Deposited research article

Characterization of the expression ratio noise structure in high-density oligonucleotide arrays

Felix Naef[1], Coleen R. Hacker[2], Nila Patil[2], and Marcelo Magnasco[1]

Addresses: [1] Mathematical physics laboratory, Center for Studies in Physics and Biology, The Rockefeller University, 1230 York Ave, NY 10021. [2]Perlegen Sciences, Inc., 2021 Stierlin Court, Mountain View, CA 94043.

Correspondence: Felix Naef. E-mail: felix@funes.rockefeller.edu

Posted: 7 December 2001

Genome ***Biology*** 2001, **3(1)**:preprint0001.1-0001.24

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2001/3/1/preprint/0001

Received: 3 December 2001

This is the first version of this article to be made available publicly. This article has been submitted to *Genome* ***Biology*** for peer review.

Characterization of the expression ratio noise structure in high-density oligonucleotide arrays

Felix Naef[1], Coleen R. Hacker[2], Nila Patil[2], and Marcelo Magnasco[1]

[1] Mathematical physics laboratory, Center for Studies in Physics and Biology, The Rockefeller University, 1230 York Ave, NY 10021.

Tel: (212) 327 8186, fax: (212) 327 8889

[2] Perlegen Sciences, Inc., 2021 Stierlin Court, Mountain View, CA 94043.

emails:

felix@funes.rockefeller.edu

chacker@perlegen.com

npatil@perlegen.com

marcelo@zahir.rockefeller.edu

Running title: Noise structure in GeneChip arrays

Abbreviations: perfect match (PM); single mismatch (MM); log-ratio (LR); standard error (SE); signal to noise (SN); least trimmed squares (LTS).

Abstract

Background

High-density oligonucleotide microarrays provide a powerful tool for assessing differential mRNA expression levels. Characterizing the noise resulting from the enzymatic and hybridization steps, called type I noise, is essential for attributing significance measures to the differential expression scores. We introduce scoring functions for expression ratios, and associated quality measures. Both the PM (Perfect Match) probes and PM-MM differentials (MM is the single MisMatch) are considered as raw intensities. We then characterize the log-ratio noise structure using robust estimates of their intensity dependent variance.

Results

We show the relationships between the obtained ratios and their quality measures. The complementarity of PM and PM-MM methods is emphasized by the probe sets signal to noise measures. Using a large set of replicate experiments, we demonstrate that the noise structure in the log-ratios very closely follows a local log-normal distribution for both the PM and PM-MM cases. Therefore, significance relative to the type I noise can be quantified reliably using the local STD. We discuss the intensity dependence of the STD and show that ratio scores >1.25 are significant in the mid- to high-intensity range.

Conclusions

The ratio noise structure inherent to high-density oligonucleotide arrays can be well described in terms of local log-normal ratio distributions with characteristic intensity

dependence. Therefore, robust estimates of the local STD of these distributions provide a simple and powerful way for assessing significance (relative to type I noise) in differential gene expression. This approach will be helpful for improving the reliability of predictions from hybridization experiments in general.

Background

As DNA microarray experiments offer a huge potential for screening possibly relevant genes, researchers often face the question of whether a particularly interesting 'hit' should be trusted or not. Considering the relative noisiness of the available technologies, this concern is more than justified and it is therefore of the utmost importance to develop finer analysis methods that allow for a better control over the noise levels inherent in the hybridization process.

In this work, we shall investigate this problem in detail in the context of GeneChip® experiments, keeping in mind that most of the techniques are also applicable to other microarray technologies. Among the different technologies, the particularity of GeneChip arrays is that each transcript is probed by many short length sequence snippets, instead of a single longer probe as in cDNA arrays [1-3]. Therefore, translating the measured probe intensities into a global gene intensity or ratio score requires a composite scoring function. In principle, the redundancy in the probes offers a way to reduce the noise level for each gene; on the other hand, finding the best estimator is a difficult problem and it is likely that the variability in the probe behavior will prevent a single estimator from being optimal in all cases. Studying the raw data reveals its great complexity and the hybridization processes underlying the measured PM vs. MM intensities proves to be hard to interpret physically. The principal difficulties stem from (i) the large number of probes with MM intensities higher than the corresponding PM, and therefore not conforming to the usual hybridization picture; (ii) the very broad intensity distributions within each probe set. The first task is therefore to design methods that can robustly handle such input data.

Once reliable measures for differential expression from two experiments are obtained, one would like to attribute significance to them. It is by now widely accepted that such discussions should be carried out in an intensity dependent fashion. There are at least two independent sources of variability to be considered: (i) the intrinsic noise levels related to the technology (type I noise); (ii) the biological variability encountered biological replicates. Previous studies have addressed significance issues in both cDNA and GeneChip arrays; they either discuss significance according to (i) [4, 5]; focus on (ii) only [6, 7]; or consider both (i) and (ii) simultaneously [8].

High-density oligonucleotide expression is based on the synthesis of oligonucleotides by photolithographic techniques. As hybridization is not expected to be specific enough for oligonucleotide sequences ≤ 25 bases in length, GeneChip uses (i) 14-20 different sequence snippets to probe each transcript; (ii) each snippet comes in two versions: the perfect match (PM) probe and the single mismatch (MM) whose middle base has been substituted to its complement and is designed to serve as a control for non-specificity. The pairs (PM, MM) are called probe pairs and the full set of probes for a given gene is called a probe set. The standard picture used to interpret the hybridization is based on the following model [2, 9, 10]:

$$PM = I_S + \beta I_{NS} + B$$
$$MM = (1-\alpha) I_S + \beta' I_{NS} + B$$

where *PM (MM) i*s the observed brightness, I_S the contribution from specific complementary binding, I_{NS} the amount from non-specific binding, and B a background of physical origin, i.e. the photodetector dark current or light reflections from the scanning process. The proper technique for estimating the background and its fluctuations

is discussed in [11]. Then, α (thought to be positive) reflects the loss of binding due to a single substitution, and β, β' are the susceptibilities for non-specific binding. In the ideal case, which is usually assumed, $\beta=\beta$' and therefore the subtraction $PM - MM = (1-\alpha)I_S$ is directly proportional to the desired signal. However, the susceptibility α can be strongly varying within a given probe set.

