The Fourth International Conference on Computational Systems Biology (ISB2010)
Suzhou, China, September 9–11, 2010

Effects of Multiple Probesets in Affymetrix GeneChips on Identifying Differentially Expressed Genes in iPS Cells

Zhi-Ping Liu[1,*] Xiang-Sun Zhang[2,†]

[1] Key Laboratory of Systems Biology, SIBS-Novo Nordisk Translational Research Centre for PreDiabetes, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China
[2] Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China

Abstract There are multiple probesets that correspond to a gene in Affymetrix GeneChip platform. Different combination and annotation methods will lead to different results of expression when we transform the information from probesets to genes. In this work, we collected seven methods to interpret the multiple probesets for representing a gene. We compared these methods and identified their effects on the identification of differential genes. Specifically, we focused on the analysis of differential genes between iPS cells and ES cells. We identified their differences by calculating the t-test p-values and Pearson correlation coefficients. Our results show that gene expression by different methods of combining the multiple probesets results in slight differences of differential genes in iPS cells. We also identified the effects on coexpression in key transcription factors of iPS cells. The results indicate that adopting different methods to handle multiple probesets is important for accurate estimation of gene expressions.

Keywords Gene expression; multiple probesets; differential expression; iPS cell.

1 Introduction

The advent of microarray technology has made it possible to monitor the gene expression levels in parallel. Microarray has been extensively used in biomedical research to address a wide variety of questions [10]. Affymetrix microarrays is one of the widely-used platforms to measure the global expression of mRNA transcripts. This technology is based on a concept of probeset. Individual probes within a probeset are originally designed to hybridize with the same unique mRNA transcript [8]. In the design of genes in Affymetrix GeneChip, gene expression is measured by extracting mRNA from the cells or interested tissues and hybridizing mRNA samples to the 25-mer probes on the microarray [12]. A GeneChip consists of a quartz wafer to which are attached some 500,000 different 25-mer deoxyoligonucletides, which are known as probes. Each expressed transcript is represented on an array by a series of probe pairs known as a probe set [6, 8, 7, 12]. In

*Email: zpliu@sibs.ac.cn
†Corresponding author. Email: zxs@amt.ac.cn

the Affymetrix GeneChip, each probeset is defined by 11 pairs of probes. Often, there are multiple probesets which correspond to a single gene. There are results which indicate that some probesets should not be considered as unique measures of transcription, because the individual probes map to more than one transcript dependent upon the biological condition [12]. About half of genes, each gene is defined by one probeset on the chip. The other half genes are defined by two or more probesets [12]. It is important to interpret the information of the probesets to the gene expression. There are various methods which have been used to meet this task [6, 4, 5, 14]. The effect of multiple probesets for describing the gene expression will be represented by these various interpretation methods [10]. Generally, the aim of most gene expression microarray experiments is to obtain a list of genes which are differentially expressed under certain condition [8]. For the cases of multiple probesets representing the same gene, gene differential expression and coexpression will be determined by various methods to combine the values of multiple probesets. This will significantly affect the final results. Dai et al. [2] have reported that the updated definition can cause as much as 30–50 % discrepancy in the genes selected as differentially expressed on a heart tissue expression profiling dataset.

Induced pluripotent stem (iPS) cells were first produced in 2006 from mouse cells [13] and in 2007 from human cells [15]. This has been cited as an important advancement in stem cell research, as it may allow researchers to obtain pluripotent stem cells which are important in research and potentially have therapeutic uses, without the controversial use of embryos [1, 9]. iPS cells are believed to be identical to natural pluripotent stem cells, such as embryonic stem (ES) cells, in many respects, i.e. the expression of certain stem cell genes and proteins, chromatin methylation patterns, doubling time, embryoid body formation, teratoma formation, viable chimera formation, and potency and differentiability [13, 15, 1]. However, the full extent of their relation to natural pluripotent stem cells is still being assessed. iPS cell is so important and it is the breakthrough technique in pluripotent stem cell. In the research of relationship of iPS and ES cells, gene expression is a promising way to identify their expression profiling and features [1]. In the analysis of the similarity and difference between ES cells and iPS cells, it will be important to develop effective methods to interpret the gene with multiple probesets.

