BIOINFORMATICS ORIGINAL PAPER

Vol. 23 no. 21 2007, pages 2897–2902
doi:10.1093/bioinformatics/btm478

Gene expression

I/NI-calls for the exclusion of non-informative genes: a highly effective filtering tool for microarray data

Willem Talloen[1,*,†], Djork-Arné Clevert[2,3,†], Sepp Hochreiter[2], Dhammika Amaratunga[4], Luc Bijnens[1], Stefan Kass[1] and Hinrich W.H. Göhlmann[1]

[1]Johnson & Johnson Pharmaceutical Research & Development, a division of Janssen Pharmaceutica n.v., Beerse, Belgium, [2]Institute of Bioinformatics, Johannes Kepler Universität Linz 4040 Linz, Austria, [3]Department of Nephrology and Internal Intensive Care, Charité University Medicine, Berlin, Germany and [4]Johnson & Johnson Pharmaceutical Research & Development, Raritan, USA

Received on April 13, 2007; revised on September 7, 2007; accepted on September 18, 2007

Advance Access publication October 5, 2007

Associate Editor: David Rocke

ABSTRACT

Motivation: DNA microarray technology typically generates many measurements of which only a relatively small subset is informative for the interpretation of the experiment. To avoid false positive results, it is therefore critical to select the informative genes from the large noisy data before the actual analysis. Most currently available filtering techniques are supervised and therefore suffer from a potential risk of overfitting. The unsupervised filtering techniques, on the other hand, are either not very efficient or too stringent as they may mix up signal with noise. We propose to use the multiple probes measuring the same target mRNA as repeated measures to quantify the signal-to-noise ratio of that specific probe set. A Bayesian factor analysis with specifically chosen prior settings, which models this probe level information, is providing an objective feature filtering technique, named informative/non-informative calls (I/NI calls).

Results: Based on 30 real-life data sets (including various human, rat, mice and Arabidopsis studies) and a spiked-in data set, it is shown that I/NI calls is highly effective, with exclusion rates ranging from 70% to 99%. Consequently, it offers a critical solution to the curse of high-dimensionality in the analysis of microarray data.

Availability: This filtering approach is publicly available as a function implemented in the R package FARMS (www.bioinf.jku.at/software/farms/farms.html).

Contact: wtalloen@prdbe.jnj.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

High-density oligonucleotide microarrays, and in particular Affymetrix GeneChip arrays (Lockhart *et al*., 1996), are now fruitfully being used in many areas of biomedical research. The wealth of information generated by this DNA microarray technology is key to its power and success, but also constitutes its major weakness. The large number of gene expression comparisons between experimental groups, combined with the commonly present noisy genes showing irrelevant variation, leads to false positives in the identification of truly differentially expressed genes (Dudoit *et al*., 2003) and increases the risk of overfitting in classification methods (Bellman, 1961). Ideally, the high-dimensionality of microarray data should be reduced before the actual analysis by excluding all the non-informative genes. This need for suitable data reduction approaches resulted in the development of many feature selection methods to separate signal from noise, i.e. the informative from the non-informative genes. Most selection algorithms are supervised like the various methods implemented within classification algorithms (Vapnik, 2000), and the ranking of genes on fold changes or test-statistics. As supervised feature selection approaches often suffer from overfitting (Varshavsky *et al*., 2006) and selection bias (Ambroise and McLachlan, 2002), unsupervised feature filtering techniques started to emerge, like ranking of features on variation (Herrero *et al*., 2003), principal components (Hastie *et al*., 2000) or SVD-entropy (Varshavsky *et al*., 2006). But still, these filtering techniques are based on assumptions that are not necessarily universally valid, and therefore still can distort the subsequent statistical analyses. This is unfortunate, as unsupervised filtering increases the significance level of the final result after multiple testing correction (Dudoit *et al*., 2003) because genes are excluded without looking at the label.