We focus at two different aspects of GeneChip data analysis. First, we describe a method to evaluate ratio scores –and associated quality measures. We shall relax the assumption that $\beta=\beta$' and consider either just the PM probes, or the usual PM-MM subtraction. These cases bracket the extremes of an ideally performing MM probe to a poor MM.

Second, we address significance measures relative to type I noise in the context of GeneChip ratio measurements. Our approach shares similarity to [8] and more remotely to [5]; both of these were developed in the context of cDNA arrays. The precise differences between our approach and that in [8] are several. First, we found no evidence for the inclusion of an additive term in linear coordinates. Judging by the data, the noise structure is very well captured by an effective multiplicative model (cf. Results). Second, the multiplicative noise component is estimated in logarithmic coordinates instead of linear, and after a local normalization. Finally, we estimate the local scale in the noise by an empirical robust fit of the local variance, with no a priori model. While it would be satisfactory to have a physical model describing the noise, our experience is that is it very hard to formulating one that accounts for the observed structure in all cases.

Results

There are two classes of approaches to composite scores: (i) the 'direct' methods [2, 11] aim at obtaining scores based on the distribution study of the probe intensities and can be

applied to a single experiment (or pair in the case of ratio estimators), and (ii) model based approaches [9, 10] that attempt to fit the susceptibilities α and β from a collection a arrays. A major difference in the two methods is that the number of experiments needed can be just one (or two for ratios) in the first case, while a large number (ideally as many as the number of probe pairs for the reduced model [9]) of homogenous quality arrays is required for the second. Model based approaches seem to be well suited for absolute intensity measures, on the other hand direct methods are not expected to be very robust for absolute intensities, due to the broad intensity distributions within probe sets [11]. However, the situation is different when considering ratios scores, as the variability in the individual probe susceptibilities cancels out when considering the expression ratios at the probe level. Ratios scores should therefore be expected to be more robust; more precisely, when comparing two conditions, we consider each probe to provide an 'independent' measurement of the real expression ratio. Our composite ratio estimation is based on a standard least trimmed square (LTS) estimator of the set of log-ratios for each probe set (details in Material and Methods). We consider the cases for which the probe log-ratios (LR) are derived either from the usual PM-MM subtracted intensities, or from the PM values alone (after background subtraction). The rationale behind proposing a variant without the subtraction lies in the empirical observation that 30% of the total probe pairs (PM, MM) are such that MM>PM. This is true for all chip series tested (mouse, human, drosophila), except for yeast, which happens to be significantly better (15%). In such probe pairs, targets systematically bind stronger to the MM, hence violating the prediction from the above model. Importantly, these 'misbehaving' pairs are not the matter of a few noisy probe sets, but are fairly homogenously distributed among the probe sets, occurring also in sets with overall high intensities [12].

Relations between 'single-gene' measurements

Figures 1 and 2 show the relationships between the ratio-related measurements, for both the PM and PM-MM cases. A collection of 12 Mu11K chips hybridized to mRNA extracts from different mouse brain regions was chosen to maximize the dynamic range of the ratios, in total, ~120,000 ratios are plotted. The arrays were normalized as explained in the Material and Methods section. The main observation is that the three quantities: the log-ratio LR, the standard error SE, and the Wilcoxon rank sign test p-value show little correlation and represent therefore relatively 'independent' indicators (see Material and Methods for the precise definitions). The lines indicating a signal-to-noise ratio of 1 clearly show that the vast majority of the measurements are well defined, especially for the larger log-ratios. The behavior of SE for small N_{good} (the colored dots) is well understood in terms of the number of residuals considered in the LTS method, i.e. genes with $N_{good}=1$ necessarily have SE=0. The only obvious correlation is that small p-values are not compatible with LR=0, as shown by the valley along LR=0 in the contour plots of LR vs. p. It is however possible to achieve very small p-values for tiny fold changes as small as 1.1. The PM and PM-MM methods show overall very similar features, the biggest difference being that p-values tend to be larger for large LRs in the PM-MM, which reflects the overall smaller number of N_{good} probes usable in that method (there are more probes lying below background after the subtraction).

The difference in the log-ratios from the PM and PM-MM methods is illustrated in Figure 3. Although there is a branch with LR~0 from the PM method when all p-values are considered, this branch rapidly disappears when focusing at smaller p-values only. Cases where the PM and PM-MM indicate regulation in opposite directions are virtually absent

when p<0.05. However, one can see a 'compression effect' in the scores from the PM method, shown by the edges with slope >1 near the regions indicated by the arrows in the bottom right panel (this can also be seen by comparing the upper two contour plots of Figures 1 and 2). This compression likely reflects cross hybridization effects that are not corrected for in the PM only method. However, if one is interested in finding significant changes more than in the LR values themselves, the determining quantity is the LR value divided by the width of the local noise. Therefore, compression in the scale is not dramatic as long as the noise envelope also shrinks (cf. Discussion).

A comparison of the LR/SE (signal-to-noise ratio SN) from both methods emphasizes the complementarity of both methods (Figure 4), as there are clearly a similar fraction of genes that have poor SN ratio (SN < 2) in one case but acceptable in the other. These are found in the top left and bottom right quadrants defined by the horizontal and vertical blue lines.

