



**HAL**  
open science

**Integrative analysis of the cancer genome atlas and cancer cell lines encyclopedia large-scale genomic databases: MUC4/MUC16/MUC20 signature is associated with poor survival in human carcinomas.**

Nicolas Jonckheere, Isabelle van Seuningen

► **To cite this version:**

Nicolas Jonckheere, Isabelle van Seuningen. Integrative analysis of the cancer genome atlas and cancer cell lines encyclopedia large-scale genomic databases: MUC4/MUC16/MUC20 signature is associated with poor survival in human carcinomas.. *Journal of Translational Medicine*, 2018, 16 (1), pp.259. <10.1186/s12967-018-1632-2>. <hal-02341297>

**HAL Id: hal-02341297**

**<https://hal.science/hal-02341297v1>**

Submitted on 31 Oct 2019

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL Authorization

1 **Integrative analysis of The Cancer Genome Atlas and Cancer Cell Lines Encyclopedia**  
2 **large-scale genomic databases: MUC4/MUC16/MUC20 signature is associated with**  
3 **poor survival in human carcinomas**

4

5 Nicolas Jonckheere and Isabelle Van Seuningen

6 Univ. Lille, Inserm, CHU Lille, UMR-S 1172 - JPARC - Jean-Pierre Aubert Research Center,

7 Team "Mucins, epithelial differentiation and carcinogenesis", F-59000 Lille, France

8

9 **Correspondence:**

10 [nicolas.jonckheere@inserm.fr](mailto:nicolas.jonckheere@inserm.fr)

11 Phone: +33 3 20 29 88 65, Fax: +33 3 20 53 85 62

12 [isabelle.vanseuningen@inserm.fr](mailto:isabelle.vanseuningen@inserm.fr)

13 Phone : +33 3 20 29 88 67, Fax : +33 3 20 53 85 62

14

15 Competing interests: Authors declare no conflict of interest

16

17

18

19 Authors' Contributions NJ conceived and designed the analysis. NJ analyzed the data. NJ

20 and IVS wrote and edited the paper.

21

22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47

## Abstract

**Background:** MUC4 is a membrane-bound mucin that promotes carcinogenetic progression and is often proposed as a promising biomarker for various carcinomas. In this manuscript, we analyzed large scale genomic datasets in order to evaluate *MUC4* expression, identify genes that are correlated with *MUC4* and propose new signatures as a prognostic marker of epithelial cancers.

**Methods** Using cBioportal or SurvExpress tools, we studied *MUC4* expression in large-scale genomic public datasets of human cancer (The cancer genome atlas, TCGA) and Cancer Cell Line Encyclopedia (CCLE).

**Results:** We identified 187 co-expressed genes for which the expression is correlated with *MUC4* expression. Gene ontology analysis showed they are notably involved in cell adhesion, cell-cell junctions, glycosylation and cell signaling. In addition, we showed that *MUC4* expression is correlated with *MUC16* and *MUC20*, two other membrane-bound mucins. We showed that MUC4 expression is associated with a poorer overall survival in TCGA cancers with different localizations including pancreatic cancer, bladder cancer, colon cancer, lung adenocarcinoma, lung squamous adenocarcinoma, skin cancer and stomach cancer. We showed that the combination of *MUC4*, *MUC16* and *MUC20* signature is associated with statistically significant reduced overall survival and increased hazard ratio in pancreatic, colon and stomach cancer.

**Conclusions:** Altogether, this study provides the link between (i) MUC4 expression and clinical outcome in cancer and (ii) MUC4 expression and correlated genes involved in cell adhesion, cell-cell junctions, glycosylation and cell signaling. We propose the MUC4/MUC16/MUC20<sup>high</sup> signature as a marker of poor prognostic for pancreatic, colon and stomach cancers.

**Keywords:** MUC4, TCGA, CCLE, patient survival, biomarker

48

**Abbreviations:**

49

50 AUROC: Area Under Receiving Operator Characteristic

51 CCLE: Cancer Cell Line Encyclopedia

52 HR: Hazard ratio

53 PDAC: Pancreatic Ductal AdenoCarcinoma

54 ROC: Receiving operator characteristic

55 TCGA: The cancer genome atlas

56

57

58 **Background**

59

60 The cancer genome atlas (TCGA) was developed by National Cancer Institute (NCI) and  
61 National Human Genome Research Institute (NHGRI) in order to provide comprehensive  
62 mapping of the key genomic changes that occur during carcinogenesis. Datasets of more  
63 than 11,000 patients of 33 different types of tumors are publically available. In parallel,  
64 Cancer Cell Line Encyclopedia (CCLE), a large-scale genomic dataset of human cancer cell  
65 lines, was generated by the Broad Institute and Novartis in order to reflect the genomic  
66 diversity of human cancers and provide complete preclinical datasets for mutation, copy  
67 number variation and mRNA expression studies [1]. In order to analyse this kind of large  
68 scale datasets, several useful online tools have been created. cBioportal is an open-access  
69 database analysis tool developed at the Memorial Sloan-Kettering Cancer Centre (MSKCC)  
70 to analyze large-scale cancer genomics data sets [2, 3]. SurvExpress is another online tool  
71 for biomarker validation using 225 datasets available and therefore provide key information  
72 linking gene expression and the impact on cancer outcome [4].

73 Mucins are large high molecular weight glycoproteins that are classified in two sub groups: (i)  
74 the secreted mucins that are responsible of rheologic properties of mucus and (ii) the  
75 membrane-bound mucins that include MUC4, MUC16 and MUC20 [5, 6]. *MUC4* was first  
76 discovered in our laboratory 25 years ago from a tracheobronchial cDNA library [7]. MUC4 is  
77 characterized by a long hyper-glycosylated extracellular domain, Epidermal Growth Factor  
78 (EGF)-like domains, a hydrophobic transmembrane domain, and a short cytoplasmic tail.  
79 MUC4 also contains NIDO, AMOP and vWF-D domains [8]. A direct interaction between  
80 MUC4 and its membrane partner, the oncogenic receptor ErbB2, alters downstream  
81 signaling pathways [9]. MUC4 is expressed at the surface of epithelial cells from  
82 gastrointestinal and respiratory tracts [10] and has been studied in various cancers where it  
83 is generally overexpressed and described as an oncomucin and has been proposed as an  
84 attractive prognostic tumor biomarker. Its biological role has been mainly evaluated in

85 pancreatic, ovarian, esophagus and lung cancers [9, 11-14]. Other membrane-bound mucins  
86 MUC16 and MUC20 share some functional features but evolved from distinct ancestors [15].  
87 *MUC20* gene is located on the chromosomal region 3q29 close to *MUC4*. MUC16 also  
88 known as the CA125 antigen is a routinely used serum marker for the diagnosis of ovarian  
89 cancer [16]. Both mucins favor tumor aggressiveness and are associated with poor overall  
90 survival and could be proposed as prognosis factors [16-18].