In this work, we propose a comparison study of the methods to interpret the multiple probesets in Affymetrix GeneChips. We compare seven methods to identify the differential expression and coexpression information in those genes corresponding to multiple probesets. The interpretation focuses on identifying the important differential genes in iPS cells. The different and common differential mapping of some transcription factors as well as their related genes are also investigated. The effects of multiple probesets are performed and evaluated in determining the gene expression. The gene expression features underlying the iPS cells are then presented. Our results provides important information and highlight the need of careful consideration when assessing whether groups of probesets are used to measure the same transcript.

2 Results

2.1 Effect on differential gene expression

The comparison of gene expression between iPS cells and ES cells was implemented by Affymetrix HT Mouse Genome 430A Array [11]. There are 15706 unique genes which

are assessed in the array, and 6219 (40%) genes are represented by multiple probesets. In microarray study, differentially expressed genes between two cases are often regarded to be associated with mechanisms behind the phenotype differences. There are many methods for detecting differential genes. Among them the t-test is one of the widely used methods. We used the t-test to detect the differentially expressed genes between iPS cells and ES cells. We implemented the seven methods, i.e. 'Average', 'Summary', 'Random', 'Mean', 'Variance', 'Correlation' and 'Entropy' (see Methods), to define the gene expression from these multiple probesets individually. Table 1 lists the number of the identified differential genes. We selected significantly differential genes by different p-values cutoffs of 0.005, 0.01 and 0.05.

Table 1: Number of differentially expressed genes in iPS cells by different methods to represent the expressions of the genes with multiple probesets.

Threshold	Average	Summary	Random	Mean	Variance	Correlation	Entropy
0.005	176	176	156	166	159	160	154
0.01	290	290	270	281	272	280	273
0.05	1081	1081	1008	1041	1010	994	987

From the total number of genes, we found that there is sight effect on the number of differential genes in the identification. In the genes represented by multiple probesets, Figure 1 shows the boxplot of the p-values of these genes. The p-values of these genes with multiple probesets transforming by the 'Random', 'Mean' and 'Entropy' methods will lead to slightly higher values than those of the other methods. In the genes with repeated probesets, we also calculated the overlapping differential genes detected by the

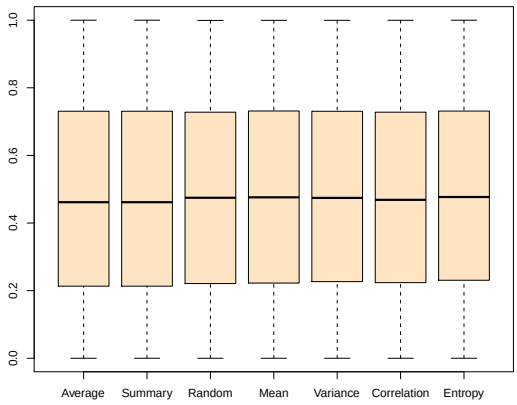

Figure 1: Box plots of differential p-values of the multiple probesets gene by the seven methods.

seven methods. Table 2 shows the numbers of the overlapped differential genes. To assess the statistical significance of these differential gene overlapped between two methods, we

used the hypergeometric probability to calculate the significance [1]. In Table 2, the diagonal part are the number of differential genes in these genes with multiple probesets by p-value 0.05. From the statistical significance, the high conservation of these genes in Table 2 provides more evidence for the slight effect of these multiple probesets by identifying differential genes in iPS cells with different mapping methods. There are different number of genes which are identified as differential genes by different methods. The 'Average' method identified more differential genes than the 'Entropy' method (about 100 genes). The result indicates that we should pay attention to these different methods when we used them to identify differential genes in those with multiple probesets. We will carefully analyze some genes which identified to be differential by one method while no significance by another method.

