Coincidence between transcriptome analyses on different microarray platforms using a parametric framework

Tomokazu Konishi[1,2], Fumikazu Konishi[2, 3], Shigeru Takasaki[2], Kohei Inoue[4], Koji Nakayama[4] and Akihiko Konagaya[2, 3]

Akita Pref. Univ.[1], RIKEN [2], Tokyo Tech.[3], Mitsubishi Chemical Safety Inst. Ltd.[4]

This study was supported by NEDO (New Energy and Industrial Technology Development Organization) Survey Research Grant: 06002357-0-1.

ABSTRACT

A parametric framework for the analysis of transcriptome data is demonstrated to yield coincident results when applied to data acquired using two different microarray platforms. Discrepancies among transcriptome studies are frequently reported, casting doubt on the reliability of collected data. The inconsistency among observations can be largely attributed to differences among the analytical frameworks employed for data analysis. The existing frameworks normalizes data against a standard determined from the data to be analyzed. In the present study, a parametric framework based on a strict model for normalization is applied to data acquired using an in-house printed chip and GeneChip. The framework is based on a common statistical characteristic of microarray data, and each data is normalized on the basis of a linear relationship with this model. In the proposed framework, the expressional changes observed and genes selected are coincident between platforms, achieving superior universality of data compared to other methods.

INTRODUCTION

The transcriptome, the contents of mRNA, determines the functions of a cell. Microarrays are currently widely used to acquire comprehensive transcriptome information, and thus have greatly facilitated transcriptome research. However, an appropriate intellectual framework [1] for systematizing the data collected using various microarrays has yet to be developed. Due to the lack of a framework that yields consistent results among different platforms, the reliability of numerous measurements in the literature may have been compromised, particularly when comparisons among different platforms have been performed, raising many questions and criticisms [2-6]. In the present study, a parametric framework is demonstrated to yield excellent coincidence when applied to data acquired using two different microarray platforms. In this framework, data is normalized with respect to a statistical characteristic common to all measurements and the data is compared to the model using a linear relationship. Superior coincidence has previously been demonstrated using this framework for data acquired using a two-colored printed chip and the GeneChip (Affymetrix, Santa Clara, CA), for which other frameworks based on MAS5 [7], RMA [8] and LOWESS [9] do not reveal any coincidence. The proposed framework thus appears to provide a means for the seamless integration of information obtained in transcriptome studies. The highly reliable data thus obtained may also provide clues for decoding the hereditary traits within the genome [10], which may in turn lead to rapid progress in the life sciences.

Achieving universality in microarray data has proved to be more problematic than may have been expected. The essential character of a transcript is determined by its concentration as the transcript acts as a template for the translation process, and the rate

of translation is linear when compared to the concentration of the template in the cytosol. However, concentrations cannot be measured using current microarray systems. Measurement of transcripts therefore requires that RNA samples be isolated from tissue, for which the collection rates and cytosol volumes are difficult to estimate. Consequently, even if the mass of each transcript in a sample can be determined, the concentrations cannot be calculated. This practical imprecision is further complicated by the variety of platforms available for microarray systems, which differ with respect to the probe sensitivity of the hybridization systems and the nucleotide sequences used. The potential errors and biases will also differ between platforms, and the level of additive noise and saturation will vary according to the measurement approach. Such noise and error contribute to further discrepancies among data sets.

To achieve the universality of data and resolve the problems associated with incompatibility, a unified intellectual framework [1] is therefore required. However, relatively little attention has been paid to the development of such a data framework [11]. Without an adequate framework that is not affected by measurement sensitivity and background, even the ratios of expression levels cannot be estimated correctly as these are framework dependent.

A parametric framework is expected to be suitable for achieving the universality of data. The framework presented here is based on thermodynamic models describing the formation of the transcriptome in a cell [10] and the detection of RNA by hybridization [11]. These models assume signal responses that are linear relative to the concentration of transcripts, and the suitability of the models has been verified by ensuring that the expected data distribution (i.e., lognormal) is obtained in actual measurements. These measurements also revealed the lowest value unaffected by additive noise and the

highest value unaffected by saturation, since the data diverge from the expected pattern at these limits [12]. Although the effects of probe sensitivity remain in the normalized data [11], such effects are cancelled when the ratios between samples are taken. Consequently, the obtained ratios are expected to represent only those ratios between concentrations in sample cells.

