Journal of Animal and Veterinary Advances 11 (24): 4593-4598, 2012
ISSN: 1680-5593

Comparison of Effect of Non-Specific Filtering Methods and their Combinations on GeneChip Data Analysis: A Case for Affymetrix Porcine Genome Microarray

[1]Ming Zhao, [2]Hong-Bo Chen, [1]Xiang-Dong Liu, [1]Chao Wang,
[1]Jing-Ying Guo, [1]Min-Hui He, [1]Shu-Hong Zhao and [1]Meng-Jin Zhu
[1]Key Lab of Agricultural Animal Genetics, Breeding and Reproduction, Ministry of Education,
Huazhong Agricultural University, 430070 Wuhan, P.R. China
[2]School of Animal Science and Nutritional Engineering,
Wuhan Polytechnic University, 430023 Wuhan, Hubei, P.R. China

Abstract: Microarray technology has been widely applied in the research area of animal genetics and breeding. In the earlier stage of microarray data analysis, the non-specific filtering is a common procedure which has been used increase the detection rate of differentially expressed genes. In this investigation, researchers use the data from Affymetrix GeneChip Porcine Genome Array to comparatively assess the effects of different non-specific filtering methods on the results of transcriptome analysis. The research results showed that the MAS5, ET and IQR filters could increase but the I/NI-calls method could decrease the number of differentially expressed genes. Here into, the IQR filter has the largest detection rate. Furthermore, the two-way combinations with IQR filter have the similar increasing effect on differentially expressed gene detection. It is concluded that not all non-specific filtering methods could increase the detection rate in microarray data analyses and the IQR filter could be considered as a preferred choice when improvement in the detection rate is needed on some occasions.

Key words: Non-specific filtering, differentially expressed gene, Affymetrix GeneChip data, pig, China

INTRODUCTION

In the post-genome era, *gene* chip has been widely applied in the research area of animal genetics and breeding. Among the commercial gene chips, Affymetrix high-density oligonucleotide microarrays have been frequently used for genome-wide gene expression measurements in a variety of fields (Auer *et al*., 2009). In the *Affymetrix* gene chip analysis, it is well known that at the low-level processing stage, there are many different preprocessing algorithms and statistical choices for raw data analysis (Mieczkowski *et al*., 2010). These low-level processing processes include quality control, background correction, PM correction, summarization, normalization, probeset filtering, etc. (Gohlmann and Talloen, 2009; Kauffmann and Huber, 2010). Hereinto, the non-specific (or unsupervised) filtering of probes, usually corresponding to genes is a specifically recommended step in the analytical pipeline for microarray data treatment (Wu and Irizarry, 2004). The execution of non-specific filtering is a routine procedure for removing or reducing the fraction of unreliable genes (probes) that poorly perform and show low overall intensity or variability across all arrays without regard to sample class label (s).

The main purpose of non-specific filtering is to simplify the microarray analysis by eliminating uninformative probes most unlikely to be of biological interest but by not interfering with formal statistical testing (Talloen *et al*., 2007; Bourgon *et al*., 2010). It has been shown by studies that the low variable probes bring an increase in the number of tests and a corresponding reduction in power and filtering out those probes without informative variation can much lessen the computational burden (Lusa *et al*., 2008) which usually can offer better solutions of the high-level processing problems for microarray data analysis. For instance, the filtering performed by detection call and variance consistently increased the number of discoveries (Hackstadt and Hess, 2009). McClintick and Edenberg (2006) also found that without filtering, large number of tests decreased the proportion of differentially expressed genes which were even truly differentially expressed.

