Supplemental Material

This document supplements the procedures and results presented in our main article “Cross Platform Comparison Of Microarray Data Using Order Restricted Inference”. Appendix A provides additional information on the quality control and data cleaning that has been performed prior to data analysis. Appendix B describes the annotation and probe matching that lead to the selection of 5927 common reporters, which are the basis of our cross platform comparison. This section also presents an assessment of the degree to which probe sequences unambiguously map to a recent version of the rat genome.

Appendix C presents additional details regarding the procedures proposed to estimate the variance components in situations with order restricted fixed effects, as well as, a performance comparison of the proposed methods with standard procedures. An additional table giving overall and specific agreement of test decisions inferred from raw and normalized data is reported in Appendix D. In Appendix E we present some further results on how agreement can be improved if low signal probes are removed from the analysis by a filtering procedure.

Appendix F presents an alternative test statistic based on the residual sum of squares of a linear model. Furthermore, differences in performance compared to the statistic based on isotonic regression, which is used for our main results, are discussed. Finally, Appendix G reports the session information of our R installation which provides detailed version information on all R-libraries used to compute the results.

A Data Cleaning

The experimenters’ comments at the ArrayExpress entry of the Illumina dataset (E-TABM-554) indicate that certain arrays from this experiment might be of inferior quality:

> IMPORTANT: 6C-1 and 1-6B1 are outlier because of poor cRNA yield. 4A-1 (low cRNA yield too) and 4D-1 seems to have been switched during processing. 2D1 and 5A3 were also mistake

(samples picked from the wrong well in the original 96 plates) as they seems to cluster with B samples (75% liver-25% Kidney).

To investigate this and other potential quality problems, we have conducted a quality assessment of the arrays used in this analysis. We visually inspected and compared the signal distributions of all chips from each platforms using boxplots and density estimates. Furthermore heatmaps and hierarchical cluster trees of the data were used to identify potential outlier arrays. This assessment confirms the experimenters comments regarding the Illumina dataset. In the Affymetrix data one array of doubtful quality was identified. The measured expressions on this array are on average twice as high with considerably higher variance than in the remaining Affymetrix samples. Hierarchical clustering using euclidean distance and complete linkage puts the array into a class of its own. The Agilent data showed no obvious outliers in terms of signal distributions. Clustering results group this dataset into the four titration groups. Consequently we have taken the following actions:

Illumina

- All arrays with cRNA yield below 10 were removed from the analysis. These are samples: 6C-1, 4D-1, 2C-2, 6B-1
- Arrays 5A-3 and 2D-1 are removed due to ambiguous labelling. We did not re-label these arrays based on the clustering results.
- Array 4A-1, however, was relabeled as a kidney sample, following the suggestion by the experimenters.

Affymetrix

- Array *NUID-0000-0064-2511* was removed from the dataset due to deviant expression distribution

B Annotation And Probe Matching

Probe annotations were based on the most recent array descriptions available at the time of annotation, which was January 23rd 2010. The file acquired from Affymetrix' website http://www.affymetrix.com had version number NA30. The annotation file found at http://www.agilent.com was called `014879_D_Genelist_20100118.zip` indicating a creation date of January 1st 2010. The chip description downloaded from Illumina's server at http://www.illumina.com was version 1.0-R3. Unfortunately, these files are available on sites that require a free user registration, so no public links can be cited.

To match probes across platforms, reporters targeting RefSeq 'NM' annotated mRNA transcripts [6] in all three platforms were chosen for analysis. Whenever more than one reporter was available for a single transcript,

	Matched				Missed	
	unambiguous		ambiguous			
Affymetrix	5,335	(90.0%)	119	(2.0%)	473	(8.0%)
Agilent	5,394	(91.0%)	514	(8.7%)	19	(0.3%)
Illumina	5,406	(91.2%)	393	(6.6%)	128	(2.2%)
Overlap	4,735	(79.9%)	76	(1.3%)	14	(0.2%)

Table 1: Alignment results for the 5,927 probes used in the main analysis. The vast majority of probes *matched* their target transcript. Some of these were also classified as *ambiguous*, indicating that they had a strong second best hit other than their target. Only a small fraction did not yield the maximum score for an alignment with the specified gene and thus *missed* its target. Altogether, 5,389 out of 5,927 probes matched their target sequences in all three platforms.

a random reporter was selected. Taking into account different microarray designs, the term 'reporter' indicated a probeset for Affymetrix and a long-oligo probe for Agilent and Illumina. A table of the selected reporters with corresponding mappings is available on the articles website at: http://statistics.msi.meduniwien.ac.at/float/cross_platform/.