Making use of domain knowledge, when available, is key in feature selection (Guyon and Elisseeff, 2003). Affymetrix microarray chips consist of probes that are designed to interrogate how much of the transcript sequence complementary to its DNA sequence is present in a sample (Lockhart *et al*., 1996). They also provide the opportunity to assess whether or not genes were detected in every array. This is because each target transcript is probed by a pair of oligonucleotides; a perfect match (PM) measuring the target mRNA concentration, and a mismatch (MM) for background measurement (Affymetrix, 2002). The difference between PM and MM is used to determine whether the transcript was

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

detected (present) or not (absent). This lies at the base of one of the most objective filtering techniques, namely absent/present calls (A/P calls, Liu *et al.*, 2002). Although this method is not very efficient in filtering (McClintick and Edenberg, 2006), it is complementary to all feature filtering techniques mentioned above and therefore one of the most commonly applied. Another feature of Affymetrix microarray chips is that each target transcript is represented by 11–20 different probe pairs. The intensities of these probes are typically summarized for each probe set to provide one expression level for the respective target transcript (Wu and Irizarry, 2004). Summarization prevents the use of the information provided by the probes on the noise level of the probe set. In this article, we use this information in a rigorous way to assess whether the probe set will be informative for subsequent analyses or not.

This article introduces the concept and applicability of informative/non-informative calls (I/NI calls). I/NI-calls is—like A/P calls—more objective than, and completely complementary to, the existing filtering techniques. We demonstrate that I/NI-calls is a very stringent gene filtering tool using a spiked-in data set and 30 real-life data sets, and illustrate the consequences of I/NI calls on tests for differential expression.

2 METHODS

2.1 The model

I/NI calls expands upon the algorithm used in factor analysis for robust microarray summarization (FARMS) (Hochreiter *et al.*, 2006). FARMS has been developed for summarization, but its excellent application properties for gene filtering remained so far undiscovered. The core of the algorithm is a factor analysis—a multivariate technique to detect a common structure in the data of multiple probes that measure the same target. The assumption is that the probe intensity measurements of the perfect matches x depend on the true mRNA concentration z via:

$$x = \lambda z + \varepsilon \tag{1}$$

with λ being the loadings for the factor analysis (Hochreiter *et al.*, 2006). In Equation (1), a $N(0, 1)$-distributed z models the common factor in the data x, while the $N(0, \psi)$-distributed ε models the independent noise in each probe of each array. In essence, model (1) is explaining the observed covariance structure of the data x by representing the data as being $N(0, \lambda\lambda^T + \psi)$-distributed with an individual noise variance ψ and signal variance $\lambda\lambda^T$. Based on the model assumption, the variance of factor z given the data x, var($z|x$), can be computed through:

$$\mathrm{var}\,(z|x) = (1 + \lambda^T \Psi^{-1} \lambda)^{-1} \tag{2}$$

This value, ranging from 0 to 1, provides a measure of how much variation in the probe set data x is explained by the factor z. The more variation in x is dominated by the signal, the more variation of z is already explained by x, so that var($z|x$) comes closer to 0. Var($z|x$) can be directly translated to a signal-to-noise ratio. A var($z|x$) of 0.5 indicates that z and ε contribute in equal parts to the total variation, corresponding with a signal-to-noise-ratio of 1. Values smaller than 0.5 indicate that there is more signal than noise and these probe sets are therefore selected for further analysis.

The estimation of the parameters of the factor analysis model is done by a Bayesian approach (3), with a prior for λ from a normal distribution with mean μ_λ and variation σ_λ (4).