In the present study, the proposed parametric framework is applied to data obtained using two different platforms. The data were obtained from a rat toxicology project study [13] in which rats were administered with various chemicals. Multi-sample RNA isolated from rat organs was then hybridized to two platforms: an in-house microarray (ToxArray III), and the GeneChip microarray in three different laboratories [14, 15]. The ToxArray III is a typical microarray on slide glass, consisting of a single 60 mer probe per gene, and two samples are measured simultaneously per chip. In contrast, the GeneChip consists of 11 perfect match (PM) 25 mer probes per gene, and a single sample is measured per chip. While differences in the measurement positions of each transcript could alter the obtained information, the overall trend in the information obtained from each of the two platforms should coincide. In this report, the coincidence of information is checked by examining the measured logarithmic ratios and gene candidates that may be affected by Safrol [15] treatment. In order to evaluate the methodology, GeneChip data were also normalized using both MAS5 [7] and RMA [8]; MAS5 is the original method described by the manufacturer and involves classification of genes into “Present”, “Marginal”, and “Absent” in addition to normalization and summarizing of data. RMA is a widely used alternative based on the quantile method. Data for ToxArray were also normalized using the LOWESS method [9].

RESULTS

Data distribution

The statistical characteristics of the data were determined using normal QQ plots (Fig. 1). Although the signal intensity follows a lognormal distribution over a certain range for both chips, there is a marked difference in the valid intensity range between the two chips. The narrower valid range for ToxArray data suggests a higher level of additive noise. The distribution of ToxArray data has a larger scale parameter than that of GeneChip data, with median values of 1.02 and 0.685, respectively. The dynamic range of signals, estimated from the ratio of the strongest to weakest signal for 10,000 measurements, is 2×10^5 for GeneChip, and 1×10^8 for ToxArray. The higher sigma value is likely to result in measurements exceeding the limits of the scanner, which usually covers a range of 10^4–10^5. Additionally, unevenness in hybridization, as observed from the pseudo images [16], was substantially higher for ToxArray (see supplemental data).

Coincidence of logarithmic ratios between chip platforms

For the genes common to GeneChip and ToxArray (4433 rat genes), the logarithmic ratios determined by different frameworks are compared in Fig. 2. In the parametric framework, the logarithmic ratios are coincident in the valid signal range (Fig. 2(a)). Outside of the valid range however, the ToxArray data become substantially divergent, while the GeneChip data remain relatively close to the $y = x$ line. This may indicate the noise reduction effect associated with the GeneChip due to the averaging of multiple PM cells for each gene. In contrast, larger differences were observed between the LOWESS- and RMA-normalized data (Fig. 2(b)). The coincidence between LOWESS and MAS5 results is very poor for data labeled “Absent” in MAS5 (Fig. 2(c)), but

improved coincidence was observed for the “Present” data, although such a relationship should not always be expected; for example, almost no coincidence was observed in other cases (supplementary data).

Coincidence in selected genes

The lists of genes exhibiting expressional changes larger than the threshold level (METHODS) are compared between the two platforms in Fig. 3. A large overlap in the lists for each platform is observed using the parametric framework. This comparison also reveals differences in the detection power of the chips. For example, 319 genes selected by GeneChip were out of the detection range of ToxArray. Comparisons between LOWESS and RMA or MAS5 methods resulted in a markedly smaller overlap of selected genes. Many genes were selected by only one of the platforms, indicating that there are substantial differences that exist between platforms that are not considered by the parametric framework. Such conflicts suggest the inclusion of more false positives than expected for the test (METHODS).

A synergy of selected genes was also observed. In estimating the physiological condition of the sample, the simultaneous selection of a group of genes indicative of a biological event is a more reliable indicator of that event than the selection of a single pertinent gene. In the present case, the parametric framework reveals an increase in genes related to proteolysis by proteasomes and metabolism of steroids, and a reduction in genes related to antigen presentation via MHC class II as major components of the gene list (Fig. 3, lower rows). In other frameworks, however, none or only a few of these genes were selected, highlighting the detection power of the parametric framework.

