ACTA ABP BIOCHIMICA POLONICA

Vol. 61, No 2/2014
333–340
on-line at: www.actabp.pl

Regular paper

Smooth muscle contamination analysis in clinical oncology gene expression research

Monika Markowska[1,2]✉, Piotr Stępniak[2], Konrad Wojdan[2,3] and Konrad Świrski[2,3]

[1]Department of Gastroenterology and Hepatology, Medical Center for Postgraduate Education, Warsaw, Poland; [2]Transition Technologies S.A., Warsaw, Poland; [3]Institute of Heat Engineering, Warsaw University of Technology, Warszawa, Poland

Gene expression profiling is one of the most explored methods for studying cancers and microarray data repositories have become a rich and important resource. The most common human cancers develop in organs that are walled by smooth muscles. The only method of sample extraction free of unintentional contamination with surrounding tissue is microdissection. Nevertheless, such an approach is implemented infrequently. In the light of the above, there is a possibility of smooth muscle contamination in a large portion of publicly available data. In this study, 2292 publicly available microarrays were analysed to develop a simple screening method for detecting smooth muscle contamination. Microarray Inspector software was used to perform the tests since it has the unique ability to use many selected genes and probesets in a single group as a tissue definition. Furthermore, the test was dataset-independent. Two strategies of tissue definition were explored and compared. The first one depended on Tissue Specific Genes Database (TiSGeD) and BioGPS web resources, which themselves were based on meta-analysis of thousands of microarrays. The second method was based on a differential gene expression analysis of a few hundred preselected arrays. The comparison of the two methods proved the latter to be superior. Among the tested samples of undefined contamination, nearly half were identified to possibly contain significant smooth muscle traces. The obtained results equip researches with a simple method of examining microarray data for smooth muscle contamination. The presented work serves as an example of how to create definitions when searching for other possible contaminations.

Key words: microarray, transcription profiling, smooth muscle contamination, tissue specific genes, cancer, data quality

Received: 17 January, 2014; revised: 13 May, 2014; accepted: 15 May, 2014; available on-line: 16 June, 2014

INTRODUCTION

According to the World Health Organization, cancer was the cause of death of 7.6 million people in 2008, representing nearly 13% of reported deaths in the world. Unfortunately, the mechanisms of tumour pathogenesis and processes leading to metastasis are still not well understood. In the field of oncology research, microarray technology is one of the most commonly used techniques in transcription profiling analysis. Although genomic technology has advanced, it is not yet a fully developed field. One of the clear findings so far is that there are different strategies in data processing which lead to discrepancies in sample handling, as well as raise incommensurability concerns between involved laboratories. Be it as it may, there has been a noticeable growth in the number of published practice guides designed to improve unity and reliability between different platforms. Nevertheless, surgical sample contamination by different tissues/cells remains a problem in a number of sample extraction processes, and is often ignored and rarely considered in the relevant literature. Most samples used in this kind of studies usually contain a mixture of cells or tissue types (Lähdesmäki *et al.*, 2005; Wang *et al.*, 2006). Although the laser capture microdissection (LCM) method enabled avoiding the interference provided by unintended tissues components (Paweletz *et al.*, 2001) LCM is still not widely used. Little data based on this technology is available for analysis. The different impacts of tissue components on cancer gene expression profiling have been discussed in the literature. For example, tissue proportions of cancer and stroma cells in surgical samples have been recently analysed (Roepman *et al.*, 2005; Wang *et al.*, 2010), and shown to have an important role in prediction of tumour invasion and metastasis processes. Meanwhile, non-cancerous material can significantly affect cancer expression profile results. Until now, the subject of smooth muscle tissue contamination in cancer samples was not considered in literature. Smooth muscles (SM) are found within the walls of cavernous organs, their mucous membranes, in gastrointestinal, respiratory and urogenital tracts, as well as in the walls of blood vessels. The most common cancers in the human population develop in the lungs, colon, stomach, cervix, prostate, pancreas, or the bladder. Surgical samples of these organs are all likely to contain smooth muscle tissue.

Recently, a new software tool for microarrays called Microarray Inspector has been developed (Stępniak *et al.*, 2013). It enables the analysis of raw microarray data files and detection of tissue cross contamination. Single biomarkers of tissues providing comparable absolute expression levels are hard, if not impossible, to find. Instead, Microarray Inspector uses the whole group of selected genes and probesets to compare against the reference set in a single trial. The reference set is also a group of genes and probesets of the same array. Usually, it is the whole present probe-sets group, but its scope and sensi-

✉e-mail: M.Markowska@tt.com.pl

Abbreviations: SM, smooth muscles; def.1, definition 1; TiSGeD, the Tissue-Specific Genes Database; SPM, specificity measure factor; LCM, laser capture microdissection

tivity can be adjusted. The Mann-Whitney-Wilcoxon U test is used to estimate the probability of a biomarker group being significantly expressed. The additional advantage of the software is that it tests a single array at a time, which makes it independent of the group of examined microarrays.

