Original Paper

Combined Gene Expression Analysis of Whole-Tissue and Microdissected Pancreatic Ductal Adenocarcinoma identifies Genes Specifically Overexpressed in Tumor Epithelia

Liviu Badea[1], Vlad Herlea[2], Simona Olimpia Dima[3], Traian Dumitrascu[3], Irinel Popescu[3,4*]*
[1]Bioinformatics group, National Institute for Research in Informatics, Bucharest 011455, Romania
[2]Department of Pathology, Fundeni Clinical Institute, Bucharest, Romania
[3]Center of General Surgery and Liver Transplantation, Fundeni Clinical Institute, Bucharest, Romania
[4]Carol Davila University of Medicine and Pharmacy, Bucharest, Romania.
*Corresponding Authors: Liviu Badea, Bioinformatics group, National Institute for Research in Informatics, 8-10 Averescu Blvd., Bucharest, Romania
Tel: +40-744-777745; E-mail: badea.liviu@gmail.com
Irinel Popescu, Carol Davila University of Medicine and Pharmacy, Bucharest, Romania
E-mail: irinel.popescu@icfundeni.ro

ABSTRACT

Background: The precise details of pancreatic ductal adenocarcinoma (PDAC) pathogenesis are still insufficiently known, requiring the use of hig-throughput methods. However, PDAC is especially difficult to study using microarrays due to its strong desmoplastic reaction, which involves a hyper-proliferating stroma that effectively "masks" the contribution of the minoritary neoplastic epithelial cells. Thus it is not clear which of the genes that have been found differentially expressed between normal and whole tumor tissues are due to the tumor epithelia and which simply reflect the differences in cellular composition. To address this problem, laser microdissection studies have been performed, but these have to deal with much smaller tissue sample quantities and therefore have significantly higher experimental noise.

Methodology: In this paper we combine our own large sample whole-tissue study with a previously published smaller sample microdissection study by Grützmann *et al.* to identify the genes that are specifically over-expressed in PDAC tumor epithelia.

Results: The overlap of this list of genes with other microarray studies of pancreatic cancer as well as with the published literature is impressive. Moreover, we find a number of genes whose over-expression appears to be inversely correlated with patient survival: keratin 7, laminin gamma 2, stratifin, platelet phosphofructokinase, annexin A2, MAP4K4 and OACT2 (MBOAT2), which are all specifically upregulated in the neoplastic epithelia, rather than the tumor stroma.

Conclusions: We improve on other microarray studies of PDAC by putting together the higher statistical power due to a larger number of samples with information about cell-type specific expression and patient survival.

KEYWORDS
Pancreatic Ductal Adenocarcinoma (PDAC); Cancer Genomics; Microarrays; Combined Gene Expression Analysis; Genes Correlated with Survival

ABBREVIATIONS:
Pancreatic Ductal Adenocarcinoma (PDAC); Normal Epithelia (NE), Tumor Epithelia (TE), Chronic Pancreatitis Stroma (CPS); Tumor Stroma (TS); Clinical Institute Fundeni (ICF) dataset (described in this paper); Grützmann-Pilarsky (GP) dataset

INTRODUCTION

Pancreatic ductal adenocarcinoma (PDAC) is among the top five causes of cancer death in the Western world, with a less than 6 months median survival and a 5-year survival rate of about 5% (1). Its very aggressive nature, rapid dissemination and resistance to conventional and targeted therapeutic agents require concentrated efforts for a more detailed molecular-level understanding of its evolution to enable the development of more effective therapeutics.

The last decade has witnessed significant advances in understanding PDAC, including important insights into the cellular origins of the disease (2), the development of ever more refined mouse models of human PDAC (3,4), a large number of high-throughput genomic studies, including gene expression studies (5-29), high-resolution array-CGH measurements of genomic amplifications and deletions (30-32), etc. However, despite these very important advances, the precise molecular and genomic-level details of the disease are still elusive, requiring the use of high-throughput methods.

PDAC is especially difficult to study due to its strong desmoplastic reaction (33, 34), with the neoplastic epithelial cells representing just a minority of the bulk tumor tissue. Although a large number of PDAC microarray studies have been published (5-29), most have quite limited numbers of samples and measure gene expression levels in whole pancreatic tumor tissue, which contains complex mixtures of various cell types. Thus it is not clear which of the genes that have been found differentially expressed between normal and tumor tissues are due to the tumor epithelia and which simply reflect the differences in cellular composition (e.g. due to a predominant stromal component in the tumors). To

Hepato-Gastroenterology 2008; 55:2015-2026

deal with this problem, laser microdissection studies have been performed (12,15,22), but these involve much smaller tissue sample quantities and therefore have significantly higher experimental noise.

In the present work we combine our own large sample whole tissue study (denoted in the following by ICF) with a previously published smaller sample microdissection study by Grützmann *et al.* and Pilarsky *et al.* (GP) (10,35) and show that although neither of the studies can adequately address all of the above-mentioned problems, their combination can be much more informative. More precisely, the larger number of samples in our whole tissue study enables a much more reliable detection of the genes that are differentially expressed between normal and tumor tissues. The increased reliability is due not only to the larger sample size, but also to the higher quality of microarray measurements for whole-tissue samples. On the other hand, as previously mentioned, the whole tissue study cannot distinguish between genes over-expressed in the tumor epithelia and those that are differentially expressed between the various cell types present in normal and respectively tumor samples.

The combination of the two studies may therefore enable a more precise determination of the genes that are specifically over-expressed in tumor epithelia. Additionally, the whole-tissue study may also identify genes over-expressed in cell types that are too difficult to isolate with laser capture microdissection.

METHODOLOGY [1]

Pancreatic tissues

Pairs of normal and tumor tissue samples were obtained at the time of surgery from resected pancreas of 36 pancreatic cancer patients and snap-frozen in liquid nitrogen. The diagnosis, histological typing, grading and staging were performed by an experienced pathologist (V.H.).

GeneChip hybridization

Gene expression was analyzed on Affymetrix U133 plus 2.0 whole genome microarrays (54675 probesets). Target preparation and microarray processing were performed as described in the Affymetrix GeneChip Expression Analysis Manual. For three of the 36 normal-tumor sample pairs we carried out replicate microarray hybridizations in order to gauge the technical measurement errors. We thus performed 78 genechip hybridizations in total. The data will be deposited in the Gene Expression Omnibus (GEO).

Scanning data analysis

Affymetrix GeneChip Operating Software (GCOS) Version 1.4 (Affymetrix Inc., Santa Clara, CA) was used for low-level scanning data processing. A patient sample pair was excluded from further analysis since one of the samples did not meet the quality controls. The microarray data was subsequently normalized using the Robust Microarray Analysis (RMA) algorithm (36).

Statistical analysis

Since the expression data is approximately log-normally distributed, we used the log-transformed data as produced by the RMA algorithm for all subsequent statistical tests. For visualization purposes, we centred the log-transformed expression data by subtracting the average probeset log-expression values. Probesets with relatively low expression (average expression values below 100 Affymetrix units) or with nearly constant expression values (standard deviation below 50) were excluded from further consideration. Of the 54675 probesets on the U133 Plus 2.0 chip, 12209 were thus retained. An unpaired t-test was used to determine the probesets (genes) that are differentially expressed between the normal and the tumor tissue samples. The first 400 probesets with the lowest t-test p-values (corresponding to a p-value cutoff of 9×10^{-12}) were selected for further analysis. We also used a more stringent fold-change constraint that excluded the probesets with log-fold change < 1 (roughly corresponding to a fold change < 2), where the log-fold change of gene g between classes N ('normal') and T ('tumor') is defined as

$$\text{with} \qquad log\text{-}fc(g, N-T) = \overline{\log_2 g(T)} - \overline{\log_2 g(N)},$$

$$\overline{\log_2 g(C)} = \frac{\sum_{S_i \in C} \log_2 g(S_i)}{|C|}$$

the average log-expression value of gene g in the samples S_i of class C. We also performed a univariate Cox proportional hazards analysis of the correlation of the individual gene expression profiles with the patient post-operative survival. A preliminary functional analysis of the set of differentially expressed genes was performed with the L2L online tool (37), as well as with the TRED Transcription Regulatory Element Database (38). The gene lists obtained were compared with gene lists from published PDAC studies (5-29). Lists (5-15) were retrieved from the Pancreatic Expression Database (www.pancreasexpression.org), while lists (16-29) were obtained from the meta-analysis of Brandt *et al.* (39).

RT-PCR validation

A number of 20 genes were selected for RT-PCR validation: SULF1, WISP1, FN1, HNT, CSPG2, NOX4, SOX4, BGN, BHLHB2, INHBA, PLAU, IGFBP5, PLAT, DKK3, PDGFC, HOP, CAPG, ETV1, LTBP1, CTHRC1. RLP13A and 18S RNA were used as controls. (RLP13A was selected by searching for the gene with the smallest std/mean ratio across all chips in the microarray study. Since 18S was observed to be much more variable in the RT-PCR data, we report

[1]More detailed descriptions of the materials and methods can be found in the Supplementary Material

measurements normalized with respect to RLP13A.) Quantitative RT-PCR was performed using an Applied Biosystems Sequence Detection System 7300.