DISCUSSION

The parametric framework appears to provide superior reproducibility, has greater testing power, and a lower false-positive rate than existing frameworks. In the present study, although the purpose and subjects of measurement were identical, the analytical results were not coincident between platforms. The degree of coincidence and conflict appears to be largely dependent on the framework employed for the analysis of acquired data. Many of the discrepancies in the information obtained from the two platforms considered here can be attributed to differences in the fundamental philosophies of the frameworks that have conventionally been applied to the respective platforms, and not to inherent differences in the capacities of the chip platforms. This is evidenced by the greater coincidence achieved between data acquired using different platforms when analyzed using the parametric framework.

Each of the frameworks normalizes and compares data using a set of hypotheses and assumptions that form the fundamental basis of the respective frameworks [1]. In the parametric framework, chip data are normalized using a distribution model as the standard, whereas in the other frameworks, a standard is sought among the data sets. For example, the standard is determined for a pair of data in LOWESS [9] and shift-log [17], and from the means of data quantiles in RMA. Consequently, the normalization of a data set in existing frameworks is affected by all of the data sets being processed at the same time. This dependency on other data sets can be expected to adversely affect the uniformity of the analysis, which becomes apparent when comparing information among different studies.

Another fundamental difference is associated with the testability of the fundamentals

of the framework. LOWESS and RMA inevitably fulfill the assumptions regarding the assumed nature of the data, that is, they take the form of stable logarithmic ratios (LOWESS) or identical data distributions (RMA). MAS5 contains numerous conditional judgments that are not based on factual knowledge. The premises of these frameworks therefore precludes effective evaluation of the model assumptions. The parametric framework, on the other hand, employs a strict model and normalization cannot be completed without coincidence between the model and chip data. Any test of validity therefore relates to the reliability of the obtained information.

The normalization process, which tests data distribution against a model (Fig. 1), is useful for identifying the likely range of data. As with other measurement systems, microarrays inevitably contain noise. With repeated measurements such as those shown in Fig. 3, the noise level can be reduced by taking the means of repeats. However, the noise contained in each measurement may still affect analyses, particularly when small numbers of repeats are available. Even in such cases, the parametric framework allows the data range that is likely to be affected by additive noise and saturation to be clearly defined (Fig. 1). The usefulness of this method for data classification is clearly shown in Fig. 2(a), and is expected to increase the reliability of analyses by reducing the false-positive rate.

The proposed parametric framework thus achieves superior universality of data and allows for the evaluation of data reliability, providing a means of integrating knowledge obtained from many different laboratories and chip platforms.

METHODS

Test animals

Male Fischer 377 rats (SPF, 5 weeks of age) were administered with 300 mg/kg/day of Safrol for up to 28 days at the Mitsubishi Chemical Safety Institute [14, 15]. RNA samples were isolated from the liver of each test animal.

Microarrays

Identical RNA samples were investigated using GeneChip (Rat Genome 230 2.0 array; Affymetrix) and a NEDO-ToxArray III ink-jet printed chip (6709 genes [13]). These microarrays share an overlap of 4433 genes.

Normalization

Parametric normalization was performed using SuperNORM (Skylight Biotech, Akita). Other normalizations were performed using *R* version 2.4 [18] with the implemented *affy* library [19] as follows. The scanner-estimated background was subtracted from the Cy3 and Cy5 data of each ToxArray chip, and the logarithmic ratios were stabilized by the LOWESS function [9] in *R*. The normalized data were separated into Cy3 and Cy5 channels [17] for gene selection. The GeneChip data were normalized and summarized using the MAS5 function or the RMA [8] function of *affy* [19].

Comparisons of selected genes

Genes that suggested larger effects in expressional changes relative to a predefined threshold were selected. The effects were estimated using treatment ($n = 4$) and control ($n = 4$) rats, all of which were measured identically. The effect (E) was estimated for each gene (g) between the logarithm forms of normalized data for treatment (T) and control (C) using data from the ith chip as follows.

$$E_g = \frac{\sum T_{g,i}}{n_{g,T}} - \frac{\sum C_{g,i}}{n_{g,C}} \quad .$$

Here, $n_{g,T}$ and $n_{g,C}$ are the numbers of available data for the treatment and the control. The threshold of the selection was determined using the mean of sample variances (s_g^2), which was measured for each gene as follows:

$$s_g^{\,2} = \frac{\sum (T_{g,i} - \overline{T_g})^2 + \sum (C_{g,i} - \overline{C_g})^2}{n_{g,T} + n_{g,C} - 2}.$$