In this study, we present the development process of a biomarker set for smooth muscle in Microarray Inspector. The proposed tissue biomarker definition set is intended for samples originating from cancers that develop in the lungs, colon, stomach, cervix, prostate, pancreas, or bladder.

MATERIALS AND METHODS

Study design. In this study, the microarray data were collected from Affymetrix HG-U133Plus, HG-U133A, or HG-U133Av2 platforms. Exclusive usage of Affymetrix arrays can be explained by the fact that Microarray Inspector currently supports only this kind of microarray files. Unfortunately, no experiments containing known smooth muscle tissue mixtures were found on Affymetrix platforms. Five types of sample data were collected: 1) fully contaminated, 2) not contaminated, 3) negative control, 4) suspected of contamination, and 5) microdissection experiments. Expression analyses of smooth muscle cells were included in the first group, as these are the main cell type present in SM tissue (Chi *et al.*, 2007). The advantage of using cell cultures is the homogeneity of the material as well as the absence of other cell types interfering with the expression found in analysis results. For this reason, transcription profiles of both cancer and normal cell lines derived from the lungs, colon, stomach, cervix, prostate, pancreas and bladder were assigned as not contaminated experiments. Expression pattern analysis of either cancerous or normal material, theoretically not contaminated with smooth muscle tissue and isolated from other body localizations, such as brain, blood, liver or lung (endothelium), was chosen as a negative control. Microarray data of tissue material derived from lung, colon, stomach, cervix, prostate, pancreas or bladder cancers were suspected of being contaminated, and therefore checked for purity. The last group includes experiments of cancer cells originating in the pancreas, lung, cervix, and colon collected by the laser microdissection method. These experiments were checked for contamination as well.

All 2292 assays in a total of 67 different experiments obtained from ArrayExpress repository (Parkinson *et al.*, 2007; http://www.ebi.ac.uk/arrayexpress) were used in the study: 192 were fully contaminated samples, 631 assays were of not contaminated material, 291 were negative control assays, 1021 were suspected samples and 157 were assays of microdissection research. Table 1 shows a short description of experiments from group numbers 1, 2 and 3, which were crucial in this research. Reference numbers of experiments from groups 1–5 can be found in Table 5.

TiSGeD and BioGPS. Tissue-specific smooth muscle genes were selected using TiSGeD (Tissue-Specific Genes Database by Xiao *et al.*, 2010; http://bioinf.xmu.edu.cn:8080/databases/TiSGeD/index.html) with specificity measure (SPM) factor greater than 0.9. SPM ranges from 0 to 1 with a high value corresponding to strong tissue specificity. In addition, the results were compared with information from the BioGPS portal (Chunlei *et al.*, 2009; http://biogps.org/). Based on the collected data, several different smooth muscle definitions were created. Table 2 shows four examples of them. All biomarkers included in the four definitions as well as their corresponding probesets in HG-U133A, HG-U133Av2 and HG-U133Plus2 Affymetrix platforms are illustrated in Table 3. The first definition (def.1) (refer to Table 2) consists of seven genes randomly selected from the tissue specific genes. The second definition (def.2) was built using six genes which do not have cytokine annotation, and the third (def.3) was composed of five cytokine/chemokine encoding genes. The fourth and last smooth muscle definition (def.4) was created by trial and error, and then applied to the conclusions of testing SM definitions 1–3 on experiment groups 1–3 (not contaminated, fully contaminated, control). Additionally, one new gene named *LRRC17* was included into the smooth muscle definition 4.

Differential expression analysis using R/Bioconductor. An alternative method used to design tissue definition is differential expression analysis. The standard approach of comparison of two groups with a t-test was applied in the R environment (R Team 2012) using the Bioconductor (Gentleman *et al.*, 2004) package genefilter (Gentleman *et al.*, 2012). Experiments representing non-contaminated samples were: E-GEOD-13309 (2 lung cancer cell lines), E-GEOD-21654 (22 pancreas cancer cell lines), E-MTAB-37 (10 bladder cancer cell lines), E-GEOD-17482 (2 prostate cancer cell lines), E-GEOD-22183 (37 gastric cancer cell lines), E-MTAB-37 (7 cervix cancer cell lines), E-GEOD-30292 (8 colon cancer cell lines and laser dissected: tumor cells, normal colonocytes, and enterocytes of ileum and jejunum). They were compared against a fully contaminated group composed of: E-GEOD-11917 (coronary artery smooth muscle cells), E-GEOD-12261 (aortic smooth muscle cells) and E-GEOD-19672 (corporal smooth muscle cells). All the experiments are based on the HG-U133 Plus 2 Affymetrix platform containing more gene probesets than any of the older available platforms, that is HG-U133A, B or Av2. The experiments were also balanced so that the number of arrays in each group did not exceed the 1:2 ratio for contaminated vs. not contaminated, as suggested in the *t*-test procedure. The total number of arrays in the groups was 124 for contaminated (64.6% of the whole group), and 194 for not contaminated (30.7% of the whole group).

