genes) correlation. The results shown in Table 3 indicate that STEPNORM performs better than other normalization methods in retaining DE signals while reducing bias.

Experiment C: Spt experiment

Figure 5 summarizes comparison results of the commonly applied normalization methods listed in Table 1, as applied to the splice array data. The X-axis shows the MSE values. The distance of points to the vertical zero line measures how well each normalization model performs, with smaller being better. The Y-axis lists the different normalization methods. Each point on the plot represents the MSE from an individual splice array with mutants being color coded. We stratify the experiments into three categories reflecting *a priori* biological knowledge regarding the

extent of differential gene expression. *ceg1-250* (shown in green), a mutant causing known splice defects, is expected to have a large number of DE genes, whereas *wt* arrays (shown in red) are a series of wildtype self-self hybridizations where no expression changes are anticipated. The rest of the mutants are expected to have only a small number of DE genes. Based on this plot, the most simplistic model, the global median shift model, performs the worst; more complex models, including STEPNORM, perform rather similarly to each other. We discuss the interpretation of this observation in more detail in Section 4.

We next examine the decomposition of MSE into squared bias and variance. Note that variance refers to the variance of the fitted values of various normalization models; more complex models tend to give rise to a higher variance of model estimates and this translates to a decreased variance in normalized log-ratios. Figure 6 plots estimated variance *vs.* squared bias for each normalization model for the mutant *spt4*Δ and the wildtype self-self hybridizations. These two experiments are chosen because they exhibit minimal/no differential expression, conforming best to the assumptions behind the comparison study. Figure 6 shows that in both experiments bias is the dominant component of MSE. The relationship between bias and variance is well illustrated by the three tRMA models. tRMA 1, 2 and 3 employ spatial median filters with respective window sizes of 3×3, 7×7 and 15×15 and, therefore, decreasing complexities. As expected for both *spt4*Δ and wildtype, tRMA 1 has the largest variance and smallest bias. However, for wildtype, tRMA 2 and 3 show sizable decreases in variance without inflating bias. That the improvement is greatest for tRMA3, the simplest model, is an indication that more elaborate normalization methods, such as tRMA 1, are over-fitting. This illustrates the importance of adapting normalization models according to array particulars.

We next use a *spt4*Δ array to illustrate changes in bias and variance during the stepwise normalization process implemented by STEPNORM. In Figure 7 black and purple points respectively indicate the variance and squared bias values of competing models within each step, with asterisks designating the models chosen by STEPNORM. As more complex normalizations are applied there is the anticipated increase in variance and decrease in bias. Note the big gains realized by *loess* intensity adjustment.

4 Discussion

Effective normalization is crucial for microarray-based research since it directly impacts the outcome of all downstream data analyses. In this paper we have presented a new normalization procedure, STEPNORM, which integrates a number of adjustment techniques into a common framework and provides tools for selecting amongst them as well as measuring overall performance. STEPNORM adjustment is applied to each individual array in an experiment facilitating array specific correction.

STEPNORM is seemingly unique in employing model selection criteria to guard against under- or over-fitting. Intensity-dependent (A) bias in log-ratios is usually the most common and dominant bias in microarray spot measurements. Typically, such bias exhibits as a nonlinear trend between M and A; the curvature can be estimated using a suitable robust scatter plot smoother, such as the `loess` procedure. Span determination is a key component of such smoothing procedures. To determin a suitable range of spans, we conducted a series of cross-validation based evaluations of

competing spans (using `loess` smoothing). In Figure 8, we plotted spans in the range of 0.01 to 1 and their corresponding prediction mean square error, obtained by 5-fold cross validation, using two slides from Experiment A. We see that there is no substantial difference in prediction error within a wide range of spans, containing our recommended default (0.4). Similar results were observed not only in other slides in Experiment A, but also in other differently dimensioned experiments (for example, the swirl data in Dudoit and Yang (2002) and the splice data in Experiment B). In conclusion, we found the the default span 0.4 performs generally well. Exploration of the impact of varying spans is computationally intensive, however it is readily done within the STEPNORM framework. In addition, we have observed that the nonlinear A bias is usually a whole-array phenomenon and doesn't localize within spots related to a specific print-tip or plate. Therefore, the current common practice of applying `loess` within each print-tip (LPT) to remove the A and S biases simultaneously likely represents over-fitting. For arrays like the QC experiment that have 48 print-tips, LPT spends about $5.5 \times 48 = 268$ degrees of freedom (for a span of 0.4). On the other hand, the adjustment selected by STEPNORM for the exemplary QC array in Section 3.1 applies `loess` for the removal of whole-array A bias, and then employs a simpler `rlm` adjustment within print-tips to remove the PT bias. This costs a total of about $5.5 + 2 \times 48 = 101.5$ degrees of freedom, far fewer than LPT.

The choice of model selection criterion is an important part of the STEPNORM procedure. We have employed BIC on account of its good theoretic properties, familiarity and computational ease. The latter concern is especially pertinent because of the high dimensional data furnished by microarray experiments. To further examine the performance of BIC we compared it to cross-validation (CV) using data from the exemplary QC array studied previously. Figure 9 depicts model selection results using 5-fold CV and BIC. The CV prediction errors for each model are shown as colored points (along with associated standard error bars) in the upper panel, and the corresponding BIC values are given in the lower panel. Purple and green curves, which connect models chosen by CV and BIC respectively in each step, have highly similar profiles. In both cases, the estimated prediction error decreases sharply following the adjustment of the A bias; and the trend becomes flattened after the correction of the PT bias.

Limitations of forward stepwise approaches have long been recognized. The interrelated concerns surround the greediness of the algorithm, instability of estimates, and inferential and model selection biases; see for example Efron et al. (2004) and Miller (2002). Here, however, these issues are mitigated by the following considerations. Search greediness is greatly reduced by our prescribing the sequence in which biases are corrected – we benefit from the fact that the ordering of magnitudes of differing bias sources is generally known. Further, bias correction is strictly an adjustment procedure. We are not interested in making inferential statements about fitted coefficients and so surrounding concerns are moot. Of course, model selection remains an issue. The above cross-validation results, at least here, provide some assurance that BIC provides a computationally feasible and reasonable means for determining number of steps.