Up to date, according to different platforms, many non-specific filtering methods have been proposed and

Corresponding Author: Meng-Jin Zhu, Key Laboratory of Agricultural Animal Genetics, Breeding and Reproduction, Ministry of Education, Huazhong Agricultural University, 430070 Wuhan, P.R. China

of which the members usually include the feature, intensity and variability-based filterings, respectively. The feature-based filtering does ordinarily winnow out the control probes or those probes without annotation ID (e.g., Entrez gene ID, GO term ID or other annotation ID from the public databases); the purpose of intensity-based filtering is to exclude the fraction of probes by detection call (e.g., MAS5, I/NI calls for 3'IVT arrays and DABG call for exon arrays) or a user-defined expression threshold (Hubbell *et al*., 2002; McClintick and Edenberg, 2006; Talloen *et al*., 2007; Okoniewski *et al*., 2007; Archer and Reese, 2010) and the variability-based filtering filters out the out-of-filter probesets by some measures of the variability across arrays (e.g., CV, SD, normality, MAD and IQR filters). In a microarray experiment while its absolute validity was still under doubt by some researchers, it has been widely accepted that an appropriate filter can increase but a poor choice of filter can actually reduce detection rate or power of gene discoveries (McClintick and Edenberg, 2006; Bourgon *et al*., 2010; Kauffmann and Huber, 2010).

Several filtering methods have been widely applied in a variety of studies while there were few studies conducted to compare the influences of them and their combinations. Despite pioneering research by Bourgon *et al*. (2010), hitherto there was no general principle that could provide a unique guideline to ensure the appropriate choice of non-specific filtering methods. When performing non-specific filtering, how to choose one or a particular combination of some of them is an ultimately unsolved but very important issue in microarray analysis. Since, the general principle having not yet been formulated, a comprehensive comparison of the commonly used non-specific filtering methods may provide an alternative solution. In this investigation, researchers use the data from Affymetrix GeneChip Porcine Genome Array to provide a comparative assessment on the effects of different non-specific filtering methods and their combinations on the transcriptome analysis.

MATERIALS AND METHODS

Microarray datasets from affymetrix porcine GeneChip®: The Affymetrix microarray datasets from the lab were used for filtering effect evaluation. The gene expression datasets came from a study of the immune stimulus by *Haemophilus parasuis* (shortened as *H. parasuis* or HPS) in porcine spleen which were generated with Affymetrix GeneChip Porcine Genome Array. The detailed information about the experimental design, animal managements, tissue collection and microarray hybridizations were earlier reported by Chen *et al*. (2009). The CELL files for raw data are accessible through GEO accession number GSE11787 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token=zlkrjsiewkoamno&acc=GSE11787).

Filtering methods and filtering parameters: The commonly used non-specific filtering methods for Affymetrix microarray data include A/M/P call filtering, non-informative call filtering, expression threshold filtering (filtering on absolute expression values) and variance-based filtering methods. Here, MAS5-calls, I/NI-calls, Expression Threshold (ET) filter and Interquartile Range (IQR) filter, corresponding to the above-mentioned filtering methods and the combined filtering methods of MAS5+IQR, I/NI+IQR and ET+IQR were involved in the investigation. All filtering operations were done on the ExpressionSet object when after converting AffyBatch object into ExpressionSet object. The var.cutoff for varFilter function was set to 0.5 in the IQR as well as in its combined filtering processes and considering the log2 transform of expression values, the filtering threshold for ET filter was set to 6.5 (approximately corresponding to a raw signal intensity value of 90). Finally, the variance estimations and the numbers of Differentially Expressed Genes (DEGs) were used to intuitionistically evaluate the effects of different filtering methods on the results of transcriptome analyses.

Software and tools: The Affymetrix technical files used in this study were downloaded from http://www.affymetrix.com/support/index.affx. All analyses including data filtering were processed and analyzed with the open source R software packages (http://www.r-project.org) and tools from the BioConductor project (http://www.bioconductor.org) (Gentleman *et al*., 2004). The basic R package stats and bioconductor packages Affy, farms and genefilter were used to perform the MAS5-calls, I/NI-calls, ET and IQR filtering, respectively. Except for the I/NI-calling data, the Robust Multichip Average algorithm (RMA) was used to preprocess the raw Affymetrix microarray data and to obtain the expression summary for each gene on each chip. For the validity of subsequent comparisons, the (pre) processing parameters for the function exp.farms were set to the same ones as in RMA, i.e., the background correction method was RMA, the PM correction method was pmonly, the normalization method was quantiles and the summarization method was medianpolish. After filtering, package limma was used to detect the differentially expressed genes.