Reprocessing of the Grützmann and Pilarsky studies

For separating the epithelial contribution from the dominant stromal signal in our dataset, we also used a PDAC microdissection dataset described in Grützmann *et al* (2004) (10) and Pilarsky *et al.* (2008) (35), which contains Affymetrix U133 A and B measurements for 17 normal epithelia (NE) samples, 20 tumor epithelia (TE), 9 chronic pancreatitis stroma (CPS) and 11 tumor stroma (TS) samples, as well as pancreatic stellate cell and PDAC cell lines. In this paper, we use the following abbreviations: GP for the Grützmann and Pilarsky study data (10,35) and ICF for our data (Clinical Institute Fundeni). We downloaded the raw expression data (CEL files) from ArrayExpress (http://www.ebi.ac.uk/microarray-as/aer/, E-MEXP-950 and E-MEXP-1121) and reprocessed the raw data using the RMA normalization algorithm (the dChip algorithm with the PM/MM setting as used in the original papers (10,35) occasionally produces negative expression values). The cell line measurements were excluded from further processing since the cell lines have been observed to have significantly different gene expression values than the surgical tissue samples, thereby potentially affecting RMA quantile normalization. **Supplementary Figure 1**[2] shows that the relative standard deviations of the Affymetrix control probesets are about two times larger in the microdissection study than in our study. The experimental noise is therefore significantly higher in the microdissection study than in the whole tissue study.

We also observed a pronounced 3' bias for the housekeeping genes (e.g. GAPDH), which is probably due to the several rounds of RNA amplification used in the GP study. However, since the majority of probesets on the U133A/B chips represent the 3' ends of transcripts, this bias may not have a major impact on the quantification of most transcripts.

We excluded samples TPK9, NPK13 and NPD15 from the dataset based on their very large scaling factors (over 39), small percentage of present probesets (< 16%) as well as very frequent outliers for the Affymetrix control probesets. Due to the smaller

[2]Supplementary Figures and Tables can be found in the Supplementary Material online.

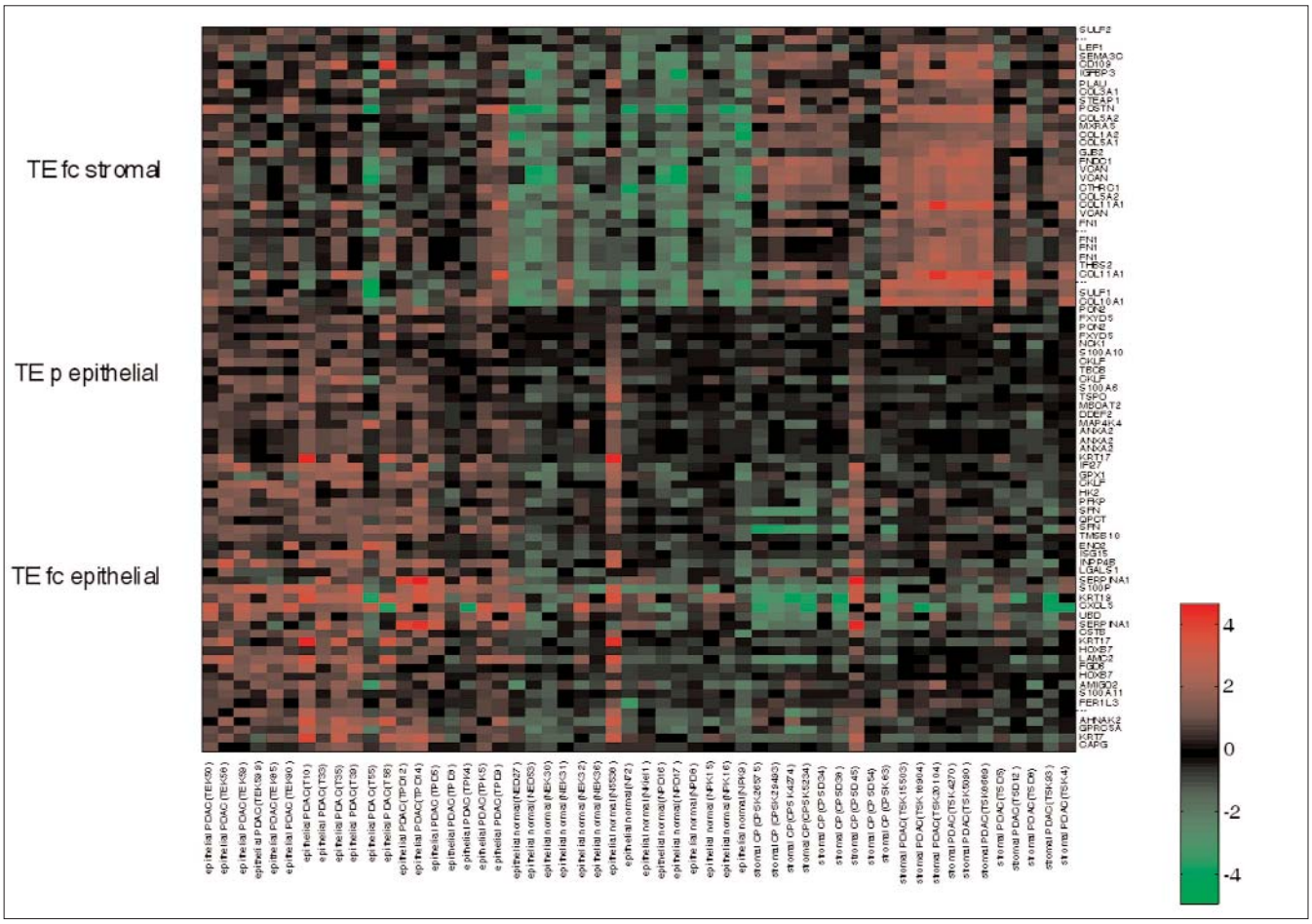

FIGURE 1 Centred log2-expression values of the genes overexpressed in tumor epithelia in the Grützmann-Pilarsky (GP) study

sample sizes as well as to the significantly higher experimental noise in the microdissection study, we used a more lenient p-value cutoff of 0.01 for all differential expression t-tests on the GP data.

Joint statistical analysis of the ICF and GP datasets

Since the normal and the tumor whole tissue samples from the ICF study contain very different cell compositions, the genes that we find overexpressed in whole tissue tumors may not necessarily be due to the tumor epithelia, but could reflect the abundant stromal proliferation in tumors due to the desmoplastic reaction. In the following, we use the following abbreviations for the various cell types: *NE* - normal epithelia, *TE* - tumor epithelia, *CPS* - chronic pancreatitis stroma, *TS* - tumor stroma. Since both the whole tissue (ICF) and the microdissection study (GP) have their advantages and drawbacks, we combined the two studies. The whole tissue study has a larger number of samples and smaller measurement errors, which enables a more reliable determination of the genes that are differentially expressed in PDAC whole tissue tumors. Unfortunately, many of these genes may be due to the stromal over-proliferation in PDAC tumors rather than the epithelial tumor cells proper. Although the microdissection study is less reliable due to its higher experimental noise and smaller number of samples, it provides crucial information about the localization of the over-expressed genes in the neoplastic epithelia, the tumor stroma, or both.

Although the two datasets (ICF and GP) have been obtained with different microarray chips (U133 Plus 2.0 and U133A/B respectively), the probesets on the U133A and B chips are also present on the U133 Plus 2.0 chip, which establishes a straight-forward correspondence between the two datasets at the probeset level.

To determine the set of genes overexpressed in ICF that are specifically upregulated in the tumor epithelia (rather than just the tumor stroma), we constructed three subsets of genes as follows (here we used a more lenient p-value cutoff of 10^{-9} for ICF, since the genes specifically overexpressed in TE are masked by the TS):

TE_fc_epithelial: $p_{ICF}(N\text{-}T)<10^{-9}$ and $log\text{-}fc_{ICF}(N\text{-}T) \geq 1$ and $log\text{-}fc_{GP}(NE\text{-}TE) \geq 1$ and $TE >_{GP} TS$.

TE_p_epithelial: $p_{ICF}(N\text{-}T)<10^{-9}$ and $log\text{-}fc_{ICF}(N\text{-}T) \geq 1$ and $p_{GP}(NE\text{-}TE)<10^{-2}$ and $p_{GP}(TE\text{-}TS)<10^{-2}$ and $TE >_{GP} TS$ and $0< log\text{-}fc_{GP}(NE\text{-}TE)<1$.

TE_fc_stromal: $p_{ICF}(N\text{-}T)<10^{-9}$ and $log\text{-}fc_{ICF}(N\text{-}T) \geq 1$ and $log\text{-}fc_{GP}(NE\text{-}TE) \geq 1$ and $TE <_{GP} TS$.

TE_fc_epithelial contains the genes that are (more or less) highly over-expressed (at least two-fold) in TE vs. NE and which are also over-expressed in TE vs. TS (otherwise, their expression in TE would be masked by the TS in the ICF tumor samples). *TE_p_epithelial* collects the genes that are only mildly (less than 2-fold) but quite consistently (with a small p-value) over-expressed in TE vs. NE. *TE_fc_stromal* also contains genes highly over-expressed in TE vs. NE, but their expression is even higher in TS, so their increase observed in the whole tissue tumor samples (in the ICF data) is predominantly due to their stromal component.

RESULTS

The genes overexpressed in PDAC whole tumor tissue contain a TGF-beta signature

We performed an expression analysis of 36 pancreatic ductal adenocarcinoma tumors and matching normal pancreatic tissue samples from pancreatic cancer patients of the Clinical Institute Fundeni (ICF) using Affymetrix U133 Plus 2.0 whole-genome chips. We started our analysis by determining the probesets (genes) that are differentially expressed between the normal and the tumor whole tissue samples in the ICF dataset. The first 400 probesets with the lowest t-test p-values (corresponding to a p-value cutoff of $9x10^{-12}$) were selected for a more in-depth analysis. Strikingly, virtually all these genes were found to be upregulated in tumors. (Only three genes out of 400, namely EIF4E3, SFXN2 and TOPORS were found downregulated - see also **Supplementary Table 2**. However, these may be false positives, as their 'normal' vs. 'tumor' fold changes are below 2 and they are not differentially expressed in the GP data).