$$p(\lambda,\psi \big| \{x\}) \propto p(\{x\} \big| \lambda,\psi) p(\lambda,\psi) \tag{3}$$

$$\lambda \sim N(\mu_\lambda, \sigma_\lambda) \tag{4}$$

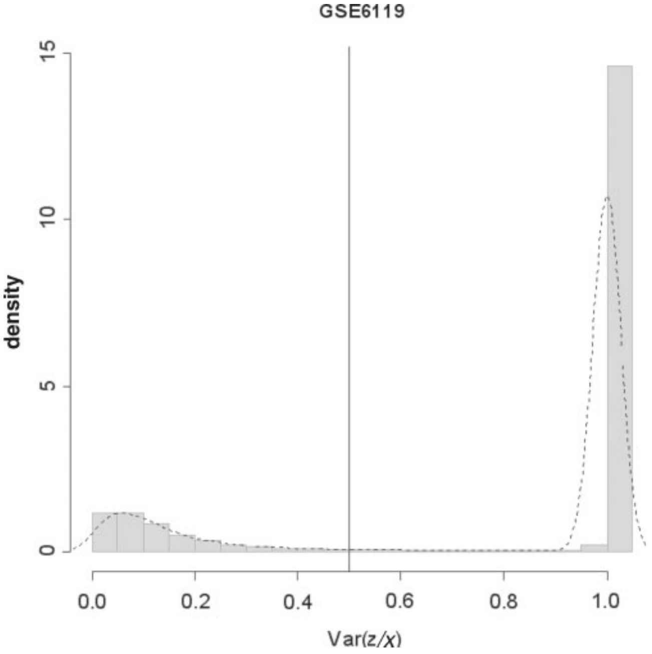

Fig. 1. Histogram of var($z|x$) for the real-life data set GSE6119 (see Supplementary Material 1 for the other data sets).

Setting μ_λ to zero makes loadings of λ equal to zero more likely. This implies that non-informative genes are more likely to be observed. Note that $\lambda = 0$ leads to $\mathrm{var}(z|x) = 1$. This means that z is not determined by an observation x when it is only explained by noise. Var($z|x$) consequently shows a clear bimodal distribution with a very distinct mode for the non-informative and informative probe sets (see Fig. 1 and Supplementary Material 1). This data-driven bimodal distribution facilitates the use of 0.5 as an objective threshold for var ($z|x$) to classify genes as informative or non-informative. Indeed, Fig. 1 (and Supplementary Material 1) show that the results are very robust against the choice of threshold value, as cut-offs between 0.3 and 0.9 would result in very similar conclusions.

Var($z|x$) is actually a multivariate measurement of the correlation between the components of x. According to model (1), the observations x are distributed according to a normal distribution with zero mean and covariance $\lambda\lambda^T + \psi$. So if the data covariance is mainly explained by λ then $x \approx \lambda z$, meaning that the noise is neglected. Then the components of x are $x_j \approx \lambda_j\ z \approx \lambda_j/\lambda_i\ x_i$, meaning that probes x_i and x_j are highly correlated. Conversely, a correlation between probes x_i and x_j is equivalent to a positive entry at position ij in the covariance matrix of x. Now, as ψ is diagonal, this entry can only be explained by $\lambda\lambda^T$. This means that highly correlated probes lead to high values of λ and low values of ψ According to (2), large λ and small ψ result in values of var($z|x$) near zero. Hence, a strong correlation among probes results in a var($z|x$) of 0, and—as can be proven analogously—a weak correlation results in a var($z|x$) of 1.

As FARMS is—like GCRMA—a multi-array summarization technique, it depends on the number of arrays being preprocessed. We show that I/NI filtering is useful when experiments have at least six arrays (see Supplementary Material 2).

2.2 Used data sets

We made use of the spike-in data set from the Affycomp website (Irizarry *et al.*, 2006) and 30 real-life data sets obtained from Gene