RESULTS AND DISCUSSION

Characterization of the raw chip data: The raw data of six Affymetrix chips came from the previous study

(Chen *et al*., 2009) in which the spleen tissues of three HPS infected piglets and three controls were separately used for the microarray hybridization. The cell files were produced by the Affymetrix GeneChip® Porcine Genome Array which contains 24,123 probesets including 20,201 genes and the control probesets (http://www.affymetrix.com). Descriptive plots of raw chip data from six samples were shown in Fig. 1 in which the boxplot, density plot,

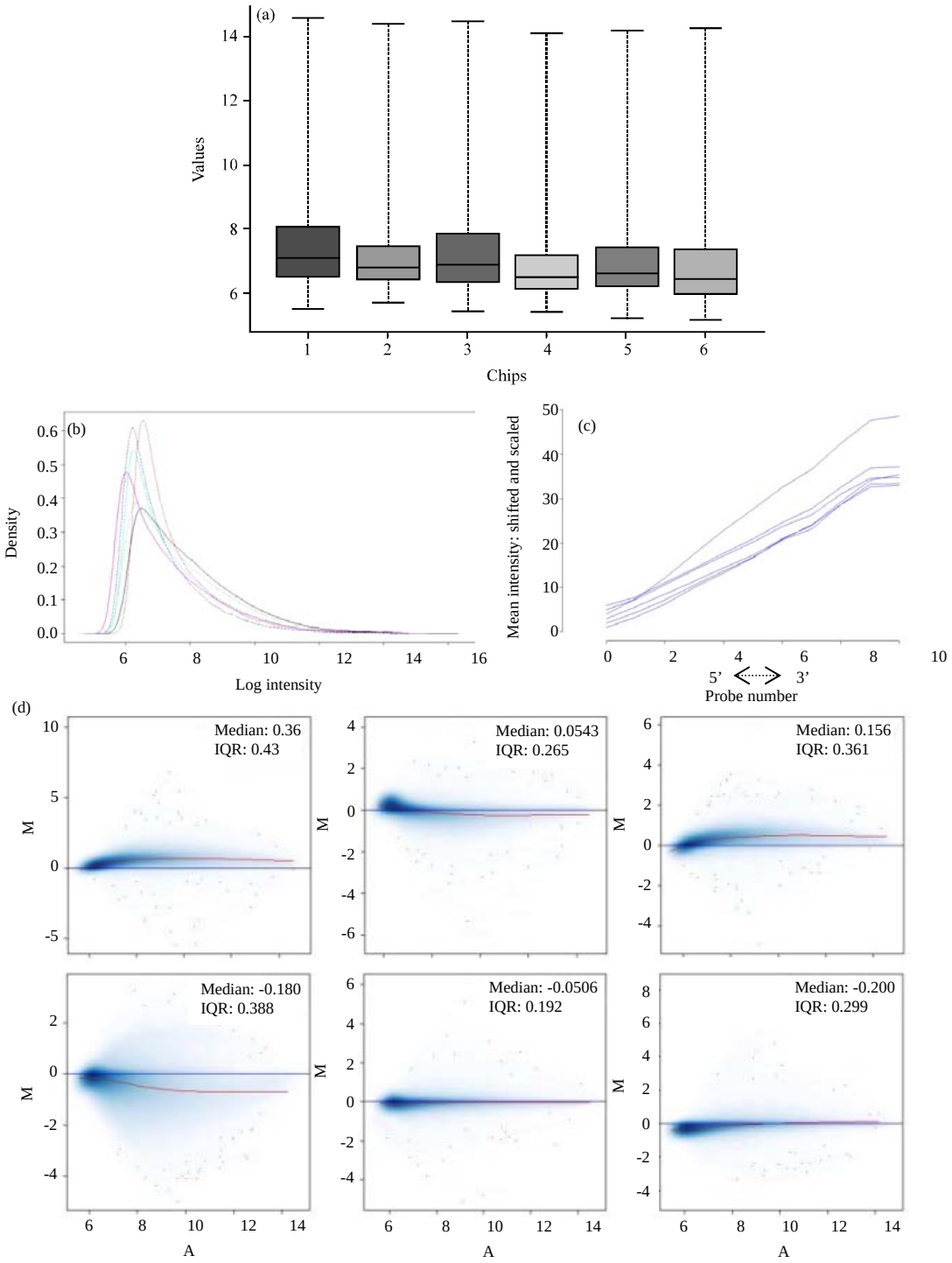

Fig. 1: Descriptive plots of raw chip data from six samples; a) boxplots of raw data for six chips; b) density plots of raw data for six chips; c) RNA degradation plots for six chips; d) MA plots of raw data for six chips