91 In this manuscript, we have used the online tools cBioportal, DAVID6.8 and SurvExpress in  
92 order to (i) evaluate *MUC4* expression in various carcinomas, (ii) identify genes that are  
93 correlated with *MUC4* and evaluate their roles and (iii) propose *MUC4/MUC16/MUC20*  
94 combination as a prognostic marker of pancreatic, colon and stomach cancers.

95

## 96 **Material and methods**

97

### 98 **Expression analysis from public datasets**

99 *MUC4* z-score expressions were extracted from databases available at cBioPortal for Cancer  
100 Genomics [2, 3]. This portal stores expression data and clinical attributes. The z-score for  
101 *MUC4* mRNA expression is determined for each sample by comparing mRNA expression to  
102 the distribution in a reference population harboring typical expression for the gene. The query  
103 “MUC4” was realized in CCLE (881 samples, Broad Institute, Novartis Institutes for  
104 Biomedical Research) [1] and in all TCGA datasets available (13 489 human samples, TCGA  
105 Research Network (<http://cancergenome.nih.gov/>)). The mRNA expression from selected  
106 data was plotted in relation to the clinical attribute (tumor type and histology) in each sample.  
107 *MUC4* expression was analyzed in normal tissues by using the Genome Tissue Expression  
108 (GTEx) tool [19, 20]. Data were extracted from GTEx portal on 06/29/17 (dbGaP accession  
109 phs000424.v6.p1) using the 4585 Entrez gene ID.

110

### 111 **DAVID6.8 identification and gene ontology of genes correlated with MUC4**

112 We established a list of 187 genes that are correlated with *MUC4* expression in CCLE  
113 dataset out of 16208 genes analyzed with cBioportal tool on co-expression tab. These genes  
114 harbor a correlation with both Pearson’s and Spearman’s higher than 0.3 or lower than -0.3.  
115 Functional annotation and ontology clustering of the complete list of genes were performed  
116 using David Functional Annotation Tool (<https://david.ncifcrf.gov/>) and Homo sapiens  
117 background [21, 22]. Enrichment scores of ontology clusters are provided by the online tool.  
118 Interaction of proteins correlated with *MUC4* was determined using String 10 tool  
119 (<https://string-db.org/>) [23]. Edges represent protein-protein associations such as known  
120 interactions (from curated databases or experimentally determined), predicted interactions

MUC4 and TCGA and CCLE databases

121 (from gene neighborhood, gene fusion or co-occurrence), text-mining, co-expression or  
122 protein homology. The network was divided in 3 clusters based on k-means clustering.

123

#### 124 **Methylation and copy number analysis**

125 Using <https://portals.broadinstitute.org/ccle>, we extracted mRNA expression of MUC4,  
126 methylation score (Reduced Representation Bisulfite Sequencing: RRBS) and copy number  
127 variations of the genes of interest. The mRNA expression of MUC4 was plotted in relation to  
128 log2 copy number or RRBS score.

129

#### 130 **SurvExpress survival analysis**

131 Survival analysis was performed using the SurvExpress online tool available in  
132 bioinformatica.mty.itesm.mx/SurvExpress (Aguire Gamboa PLoS One 2013). We used the  
133 optimized algorithm that generates risk group by sorting prognostic index (higher value of  
134 MUC4 for higher risk) and split the two cohorts where the p-value is minimal. Hazard ratio  
135 [95% confidence interval (CI)] was also evaluated. The tool also provided a box plot of genes  
136 expression and the corresponding p value testing the differences.

137

#### 138 **Gene Expression Omnibus microarray**

139 GSE28735 and GSE16515 pancreatic cancer microarrays were analysed from the NCBI  
140 Gene Expression Omnibus (GEO) database (<http://www.ncbi.nlm.nih.gov/geo/>). GSE28735  
141 is a dataset containing 45 normal pancreas (adjacent non tumoral, ANT) and 45 tumor (T)  
142 tissues from PDAC cases. GSE16515 contains 52 samples (16 had both tumor and normal  
143 expression data, and 20 only had tumor data. Data were analysed using GEO2R software.  
144 The dataset GSE28735 used Affymetrix GeneChip Human Gene 1.0 ST array. The dataset

145 GSE16515 used the Affymetrix Human Genome U133 Plus 2.0 Array. GSE13507 contains  
146 165 bladder cancer and 58 ANT samples. GSE30219 contains 14 normal lung, 85  
147 adenocarcinomas and 61 squamous cancer samples. GSE40967 contains 566 colorectal  
148 cancers and 19 normal mucosae. GSE27342 contains 80 tumors and 80 paired ANT tissues.  
149 GSE4587 contains 2 normal, 2 melanomas and 2 metastatic melanomas. GSE14407  
150 contains 12 ovarian adenocarcinomas and 12 normal ovary samples.

151

## 152 **Statistical analysis**

153 For *MUC4* expression analysis, paired and unpaired t test statistical analyses were  
154 performed using the Graphpad Prism 6.0 software (Graphpad softwares Inc., La Jolla, CA,  
155 USA).  $P < 0.05$  was considered as statistically significant. Receiving operator characteristic  
156 (ROC) curves and areas under ROC (AUROC) were evaluated by comparing tumor and ANT  
157 values. cBioportal provided Pearson and Spearman tests were performed to analyze  
158 correlation of other genes, RRBS score and log<sub>2</sub> copy number with *MUC4* expression.  
159 DAVID tool provided p value of each ontology enrichment score. SurvExpress tool provided  
160 statistical analysis of hazard ratio and overall survival. A Log rank testing evaluated the  
161 equality of survival curves between the high and low risk groups.