The threshold level (L) was then determined using the expression

$$L = 3 \times \sqrt{\frac{\sum s_g^{\,2}}{n_{genes}}},$$

where n_{genes} is the number of the gene. Genes that display $|E_g|$ values larger than L were selected. In the parametric framework, only data in the valid range were used. In cases where $n_{g,T}$ or $n_{g,C}$ was less than 3, the gene was excluded. In MAS5, only “Present” data were used. Assuming the effects of biological conditions and experimental error are normally distributed, the expected false positive rate is 0.27% (12 cases) of the gene comparisons. Standard statistical tests such as the Student’s *t* test or ANOVA were not employed because both estimate the threshold in a gene-wise manner and the number of total measurements was considered insufficient. Such tests may also be unfavorable for toxicological applications due to oversensitivity when variances are small. If small differences can be detected using a sufficiently large data set and a reference pattern of transcriptome changes can be established, it may be possible to conduct experiments on-site using a limited number of chips. However, comparison of on-site data with a reference pattern will be difficult because the information obtained using a small number of measurements may contain fluctuations attributable to noise, resulting in an

elevated false-positive rate.

1. Popper, S.K.R. Myth of the Framework: In Defence of Science and Rationality. (Routledge, London; 1994).
2. Frantz, S. An array of problems. *Nat Rev Drug Discov* **4**, 362-363 (2005).
3. Harris, L. Microarray data stands up to scrutiny. *The Scientist* **19**, 32 (2005).
4. Abdullah-Sayani, A., Bueno-de-Mesquita, J. & van de Vijver, M. Technology Insight: tuning into the genetic orchestra using microarrays--limitations of DNA microarrays in clinical practice. *Nat Clin Pract Oncol* **3**, 501-516 (2006).
5. Draghici, S., Khatria, P., Eklundb, A.C. & Szallasic, Z. Reliability and reproducibility issues in DNA microarray measurements. *Trends Genet* **22**, 101-109 (2006).
6. Stafford, P. & Brun, M. Three methods for optimization of cross-laboratory and cross-platform microarray expression data. *Nucleic Acids Res* **35**, e72 (2007).
7. Hubbell, E., Liu, W. & Mei, R. Robust estimators for expression analysis. *Bioinformatics* **18**, 1585-1592 (2002).
8. Bolstad, B., Irizarry, R., Astrand, M. & Speed, T. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185-193 (2003).
9. Yang, Y. et al. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res* **30**, e15 (2002).
10. Konishi, T. A thermodynamic model of transcriptome formation. *Nucleic Acids Res* **33**, 6587-6592 (2005).

11. Konishi, T. Data distribution of short oligonucleotide expression arrays and its application to the construction of a generalized intellectual framework. *submitted*.

12. Konishi, T. Three-parameter lognormal distribution ubiquitously found in cDNA microarray data and its application to parametric data treatment. *BMC Bioinformatics* **5**, 5 (2004).

13. Nakayama, K. et al. Differences in gene expression profiles in the liver between carcinogenic and non-carcinogenic isomers of compounds given to rats in a 28-day repeat-dose toxicity study. *Toxicol Appl Pharmacol* **217**, 299-307 (2006).

14. Sumida, K. et al. Optimization of an animal test protocol for toxicogenomics studies (I); requirement study of a protocol. *J Toxicol Sci* **32**, 19-32 (2007).

15. Sumida, K. et al. Optimization of an animal test protocol for toxicogenomics studies (ii); a cross-laboratory gene expression analysis. *J Toxicol Sci* **32**, 33-45 (2007).

16. Konishi, T. Detection and Restoration of Hybridization Problems in Affymetrix GeneChip Data by Parametric Scanning. *Genome Inform* **17**, 100-109 (2006).

17. Kerr, M.K. et al. Statistical analysis of a gene expression microarray experiment with replication. *Statistica Sinica* **12**, 203-217 (2002).

18. R Development Core Team. R: A Language and Environment for Statistical Computing, Vol. ISBN 3-900051-07-0. (Vienna, Austria; 2007).