We compared several commonly applied normalization models in the context of bias and variance using a set of specially constructed splicing array data. Figure 5 indicates that most of the normalization methods under comparison including STEPNORM performed equally competitively. This small difference may be a result of the properties of the data sets; that is, our current data are not varied enough to better discriminate the various models and illustrate the effectiveness of STEP-

NORM. Future collection of more experiments from the Spt study with similar array layout may provide clearer comparison results. Another interesting observation is the clear differences between the three categories of mutants. This further illustrates the importance of assumption behind the various models. In situations where we expect a large number of biological changes, commonly used normalization methods based on all genes do not perform well. Therefore, care needs to be taken at the design stages of the array to ensure the inclusion of sufficient control probes.

We have applied STEPNORM to four different experiments with 41 two-channel spotted arrays (Xiao et al. (2004a)) and in each case obtained good results as assessed via diagnostic plots and/or criteria based on variance and bias when available. We used the four-step procedure as outlined in Figure 2. However, the STEPNORM framework is flexible and can be easily extended or modified. The user can include new normalization models as long as measures of model fit and complexity can be obtained; these being needed for the calculation of BIC. Differing selection criteria may require alternate arguments. Furthermore, the user can modify the number and sequence of adjustment steps. However, we believe our default specifications here generally reflect the order and maximal extent of microarray biases.

The STEPNORM procedure currently addresses within-array normalization issues. This is equivalent to a between-array location normalization and is an essential step in two-channel microarray normalization. A natural extension will be incorporating between-array scale or distributional normalization into the STEPNORM framework. Finally, the STEPNORM procedure is implemented as an `R` package (Ihaka and Gentleman (1996)) `stepNorm`, which may be downloaded from the Bioconductor website (`http://www.bioconductor.org/`).

Appendix

Below is an outline of the STEPNORM algorithm as depicted in Figure 2.

For microarray data $\mathcal{D}_0$, let M_i $(i = 1, \ldots, N)$ be the log-ratio for the ith gene. Apply the following steps.

Step A

1. Apply normalization model f_j $(j = 1, \ldots, 3)$ to $\mathcal{D}_0$ yielding normalized log-ratios M'_{ij}. Here, in this step competing normalization models include {median shift, *rlm*, *loess*}.

2. For the jth model with $p$df, compute the BIC_j value as follows (see Equation 3)

$$BIC_j = N \cdot log(\sum_{i=1}^{N} M'^2_{ij}/N) + p \cdot log(N)$$

   Compute the BIC_0 value for the null model as follows:

$$BIC_0 = N \cdot log(\sum_{i=1}^{N} M_i^2/N) + p \cdot log(N)$$

3. Obtain the best normlization model f_{j^*} by choosing $j^* = argmin_j\{BIC_j, j = 1, \ldots, 3\}$

4. Compare $\{BIC_{j^*}, BIC_0\}$, if $BIC_{j^*} < BIC_0$, normalize $\mathcal{D}_0$ using the chosen normalization model f_{j^*} to obtain normalized data $\mathcal{D}_1$; if $BIC_{j^*} > BIC_0$, no normalization is necessary, and $\mathcal{D}_0$ becomes $\mathcal{D}_1$.

Step PT

1. Apply normalization model f_j $(j = 1, \ldots, 3)$ *within each print-tip-group* to $\mathcal{D}_1$ yielding normalized log-ratios M'_{ij}. Here, in this step competing normalization models include {median shift, *rlm*, *loess*}.

2. Apply steps 2-3) in **Step A** analogously.

3. Compare $\{BIC_{j^*}, BIC_0\}$, if $BIC_{j^*} < BIC_0$, normalize $\mathcal{D}_1$ using the chosen normalization model f_{j^*} to obtain normalized data $\mathcal{D}_2$; if $BIC_{j^*} > BIC_0$, no normalization is necessary, and $\mathcal{D}_1$ becomes $\mathcal{D}_2$.

Step PL and **Step 2D Spatial** can be analogously applied to yield the normalized data $\mathcal{D}^*$ by STEPNORM.

References

H. Akaike. Information measures and model selection. In *Proceedings of the 44th Session of the International Statistical Institute*, 1983.

A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, et al. Different types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403:503–511, 2000.

A. Barczak, M. W. Rodriguez, K. Hanspers, L. L. Koth, Y. C. Tai, B. M. Bolstad, T. P. Speed, and D. J. Erle. Spotted long oligonucleotide arrays for human gene expression analysis. *Genome Research*, 13:1775–1785, 2003.

J. D. Barrass and J. D. Beggs. Splicing goes global. *Trends in Genetics*, 19:295–298, 2003.

L. Breiman, J. H. Friedman, R. Olshen, and C. J. Stone. *Classification and regression trees*. The Wadsworth statistics/probability series. Wadsworth International Group, 1984.

M. J. Callow, S. Dudoit, E. L. Gong, T. P. Speed, and E. M. Rubin. Microarray expression profiling identifies genes with altered expression in HDL deficient mice. *Genome Research*, 10 (12):2022–2029, 2000.

R. J. Cho, M. Huang, M. J. Campbell, H. Dong, L. Steinmetz, L. Sapinoso, G. Hampton, S. J. Elledge, R. W. Davis, and D. J. Lockhart. Transcriptional regulation and function during the human cell cyle. *Nature Genetics*, 27(1):48–54, 2001.

T. A. Clark, C. W. Sugnet, and M. Ares. Genomewide analysis of mrna processing in yeast using splicing-specific microarrays. *Science*, 296:907–910, 2002.

W. S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836, 1979.

W. S. Cleveland and S. J. Devlin. Locally-weighted regression: An approach to regression analysis by local fitting. *American Statistical Association*, 83:590–610, 1988.

J. DeRisi, L. Penland, P. O. Brown, M. L. Bittner, P. S. Meltzer, M. Ray, Y. Chen, Y. A. Su, and J. M. Trent. Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nature Genetics*, 14:457–460, 1996.

D. Dudoit and Y. H. Yang. Bioconductor R packages for exploratory analysis and normalization of cDNA microarray data. In G. Parmigiani, E. S. Garret, R. A. Irizarry, and S. L. Zeger, editors, *The Analysis of Gene Expression Data: Metods and Softwareg Manual*, New York, 2002. Springer.