Careful considerations lead to this selection: Relevant comparison needs to match common use as closely as possible. Typical microarray users rely on the annotation provided by the manufacturer and do not have probe reannotation tools at their disposal. Moreover, the RefSeq 'NM' transcripts have been well established, providing a fair basis for probe design to all platforms.

Holloway *et al.* [3] have pursued a similar strategy by selecting for UniGene IDs common to all platforms. They have assessed potential changes to their results due to more elaborate probe sequence matching, as unlikely. We agree for two reasons.

Firstly, the vast majority of probes are indeed correctly annotated, as the reanalysis in Table B shows. Only 14 genes were in fact missed by all platforms, while 5,389 out of 5,927 (91%) matched the right transcript in all platforms.

Secondly, to affect analysis, incorrectly assigned probes would need to have a strong bias for expression in either kidney or liver. That is in fact unlikely. Even more so, when cross-hybridizing probes are considered, that indistinguishably report different transcripts. In this case, all involved transcripts would need to have a bias in the same direction.

Methodological details

To reconfirm the validity of the probe annotation files used and to assess the probe matching procedure performed, the 5,927 reporters for 'NM' transcripts present in all three platforms were BLAST-ed (http://blast.ncbi.nlm.nih.gov) against the Reference mRNA Sequences database (refseq_rna, update from October 10th 2010).

The search was limited to the organism in question *Rattus norvegicus* (taxid: 10116), models ('XM' identifiers) and uncultured/environmental sequences were excluded. The blastn algorithm was applied [1], in order not to limit the results to hits of highly similar or perfect match sequences, but to equally allow for partially matching sequences. This is especially relevant for long-oligo probes. Short and long-oligos were BLAST-ed separately, such that the long-oligo queries were unaffected by BLAST's automatic parameter adjustment for short input. Default algorithm parameters were used and a maximum of five aligned sequences was obtained for each queried probe.

It was then checked, if the RefSeq ID of the top ranked BLAST hit had an E-value of less than 0.001 and if it matched the transcript ID claimed by the manufacturer. A probe meeting these criteria was considered a *match* and a *miss* otherwise. To further assess the amount of cross-hybridization for each platform, we evaluated the second best hits' potential to interfere with the target transcript measurement [4]. To this end, it was determined for which probes the second best hit would equally have an E-value of no more than 0.001. This subgroup of *matched* probes was denoted *ambiguous* probes.

Affymetrix probesets were classified as *matches*, if more than half of their probes were *matches*, while probe sets were classified as *misses*, if more than half of their probes were *misses*, and as *ambiguous* matches otherwise.

At this point, we would again like to emphasize, the difficulty in choosing a comparable set of probes across platforms. On the one hand, one would have to exclude all potentially cross-hybridizing reporters from the analysis, in order not to confound the biophysical properties of a given platform with the quality of its probe design. On the other hand, valid probe design is an inherent quality feature of the platform, that needs to be addressed in a separate comparison, which goes beyond the scope of this study.

Regardless, for the EMERALD data set analysed here, the fraction of potentially cross-hybridizing probes affects only 2.0–8.7% of the reporters, and are thus clearly not prevalent.

Table 2 shows percent specific agreement (see Section 2.5 of the main

	all:up	all:dn	all:dc	hit:up	hit:dn	hit:dc	miss:up	miss:dn	miss:dc
Affy-Agil:None	82.22	77.69	0.12	83.58	78.49	0.00	76.43	74.31	0.59
Affy-Agil:Base	82.74	73.67	1.87	84.17	75.60	1.28	76.43	66.17	4.20
Affy-Agil:Quan	81.88	75.39	1.75	83.46	76.96	1.35	75.03	68.77	3.38
Affy-Illu:None	84.72	84.25	0.36	86.10	85.20	0.22	78.65	79.67	0.96
Affy-Illu:Base	83.47	69.83	1.10	85.05	72.14	0.67	76.32	60.49	2.85
Affy-Illu:Quan	81.65	72.10	1.64	83.59	73.80	1.48	72.83	64.72	2.30
Agil-Illu:None	82.10	78.22	0.44	83.58	79.16	0.34	75.87	73.98	0.81
Agil-Illu:Base	80.60	62.98	2.49	81.93	64.68	1.96	74.73	56.29	4.61
Agil-Illu:Quan	79.37	76.13	1.69	80.92	77.13	1.58	72.53	71.83	2.18

Table 2: Specific agreement for probes of different annotation quality. First three columns give specific agreement achieved when all common probes are considered. Columns 4 to 6 give respective figures when only probes classified as unambiguous hits (across all platforms) are considered. The last three columns show results for the remaining probes

article for definition) obtained from different subsets of reporters. The first three columns give figures for all nine comparisons between platforms and normalizations when all 5927 common reporters are used to compute specific agreement. These figures are identical to those shown in Figure 6 of the article. The remaining six columns show results for the same comparisons either using only reporters that could be consistently mapped across all three platforms or only the remaining missing or ambiguous reporters. As can be seen, using only reporters of high annotation quality, improves agreement by approximately 1 percentage point. Agreement on the remaining reporters is around ten percentage points lower. This clearly indicates that badly annotated probes are a source of disagreement between platforms. Agreement of normalized data is worse compared to non-normalized data regardless of annotation quality. This excludes annotation problems as a possible source of the decrease in agreement observed on normalized data.