Table 1 presents the list of 65 genes with p-values below 10^{-14} and a log-fold change above 1 (roughly corresponding to at least 2-fold over-expression in tumors). **Supplementary Table 1** contains the full list of 319 probesets (239 unique genes) with p-values below the above-mentioned threshold of $9x10^{-12}$ and a log-fold change above 1[3].

We have compared the list of over-expressed genes obtained in our study with 25 other microarray studies of pancreatic cancer, as shown in **Supplementary Table 3**. The overlap is significant: 203 of the 319 probesets (135 of the 239 unique genes) selected in our study were mentioned in at least one of the other 25 studies. The fluctuating inter-study overlap shows the difficulty of determining the genes differentially expressed in PDAC. A large number of factors are responsible for this: the limited numbers of samples (as far as we know, our dataset has the largest number of samples among the published studies), the genomic heterogeneity of the disease, the large inter-platform variability, the complex cellular compositions of the tumor samples, the different gene selection methods and/or thresholds used in the various studies, etc.

A univariate Cox proportional hazards analysis of the correlation of the individual gene expression profiles with the patient survival produced a number of genes whose over-expression is inversely correlated with patient survival: ANXA2, ANXA2P2, KRT7, MAP4K4, LAMC2, OACT2, SLC16A3, TM4SF1 and KYNU. The associated p-values are below 0.01, indicating marginal correlation with survival (of the 9 genes mentioned, about 3 are expected to be correlated by chance).

[3]The multiple occurrences of the certain genes in the Supplementary Tables is due to the fact that Affymetrix chips have several probesets for many genes, which we do not aggregate. In Table 1 we only list the probeset with the smallest p-value for each gene.

TABLE 1 The Genes Overexpressed in whole Tumor Tissue vs. Normal Pancreatic Tissue ($p < 10^{-14}$ and *log-fc* ≥ 1).

Genes (marginally) correlated with survival are in bold font. Genes involved in the TGF-beta pathway or influenced by it transcriptionally are indicated with an X in the last column

Gene Symbol	Gene Title	ProbeSetID	*p*-value	log-fc	survival (Cox p)	TGF-beta signature
INHBA	Inhibin, beta A (activin A, activin AB alpha polypeptide)	227140_at	<2.2E-16	5.15	0.6706	X
COL10A1	collagen, type X, alpha 1(Schmid metaphyseal chondrodysplasia)	217428_s_at	<2.2E-16	5	0.9609	X
SULF1	sulfatase 1	212353_at	<2.2E-16	4.63	0.8061	
COL8A1	Collagen, type VIII, alpha 1	226237_at	<2.2E-16	4.6	0.6288	X
HNT	neurotrimin	227566_at	<2.2E-16	3.9	0.9306	
NOX4	NADPH oxidase 4	219773_at	<2.2E-16	3.02	0.5473	X
SRPX2	sushi-repeat-containing protein, X-linked 2	205499_at	<2.2E-16	2.44	0.6478	
---	CDNA FLJ10196 fis, clone HEMBA1004776	226997_at	<2.2E-16	2.41	0.5994	
CAPG	capping protein (actin filament), gelsolin-like	201850_at	<2.2E-16	2.25	0.9151	
TMEPAI	transmembrane, prostate androgen induced RNA	222449_at	<2.2E-16	2.19	0.0187	X
LTBP1	latent transforming growth factor beta binding protein 1	202729_s_at	<2.2E-16	1.81	0.8294	X
ITGB5	integrin, beta 5	201125_s_at	<2.2E-16	1.41	0.1833	X
RASAL2	RAS protein activator like 2	222810_s_at	<2.2E-16	1.23	0.3589	
LOC162073	Hypothetical protein LOC162073	1568619_s_at	<2.2E-16	1.03	0.5727	
COL11A1	collagen, type XI, alpha 1	37892_at	2.2E-16	4.4	0.772	X
THBS2	thrombospondin 2	203083_at	2.2E-16	3.97	0.8929	X
FN1	fibronectin 1	212464_s_at	2.2E-16	3.69	0.2846	X
---	---	238617_at	2.2E-16	2.87	0.598	
AEBP1	AE binding protein 1	201792_at	2.2E-16	2.68	0.8476	X
DCBLD1	discoidin, CUB and LCCL domain containing 1	226609_at	2.2E-16	1.66	0.974	
TMEM16A	transmembrane protein 16A	218804_at	4.4E-16	2.9	0.5386	
TPBG	trophoblast glycoprotein	203476_at	4.4E-16	1.63	0.0945	
COMP	cartilage oligomeric matrix protein	205713_s_at	6.7E-16	3.53	0.4253	X
SOX4	SRY (sex determining region Y)-box 4	201416_at	8.9E-16	1.76	0.2043	X
FLJ12442	hypothetical protein FLJ12442	218051_s_at	8.9E-16	1.6	0.4759	
TREM2	triggering receptor expressed on myeloid cells 2	219725_at	1.1E-15	1.62	0.1062	
CSPG2	chondroitin sulfate proteoglycan 2 (versican)	204619_s_at	1.3E-15	3.87	0.9914	X
COL1A2	Collagen, type I, alpha 2	229218_at	1.3E-15	3.16	0.6324	X
SLPI	secretory leukocyte peptidase inhibitor	203021_at	1.3E-15	2.65	0.0642	X
MARVELD1	MARVEL domain containing 1	223095_at	1.6E-15	1.2	0.8907	
ITGA2	Integrin, alpha 2 (CD49B, alpha 2 subunit of VLA-2 receptor)	227314_at	2.0E-15	2.58	0.2731	X
PGM2L1	phosphoglucomutase 2-like 1	229553_at	2.0E-15	1.63	0.5814	
ANTXR1	anthrax toxin receptor 1	224694_at	2.2E-15	2.82	0.9509	
OLR1	oxidised low density lipoprotein (lectin-like) receptor 1	210004_at	2.4E-15	3.03	0.4339	X
MICAL2	microtubule associated monoxygenase, calponin and LIM domain containing 2	212473_s_at	2.7E-15	1.52	0.2504	
MFAP2	microfibrillar-associated protein 2	203417_at	2.9E-15	1.97	0.9108	
ANXA2	**annexin A2**	**210427_x_at**	**2.9E-15**	**1.31**	**0.0073**	**X**
KRT7	**keratin 7**	**209016_s_at**	**3.1E-15**	**3.13**	**0.0051**	**X**
MXRA8	matrix-remodelling associated 8	213422_s_at	3.1E-15	1.69	0.8417	
PDGFC	platelet derived growth factor C	218718_at	3.3E-15	1.36	0.7998	X
POSTN	periostin, osteoblast specific factor	1555778_a_at	3.6E-15	4.92	0.4624	X
IRS1	insulin receptor substrate 1	204686_at	3.6E-15	1.24	0.0553	
COX7A1	Cytochrome c oxidase subunit VIIa polypeptide 1 (muscle)	228481_at	3.8E-15	2.77	0.725	
ADAMTS2	ADAM metallopeptidase with thrombospondin type 1 motif, 2	226311_at	3.8E-15	2.76	0.5509	X
SPON2	spondin 2, extracellular matrix protein	218638_s_at	3.8E-15	1.8	0.6711	
FBXO32	F-box protein 32	225803_at	3.8E-15	1.69	0.6703	
FAP	fibroblast activation protein, alpha	209955_s_at	4.2E-15	3.37	0.9252	
LY6E	lymphocyte antigen 6 complex, locus E	202145_at	4.2E-15	1.53	0.0856	
LOXL1	lysyl oxidase-like 1	203570_at	4.4E-15	2.3	0.845	X
SH3MD1	SH3 multiple domains 1	224817_at	4.4E-15	1.59	0.3483	
MFGE8	milk fat globule-EGF factor 8 protein	210605_s_at	4.7E-15	1.21	0.1058	
SLC24A3	solute carrier family 24 (sodium/potassium/calcium exchanger), member 3	57588_at	4.9E-15	1.49	0.514	
C5orf13	chromosome 5 open reading frame 13	201310_s_at	5.1E-15	2.07	0.9133	X
S100A11	S100 calcium binding protein A11 (calgizzarin)	208540_x_at	5.1E-15	1.65	0.3704	X
EDNRA	endothelin receptor type A	216235_s_at	5.6E-15	1.82	0.8131	X
COL5A2	collagen, type V, alpha 2	221729_at	5.8E-15	3.38	0.8486	X
PLXDC2	plexin domain containing 2	227276_at	6.0E-15	2.16	0.3087	
NUAK1	NUAK family, SNF1-like kinase, 1	204589_at	6.0E-15	1.67	0.8995	
DACT1	dapper, antagonist of beta-catenin, homolog 1 (Xenopus laevis)	219179_at	7.1E-15	2.36	0.8479	
IER5L	immediate early response 5-like	226552_at	7.1E-15	1.31	0.5595	
CTHRC1	collagen triple helix repeat containing 1	225681_at	7.3E-15	4.38	0.9419	X
ITGB1	integrin, beta 1 (fibronectin receptor, beta polypeptide, antigen CD29 includes MDF2, MSK12)	211945_s_at	8.4E-15	1.12	0.0335	X
COL1A1	collagen, type I, alpha 1	202311_s_at	8.7E-15	4.12	0.8808	X
COL16A1	collagen, type XVI, alpha 1	204345_at	9.1E-15	1.5	0.3062	X
WISP1	WNT1 inducible signaling pathway protein 1	229802_at	9.5E-15	3.8	0.686	X

The microarray results were validated with quantitative RT-PCR (using the same total RNA extractions) for the following 20 genes: SULF1, WISP1, FN1, HNT, CSPG2, NOX4, SOX4, BGN, BHLHB2, INHBA, PLAU, IGFBP5, PLAT, DKK3, PDGFC, HOP, CAPG, ETV1, LTBP1, CTHRC1. **Supplementary Table 5** presents a detailed comparison of the RT-PCR and microarray measurements, including Pearson correlations between the two techniques, as well as p-values of the differential expression tests. The average correlation (on log2-transformed data) between RT-PCR and microarrays was very high: 0.93 (standard deviation 0.0538). We also noted very low p-values of the t-tests, showing a significant differential expression of the 20 genes between normal and tumor tissues.