RNA degradation plot and MAplot were included, respectively. Seen from them, all chips in the plots had a similar appearance and it can be perorated that the microarray experiments were technically reliable. Table 1 also displayed the descriptive statistical parameters of raw chip data from six samples. In Table 1, arrays 1-3 are three control samples and arrays 4-6 are three HPS-infected samples. Except for the mean (average value) and the third quantile of array 2 and the maxima of arrays 2 and 3, all of the estimations of the descriptive statistical parameters from the control samples were larger than the ones from the HPS-infected samples.

Assessment on the effect of different non-specific filtering methods: Four individual non-specific filtering methods and three two-way combined methods were used to comparatively perform the non-specific filtering on chip data. Because the algorithms of ET, MAS5 and I/NI calls are incompatible, their combinations were not considered. Given this only three combinatory methods including MAS5+IQR, I/NI+IQR and ET+IQR were proposed.

Here two indicators, variance and number of differentially expressed genes were used to evaluate the effects of different non-specific filtering methods on the transcriptome analysis. First, the comparison of estimated variance of gene expressions on each microarray among different non-specific filtering methods and their combinations was shown in Table 2 in which the unfiltered raw data had the largest estimation of variance. It can be found that in the filtered data, MAS5 had the largest variance and the ET filter and its combined filter (ET+IQR) had the smallest ones. In the statistical principle, ET filter uses one-sided truncation strategy (one-tailed cut-off) according to the gene expression threshold which usually results in a skew distribution of the gene expression data. This means that ET filter is not an appropriate method for non-specific filtering of chip data.

To evaluate the concrete effects of different non-specific filtering methods on the transcriptome analysis, the number of differentially expressed genes (or transcripts) was further investigated. The detection of differentially expressed genes was achieved through the empirical Bayes method by package limma of Bioconductor (http://www.bioconductor.org/packages/release/bioc/html/limma.html). The results were shown in Table 3. Researchers found that the number of differentially expressed genes differed much between different non-specific filtering methods. In total, the data produced by IQR filter detected the largest number of differentially expressed genes and I/NI-calls method had the smallest number. Furthermore, it is interesting that the combined utilization of IQR filter with the other filters can increase the number of differentially expressed genes. For example, compared with MAS5 calls method, the MAS5+IQR combined filtering method can increase 14 differentially expressed genes (41 vs. 55 in Table 3). To make an integrated consideration of all filtering results, the IQR filter method could increase the detection sensitivity of differentially expressed genes in the microarray data analyses. The ET filter and I/NI calls method seem to produce an over-filtering result and the data filtered by them would lose some useful information and decrease detection rate of differentially expressed genes.

The transcriptome analysis, one important issue of agricultural animal genomics can be used to reveal the molecular basis and mine the candidate genes underlying the economically important traits of agricultural animals. There is an increasing trend of applications of transcriptome analysis in the research field of animal genetics and breeding. Up to date, the expression gene chips have provided one of the major approaches to perform transcriptome analysis in this field. The GeneChip data analysis is a complicated process that is associated with many statistical techniques and tools. To increase the sensitivity and power for microarray data analyses or

Table 1: Descriptive statistics of raw chip data for six samples

Statistical parameters*	Array1	Array2	Array3	Array4	Array5	Array6
Min.	40.0	44.0	37.0	36.0	28.0	32.0
1st Qu	90.0	88.0	80.0	70.0	74.0	63.0
Median	137.0	111.0	119.0	90.0	98.0	88.0
Mean	424.4	280.7	372.5	227.2	285.2	279.5
3rd Qu	279.0	177.0	238.0	146.0	177.0	167.0
Max.	65532.0	30255.0	32901.0	27045.0	28390.0	33018.0