Noise structure

In Figure 5, we show generic scatterplots for increasing levels of overall differential regulation, from duplicates to strong regulation (left to right panel). We shall always refer to duplicates as experiments where the enzymatic steps in the target sample preparation have been performed independently. In idealistic terms, the scatter cloud is thought of as consisting of two components: one that is just noise from the enzymatic and hybridization steps affecting all the genes; and a component that reflects true sample differences. For illustration, we show in Figure 5 three prototypical cases; there, the second component increases gradually from zero (leftmost plot, duplicates), as visible in the quantile-quantile (QQ) plots in panel B. The nearly straight leftmost QQ plot indicates that

variable $LR/\sigma(\eta)$ follows closely a normal distribution. As the amount of true regulation increases, longer tails develop that depart from normal behavior.

The fact that the noise component behaves as local log-normal distributions is not dictated by the choice of the variable $LR/\sigma(\eta)$, on the contrary, it emerges as a rather pleasant feature of GeneChip experiments. In Figure 6, we demonstrate how systematic this log-normality occurs: we show the two best and two worst from 40 independent human HG-U95A duplicates collected in a study of rhumatoid arthritis. The majority of the cases look closer to the first two examples; and the PM and PM-MM methods lead overall to equally good log-normal distributions. The mean and variation in the local STD for both the PM and PM-MM methods are shown in Figure 7 and reflect the characteristic contraction of the noise envelope with increasing coordinate along the diagonal. It is obvious that the PM noise envelope is thinner than that of the PM-MM at low intensity; on the other hand, both methods lead to comparable local σ in the 'flat' mid to high-intensity domains. There, σ is approximately 0.15 such that $2*\sigma$ corresponds to a ratio of ~1.25. This in turn means that ~95% of the measurements in the mid- to high-intensity range are reproducible within a factor of 1.25.

Discussion

Our experience is that despite constant improvements, current incarnations of the arrays still behave fairly inhomogenously as far as their PM and MM hybridization properties are concerned. This is likely the consequence of various sequence dependent effects: (i) the difference in stacking energies of single-stranded snippets between the PM and the MM sequences can easily be in the range of the gain in binding energies; (ii) there are certainly kinetic effects as the hybridizations are not carried to complete equilibrium; and

(iii) there is always the possibility of sequence dependent synthesis efficiencies. The wide range of probe set behavior is best seen in the SN ratios in Figure 4. For these reasons, we believe that currently the most reliable approach consists in integrating the results from both the PM only and the PM-MM methods. For instance, considering the intersection (or union) of genes predicted by either method would minimize the false positive (or negative) rates. In addition, there seems to be a significant variation in the hybridization properties across different chip series, as can be observed from a simple statistics on the number of probe pairs with MM>PM (cf. Results). The mentioned superiority of the yeast chip may of course be related to the relative simplicity of the yeast genome as compared to that of the higher organisms.

Another point worth mentioning is that the values of the ratios scores are not expected to necessarily correlate well with the real mRNA concentration ratios, due to various effects such as non-linearities in the probe binding affinities. This emphasizes the importance of measuring differential expression relative to the local noise; only then can we decide whether a given ratio score is to be considered as potentially real or not.

Finally, the question of handling significance across replicated experiments comes as a second step to be built on top of the analysis presented above. The most reasonable approach would be to follow [7], namely to consider the t-statistics of the expression ratios across the samples. However, one would also want to weigh the average according to the noise content in each of the samples, in a manner similar to that discussed in [8].

Materials and methods

Regression of the log-ratios, SE

This section explains how we compute expression ratios, for genes measured in two separate arrays. Let (x_i, y_i) denote the brightness measurements for one probe set (the index i ranges from 1 to the number of probe pairs for the particular probe set, 14-20 depending on the chip series) taken in two different hybridization arrays X and Y. We investigate both cases where the intensities (x_i, y_i) are either the intensities of the background subtracted PM cells, or the PM-MM values (which need no background correction). Only N_{good} 'good' probe pairs are retained for determining the ratio and associated quantities. We discard probes that are saturated in both X and Y, or probe pairs such that PM-MM < $3\sqrt{2}\,\sigma$ or PM < 3σ in both X or Y. Here, σ corresponds to the standard deviation of the fluctuations in the background intensity. Not considering such probes prevents contamination of the ratio estimates from noisy low-intensity probes. After identifying the probe pairs allowed into the analysis, the differential expression score LR for the gene in question is obtained from a LTS robust regression of $LR_i = \log_2(x_i/y_i)$ to an intercept α=LR. LTS regression corresponds to minimizing $\chi^{LTS} = \frac{1}{Ns}\sum_{i=1}^{Ns}(LR_i - \alpha)^2$: the sum of the Ns smallest squared residuals [13, 14]. We used the default Ns = $[N_{good}/2]+1$ and this parameter can be adjusted in our scripts; however, we found no evidence for changing the default. An estimate of the standard error (SE) for α is given by $SE = \sqrt{\chi^{LTS}}$. Composite absolute intensities for the gene in each experiment can be obtained via geometric means of the (x_i, y_i) probes kept

in the LTS regression, however, these are only indicative measures as the method was designed primarily for expression ratios.

Wilcoxon statistics, number of cells used

In addition to the SE reporting a quantitative estimate of the error in the log-ratio measurement, it is also instructive to report a p-value from a paired Wilcoxon rank sign test of the LR_i values. Casually speaking, this value is related to the portion of the probes indicating regulation in the same direction: the theoretical minimum p-value p_{min} is achieved when all probe ratios agree on the same direction of regulation. Moreover, the test is non-parametric since operating in rank space, and p therefore also incorporates information about the number of probes pairs used N_{good}. Namely, the Wilcoxon p-value has a lower bound that decreases with increasing sample size. For instance p_{min}=1/4 for N_{good}=3 and 1/8 for N_{good}=4, so that small p scores can only be reached when enough probes are used. However, the converse is not true as a gene that is not differentially expressed can have a p-value that is close to 1 even if all the probe pairs are 'good' in the above sense.