All arrays were normalized together using the GCRMA algorithm from the package gcrma (Wu *et al.*, 2012). Next, all probesets with general low intensity and variability in all samples were discarded. This was achieved by filtering out probesets that did not present a log2-based expression value higher than 6.64 (intensity higher than 100) in at least 25% of all the arrays. Probesets showing an interquartile range lower than 0.5 were also discarded. Following the filtration, the *t*-test was applied. It yielded *p*-values describing how likely the corresponding probesets are to emerge as differentially expressed by chance. *P*-values were next adjusted using the Benjamini & Hochberg method (Benjamini & Hochberg, 1995). The top 100 probesets with best (lowest) *p*-values were annotated using the Bioconductor packages annotate (Gentleman, 2012), KEGG (Carlson, 2012), GO (Carlson, 2012), annaffy (Smith, 2010) and XML (Lang, 2012). The list of results was investigated to select probesets for tissue definition. Genes with low expression in the contaminant group or genes related to cancer were omitted. The exception was the *FGF5* gene, which was included in definition 5 (def.5), despite it being previously reported to be overexpressed in human

Table 1. Description of experiments used in the study (fully contaminated, not contaminated and control), obtained from ArrayExpress repository at http://www.ebi.ac.uk/arrayexpress.

Group	Experiment	Material	Description
1	E-GEOD-11917	smooth muscle	Calcification of human coronary artery smooth muscle cells in the presence of vitamin D
1	E-MEXP-569	smooth muscle	Aortic vascular smooth muscle cell response to cyclical mechanical strain
1	E-GEOD-21363	smooth muscle	Expression profile of HITC6 smooth muscle cells during their cord morphogenesis
1	E-GEOD-12261	smooth muscle	Analysis of 2-methoxyestradiol effect on smooth muscle cell hyperproliferation and vascular remodeling in atherosclerosis
1	E-GEOD-19672	smooth muscle	Transcription profiling of human corporal smooth muscle cells after MaxiK potassium channel silencing
1	E-GEOD-13168	smooth muscle	Human airway smooth muscle cells after treatment with glucocorticoids and Protein Kinase A — transcription profiling analysis
2	E-GEOD-11839	bladder	Gene expression analysis of human female urothelial cell cultures, differentiated vs. non-differentiated and interstitial cystitis *vs*. control
2	E-MTAB-37	bladder	Transcriptomic profiles of various cancer cell lines (only bladder assays chosen)
2	E-GEOD-26828	bladder	Transcription expression analysis of six cadmium-transformed UROtsa cell isolates
2	E-GEOD-21654	pancreas	Gene expression patterns from untreated 22 commercially available pancreatic cancer cell lines
2	E-GEOD-22973	pancreas	Differential gene expression analysis of primary and metastatic pancreatic cancer cell lines after induction by PMA (a known inducer of invasion)
2	E-GEOD-37645	pancreas	MRK-003 inhibitor attenuating effect on pancreatic ductal adenocarcinoma cell growth
2	E-GEOD-22337	pancreas	UDP-GlcNAc 2-epimerase/ManNAc kinase (GNE) inducing effect on Capan-1 pancreatic carcinoma cells apoptosis
2	E-GEOD-22183	gastric	Global expression profiles of 37 unique gastric cancer cell lines
2	E-GEOD-32540	gastric	E-cadherin/CDH1 intron2-initiated transcript influence on gastric cancer cell invasion and angiogenesis
2	E-GEOD-20058	gastric	Transcription profiling of the side population of gastric cancer cell lines
2	E-GEOD-35830	cervix	Gene expression analysis of human Ect1 ectocervical epithelial cells treated with seminal plasma or transforming growth factor-β
2	E-MTAB-37	cervix	Transcriptomic profiles of various cancer cell lines (only cervix assays chosen)
2	E-MEXP-1078	cervix	Gene expression analysis of HeLa cells treated with the major listerial toxin listeriolysin
2	E-GEOD-13309	lung	Human lung cancer cells after treatment with tobacco smoke condensate — microarray expression analysis
2	E-GEOD-4824	lung	Global expression profiles of human lung cancer cell lines
2	E-GEOD-30660	lung	The effect of repeated whole cigarette smoke challenge on three-dimensional human lung airway epithelial cultures
2	E-GEOD-18454	lung	Gene expression analysis of normal human bronchial epithelial and human small airway epithelial cells treated with 5-aza-dC
2	E-GEOD-36176	lung	Lung cancer cells exposed to Notch inhibitor — global gene expression analysis
2	E-GEOD-29008	colon	The effect of Clostridium difficile Toxins A and B on human colon epithelial cells (HCT-8) — transcriptome analysis
2	E-GEOD-13367	colon	Global gene expression profiles of mucosal colonic biopsies and isolated colonocytes (only colonocytes choosen)
2	E-GEOD-15132	colon	Analysis of riboflavin deficiency in a human intestinal epithelial cells
2	E-MEXP-2010	colon	Expression data from human colon cancer cells treated with docosahexaenoic acid
2	E-GEOD-15799	colon	Gene expression profiling of NS398-treated HT29 colon adenocarcinoma cell line
2	E-GEOD-10650	colon	Human colon carcinoma cell line HCT116 and PCLKC — transcriptome analysis
2	E-GEOD-35973	colon	Expression data from colon cancer cells with mutant K-ras under hypoxic conditions
2	E-GEOD-30292	colon	Comparative analysis of differences between various intestinal colon carcinoma cell lines and normal intestinal epithelium — selecting relevant intestinal tumor model
2	E-GEOD-30304	prostate	Prostate epithelial cells response to tissue contextual changes
2	E-TABM-948	prostate	Normoxia- and hypoxia-treated prostate tumor cell lines and primary prostate epithelial cells — global gene expression analysis

