Comparison of Preprocessing Procedures for Oligo-nucleotide Micro-arrays by Parametric Bootstrap Simulation of Spike-in Experiments

J. Freudenberg[1], H. Boriss[1], D. Hasenclever[2]
[1]Interdisciplinary Center of Bioinformatics (IZBI), University of Leipzig, Germany
[2]Institute of Medical Informatics, Statistics and Epidemiology (IMISE), University of Leipzig, Germany

Summary

Objective: Due to scarcity of calibration data for micro-array experiments, simulation methods are employed to assess preprocessing procedures. Here we analyze several procedures' robustness against increasing numbers of differentially expressed genes and varying proportions of up-regulation.

Methods: Raw probe data from oligo-nucleotide micro-arrays are assumed to be approximately multivariate normally distributed on the log scale. Chips can be simulated from a multivariate normal distribution with mean and variance-covariance matrix estimated from a real raw data set.

A chip effect induces strong positive correlations. In reverse, sampling from a normal distribution with strong correlation variance-covariance matrix generates data exhibiting a chip effect. No explicit model of chip-effect is needed. Differences can be artificially spiked-in according to a given distribution of effect sizes.

Thirty preprocessing procedures combining background correction, normalization, perfect match correction and summarization methods available from the BioConductor project were compared.

Results: In the symmetrical setting "50% differentially expressed genes, 50% of which up-regulated" background correction reduces bias, but inflates low intensity probe variance as well as the mean squared error of the estimates. Any normalization reduces variance and increases sensitivity with no clear winner. Asymmetry between up and down regulation causes bias in the effect-size estimate of non-differentially expressed genes. This markedly inflates the false positive discovery rates. Variance stabilizing normalization (VSN) behaved best.

Conclusion: A simple parametric bootstrap was used to simulate oligo-nucleotide micro-array raw data. Current normalization methods inflate the false positive rate when many genes show an effect in the same direction.

Keywords

Gene expression profiling, oligo-nucleotide array, normalization

Methods Inf Med 2004; 43: 434–8

Introduction

High-density oligo-nucleotide micro-arrays by Affymetrix contain 11 to 20 short (25-mer) perfect match (PM) nucleotide sequences per transcript. In addition, every PM sequence is accompanied by a mismatch (MM) sequence where the middle nucleotide is replaced by its complement [1]. Raw probe level data of such arrays require a four-step preprocessing in order to obtain transcript-wise expression values: Background correction (to eliminate stray signals separately for each chip), normalization (to account for chip effects caused by variance in total amount of RNA or scanning conditions etc.), PM correction (making use of the probe-pair design of the micro-array technology in an attempt to eliminate non-specific hybridization), and summarization (synthesizing one expression value per gene from the multiple probes addressing it). For each of these steps multiple options have been developed most of which are readily available within the BioConductor project [2]. This generates a confusing plethora of combinations.

For the purpose of evaluating and validating statistical procedures a few calibration data sets with known underlying parameters have been made publicly available (e.g. GeneLogic, http://qolotus02.genelogic.com/datasets.nsf/; Affymetrix, http://www.affymetrix.com/support/datasets.affx). They offer a valuable insight on the precision and accuracy of expression measures as Cope et al. show in a recent paper where they describe a graphical web tool to automatically assess the performance of preprocessing methods based on theses data sets [3].

However, the available calibration data sets either contain only a handful of differentially expressed genes (spike-in experiments) or all genes present are differentially expressed with effect in the same direction (dilution series experiments). In addition, they tend to exhibit an unusually low noise level. For both reasons they do not represent a typical experiment setting. Furthermore, as the available calibration data sets are repeatedly used in the development of new methods for the aforementioned preprocessing steps a certain degree of over-fitting may be suspected.

To circumvent some of these obstacles and to further improve research on preprocessing strategies we developed a technique to simulate artificial spike-in chip raw data. In particular, we are now able to study the influence of the proportion of differentially expressed genes and the influence of the proportion up-regulated among differentially expressed genes. This has not been systematically investigated before.

Simulation Methods

We use the parametric bootstrap [4] to simulate chips based on a given real raw data set assuming that raw probe data from oligo-nucleotide micro-arrays are approximately multivariate normally distributed on the log scale. The log scale is used because errors [6] (at least for larger signals) and chip effects tend to be multiplicative.

It may surprise that we do not need to specify a particular model for the chip-effect. Note that the existence of a chip effect induces strong positive correlations between all probes in the observed variance-

covariance matrix. A chip effect can be understood as just adding a chip-specific term on the log scale to all probes of a given micro-array. In three un-normalized data sets we found a median correlation coefficient of 0.72, 0.82 and 0.80 with 82.3, 91.3 and 97.2% over 0.5, respectively.