Our method does not primarily aim at quantifying the presence or absence of a gene in a particular sample. Nevertheless, we report the number N_{above} of probes with intensity larger than $3*\sigma_{eff}$ ($\sigma_{eff} = \sqrt{2} *\sigma$ for the PM-MM case) for both samples X and Y. Using enough data, one could compute a probability of presence depending on N_{above}, and it is likely that this calibration would be dependent on the chip series.

Normalization and noise characterization

The measures described above are all single-gene properties; they can be computed when given just the intensities gathered for a single gene. In contrast, correcting for systematic

trends (also called 'data massage') and more important, classifying expression ratio according to their significance requires measures that involve the entire gene population on the arrays. We stress that these techniques are meaningful only when the number of genes probed is sufficiently large, and under the assumption that a large fraction of them do not show differential regulation between the two tested samples. These requirements are usually met in GeneChip experiments. Normalization aims at correcting for systematic trends (i.e. due to dye efficiencies and amplification, sample concentration, photo-detector efficiency) so as to make a collection of arrays directly comparable. One must distinguish global from local normalization, in the first, the intensities of all the probes on the array are scaled by a constant factor; in the second, the normalization factor can be intensity dependent. Local normalization techniques are mostly discussed in the context of cDNA arrays [15, 16], where the intensity dependence can be severe. High-density oligonucleotide arrays suffer less from 'bent' noise structures; nevertheless, local normalization has also been introduced for them [17, 18]. Although attractive, local normalization should not be applied blindly as it can hide real failures in the data and create its own artifacts. Our approach to normalization is based on centering the log-ratio distribution either globally, or locally as in [15]. We always normalize an array with respect to another one, and we found it more accurate to do so at the gene rather than at the probe level (we normalize the scores a posteriori).

Turning to the noise structure, significance of regulation is quantified from a local robust regression $\phi(\eta)$ of the variable LR^2 versus η, where $LR = \log(I_X/I_Y)$ and $\eta = \log(I_X*I_Y)/2$. I_X and I_Y denote the intensities of the genes in channel X and Y. The local STD is then given by $\sigma(\eta) = \sqrt{\phi(\eta)}$. We used the R routine loess for the fitting [14]. The idea is then that if the variable $LR/\sigma(\eta)$ follows a good Normal distribution in the case of

replicate (pure noise) experiments, it can be used reliably for assessing the significance of a ratio score.

Scripts will be made available at http://asterion.rockefeller.edu/felix/Affyscripts.

References

1. Chee M, Yang R, Hubbell E, Berno A, Huang XC, Stern D, Winkler J, Lockhart DJ, Morris MS ,Fodor SP: **Accessing genetic information with high-density DNA arrays**. Science 1996, **274**:610-614.
2. Lipshutz RJ, Fodor SP, Gingeras TR ,Lockhart DJ: **High density synthetic oligonucleotide arrays**. Nat Genet 1999, **21**:20-24.
3. Lockhart DJ ,Winzeler EA: **Genomics, gene expression and DNA arrays**. Nature 2000, **405**:827-836.
4. Kerr MK, Martin M ,Churchill GA: **Analysis of variance for gene expression microarray data**. J Comput Biol 2000, **7**:819-837.
5. Newton MA, Kendziorski CM, Richmond CS, Blattner FR ,Tsui KW: **On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data**. J Comput Biol 2001, **8**:37-52.
6. Tusher VG, Tibshirani R ,Chu G: **Significance analysis of microarrays applied to the ionizing radiation response**. Proc Natl Acad Sci U S A 2001, **98**:5116-5121.
7. **Speed Group Microarray Page**: [http://www.stat.berkeley.edu/users/terry/zarray/Html/matt.html].
8. Hughes TR, et al.: **Functional discovery via a compendium of expression profiles**. Cell 2000, **102**:109-126.
9. Li C ,Wong WH: **Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection**. Proc Natl Acad Sci U S A 2001, **98**:31-36.
10. Li C ,Hung Wong W: **Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application**. Genome Biol 2001, **2**:RESEARCH0032.
11. Naef F, Lim DA, Patil N ,Magnasco M: **From features to expression: high-density oligonucleotide arrays analysis revisted**. Proceedings of the DIMACS Workshop on Analysis of Gene Expression Data, October 24 - 26, 2001 2001.
12. **arXiv.org e-Print archive**:[http://xxx.lanl.gov/abs/physics/0111199].
13. Rousseew PJ ,M. LA: **Robust Regression and Outlier Detection**. Wiley 1987.
14. Ihaka R ,Gentleman R: **R: A language for data analysis and graphics**. Journal of Computational and Graphical Statistics 1996, **5**:299-314.
15. Tseng GC, Oh MK, Rohlin L, Liao JC ,Wong WH: **Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects**. Nucleic Acids Res 2001, **29**:2549-2557.

16. **Speed Group Microarray Page**:
[http://oz.berkeley.edu/users/terry/zarray/Html/normspie.html].
17. Schadt EE, Li C, Su C ,Wong WH: **Analyzing high-density oligonucleotide gene expression array data**. J Cell Biochem 2000, **80**:192-202.
18. Hill AA, Brown EL, Whitley MZ, Tucker-Kellog G, Hunter CP ,Slonim DK: **Evaluation of normalization procedures for oligonucleotide array data based on spiked cRNA controls**. Genome Biol 2001, **2**:RESEARCH0055.

