ACTA ABP BIOCHIMICA POLONICA

Vol. 60, No 4/2013
647–655
on-line at: www.actabp.pl

Regular paper

Microarray Inspector: tissue cross contamination detection tool for microarray data*

Piotr Stępniak[1]✉, Matthew Maycock[1], Konrad Wojdan[1,2], Monika Markowska[1,3], Serhiy Perun[1,4], Aashish Srivastava[5], Lucjan S. Wyrwicz[5] and Konrad Świrski[1,2]

[1]Transition Technologies S.A., Warszawa, Poland; [2]Institute of Heat Engineering, Warsaw University of Technology, Warszawa, Poland; [3]Department of Gastroenterology and Hepatology, Medical Center for Postgraduate Education, Warsaw, Poland; [4]Institute of Physics PAS, Warszawa, Poland; [5]Laboratory of Bioinformatics and Biostatistics, Maria Skłodowska-Curie Memorial Cancer Center and Institute of Oncology, Warszawa, Poland

Microarray technology changed the landscape of contemporary life sciences by providing vast amounts of expression data. Researchers are building up repositories of experiment results with various conditions and samples which serve the scientific community as a precious resource. Ensuring that the sample is of high quality is of utmost importance to this effort. The task is complicated by the fact that in many cases datasets lack information concerning pre-experimental quality assessment. Transcription profiling of tissue samples may be invalidated by an error caused by heterogeneity of the material. The risk of tissue cross contamination is especially high in oncological studies, where it is often difficult to extract the sample. Therefore, there is a need of developing a method detecting tissue contamination in a post-experimental phase. We propose Microarray Inspector: customizable, user-friendly software that enables easy detection of samples containing mixed tissue types. The advantage of the tool is that it uses raw expression data files and analyses each array independently. In addition, the system allows the user to adjust the criteria of the analysis to conform to individual needs and research requirements. The final output of the program contains comfortable to read reports about tissue contamination assessment with detailed information about the test parameters and results. Microarray Inspector provides a list of contaminant biomarkers needed in the analysis of adipose tissue contamination. Using real data (datasets from public repositories) and our tool, we confirmed high specificity of the software in detecting contamination. The results indicated the presence of adipose tissue admixture in a range from approximately 4% to 13% in several tested surgical samples.

Key words: microarray, transcription profiling, contamination analysis, adipose tissue, cancer, data quality

Received: 21 September, 2013; revised: 25 November, 2013; accepted: 15 December, 2013; available on-line: 29 December, 2013

INTRODUCTION

Microarrays have provided a tremendous amount of interesting data, but as a tool they pose technical difficulties in experiment execution, which raises certain questions pertaining to the method's reliability (Tan, 2003; Shi *et al*., 2004; Michiels *et al*, 2005; Ioannidis, 2005; Dupuy & Simon, 2007). Large scale microarray quality assessment projects such as MAQC (Shi *et al*. 2006), have confirmed that it is possible to obtain reproducible results, even between various platforms (Chen *et al*. 2007). Nevertheless, common standards for both experiment preparation and data analysis are required (Tan, 2003; Dupuy, 2007; Ji *et al*., 2006; Shi *et al*., 2010). Still, some side effects are to be expected (Chen *et al*., 2007), as each step of the procedure (sample extraction, storage, preparation, hybridization, washing) can introduce an error (Ji *et al*., 2006). On the other hand, a positive result in MAQC-II study reports that different data analysis approaches produce comparable predictive models for a given dataset, thus confirming that applying enough quality measures can yield reliable data for analysis.

Quality control in data analysis is usually integrated with normalisation (Affymetrix; Irrizarry, 2003) and considers mainly RNA integrity and technical problems of hybridization such as spatial and probe effects. The methods and tools provided by microarray suppliers and the research community, allow conducting tests of signal intensity, average background noise, percent of present calls verification, RNA fragmentation assessment, and in some cases, sample and lab effects (Affymetrix; Affymetrix I, 2002; Irizarry, 2003; Bolstad *et al*., 2004; McCall *et al*., 2011). Surprisingly, the problem of tissue contamination and its consequences on data quality is usually ignored. Nevertheless, as already emphasised by researchers (McCall *et al*., 2011), there is a great need for dataset independent (single array) quality assessment methods, especially considering that 10% of publicly available datasets are estimated to be corrupted.

The general disadvantage of the above methods is that biologically atypical samples can be incorrectly considered to be of poor quality. The methods test only if a sample deviates from other samples in the experiment, and provide no information or explanation of the cause. This situation leads to two classes of errors: 1. removing from further analysis results that are proper, but which represent an atypical biological image, and 2. forming a set of results that were classified as "good/clean", but can in fact represent a homogeneous set of contaminated results (contaminated by other tissues). For example, the extracted RNA was of good quality, but came from

✉e-mail: P.Stepniak@tt.com.pl

*The software has been presented as posters on the following conferences: Advances in Microarray Technology, 5–6.03.2013, Barcelona, Spain. RECOMB 2013, 17th Annual International Conference on Research in Computational Molecular Biology, 7-10.04.2013, Beijing, China.

Abbreviations: MAQC, microarray quality control; qRT-PCR, quantitative reverse transcription polymerase chain reaction; TiSGeD, the tissue-specific genes database; GEO, gene expression omnibus

two or more cell types, e.g., tumour and surrounding adipose tissue.

The current paradigm deems it necessary to confirm tissue sample integrity before hybridization and examining its morphology after dissection (Skrzypczak *et al.*, 2010). This step is time-consuming and, therefore, sometimes omitted. False microarray findings can be identified after hybridization, when attempting to confirm them with qRT-PCR. This procedure, however, is resource-consuming and thus limited to a small number of genes of utmost interest. Furthermore, both morphological examination and qRT-PCR require the same sample which was used for hybridization. Considering that in clinical and diagnostic studies, the entire sample is often used for the microarray, it is not possible to perform such tests. Similarly, when file sets from public repositories are to be analysed, computational techniques remain the only available test for the quality of microarray samples.