A number of 4-5 "normal" samples depart signifi-

TABLE 2 Genes Specifically over-expressed in tumor epithelia according to the ICF and GP datasets. Genes (marginally) correlated with survival are in bold font.

Gene Symbol	Gene Title	Probe Set ID	N-T (ICF)			NE-TE (GP)		TE-TS (GP)	
			p-value	log-fc	Cox p	*p*-value	log-fc	*p*-value	log-fc
TE_fc_epithelial									
CAPG	capping protein (actin filament), gelsolin-like	201850_at	0.0E+0	2.25	0.9151	1.4E-4	1.01	2.0E-3	-0.8
KRT7	keratin 7	209016_s_at	3.1E-15	3.13	**0.0051**	1.2E-2	1.24	4.3E-6	-2.22
GPRC5A	G protein-coupled receptor, family C, group 5, member A	203108_at	1.3E-14	2.97	0.0545	3.0E-4	1.61	3.1E-4	-1.31
C14orf78	chromosome 14 open reading frame 78	212992_at	1.7E-14	2.49	0.0179	2.0E-5	2.25	1.8E-5	-1.82
FER1L3	fer-1-like 3, myoferlin (C. elegans)	201798_s_at	2.0E-14	2.32	0.3854	4.3E-4	1.13	1.8E-1	-0.39
S100A11	S100 calcium binding protein A11 (calgizzarin)	200660_at	2.2E-14	2.03	0.5039	1.3E-5	1	4.1E-4	-0.82
AMIGO2	adhesion molecule with Ig-like domain 2	222108_at	1.2E-13	1.79	0.1869	3.7E-3	1.36	2.1E-1	-0.56
HOXB7	homeo box B7	216973_s_at	1.3E-13	1.23	0.522	1.8E-6	1.13	1.3E-3	-0.64
FGD6	FYVE, RhoGEF and PH domain containing 6	219901_at	1.5E-13	1.42	0.0717	2.8E-3	1.01	6.1E-4	-1.1
LAMC2	laminin, gamma 2	202267_at	2.9E-13	2.76	**0.0097**	6.3E-4	1.87	1.5E-6	-2.21
KRT17	keratin 17	212236_x_at	4.5E-13	2.49	0.0122	2.9E-2	1.16	1.1E-3	-1.64
CSTB	cystatin B (stefin B)	201201_at	8.4E-13	1.21	0.0831	2.1E-4	1.35	9.4E-5	-1.18
SERPINA1	serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 1	211429_s_at	2.3E-12	2.31	0.9265	1.6E-2	1.32	5.5E-3	-1.74
UBD	ubiquitin D	205890_s_at	2.6E-12	2.9	0.2384	1.5E-2	1.01	1.6E-2	-1.23
CXCL5	chemokine (C-X-C motif) ligand 5	214974_x_at	3.5E-12	3.73	0.0835	1.2E-1	1.04	2.2E-3	-2.68
KRT19	keratin 19	201650_at	7.6E-12	3.7	0.0131	7.2E-2	1.05	1.2E-4	-2.93
S100P	S100 calcium binding protein P	204351_at	8.9E-12	3.66	0.6669	4.1E-6	3.12	2.5E-4	-2.08
LGALS1	lectin, galactoside-binding, soluble, 1 (galectin 1)	201105_at	1.9E-11	2.21	0.1631	3.0E-3	1.06	7.6E-1	-0.09
INPP4B	inositol polyphosphate-4-phosphatase, type II, 105kDa	205376_at	2.0E-11	1.62	0.0568	1.0E-5	1.91	1.3E-8	-2.46
G1P2	Interferon, alpha-inducible protein (clone IFI-15K)	205483_s_at	3.0E-11	1.86	0.0575	6.7E-3	1.13	8.2E-2	-0.66
ENO2	enolase 2 (gamma, neuronal)	201313_at	5.9E-11	1.61	0.016	2.4E-3	1.02	1.8E-1	-0.45
TMSB10	thymosin, beta 10	217733_s_at	1.1E-10	1.42	0.0299	2.5E-7	1.38	1.7E-5	-0.92
SFN	Stratifin	33323_r_at	1.3E-10	2.68	**0.0085**	3.5E-6	1.36	8.0E-4	-1.73
QPCT	glutaminyl-peptide cyclotransferase (glutaminyl cyclase)	205174_s_at	1.4E-10	1.55	0.7758	9.2E-4	1.02	6.6E-2	-0.56
PFKP	phosphofructokinase, platelet	201037_at	2.1E-10	1.52	**0.0094**	2.6E-3	1.07	1.9E-4	-1.19
HK2	hexokinase 2	202934_at	2.2E-10	1.52	0.012	1.5E-4	1.21	1.7E-2	-0.83
CKLF	chemokine-like factor	219161_s_at	2.5E-10	1.52	0.2197	5.2E-5	1.41	1.6E-7	-1.79
GPX1	glutathione peroxidase 1	200736_s_at	5.9E-10	1.25	0.9274	5.0E-3	1.25	7.0E-3	-0.95
IFI27	Interferon, alpha-inducible protein 27	202411_at	7.0E-10	2.16	0.3499	6.3E-5	1.91	1.3E-4	-1.56
TE_p_epithelial									
ANXA2	annexin A2	210427_x_at	2.9E-15	1.31	**0.0073**	1.8E-3	0.7	1.2E-5	-0.82
MAP4K4	mitogen-activated protein kinase kinase kinase kinase 4	206571_s_at	1.0E-13	1.35	**0.0079**	3.9E-3	0.89	1.3E-6	-1.42
DDEF2	development and differentiation enhancing factor 2	206414_s_at	2.7E-13	1.22	0.0483	6.5E-3	0.76	1.0E-4	-1.11
OACT2	O-acyltransferase (membrane bound) domain containing 2	226726_at	4.3E-13	1.52	**0.0025**	4.6E-6	1	5.9E-4	-0.59
BZRP	benzodiazapine receptor (peripheral)	202096_s_at	9.9E-13	1.01	0.2967	8.7E-3	0.95	9.7E-6	-1.23
S100A6	S100 calcium binding protein A6 (calcyclin)	217728_at	9.3E-12	2.46	0.0928	7.2E-4	0.95	1.6E-6	-1.22
CKLF	chemokine-like factor	223451_s_at	2.0E-11	1.6	0.2498	6.8E-3	0.98	1.3E-4	-1.39
CKAP1	cytoskeleton associated protein 1	216194_s_at	2.2E-11	1.04	0.129	3.6E-3	0.89	3.7E-3	-0.62
S100A10	S100 calcium binding protein A10 (annexin II ligand, calpactin I, light polypeptide (p11))	200872_at	3.2E-11	1.52	0.3762	2.7E-4	0.7	4.6E-5	-0.69
NCK1	NCK adaptor protein 1 /// NCK adaptor protein 1	211063_s_at	4.9E-11	1.01	0.318	1.5E-3	0.83	9.6E-3	-0.78
FXYD5	FXYD domain containing ion transport regulator 5	224252_s_at	1.0E-10	1.63	0.1072	9.1E-4	0.74	9.8E-4	-0.72
PON2	paraoxonase 2	201876_at	2.9E-10	1.05	0.0696	1.1E-4	0.84	6.4E-3	-0.9
TE_fc_stromal									
COL10A1	collagen, type X, alpha 1(Schmid metaphyseal chondrodysplasia)	217428_s_at	0.0E+0	5	0.9609	1.3E-3	1.86	1.1E-2	1.6
SULF1	sulfatase 1	212353_at	0.0E+0	4.63	0.8061	2.3E-2	1.2	2.1E-2	1.14
COL8A1	Collagen, type VIII, alpha 1	226237_at	0.0E+0	4.6	0.6288	3.5E-2	1.17	1.2E-5	2.61
COL11A1	collagen, type XI, alpha 1	37892_at	2.2E-16	4.4	0.772	1.2E-4	2.08	1.9E-3	2.13
THBS2	thrombospondin 2	203083_at	2.2E-16	3.97	0.8929	7.0E-3	1.03	3.4E-5	1.72
FN1	fibronectin 1	212464_s_at	2.2E-16	3.69	0.2846	1.5E-5	1.84	1.3E-1	0.66
---	---	238617_at	2.2E-16	2.87	0.598	1.4E-3	1.11	2.4E-3	1.14
CSPG2	chondroitin sulfate proteoglycan 2 (versican)	204619_s_at	1.3E-15	3.87	0.9914	2.5E-3	1.52	3.4E-3	1.24
COL5A2	collagen, type V, alpha 2	221729_at	5.8E-15	3.38	0.8486	1.4E-3	1.4	1.2E-3	1.37
CTHRC1	collagen triple helix repeat containing 1	225681_at	7.3E-15	4.38	0.9419	1.3E-3	1.66	2.4E-4	1.4
FNDC1	fibronectin type III domain containing 1	226930_at	1.3E-14	2.71	0.2714	1.4E-3	1.08	7.0E-5	2.14
GJB2	gap junction protein, beta 2, 26kDa (connexin 26)	223278_at	2.0E-14	3.67	0.0301	6.9E-5	1.84	3.0E-1	0.5
COL5A1	collagen, type V, alpha 1	212489_at	2.5E-14	3.42	0.9691	5.2E-3	1.12	1.2E-4	1.59
COL1A2	collagen, type I, alpha 2	202404_s_at	2.7E-14	4.17	0.6707	1.9E-4	2.17	7.0E-3	0.86
MXRA5	matrix-remodelling associated 5	209596_at	3.0E-14	2.34	0.5318	3.9E-4	1.2	1.7E-2	0.77
COL5A2	collagen, type V, alpha 2	221730_at	3.9E-14	3.69	0.8346	1.7E-3	1.09	4.0E-3	1.64
POSTN	periostin, osteoblast specific factor	210809_s_at	7.3E-14	5.13	0.3573	1.2E-5	3.29	1.2E-1	0.87
STEAP1	six transmembrane epithelial antigen of the prostate 1	205542_at	8.3E-13	2.17	0.4539	2.2E-5	1.19	3.2E-1	0.31
COL3A1	collagen, type III, alpha 1 (Ehlers-Danlos syndrome type IV, autosomal dominant)	211161_s_at	1.9E-12	3.23	0.6954	9.7E-3	1.02	6.2E-3	0.87
PLAU	plasminogen activator, urokinase	205479_s_at	2.4E-12	2.04	0.0967	1.4E-3	1.44	7.3E-1	0.15
IGFBP3	insulin-like growth factor binding protein 3	210095_s_at	2.6E-12	3.01	0.9378	4.4E-2	1.13	2.8E-3	1.48
CD109	CD109 antigen (Gov platelet alloantigens)	226545_at	4.9E-12	2.35	0.0133	3.8E-5	1.98	6.3E-1	0.26
SEMA3C	sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3C	203789_s_at	3.0E-11	3.01	0.6771	1.3E-3	1.47	1.7E-2	1.37
LEF1	lymphoid enhancer-binding factor 1	221558_s_at	1.9E-10	2.73	0.5167	1.2E-4	1.18	1.9E-2	0.88
ANTXR2	Anthrax toxin receptor 2	228573_at	3.7E-10	1.47	0.5301	2.4E-3	1.04	8.0E-1	0.07
SULF2	sulfatase 2	224724_at	4.0E-10	2.44	0.8042	1.6E-4	1.24	3.9E-1	0.27