Figures

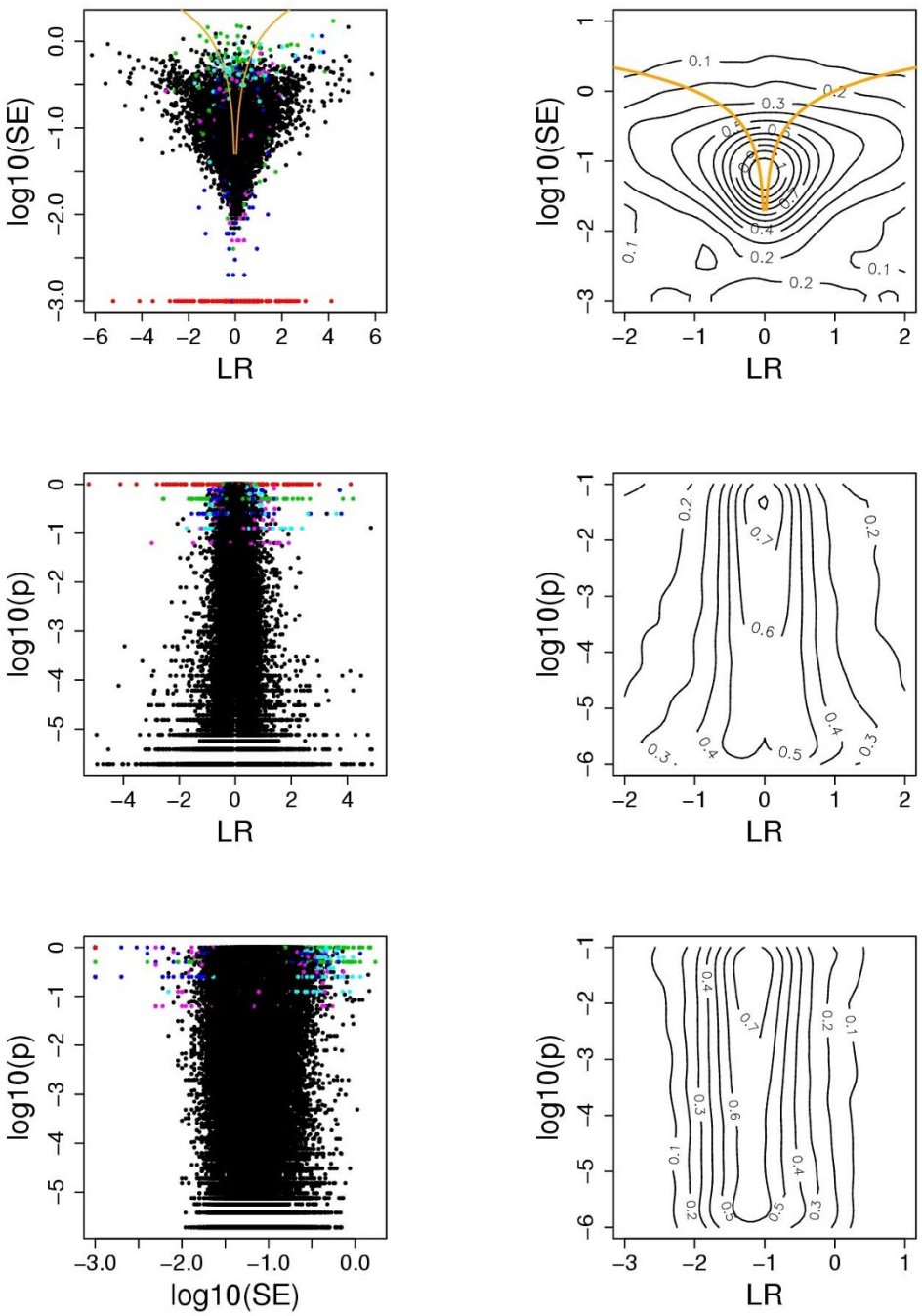

Figure 1. Relationships between the LR, SE and p measures for the PM only scores. The right panels show contour plot of the densities raised to the power 0.2, so as to resolve structure in the low-density

regions. The red, green, blue, cyan and magenta dots refer to N=1, 2, 3, 4, 5 'good' cell pairs. Black dots have >6 good cell pairs. The orange line in the top left panel represent |LR|=SE, so that points below the line have signal-to-noise ratios > 1. The SE for the red dots (N=1) is artificially set to 0.001 for plotting purposes; it should be strictly 0.