2	E-GEOD-17482	prostate	Comparative analysis of gene expression profiles regulated by Stat5a/b vs. Stat3 in human prostate cancer cell lines
3	E-GEOD-25014	lung	Transcription profiling of human pulmonary artery endothelial cells and microvascular endothelial cells induced by heme
3	E-GEOD-18269	liver	Comparative analysis of expression profiles of HepaRG, HepG2 to primary human hepatocytes and liver
3	E-GEOD-25155	kidney, gastric	Gene expression profile of human kidney and gastrointestinal epithelial cells treated with Helicobacter pylori lipopolysaccharide
3	E-MTAB-274	blood	Global gene expression analysis of blood from participants with subjective memory decline
3	E-GEOD-15932	blood	Differential gene expression analysis in peripheral blood specific to pancreatic cancer associated diabetes
3	E-GEOD-16249	skin	Knockdown of MITF and PAX3 in metastatic melanoma cell lines mediated by siRNA
3	E-GEOD-21354	brain	Transcription profiling analysis of three types of grade II gliomas
3	E-GEOD-40968	breast	ACSL4 gene expression effects in breast cancer cell lines
3	E-MEXP-2957	blood	Comparative microarray profiling of male and female patients with Chronic Lymphocytic Leukemia
3	E-GEOD-18864	breast	Determining efficacy of neoadjuvant Cisplatin in sporadic triple negative breast cancers

pancreatic cancer (Kornmann *et al*., 1997). The examined data confirmed that *FGF5* expression in a healthy pancreas is very low, unlike that of a cancerous pancreas. However, the risk of falsely marking cancer as smooth muscle is minimized by the other probesets in the tissue definition. Furthermore, definition 5 was more prone to false positive results in experiments from the control group due to the lack of the *FGF5* probeset. The probesets selected by this method corresponding to smooth muscle biomarkers in Microarray Inspector are presented in Table 2.

Evaluation of created smooth muscle definitions. Using the TiSGeD database and BioGPS portal information, four different definitions were built: SM definition 1 to 4, corresponding to fifteen genes. Based on differential expression analysis, eighteen probesets corresponding to eleven genes were selected and included in SM definition 5 (Table 2). Five definitions were checked for quality and smooth muscle tissue specificity by verifying the purity of fully contaminated, not contaminated, and control experiments (groups 1–3 from Study design). All results including contamination analysis of suspected and microdissection experiments are presented in Table 4. More detailed information is available in Table 5.

Table 2. SmoothMuscle definitions and number of probesets available on Affymetrix platforms.

Based on TiSGeD and BioGPS				Based on Differential Gene Expression Analysis
Def.1	Def.2	Def.3	Def.4	Def.5
CCL7	PXDN	FGF2	GFPT2	BGN
CSF3	MMP1	CCL8	THBS2	IFFO1
CXCL1	PLOD2	IL6	CXCL6	FGF5
CXCL3	THBS2	CSF3	CXCL3	*COL1A1* (202311_s_at)
CXCL6	GFTP2	CXCL3	IL6	*ITGA4* (205884_at, 205885_s_at)
GFTP2	STC1		LRRC17	*PRRX1* (205991_s_at)
STC1			FGF2	*ARHGAP22* (206298_at)
				DCN (211813_x_at, 211896_s_at)
				PAMR1 (213661_at)
				ELTD1 (219134_at)
				C1orf54 (219506_at)
Total number of probesets HG-U133A, HG-U133Av2, HG-U133plus2				
9	10	6	8 or 9 (HG-U133plus2)	18

RESULTS

Microarray Inspector analysis of six fully contaminated experiments indicates that only three definitions, SM definition 2, 4 and 5, demonstrated an average contamination in the 95–100% range, while SM definition 1 and 3 returned only 79.7% and 87%, respectively (Table 4). The lowest average level of contamination (0.05%) was achieved by subjecting twenty nine not contaminated experiments with definition 5. The next lowest values were 8% and 7% for definition 3 and definition 4, respectively. Definition 2 provided the highest value of 37.3%. All definitions apart from SM definition 2 demonstrated an average percentage of contamination below or close to 5% in negative control experiments. When suspected sample experiments (group number 4 from the Study design) were analysed, it was found that definition 1, 3, 4 and 5 gave the average results in the 42–54% range, while definition 2 nearly twice as high at 88.1%. The analysis of experiments with cancer cells collected by laser microdissection showed that the application of definitions 3, 4 and 5 was obtained with the average contamination

Table 3. Smooth muscle tissue-specific genes:
A — selected from TiSGeD database and BioGPS portal, B — selected by differential expression analysis using R/Bioconductor. Annotation highlights based on GeneCards (Safran *et al.*, 2010; http://www.genecards.org/) and Atlas of Genetics and Cytogenetics in Oncology and Haematology (Huret *et al.*, 2013; http://atlasgeneticsoncology.org/).