Table 1. The real-life datasets used for the assessment of I/NI calls

Accession number	Chip	Total	I/NI calls	A/P calls
E-MEXP-101	hgu133a	22 283	1726	12 898
E-MEXP-120	hgu133a	22 283	5027	13 850
E-MEXP-121	hgu133a	22 283	5105	16 574
E-MEXP-714	hgu133a	22 283	1242	13 711
E-MEXP-72	hgu133a	22 283	4385	13 801
Spike-in U133	hgu133a	22 300	113	12 869
E-MEXP-882	hgu133plus2	54 675	16 022	41 355
E-TABM-127	hgu133plus2	54 675	4962	41 022
E-TABM-34	hgu133plus2	54 675	12 810	35 162
E-TABM-84	hgu133plus2	54 675	6781	38 258
GSE3744	hgu133plus2	54 675	10 673	42 625
E-MEXP-834	Mouse430_2	45 101	8067	26 382
E-MEXP-835	Mouse430_2	45 101	5247	26 891
E-MEXP-839	Mouse430_2	45 101	8107	28 485
E-MEXP-842	Mouse430_2	45 101	1756	27 945
E-TABM-102	Mouse430_2	45 101	8858	29 934
E-MEXP-856	Mouse430A_2	22 690	5014	16 569
GSE2867	Mouse430A_2	22 690	3027	16 412
GSE2882	Mouse430A_2	22 690	4080	15 035
GSE3858	Mouse430A_2	22 690	2801	14 379
GSE4065	Mouse430A_2	22 690	984	12 181
E-MEXP-553	Rat230_2	31 099	3255	19 261
E-MEXP-920	Rat230_2	31 099	954	22 725
E-MEXP-948	Rat230_2	31 099	4080	19 378
GSE5606	Rat230_2	31 099	2723	20 626
GSE6119	Rat230_2	31 099	7449	22 030
GSE1491	ATH1-121501	22 810	3138	17 855
GSE3326	ATH1-121501	22 810	8186	17 827
GSE3350	ATH1-121501	22 810	5716	16 646
GSE3416	ATH1-121501	22 810	4635	15 159
GSE431	ATH1-121501	22 810	3593	15 653

The Accession number from either GEO or ArrayExpress is mentioned, together with the used chip type and the number of probe sets (total number on the array, and number of probe sets filtered using I/NI calls and A/P calls).

Expression Omnibus (GEO) (www.ncbi.nlm.nih.gov/geo/) and ArrayExpress (www.ebi.ac.uk/arrayexpress/). The 30 publicly available data sets were selected to cover six of the most commonly used Affymetrix gene chips, namely human genome chips (HGU133plus2 and HGU133A), mice genome chips (Mouse430_2 and Mouse430A_2), rat genome chips (Rat230_2) and Arabidopsis genome chips (ATH1-121501). See Table 1 for GEO and ArrayExpress accession numbers, a brief description of the data and the actual numbers of filtered probe sets.

3 RESULTS

3.1 Calling a probe set informative or non-informative

As the different probes of a probe set are designed to measure the same target transcript, most of them should be correlated if there is meaningful variation in the concentration of this target transcript across the arrays in the experiment. We call a probe set informative when many of its probes reflect the same increase or decrease in mRNA concentration across arrays. No common probe pattern across arrays indicates that the variation in probe expression values among arrays did not exceed the noise within a probe set, and suggests therefore the exclusion of this probe set. We call such a gene non-informative, as opposed to an undetectable gene, which is a gene that was called absent in all arrays using A/P calls (Liu *et al.*, 2002). The scatterplots in Figure 2 illustrate how 11 probes of a probe set are correlated for a non-informative (red) and an informative (green) probe set. In an informative probe set, the variation in mRNA concentration across arrays is apparent in all its probes, making these probes highly correlated. A non-informative probe set, on the other hand, has typically no consistent probe behavior. Here, increased expression values in certain arrays do not coincide in any of the joint probes. Empirical and simulated data show that probe sets with an intermediate behavior between these two clear examples are called informative as soon as at least half of their probes are correlated (see Supplementary Material 3).