*Min. = Minimum, 1st Qu = The first Quantile, 3rd = The third quantile and Max. = maximum

Table 2: Variance of expression values after filtering by different methods and their combinations

Filtering methods	Array1	Array2	Array3	Array4	Array5	Array6
Raw data	5.004936	4.876620	5.014238	4.566523	5.047146	5.062504
MAS5 calls	4.620724	4.549196	4.664633	4.363491	4.647360	4.642044
I/NI-calls	3.843309	3.772213	3.980755	4.216792	3.734367	3.657621
ET filter	2.105835	2.110668	2.090716	2.500279	2.088920	2.127427
IQR filter	4.306315	4.164188	4.310459	3.784453	4.355271	4.372256
MAS5+IQR	3.789337	3.711531	3.829515	3.534656	3.796335	3.788257
I/NI+IQR	3.258514	3.178498	3.366006	3.837222	3.097423	3.045961
ET+IQR	1.843818	1.849387	1.825520	2.426095	1.826146	1.872390

Table 3: Numbers of differentially expressed genes for different non-specific filtering methods and their combinations

Methods	No. of retained probesets	No. of DEGs
Unfiltered data	24,123	34
MAS5 calls	17,545	41
I/NI-calls	8,191	21
ET filter	9,293	37
IQR filter	12,061	58
MAS5+IQR	8,772	55
I/NI+IQR	4,095	33
ET+IQR	4,646	41

even other analysis (Jiang and Gentleman, 2007), the non-specific filtering methods have been commonly used at the low-level processing stage. But in the research field of agricultural animal genomics, the appropriate applications of non-specific filtering methods have been not taken into account on most occasions.

The non-specific filtering is a statistical process for excluding or including a subset of genes or probes which does not utilize the sample class labels (or phenotypic covariates). This process could remove the invalid genes (probes) that commonly have the lowest between-array variability. An appropriate application of non-specific filtering could increase the sensitivity and detection rate of differentially expressed genes through improvement to p-value adjustment while a poor choice of filter statistic could reduce the detection power (Bourgon *et al*., 2010). Thus, to increase the discoveries of *DE* genes, it is desirable to filter out the probes (genes) that can't provide valid information (Calza *et al*., 2007). In this study, four individual non-specific filtering methods (MAS5-calls, I/NI-calls, ET filter and IQR filter) and three two-way combined methods (MAS5+IQR, I/NI+IQR and ET+IQR) were considered. Researchers have provided a comparative assessment on their effects on the transcriptome analysis of Affymetrix porcine genome microarray in which the consideration of different combination of non-specific filtering methods is a new attempt and provides a uniqueness of the investigation.

The research results showed that except for the I/NI-calls method, all of the investigated non-specific filtering methods could increase the detection rate of differentially expressed genes. The I/NI method uses the Informative/Non-Informative (I/NI) calls to perform filtering operation on the raw microarray data (Talloen *et al*., 2007) and in the study the results indicated that the Informative/Non-Informative (I/NI) calls had the smallest detection rate of differentially expressed genes. Among four non-specific filtering methods, the IQR filter has the largest detection rate. The IQR filter essentially belongs to the variance-based filters. In fact, the variance-based filters are one of the most commonly used methods in the expression analyses because after filtering, they usually do not bias the distribution of microarray data (Prieto *et al*., 2008). It is also interesting that being compared with individual methods, the two-way combinations with IQR filter could increase the discoveries of differentially expressed genes. This means that, if we can control the false positive rate, the IQR filter is an ideal filter to increase the detection rate of differentially expressed genes.

CONCLUSION

It is concluded from the research results that it is not all non-specific filtering methods could increase the sensitivity and power for microarray data analyses. The non-specific filters that use the Informative/Non-Informative (I/NI) calls or bias the microarray data could decrease the sensitivity of detection of differentially expressed genes. Among the investigated filters, the IQR filter is an ideal filter to increase the discoveries of differentially expressed genes. When the number of differentially expressed genes is relatively small and improvement in the detection rate is needed on some occasions, the IQR filter could be considered as a preferred non-specific filtering method at the low-level processing stage of microarray data analysis.