Gene		Probesets			Annotation highlights
		HG-U133A	HG-U133Av2	HG-U133plus2	
A	CCL7	208075_s_at	208075_s_at	208075_s_at	chemotactic attractor of monocytes and eosinophils
	CCL8	214038_at	214038_at	214038_at	chemotactic attractor of monocytes, lymphocytes, basophils and eosinophils
	CSF3	207442_at	207442_at	207442_at	granulocyte colony-stimulation
	CXCL1	204470_at	204470_at	204470_at	chemotactic attractor of neutrophils
	CXCL3	207850_at	207850_at	207850_at	inflammatory response
	CXCL6	206336_at	206336_at	206336_at	chemotactic attractor of neutrophils
	FGF2	204421_s_ at204422_s_at	204421_s_at 204422_s_at	204421_s_at 204422_s_at	pleiotropic signaling molecule, associated with many cancers
	GFPT2	205100_at	205100_at	205100_at	controls the flux of glucose into the hexosamine pathway
	IL6	205207_at	205207_at	205207_at	inflammatory response, implicated in various cancers
	LRRC17	205381_at	205381_at	205381_at 232924_at	osteoblast differentiation and proliferation
	MMP1	204475_at	204475_at	204475_at	matrix metallopeptidase
	PLOD2	202619_s_at 202620_s_at	202619_s_at 202620_s_at	202619_s_at 202620_s_at	procollagen-lysine, role in stability of the collagen cross-links
	PXDN	212012_at 213013_at	212012_at 213013_at	212012_at 213013_at	peroxidasin homolog, extracellular matrix formation
	STC1	204595_s_a 204596_s_at 204597_x_at	204595_s_at 204596_s_at 204597_x_at	204595_s_at 204596_s_at 204597_x_at	stimulation of renal phosphate reabsorption
	THBS2	203083_at	203083_at	203083_at	stimulation of chemotaxis, implicated in various cancers
B	BGN	201261_x_at 201262_s_at 213905_x_at	201261_x_at 201262_s_at 213905_x_at	201261_x_at 201262_s_at 213905_x_at	may be involved in collagen fiber assembly
	FGF5	208378_x_at 210310_s_at 210311_at	208378_x_at 210310_s_at 210311_at	208378_x_at 210310_s_at 210311_at	regulation of proliferation and differentiation of hair cells
	IFFO1	209721_s_at 36030_at	209721_s_at 36030_at	209721_s_at 36030_at	cytoskeleton, nuclear envelope
	COL1A1	202311_s_at	202311_s_at	202311_s_at	collagen, type I, alpha 1, implicated in skin tumours
	ITGA4	205884_at 205885_s_at	205884_at 205885_s_at	205884_at 205885_s_at	regulation of immune response
	PRRX1	205991_s_at	205991_s_at	205991_s_at	regulator of muscle creatine kinase
	ARHGAP22	206298_at	206298_at	206298_at	signal transduction
	DCN	211813_x_at 211896_s_at	211813_x_at 211896_s_at	211813_x_at 211896_s_at	connective tissue, suppression of tumour growth
	PAMR1	213661_at	213661_at	213661_at	muscle regeneration
	ELTD1	219134_at	219134_at	219134_at	could be involved in cardiac development
	C1orf54	219506_at	219506_at	219506_at	uncharacterized protein

Table 4. Average percentage of smooth muscle contamination in all five groups.

Group		Assays	Experiments	Expected contamination	Smooth muscle				
					Def.1 (%)	Def.2 (%)	Def.3 (%)	Def.4 (%)	Def.5 (%)
1	fully contaminated	192	6	100	79.7	99	87	95.3	100
2	not contaminated	631	29	0	18.2	37.2	8	7	0.05
3	control	291	11	0	0.7	46.7	2.4	5.5	3
4	suspected of contamination	1021	16	0–100	46.5	88.1	42.9	50.1	53.9
5	microdissection	157	5	0	9	42.7	1.3	4.5	0.6

Table 5. Detailed percentage of smooth muscle contamination in each experiment from all five groups detected by each Smooth Muscle definition.