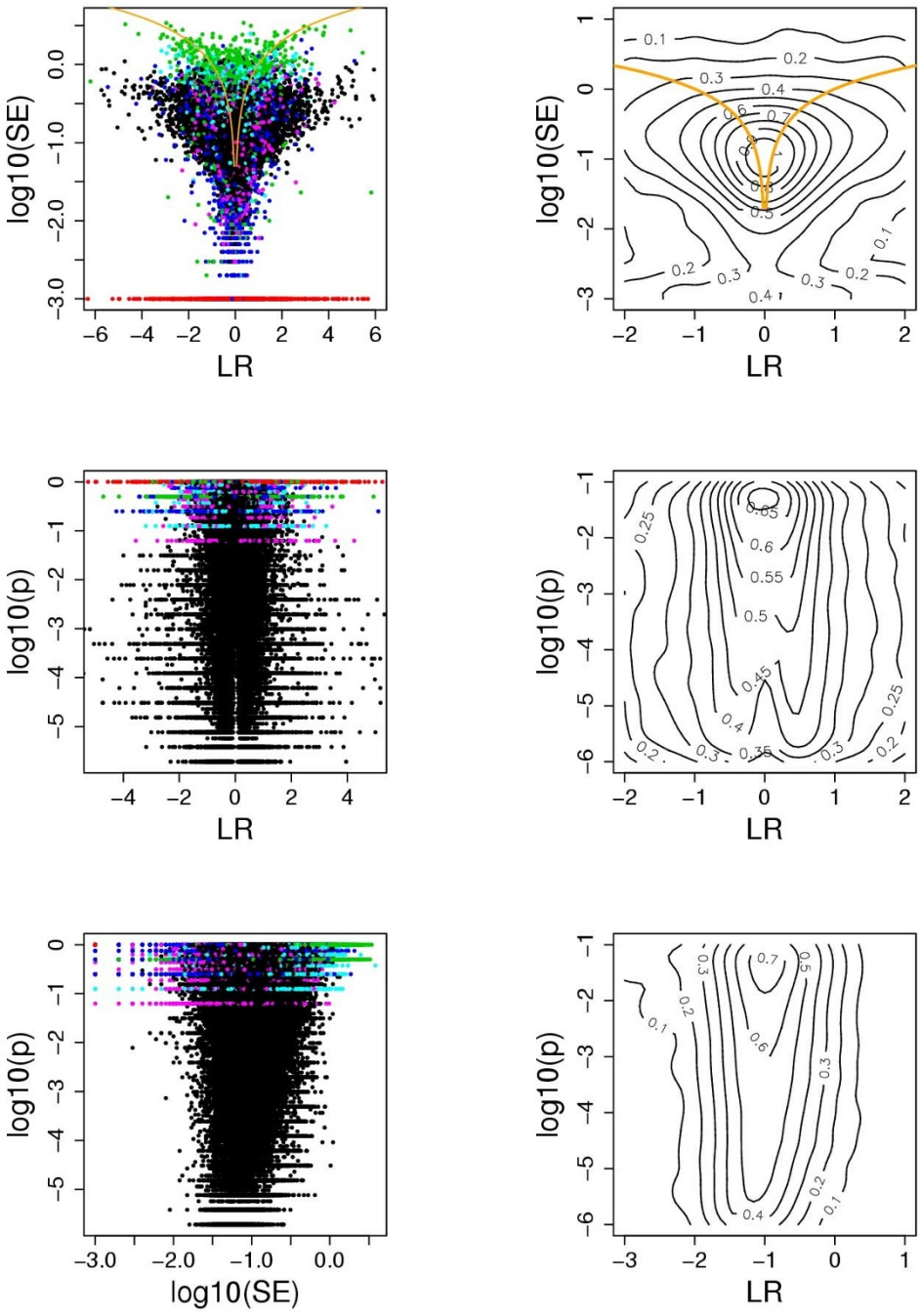

Figure 2. Same as Figure 1 but for the PM-MM method.

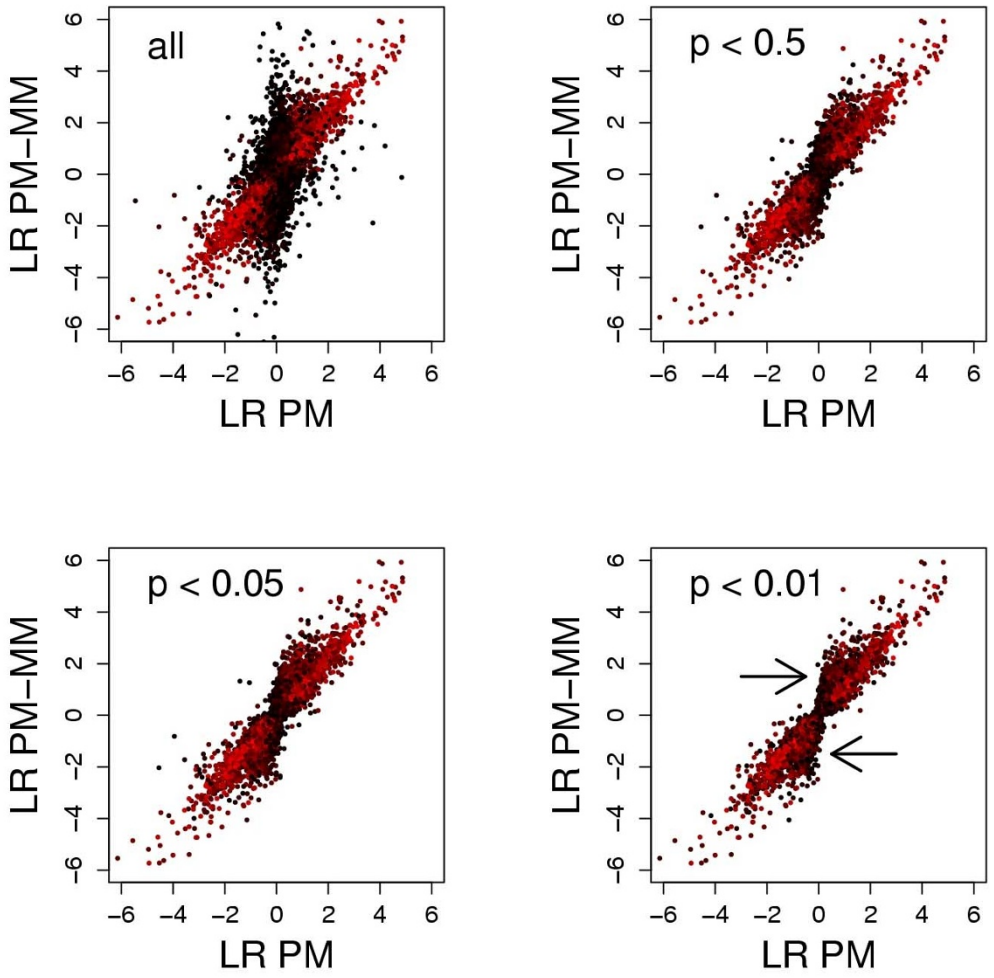

Figure 3. Comparison between the ratios scores from the PM only and the PM-MM methods. Only genes with p-values (from the PM-MM method) smaller than indicated are plotted. Brightest red dots have the smallest p-values (~10^{-6}) whereas black correspond to p=1 (upper left) to p=0.01 (lower right).