cantly from the gene expression profiles of the other normal samples, being closer to the tumor samples in terms of gene expression. Histopathologic analysis revealed a significant fibrosis and/or tumoral infiltration in these samples, which makes them difficult to discriminate from the strong desmoplastic reaction of the tumor tissue. The samples in question are listed in **Supplementary Table 6** and their expression profiles can be seen in **Figure 2** and **Supplementary Figure 2**. The RT-PCR measurements confirm the microarray measurements for these anomalous samples, supporting our interpretation.

For a more comprehensive biological interpretation of the set of genes overexpressed in PDAC, we looked for enrichment in known biological annotations using the L2L Microarray Analysis Tool (37). The set was enriched in TGF-beta target genes ("tgfbeta_all_up" with p-value 4.31×10^{-29}, "tgfbeta_early_up" with p-value 3.18×10^{-27}), as well as in genes expressed in fibroblast or the epithelial-mesenchymal transition (**Supplementary Table 7A**). We also observed enrichment in Gene Ontology (40) Biological Process annotations such as: cell adhesion, response to wounding, integrin-mediated signaling, transforming growth factor beta receptor signaling, cell migration, focal adhesion formation, inflammatory and defense responses (see also **Supplementary Table 7B**). All of this is in line with the observed phenotype of PDAC, which involves an overproliferation of the extracellular matrix (fibrosis, "desmoplastic reaction") and inflammation, supporting the view of cancer as an abnormal response to wounding.

Using the Transcriptional Regulatory Element Database TRED (38), we found that many of the genes

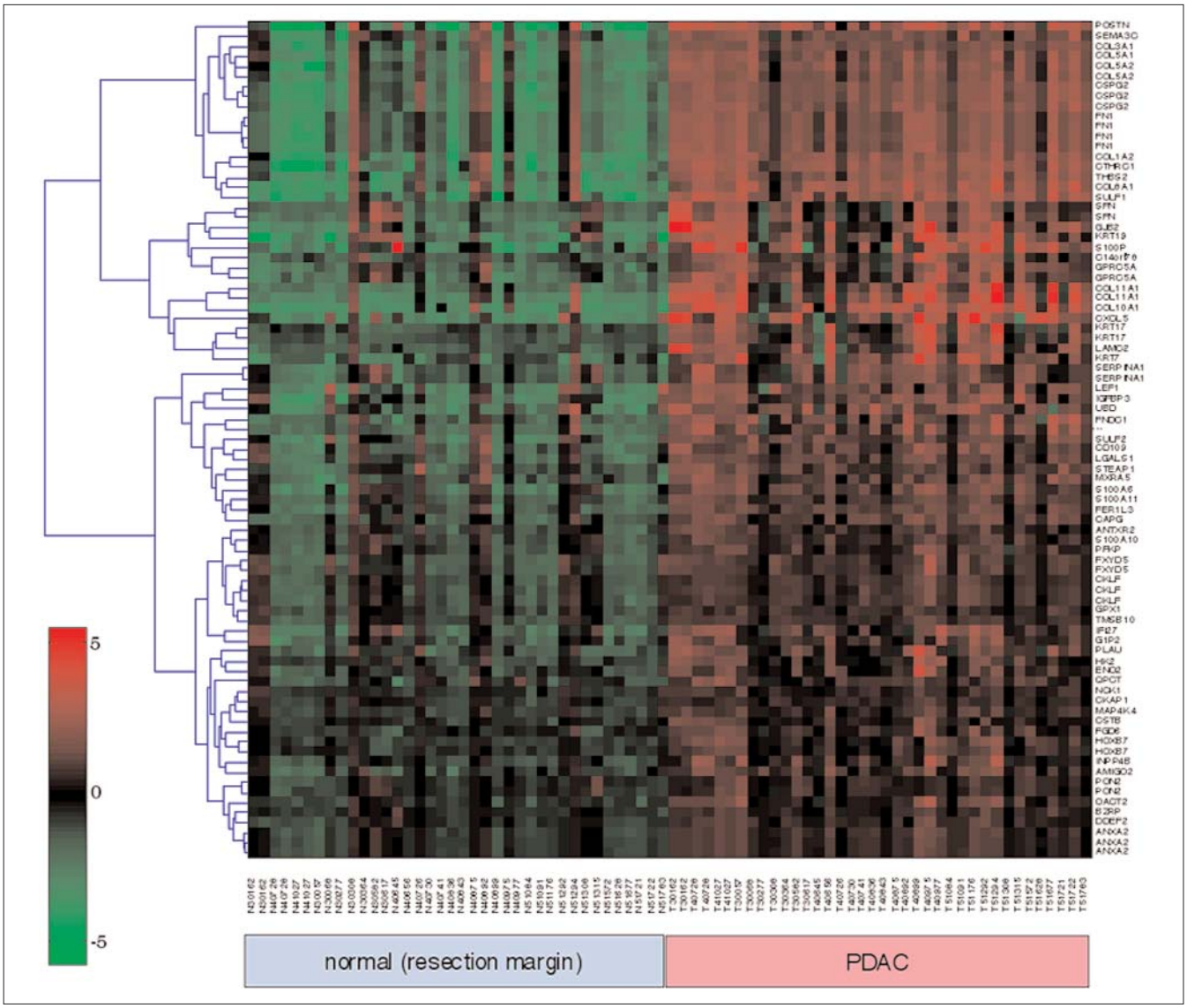

FIGURE 2 Centred log2-expression values of the genes overexpressed in tumor epithelia in our whole-tissue study (ICF). Note the anomalous "normal samples" (red columns in the green area)

in our set are controlled by combinations of large numbers of transcription factors including: SP1, AP2, AP1, JUN, E2F, SMAD, CEBP, etc. (Many direct and indirect TGF-beta targets are controlled by combinations of these factors.)

The preliminary analysis of the genes over-expressed in PDAC emphasizes the essential role of the TGF-beta pathway in PDAC (41). Remarkably, the TGF-beta pathway links the two observed phenotypes of PDAC: fibrosis/extracellular matrix proliferation and its aggressive metastatic potential, the latter being due to the fact that TGF-beta controls the so-called epithelial-mesenchymal transition (EMT) and induces an enhanced migratory ability of the cells. TGF-beta family cytokines are normally produced both in the tumor cells proper and in the stromal microenvironment. The TGF-beta superfamily cytokine most highly over-expressed in our data is INHBA. Interestingly, INHBA upregulation is accompanied by the over-expression of its receptor, ACVR1, which is consistent with a previous report mentioning the co-expression of INHBA and ACVR1 in pancreatic cancer (42).

Since the bulk of PDAC tumor volume consists of non-neoplastic fibroblastic, vascular, and inflammatory cells surrounded by huge quantities of extracellular matrix (33,34), it is very likely that the observed TGF-beta signature is mainly of stromal origin. As can be seen from **Table 1** and **Supplementary Table 1**, the genes involved in the TGF-beta pathway or affected by it form a significant fraction of the genes overexpressed in PDAC whole tissue tumors. A precise identification of the subset of genes specifically over-expressed in the neoplastic epithelia is therefore essential.