Thus, we see a need to develop a method that is able to recognize tissue contamination in a post-experimental phase. Failing to detect contamination at the early stage of microarray data analysis can lead to non-representative conclusions and, in consequence, further costly research which produces inaccurate results. In this paper, we present a method applied in a flexible, user-friendly software called Microarray Inspector. It is available for free for non-commercial research at our site: http://bioinformatics.tt.com.pl/.

MATERIALS AND METHODS

Contaminant tissue definition. The basis of an accurate contamination analysis is the correct evaluation of which tissues constitute the contamination in a given microarray experiment, as well as the correct identification of which biomarkers are related to such tissues (Table 1). Microarray Inspector provides a definition of adipose tissue composed of selected biomarkers based on tissue-specific and tissue-enriched genes (Table 2). The adipose tissue definition is designed for Affymetrix HG-U133A, HG-U133Av2, HG-U133plus2 array types, however, the software itself allows new tissue definitions, provided the array is annotated in AnnotationDbi (Pages *et al.*, 2008). The Microarray Inspector tool analyses a set of contaminant biomarkers against a reference set. Technically, a biomarker is either a named gene that will be mapped to a list of probe sets, or an individual named probe set. The contamination set is formed via a collection of probe sets mapped from selected biomarkers. It is desirable for such biomarkers to have a high level of expression in the defined type of tissue (Chunlei *et al.*, 2009; She *et al.*, 2009; Xiao *et al.*, 2010). The choice of a proposed set of adipose biomarkers was made based on differential gene expression analysis of a few hundred preselected arrays (183 of contaminated assays to 217 not contaminated).

Algorithm. The Microarray Inspector algorithm (Fig. 1) currently uses only raw expression data from Affymetrix CEL files. Each file is being normalized using MAS5 algorithm implemented in R (RC Team 2012) Bioconductor (Gentleman *et al.,* 2004) package affy (Gautier *et al.*, 2004). MAS5 algorithm has been selected among others for several reasons. First and foremost, it uses the Wilcoxon test and is therefore adjusted for it, which is convenient for further calculations. Moreover, MAS5 normalizes each CEL file separately, whereas other algorithms, like RMA or GCRMA, use information from all the CEL files loaded, thus giving dataset-dependent results. Additionally, normalizing and analysing one file at a time is also much less computationally expensive.

After normalization, the base-2 log of the normalized MAS5 expressions of the sample are calculated and initially put on a scale of 500 (Bioconductor defaults — Gentleman *et al.,* 2004). Expression values are mapped to probe sets from the two analysed tissue sets (test and reference), yielding two lists of real numbers and allowing a statistical analysis to be performed. Our goal is to determine if there is a reason to believe in significant ex-

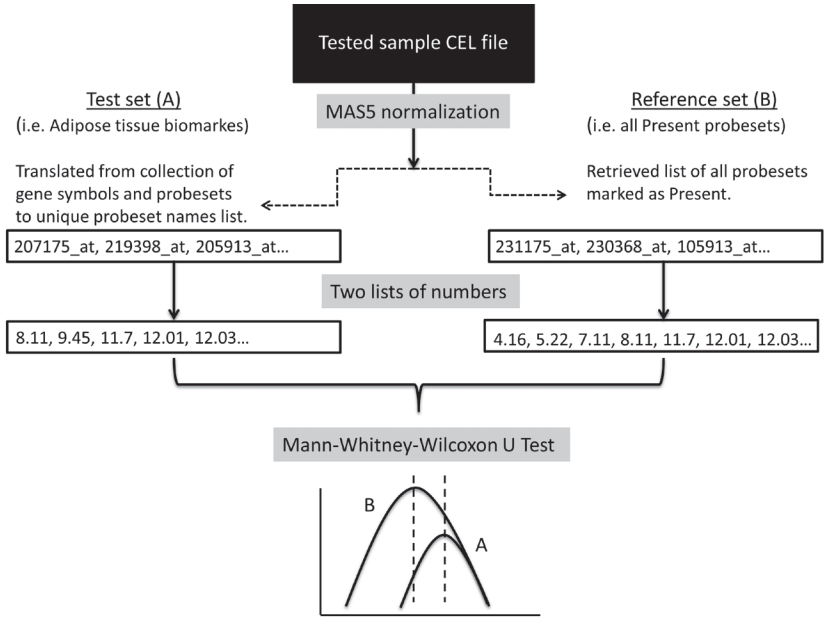

Figure 1. The Microarray Inspector algorithm.
The basic component of the method is the Mann-Whitney-Wilcoxon U test which compares two sets of numbers (normalized expression values of biomarker A and reference B sets), yielding a p-value reflecting the probability that the location (the dotted line) of A is not shifted towards higher values from B. If a resulting p-value is lower than the present significance level (by default 0.05), then the test shows lack of A in B.

Table 1. Possible tissue sample contamination in oncological experiments.

Cancer type	Possible contamination with
Breast cancer	Adipose, muscle, fibroblasts, vasculatory or inflammation tissues
Colorectal cancer	Muscle, fibroblasts, vasculatory or inflammation tissues
Ovulary cancer	Adipose, fibroblasts, vasculatory or inflammation tissues
Eye cancer	Fibroblasts, vasculatory or inflammation tissues
Brain cancer	Fibroblasts, vasculatory or inflammation tissues

pression of contaminating tissue biomarkers in the sample. The decision on contamination detection is based on the analysis of the biomarker probe sets' expression in the context of the expression of microarray probe sets related to the reference set. By default, the reference set consists of all probe sets that, after MAS5 (Affymetrix) normalization, obtained "Present" status in the Wilcoxon test (Wilcoxon, 1945). However, references can be also formed from "Marginal" or "Absent" probe sets, or even from another set of biomarkers (detailed information in Additional options section).