Table 2: Overlapping differential genes with multiple probesets identified by different methods in iPS cells. Diagonal values are the numbers of identification by threshold p-value of 0.05. Upper diagonal values are the overlapped numbers and lower diagonal values are their corresponding significance.

Method	Average	Summary	Random	Mean	Variance	Correlation	Entropy
Average	472	472	211	201	207	212	236
Summary	0.00e-000	472	211	201	207	212	236
Random	4.82e-145	5.32e-197	399	238	213	236	210
Mean	8.87e-123	1.64e-167	2.50e-180	432	228	253	170
Variance	5.47e-139	8.66e-211	2.40e-228	8.82e-057	401	122	264
Correlation	7.15e-151	2.43e-170	6.44e-106	9.32e-262	3.36e-162	385	202
Entropy	1.45e-189	4.82e-145	8.87e-123	5.47e-139	1.18e-176	1.45e-189	378

2.2 Effect on gene Coexpression

We also identified the effects on gene coexpression profiles in iPS cells. To explore the relationship among these key transcriptional factors (TFs) of iPS cells, such as Pou5f1, Sox2, Klf4, c-Myc, Nanog and Lin28, which can be achieved by overexpression of these TFs by direct reprogramming of somatic cells, we collected such 13 TFs from literatures [1, 13, 15]. We identified the gene correlations for deciphering their relationships with the other genes and identifying the effect of different methods to represent the expression of genes with multiple probesets. For simplicity, we considered the 900 differential genes with multiple probesets in the former section identified by at least one method. Figure 2 shows the boxplots of the correlation value between these TFs and differential expressed genes. From Figure 2, we identified the correlations of the differential genes with the TFs. There are few differences in these correlations which are detected from the gene expression transformed from the probesets by different methods. 'Random', 'Variance' and 'Entropy' methods lead to slightly smaller correlation values than that of other methods.

We also identified the differences underlying the weighted gene coexpression networks. We used the correlation test to identify the significant correlations between the TFs by a threshold of p-value 0.05. The correlations between these transcription factors are measured by the same scheme. Figure 3 shows the dynamics of these constructed gene coexpression networks in these genes by the seven methods. In the TFs' coexpression networks, there are 5 genes which are represented by multiple probesets. Most of the significant correlation are identified by 'All' the methods. By different methods, there are slight differences of these correlations. For instance, the correlation between 'Lin28'

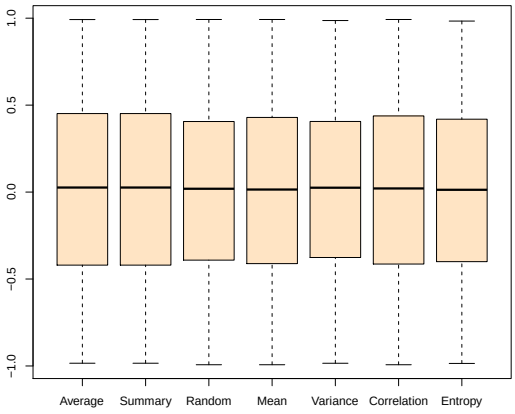

Figure 2: Box plots of correlation values between some transcriptional factors and differential genes by different methods to interpret the multiple probesets.

and ‘Nanog’ are significant by the methods ‘Random’, ‘Correlation’ and ‘Entropy’. From the effects on the gene coexpression of these key TFs in iPS cells, we can find that there are certain effects of these multiple probesets in decision of their gene coexpressions by different methods.