B. Efron, T. J. Hastie, I. Johnstone, and R. J. Tibshirani. Least angle regression. *Annals of Statistics*, 32(2):407–451, 2004.

J. Fan, P. Tam, G. V. Woude, and Y. Ren. Normalization and analysis of cdna microarrays using within-array replications applied to neuroblastoma cell response to a cytokine. *PNAS*, 101(5): 1135–1140, 2004.

D. B. Finkelstein, J. Gollub, R. Ewing, F. Sterky, S. Somerville, and J. M. Cherry. Iterative linear regression by sector: renormalization of cDNA microarray data and cluster analysis weighted by cross homology. In *CAMDA 2000*, 2000.

D. Foster and E. George. The risk inflation criterion for multiple regression. *Annals of Statistics*, 22:1947–1975, 1994.

T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. Springer series in statistics. Springer, 2001.

T. J. Hastie and R. J. Tibshirani. *Generalized Additive Modles*. Chapman and Hall, New York, 1990.

W. Huber, A. von Heydebreck, H. Sültmann, A. Poustka, and M. Vingron. Variance stabilization applied to microarray data calibration and to the quantification of differetial expression. *Bioinformatics*, 1(1):1–9, 2002.

R. Ihaka and R. Gentleman. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5:299–314, 1996.

T. B. Kepler, L. Crosby, and K. T. Morgan. Normalization and analysis of DNA microarray data by self-consistency and local regression. Technical Report 00-09-055, Santa Fe Institute, 2000.

D. J. Lockhart, H. L. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Wang, M. Kobayashi, and H. Horton. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, 14:1675–1680, 1996.

A. Miller. *Subset Selection in Regression*. Chapman and Hall, London, 2002.

J. Quackenbush. Microarray data normalization and transformation. *Nature Genetics*, 32(4):496–501, 2002. Supplement to *Nature Genetics*.

M. Schena, editor. *DNA Microarrays : A Practical Approach*. Oxford University Press, 1999.

A. Schulze and J. Downward. Navigating gene expression using microarrays – a technology review. *Nature Cell Biology*, 3(2):E190–E195, 2001.

G. Schwartz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1979.

K. F. Sellers, J. Miecznikowski, and W. F. Eddy. Removal of systematic variation in genetic microarray data. Technical report, Department of Statistics, Carnegie Mellon University, 2003.

R. Tibshirani and K. Knight. The covariance inflation criterion for adaptive model selection. *Journal of the Royal Statistical Society, Series B*, 61:529–546, 1999.

Y. Wang, J. Lu, R. Lee, Z. Gu, and R. Clarke. Iterative normalization of cdna microarray data. *IEEE Trans Inf Technol Biomed*, 6(1):29–37, 2002.

D. L. Wilson, M. J. Buckley, C. A. Helliwell, and I. W. Wilson. New normalization methods for cdna microarray data. *Bioinformatics*, 19:1325–1332, 2003.

Y. Xiao, C. A. Hunt, M. R. Segal, and Y. H. Yang. A novel stepwise normalization method for two-channel cDNA microarrays. The IEEE Engineering in Medicine and Biology Society Conference 2004, 2004a. Accepted.

Y. Xiao, Y. H. Yang, T. A. Burckin, L. Shiue, G. A. Hartzog, and M. R. Segal. Analysis of a splice array experiment. Submitted to *Genome Biology*, `http://itsa.ucsf.edu/~yxiao/Research/Splice.htm`, 2004b.

Y. H. Yang, M. J. Buckley, S. Dudoit, and T. P. Speed. Comparison of methods for image analysis on cDNA microarray data. *Journal of Computational and Graphical Statistic*, 11(1):108–136, 2002a.

Y. H. Yang, S. Dudoit, P. Luu, D. M. Lin, V. Peng, J. Ngai, and T. P. Speed. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*, 30(4):e15, 2002b.

Y. H. Yang and N. P. Thorne. Normalization for two-color cdna microarray data. In D. R. Goldstein, editor, *Science and Statistics: A Festschrift for Terry Speed*, volume 40 of *LMS Lecture Notes – Monograph Series*, pages 403–418. 2003.

Table 1: Description of normalization methods used in the comparison of bias and variance. Model complexity (df) is approximated according to the splice data, which has 5760 data points and 16 print-tips.

Models	Description	Complexity (df)
median shift	simplest normalization that shifts the median of all log-ratios to zero	1
loess	A-dependent normalization using the scatter plot smoother *loess*	5 (span=0.4)
vsn	variance stabilizing transformation	4
quantile	single-channel intensity normalization	NA
LPT	A-dependent *loess* normalization conducted within each print-tip-group	$5 \times 16 = 80$
tRMA1	A-dependent *loess* normalization followed by spatial median filtering 3×3	$5 + \frac{5760}{3 \times 3} = 645$
tRMA2	A-dependent *loess* normalization followed by spatial median filtering 7×7	$5 + \frac{5760}{7 \times 7} = 123$
tRMA3	A-dependent *loess* normalization followed by spatial median filtering 15×15	$5 + \frac{5760}{15 \times 15} = 31$
STEPNORM	stepwise normalization	dependent upon arrays

Table 2: Results of application of STEPNORM on a QC array. BIC is employed as the model selection criterion. Models chosen by STEPNORM along the normalization steps are in bold fonts.

Bias	Models	K	$-2log\hat{L}(\times 10^{-4})$	**BIC**
A	null	0	-0.960	-0.960
	median shift	1	-0.983	-0.982
	rlm	2	-2.250	-2.248
	loess	5.51	-3.065	**-3.059**
PT	null	+0	-3.065	-3.059
	median shift	+48	-3.237	-3.184
	rlm	+96	-3.270	**-3.192**
	loess	+266	-3.253	-2.981
PL	**null**	+0	-3.270	**-3.192**
	median shift	+58	-3.326	-3.165
	rlm	+116	-3.387	-3.168
	loess	+320	-3.341	-2.917
Spatial 2D	**null**	+0	-3.270	**-3.192**
	rlm	+4	-3.294	-3.188
	loess	+13.6	-3.297	-3.182
	median filter (9×9)	+184	-3.366	-3.079
	ANOVA	+359	-3.394	-2.931

Table 3: Comparison of pairwise correlation of log-ratios between replicate arrays using the ApoAI data set.