C Estimation Of Variance Components

C.1 Introduction

Our methods for estimating the variance components are based on the idea to estimate the variance components using a model matrix that is conditional on the estimated isotonic regression. Based on this approach, we suggest using estimates from either Henderson's Method III or Restricted Maximum Likelihood procedures (REML) [7]. Both methods can deal with imbalanced designs that arise when groups are pooled to fulfill the monotonicity restriction.

Section C.2 summarizes the model definition used in both approaches, which is the same as in the article. Specifics regarding monotonic regression and how the model matrix is conditionally manipulated are presented in Section C.3. Section C.4 outlines the approach based on Henderson's Method III and Section C.5 that based on REML respectively. Section C.6 finally presents results from simulation experiments, that demonstrate the gain in efficiency achieved over unrestricted procedures together with a performance comparison between the two approaches.

C.2 Model Definition

We suppose the following mixed effects model:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk} \tag{1}$$

$$\begin{aligned} \mu_g, \alpha_i & \quad \text{fixed effect} && (2) \\ \beta_j & \sim N(0, \sigma_\beta) && (3) \\ \gamma_{ij} & \sim N(0, \sigma_\gamma) && (4) \\ \epsilon_{ijk} & \sim N(0, \sigma_\epsilon), && (5) \end{aligned}$$

which implies a sum to zero constraint on the α_i ($i \in \{1, \ldots, a\}$) (*i.e.* $\sum_{i=1}^{a} \alpha_i = 0$). Furthermore we impose the following order restriction:

$$\alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_i \leq \alpha_{i+1} \leq \cdots \leq \alpha_a \tag{6}$$

or,

$$\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_i \geq \alpha_{i+1} \geq \cdots \geq \alpha_a \tag{7}$$

The aim of our procedures is to estimate the variance terms σ_i, σ_γ and σ_ϵ from (2-5).

C.3 Monotonic Regression

Monotonic regression is defined as a procedure that finds the best monotonic fit to a vector of dimension larger than 1. For a given direction (*i.e.* if it is known which restriction in (6 - 7) is to be used) isotonic regression [2] using the 'Pool Adjacent Violators Algorithm' minimizes the squared error of the regression fit, through successive pooling of adjacent groups with order violating means. For a vector of dimension a this returns a partition of the index set $\{1, ..., a\}$ into subsets of subsequent indices so that the means of these aggregated groups fulfill the order restriction. Define the functions $I^{up}(i), I^{down}(i)$ that for each $i \in \{1, ..., a\}$ return the corresponding subsets of indices in the partition given by the isotonic regression assuming an upward or downward trend respectively.

Using model (1), isotonic regressions, for both order restrictions (6,7), are fit to the means $\overline{y}_i = \frac{1}{bn}\sum_j \sum_k y_{ijk}\ i \in \{1, ..., a\}$. The corresponding isotonic regression fits for any given i are then given by:

$$y_i^{\star up} = \frac{1}{nb|I^{up}(i)|} \sum_{i' \in I(i)} \sum_{jk} y_{i'jk}, \tag{8}$$

and

$$y_i^{\star down} = \frac{1}{nb|I^{down}(i)|} \sum_{i' \in I(i)} \sum_{jk} y_{i'jk} \tag{9}$$

respectively. The directional decision is then based on which direction gives the fit with the lowest residual sum of squares. The index set partition corresponding to monotonic regression for any given index i is therefore given by

$$I(i) = I^{\text{argmin}_{up,down}\left\{\sum_{i'}(\overline{y}_{i'}-y_{i'}^{\star up})^2,\sum_{i'}(\overline{y}_{i'}-y_{i'}^{\star down})^2\right\}}(i) \tag{10}$$

and the monotonic regression mean of group i by

$$y_i^\star = y_i^{\star\text{argmin}_{up,down}\left\{\sum_{i'}(\overline{y}_{i'}-y_{i'}^{\star up})^2,\sum_{i'}(\overline{y}_{i'}-y_{i'}^{\star down})^2\right\}}. \tag{11}$$

Conditional on the monotonic regression we then redefine the levels of the fixed effects α_i according to the partition of the index set $\{1, ..., a\}$.