162

163

164 **Results**

165

166 **MUC4 expression analysis in databases**

167 *MUC4* expression was analyzed from databases available at cBioPortal for Cancer  
168 Genomics [2, 3]. We queried for *MUC4* mRNA expression in the 881 samples from CCLE [1]  
169 (Figure 1). The oncoprint showed that *MUC4* was altered in 195 samples out of 881 (22%).  
170 188 were amplification (n=120) or mRNA upregulation (n=88) (Supplemental Figure 1).  
171 Results were sorted depending on the tumor type. We mainly observed an important z-score  
172 expression of *MUC4* in carcinoma samples (n= 538 samples, p=0.001) (Figure 2A). *MUC4*  
173 Expression scores were subsequently sorted depending of the organ (Figure 2B). As  
174 expected, pancreatic cancer cell lines harbor the highest *MUC4* expression (n=35, z-  
175 score=2.166, p=0,0006 against theoretical control median=0). Other cell lines from different  
176 tissues (lung NSC, esophagus, bile duct, stomach, upperdigestive, colorectal, ovary, and  
177 urinary tract) showed statistically significant alteration. We also performed a similar analysis  
178 on 13 489 human samples retrieved from TCGA by using the cBioportal platform. An  
179 important *MUC4* expression z-score was observed in bladder urothelial carcinoma, cervical  
180 squamous cell carcinoma/endocervical adenocarcinoma, colorectal carcinoma, esophageal  
181 carcinoma, head and neck squamous cell carcinoma, lung adenocarcinoma, lung squamous  
182 cell carcinoma, ovarian serous cystadenocarcinoma, pancreatic adenocarcinoma, prostate  
183 adenocarcinoma, stomach adenocarcinoma and uterine corpus endometrial carcinoma  
184 (Figure 3). Expression of *MUC4* in normal tissues was analyzed using the GTEX project tool,  
185 *MUC4* was expressed in lung, testis, small intestine, terminal, ileum, prostate, vagina, minor  
186 salivary gland and esophagus mucosa and transverse colon (supplemental figure2).  
187 Altogether, this shows that *MUC4* high expression is observed in carcinoma and notably in  
188 pancreatic cancer.

189

**MUC4 co-regulated genes**

Using the co-expression tool on expression data extracted from the 881 samples of CCLE [1], we obtained a list of genes that are co-expressed with *MUC4*. Genes that harbor a correlation with both Pearson's and Spearman's higher than 0.3 or lower than -0.3 were selected. 187 genes are positively (n=178) or negatively (n=9) correlated with *MUC4* expression. The better correlated genes were Adhesion G Protein-Coupled Receptor F1 (*ADGRF1*, Pearson's correlation=0.56) and Lipocalin2 (*LCN2*, Pearson's correlation=0.54) (table 1). We also observed that expression of other membrane bound mucins *MUC16* and *MUC20* are positively correlated with *MUC4*. Correlation between *MUC16* and *MUC20* was also observed (not shown). Only few genes were negatively correlated such as *ZEB1* transcription factor or *ST3 Beta-Galactoside Alpha-2,3-Sialyltransferase 2 (ST3GAL2)* (table 2).

Functional Annotation of the complete list of genes and ontology clustering were performed using David Functional Annotation Tool. The gene clustering analysis is presented in table 3. The complete gene ontologies that are statistically significant are provided in supplemental table 1. We observed the highest enrichment scores in gene clusters involved in cell adhesion (7.08) and tight junction (5.44) (table 3). Notably, we observed the correlation of expression of *MUC4* with genes encoding integrins (*ITGB4* and *ITGB6*) and cadherin-type proteins such as *CDH1*, *CDH3*, Desmocollin 2 (*DSC2*). A strong enrichment of 91 transmembrane proteins was observed including EPH Receptor A1 (*EPHA1*), Epithelial cell adhesion molecule (*EPCAM*), Carcinoembryonic Antigen Related Cell Adhesion Molecule-5 and -6 (*CEACAM5* and *CEACAM6*), C-X-C motif chemokine ligand 16 (*CXCL16*) and ATPase Secretory Pathway Ca<sup>2+</sup> Transporting 2 (*ATP2C2*). As *MUC4* is a glycoprotein, it is interesting to also note the correlated expression of enzymes involved in different steps of glycosylation such sialyltransferases (*ST3GAL2*, *ST6GALNAC1*), beta-1,3-N-acetylglucosaminyltransferases (*B3GNT5*, *B3GNT3*), fucosyltransferases (*FUT3*, *FUT2*), and UDP-GalNAc transferase (*GALNT3*). *MUC4* was also associated with genes associated with

217 cell signaling containing SH2 domain (Cbl proto-oncogene C (*CBLC*), signal transducing  
 218 adaptor family member 2 (*STAP2*), dual adaptor of phosphotyrosine and 3-phosphoinositides  
 219 1 (*DAPP1*), SH2 domain containing 3A (*SH2D3A*), protein tyrosine kinase 6 (*PTK6*), growth  
 220 factor receptor bound protein 7 (*GRB7*), fyn related Src family tyrosine kinase (*FRK*), tensin 4  
 221 (*TNS4*) or SH3 domains (MET transcriptional regulator (*MACC1*), Rho GTPase activating  
 222 protein 27 (*ARHGAP27*), tight junction protein 2 (*TJP2*), Rho guanine nucleotide exchange  
 223 factor-5 and -16 (*ARHGEF5*, *ARHGEF16*), protein tyrosine kinase 6 (*PTK6*), EPS8 like 1  
 224 (*EPS8L1*), tight junction protein 3 (*TJP3*) and *FRK*). Finally, several genes encoding proteins  
 225 with a SEA domain (*ADGRF1*, *ST14*, *MUC16*) were correlated with *MUC4* expression.  
 226 Additionally, we analyzed protein-protein interactions of differentially expressed proteins with  
 227 *MUC4* with the String 10 tool. We showed that *MUC4* is directly related with *CEACAM5*,  
 228 *CEACAM6*, *MUC16*, *MUC20* and glycosylation enzymes (*ST3GAL2*, *B3GNT3*, *B3GNT5* and  
 229 *GALNT3*) (Supplemental Figure 3). Altogether, we have identified genes with expression  
 230 correlated with *MUC4* involved notably in cell adhesion, cell-cell junctions, glycosylation and  
 231 cell signaling. In order to understand the association between the observed aberrant  
 232 expression of *MUC4* and other molecular events, we explored the correlation between *MUC4*  
 233 expression in CCLE and DNA methylation (RRBS) of the top genes correlated with *MUC4*.  
 234 We observed that *MUC4* expression is negatively correlated with the methylation score of 16  
 235 out of 20 of the top genes (*LCN2*, *MUC20*, *STEAP4*, *WFDC2*, *GJB3*, *SH2D3A*, *RNF39*,  
 236 *PRSS22*, *HS3ST1*, *GPR87*, *TACST2*, *FAM83A*, *LAMC2*, *B3GNT3*, *CLDN7*) (Figure 4)  
 237 suggesting that the association of *MUC4* and the correlated genes could be mediated by  
 238 methylation regulation. Only *ADGRF1* RBBS is not correlated with *MUC4* mRNA level.  
 239 *MUC16*, *SCEL* and *C1ORF116* scores were not available. Additionally we also evaluated the  
 240 copy number variation association of the top genes with *MUC4* expression. We only  
 241 observed a weak amplification of *MUC20* copy number (Pearson's correlation = 0.13) and a  
 242 weak deletion of *MUC16* copy number (Pearson's correlation = -0.14) suggesting that the  
 243 relationship between *MUC4* expression and copy number variation of top genes is unlikely  
 244 (Supplemental Figure 4).