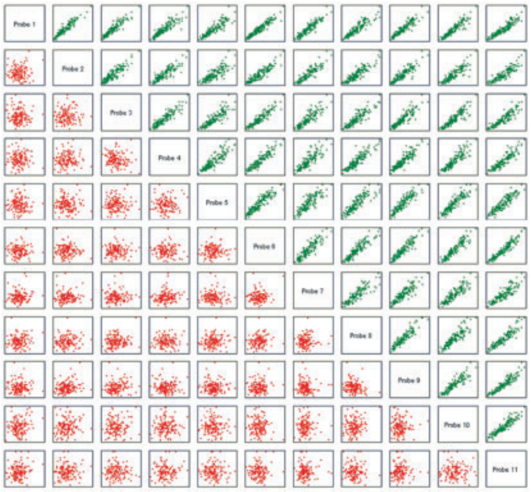

Fig. 2. Probe level patterns for an informative and a non-informative probe set. This scatterplot matrix shows all pair-wise correlations among the 11 probes of the same probe set across arrays for (1) an informative probe set (colored in green in the upper right panel) and for (2) a non-informative probe set (colored in red in the bottom left panel). Each dot represents an array.

3.2 Exclusion rates of I/NI calls

For the spike-in data set (Irizarry *et al.*, 2006) and 30 real-life data sets, on average 84 (±1.5)% of all probe sets could be excluded using I/NI calls, while A/P calls excluded only 33 (±1)%. This significant difference in filtering efficiency (paired *t*-test, $t_{30} = 37$, $P < 0.0001$) was apparent in all the different Affymetrix chips under study (Fig. 3). Such high exclusion rates generated by I/NI calls are expected when using high-content genome arrays where most probe sets are irrelevant for the interpretation of the experiment. In the spike-in data set, I/NI filtering excluded 99.5% of the probe sets. The remaining 0.5% included all spiked-in probe sets. In addition to this confirmed absence of false negatives, we have never observed—in all biological data sets examined so far—the exclusion of a gene that was proven to be biologically meaningful. On the contrary, instead of being too stringent,

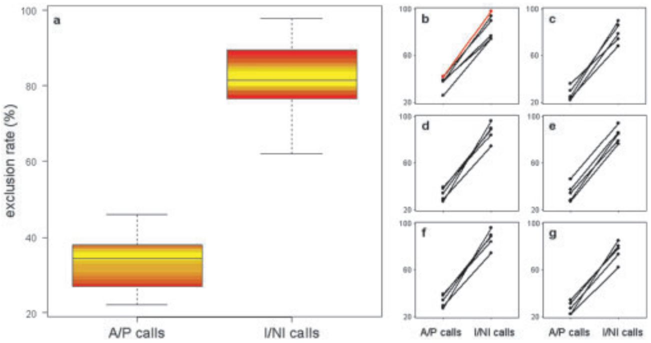

Fig. 3. Graphical comparison of exclusion rates between informative/non-informative (I/NI) calls and absent/present (A/P) calls. (**a**) Box-plots showing the distribution of the exclusion rates of both filtering techniques. The color gradient reflects the distribution within the interquartile range, going from yellow (=50%) to red (=25% and 75%). On the right, the exclusion rates of both filtering techniques are connected for each data set for (**b**) the hgu133a chip with the spiked-in data colored red, (**c**) the hgu133plus2 chip, (**d**) the Mouse430_2 chip, (**e**) the Mouse430A_2 chip, (**f**) the Rat230_2 chip and (**g**) the ATH1-121501 chip. See Table 1 for a description of the used data sets, which are obtained from GEO (www.ncbi.nlm.nih.gov/geo/) and ArrayExpress (www.ebi.ac.uk/arrayexpress/).

filtering using I/NI calls is too conservative as it still selects probe sets with low variation like a number of background probe sets in the spike-in data set. However, as I/NI calls in the current setting already results in exceptionally strong filtering, we suggest its use in this slightly conservative setting to prevent the exclusion of potentially interesting genes.