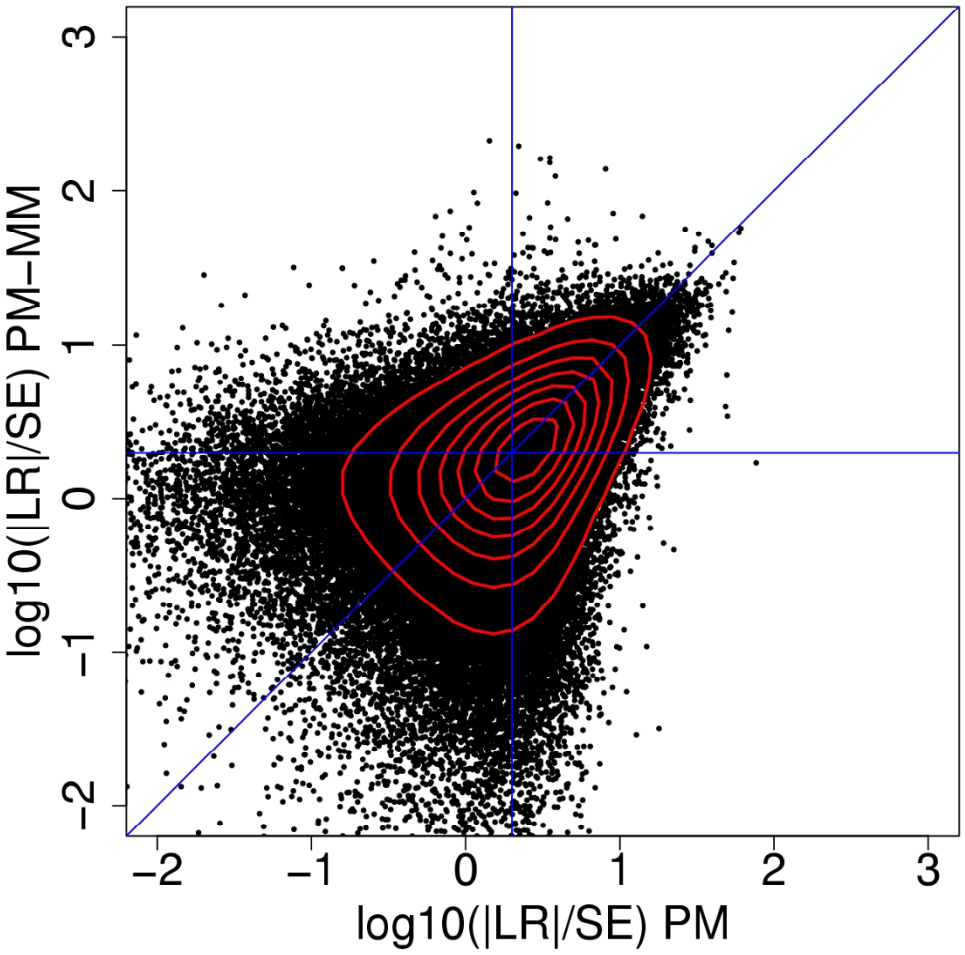

Figure 4. Comparison of the signal-to-noise ratios (SN) for the PM and PM-MM methods. Only genes with Ngood > 4 in the PM-MM method are plotted. The red lines are contour lines and the blue lines denote the diagonal and SN values of 2 on either axis.

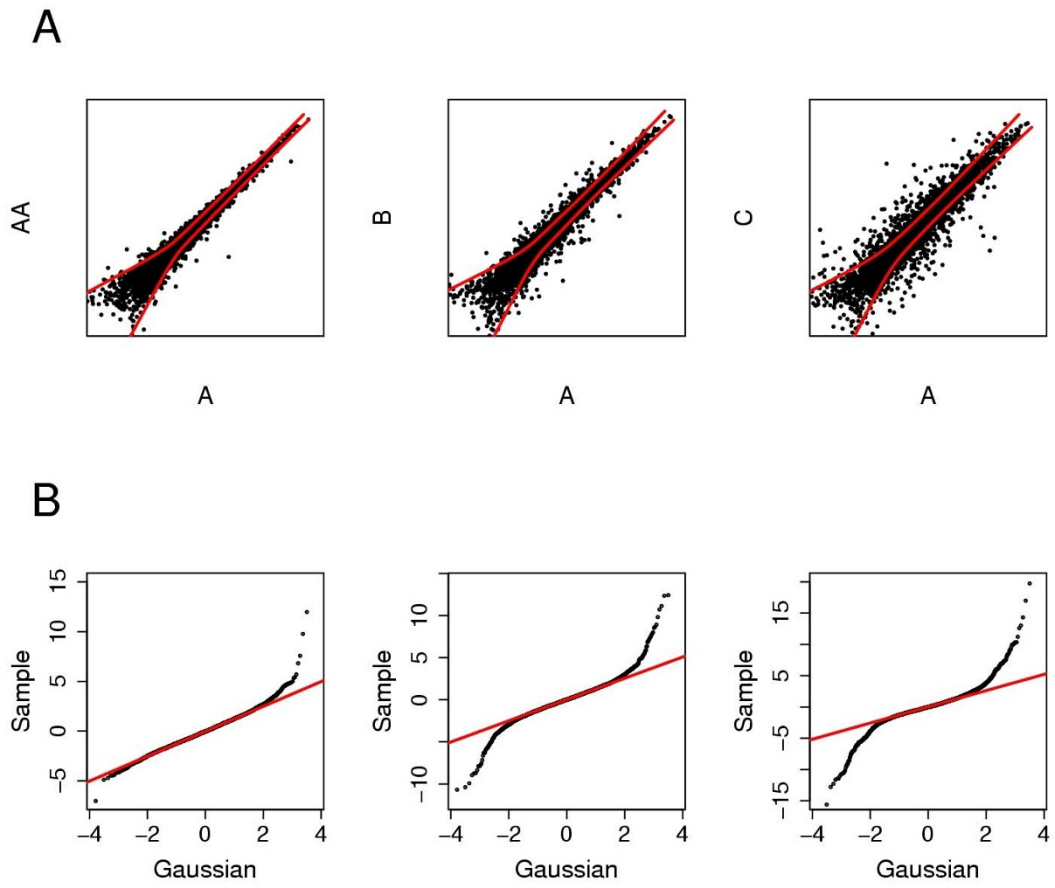

Figure 5. (A) Scatterplot showing increasing levels of differential regulation, together with the 2*σ noise envelopes. There are 3 different conditions A, B and C. AA is a duplicate of A. (B) Associated QQ-plots of the LR/σ(η).