19. Gautier, L., Cope, L., Bolstad, B. & Irizarry, R. affy--analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* **20**, 307-315 (2004).

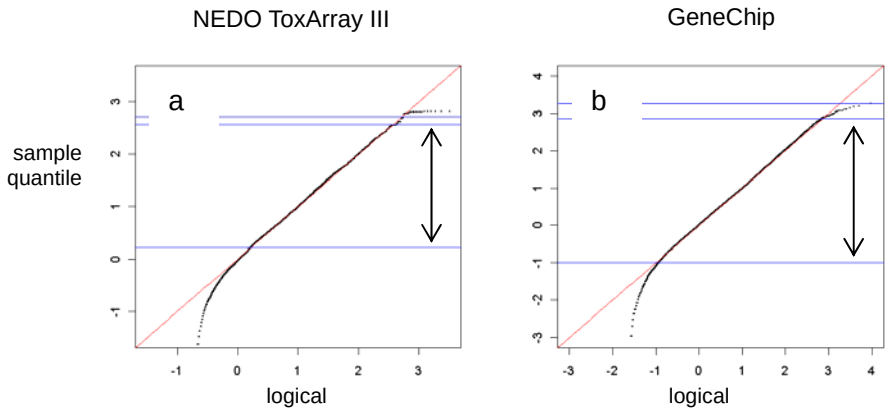

Figure 1. Distribution of normalized data. Signals obtained for the same RNA sample are shown in normal QQ plots. (a) Spots other than controls from NEDO ToxArray III, and (b) PM data from GeneChip. Red line denotes $y = x$. Arrows denote valid range of data with respect to model fit.

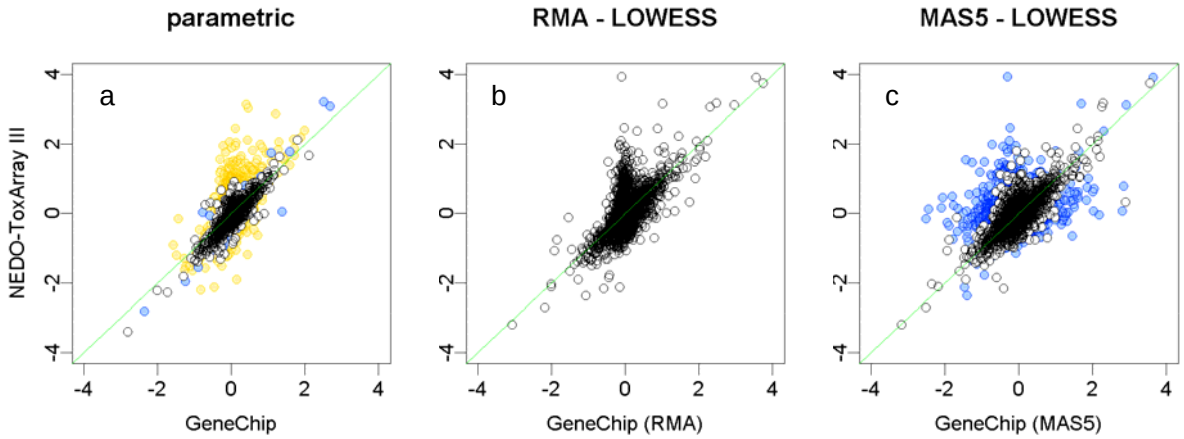

Figure 2. Coincidence in $\log_2$ ratio between chip platforms. Differences in transcripts between Safrol-administrated sample and mock control are measured using GeneChip and NEDO ToxArray III. Data plotting on $y = x$ line (green) are coincident between platforms. Colored plots denote "out of detection" judgment by the respective frameworks. (**a**) Parametric framework. Colored plots denote signals out of the valid range of the parametric model (GeneChip, blue; ToxArray, orange). Residual root mean square (RMS) between platforms in valid range (average difference) is 0.210 (1.16 fold). (**b**) LOWESS (ToxArray) vs. RMA (GeneChip). Residual RMS is 0.357 (1.28 fold). (**c**) LOWESS (ToxArray) vs. MAS5 (GeneChip). Blue plots denote "Absent" MAS5 data. Residual RMS for "Present" data is 0.283 (1.21 fold).