Identification of genes specifically over-expressed in tumor epithelia by correlation with a published microdissection study

Normal and tumor whole tissue samples have significantly different cell populations. While normal samples consist mainly of acinar and ductal epithelia, as well as endocrine islets, the PDAC tumors contain neoplastic epithelia, but also a hyper-proliferating tumor stroma, comprising carcinoma associated myofibroblasts, pancreatic stellate cells, vascular cells (endothelial cells and pericytes), as well as various inflammatory/immune cells, etc. (33). Gene expression changes between tissues of such different compositions of cell types may not necessarily be due to genes over-expressed in the neoplastic epithelia, but could also originate from genes specifically expressed in a different cell type (e.g. in the tumor-associated stroma).

Laser microdissection has previously been used for isolating the gross histological structures of PDAC, thereby allowing a more specific comparison of gene expression levels [e.g. (10,35)]. Unfortunately, it appears that the quality of microarray measurements on microdissected samples is significantly lower than that for whole tissue samples, due to the smaller sample quantities. (More precisely, **Supplementary Figure 1** shows that the relative standard deviations of the Affymetrix control probesets are about two times larger in the GP microdissection study than in our study. The experimental noise is therefore significantly higher in the microdissection study than in our whole tissue study.)

Both the whole tissue (ICF) and the microdissection study (GP) thus seem to have their advantages and drawbacks. The whole tissue study has a larger number of samples and higher quality, which enables a more reliable determination of the genes that are differentially expressed in PDAC whole tissue tumors. Unfortunately, many of these genes may be due to the stromal over-proliferation in PDAC tumors rather than the epithelial tumor cells proper. Although the microdissection study is less reliable due to its higher experimental noise and smaller number of samples, it proves essential for indicating whether the genes found overexpressed in the ICF data are upregulated in the neoplastic epithelia, the tumor stroma, or both.

By combining the two studies we can therefore reliably identify the genes that are highly and specifically over-expressed in neoplastic epithelia.

In the following, we make the simplifying assumption that the normal whole tissue samples consist mainly of normal epithelia (NE), while the tumor samples contain a mixture of tumor epithelia (TE) and tumor stroma (TS), the latter being typically more abundant than the former (TS >> TE). Therefore only the genes that are highly (or very consistently) overexpressed in TE (vs. NE) will be found overexpressed in whole tissue tumors, which have a predominant stromal compartment. According to this assumption, the genes underexpressed in TE (vs. NE) would be expected to be masked in the whole tissue tumor sample by the much more abundant tumor stroma. This is exactly what we observed in the ICF study, where we found virtually no genes underexpressed in whole tissue tumors (vs. the normal samples).

To determine the set of genes overexpressed in whole tumor tissues that are specifically upregulated in the tumor epithelia (rather than just the tumor stroma), we considered the genes that are either highly (*TE_fc_epithelial*) or consistently (*TE_p_epithelial*) overexpressed[4] in TE vs. NE, while also being more highly expressed in the TE than in the TS (this latter condition ensures that these genes are not masked by the TS in the whole tissue tumor samples). We also determined the genes that are highly over-expressed in TE vs. NE, but have an even higher expression in the TS (*TE_fc_stromal*), for which the increase observed in the whole tissue tumor samples (ICF) is predominantly due to their stromal component. **Table 2** displays the three above-mentioned lists of genes, while **Figures 1** and **2** show their expression levels in the microdissection (GP) and whole tissue (ICF) study respectively. (Larger versions of the Figures can be found in the Supplementary Material.)

[4]See the Methodology section for a precise definition of gene sets *TE_fc_epithelial* and *TE_p_epithelial*.

It is very interesting to note that 5 of the 9 genes (marginally) correlated with survival in the ICF data (mentioned in the previous Section, **Table 1** and **Supplementary Table 1**) are specifically over-expressed in tumor epithelia: keratin 7 (KRT7) and laminin gamma 2 (LAMC2) in *TE_fc_epithelial*, as well as annexin A2 (ANXA2), MAP4K4 and OACT2 (MBOAT2) in *TE_p_epithelial*. Two additional genes, stratifin (SFN) and platelet phosphofructokinase (PFKP) appear in *TE_fc_epithelial* as correlated with survival due to the lower p-value thresholds used for the ICF data when combining the two datasets. Moreover, all of the remaining 4 genes (correlated with survival in ICF but not passing the thresholds used for the GP data) ANXA2P2, SLC16A3, TM4SF1 and KYNU have higher expression in TE than TS. None of the genes found correlated with survival had higher expression values in the TS than the TE. *This epithelial localization of the over-expressed genes found correlated with survival is consistent with a model in which the patient survival ultimately depends on the gene expression profile of the neoplastic epithelia.*

Supplementary Figure 2 shows the Kaplan-Meier curves for the patients with high levels of expression of the differentially expressed genes found marginally correlated with survival compared to those for all patients.

Supplementary Table 4 presents a comparison of our list of genes (found over-expressed in tumor epithelia) with other microarray studies of pancreatic cancer.

We also performed an extensive literature search for mentions of these genes in connection with pancreatic cancer or other cancers. For brevity, **Supplementary Table 7C** lists a single mention for each gene (with preference for pancreatic cancer mentions). Of the 64 unique named genes, at least 26 are known to be involved in pancreatic cancer and 16 in other cancers (42 in total).

For an initial biological interpretation of this set of genes, we looked for enrichment in known biological annotations using the L2L Microarray Analysis Tool (34). The set specifically over-expressed in epithelia (*TE_fc_epithelial* + *TE_p_epithelial*) was modestly enriched in genes over-expressed in ER-negative breast cancer tumors (*"brca_er_neg"* with p-value 6.22×10^{-8}), while TE_fc_stromal was enriched, as expected, in TGF-beta target genes (*"tgfbeta_early_up"* with p-value 2.13×10^{-14} and *"tgfbeta_all_up"* with p-value 2.69×10^{-14}), thereby confirming the primary role of the TGF-beta pathway in stromal cells.

DISCUSSION

By combining our whole-tissue study of PDAC with the Grützmann-Pilarsky microdissection study, we have been able to identify a robust list of genes specifically upregulated in the neoplastic epithelia of PDAC. The overlap of this list of genes with other microarray studies of pancreatic cancer as well as with the published literature is very significant. Moreover, we have observed a number of genes whose expression shows a (marginal) inverse correlation with patient survival - it is remarkable that all of these are specifically upregulated in the neoplastic epithelia, rather than the tumor stroma.

On the other hand, the main drawback of this approach is that the whole tissue study will not be usable for confirming the genes that are found to be down-regulated in the slightly less reliable microdissection study. Note however, that the whole-tissue study may also identify genes over-expressed in cell types that are too hard to isolate even with laser microdissection, although it will not be able to determine the corresponding cell types.

A detailed description of the known biological functions of the genes over-expressed in PDAC tumor epithelia is beyond the scope of this paper. Briefly, we note the over-expression of several genes known to be involved in the epithelial-mesenchymal transition (EMT), namely LAMC2, HOXB7, MAP4K4, FN1, PLAU, POSTN and LEF1. In our study, laminin gamma 2 (LAMC2) expression was also inversely correlated with patient survival. This is consistent with the observation that "the level of circulating LN gamma2 NH(2)-terminal fragment (G2F) is a new, prognostic, tumor-characterizing marker for estimating the invasiveness and malignancy of epithelial carcinomas" (43).

The involvement of the developmental transcription factor HOXB7 is intriguing, especially since it has been previously linked to breast cancer epithelial-mesenchymal transition (EMT) (44). MAP4K4 has been found to be a promigratory kinase (45).

The transcription factor LEF1 is known to cooperate with beta-catenin for inducing EMT. Interestingly, LEF1 has been observed to induce EMT during mouse palate development even in the absence of beta-catenin (but in the presence of Smad factors) (46) thereby suggesting a potential link between the TGF-beta pathway and the EMT observed in PDAC (LEF1 is overexpressed both in tumor epithelia and stroma).

Up-regulation of the actin-capping protein CAPG in pancreatic cancer and its ability to modulate cell motility in vitro suggest its potentially important role in pancreatic cancer cell motility and consequently dissemination (47).

The G protein-coupling receptor GPRC5A has been suggested to play a role in tumor growth and as well as in embryonic development and epithelial cell differentiation (48).

We found a number of S100 family proteins to be upregulated in PDAC epithelia. S100P, the most frequently encountered PDAC marker in microarray studies (see **Supplementary Tables 3** and **4**) is known to promote pancreatic cancer growth, survival, and invasion (49). S100A11 plays a dual role in growth regulation of epithelial cells (50). S100A10 acts together with annexin A2 (ANXA2) as a plasminogen receptor, regulates plasmin production and thereby tumor growth and invasion (51). Moreover, AHNAK

interaction with the annexin 2/S100A10 complex regulates cell membrane cytoarchitecture, for example in case of cellular wounding (52). (C14orf78 is a synonym of AHNAK2.) It therefore appears that the various components of an insufficiently known biological process are simultaneously upregulated in PDAC.

Yet another S100 family member, S100A6 is an early event in pancreatic cancer development (53).

CXCL5, an inflammatory chemokine, was found overexpressed in colorectal cancer (54), as well as in late stage gastric cancer, being involved in its progression and in lymph node metastasis (55).

Galectin-1 (LGALS1) is known to be an inductor of pancreatic stellate cell activation (56) and to play a role in tumor immune privilege (57).

Dysadherin (FXYD5) expression seems to reflect tumor aggressiveness and to be a positive marker of poor prognosis (58). Dysadherin is able to modulate actin structures, stimulate cell motility, and contribute directly to the metastatic potential of human pancreatic cancer cells (59).