3.3 Impact on performance of statistical tests

Applying I/NI call selection prior to tests for differential expression has two major implications. First—apart from multiple testing correction—the list of significant genes shortens as some probe sets that would otherwise have been called significant have now been excluded. To illustrate this, we used two groups of three arrays (triplicates) that were spiked in at different concentrations [Experiments 5 and 6 of the spiked-in data set (Irizarry *et al.*, 2006)]. We tested for differential expression between these two groups with a *t*-test after filtering using both A/P and I/NI calls, using GCRMA summarized data as an independent comparison platform. After A/P filtering, the so-called significantly differentially expressed genes ($n = 740$) contained many false positives, i.e. probe sets that were not spiked-in (Fig. 4), while the list of significant genes is much shorter ($n = 36$) due to a much smaller number of false positives (Fig. 4). In various data sets, I/NI calls indeed weeded genes out that were statistically significant but had a biological function that seemed irrelevant in the respective experimental framework. Hence, filtering based on I/NI calls makes gene lists more interpretable as it seems to help excluding false positives.

A second implication of I/NI filtering on tests for differential expression is that multiple testing becomes less problematic, because the number of tests dramatically decreases. Of the 35 spiked-in probe sets that were initially called significant, only 14 remained after FDR correction (Benjamini and Hochberg, 1995) with A/P filtering while the number of significantly called genes remained unaffected after I/NI calls.

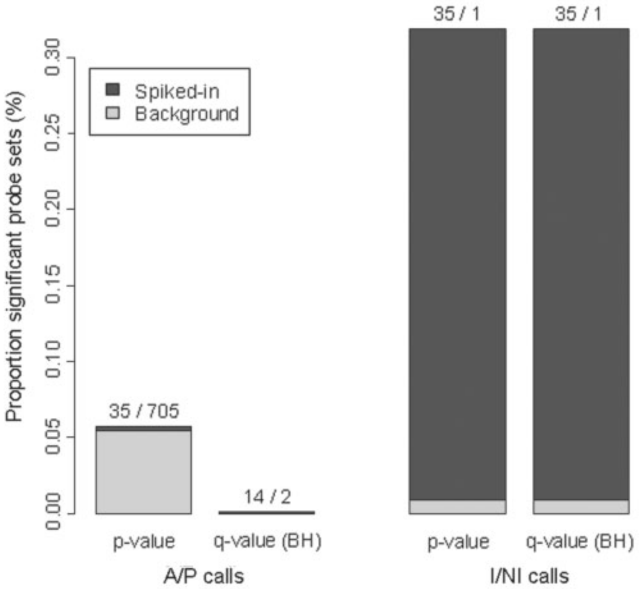

Fig. 4. Effect of gene filtering on tests for differential expression. Two differently spiked-in arrays, each done in triplicate (Experiments 5 and 6; Irizarry et al., 2006) were tested for differential expression with a *t*-test after filtering using both A/P and I/NI calls, using GCRMA summarized data. The proportion of significant probe sets ($\alpha = 0.05$) is given for the two filtering techniques before and after multiple testing correction with an FDR of 10% (Benjamini and Hochberg, 1995).

To illustrate the biological relevance of I/NI calls, we compared the genes called informative in two of the studied data sets with the conclusions of their respective papers. Nishiruma *et al.* (2003), describing public data set GSE431, provide a list containing the 684 significant probe sets, sorted by their fold change. Of the top ranked 50 genes of this list (i.e. the 50 significant genes with the highest fold change), 49 were called informative (98%), indicating that I/NI calls indeed filters the relevant genes. This high proportion of informative genes decreased gradually to 75% when including more genes with smaller fold changes (see Supplementary Material 4). This is in line with the expectation that significant genes with smaller fold-changes are more likely to be false positives, and suggests therefore that I/NI calls is capable to identify these false positives. In another paper, Glyn-Jones *et al.* (2007; data set GSE5606) compared animals with and without a treatment that induces diabetes. They conclude that the genes that were differentially expressed between the treatments were often related to proteins in the mitochondria and to genes regulating fatty acid metabolism (see Supplementary Material 4). A pathway analysis of the genes called informative using I/NI calls resulted in highly significantly affected pathways like 'Mitochondrial long chain fatty acid beta-oxidation' ($P = 1E - 14$) and 'Mitochondrial unsaturated fatty acid beta-oxidation' ($P = 3E - 12$). In contrast, an identical pathway analysis using non-informative genes resulted in much less significant pathways that seemed to be irrelevant in the context of the article. Clearly, I/NI-calls made the analysis more focused on the relevant expression changes.