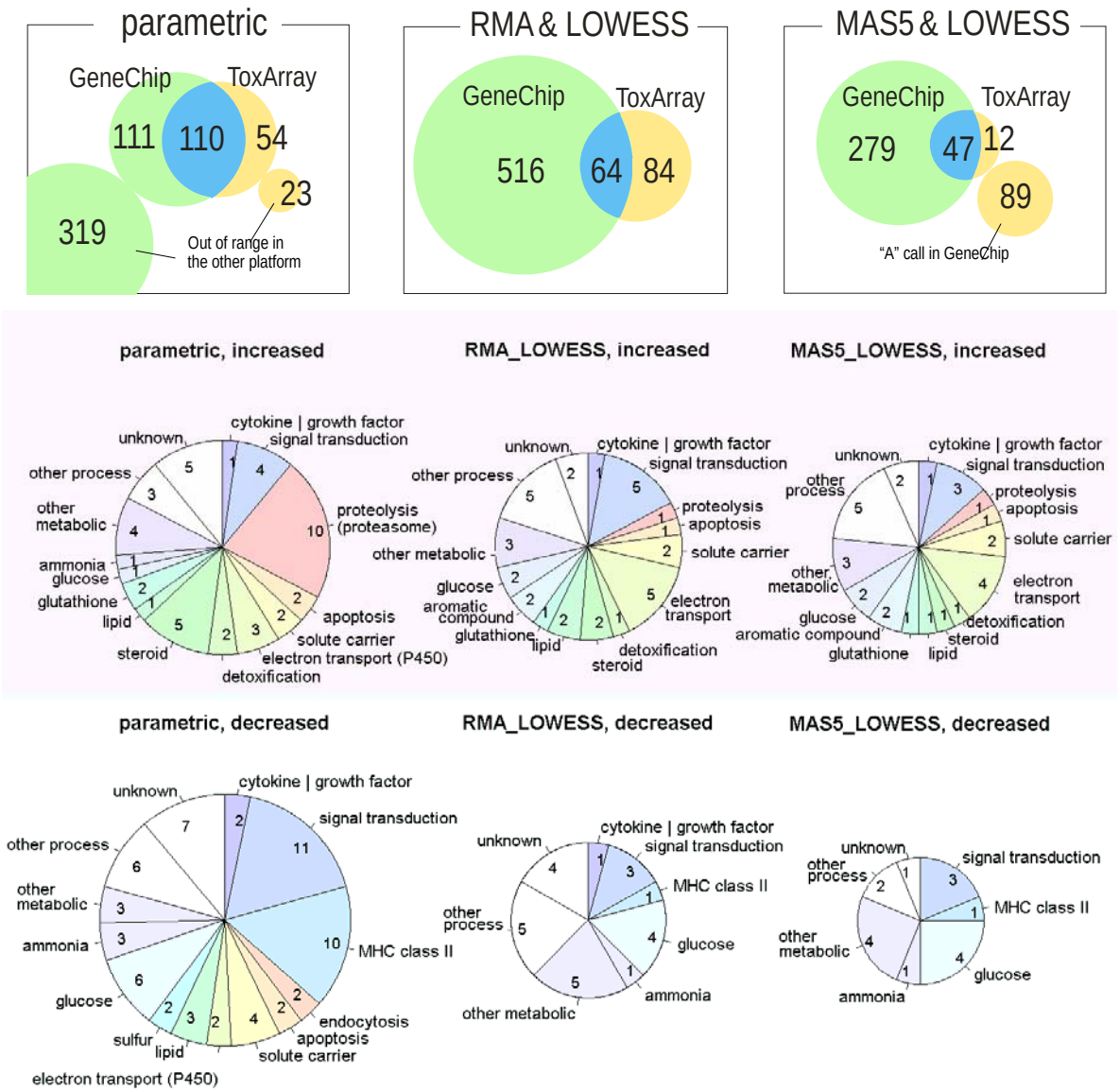

Figure 3. Summary of genes selected using an average level of estimated noise (METHODS). (Upper) Coincidence between selected gene lists. Values denote number of selected genes. Signals outside of the valid model range are omitted as shown. Comparisons of logarithmic ratios are shown in the supplementary data. (Lower) Functions of cross-selected genes from gene titles and biological processes of gene ontology provided by the chip manufacturer.