245

246 **MUC4 and patient survival**

247 To establish a correlation between *MUC4* expression and patient survival, we have  
248 compared survival analysis and hazard ratio in population designated as MUC4 high risk and  
249 low risk in every organ from TCGA datasets (table 4). We have used SurvExpress optimized  
250 algorithm that generates risk group by sorting prognostic index (higher value of MUC4 for  
251 higher risk). The algorithm splits the population where the p-value testing the difference of  
252 MUC4 expression is minimal [4]. Pancreatic cancer presented the most important hazard  
253 ratio for MUC4 (HR= 3.94 [CI, 1.81-8.61] p=0.0005756) (Figure 5A). MUC4 high risk was  
254 also significantly associated with survival in bladder cancer (HR= 1.48), colon cancer (HR=  
255 2.1), lung adenocarcinoma (HR= 1.7), lung squamous carcinoma (HR= 1.69), ovarian cancer  
256 (HR= 1.33), skin cancer (HR= 1.87) and stomach cancer (HR= 1.58) (Figure 5A). Acute  
257 myeloid leukemia (HR= 1.59) and liver cancer (HR= 1.4) almost reach statistical significance.  
258 Other datasets did not show any statistically significant differences.

259 A significant reduction in patient's survival was observed in bladder cancer (p=0.01135),  
260 colon cancer (p=.00891), lung adenocarcinoma (p=0.008187), lung squamous carcinoma  
261 (p=0.03586), ovarian cancer (p=0.0186), pancreatic cancer (p=0.000219), skin cancer  
262 (p=0.02384) and stomach cancer (p=0.04751) as illustrated in Kaplan Meier curves (Figure  
263 5B). Strikingly, pancreatic median survival was 593 days in *MUC4*<sup>high</sup> cohort (n=149) whereas  
264 the 50% survival was not reached in *MUC4*<sup>low</sup> cohort (n=27). In lung squamous carcinoma,  
265 the median survival of *MUC4*<sup>high</sup> cohort (n=116) was 1067 days whereas *MUC4*<sup>low</sup> cohort  
266 (n=59) presented a 2170 days median survival. It is interesting to note that the algorithm  
267 splits the population in two parts that were characterized as the most different regarding  
268 *MUC4* expression. Therefore, there are a modest number of *MUC4*<sup>low</sup> PDAC or lung  
269 adenocarcinoma patients and a low number of *MUC4*<sup>high</sup> colon or stomach cancer patients. A  
270 similar survival analysis was performed on pancreatic cancer by dividing the patient

271 population in two equal parts (88 vs 88), *MUC4*<sup>high</sup> harbored a decreased survival that was  
272 close to statistical significance ( $p=0.06784$ ) (not shown). Therefore, *MUC4* expression is  
273 associated with a poorer overall survival in different cancers including pancreatic cancer.

274 We also compared the survival and hazard ratio, in the same cancers whose survival is  
275 associated with MUC4 (bladder cancer, colon cancer, lung adenocarcinoma, lung squamous  
276 carcinoma, ovarian cancer, pancreatic cancer, skin cancer and stomach cancer), according  
277 to gene signatures corresponding to the five first gene ontology term from supplemental table  
278 1 (GO 0031424: keratinization, GO 0007155: cell adhesion, GO 0019897: extrinsic  
279 component of plasma membrane, GO 0016323 : basolateral plasma membrane and GO  
280 0016324: apical plasma membrane) (Figure 6A, supplemental table 2). These gene  
281 signatures were all significantly associated with survival in the TCGA dataset tested. The  
282 “keratinization” (GO 0031424) and “cell adhesion” (GO 0007155) signature are associated  
283 with HR comprised between 1.65 and 3.76 and between 2.15 and 3.23, respectively. The GO  
284 0019897 signature is associated with weaker HR (1.55-2.30). “basolateral” (GO 0016323)  
285 and “apical plasma membrane” (GO 0016324) signatures harbors more increased HR (2.21-  
286 4.5 and 1.77-4.42, respectively) in these datasets.

287 We performed a similar analysis according to the top genes (*ADGRF1*, *LCN2*, *MUC20*,  
288 *C1ORF116*, *SCEL*, *STEAP4*) that harbored Pearson’s correlation with MUC4 superior to 0.5  
289 (Figure 6B, supplemental table 3). This signature is associated with survival in all TCGA  
290 dataset tested (HR comprised between 1.91 and 8.77). Notably, pancreatic cancer harbored  
291 the strongest association with survival according to this signature (HR=8.77 [CI, 2.15-35.83]).  
292 Overall, these bigger signatures harbored higher hazard ratio compared to MUC4 alone.