Model	mean pairwise control slide correlation	adjusted mean DE gene correlation
raw	0.58	0.30
median shift	0.58	0.30
rlm	0.48	0.43
loess	0.41	0.46
LPT	0.29	0.57
vsn	0.40	0.47
tRMA1	0.25	0.59
step	**0.23**	**0.63**

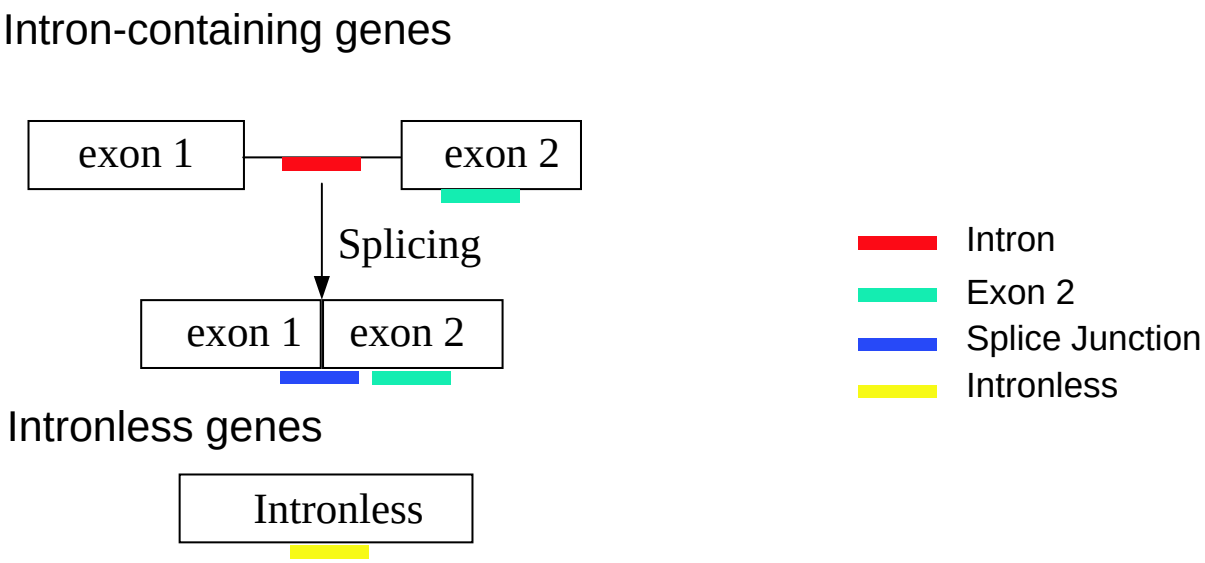


Figure 1: Probe design of the splicing arrays described in Experiment B. There are three oligonucleotide probes for each intron-containing gene: intron (red), splice-junction (blue) and exon (green). In addition, there are approximately 800 probes representing intronless genes (yellow). This figure is modified from Clark et al. (2002).

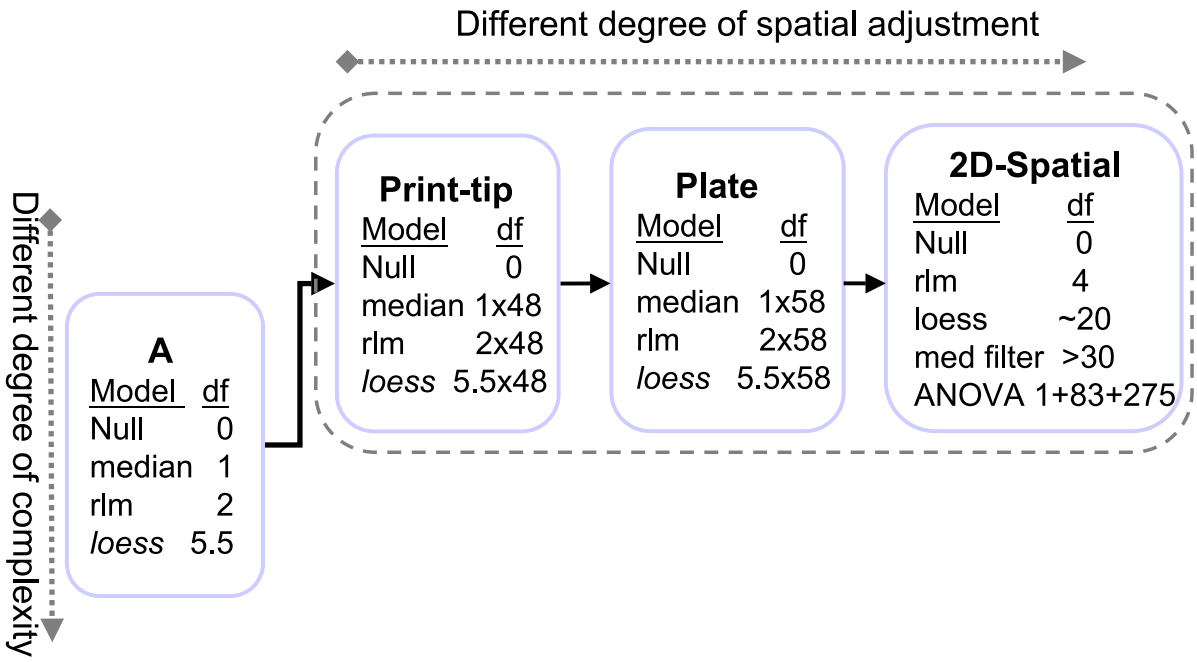


Figure 2: Stepwise normalization procedure using the example of the QC array in Experiment A. This QC array has 48 print-tip-groups and 58 well plates.