As example, consider the vector $(\overline{y}_1, ..., \overline{y}_4) = (1, 3, 2, 4)$ and assume an upward trend. The isotonic regression would result in the partition of $\{1\}, \{2, 3\}, \{4\}$ and the corresponding fit is $(y_1^{\star up}, ..., y_4^{\star up}) = (1, 2.5, 2.5, 4)$. Assuming a downward trend all groups are pooled resulting in the partition $\{1, 2, 3, 4\}$ and fit $(y_1^{\star down}, ..., y_4^{\star down}) = (2.5, 2.5, 2.5, 2.5)$. Based on the residuals we decide for an upward trend in this case. Consequently, α_i which originally had 4 levels in this example is reduced to the levels 1, $\{2, 3\}$, and 4

C.4 Estimation Using Henderson's Method III

As can be seen in the examples of Section C.3 the model matrix becomes unbalanced ruling out the typical ANOVA sums of squares decomposition for estimation of the variance components. Henderson's Method III is a generalized ANOVA method that provides variance estimates in situations for imbalanced designs. For a detailed outline of the method see for example Chapter *5.5* of [7], with the special case of the 'with interaction mixed model' being treated in Chapter *5.6.d.* With Henderson's procedure the variance component estimates are obtained as solution of a system of linear equations involving sums of squares from several submodels. The downside of this approach is that in certain cases the solutions can become negative. To avoid negative variance estimates such solutions have to be set to zero. The corresponding estimation procedures were implemented in R and are available on request from the authors.

C.5 Estimation Using Restricted Maximum Likelihood

Another possibility to compute estimates for the variance components, is to use a restricted maximum likelihood procedure [7]. The pooling of order violating groups in this case is incorporated into the model matrix which is passed to the estimating procedure. We use the methods implemented in the R package `nlme` [5] to acquire estimates using this approach.

C.6 Simulation Results

To assess the potential improvement of order restricted estimation, we compare the following methods to estimate the variance components in our simulations:

ANOVA The ANOVA estimate (based on the factor levels α_i) for balanced designs. Negative variance estimates are set to zero.

Isotonic ANOVA Henderson's Method III (based on the factor levels obtained from monotonic regression) as outlined in Sections C.3 and C.4.

REML The variance components are estimated based on the original factor levels α_i using the restricted maximum likelihood approach as implemented in the R package `nlme`.

Isotonic REML The variance components are estimated based on the factor levels obtained from monotonic regression (as defined in Section C.3) using the restricted maximum likelihood approach as detailed in Section C.5.

Improvements in efficiency through restricted estimation can be expected for situations where order violations occur frequently. By pooling those, less group means have to be estimated, enabling degrees of freedom to be retained. Cases where this is to be expected are situations of no or little trend across groups.

We show by simulation that our methods lead to a considerable improvement in estimation efficiency. This comes at the price of a slightly increased bias which does, however, converge to zero for increasing numbers of independent samples.

Our simulation scenarios are all derived from a parameter setup close to what we observe in the EMERALD dataset. It is characterized by a relatively large residual error, in comparison with small to medium animal- and interaction variance, of approximately comparable range. In the reference scenario the corresponding parameters are set to $\sigma_\beta = \sigma_\gamma = \frac{1}{2}$ and $\sigma_\epsilon = 1$. The sample size parameters $a = 4$, $b = 6$ and $n = 3$ were chosen to replicate the EMERALD experiment's design. The reference setup represents a situation with constant expression across titration groups (*i.e.* no trend). Based on this reference scenario we let one or more parameters vary while keeping the rest fixed. For each parameter constellation the root mean squared error and bias of the variance component estimates, using all four methods, were computed using simulated datasets (10^4 runs).

For each of the scenarios we show plots of the bias and root mean squared error in the estimate of each components variance across the range of the variable parameter. The plots are structured in two rows and three columns. The first row shows the bias and the second the root mean squared error. The three columns are dedicated to the estimated variance components of the model. The first corresponds to the individual (Animal) error σ_β the second to the interaction term σ_γ and the third to the residual error σ_ϵ. Results from the four investigated methods are discriminated by different line types where color differentiates between the REML (red) and ANOVA (black) based methods and solid and dashed lines between unrestricted (dashed) and order restricted (solid) procedures.