293

## 294 **MUC4, MUC16 and MUC20 signature in cancer**

295 Mucins have been proposed as potential biomarkers for carcinoma. Notably, previous work  
296 suggested that combination of mucins expression may be useful for early detection and

297 evaluation of malignancy of pancreatobiliary neoplasms [24]. Moreover, MUC16/CA125  
298 antigen is an already routinely used serum marker for the diagnosis of ovarian cancer [16].  
299 Therefore, we decided to intentionally focus on the two other membrane bound mucins  
300 *MUC16* and *MUC20* that were correlated with expression of *MUC4*. We analyzed the survival  
301 curves of the high risk group (*MUC4/MUC16/MUC20*<sup>high</sup>, n= 159) and low risk group  
302 (*MUC4/MUC16/MUC20*<sup>low</sup>, n=17) from the pancreas TCGA dataset. The  
303 *MUC4/MUC16/MUC20*<sup>high</sup> risk group was associated with an increased hazard ratio (HR=6.5  
304 [2.04-20.78], p=0.001582) and a shorter overall survival (p=0.0003088) (Figure 7A). Median  
305 survival was similar as in *MUC4*<sup>high</sup> cohort (593 days). The *MUC4/MUC16/MUC20*<sup>high</sup> group  
306 harbored a statistically significant increase of *MUC4*, *MUC16* and *MUC20* expression (Figure  
307 7B). We also analyzed overall survival in every other PDAC database available in  
308 Surexpress. We show that *MUC4*<sup>high</sup> group was associated with a statistically significant  
309 reduced overall survival and increased hazard ratio in both ICGC and Stratford (GSE21501)  
310 cohorts (Figure 7C). In Zhang cohort (GSE28735), *MUC4*<sup>high</sup> group was associated with a  
311 reduced overall survival that was close to statistical significance (p=0.08971). In other  
312 organs, the *MUC4/MUC16/MUC20*<sup>high</sup> group was associated with an increased hazard ratio  
313 and reduced overall survival in bladder cancer, colon cancer, lung adenocarcinoma, lung  
314 squamous adenocarcinoma, skin cancer, stomach cancer (supplemental figure 5A). Notably,  
315 the *MUC4/MUC16/MUC20*<sup>high</sup> group in colon cancer (HR=2.26 [1.51-3.4]) showed a median  
316 survival of 1741 days whereas the low risk group did not reach the 50% survival. Similarly,  
317 the *MUC4/MUC16/MUC20*<sup>high</sup> group in stomach cancer showed a median survival of 762  
318 days whereas the low risk had a median survival of 1811 days. No significant difference was  
319 observed for ovarian cancer (p=0.2081). Moreover, a reduced overall survival was observed  
320 in liver cancer (p=0.04789) and acute myeloid leukemia (AML) (p=0.02577) (supplemental  
321 Figure 5B) in which we did not show any statistical difference when sorting the patients for  
322 *MUC4* alone. Overall, we observed that *MUC4/MUC16/MUC20* signature harbored an  
323 increased hazard ratio compared with *MUC4* alone for pancreatic cancer and to a lower  
324 extent in bladder cancer, colon cancer, lung squamous cancer and stomach cancer.

325 We analyzed MUC4, MUC16 and MUC20 expression in pancreatic tumor (T) and paired  
326 adjacent non tumoral tissues (ANT) from GSE28735 (Figure 6) and GSE16515 (not shown)  
327 datasets [25, 26]. We confirmed *MUC4* overexpression in tumor tissues ( $p<0.0001$ ). *MUC16*  
328 and *MUC20* mRNA level were also increased ( $p<0.0001$  and  $p=0.0062$ ) in tumor samples  
329 (Figure 8A). As previously observed in CCLE dataset, MUC4 expression was correlated with  
330 MUC16 ( $p=0.0006$ ) and MUC20 ( $p=0.0621$ ) in GSE28735 (Supplemental Figure 6). We also  
331 analyzed MUC4, MUC16 and MUC20 expression in datasets of other cancers (supplemental  
332 Figure 7). MUC4 expression is increased in bladder cancer vs ANT (GSE13507,  $p<0.01$ ).  
333 MUC20 is increased in lung adenocarcinoma vs normal samples (GSE30219,  $p<0.05$ ).  
334 MUC4 and MUC20 expression is increased in colorectal cancer vs normal mucosae  
335 (GSE40967,  $p<0.01$ ). MUC16 and MUC20 relative expression is increased in ovarian  
336 adenocarcinoma (GSE14407,  $p<0.01$  and  $p<0.05$  respectively). ROC curves of MUC4,  
337 MUC16, MUC20 and MUC4+MUC16+MUC20 combination were established using  
338 GSE28735 dataset. The combination of MUC4+MUC16+MUC20 produced a high specificity  
339 of 97.78% (88.23-99.94) and a mild sensitivity of 55.56% (40-70.36) (likelihood ratio = 25)  
340 (Figure 8B). Similar results were obtained for GSE16515 with 93.75% specificity and 69.44%  
341 sensitivity (LR+=11.11) (not shown). MUC16 AUROC was similar to that of  
342 MUC4+MUC16+MUC20 in GSE28735 dataset but harbors a lower specificity/sensitivity in  
343 GSE16515.

344 Altogether, this suggests that MUC4/MUC16/MUC20<sup>high</sup> signature would be useful in  
345 stratification of patients with worst prognosis in several carcinoma and notably pancreatic,  
346 stomach and colon cancers.

347

348 **Discussion**

349

350 The TCGA and the CCLE have provided a tremendous amount of publicly available data  
351 combining gene expression information related to clinical outcome. Web-based tools allow  
352 the scientific community to perform powerful large scale genomic analysis and propose new  
353 biomarkers or new therapeutic targets. In the present report, we analyzed *MUC4* expression  
354 systematically in all organs and confirmed its aberrant expression in associated carcinoma.  
355 We identified 187 genes for which the expression is correlated with *MUC4* expression. These  
356 genes are involved in cell adhesion, cell-cell junctions, glycosylation and cell signaling.  
357 *MUC4* was also correlated with *MUC16* and *MUC20* membrane bound mucins. This  
358 combination is associated with a poorer overall survival in different cancers including  
359 pancreatic, colon and stomach cancers suggesting *MUC4/MUC16/MUC20* as a poor  
360 prognostic signature for these cancers.

361 Previous works have showed that *MUC4* is altered in normal, premalignant and malignant  
362 epithelia of the digestive tract [27]. The mechanisms underlying this alteration of expression  
363 are diverse and involve regulators such as growth factors, cytokines, demethylation of  
364 promoters and miRNA [28-32]. In the present manuscript we also observe that *MUC4* gene is  
365 amplified in 13% of cancer cell lines. We also found a mild correlation between alteration of  
366 *MUC4* copy number and *MUC4* expression suggesting that gene amplification could also  
367 mediate this *MUC4* aberrant expression. This kind of regulation is scarcely described in the  
368 literature. In TCGA, We confirmed that *MUC4* expression was observed mainly in human  
369 carcinomas including bladder, cervix, head and neck, lung, ovarian, pancreatic, prostate,  
370 stomach carcinomas. For most of these organs, *MUC4* high expression was associated with  
371 a poorer overall survival. *MUC4* is one of the most differentially expressed genes in  
372 pancreatic cancer that are thought to be potential clinical targets [33]. Recently, a meta-  
373 analysis based on 1900 patients from 18 studies showed that *MUC4* overexpression was