3.4 Properties of probe sets excluded by I/NI and by A/P calls

The expression values of the probe sets excluded by I/NI and A/P calls have different distributional properties (Fig. 5 and Supplementary Material 5 for all the 31 data sets under study). Most probe sets excluded with A/P calls have average expression values below 5 and variances of 0.1 or lower (Fig. 5b). This is because filtering based on A/P calls selects for probe sets that were called at least once present, making it dependent on the average expression value (the lower, the more likely absent) and on variation across arrays (the higher, the more likely at least one array is called present). This is however not a general pattern, as some probe sets with low average expression values and low variances are still filtered (Fig. 5a). Probe sets excluded by I/NI calls are—like A/P calls—also less variable (Fig. 5d), but can have either low or high expression values. The low-expressed probe sets excluded by I/NI calls are mostly probe sets where the technical noise was as high as the variation across arrays. The highly expressed, but excluded, probe sets code for transcripts with equally high concentrations in all arrays. These probe sets are present—and therefore selected by A/P calls—but not variable across arrays. Hence, as microarray experiments in principle try to discover differences between conditions, these genes are mostly regarded as being non-informative. Besides, e.g. house-keeping genes, these probe sets also include genes expressed at saturation levels (>13, see Fig. 5c). Figure 5c also indicates that lower expressed genes need to be more variable to be called informative by I/NI calls. This is because background noise increases with decreasing average expression levels. The signal, i.e. the true variation in gene expression across arrays, therefore needs to increase as well in order to call these genes informative. This is an objective approach similar to current common practice where people rather subjectively rely less on differentially expressed genes at lower intensity values. As these have indeed a higher potential of being false positives due to background noise, microarray users often ignore them when their fold change is rather low. Another common practice in microarray analysis is to select the most variable genes after deleting the always-absent ones. This approach not only involves arbitrary threshold choices like for instance the number of variable genes, but it also hampers the detection of truly differentially expressed genes at relatively small fold changes when they coincide with other, quite noisy—and therefore variable—genes. Hence, I/NI calls are providing a better alternative as they serve the same purpose and are based on the same reasoning as filtering on variance or on coefficient of variation, but have three main improvements: First, they do not use a general measure of probe set variation, but disentangle biological variation from variation due to technical noise, and use the mutual proportion between them as a kind of selection criterion. Second, they avoid the need of several decision steps (A/P calls, filtering on minimum variation and so forth), but incorporate all information into a single analysis. And third, no arbitrary threshold choices or assumptions have to be taken.

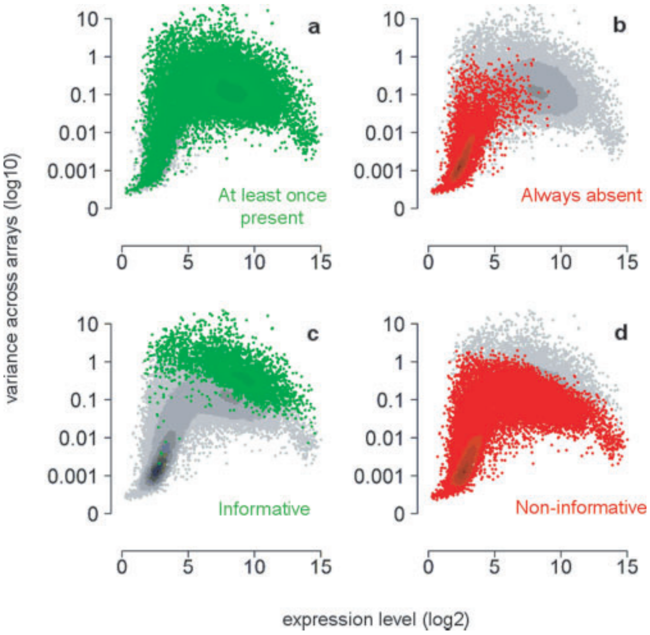

Fig. 5. Distributional properties (variance and mean) of GCRMA summarized genes selected by A/P calls and I/NI calls. Variance and mean are calculated per probe set across arrays using GCRMA summarization, and are plotted against each other. All probe set values are plotted in gray in the background and are superimposed by probe sets called at least once present (colored green in **a**), always absent (colored red in **b**), informative (colored green in **c**) and non-informative (colored red in **d**).