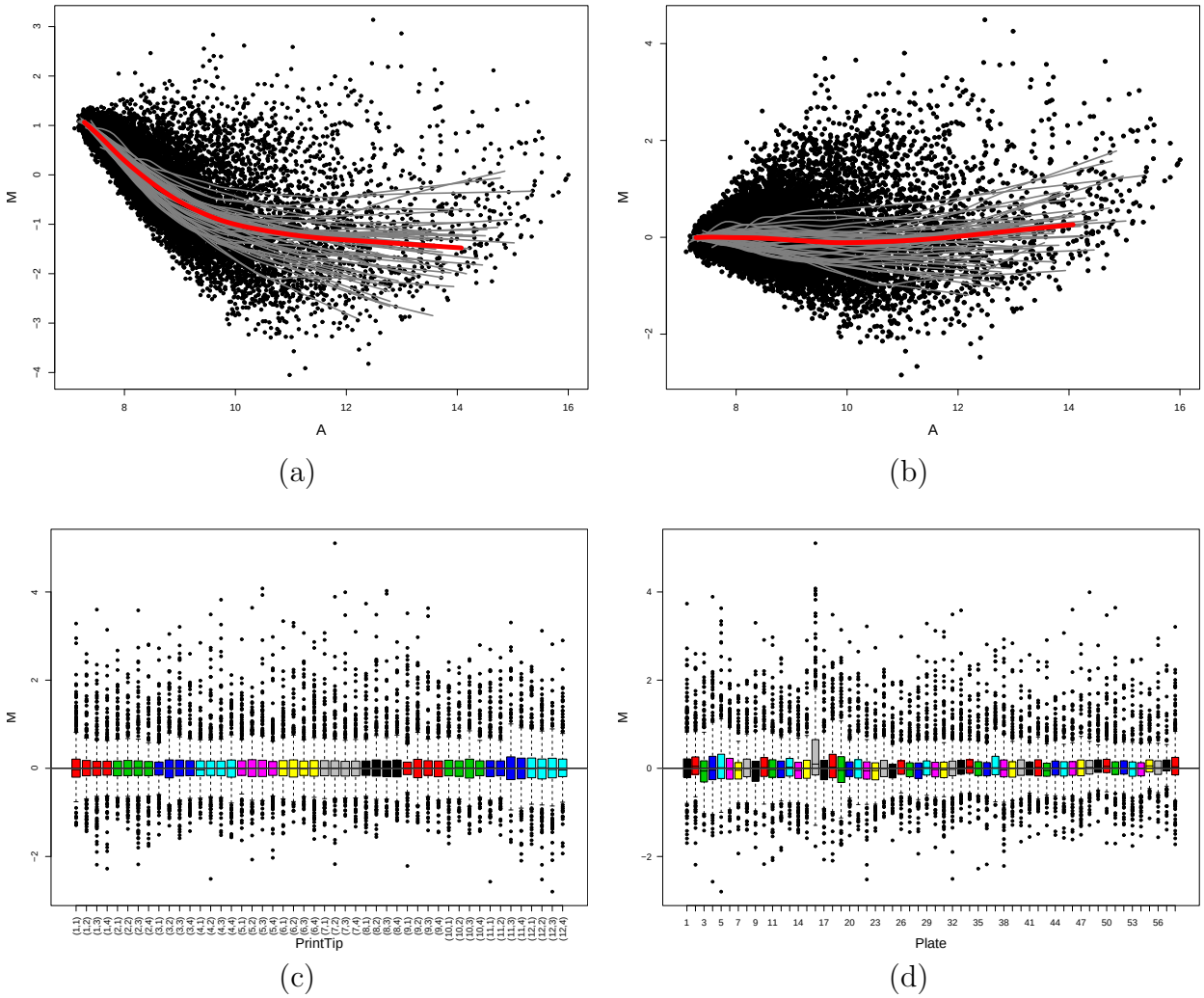
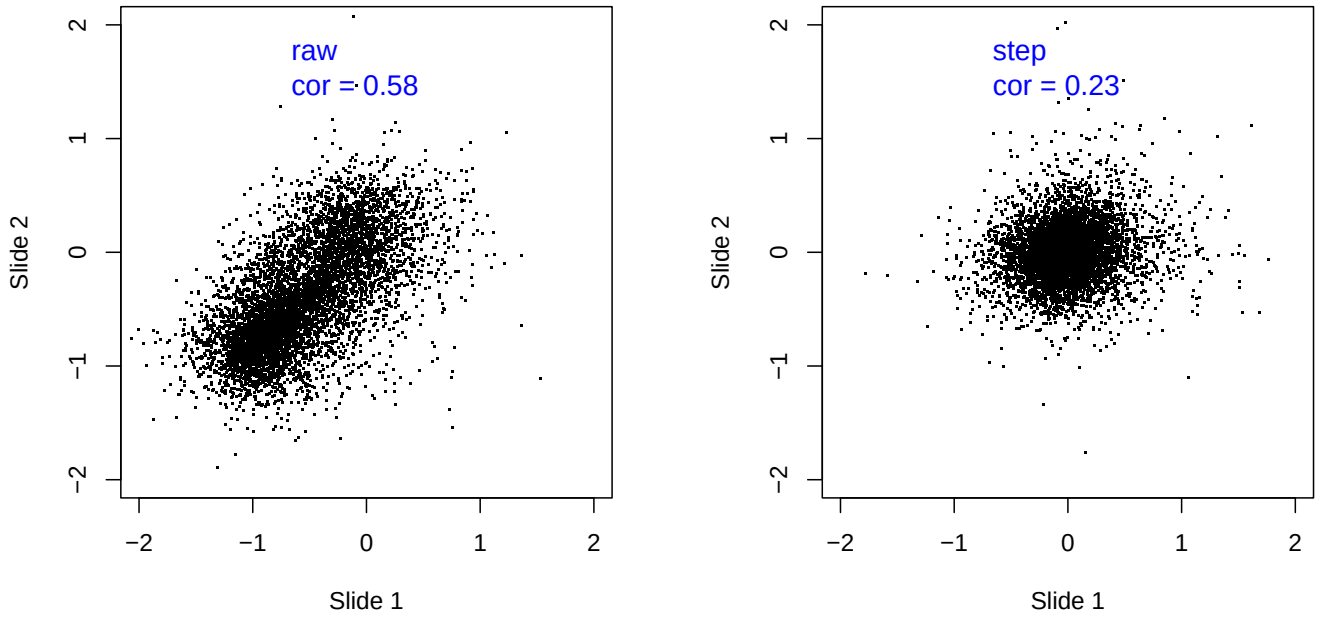


Figure 3: Graphical display of biases in log-ratios of a QC array before and after stepwise normalization. (a)*MA*-plots with *loess* fit for the whole array (red) and for each of the 48 print-tip-groups (gray) before normalization; (b)*MA*-plots after stepwise normalization for the removal of the A bias; (c) Boxplot stratified by the 48 print-tip-groups after stepwise normalization for the removal of the PT bias; (d) Boxplot stratified by the 58 well plates after stepwise normalization for the removal of the PT bias.