374 associated with tumor stage, tumor invasion and lymph node metastasis [34]. A worse  
375 overall survival was observed in MUC4-overexpressing patients with biliary tract carcinoma  
376 (HR 2.41), pancreatic cancer (HR 2.01), and colorectal cancer (HR 1.73). Using the TCGA  
377 cohorts, we extended this finding on lung adenocarcinoma, lung squamous carcinoma,  
378 ovarian cancer, skin cancer and stomach cancer. The authors noted that a limit of this meta-  
379 analysis was insufficient statistical power of some eligible studies. The large scale genomic  
380 approach of TCGA helps us to overcome this limitation. Based on available TCGA datasets,  
381 mucin mutation map was generated by cBioPortal Mutation Mapper [35]. *MUC4* mutations  
382 were notably observed in Kidney Clear Cell Renal Carcinoma (20-45%) and were correlated  
383 with survival outcomes. Rare mutations were described in the main overexpressing model  
384 that is pancreatic cancer. Because of the very large size of *MUC4* gene, probability of  
385 acquiring mutation could be increased. MUC4 belongs to the most mutated genes upon  
386 stress exposure such as nicotine treatment or aging [36, 37]. The enrichment of mutation of  
387 MUC4 could be related with the fact that the first risk factor of kidney cancer is smoking [38]  
388 and that kidney cancer diagnosis is occurring at elder ages (65 years) [39]. Pancreatic  
389 cancer shares these characteristics but harbors a very rare mutation occurrence (3%)  
390 suggesting that aging could be specific of cancers such as kidney or lung and that  
391 overexpression is more important for other cancers. So far, functional consequences of  
392 MUC4 mutation remain to be elucidated.

393 We and others have investigated MUC4 biological roles in various cancers such as  
394 pancreatic, ovarian, esophagus and lung cancers. MUC4 was shown to promote  
395 aggressiveness of tumors as it induces proliferation, migration, invasion, EMT, cell stemness  
396 and chemoresistance [9, 11-14]. In the present work, we showed that *MUC4* expression was  
397 correlated with genes, such as integrins cadherin-type proteins, involved in cell adhesion and  
398 cell-cell junctions. As a membrane-bound mucin, MUC4 is thought to act on cell-cell and cell-  
399 MEC interaction. Because of its huge extracellular domain that profoundly modifies steric  
400 hindrance, MUC4 may alter migration, invasion and adherence properties [40]. Rat

401 homologue of MUC4, sialomucin complex (SMC), overexpression leads to suppression of  
402 cell adhesion [41]. Notably, MUC4 overexpression disrupts the adherens junctions and leads  
403 to partial delocalization of E-cadherin to the apical surface of the cell causing loss of cell  
404 polarity [42]. Moreover, interactions between MUC4 glycans and galectin-3 were shown to  
405 also mediate docking of circulating tumor cells to the surface of endothelial cells [43]. The  
406 alteration of cell adhesion induced by MUC4 is one of the first steps toward the metastatic  
407 process. MUC4 expression was also correlated with several genes encoding glycosylation  
408 enzymes or glycoproteins. This essential set of genes is involved in a wide set of cellular  
409 function including cell adhesion, barrier role, interaction with selection of endothelial cells or  
410 regulation of cell signaling [5, 44]. The glycan-associated antigens are commonly associated  
411 with patient survival of gastrointestinal cancer [45]. Alteration of MUC4 glycosylation is  
412 proposed to play a substantial role in binding properties mediated by the extracellular subunit  
413 of MUC4 and the NIDO domain [46]. One should note that the expression of these genes is  
414 correlated with MUC4. However, a direct regulatory mechanism remains to be demonstrated  
415 in future studies.

416 In order to regulate these major biological properties, MUC4 has been commonly associated  
417 with cell signaling alteration and notably MAPK, NF- $\kappa$ B, or FAK signaling pathways.  
418 Interestingly, we observed that MUC4 expression is highly correlated with proteins containing  
419 Src Homology 2 (SH2) domain or Src Homology 3 (SH3) domains. Intracellular adaptor  
420 signaling proteins family is characterized by one SH2 and at least one SH3 domain and is  
421 crucial for effective integrating of intracellular and extracellular stimuli [47].

422 It is interesting to note that MUC4 expression is not correlated with MUC1 that is a major  
423 membrane-bound mucin commonly overexpressed in cancer [48, 49]. In the US, it was  
424 estimated that 900 000 cancers, out of 1 400 000, harbor overexpression of MUC1  
425 highlighting its attractiveness as a therapeutic target. This could be explained by different  
426 regulatory mechanisms such as different signaling pathways or different miRNA regulating  
427 the two mucins.

428 MUC16 is the peptide part to the CA125 serum marker for ovarian cancer [50]. MUC16 is a  
429 very large mucin (22 000 amino acid (aa)) that is heavily glycosylated and facilitates ovarian  
430 cancer. MUC20 is a small mucin (500 aa) mostly expressed in renal proximal tube and that is  
431 deregulated in several cancers such as colorectal or ovarian cancers where it favors  
432 aggressiveness [17, 18]. MUC16/CA125 is routinely used in clinics unlike MUC4 and  
433 MUC20. In the present manuscript, we showed that expression of *MUC16* and *MUC20* are  
434 positively correlated with *MUC4* and that the *MUC4/MUC16/MUC20*<sup>high</sup> combinatory  
435 expression is associated with an increased hazard ratio and reduced overall survival  
436 suggesting a potential for this signature as a prognostic marker for several carcinomas and  
437 notably pancreatic, stomach and colon cancer. Biomarkers for pancreatic cancer are needed  
438 for detection and evaluation of response to therapy [51]. Unfortunately, the marker currently  
439 used (CA19.9) lacks sensitivity or specificity to be used in cancer diagnosis. Similarly  
440 established biomarker with adequate sensitivity and specificity are lacking for gastric cancer  
441 [52]. The need of biomarkers is less urgent for colorectal cancer since several  
442 predictive/prognostic/diagnostic biomarkers have been described [53].