Group	Experiment	Platform	Assays	Material	Def.1	Def.2	Def.3	Def.4	Def.5
1	E-GEOD-11917*	Plus2	105	smooth muscles	85.7	100	100	100	100
1	E-MEXP-569	A	8	smooth muscles	100	100	100	100	100
1	E-GEOD-21363	only A	6	smooth muscles	100	100	100	100	100
1	E-GEOD-12261*	Plus2	12	smooth muscles	100	100	100	100	100
1	E-GEOD-19672*	Plus2	7	smooth muscles	100	100	100	100	100
1	E-GEOD-13168	A	54	smooth muscles	55.6	94.5	53.7	83.3	100
2	E-GEOD-11839	Plus2	12	bladder	66.7	100	16.7	25	0
2	E-MTAB-37*	Plus2	30	bladder	20	36.7	30	10	0
2	E-GEOD-26828	Plus2	9	bladder	0	100	0	0	0
2	E-GEOD-21654*	Plus2	22	pancreas	40.9	40.9	13.6	13.6	0
2	E-GEOD-22973	Plus2	12	pancreas	25	33.3	16.7	0	0
2	E-GEOD-37645	Plus2	18	pancreas	22.2	22.2	0	0	0
2	E-GEOD-22337	Plus2	12	pancreas	0	0	0	0	0
2	E-GEOD-22183*	Plus2	37	gastric	18.9	16.2	0	5.4	0
2	E-GEOD-32540	Plus2	9	gastric	0	0	0	0	0
2	E-GEOD-20058	Plus2	10	gastric	0	0	0	0	0
2	E-GEOD-35830	Plus2	12	cervix	0	100	25	0	0
2	E-MTAB-37*	Plus2	21	cervix	42.9	57.1	0	0	0
2	E-MEXP-1078	Plus2	6	cervix	0	100	0	0	0
2	E-GEOD-13309*	Plus2	24	lung	0	0	0	0	0
2	E-GEOD-4824	only A	85	lung	8.2	34.1	4.7	5.9	0
2	E-GEOD-30660	Plus2	8	lung	0	0	0	0	0
2	E-GEOD-18454	Plus2	12	lung	66.7	100	33.3	41.7	0
2	E-GEOD-36176	Plus2	32	lung	100	78.1	50	50	0
2	E-GEOD-29008	Plus2	20	colon	0	0	0	0	0
2	E-GEOD-13367	Plus2	28	colon	0	0	3.6	0	0
2	E-GEOD-15132	Plus2	18	colon	0	0	0	0	0
2	E-MEXP-2010	Plus2	22	colon	0	0	0	0	0
2	E-GEOD-15799	Plus2	6	colon	0	0	0	0	0
2	E-GEOD-10650	Plus2	6	colon	0	0	0	0	0
2	E-GEOD-35973	Plus2	8	colon	0	0	0	0	0
2	E-GEOD-30292*	Plus2	43	colon	4.7	25.6	4.7	7	7
2	E-GEOD-30304	Plus2	18	prostate	0	61.1	0	0	0
2	E-TABM-948	Plus2	73	prostate	24.7	72.6	0	2.7	0
2	E-GEOD-17482*	Plus2	18	prostate	11.1	50	22.2	0	0
3	E-GEOD-25014	Plus2	24	lung	0	100	16.7	37.5	0
3	E-GEOD-18269	Plus2	15	liver	0	20	0	0	0
3	E-GEOD-25155	Plus2	28	kidney, gastric	0	7.1	0	0	0
3	E-MTAB-274	Plus2	40	blood	0	0	0	0	0
3	E-GEOD-16249	Plus2	8	skin	12.5	12.5	0	0	0
3	E-GEOD-21354	Plus2	18	brain	5.6	38.9	16.7	38.9	0
3	E-GEOD-40968	Plus2	18	breast	0	100	0	0	0
3	E-GEOD-15932	Plus2	32	blood	0	0	0	0	0
3	E-MEXP-2957	A, Plus2	24	blood	0	0	0	0	0
3	E-MEXP-118864	Plus2	84	breast	0	96.4	0	0	10.7
4	E-GEOD-5287	A	30	bladder	6.7	86.7	0	10	20

4	E-GEOD-7476	Plus2	12	bladder	8.3	83.3	16.7	16.7	33.3
4	E-GEOD-31684	Plus2	93	bladder	30.1	93.5	14	24.7	40.9
4	E-GEOD-32676	Plus2	32	pancreas	71.9	96.9	37.5	78.1	84.4
4	E-GEOD-22780	Plus2	16	pancreas	43.8	93.8	18.8	68.8	68.8
4	E-GEOD-14208	Av2	123	gastric	37.4	81.3	26	29.3	10.6
4	E-GEOD-22377	Plus2	43	gastric	46.5	97.7	27.9	58.1	86
4	E-GEOD-35809	Plus2	70	gastric	58.6	100	65.7	67.1	78.6
4	E-GEOD-5787	Plus2	33	cervix	48.5	84.8	33.3	36.4	18.2
4	E-GEOD-18842	Plus2	91	lung	67	92.3	59.3	62.6	82.4
4	E-GEOD-19188	Plus2	156	lung	45.5	74.4	48.7	47.4	57.1
4	E-GEOD-4183	Plus2	53	colon	49.1	67.9	34	30.2	26.4
4	E-GEOD-23878	Plus2	59	colon	30.5	78	33.9	20.3	13.6
4	E-GEOD-31595	Plus2	37	colon	46	100	54.1	48.6	32.4
4	E-GEOD-3325	Plus2	19	prostate	0	89.5	10.5	15.8	10.5
4	E-GEOD-17951	Plus2	154	prostate	64.3	100	76	96.1	99.4
5	E-GEOD-19650	Plus2	22	pancreas	31.8	63.6	4.5	18.2	4.5
5	E-GEOD-27716	Plus2	40	lung	5	62.5	2.5	5	0
5	E-GEOD-7803	A	41	cervix	12.2	61	0	2.4	0
5	E-MEXP-383	A	36	colon	0	8.3	0	0	0
5	E-GEOD-15960	Plus2	18	colon	0	0	0	0	0