To demonstrate the improvement in efficiency achieved by restricted estimation, we first show scenarios with varying degrees of a trend throughout the titration groups. For this purpose we add a linear trend, parametrized by its slope (*i.e.* $(\alpha_1, ..., \alpha_4) = d * (1, 2, 3, 4)$) to the reference setup. Figure 1values in the interval -1 to 1 and is shown on the x-axis. If there is no trend (*i.e.* $d = 0$) or in situations of moderate trends, all restricted procedures lead to a gain in efficiency compared to their unrestricted counterparts. Mean squared error is reduced most effectively in the estimation of the interaction effect. As expected for increasing slopes the performance of the restricted procedures converges towards that of the unrestricted procedure.

The absolute bias and MSE of the order restricted procedures decreases for increased sample size when we let the number of independent samples b vary, ranging from 12 to 60. The results of this simulation are shown in

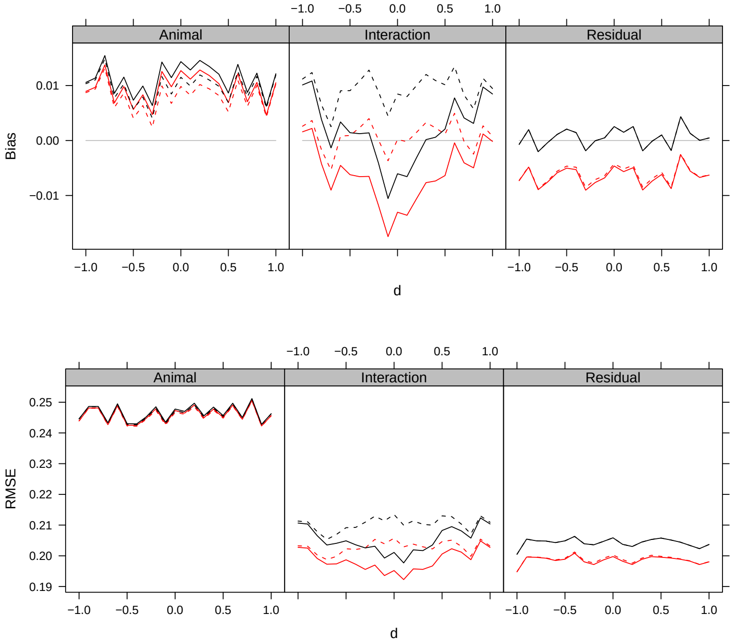

Figure 1: Bias and root mean squared error in estimates of each components variance achieved with linear trends of varying slope. Red lines show values for REML based procedures; black lines for ANOVA based procedures. Dashed lines represent unrestricted procedures; solid lines order-restricted procedures.

Figure 2. As the number of animals (index $_j$) gets larger the bias and MSE in estimating the variance components decreases. This is a good indication for asymptotical consistency of our methods.

Finally, to illustrate a key difference between the methods based on REML and Henderson's Method III, we let the animal and interaction variance simultaneously take values in the interval from 0 to $\frac{1}{2}$ ($i.e.$ $\sigma_\beta = \sigma_\gamma \in [0, \frac{1}{2}]$). Configurations with relatively small variance components are especially prone to generate data with negative solutions in the procedures derived from Henderson's method. Setting these estimates to zero, in order to avoid negative variance estimates, decreases the accuracy of this method compared to the REML based procedures. Figure 3 shows that the advantage of REML is most apparent in situations where the animal and interaction variance are close to zero and decreases for larger values.

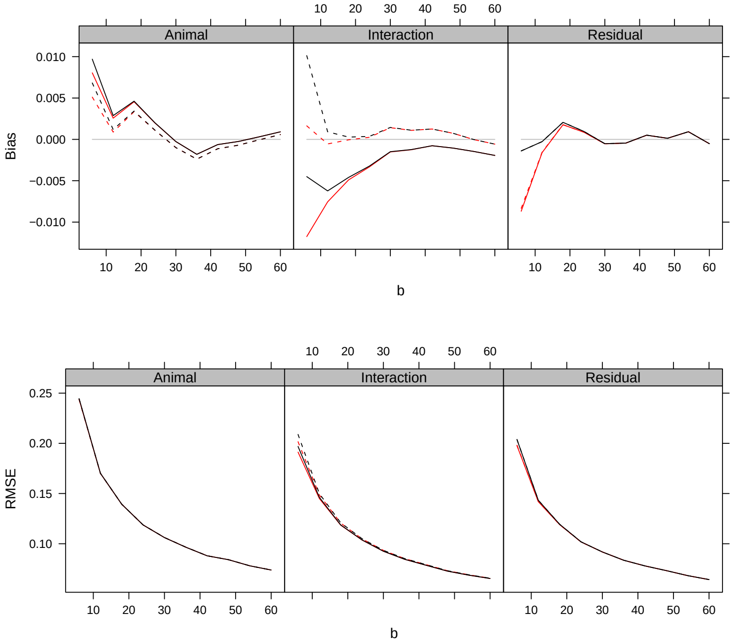

Figure 2: Bias and root mean squared error of estimates of each components variance with different sample sizes (number of animals). Red lines show values for REML based procedures; black lines for ANOVA based procedures. Dashed lines represent unrestricted procedures; solid lines order-restricted procedures.