443 The present work highlights the relationship between MUC4/MUC16/MUC20 expression and  
444 overall survival. This signature could be proposed as a prognostic marker. Moreover, MUC4  
445 is expressed in the earliest stage (PanIN1A) of pancreatic cancer but is not specific enough.  
446 The potential of the combination *MUC4/MUC16/MUC20* as a diagnosis marker is not known  
447 and remains to be investigated in the future. Moreover, development of unsupervised  
448 algorithm will allow the identification of new non intentional bigger signatures leading to  
449 better prognostic and predictive performances. Genome wide computational unsupervised  
450 procedure from discovery dataset will help to determine hypothesis signature. The signature  
451 will be subsequently validated on a number of independent datasets. Thus, multi-platform  
452 analysis using TCGA datasets helped to characterize the complex molecular landscape of  
453 PDAC [54]. Another meta-analysis approach based on PC datasets allowed the identification  
454 of a 5 genes classifier signature (TMPRSS4, AHNAK2, POSTN, ECT2, SERPINB5) with

MUC4 and TCGA and CCLE databases

455 95% sensitivity and 89% specificity in discriminating PDAC from non-tumor samples [55].

456 Interestingly, TMPRSS4 and SERPINB5 are two genes belonging to the gene list correlated  
457 with MUC4 expression.

458

459 **Conclusion**

460

461 We analyzed MUC4 expression systematically in all organs in TCGA and CCLE large scale  
462 databases and confirmed its aberrant expression in associated carcinoma and the MUC4  
463 impact on patient's survival. Moreover, 187 genes (involved in cell adhesion, cell-cell  
464 junctions, glycosylation and cell signaling) were correlated with MUC4. Among them, *MUC16*  
465 and *MUC20* membrane bound mucins and their combination *MUC4/MUC16/MUC20* is  
466 associated with a poorer overall survival in different cancers including pancreatic, colon and  
467 stomach cancers suggesting *MUC4/MUC16/MUC20* as a poor prognostic signature for these  
468 cancers. This potential as new biomarkers remains to be investigated in the future.

469

470

471 **Declarations section:**

472

473 Ethics approval and consent to participate: not applicable

474 Consent to publish: not applicable

475 Availability of data and materials: All data are available and are based upon public data  
476 extracted from the TCGA Research Network (<http://cancergenome.nih.gov/>), Genome Tissue  
477 Expression (GTEx) project (<http://www.GTEXportal.org/>) and Gene Expression Omnibus  
478 (GEO) database (<http://www.ncbi.nlm.nih.gov/geo/>).

479 Competing interests: Authors declare no conflict of interest

480 Funding: Our work is supported by grants from la Ligue Nationale Contre le Cancer (Comités  
481 59, 62, 80, IVS, NJ), from SIRIC ONCOLille, Grant INCaDGOS-Inserm 6041 (IVS, NJ) and  
482 from région Nord-Pas de Calais “Contrat de Plan Etat Région” CPER Cancer 2007-13 (IVS).

483 Authors' Contributions NJ conceived and designed the analysis. NJ analyzed the data. NJ  
484 and IVS wrote and edited the paper.

485 Acknowledgements: We are grateful to M. Foster and A. Turner for helpful contributions and  
486 Dr B Neve, Dr A. Vincent, Dr R. Vasseur (Inserm UMR-S1172, Lille) for their critical reading  
487 of the manuscript.

488

489

490

491

492 **References**

493

- 494 1. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S *et al*: The Cancer Cell  
 495 Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 2012,  
 496 **483**(7391):603-607.
- 497 2. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA *et al*: The cBio cancer genomics  
 498 portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov*  
 499 2012, **2**(5):401-404.
- 500 3. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO *et al*: Integrative analysis of  
 501 complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* 2013,  
 502 **6**(269):pl1.
- 503 4. Aguirre-Gamboa R, Gomez-Rueda H, Martinez-Ledesma E, Martinez-Torteya A, Chacolla-  
 504 Huaranga R, Rodriguez-Barrientos A *et al*: SurvExpress: an online biomarker validation tool  
 505 and database for cancer gene expression data using survival analysis. *PLoS One* 2013,  
 506 **8**(9):e74250.
- 507 5. Corfield AP: Mucins: a biologically relevant glycan barrier in mucosal protection. *Biochim*  
 508 *Biophys Acta* 2015, **1850**(1):236-252.
- 509 6. Dekker J, Rossen JW, Buller HA, Einerhand AW: The MUC family: an obituary. *Trends Biochem*  
 510 *Sci* 2002, **27**(3):126-131.
- 511 7. Porchet N, Nguyen VC, Dufosse J, Audie JP, Guyonnet-Duperat V, Gross MS *et al*: Molecular  
 512 cloning and chromosomal localization of a novel human tracheo-bronchial mucin cDNA  
 513 containing tandemly repeated sequences of 48 base pairs. *Biochem Biophys Res Commun*  
 514 1991, **175**(2):414-422.
- 515 8. Jonckheere N, Skrypek N, Frenois F, Van Seuning I: Membrane-bound mucin modular  
 516 domains: from structure to function. *Biochimie* 2013, **95**(6):1077-1086.
- 517 9. Jonckheere N, Skrypek N, Merlin J, Dessein AF, Dumont P, Leteurtre E *et al*: The mucin MUC4  
 518 and its membrane partner ErbB2 regulate biological properties of human CAPAN-2  
 519 pancreatic cancer cells via different signalling pathways. *PLoS One* 2012, **7**(2):e32232.
- 520 10. Jonckheere N, Skrypek N, Van Seuning I: Mucins and pancreatic cancer. *Cancers (Basel)*  
 521 2010, **2**(4):1794-1812.
- 522 11. Bruyere E, Jonckheere N, Frenois F, Mariette C, Van Seuning I: The MUC4 membrane-  
 523 bound mucin regulates esophageal cancer cell proliferation and migration properties:  
 524 Implication for S100A4 protein. *Biochem Biophys Res Commun* 2011, **413**(2):325-329.
- 525 12. Skrypek N, Duchene B, Hebbar M, Leteurtre E, van Seuning I, Jonckheere N: The MUC4  
 526 mucin mediates gemcitabine resistance of human pancreatic cancer cells via the  
 527 Concentrative Nucleoside Transporter family. *Oncogene* 2013, **32**(13):1714-1723.
- 528 13. Bafna S, Kaur S, Momi N, Batra SK: Pancreatic cancer cells resistance to gemcitabine: the role  
 529 of MUC4 mucin. *Br J Cancer* 2009, **101**(7):1155-1161.
- 530 14. Kaur S, Kumar S, Momi N, Sasson AR, Batra SK: Mucins in pancreatic cancer and its  
 531 microenvironment. *Nat Rev Gastroenterol Hepatol* 2013, **10**(10):607-620.
- 532 15. Duraisamy S, Ramasamy S, Kharbanda S, Kufe D: Distinct evolution of the human carcinoma-  
 533 associated transmembrane mucins, MUC1, MUC4 AND MUC16. *Gene* 2006, **373**:28-34.
- 534 16. Bafna S, Kaur S, Batra SK: Membrane-bound mucins: the mechanistic basis for alterations in  
 535 the growth and survival of cancer cells. *Oncogene* 2010, **29**(20):2893-2904.
- 536 17. Chen CH, Wang SW, Chen CW, Huang MR, Hung JS, Huang HC *et al*: MUC20 overexpression  
 537 predicts poor prognosis and enhances EGF-induced malignant phenotypes via activation of  
 538 the EGFR-STAT3 pathway in endometrial cancer. *Gynecol Oncol* 2013, **128**(3):560-567.
- 539 18. Xiao X, Wang L, Wei P, Chi Y, Li D, Wang Q *et al*: Role of MUC20 overexpression as a predictor  
 540 of recurrence and poor outcome in colorectal cancer. *J Transl Med* 2013, **11**:151.