Next, the Mann-Whitney-Wilcoxon U Test (Mann & Withney, 1947) is used to determine if the contaminating probe sets are, as a whole, less expressed than the reference set. There are two reasons to use the Mann-Whitney-Wilcoxon U Test. First, it is a non-parametric test that can compare datasets of different sizes. The second and more important reason, is that the test assesses whether or not one set of numbers has larger values than the

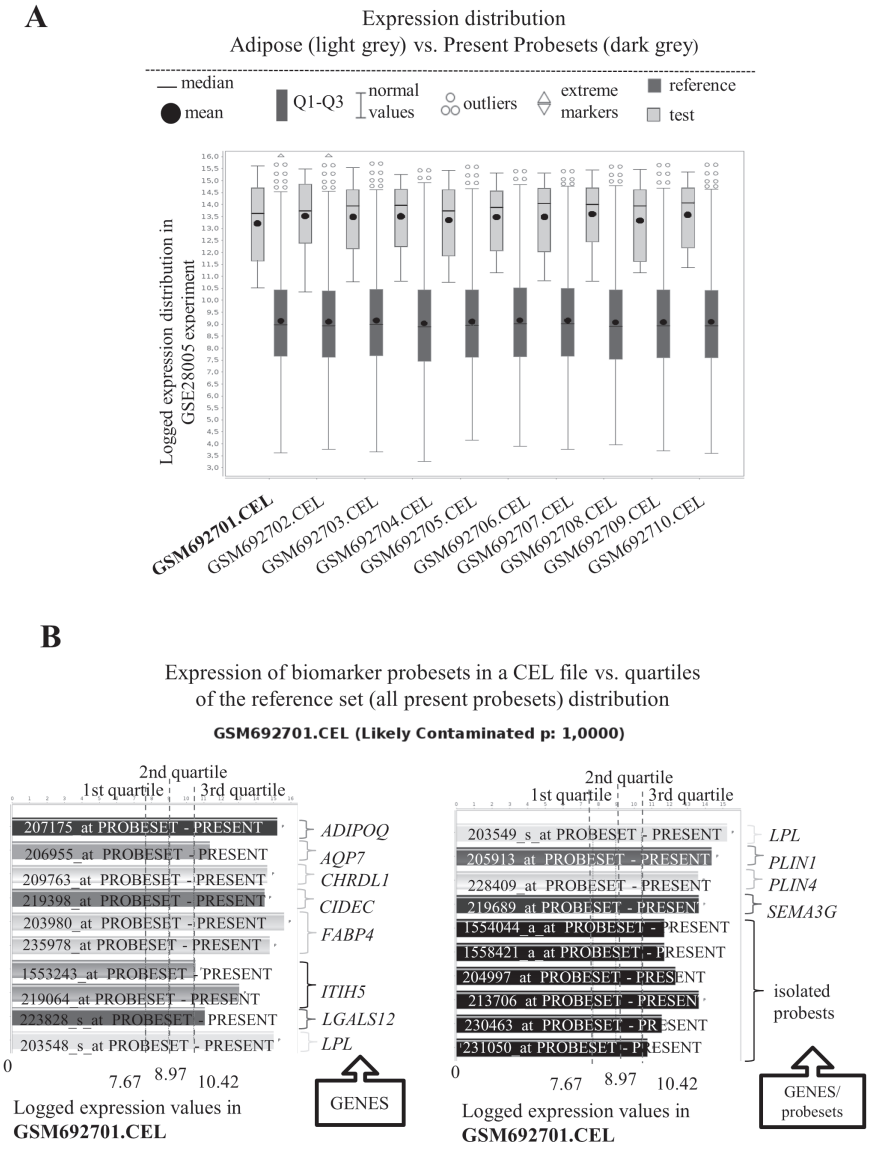

Figure 2. Visual comparison of biomarker and reference expression results of a fully contaminated experiment (GSE28005) (default settings).
(**A**) The whisker plots of the experiment's ten example assays show the distribution of the probe sets' expression values of tested tissue (light grey) and the distribution of the reference set (dark grey). (**B**) Probe sets' expression charts of the first contaminated assay GSM692701.CEL sample (p>0.05). The dotted lines represent the reference quartile's expression values.

Table 2. Details pertaining to biomarkers of a contaminating tissue.

Biomarkers (gene symbols)	Probe sets platform HG-U133plus2	Probe sets platforms: HG-U133A, HG-U133Av2
ADIPOQ	207175_at	207175_at
AQP7	206955_at	206955_at
CHRDL1	209763_at	209763_at
CIDEC	219398_at	219398_at
FABP4	203980_at, 235978_at	203980_at, 235978_at
ITH5	1553243_at, 219064_at	1553243_at, 219064_at
LGALS12	223828_s_at	N/A
LPL	203548_s_at, 203549_s_at	203548_s_at, 203549_s_at
PLIN1	205913_at	205913_at
PLIN4	228409_at	N/A
SEMA3G	219689_at	219689_at
	1554044_a_at	N/A
	1558421_a_at	N/A
	204997_at	204997_at
	213706_at	213706_at
	230463_at	N/A
	231050_at	N/A

other, what is required when trying to compare the expression of a possible contaminant against a reference set.

Our null hypothesis is that the location (a pseudo-median, Hollander & Wolfe, 1973) of the expression values of the contamination set is greater or equal to the location of the expression values of the reference set. The alternative hypothesis says, that the location of the expression values of the contaminant is smaller than the location of the expression values of the reference set. The test yields no information pertaining to the magnitude of the difference when the null hypothesis is rejected. If, with a given significance level, the null hypothesis is not rejected for a given sample (i.e. we do not accept the alternative hypothesis), then the sample will be marked as contaminated with the given set of biomarkers (Fig. 2). However, if with a given significance level the null hypothesis is rejected, then the sample will not be marked as contaminated with the given set of biomarkers (Fig. 3); any determination of a possible contamination is left for further investigation by the involved scientists.