(a) (b)

Figure 4: Scatter plots of log-ratios (a) before and (b) after STEPNOM normalization between two randomly selected ApoAI arrays .

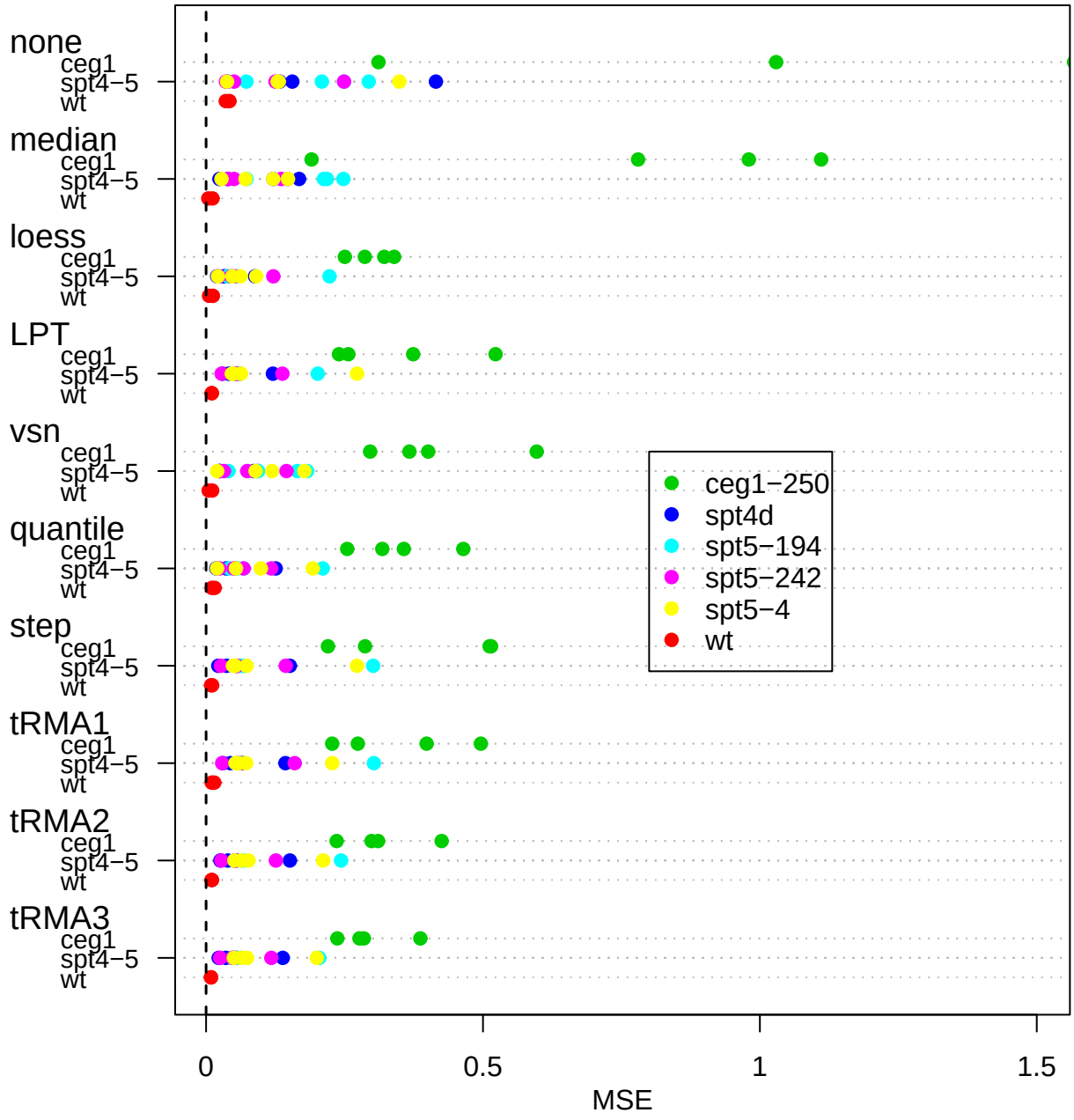
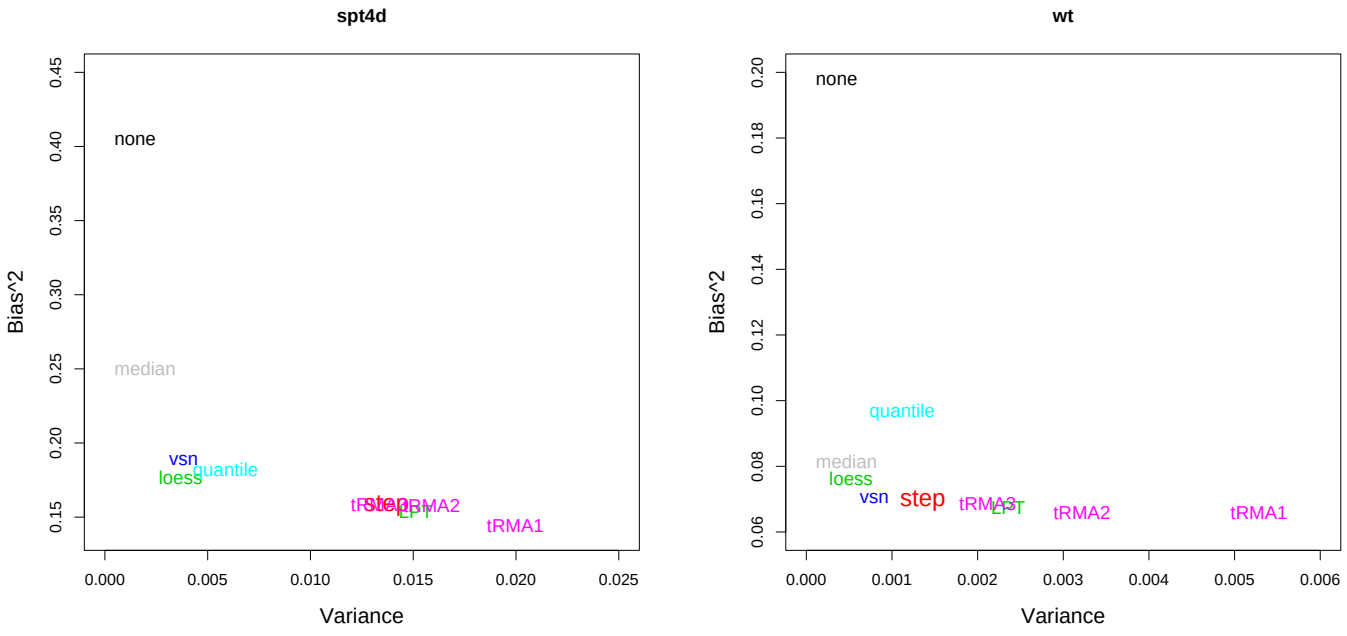