Among the genes overexpressed both in the tumor epithelia and the stroma, we noted several extracellular matrix proteins such as collagens (COL10A1, COL8A1, COL11A1, COL5A2, COL5A1, COL1A2, COL3A1) and fibronectin (FN1). Colorectal carcinogenesis was observed to be associated with stromal expression of both COL11A1 and COL5A2 (60).

Thrombospondin 2 (THBS2), a modulator of fibrogenesis, may be produced by Pancreatic Stellate Cells (PSCs) in response to injury and play a role in the development of pancreatic fibrosis that characterizes chronic pancreatitis (61) and PDAC.

Periostin (POSTN), over-expressed both in the tumor epithelia and stroma, creates a tumor-supportive microenvironment in the pancreas by sustaining fibrogenic stellate cell activity. Once stimulated by cancer cells, PSCs remain active via an autocrine periostin loop even under radiotherapy and produce excessive extracellular matrix proteins, creating a tumor-supportive microenvironment (62). POSTN promotes the invasiveness of tumour cells by increasing the motility of cells and enhances the survival of tumour cells exposed to hypoxic conditions (63). An active role for periostin in the epithelial-mesenchymal transformation and metastasis has been suggested (64). POSTN over-expression has also been cited to promote metastatic growth of colon cancer by augmenting cell survival via the Akt/PKB pathway, as well as by enhancing invasion and angiogenesis (65).

SULF1 is known to regulate growth and invasion of pancreatic cancer cells by interfering with heparin-binding growth factor signalling. Consistently with our data, pancreatic cancer samples analyzed in (66) expressed significantly (22.5-fold) increased SULF1 mRNA levels compared to normal controls, and SULF1 mRNA was localized in the cancer cells themselves as well as in peritumoral fibroblasts. (This confirms our classification of SULF1 in *TE_fc_stromal*.)

The large number of high-throughput studies of PDAC have identified most of the genes involved in pancreatic oncogenesis and cancer progression. Some of these genes have been the subject of intense experimental investigations, others are less well studied and await future more in-depth analysis. It is somewhat surprising that although very few (2-3) gene perturbations are enough to trigger PDAC in experimental animals (3,4), pancreatic cancer seems to involve changes in the expression levels of hundreds of genes, a large number of signaling pathways and surprisingly many distinct biological processes which act in a synergistic manner. Thus, it is likely that knowledge of the triggering events will not be enough to revert the transformed cells to a normal phenotype. Future work may therefore need to shift from the discovery of individual PDAC-related genes to the elucidation of the complex web of relationships and interactions between these, as well as to mapping these genes to higher level biological processes and to the observed phenotypes.

The list of carefully selected genes identified in this work using two complementary microarray studies may be a good starting point for such more in-depth investigations.

Acknowledgments

This study was supported by the Research of Excellence Project GENOPACT (CEEX 56/2005). We are deeply grateful to the authors of the microdissection studies (10,35) for making their data publicly available, as well as to Stefan Constantinescu and Victor Velculescu for the very useful comments and suggestions for improving the paper. L.B. acknowledges the help of Doina Tilivea and Anca Hotaran for the comparison with other microarray studies.

Supplementary information

Supplementary information can be found online at www.ai.ici.ro/HGE08/

References

1 **Hezel AF, Kimmelman AC, Stanger BZ, Bardeesy N, DePinho RA:** Genetics and biology of pancreatic ductal adenocarcinoma. Genes Dev 2006; 20:1218-1249.

2 **Hermann PC, Huber SL, Herrler T, Aicher A, Ellwart JW, Guba M et al.:** Distinct populations of cancer stem cells determine tumor growth and metastatic activity in human pancreatic cancer. Cell Stem Cell 2007; 1(3):313-323.

3 **Bardeesy N, Cheng KH, Berger JH, Chu GC, Pahler J, Olson P et al.:** Smad4 is dispensable for normal pancreas development yet critical in progression and tumor biology of pancreas cancer. Genes Dev 2006; 20(22):3130-3146.

4 **Ijichi H, Chytil A, Gorska AE, Aakare ME, Fujitani Y, Fujitani S et al.:** Aggressive pancreatic ductal adenocarcinoma in mice caused by pancreas-specific blockade of transforming growth factor-beta signaling in cooperation with active Kras expression. Genes Dev 2006; 20(22):3147-3160.

5 **Van Heek NT, Maitra A, Koopmann J, Iacobuzio-Donahue CA, Yeo CJ, Cameron JL et al.:** Gene expression profiling identifies markers of ampullary adenocarcinoma. Cancer Biol Ther 2004; 3(7):651-656.

6 **Segara D, Biankin AV, Kench JG, Henshall SM:** Expression of HOXB2, a retinoic acid signaling target in pancreatic

cancer and pancreatic intraepithelial neoplasia. Clin Cancer Res 2005; 11(9):3587-3596.

7 **Nakamura T, Furukawa Y, Nakagawa H, Tsunoda T, Katagiri T:** Genome-wide cDNA microarray analysis of gene expression profiles in pancreatic cancers using populations of tumor cells and normal ductal epithelial cells selected for purity by laser microdissection. Oncogene 2004; 23(13):2385-2400.

8 **Maitra A, Hansel DE, Argani P, Ashfag R, Yeo CJ:** Global expression analysis of well-differentiated pancreatic endocrine neoplasms using oligonucleotide microarrays. Clin Cancer Res 2003; 9(16 Pt 1):5988-5995.

9. **Lu Z, Hu L, Evers S, Chen J, Shen Y:** Differential expression profiling of human pancreatic adenocarcinoma and healthy pancreatic tissue. Proteomics 2004; 4(12):3975-3988.

10 **Grutzmann R, Pilarsky C, Ammerpohl O, Sager HD.** Gene expression profiling of microdissected pancreatic ductal carcinomas using high-density DNA microarrays. Neoplasia 2004; 6(5):611-622.

11 **Friess H, Ding J, Kleeff J, Fenkell L, Rosinski JA, Guweidhi A et al.:** Microarray-based identification of differentially expressed growth and metastasis associated genes in pancreatic cancer (gene lists retrieved from Pancreatic Expression Database www.pancreasexpression.org). Cell Mol Life Sci 2003; 60(6):1180-1199.

12 **Crnogorac-Jurcevic T, Efthimiou E, Nielsen T, Loader J, Terris B, Stamp G et al.:** Expression profiling of microdissected pancreatic adenocarcinomas. Oncogene 2002; 21(29):4587-4594.

13 **Crnogorac-Jurcevic T, Efthimiou E, Capelli P, Blaveri E, Baron A, Terris B et al.:** Gene expression profiles of pancreatic cancer and stromal desmoplasia. Oncogene 2001; 20(50):7437-7446.

14 **Crnogorac-Jurcevic T, Gangeswaran R, Bhakta V, Capurso G, Latimore S, Akada M et al.:** Proteomic analysis of chronic pancreatitis and pancreatic adenocarcinoma. Gastroenterology 2005; 129(5):1454-1463.

15 **Buchholz M, Braun M, Heidenblut A, Kestler HA, Kloppel G, Schmiegel W et al.:** Transcriptome analysis of microdissected pancreatic intraepithelial neoplastic lesions. Oncogene 2005; 24(44):6626-6636.

16 **Friess H, Ding J, Kleeff J, Liao Q, Berberat PO, Hammer J et al.:** Identification of disease-specific genes in chronic pancreatitis using DNA array technology. Ann Surg 2001; 234(6):769-779.

17 **Han H, Bearss DJ, Browne LW, Calaluce R, Nagle RB, Von Hoff DD.** Identification of differentially expressed genes in pancreatic cancer cells using cDNA microarray. Cancer Res 2002; 62(10):2890-2896.

18 **Iacobuzio-Donahue CA, Maitra A, Shen-Ong GL, Van Heek NT, Cameron JL, Yeo CJ et al.:** Discovery of novel tumor markers of pancreatic cancer using global gene expression technology. Am J Pathol 2002; 160(4):1239-1249.

19 **Tan ZJ, Hu XG, Cao GS, Tang Y:** Analysis of gene expression profile of pancreatic carcinoma using cDNA microarray. World J Gastroenterol 2003; 9(4):818-823.

20 **Iacobuzio-Donahue CA, Maitra A, Olsen M, Van Heek NT, Yeo CJ, Cameron JL et al.:** Exploration of global gene expression patterns in pancreatic adenocarcinoma using cDNA microarrays. Am J Pathol 2003; 162(4):1151-1162.

21 **Logsdon CD, Simeone DM, Binkley C, Arumugam T, Hanash S:** Molecular profiling of pancreatic adenocarcinoma and chronic pancreatitis identifies multiple genes differentiatelly regulated in pancreatic cancer. Cancer Res 2003; 63(10): 2649-2657.

22 **Grutzmann R, Foerder M, Alldinger I, Staub E, Pilarsky C:** Gene expression profiles of microdissected pancreatic ductal adenocarcinoma. Wirchows Arch 2003; 443(4):508-517.

23 **Crnogorac-Jurcevic T, Missiaglia E, Blaveri E, Gangeswaran R, Jones M, Terris B et al.:** Molecular alterations in pancreatic carcinoma: expression profiling shows that dysregulated expression of S100 genes is highly prevalent. J Pathol 2003; 201(1):63-74.

24 **Friess H, Ding J, Kleeff J, Fenkell L, Rosinski JA, Guweidhi A et al.:** Microarray-based identification of differentially expressed growth and metastasis associated genes in pancreatic cancer (gene lists retrieved from meta-analysis by Brandt et al.). Cell Mol Life Sci 2003; 60(6):1180-1199.

25 **Iacobuzio-Donahue CA, Ashfag R, Maitra A, Adsay NV, Cameron JL, Yeo CJ et al.:** Highly expressed genes in pancreatic ductal adenocarcinomas: a comprehensive characterization and comparison of the transcription profiles obtained from three major technologies. Cancer Res 2003; 63(24):8614-8622.