The statistical test relies on a simplified assumption that the probe sets' expression values are independent, and furthermore, that the distributions of the two sets (test and reference) are of the same type, but shifted from each other.

The main calculation parameter in this test is the significance level α, which has a default value of 0.05. Thus α is the threshold that Microarray Inspector will compare against the *p*-value returned by the Mann-Whitney-Wilcoxon U Test. This test gives a *p*-value assessment of whether the values in the contaminant list are at least as large as those in the reference set. A sample is marked as contaminated, when the expression values from the contaminant set are at least as large as those from the reference set. This happens when the yielded *p*-value is greater than the significance level α. It can then be said with $(1-\alpha)*100\%$ confidence, the unmarked sample is not contaminated.

Additional options. To enable the experienced user to flexibly apply their own expert knowledge, Microarray Inspector allows for several parameter adjustments: changing MAS5 parameters (τ, $\alpha 1$, $\alpha 2$), setting the Mann-Whitney-Wilcoxon U Test significance level (α), selecting reference sets, trimming reference set probe sets by upper and/or lower percentage. For instance, tuning α can easily add either flexibility or rigidity to the analysis. Higher α will cause less samples to be marked as contaminated, but the confidence of cleanliness estimation will drop. Smaller α yields more results marked as contaminated, but samples are estimated not to be contaminated with a higher degree of confidence. The reference set could also be trimmed, which affects the location of reference. Such tuning enables focusing the contamination test in either higher or lower expressed genes.

The user can also choose between reference sets. Using "Present" probe sets imposes the strictest standards when potentially marking a sample as contaminated. The location (or the aforementioned pseudo-median) of the expression of "Present" probe sets will be higher than that of "Marginal", which itself will be higher than that of "Absent". Hence the contaminant will have to be "more expressed" when tested against the "Present" probe sets in order to be marked as contaminated. Likewise, using the "Absent" probe sets as reference will mark much more of the samples as possibly contaminated. It may be desired, if the researcher prefers, to consider even the lowest possibility of contamination for the sake of caution. Similar effects may be achieved by using a reference tissue composed of user defined biomarkers. In this case, the tested contamination biomarkers distribution location must be at least as high as the location of the reference biomarker set.

The trimming of reference set values from the top and bottom is designed to modify the location of the reference. Trimming the top to a greater extent than

Table 3. Summary of tested experiments.
The first column provides experiment numbering for convenient referencing in the text. The second column indicates the Affymetrix platform, the third provides GEO series code of the experiment, the fourth column presents the number of assays, followed by the material used in transcription profiling in the fifth column. The last two columns show the percentage of expected contamination (number of contaminated assays to all assays from experiment) and the result of analysis using Microarray Inspector.

No.	Platform	Experiment	Assays		Material	Contamination [%]	
						Expected	Results
1.	HG-U133_Plus_2	GSE28005	38	318	adipose tissue	100	100
2.	HG-U133_Plus_2	GSE26637	20		adipose tissue	100	100
3.	HG-U133A	GSE5090	17		adipose tissue	100	100
4.	HG-U133_Plus_2	GSE28603	12		adipocytes	100	100
5.	HG-U133_Plus_2	GSE27657	18		adipose tissue	100	100
6.	HG-U133A	GSE35710	48		adipose tissue	100	100
7.	HG-U133_Plus_2	GSE35411	26		adipose tissue	100	100
8.	HG-U133_Plus_2	GSE29410	6		adipose tissue	100	100
9.	HG-U133_Plus_2	GSE24422	24		adipocytes	100	100
10.	HG-U133_Plus_2	GSE20950	39		adipose tissue	100	100
11.	HG-U133_Plus_2	GSE41168	70		adipose tissue	100	100
12.	HG-U133_Plus_2	GSE29721	20	1298	hepatic cellular carcinoma or normal liver (microdissection)	0	0
13.	HG-U133A_2	GSE10797	66		breast epithelium and stroma (microdissection)	0	0
14.	HG-U133_Plus_2	GSE25155	28		kidney or gastric cell lines	0	0
15.	HG-U133_Plus_2	GSE11919	9		skin fibroblasts	0	0
16.	HG-U133_Plus_2	GSE11917	105		coronary artery smooth muscle cells	0	0
17.	HG-U133_Plus_2	GSE21654	22		pancreatic cancer cell lines	0	0
18.	HG-U133_Plus_2	GSE40968	18		breast cancer cell lines	0	0
19.	HG-U133_Plus_2	GSE16249	8		melanoma cell lines	0	0
20.	HG-U133_Plus_2	E-MTAB-37	950		various cancer cell lines	0	0
21.	HG-U133_Plus_2	E-MTAB-274	40		blood	0	0
22.	HG-U133_Plus_2	GSE15932	32		blood	0	0
23.	HG-U133_Plus_2	GSE41168	70	340	skeletal muscles	0–100	5.7
24.	HG-U133_Plus_2	GSE7117	8		liver	0–100	0
25.	HG-U133_Plus_2	GSE30718	47		kidney	0–100	12.8
26.	HG-U133_Plus_2	GSE7821	40		intestine	0–100	0
27.	HG-U133_Plus_2	GSE18864	84		breast tumour	0–100	3.6
28.	HG-U133A	GSE42822	91		breast cancer	0–100	7.7

the bottom, should result in a reference set's location with lower expression values and possibly more samples will be marked as contaminated. Conversely, trimming the bottom significantly more than the top ought to result in a reference set's location with higher expression values and possibly fewer samples will be marked as contaminated. To be more descriptive, trimming the top 50% of values will result in using the location as a pseudo first quartile instead of a pseudo-median. Likewise, trimming the bottom 50% of values will result in using the location as a pseudo third quartile instead of a pseudo-median. Lastly, choosing to trim the top and bottom values with an equal amount should not significantly affect the location of the reference set. Such flexibility in trimming is desired for some experiments. In some cases, even relatively low expression of contaminant biomarkers can represent considerable contamination, while in other cases quite the opposite.