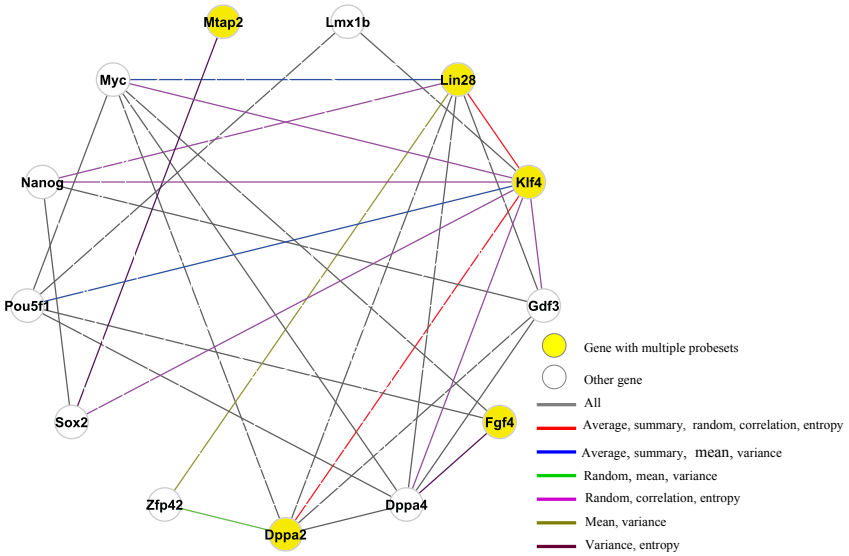

Figure 3: Gene coexpression networks constructed by the significant correlation between the TFs by different methods.

2.3 Effect of multiple probesets on expression profiles in iPS cells

We analyzed the effects of multiple probesets to identify the differential genes from Affymetrix GeneChips in iPS cells. There are slight differences in the results from the different methods. We have tested seven methods to interpret the information of multiple probesets to genes. In the dataset, there are generally weak effects on the global performance about the whole chip. However, when we focus on a specific expression profile of individual genes, there are certain effects. There are some genes whose profiles are significantly affected by the methods to transform the multiple probesets to genes. Figure 4 (a) shows different p-values of some genes from the expressions by the different methods. The profiles of some genes, such as 'Atg5' and 'Tapppc4', are significantly affected by these methods. Figure 4 (b) is the p-values of the known TFs. There are 5 genes with multiple probesets in these TFs, which are slightly affected by different methods.

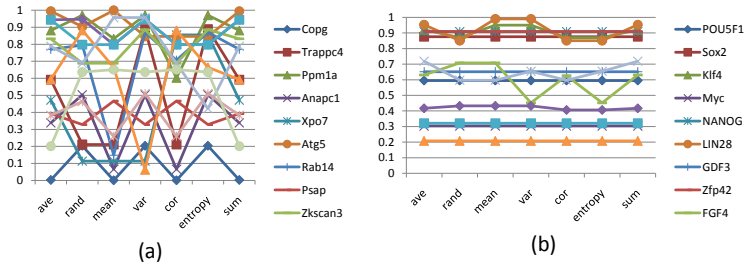

Figure 4: Differential p-values of some genes. (a) 9 randomly selected genes. (b) the known TFs.

For the seven methods, Figure 5 shows the hierarchical clustering of these methods based on their distances of p-values of gene differential expressions. From the similarity between these methods of 'Average' and 'Summary', we can recognize that their effects are identical in identifying differential genes. 'Correlation' method is also very close to 'Mean' method. This indicates that the probeset with maximum correlation with others are close to the probeset with maximum mean value. The similarity and difference between these methods are presented in the figure.

For analyzing the GO enrichment, we also identified the difference of functional enrichments of these differential genes detected by different methods. Table 3 lists the enriched GO terms by the seven methods. From the significant GO (p-value of hypergeometric test p-value threshold of $10^{-}3$, level threshold of 4), we can identify the special features of functional enrichments in iPS cells with comparison with ES cells. We identified these enriched biology process ('BP'), molecular function ('MF') and cellular component ('CC') in every differential gene sets detected by the methods respectively. We found that the effect on the decision of functional enrichment in differential genes will be obvious in the analysis. By the thresholds, there are no significant GO terms in the gene sets of 'Correlation' method. In Table 3, the differential genes related to 'mitochondrion' are differential expressed as well as some regulation processes of metabolism.