In reverse, the sampling of chips from a multivariate normal distribution with such strong positively correlated variables results in data that exhibit all the characteristics that are usually regarded as indicative to the presence of a chip effect.

Given the Cholesky decomposition of the empirical variance-covariance matrix it is standard to generate a corresponding multivariate normal random vector using a univariate normal random generator. As is well known, the empirical variance-covariance matrix by far doesn't have full rank. Instead of artificially regularizing it by adding small values to the diagonal [5] before sampling, we apply the Cholesky decomposition algorithm to the singular positive semi-definite matrix. Notice that the resulting lower triangular matrix has only r non-zero columns where r is the rank of the original empirical variance-covariance matrix. Thus we are able to massively reduce both computer time and workspace requirements. This procedure has only a small numerical impact on the sampled data when comparing to data sampled from a regularized matrix. With this simplification we can simulate micro-arrays on the probe level (dimension ~ 400,000).

In order to investigate the impact of an increasing number of differentially expressed genes on the performance of the normalization procedures we implemented a method to artificially spike-in differences according to a given distribution of effect sizes. The naive approach, i.e. just shifting the means by the "true" effect size, leads to data that deviates from the pattern seen in real data sets at the low intensity range. The reason is that measurement errors are both additive and multiplicative on the original scale [6]. In the low intensity range the additive error component must not be inflated by the "true" fold change spiked in. We developed a spike-in model correcting roughly for this phenomenon by first estimating the additive component using the RMA background estimation method [7]. We then apply the fold-change only to the background corrected signal and add the background again. (Details on the simulation are described in the diploma thesis of JF, which is available from the authors.)

For simplicity, we assume every probe of a given probe set to have the same effect size. This model seems to be reasonable for the PM probes as they all assess the same transcript sequence but it may not apply to MM probes which are designed to only measure unspecific binding. However, Chudin et al. point out that the MM probes also pick up a specific signal similar to the PM signal albeit not as sensitive to the true transcript abundance [8]. Thus, simulation results may be biased for preprocessing methods that make use of MM signals.

Design of Simulation Study

Based on the simulation procedure described above we simulated 28 gene expression experiments using two original data sets. The larger data set concerning 127 samples of adeno-carcinoma of the lung has been published by Bhattacharjee et al. [9] and is available on the intenet(http://www.camda.duke.edu/camda03/data sets/); the other data set consisting of 15 samples from hot thyroid nodules was published by Eszlinger et al. [10] and was kindly provided by the authors. Both data sets are based on Affymetrix' HG-U95Av2 arrays which contain approximately 400,000 probes in 12,625 probe sets. We used the larger data set to report results in this paper but we found similar results using the smaller set.

In order to check the assumption that the log-transformed raw data is normally distributed we performed probe-wise Shapiro-Wilk tests for normality. The resulting p-values were approximately uniformly distributed in both data sets as expected under the null hypothesis.

The simulated experiments varied in the number of differentially expressed genes, the proportion of up-regulated genes, and the sample size. The simulated experiments were preprocessed in 30 different ways of combining a background correction, a normalization, a PM correction, and a summarization method available within the R-based BioConductor project [2].

For background correction we used the following options: Affymetrix' Microarray Suite (MAS) 5.0 [11], RMA [7], none. For normalization we chose among the possibilities: quantiles [12], constant (MAS 5.0) [11], VSN (13), invariant set [14], none. For PM correction we chose to either use the PM signals only (see [7]) or to compute the ideal mismatches as in MAS 5.0 [11]. The former option was then combined with median-polish [7] and the latter with Tukey's bi-weight algorithm [11]. Currently, these are the two most common summarization methods. Altogether, this results in 30 different preprocessing protocols which is of course only a small yet still manageable fraction of the large number of possible combinations available in BioConductor and elsewhere.

Unless otherwise specified we consider a two-sample comparison situation with 50% of the genes being differentially expressed with additive absolute effect sizes (on the log scale) randomly chosen from a half normal distribution centered at zero.

We then compared resulting probe signals and expression values by applying several criteria namely:

1) log variance ratio to quantify the decrease/increase of precision resulting from background correction and normalization,
2) the slope of the estimated versus true effect size regression line (on the log scale) as a measure for accuracy,
3) mean squared error of the effect size estimates, a statistic that addresses both accuracy and precision,
4) sensitivity and specificity to detect differentially expressed genes using univariate Welch-t-tests and the Benjamini-Hochberg procedure to control the false discovery rate at the 5% level.