Independent from differences in efficiency, an apparent advantage of the ANOVA based methods over REML procedures are a much quicker computation time and closed solutions for the estimates. In our implementation the iterative REML procedures are slower by approximately a factor 100. Furthermore, we encountered failing software due to problems with convergence and computational singularity in a small proportion of our simulations. In most cases these problems could be overcome by readjustment of certain control parameters. This however, required human intervention during the otherwise automatic code execution. Nevertheless, on the real dataset no such case was encountered and the size of the investigated data was small enough to feasibly use REML procedures.

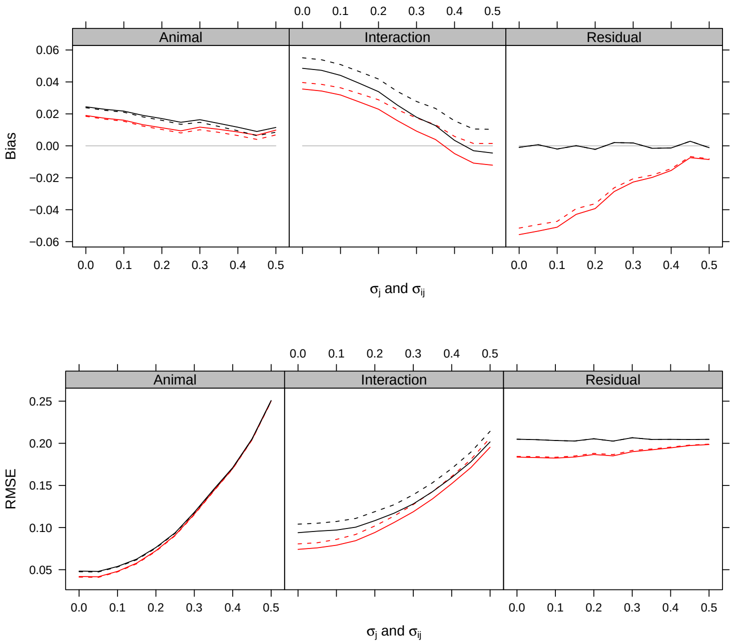

Figure 3: Bias and root mean squared error achieved in situations of varying animal and interaction variance. Red lines show values for REML based procedures; black lines for ANOVA based procedures. Dashed lines represent unrestricted procedures; solid lines order-restricted procedures.

D Additional Table: Agreement Between Normalizations

Table 3 shows overall and specific agreement, as defined in Section 2.5 of the main article, for pairwise comparisons between all combinations of test decisions inferred from raw or normalized data.

	None-Base				None-Quan				Base-Quan			
	all	up	none	down	all	up	none	down	all	up	none	down
Affy	0.77	0.92	0.7	0.59	0.64	0.85	0.51	0.48	0.81	0.91	0.71	0.82
Agil	0.56	0.78	0.29	0.44	0.63	0.78	0.5	0.53	0.72	0.88	0.54	0.7
Illu	0.84	0.91	0.81	0.74	0.71	0.82	0.66	0.6	0.77	0.87	0.72	0.72

Table 3: Overall and specific agreement for pairwise comparisons between normalizations.

E Agreement Of Test Decisions Stratified By Overall Expression

As shown in Section 3.3 of the main article, Normalization has negative effects on the agreement of differential expression analysis across platforms. In this section we examine the question whether these effects are homogeneous across the whole signal range or only apply to genes with low (on average) expression values. For this purpose we recalculated the agreement measures using probes for which the average (over all samples) expression exceeded a certain threshold. The threshold was chosen as such that the median expression of the included probes had to be larger than a given percentile of the data distribution of median expressions in at least one of the three platforms. For example: At the 10 percent threshold we removed all probes for which the median expression did not exceed the 10 percent smallest median expressions in any of the three platforms. In doing so, we mimic an unspecific filtering procedure which removes probes of low signal from the final analysis in order to achieve a better power for differential expression inference. Such filtering is often applied, as it is assumed that the proportion of true alternatives is larger within genes of higher overall expression.

Figure 4 shows specific agreement of up and downward trends as well as the proportion of test decision with contradicting results between different platforms for all pairwise comparisons of the three platforms. The x-axis of each panel gives the percentile that the overall per platform expression median has to exceed in at least one of the three platforms. We observe that agreement increases both for up and downward trends if only highly expressed genes are examined. At the same time the proportion of genes for which a contradicting test decision was found between two platforms decreases.