4 CONCLUSIONS

By incorporating probe level information to assess the noisy nature of probe sets, I/NI calls provide a highly powerful and objective tool for gene filtering. Consequently, I/NI calls offer a key solution to the main problem in the analysis of high-dimensional microarray data, being the high recurrence of false positive results because of multiple testing and overfitting. We therefore suggest that I/NI calls be used more routinely in combination with summarization techniques like FARMS (Hochreiter *et al.*, 2006) or GCRMA (Wu *et al.*, 2004).

ACKNOWLEDGEMENTS

We are grateful to Jeroen Aerssens and An De Bondt for reviewing and discussing earlier versions of this manuscript. We also want to explicitly acknowledge the current possibility to make easy use of publicly available microarray data sets. This is thanks to invaluable efforts of the people maintaining ArrayExpress and GEO, and in the first place thanks to the scientists who are willing to submit their data to the public domain.

Conflict of Interest: none declared.

REFERENCES

Affymetrix (2002) Statistical Algorithms Description Document. Available from www.affymetrix.com

Ambroise,C. and McLachlan,G.J. (2002) Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl Acad. Sci. USA*, **99**, 6562–6566.

Bellman,R.E. (1961) *Adaptive Control Processes*. Princeton University Press, Princeton, NJ.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B Met.*, **57**, 289–300.

Dudoit,S. *et al.* (2003) Multiple hypothesis testing in microarray experiments. *Stat. Sci.*, **18**, 71–103.

Glyn-Jones,S. *et al.* (2007) Transcriptomic analysis of the cardiac left ventricle in a rodent model of diabetic cardiomyopathy: molecular snapshot of a severe myocardial disease. *Physiol. Genomics*, **28**, 284–293.

Guyon,I. and Elisseeff,A. (2003) An introduction to variable and feature selection. *JMLR*, **3**, 1157–1182.

Hastie,T. *et al.* (2000) 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biol.*, **1**, RESEARCH0003.

Herrero,J. *et al.* (2003) Gene expression data preprocessing. *Bioinformatics*, **19**, 655–656.

Hochreiter,S. *et al.* (2006) A new summarization method for Affymetrix probe level data. *Bioinformatics*, **22**, 943–949.

Irizarry,R.A. *et al.* (2006) Comparison of Affymetrix GeneChip expression measures. *Bioinformatics*, **22**, 789–794.

Liu,W.M. *et al.* (2002) Analysis of high density expression microarrays with signed-rank call algorithms. *Bioinformatics*, **18**, 1593–1599.

Lockhart,D.J. *et al.* (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.*, **14**, 1675–1680.

McClintick,J.N. and Edenberg,H.J. (2006) Effects of filtering by present call on analysis of microarray experiments. *BMC Bioinformatics*, **7**, 49.

Nishimura,M.T. *et al.* (2003) Loss of a callose synthase results in salicylic acid-dependent disease resistance. *Science*, **301**, 969–972.

Vapnik,V.N. (2000) *The Nature of Statistical Learning Theory*. Springer Verlag, New York.

Varshavsky,R. *et al.* (2006) Novel unsupervised feature filtering of biological data. *Bioinformatics*, **22**, e507–e513.

Wu,Z. and Irizarry,R.A. (2004) Preprocessing of oligonucleotide array data. *Nat. Biotechnol.*, **22**, 656–658.

Wu,Z. *et al.* (2004) A model based background adjustment for oligonucleotide expression arrays. *J. Am. Stat. Assoc.*, **99**, 909–917.