Figure 5: Dot plot comparing normalization methods based on the Spt experiment. Each dot represents the MSE of the normalization factor; see Equation 4. For a better illustration of differences between normalization methods, two of the *ceg1-250* arrays that have MSEs larger than 1.5 in the "none" method were removed from this figure.



(a) (b)

Figure 6: Scatter plot of squared bias versus variance of various normalization models for (a) the *spt4*Δ arrays and (b) the wildtype self self hybridization arrays.

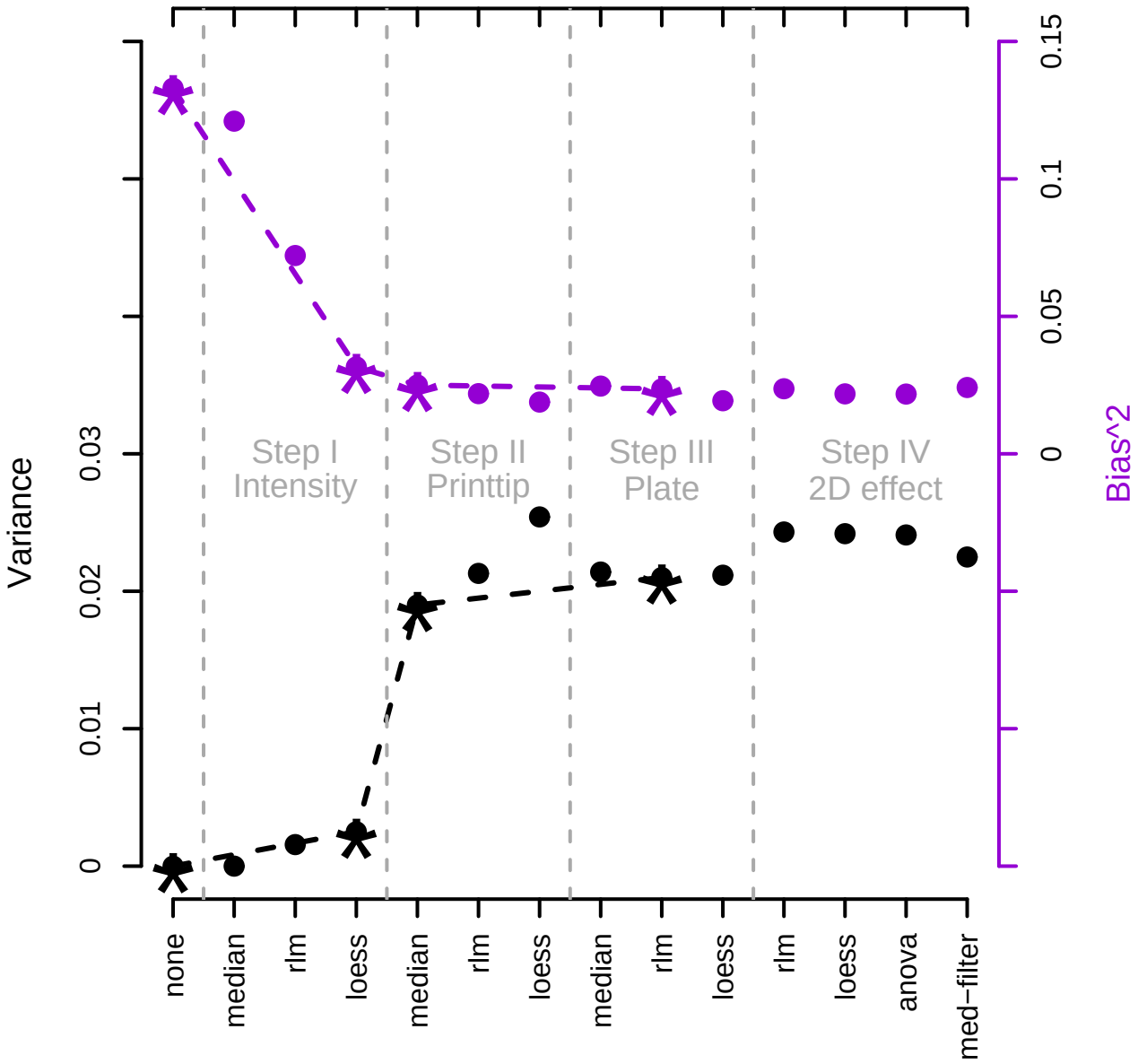


Figure 7: Plot of variance and squared bias for the stepwise normalization of a *spt4*Δ array. Colored points represent the variance (black) or the squared bias (purple) of each model within a normalization step. Asterisks highlight models selected by STEPNORM.