Figure 5 looks at the proportion of genes rejected at each threshold value. Overall one can see that the proportion of genes rejected among the selected, increases (column furthest left), and that the proportion of genes selected with an upward trend, decreases. Assuming a higher mRNA concentration in the kidney samples provides an explanation for this phenomenon as it amplifies any upward trend and attenuates the downward trends. It is hence reasonable to believe that the available power to reject an upward trend is generally very high and cannot be to a great extent improved by the filtering procedure. Improving the power to detect a downward trend thereby shifts the proportion of directional decisions to the advantage of downward trends.

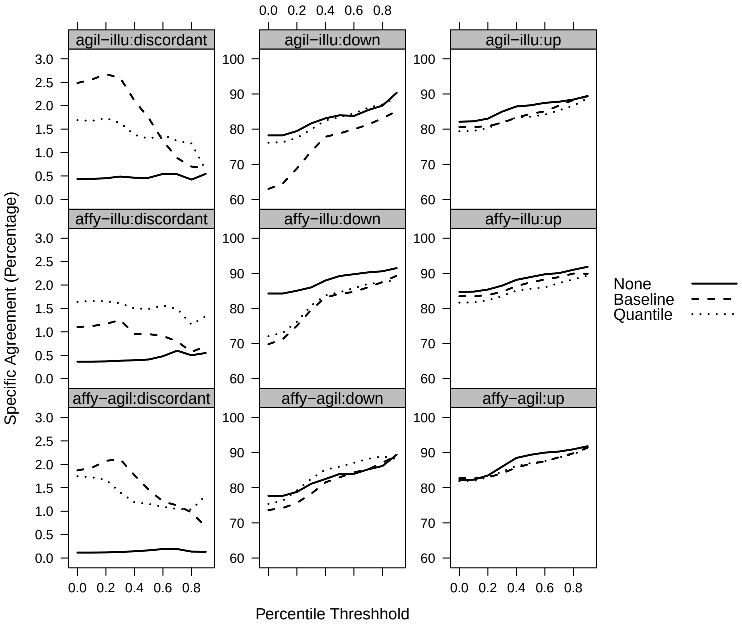

Figure 4: Specific agreement and contradicting decisions depending on a threshold that the median signal of a gene has to exceed in order to be included in the analysis. First column shows the proportion of discordant probes (as defined in Section 2.5 of the main article) for the three pairwise comparisons, if the analysis were restricted to only those measurements exceeding the threshold. Columns two and three show specific agreement as a function of the inclusion threshold for up and downward trends respectively.

F Linear Trend Test

To provide a comparison between our pattern based approach to an approach based on a linear model we have computed an alternative test based on permutations of the residual sum of squares of a linear regression fit to each gene according to the following model:

$$y_i = T_L \lambda_i + T_K \kappa_i + \epsilon_i,$$

where y_i stands for the vector of measurments from gene i. T_L and T_K are vectors with the titration contents of liver and kidney material ($e.g.\ 0, \frac{1}{4}, \frac{3}{4}, 1$) in the corresponding samples. λ_i and κ_i are unknown parameters for the expression level in liver and kidney that are estimated by ordinary least

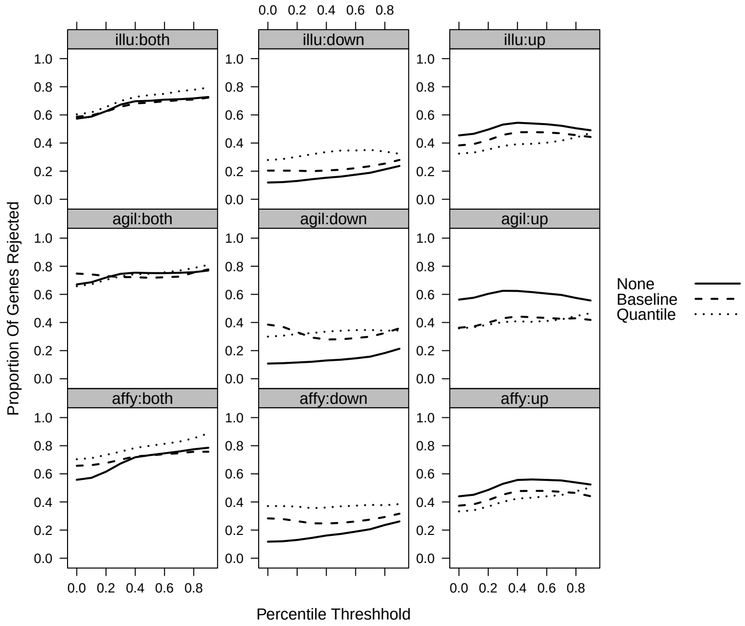

Figure 5: Proportion of genes found with a significant trend at each selection threshold. The first column shows the proportion of genes with a significant trend regardless of direction. The second and third column, show the proportions of significant upward and downward trends respectively. Each row gives results for a different platform.

squares regression. Finally ϵ_i stands for the residual error. The ratio between the residual and total sum of squares defines a statistic that can directly be used, instead of E^2, with the resampling based multiple testing procedure outlined in Section 2.5 of the main article.