An easy interface for self-definition of test and reference tissues, using any genes and probe sets, is also provided.

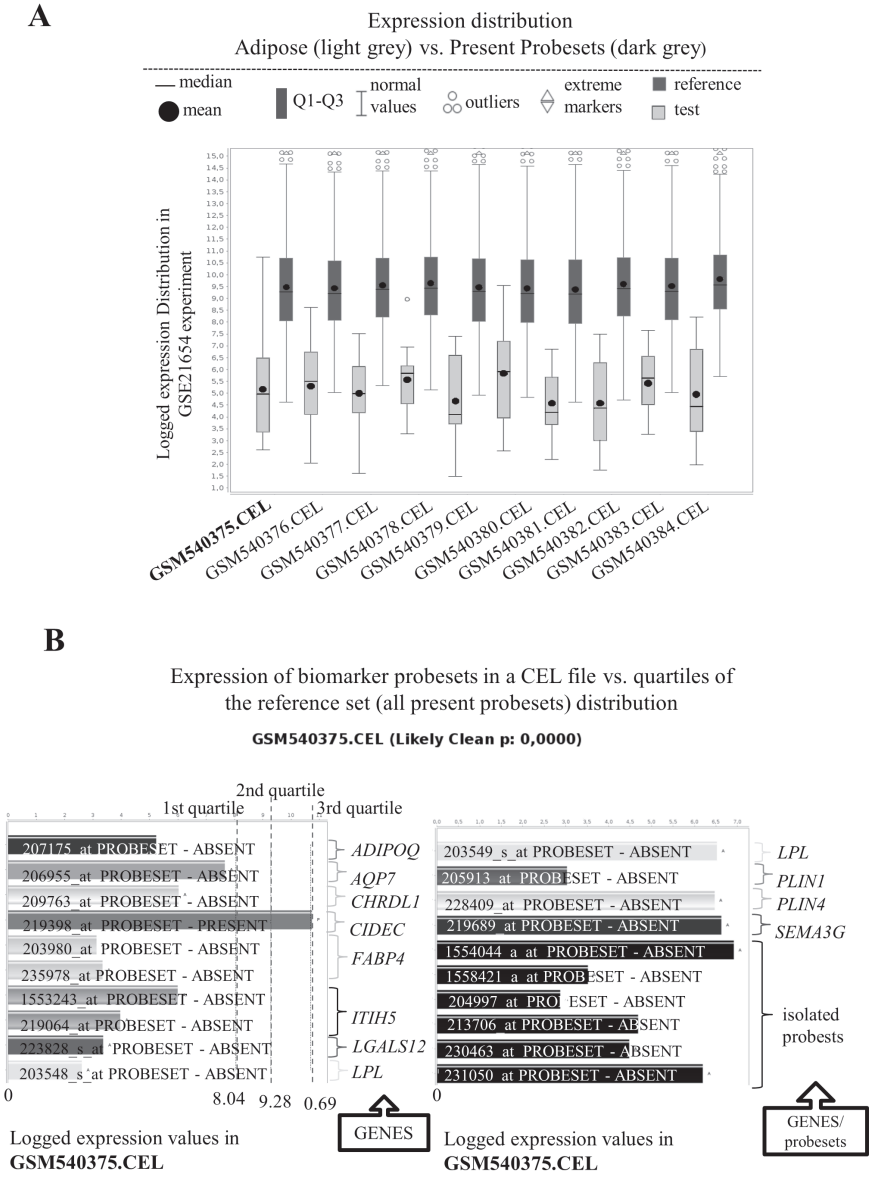

Figure 3. Visual comparison of biomarker and reference expression results of a clean experiment (GSE21654) (default settings). (**A**) The whisker plots of the experiment's ten example assays show the distribution of the probe sets' expression values of tested tissue (light grey) and the distribution of the reference set (dark grey). (**B**) Probe sets' expression charts of the example's not contaminated assay GSM540375.CEL sample ($p<0.05$). The dotted lines represent the reference quartile's expression values.

Datasets. Firstly, to analyse the specificity of our method, we tested 11 experiment datasets (total of 318 assays) for the presence of adipose tissue/cells. These experiments include transcription profiling analysis of adipose tissue or adipocytes. We expected to produce close to 100% positive "contamination" reports. The other 11 datasets (total of 1298 assays) were selected to reflect the expected 0% "contamination". They were both different cancer or normal cell lines (apart from adipocytes), as well as microdissected material derived from various organs or blood. To check the purity of the samples from public repositories, we selected six experiment datasets (total of 340 assays) expected to be contaminated. These were surgical biopsies of, for example, skeletal muscles, liver, kidney, intestine, or breast.

All above datasets were downloaded from GEO database (Edgar *et al.*, 2002; Barrett *et al.*, 2011) and analysed using Microarray Inspector.