*experiments included in differential expression analysis; Platform: Plus2 = HG-U133plus2, A = HG-U133A, Av2 = HG-U133Av2; Def. 1–5 (%) — number of contaminated assays compared to all assays from experiment; data obtained from ArrayExpress repository at http://www.ebi.ac.uk/arrayexpress.

level under 5%. For definition 1, the average contamination was raised to 9%, whereas the maximum value was obtained for definition 2.

DISCUSSION

In this study, two strategies of designing smooth muscle definitions were shown. The tissue definition itself was not just a simple set of biomarkers. Instead, it was a collection of probesets that, only when taken together, enable identification of smooth muscle presence in the sample. Single probesets composing the definitions were not sufficient to perform the test. The whole expression of tissue definition was important in detection of contamination.

The first way to create smooth muscle tissue definition was based on the TiSGeD database and the BioGPS portal. Surprisingly, none of the four definitions created that way, was optimal for verifying the purity of the sample. Smooth muscle definition 4, although not perfect, seemed to be the closest to expected results. Unfortunately, definitions 1 and 3 were not good enough either — they provided a high percentage of contamination in the not contaminated experiments, and also gave the lowest value of contamination in the fully contaminated group (below 90%). SM definition 2 in both not contaminated and control experiments demonstrated the highest average percentage of reported contamination. Some of the smooth muscle tissue-specific genes selected from public databases are either expressed at high levels in other tissues, or at too low levels in smooth muscle tissue. The authors of the TiSGeD database claimed that at times, the assignment of gene expression tissue specificity was inconsistent. According to them, one of the reasons for this was the difference in the tissue scale between experiments (Xiao *et al.*, 2010). What this means is that the material used in transcription profiling analysis was in some cases contaminated with other tissues or cells in different degrees, which affected gene expression results. This might explain why many genes selected with the TiSGeD-BioGPS strategy were associated with inflammatory response (*CCL7*, *CCL8*, *CSF3*, *CXCL1*, *CXCL3*, *CXCL6*, *IL6*, *THBS2*) and some were implicated in various cancers (*FGF2*, *IL6*, *THBS2*). We should also take into account that the results from both databases were based on integration studies of independent microarray data.

Smooth muscle definition 5 created by differential expression analysis has proven to be the optimal definition. All the results were in line with expectations including analysis of the sample purity from group number 5, where the material was obtained by laser microdissection (assuming a 5% error). The differential expression analysis method used in designing tissue definition appeared to be working better. The reason for this, is that this method was based on the microarray data from only one platform — HG-U133Plus2. It was previously reported that the most reliable quantitative results of integrated analysis were obtained from the same platform (Shi *et al.*, 2006). Another advantage of this method is the utilization of 88 different cancer cell lines of seven human organs and several different smooth muscle cells. Smooth muscle cells from different anatomical locations have many common morphological and molecular features. Nevertheless, they also have individual properties and functions. For instance, colon smooth muscle cells are responsible for moving food in the digestive system, whereas vascular SM regulate the flow of blood through the blood vessels. As it was reported, SMs have distinct expression patterns associated with the anatomical location (vascular system, visceral organs, or bronchi) (Chi

et al., 2006). For this reason, several different smooth muscle cells derived from various body locations, such as coronary artery, aorta, and penis were included in the differential gene expression analysis.

Genes composing SM definition 5 came from a wider variety of functions. Several of them were reported to be involved in collagen formation (*BGN*, *COL1A1*) and muscle functioning (*PRRX1*, *PAMR1*). *COL1A1* was also implicated in skin cancers and *ITGA4* was reported as involved in regulation of the immune response, which was similar to most of the genes' functions in definitions 1–4. The function of *ELTD1* and *C1orf54* is undetermined, and in the light of this work, could be an interesting target for future studies.

The results obtained in this research indicated that about 50% of cancer tissue samples derived from seven listed organs were likely contaminated with smooth muscle tissue. Possible smooth muscle contamination was already detected during expression profiling of the human bladder (E-GEOD-7476) and gastric (E-GEOD-22377) cancer, as mentioned by the authors in (Mengual *et al.*, 2009; Förster *et al.*, 2011). The contamination in the first experiment was identified mostly in control samples causing complications in interpretation of the results. Interestingly, gene expression analysis of muscle-invasive bladder cancers (microdissected cancer cells) resulted in smooth muscle contamination being detected (E-GEOD-31684 by Riester *et al.*, 2012). In this case, the presence of smooth muscle tissue could indicate the presence of contamination, as well as cancer progression and invasion of smooth muscle tissue. Perhaps when equipped with SM definition 5, Microarray Inspector could be useful in prediction of muscle invasive cancers.