We avoid criteria based on pure fold change measures, as they do not account for any statistical uncertainty caused by various sources of variation.

Table 1 Medians of mean squared errors (MSE) of effect size estimates per preprocessing procedure. The median MSEs were averaged over 14 simulated experiments.

Normalization method	No BG correction		RMA BG correction		MAS 5.0 BG correction	
	MPa)	TBIb)	MP	TBI	MP	TBI
Quantiles	0.05	0.07	0.15	0.31	0.16	0.22
Constant	0.17	0.13	0.29	0.27	0.23	0.27
Global Loess	0.05	0.07	0.16	0.35	0.16	0.22
VSN	0.08	0.12	0.23	0.45	0.08	0.11
Invariant Set	0.05	0.06	0.37	0.53	0.16	0.23
No normalization	0.33	0.36	1.36	2.11	0.84	0.89

a) PM only + median polish (RMA summary method)
b) Ideal Mismatch + Tukey Bi-Weight (MAS 5.0 summary method)
BG = background

Results

Any investigated normalization method considerably reduces the variance on the probe level (median up to 8-fold in our data sets). Any background correction method leads to marked variance inflation in the lower intensity range (data not shown).

The slope of the estimated versus true effect size regression line (on the log scale) should be near one with an unbiased estimator. Without background correction all normalization methods lead to biased estimators with slopes in the order of 0.5 where VSN (which includes some background correction in the underlying model) is in the lead (0.69). With either background method the bias is essentially removed (slopes close to 1).

We use the mean squared error $MSE(\theta, \hat{\theta}) = E((\hat{\theta} - \theta)^2) = Var(\hat{\theta}) + (E(\hat{\theta} - \theta))^2$ as criterion to trade off increased variance against decreased bias. Table 1 shows that the MSE is lowest without background correction and some normalization. There are only minor differences between normalization methods and median-polish shows a minor advantage as a summary measure.

Table 2 summarizes the sensitivity of a typical test procedure (Welch t-test and Benjamini-Hochberg procedure to control the false discovery rate at 5%) to detect spiked-in differentially expressed genes. Normalization leads to a sensitivity of up to 59% while only a few genes are detected when no normalization is performed. Background adjustment has a minor influence on sensitivity. Summary measures obtained with median-polish appear to have a somewhat higher sensitivity than those obtained with Tukey bi-weight. These results were obtained with sample size N = 15, but qualitatively appear not to depend on the sample size (range N = 3-30).

Varying the proportion of differentially expressed genes (in the symmetrical setting with 50% up-regulated) we only found a moderate increase in MSE for all normalization methods investigated and no apparent effect on sensitivity.

Our major finding concerns the effect of an increasing asymmetry in the proportion up-regulated among differentially expressed genes. Figure 1 shows that asymmetry shifts the median of the estimated effect sizes in non-differentially expressed genes away from zero thus introducing bias.

This phenomenon compromises specificity and inflates the rate of false positive findings: Figure 2 describes the true false discovery rate in selection lists obtained by the Welch t-test and the Benjamini-Hochberg procedure to control the false discovery rate at a 5% level. Due to the bias, the true false discovery rate is well above the pre-specified level. VSN turned out to be relatively more robust to asymmetry than the other normalization procedures investigated.

Table 2 Sensitivity to detect differentially expressed genes by preprocessing procedure. In this example 50% of the genes were differentially expressed, 50% of which were up-regulated. Each group included 15 arrays. In order to obtain a list of differentially expressed genes we tested the difference in mean log expression against zero for every gene using univariate t-tests. The resulting p-values were then adjusted by the Benjamini-Hochberg procedure and a 5%-level cutoff was applied.

Normalization method	No BG correction		RMA BG correction		MAS 5.0 BG correction	
	MPa)	TBIb)	MP	TBI	MP	TBI
Quantiles	0.59	0.52	0.58	0.52	0.59	0.53
Constant	0.37	0.39	0.53	0.48	0.54	0.48
Global Loess	0.50	0.40	0.52	0.43	0.52	0.44
VSN	0.59	0.52	0.58	0.50	0.59	0.52
Invariant Set	0.59	0.51	0.46	0.44	0.59	0.51
No normalization	0.00	0.00	0.03	0.04	0.10	0.08

a) PM only + median-polish (RMA summary method)
b) Ideal Mismatch + Tukey Bi-Weight (MAS 5.0 summary method)
BG = background

Discussion

We developed and implemented a method to simulate oligo-nucleotide micro-arrays in order to compare preprocessing procedures in contexts which up to now were not yet analyzable. We use a parametric bootstrap approach to sample chips similar to an empirically given raw data set. Chip effects which normalization tries to eliminate induce strong positive correlations between probe intensities. We need not explicitly model the form of these chip effects since the positive correlations present in the raw data in reverse give rise to "pseudo"-chip effects.