RESULTS AND DISCUSSION

Specificity

Our results proved that all analysed adipose tissue/cell arrays were reported as "contaminated" when using the following default sensitivity setup: MAS5 parameters ($\tau = 0.015$, $\alpha 1 = 0.04$, $\alpha 2 = 0.06$), significance level 0.05 and no trimming of reference sets. It is important to note that with the same parameters, none of the theoretically clean samples returned as contaminated after the tests. Summary of the results is presented in Table 3. Occasional high values of biomarkers in "unspecific" tissue samples are to be expected. Most experiments are designed to change the natural balance of sample gene expression patterns. In particular, cancer cells deviate from their standard counterparts. However, our bio-

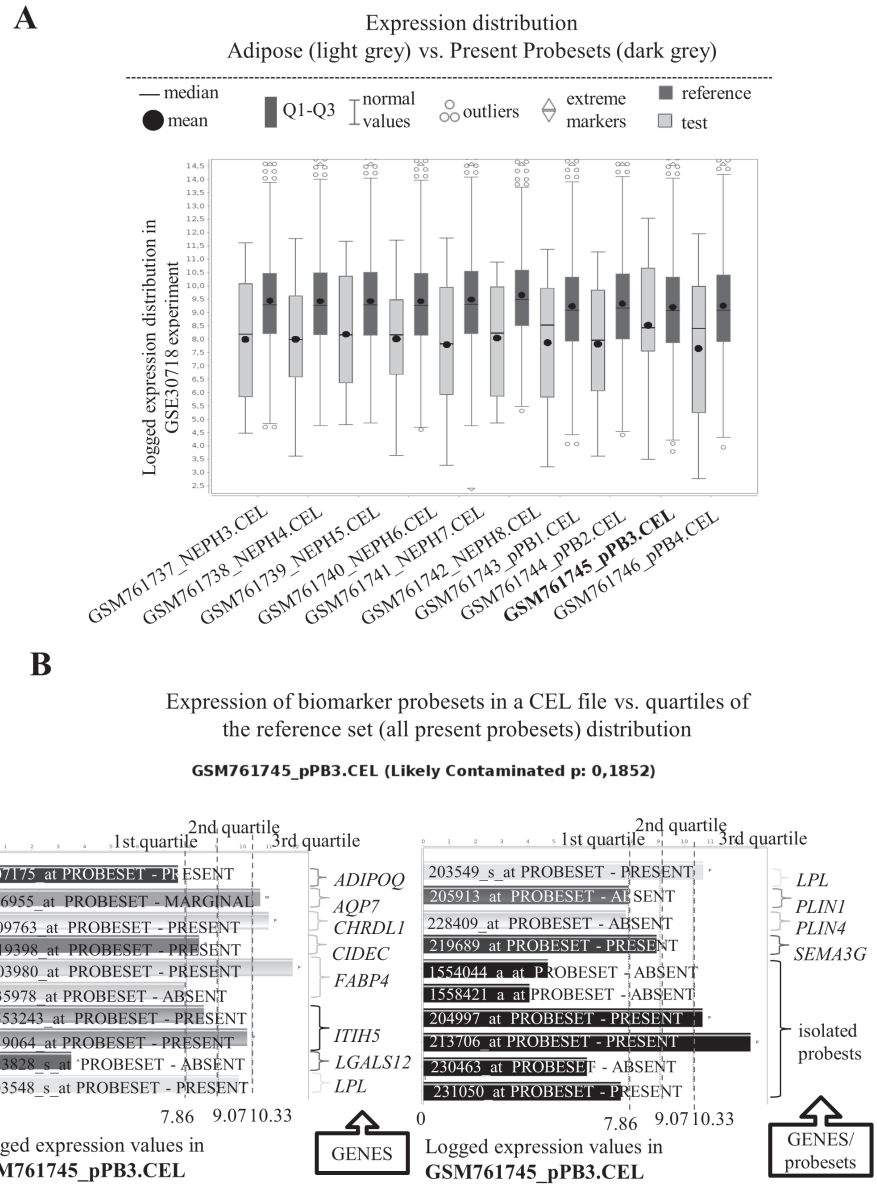

Figure 4. Visual comparison of biomarker and reference expression results of the tested experiment GSM30718 (default settings). (**A**) The whisker plots of the experiment's ten example assays show the distribution of the probe sets' expression values of tested tissue (light grey) and the distribution of the reference set (dark grey). (**B**) Probe sets' expression charts of the example's contaminated assay GSM761745_pPB3.CEL sample ($p>0.05$). The dotted lines represent the reference quartile's expression values.

marker set has been selected to minimize the possibility of false positive results under "unspecific" conditions, which could increase the expression of biomarkers.

Purity of the samples and the trimming effect

We examined six experiments that could possibly contain sample tissue cross-contamination (Table 3, no. 23–28). Four of the six datasets showed adipose tissue contamination ranging from approximately 4% to 13%. Taking into consideration that usual data analysis methods filter out low expression probe sets, we trimmed the bottom 20% of present reference probe sets, thus lowering the sensitivity of the contamination detection. For instance, in experiment GSE30718 we identified a contamination level of 12.8%. A total 6 assays out of 47 showed adipose contamination with default sensitivity settings. After trimming the reference by the bottom 20% of values, four of the previously contaminated arrays were marked as not contaminated. Figures 4 and 5 present a clear example of a trimming effect. The length of low whiskers of references (in red) is visibly shorter in the trimmed experiment chart (Fig. 5) than in the same experiment but with default setting (Fig. 4). After trimming of the reference set, some biomarker probe sets fell to lower expression quartiles, thus changing their labelling to contamination-free. This feature of the test can be applied for instance when the further analyses on the tested microarrays will focus only on highly expressed genes, making it reasonable to ignore probe sets with low expression values.