Results from applying this method to our dataset show that neither statistic is uniformly more powerful across platforms and normalization methods. For non-normalized data the linear test is slightly better (identifying between 2 and 4 percent more significant trends). On baseline normalized data the advantage of the linear test is decreased in all platforms. Regarding the Agilent data, Barlow's test identifies 2 percent more significant trends. However, the statistic based on isotonic regression, outperforms the linear test on quantile normalized data, identifying 2 to 4 percent more significant trends between all platforms. We see that the linear trend test has a slight advantage on non-normalized data. For the normalization that provides the most rejections the isotonic regression based statistic is more favorable across all platforms.

It is theoretically clear that the linear test is more powerful against linear alternatives, whereas Barlow's test is more powerful against non-linear trends. Exploratory analysis of the measurements show that those trends only found significant by the linear trend test, have an average difference between liver and kidney three to five times smaller, than those genes found significant by both genes. Genes found significant only by Barlow's test, show considerable differences between slopes from either L to $M1$ or $M2$ to K compared to the slope from L to K, which indicates a deviation from linearity. Saturation as one source of non-linearity may lead to such a distribution of alternatives where the majority of the non-linear trends are those genes with large effects making them easy to detect by either statistic. If linearity is provided for genes with smaller differences between the tissues then the linear trend test can have an advantage over Barlow's test.

	linear	E2
Affy:None	3399	3306
Agil:None	4124	3975
Illu:None	3529	3399
Affy:Base	3918	3896
Agil:Base	4335	4433
Illu:Base	3585	3479
Affy:Quan	4044	4171
Agil:Quan	3788	3899
Illu:Quan	3431	3579

Table 4: Number of genes rejected by test statistic. Left column provides numbers of genes found significant using the linear model based statistic. Right column shows figures for Barlow's statistic

G R Session Info

Finally, we load all procedures into R and print the session info to provide details on the R-packages used to compute the results in our manuscript:

```
> source(file.path(wd, Rpath, "testProcedures.R"))
> source(file.path(wd, Rpath, "varCompProcedures.R"))
> source(file.path(wd, Rpath, "utilities.R"))
> source(file.path(wd, Rpath, "dedicatedPlotting.R"))
> sessionInfo()

R version 2.10.1 (2009-12-14)
x86_64-pc-linux-gnu

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
 [5] LC_MONETARY=C              LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] grid      stats     graphics  grDevices utils     datasets  methods
[8] base

other attached packages:
 [1] fBasics_2100.78    timeSeries_2110.87 timeDate_2110.87   MASS_7.3-5
 [5] xtable_1.5-5       Biobase_2.4.1      latticeExtra_0.6-5 lattice_0.18-3
 [9] RColorBrewer_1.0-2 limma_3.2.3        orQA_0.2.0         nlme_3.1-96
[13] genefilter_1.28.2  Rcpp_0.8.6         gtools_2.6.1       gdata_2.7.1
[17] multicore_0.1-3

loaded via a namespace (and not attached):
[1] annotate_1.22.0     AnnotationDbi_1.6.1 DBI_0.2-5
[4] RSQLite_0.9-2       splines_2.10.1      survival_2.35-8
```

References

[1] S.F. Altschul, T.L. Madden, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389, 1997.

[2] Richard E. Barlow. *Statistical Inference Under Order Restrictions.* John Wiley and Sons Ltd, 1972.

[3] A. J Holloway, A. Oshlack, et al. Statistical analysis of an RNA titration series evaluates microarray precision and sensitivity on a whole-array basis. *BMC Bioinformatics*, 7(1):511, 2006.

[4] G.G. Leparc, T. Tuchler, et al. Model-based probe set optimization for high-performance microarrays. *Nucleic Acids Research*, 37(3):e18, 2009.

[5] Jose Pinheiro, Douglas Bates, et al. *nlme: Linear and Nonlinear Mixed Effects Models*, 2009. R package version 3.1-93.

[6] K.D. Pruitt et al. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*, 35:D61–65, 2006.

[7] S. R Searle, G. Casella, et al. *Variance components.* Wiley New York, 1992.