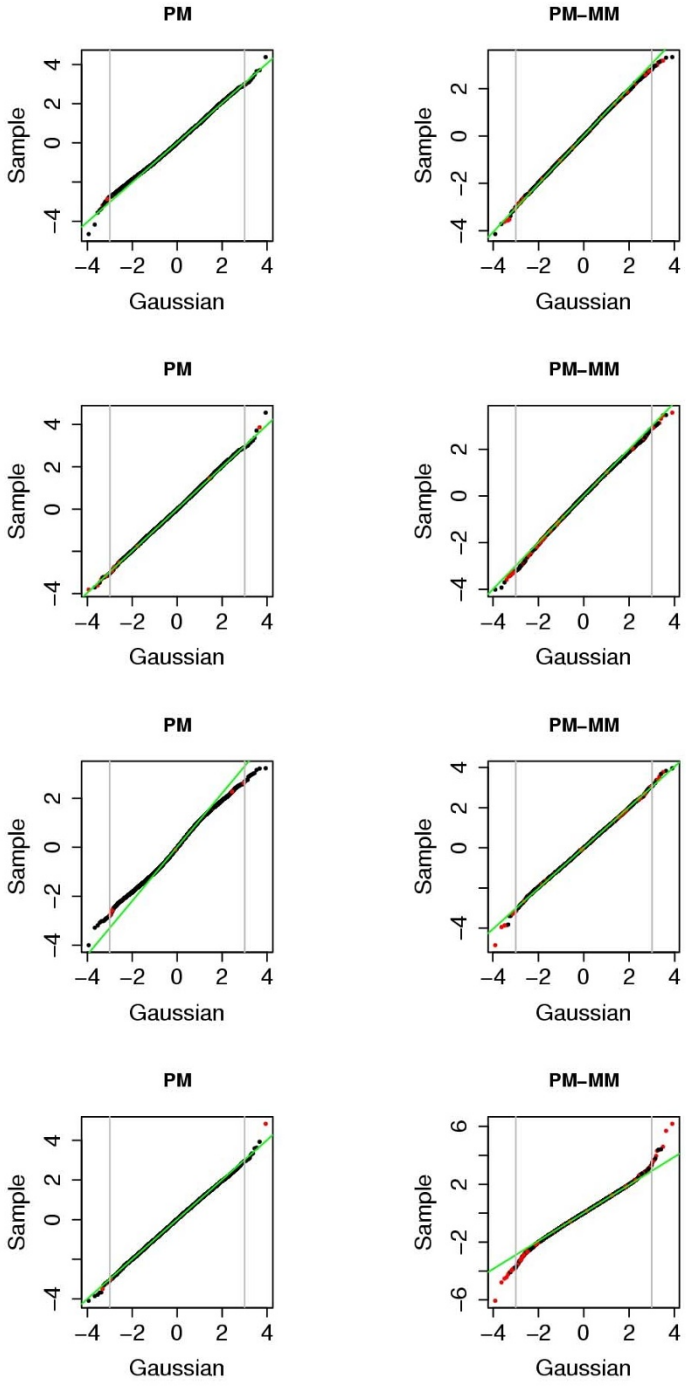

Figure 6. Quantile-quantile plots of LR/σ(η) vs. Normal distributions in replicate HG-U95A experiments. The examples shown are the two best and two worst of 40 independent replicates. The region in-between the gray lines contains 99.9% of the data. Red dots represent ratios with Ngood < 5.

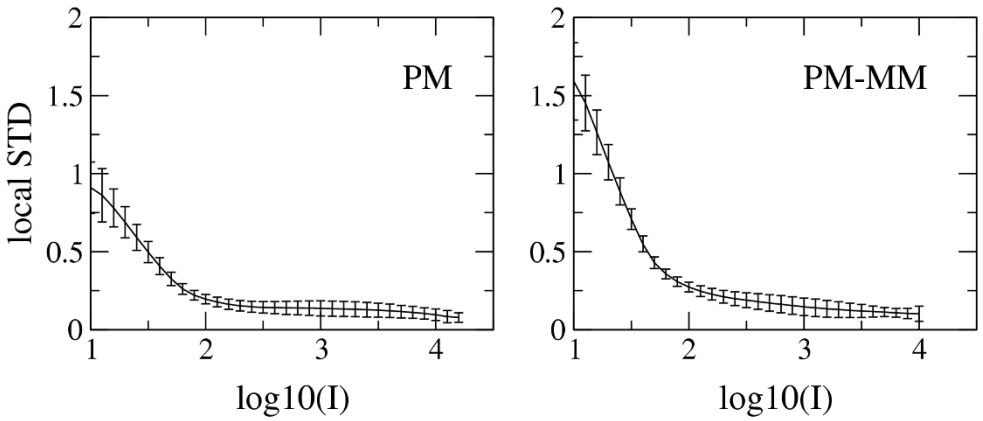

Figure 7. Intensity dependence of the noise envelope: STD of the log2-ratios from 40 pairs of replicate HG-U95A experiments. The errors represent the variation in local STD from the 40 replicates. The intensity $I=(I_X*I_Y)^{1/2}$ measures the coordinate along the diagonal of the scatterplots.