Microarray Inspector and other similar methods/tools

We propose simple, user friendly software that enables screening analysis of adipose contamination in hundreds of arrays simultaneously. Our software predicts *in silico* the purity of the sample, similarly to the method

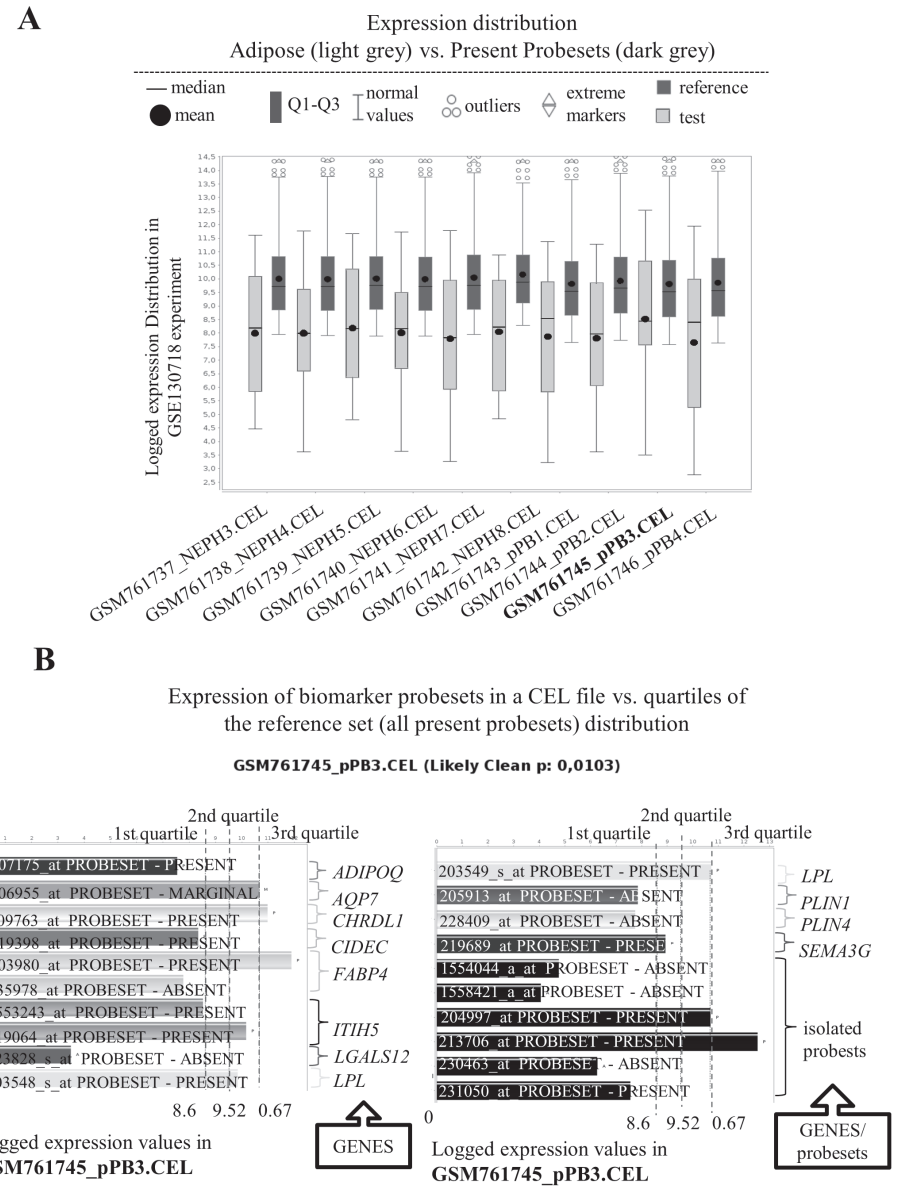

Figure 5. Trimming effect of bottom 20% values in the tested experiment GSE30718.
(**A**) The length of low whiskers of references (dark grey) is shorter in the trimmed chart than in the same experiment when using defaults settings (Fig. 4A). (**B**) Detailed probe sets' expression of file GSM761745_pPB3.CEL. The shift of reference expression values quartiles (dotted lines) results in marking the sample as not contaminated ($p<0.05$).

presented by Wang *et al*., 2010. Contrary to that method, however, our tool does not require histological evaluation to build a prediction model that is necessary in the cited article. Instead, we use predetermined experiments with 0% (cell lines, apart from adipocytes experiments) and 100% of contamination (adipose tissue profiling assays). The aim is to answer the following question: is the experiment contaminated or not? Several approaches have already focused on determining the proportion of admixture (Venet *et al*., 2001; Lu *et al*., 2003; Lähdesmäki *et al*., 2005; Wang *et al*., 2006; Clarke *et al*., 2010). However, regardless of the proportion value of admixture in the sample, the results and data interpretation may be inconsistent. Furthermore, the methods that are based on histological evaluation of admixture percentage do not take into account the amount of mRNA. Particular attention should be brought to the fact that cancer cells generate much more RNA than normal cells, which could cause the result discrepancies. Several available methods (Venet *et al*., 2001; Lähdesmäki *et al*., 2005; Wang *et al*., 2006) require expression profiles of purified reference tissue i.e. microdissected material, to calculate the proportion of contamination. Unfortunately, reference hybridizations for many tissue types are unavailable. In addition, laser capture microdissection (LCM), which is used to purify the samples, remains a real challenge when attempting to extract high quality mRNA. This drawback is caused by mRNA's inherent instability. We try to avoid all of the above problems and design a method useful for screening analysis of a lot of data at the same time, with high specificity and sensitivity controlled by the user. Another advantage of our method is that it enables the user to build necessary tissue definitions based on manually preselected existing data from public repositories. Although confirmatory experiments are always preferred, they are time-consuming and costly. In our case this is not neces-

sary, provided the researcher can access high quality data which underwent a thorough selection process in order to build a contamination biomarker set. Besides using Microarray Inspector, the user can check the method's reliability on large portion of data in a relatively short time.

CONCLUSIONS

As presented in the results section, our method provides a unique insight into microarray experiments. Microarray Inspector helps researchers decide whether or not the results are reliable, should the contaminated samples be discarded, or if the analysis procedures have to be modified to provide more strict filtration and thresholds. For example, in our model case of adipose contamination, considering potentially discovered contamination after the analysis, the findings may be reassessed, and false results relating to contamination could be identified and subsequently discarded.