CONCLUSIONS

The heterogeneity of sample composition in microarray analysis causes differences in the results obtained by different laboratories. We propose methods that prove useful when verifying the purity of the test material with a high possibility of smooth muscle contamination, which was derived from most common cancers in the human population. With the information provided in this paper, the users of Microarray Inspector will be able to use our definition or to design other tissue definitions crafted for their own needs. We believe that a proper verification of tissue sample contamination enables avoiding incorrect conclusions from obtained results.

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].

REFERENCES

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

Carlson M (2012) GO.db: A set of annotation maps describing the entire Gene Ontology.

Carlson M (2012) KEGG.db: A set of annotation maps for KEGG.

Chi J-T, Rodriguez EH, Wang Z, Nuyten DSA, Mukherjee S, Van de Rijn M, Van de Vijver MJ, Hastie T, Brown PO (2007) Gene expression programs of human smooth muscle cells: tissue-specific differentiation and prognostic significance in breast cancers. *PLoS Genet* **3**: 1770–1784.

Chunlei CO, 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.

Förster S, Gretschel S, Jöns T, Yashiro M, Kemmner W (2011) THBS4, a novel stromal molecule of diffuse-type gastric adenocarcinomas, identified by transcriptome-wide expression profiling. *Mod Pathol* **24**: 1390–403.

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.

Gentleman R, Carey V, Huber W, Hahne F (2012) Genefilter: methods for filtering genes from microarray experiments.

Gentleman R (2012) annotate: Annotation for microarrays. (http://www.bioconductor.org/packages/release/bioc/html/annotate.html).

Huret JL, Ahmad M, Arsaban M, Bernheim A, Cigna J, Desangles F, Guignard JC, Jacquemot-Perbal MC, Labarussias M, Leberre V, Malo A, Morel-Pair C, Mossafa H, Potier JC, Texier G, ViguiÓ F, Yau Chun Wan-Senon S, Zasadzinski A, Dessen P (2013) Atlas of genetics and cytogenetics in oncology and haematology in 2013. *Nucleic Acids Res* **41** (Database issue): D920–D924.

Kornmann M, Ishiwata T, Beger HG, Korc M (1997) Fibroblast growth factor-5 stimulates mitogenic signaling and is overexpressed in human pancreatic cancer: evidence for autocrine and paracrine actions. *Oncogene* **15**: 1417–1424.

Lang DT (2012) XML: Tools for parsing and generating XML within R and S-Plus. (http://cran.r-project.org/web/packages/XML/index.html).

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.

Mengual L, Burset M, Ars E, Lozano JJ, Villavicencio H, Ribal MJ, Alcaraz A (2009) DNA microarray expression profiling of bladder cancer allows identification of noninvasive diagnostic markers. *J Urol* **182**: 741–748.

Parkinson H, Kapushesky M, Shojatalab M, Abeygunawardena N, Coulson R, Farne A, Holloway E, Kolesnykov N, Lilja P, Lukk M, Mani R, Rayner T, Sharma A, William E, Sarkans U, Brazma A (2007) ArrayExpress — a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res* **35**: D747–D750.

Paweletz CP, Liotta LA, Petricoin EF (2001) New technologies for biomarker analysis of prostate cancer progression: Laser capture microdissection and tissue proteomics. *J Urol* **57**: 160–163.

Riester M, Taylor JM, Feifer A, Koppie T, Rosenberg JE, Downey RJ, Bochner BH, Michor F (2012) Combination of a novel gene expression signature with a clinical nomogram improves the prediction of survival in high-risk bladder cancer. *Clin Cancer Res* **18**: 1323–1333.

Roepman P, De Koning E, Van Leenen D, De Weger RA, Kummer JA, Slootweg PJ, Holstege FCP (2006) Dissection of a metastatic gene expression signature into distinct components. *Genome Biol* **7**: R117.

Stępniak P, Maycock M, Wojdan K, Markowska M, Perun S, Srivastava A, Wyrwicz LS, Świrski K (2013) Microarray Inspector: tissue cross contamination detection tool for microarray data. *Acta Biochim Pol* **60**: 647–655.

Safran M, Dalah I, Alexander J, Rosen N, Iny Stein T, Shmoish M, Nativ N, Bahir I, Doniger T, Krug H., Sirota-Madi A, Olender T, Golan Y, Steltzer G, Harel A, Lancet D (2010) GeneCards Version 3: the human gene integrator. Database (Oxford) 2010: baq020.

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

Smith CA (2010) Annaffy: Annotation tools for Affymetrix biological metadata. (http://www.bioconductor.org/packages/devel/bioc/html/annaffy.html).

Team RC (2012) R: A Language and Environment for Statistical Computing. (http://www.r-project.org/).

Wang M, Master SR, Chodosh LA (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.

Wu J, with contributions from James MacDonald Jeff Gentry RI (2012) gcrma: Background Adjustment Using Sequence Information. (http://bioconductor.wustl.edu/bioc/html/gcrma.html).

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