In a simulation study, we compared various combinations of background correction, normalization, PM correction and summary methods available within the BioConductor project in the situation where 50% of all genes are differentially expressed.

With symmetry between up and down regulation, we essentially confirm known results on variance and bias from a setting with very low proportion of differentially expressed genes [7, 12, 15] also for our context. In the symmetric case there is no clear winner among normalization methods, but it seems to be advantageous to have no background correction and use a summary measure taking into the account the differences in probe affinity (e.g. median-polish).

A pronounced asymmetry between up and down regulation causes a bias in the effects size estimate of non-differentially expressed genes. This inflates the false positive detection rates. The problem concerns all normalization procedures investigated, but the VSN method appears to be the relatively most robust method. This is not surprising since the VSN model is fitted using least trimmed sum of squares regression based only on the 50% smallest residues; thus normalization is mainly based on non-differentially expressed genes.

We currently investigate using iterative selection of a subset of genes with high probability of not being differential and basing normalization only on this subset as a remedy to this problem. Although the invariant set method [14] did not solve the problem, first results using genes with small variance across samples look promising and will be published elsewhere.

The situation in which many genes are differentially expressed in one direction is by no means artificial. This situation may even be common in certain cell-line experiments. For instance, Lemon et al. [16] stimulated starved fibroblasts and reported massive up-regulation of many genes.

In conclusion, bootstrap simulation can be used to compare preprocessing methods and massive up-regulation as encountered in certain biological experiments poses a problem for currently used normalization methods.

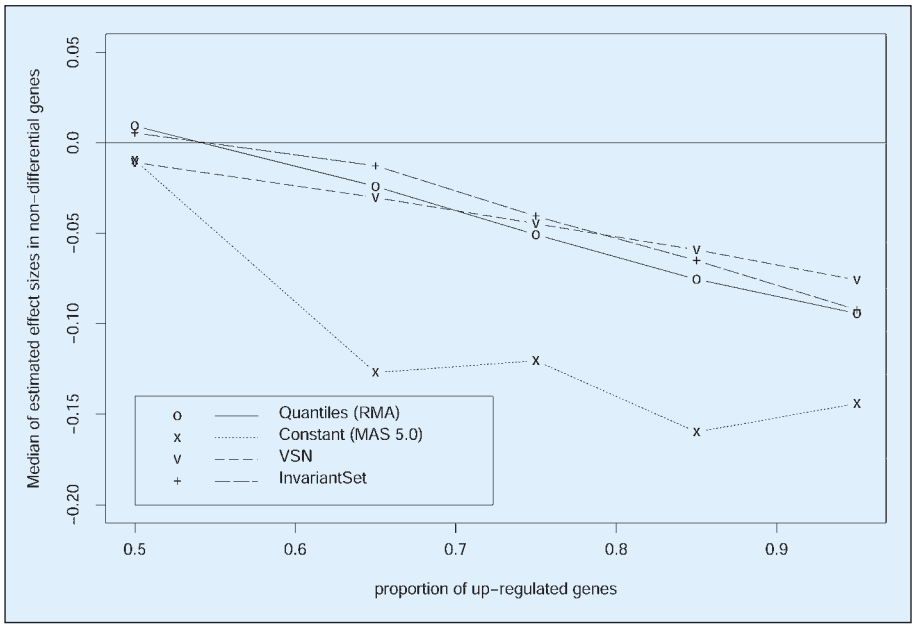

Fig. 1 Median estimated effect sizes in those genes simulated non-differentially vs. the proportion of up-regulated genes per sample. In all cases approximately 50% of the genes were differentially expressed. No background correction, the respective normalization and median-polish was performed. Most normalization methods lead to deviation from zero thus introducing bias.