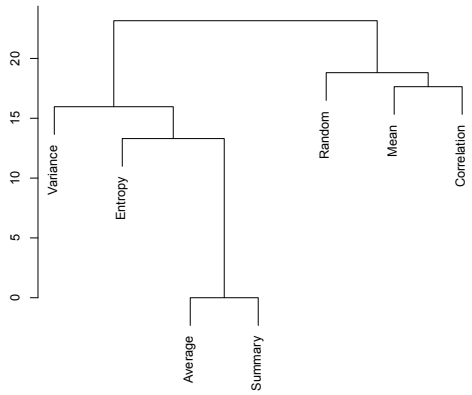

Figure 5: Hierarchical clustering of these methods to interpret multiple probesets.

Table 3: GO enrichments in differential genes identified by the seven methods.

Method	Term	P-value	Ontology	Description
Average/Summary	GO:0005634	1.67E-08	CC	nucleus
	GO:0048522	4.81E-07	BP	positive regulation of cellular process
	GO:0080090	6.86E-07	BP	regulation of primary metabolic process
	GO:0051246	8.04E-06	BP	regulation of protein metabolic process
	GO:0060255	9.77E-06	BP	regulation of macromolecule metabolic process
	GO:0009893	1.96E-05	BP	positive regulation of metabolic process
	GO:0031625	2.79E-05	MF	ubiquitin protein ligase binding
Random	GO:0005739	1.72E-06	CC	mitochondrion
Mean	GO:0005739	2.06E-05	CC	mitochondrion
Variance	GO:0048522	1.41E-06	BP	positive regulation of cellular process
	GO:0009893	5.21E-06	BP	positive regulation of metabolic process
	GO:0031325	1.65E-05	BP	positive regulation of cellular metabolic process
	GO:0010646	2.70E-05	BP	regulation of cell communication
Entropy	GO:0048522	2.91E-05	BP	positive regulation of cellular process
	GO:0005739	3.97E-06	CC	mitochondrion

3 Conclusion

In this work, we presented a comparison study of seven methods for transforming the information of multiple probesets into gene expression. We found that the global effects on the analysis of differential expression and coexpression are minor. However, the methods for interpreting these multiple probesets to gene level of specific genes significantly affect the results in iPS cells. This implies that it is crucial to assess the expressions of the selected genes when they have multiple probesets. The available methods to handle the multiple probesets were summarized. The results also suggest the importance to interpret the multiple probesets to gene expression which provides more information about the gene expression when we implement these different methods.

4 Methods

4.1 Data sources

We used the data of gene expression profiling research on iPS cells and ES cells [11]. The gene expression datasets were downloaded from NCBI Gene Expression Omnibus

(GEO, http://www.ncbi.nlm.nih.gov/geo/) database (ID:GSE20576). The 4 samples of ES cell and first 8 samples of iPS cell were transformed to absolute expression value by RMA algorithm in R Bioconductor. Probesets were mapped to NCBI entrez genes using DAVID [3]. The expression data contained 22716 probes and resulted in 15706 genes.

4.2 Multiple probesets

There are 6219 genes which correspond to multiple probesets. If there are multiple probesets corresponding to the same genes, we will use the following methods to decipher the expression individually from probe level to gene level.

- Average: In many papers of gene expression analysis, when there are multiple probesets corresponding to a gene. These probesets are averaged to represent the gene expression.
- Summary: We also can summarize the probesets corresponding to one gene together to represent the gene expression.
- Random: An alternative method is to randomly choose one of these probesets and use the selected one as the representation of the gene expression.
- Mean: We can choose the probesets with maximum mean value as the gene expression.
- Variance: The variance indicates the dynamics of the expression in the whole list of gene expressions. We can choose the one that corresponds to that with maximum variance in these multiple probesets.
- Correlation: We calculate the correlation value of one probeset with the others and then summarize these correlations. The one with largest summary of correlation is interpreted as the gene expression.
- Entropy: We identify the probeset with maximum entropy in these candidates. The entropy is defined as $E = -\sum_i p_i log_2 p_i$, where p_i is the percentage of the expression in all the samples.