ACKNOWLEDGEMENTS

This research was supported by the National Natural Science Foundation of China (30901021) and the National High Technology Research and Development Program of China (863 Program) (2013AA102502).

REFERENCES

Archer, K.J. and S.E. Reese, 2010. Detection call algorithms for high-throughput gene expression microarray data. Brief Bioinfo., 11: 244-252.

Auer, H., D.L. Newsom and K. Kornacker, 2009. Expression profiling using affymetrix genechip microarrays. Methods Mol. Biol., 509: 35-46.

Bourgon, R., R. Gentleman and W. Huber, 2010. Independent filtering increases detection power for high-throughput experiments. Proc. Natl. Acad. Sci. USA., 107: 9546-9551.

Calza, S., W. Raffelsberger, A. Ploner, J. Sahel and T. Leveillard *et al*., 2007. Filtering genes to improve sensitivity in oligonucleotide microarray data analysis. Nucleic. Acids. Res., Vol. 35. 10.1093/nar/gkm537.

Chen, H., C. Li, M. Fang, M. Zhu and X. Li *et al*., 2009. Understanding Haemophilus parasuis infection in porcine spleen through a transcriptomics approach. BMC Gen., Vol. 10. 10.1186/1471-2164-10-64.

Gentleman, R.C., V.J. Carey, D.M. Bates, B. Bolstad and M. Dettling *et al*., 2004. Bioconductor: Open software development for computational biology and bioinformatics. Genome Biol., Vol. 5. 10.1186/gb-2004-5-10-r80.

Gohlmann, H. and W. Talloen, 2009. Gene Expression Studies Using Affymetrix Microarrays. Chapman, Hall, London, ISBN-13: 978-1420065152, Pages: 359.

Hackstadt, A.J. and A.M. Hess, 2009. Filtering for increased power for microarray data analysis. BMC Bioinfo., Vol. 10. 10.1186/1471-2105-10-11.

Hubbell, E., W.M. Liu and R. Mei, 2002. Robust estimators for expression analysis. Bioinformatics, 18: 1585-1592.

Jiang, Z. and R. Gentleman, 2007. Extensions to gene set enrichment. Bioinformatics, 23: 306-313.

Kauffmann, A. and W. Huber, 2010. Microarray data quality control improves the detection of differentially expressed genes. Genomics, 95: 138-145.

Lusa, L., E.L. Korn and L.M. McShane, 2008. A class comparison method with filtering-enhanced variable selection for high-dimensional data sets. Stat Med., 27: 5834-5849.

McClintick, J.N. and H.J. Edenberg, 2006. Effects of filtering by Present call on analysis of microarray experiments. BMC Bioinf., Vol. 7. 10.1186/1471-2105-7-49.

Mieczkowski, J., M.E. Tyburczy, M. Dabrowski and P. Pokarowski, 2010. Probe set filtering increases correlation between Affymetrix GeneChip and qRT-PCR expression measurements. BMC Bioinfo., Vol. 11. 10.1186/1471-2105-11-104.

Okoniewski, M.J., Y. Hey, S.D. Pepper and C.J. Miller, 2007. High correspondence between Affymetrix exon and standard expression arrays. Biotechniques, 42: 181-185.

Prieto, C., A. Risueno, C. Fontanillo and J. de las Rivas, 2008. Human gene coexpression landscape: Confident network derived from tissue transcriptomic profiles. PLoS One, Vol. 3. 10.1371/journal.

Talloen, W., D.A. Clevert, S. Hochreiter, D. Amaratunga, L. Bijnens, S. Kass and H.W. Gohlmann, 2007. I/NI-calls for the exclusion of non-informative genes: A highly effective filtering tool for microarray data. Bioinformatics, 23: 2897-2902.

Wu, Z. and R.A. Irizarry, 2004. Preprocessing of oligonucleotide array data. Natr. Biotechnology, 22: 656-658.