26 **Terris B, Blaveri E, Crnogorac-Jurcevic T, Jones M, Missiaglia E, Sauvanet A et al.:** Characterization of gene expression profiles in intraductal papillary-mucinous tumors of the pancreas. Am J Pathol 2002; 160(5):1745-1754.

27 **Jin G, Hu XG, Ying K, Tang Y, Liu R, Mao YM:** Discovery and analysis of pancreatic adenocarcinoma genes using cDNA microarrays. World J Gastroenterol 2005; 11(41): 6543-6548.

28 **Grutzmann R, Boriss H, Ammerpohl O, Sager HD, Pilarsky C:** Meta-analysis of microarray data on pancreatic cancer defines a set of commonly dysregulated genes. Oncogene 2005; 24(32):5079-5088.

29 **Hustinx SR, Cao D, Maitra A, Sato N, Iacobuzio-Donahue CA, Cameron JL et al.:** Differentially expressed genes in pancreatic ductal adenocarcinomas identified through serial analysis of gene expression. Cancer Biol Ther 2004; 3(12):-1254-1261.

30 **Bashyam MD, Bair R, Kim YH, Wang P, Hernandez-Boussard T, Karikari CA et al.:** Array-based comparative genomic hybridization identifies localized DNA amplifications and homozygous deletions in pancreatic cancer. Neoplasia 2005; 7(6):556-562.

31 **Heidenblad M, Schoenmakers EF, Jonson T, Gorunova L, Veltman JA, Van Kessel AG et al.:** Genome-wide array-based comparative genomic hybridization reveals multiple amplification targets and novel homozygous deletions in pancreatic carcinoma cell lines. Cancer Res 2004; 64(9):3052-3059.

32 **Heidenblad M, Lindgren D, Veltman JA, Jonson T, Gorunova L, Hoglund M:** Microarray analysis reveal strong influence of DNA copy number alterations on the transcriptional patterns in pancreatic cancer: implications for the interpretation of genomic amplifications. Oncogene 2005; 24(10):1794-1801.

33 **Chu GC, Kimmelman AC, Hezel AF, DePinho RA:** Stromal biology of pancreatic cancer. J Cell Biochem 2007; 101(4): 887-907.

34 **Mahadevan D, Von Hoff DD:** Tumor-stroma interactions in pancreatic ductal adenocarcinoma. Mol Cancer Ther 2007; 6(4):1186-1197

35 **Pilarsky C, Ammerpohl O, Sipos B, Dahl E, Friess H, Grutzmann R:** Activation of Wnt signalling in stroma from pancreatic cancer identified by gene expression profiling. J Cell Mol Med 2008; Feb 24.

36 **Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP:** Summaries of Affimetrix GeneChip probe level data. Nucleic Acids Res 2003; 31(4):e15.

37 **L2L, L2L Microarray Analysis Tool:** http://depts.washington.edu/l2l/ .

38 **TRED.Transcriptional Regulatory Element Database:** http://rulai.cshl.edu/cgi-bin/TRED/tred.cgi?process=home

39 **Brandt R, Grutzmann R, Bauer A, Jesnowski R, Ringel J, Lohr M et al.:** DNA microarray analysis of pancreatic malignancies. Pancreatology 2004; 4(6):587-597.

40 **Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM et al.:** Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000; 25(1):25-29.

41 **Truty MJ, Urrutia R:** Basics of TGF-beta and pancreatic cancer. Pancreatology 2007; 7(5):423-435.

42 **Kleeff J, Ishiwata T, Friess H, Buchler MW, Korc M:** Concomitant over-expression of activin/inhibin beta subunits and their receptors in human pancreatic cancer. Int J Cancer 1998; 77(6):860-868.

43 **Katayama M, Sanzen N, Funakoshi A, Sekiguchi K:** Laminin

gamma2-chain fragment in the circulation: a prognostic indicator of epithelial tumor invasion. Cancer Res 2003; 63(1):222-229.

44 **Wu X, Chen H, Parker B, Rubin E, Sukumar S:** HOXB7, a homeodomain protein, is overexpressed in breast cancer and confers epithelial-mesenchymal transition. Cancer Res 2006; 66(19):9527-9534.

45 **Collins CS, Hong J, Sapinoso L, Zhou Y, Liu Z, Micklash K et al.:** A small interfering RNA screen for modulators of tumor cell motility identifies MAP4K4 as promigratory kinase. Proc Natl Acad Sci U S A 2006; 103(10):3775-3780.

46 **Nawshad A, Hay ED:** TGFbeta3 signaling activates transcription of the LEF1 gene to induce epithelial mesenchymal transformation during mouse palate development. J Cell Biol 2003; 163(6):1291-1301.

47 **Thompson CC, Ashcroft FJ, Patel S, Saraga G, Lemoine NR, Crnogorac-Jurcevic T et al:** Pancreatic cancer cells overexpress gelsolin family-capping proteins, which contribute to their cell motility. Gut 2007; 56(1):95-106.

48 **Wu Q, Ding W, Mirza A, Van Arsdale T, Liu S:** Integrative genomics revealed RAI3 is a cell growth-promoting gene and a novel P53 transcriptional target. J Biol Chem 2005; 280(13):12935-12943

49 **Arumugam T, Simeone DM, Van Golen K, Logsdon CD:** S100P promotes pancreatic cancer growth, survival and invasion. Clin Cancer Res 2005; 11(15):5356-5364.

50 **Sakaguchi M, Sonegawa H, Murata H, Kitazoe M, Huh NH:** S100A11, a dual mediator for growth regulation of human keratinocytes. Mol Biol Cell 2008; 19(1):78-85.

51. **Kwon M, MacLeod TJ, Zhang Y, Waisman DM:** S100A10, annexin A2 and annexin a2 heterotetramer as candidate plasminogen receptors. Front Biosci 2005; 10:300-325.

52 **Benaud C, Gentil BJ, Assard N, Court M, Garin J, Delphin C et al.:** AHNAK interaction with the annexin 2/S100A10 complex regulates cell membrane cytoarchitecture. J Cell Biol 2004; 164(1):133-144.

53 **Vimalachandran D, Greenhalf W, Thompson CC, Crnogorac-Jurcevic T, Lemoine NR, Neoptolemos JP et al.:** High nuclear S100A6 (Calcyclin) is significantly associated with poor survival in pancreatic cancer patients. Cancer Res 2005; 65(8):3218-3225.

54 **Dimberg J, Dienus O, Lofgren S, Hugander A, Wagsater D:** Expression and gene polymorphisms of the chemokine CXCL5 in colorectal cancer patients. Int J Oncol 2007; 31(1):97-102.

55 **Park JY, Park KH, Bang S, Kim MH, Lee JE, Song SY:** CXCL5 overexpression is associated with late stage gastric cancer. J Cancer Res Clin Oncol 2007; 133(11):835-840.

56. Fitzner B, Walzel H, Sparmann G, Emmrich J, Liebe S, Jaster R. Galectin-1 is an inductor of pancreatic stellate cell activation. Cell Signal 2005; 17(10):1240-1247.

57 **Jung EJ, Moon HG, Cho BI, Jeong CY, Lee JS, Park ST:** Galectin-1 expression in cancer-associated stromal cells correlates tumor invasivness and tumor progression in breast cancer. Int J Cancer 2007; 120(11):2331-2338.

58 **Shimamura T, Sakamoto M, Ino Y, Sato Y, Hirohashi S:** Dysadherin overexpression in pancreatic ductal adenocarcinoma reflects tumor aggressiveness: relationship to e-cadherin expression. J Clin Oncol 2003; 21(4):659-667.

59 **Shimamura T, Yasuda J, Ino Y, Hirohashi S:** Dysadherin expression facilitates cell motility and metastatic potential of human pancreatic cancer cells. Cancer Res 2004; 64(19): 6989-6995.

60 **Fischer H, Stenling R, Rubio C, Lindblom A:** Colorectal carcinogenesis is associated with stromal expression of COL11A1 and COL5A2. Carcinogenesis 2001; 22(6):875-887.

61 **Neuschwander-Tetri BA, Talkad V, Otis Stephen F:** Induced thrombospondin expression in the mouse pancreas during pancreatic injury. Int J Biochem Cell Biol 2006; 38(1):102-109.

62 **Erkan M, Kleeff J, Gorbachevski A, Reiser C, Mitkus T, Esposito I et al.:** Periostin creates a tumor-supportive microenvironment in the pancreas by sustaining fibrogenic stellate cell activity. Gastroenterology 2007; 132(4):1447-1464.

63 **Baril P, Gangeswaran R, Mahon PC, Caulee K, Kocher HM, Harada T et al.:** Periostin promotes invasiveness and resistance of pancreatic cancer cells to hypoxia-induced cell death: role of the beta4 integrin and the PI3k pathway. Oncogene 2007; 26(14):2082-2094.

64 **Yan W, Shao R:** Transduction of a mesenchyme-specific gene periostin into 293T cells induces cell invasive activity trough epithelial-mesenchymal transformation. J Biol Chem 2006; 281(28):19700-19708.

65 **Bao S, Ouyang G, Bai X, Huang Z, Ma C, Liu M et al.:** Periostin potently promotes metastatic growth of colon cancer by augmenting cell survival via the Akt/PKB pathway. Cancer Cell 2004; 5(4):329-339.

66 **Abiatari I, Kleeff J, Li J, Felix K, Buchler MW, Friess H:** Hsulf-1 regulates growth and invasion of pancreatic cancer cells. J Clin Pathol 2006; 59(10):1052-1058.