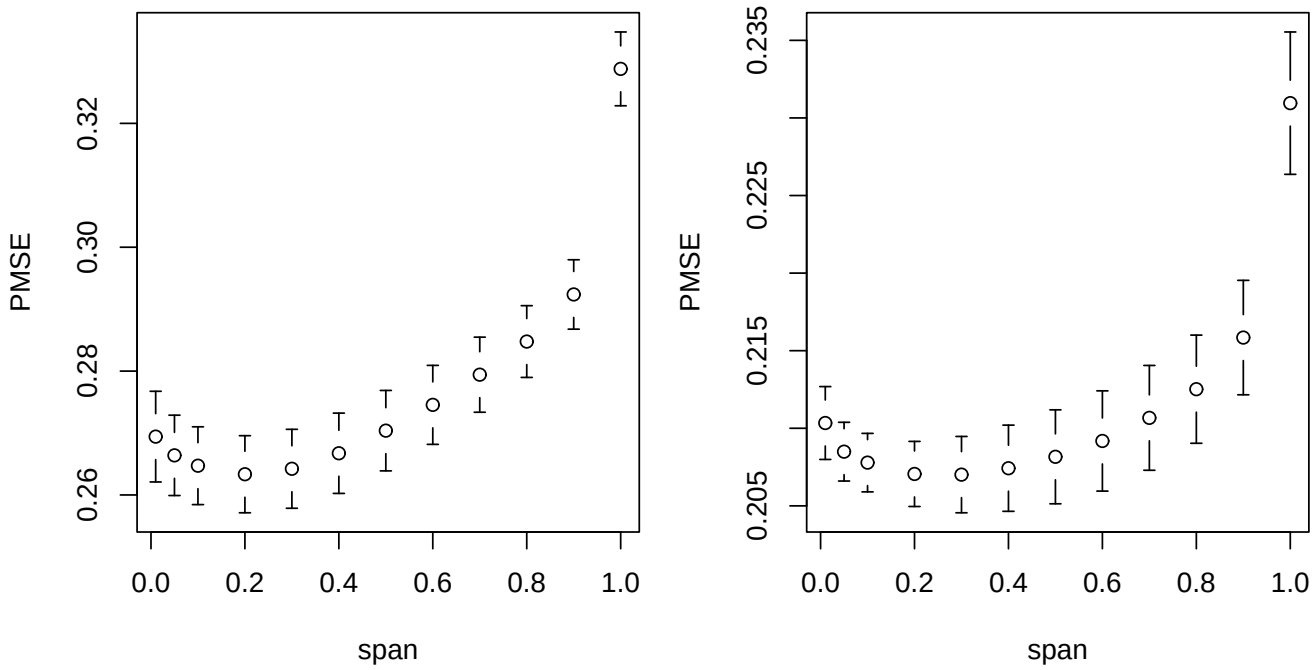
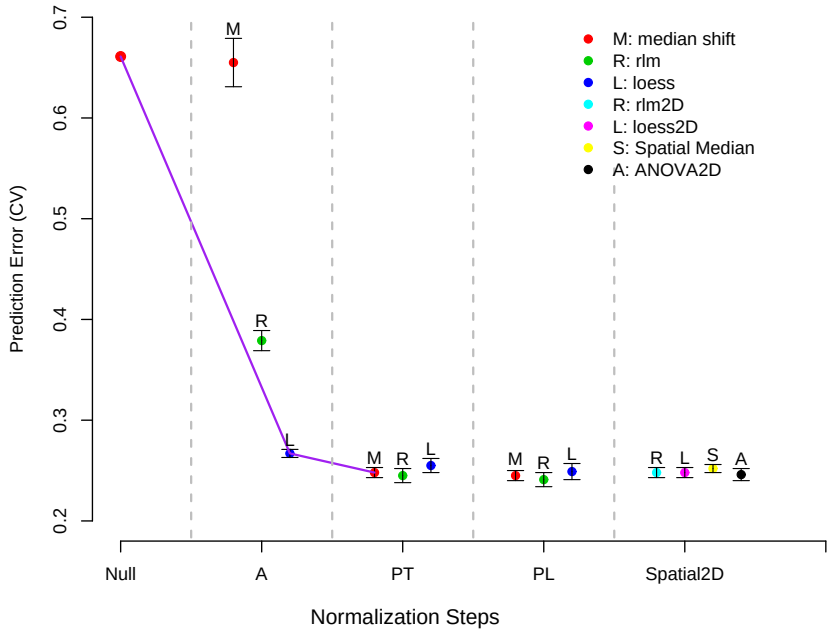
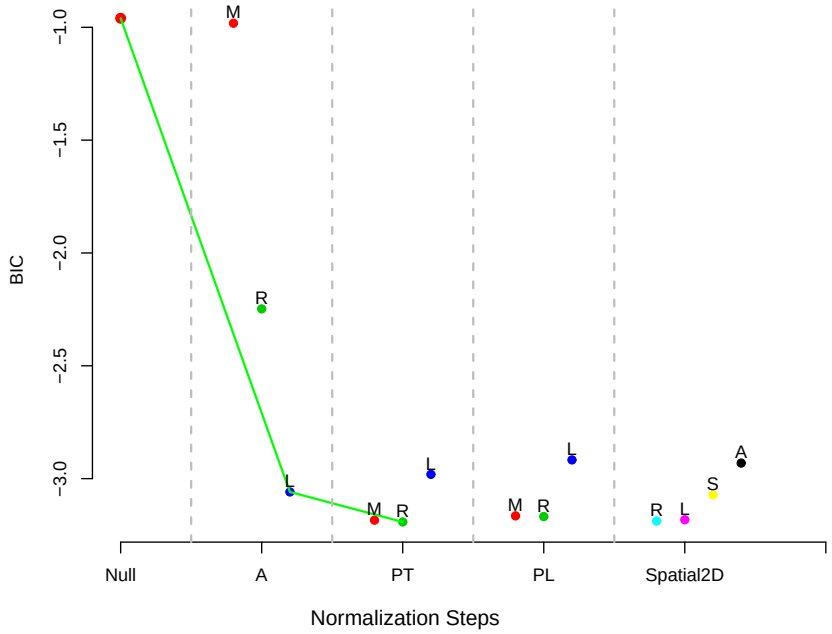


Figure 8: Plot of prediction mean square error (PMSE) by 5-fold cross validation versus spans, using two arrays from Experiment A.



(a)



(b)

Figure 9: Comparison of different model selection criteria based on a QC array. Panel (a) and (b) show plots of prediction errors at different normalization adjustments. (a) 5-fold CV (1-SE rule; Breiman et al. (1984); Hastie et al. (2001)) is employed as the model selection criterion. Plotted are CV prediction errors with error bars for models (colored spots) within each normalization step. (b) BIC is employed as the model selection criterion. Purple and green curves connect models selected by CV and BIC along the normalization steps respectively.