4.3 Detecting differential expression and coexpression

After we fixed the expression profiling for each gene, we identified the differentially expressed p-values by comparing the gene expressions of iPS cells with those of ES cells by Welch's two-tailed t-test. We identified the differential information of gene expressions from the seven methods respectively. The gene coexpression information in iPS cells was calculated by Pearson correlation coefficient (PCC). We also calculated the PCC value of pairwise genes in the seven generated gene expressions.

4.4 Deciphering the effects of multiple probesets

To detect the effects of multiple probesets, we identified the effects of these multiple probesets by applying different methods to transform the probeset information to genes. Firstly, we compared these differential p-values by different methods. Then we compared the different correlation coefficient values with the 13 known transcriptional factors related pluripotent cells listed in literatures [13, 15, 1, 9].

Acknowledgements

This work was supported by the Chief Scientist Program of Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences with Grant No. 2009CSP002, and

supported by the National Natural Science Foundation of China (NSFC) under Grant No. 10801131 and Grant No. 60873205.

References

[1] Chin MH, Mason MJ, Xie W, et al.: Induced pluripotent stem cells and embryonic stem cells are distinguished by gene expression signatures. Cell Stem Cell, 5:111–123, 2009.

[2] Dai M, Wang P, Boyd AD, et al.: Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. Nucl Acids Res, 33:e175, 2005.

[3] Dennis G Jr, Sherman BT, Hosack DA, et al.: DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol, 4:P3, 2003.

[4] Gautier L, Moller M, Friis-Hansen L, Knudsen S: Alternative mapping of probes to genes for Affymetrix chips. BMC Bioinformatics, 5:111, 2004.

[5] Liu H, Zeeberg BR, Qu G, et al.: AffyProbeMiner: a web resource for computing or retrieving accurately redefined Affymetrix probe sets. Bioinformatics, 23:2385–2390, 2007.

[6] Liu G, Loraine AE, Shigeta R, et al.: NetAffx: Affymetrix probesets and annotations. Nucleic Acids Res, 31:82–86, 2003.

[7] Lu X, Zhang X: The effect of GeneChip gene definitions on the microarray study of cancers. Bioessays, 28:739–746, 2006.

[8] MAQC Consortium: The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat Biotechnol, 24:1151–1161, 2006.

[9] Newman AM, Cooper JB: AutoSOME: a clustering method for identifying gene expression modules without prior knowledge of cluster number. BMC Bioinformatics, 11:117, 2010.

[10] Quackenbush J: Microarray data normalization and transformation. Nature Genetics, 32:496–501, 2002.

[11] Stadtfeld M, Apostolou E, Akutsu H, Fukuda A, et al.: Aberrant silencing of imprinted genes on chromosome 12qF1 in mouse induced pluripotent stem cells. Nature, 465, 175–181.

[12] Stalteri MA, Harrison AP: Interpretation of multiple probe sets mapping to the same gene in Affymetrix GeneChips. BMC Bioinformatics, 8:13, 2007.

[13] Takahashi K, Yamanaka S: Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell. 126:663–676, 2006.

[14] Yin J, McLoughlin S, Jeffery IB, Glaviano A, Kennedy B, Higgins DG: Integrating multiple genome annotation databases improves the interpretation of microarray gene expression data. BMC Genomics, 11:50, 2010.

[15] Yu J, Hu K, Smuga-Otto K, et al.: Human induced pluripotent stem cells free of vector and transgene sequences. Science, 324:797–801, 2009.