The default parameter settings should cover most experimental conditions, although we emphasise that the user has the means to control more than just the main test parameter α to tune the test as described above. Letting the user apply his expert knowledge to the analysis, is the chief goal. We hope the method implemented in our software tool will fill a significant gap in post experimental data analysis and will enable researchers to easily validate sample compositions. Currently, the software is limited to examine Affymetrix CEL files, however we expect to extend it to other platforms in the near future.

Conflict of interests

The authors declare that they have no conflict of interests.

Acknowledgements

This work was supported by the Polish Agency for Enterprise Development [UDA-POIG.01.04.00-14-001/10-00]. AS was supported by the MPD [MPD/2009/5/styp11].

REFERENCES

Affymetrix I: Statistical algorithms description document. *Technical paper* 2002.

Affymetrix Statistical Algorithms Reference Guide.

Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall Ka, Phillippy KH, Sherman PM, Muertter RN, Holko M, Ayanbule O, Yefanov A, Soboleva A (2011) NCBI GEO: archive for functional genomics data sets — 10 years on. *Nucleic Acids Res* 39: D1005–D1010.

Bolstad BM, Collin F, Simpson KM, Irizarry RA, Speed TP (2004) Experimental design and low-level analysis of microarray data. *Int Rev Neurobiol* **60**: 25–58.

Chen JJ, Hsueh H-M, Delongchamp RR, Lin C-J, Tsai C-A (2007) Reproducibility of microarray data: a further analysis of microarray quality control (MAQC) data. *BMC Bioinformatics* **8**: 412.

Chunlei Wu CO et al., Wu C, Orozco C, Boyer J, Leglise M, Goodale J, Batalov S, Hodge CL, Haase J, Janes J, Huss JW, Su AI (2009) BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol* **10**: R130.

Clarke J, Seo P, Clarke B (2010) Statistical expression deconvolution from mixed tissue samples. *Bioinformatics (Oxford, England)* **26**: 1043–1049.

Dupuy A, Simon RM (2007) Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *J Nat Cancer Institute* **99**: 147–157.

Edgar R, Domrachev M, Lash AE (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* **30**: 207–210.

Gautier L, Cope L, Bolstad BM, Irizarry Ra (2004) affy — analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics (Oxford, England)* **20**: 307–315.

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JYH, Zhang J (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5**: R80.

Hollander M, Wolfe DA (1973) *Nonparametric Statistical Methods*. 2nd (1999) edition. New York: John Wiley & Sons.

Ioannidis JPA (2005) Microarrays and molecular research: noise discovery? *Lancet* **365**: 454–455.

Irizarry R a (2003) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res* **31**: 15e–15.

Ji H, Davis RW (2006) Data quality in genomics and microarrays. *Nature Biotechnol* **24**: 1112–1113.

Lähdesmäki H, Shmulevich L, Dunmire V, Yli-Harja O, Zhang W (2005) *In silico* microdissection of microarray data from heterogeneous cell populations. *BMC Bioinformatics* **6**: 54.

Lu P, Nakorchevskiy A, Marcotte EM (2003) Expression deconvolution: a reinterpretation of DNA microarray data reveals dynamic changes in cell populations. *Proc Natl Acad Sci USA* **100**: 10370–1035.

Mann HB, Whitney DR (1947) On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics* **18**: 50–60.

McCall MN, Murakami PN, Lukk M, Huber W, Irizarry R a (2011) Assessing affymetrix GeneChip microarray quality. *BMC Bioinformatics* **12**: 137.

Michiels S, Koscielny S, Hill C (2005) Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet* **365**: 488–492.

Pages H, Carlson M, Falco S, Li N (2008) AnnotationDbi: Annotation Database Interface. *R package version 1.16.18*.

She X, Rohl C a, Castle JC, Kulkarni A V, Johnson JM, Chen R (2009) Definition, conservation and epigenetics of housekeeping and tissue-enriched genes. *BMC Genomics* **10**: 269.

Shi L, Tong W, Goodsaid F, Frueh FW, Fang H, Han T, Fuscoe JC, Casciano DA (2004) QA/QC: challenges and pitfalls facing the microarray community and regulatory agencies. *Expert Rev Mol Diagnostics* **4**: 761–777.

Shi L, Reid LH, Jones WD *et al.* (2006) The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnol* **24**: 1151–1161.

Shi L, Campbell G, Jones W, Campagne F (2010) The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nature Biotechnol* **28**: 827–838.

Skrzypczak M, Goryca K, Rubel T, Paziewska A, Mikula M, Jarosz D, Pachlewski J, Oledzki J, Ostrowski J, Ostrowsk J (2010) Modeling oncogenic signaling in colon tumors by multidirectional analyses of microarray data directed for maximization of analytical reliability. *PloS One* 5.

Tan PK (2003) Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res* **31**: 5676–5684.

Team RC (2012) *A Language and Environment for Statistical Computing*.

Wang M, Master SR, Chodosh L a (2006) Computational expression deconvolution in a complex mammalian organ. *BMC Bioinformatics*, **7**: 328.

Wang Y, Xia X-Q, Jia Z, Sawyers A, Yao H, Wang-Rodriquez J, Mercola D, McClelland M (2010) *In silico* estimates of tissue components in surgical samples based on expression profiling data. *Cancer Res* **70**: 6448–6455.

Wilcoxon F (1945) Individual Comparisons by Ranking Methods. *Individual Comparisons by Ranking Methods* **1**: 80–83.

Venet D, Pecasse F, Maenhaut C, Bersini H (2001) Separation of samples into their constituents using gene expression data. *Bioinformatics (Oxford, England)* **17** (Suppl 1): S279–S287.

Xiao S-J, Zhang C, Zou Q, Ji Z-L (2010) TiSGeD: a database for tissue-specific genes. *Bioinformatics (Oxford, England)* **26**: 1273–1275.