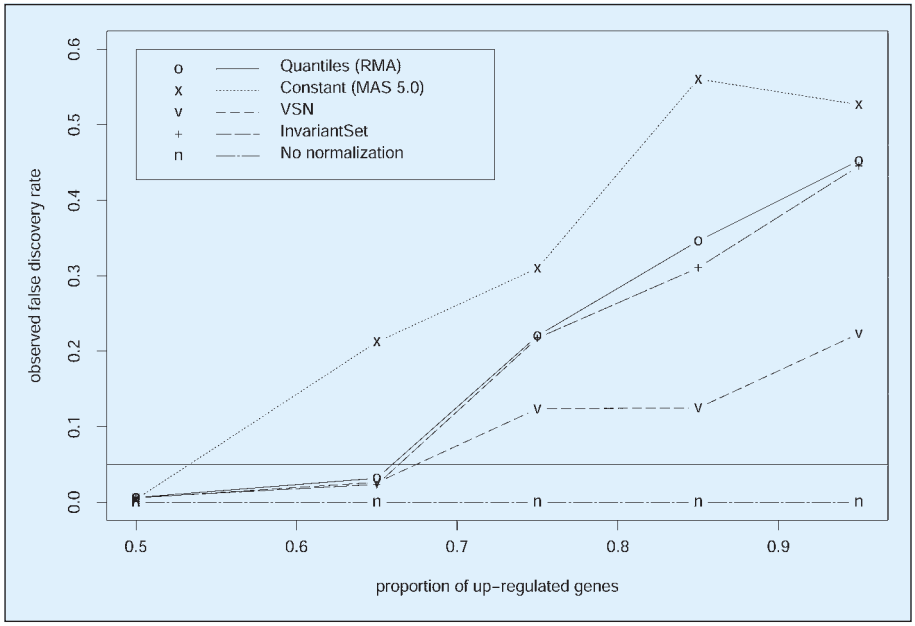

Fig. 2 Observed false discovery rate vs. proportion of up-regulated genes for different normalization methods. The lists of differentially expressed genes were obtained by including all genes having a Benjamini-Hochberg adjusted p-value from a univariate t-test lower than 5%. The investigated normalization methods lead to a failure in controlling the specified false discovery rate if approximately 70% of the genes or more are up-regulated. Not to normalize avoids false positives but does not produce many true positives either.

References

1. Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, Brown EL. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol 1996; 14 (13): 1675-80.
2. Ihaka R, Gentleman RR. A Language for Data Analysis and Graphics. Journal of Computational and Graphical Statistics 1996; 5 (3): 299-314. http://www.r-project.org, http://www.bioconductor.org.
3. Cope LM, Irizarry RA, Jaffeee H, Wu Z, Speed TP. A Benchmark for Affymetrix GeneChip Expression Measures. Bioinformatics 2004; 20 (3): 323-31.
4. Efron B, Tibshirani RJ. An Introduction to the Bootstrap. Chapman & Hall, New York, NY, USA, 1993.
5. van der Laan MJ, Bryan J. Gene expression analysis with the parametric bootstrap. Biostatistics 2001; (4): 445-61.
6. Rocke DM, Durbin B. A Model for Measurement Error for Gene Expression Arrays. Journal of Computational Biology 2001; 8 (6): 557-69.
7. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 2003; 4 (2): 249-64.
8. Chudin E, Walker R, Kosaka A, Wu SX, Rabert D, Chang TK, Kreder DE. Assessment of the relationship between signal intensities and transcript concentration for Affymetrix GeneChip arrays. Genome Biology 2001; 3 (1).
9. Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, Loda M, Weber G, Mark EJ, Lander ES, Wong W, Johnson BE, Golub TR, Sugarbaker DJ, Meyerson M. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc Natl Acad Sci USA 2001; 98 (24): 13790-5. Epub 2001; Nov 13.
10. Eszlinger M, Krohn K, Frenzel R, Kropf S, Tonjes A, Paschke R. Gene expression analysis reveals evidence for inactivation of the TGF-beta signaling cascade in autonomously functioning thyroid nodules. Oncogene 2004; 23 (3): 795-804.
11. Affymetrix. Statistical Algorithms Description Document. Affymetrix, Inc., Santa Clara, CA, 2002. http://www.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf.
12. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 2003; 19 (2): 185-93.
13. Huber W, von Heydebreck A, Sültmann H, Poustka A, Vingron M. Parameter estimation for the calibration and variance stabilization of microarray data. Statistical Applications in Genetics and Molecular Biology 2003; 2 (1).
14. Li C, Wong WH. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. Genome Biology 2001; 2 (8).
15. Bolstad BM. Comparing the effects of background, normalization and summarization on gene expression estimates. http://www.stat.berkeley.edu/users/bolstad/stuff/components.pdf. Unpublished manuscript 2002.
16. Lemon WJ, Palatini JJT, Krahe R, Wright FA. Theoretical and experimental comparisons of gene expression indexes for oligonucleotide arrays. Bioinformatics 2002; 18 (11): 1470-6.

Correspondence to:
Dr. Dirk Hasenclever
Institute of Medical Informatics, Statistics and Epidemiology (IMISE)
University of Leipzig
Liebigstr. 27
04103 Leipzig
Germany
E-mail: Dirk.Hasenclever@IMISE.uni-Leipzig.de