- 541 19. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S *et al*: The Genotype-Tissue  
542 Expression (GTEx) project. *Nat Genet* 2013, **45**(6):580-585.
- 543 20. Ardlie KG, Deluca DS, Segrè AV, Sullivan TJ, Young TR, Gelfand ET *et al*: Human genomics. The  
544 Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans.  
545 *Science* 2015, **348**(6235):648-660.
- 546 21. Huang da W, Sherman BT, Lempicki RA: Systematic and integrative analysis of large gene lists  
547 using DAVID bioinformatics resources. *Nat Protoc* 2009, **4**(1):44-57.
- 548 22. Huang da W, Sherman BT, Lempicki RA: Bioinformatics enrichment tools: paths toward the  
549 comprehensive functional analysis of large gene lists. *Nucleic Acids Res* 2009, **37**(1):1-13.
- 550 23. Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M *et al*: The STRING database  
551 in 2017: quality-controlled protein-protein association networks, made broadly accessible.  
552 *Nucleic Acids Res* 2017, **45**(D1):D362-D368.
- 553 24. Yonezawa S, Higashi M, Yamada N, Yokoyama S, Kitamoto S, Kitajima S *et al*: Mucins in  
554 human neoplasms: clinical pathology, gene expression and diagnostic application. *Pathol Int*  
555 2011, **61**(12):697-716.
- 556 25. Pei H, Li L, Fridley BL, Jenkins GD, Kalari KR, Lingle W *et al*: FKBP51 affects cancer cell  
557 response to chemotherapy by negatively regulating Akt. *Cancer Cell* 2009, **16**(3):259-266.
- 558 26. Zhang G, Schetter A, He P, Funamizu N, Gaedcke J, Ghadimi BM *et al*: DPEP1 inhibits tumor  
559 cell invasiveness, enhances chemosensitivity and predicts clinical outcome in pancreatic  
560 ductal adenocarcinoma. *PLoS One* 2012, **7**(2):e31507.
- 561 27. Jonckheere N, Van Seuning I: The membrane-bound mucins: From cell signalling to  
562 transcriptional regulation and expression in epithelial cancers. *Biochimie* 2010, **92**(1):1-11.
- 563 28. Andrianifahanana M, Singh AP, Nemos C, Ponnusamy MP, Moniaux N, Mehta PP *et al*: IFN-  
564 gamma-induced expression of MUC4 in pancreatic cancer cells is mediated by STAT-1  
565 upregulation: a novel mechanism for IFN-gamma response. *Oncogene* 2007, **26**(51):7251-  
566 7261.
- 567 29. Jonckheere N, Perrais M, Mariette C, Batra SK, Aubert JP, Pigny P *et al*: A role for human  
568 MUC4 mucin gene, the ErbB2 ligand, as a target of TGF-beta in pancreatic carcinogenesis.  
569 *Oncogene* 2004, **23**(34):5729-5738.
- 570 30. Vincent A, Ducourouble MP, Van Seuning I: Epigenetic regulation of the human mucin  
571 gene MUC4 in epithelial cancer cell lines involves both DNA methylation and histone  
572 modifications mediated by DNA methyltransferases and histone deacetylases. *Faseb J* 2008,  
573 **22**(8):3035-3045.
- 574 31. Yamada N, Nishida Y, Tsutsumida H, Goto M, Higashi M, Nomoto M *et al*: Promoter CpG  
575 methylation in cancer cells contributes to the regulation of MUC4. *Br J Cancer* 2009,  
576 **100**(2):344-351.
- 577 32. Lahdaoui F, Delpu Y, Vincent A, Renaud F, Messenger M, Duchene B *et al*: miR-219-1-3p is a  
578 negative regulator of the mucin MUC4 expression and is a tumor suppressor in pancreatic  
579 cancer. *Oncogene* 2015, **34**(6):780-788.
- 580 33. Iacobuzio-Donahue CA, Ashfaq R, Maitra A, Adsay NV, Shen-Ong GL, Berg K *et al*: Highly  
581 expressed genes in pancreatic ductal adenocarcinomas: a comprehensive characterization  
582 and comparison of the transcription profiles obtained from three major technologies. *Cancer*  
583 *Res* 2003, **63**(24):8614-8622.
- 584 34. Huang X, Wang X, Lu SM, Chen C, Wang J, Zheng YY *et al*: Clinicopathological and prognostic  
585 significance of MUC4 expression in cancers: evidence from meta-analysis. *Int J Clin Exp Med*  
586 2015, **8**(7):10274-10283.
- 587 35. King RJ, Yu F, Singh PK: Genomic alterations in mucins across cancers. *Oncotarget* 2017.
- 588 36. Bavarva JH, Tae H, Mclver L, Garner HR: Nicotine and oxidative stress induced exomic  
589 variations are concordant and overrepresented in cancer-associated genes. *Oncotarget* 2014,  
590 **5**(13):4788-4798.
- 591 37. Bavarva JH, Tae H, Mclver L, Karunasena E, Garner HR: The dynamic exome: acquired variants  
592 as individuals age. *Aging (Albany NY)* 2014, **6**(6):511-521.

- 593 38. Hunt JD, van der Hel OL, McMillan GP, Boffetta P, Brennan P: Renal cell carcinoma in relation  
594 to cigarette smoking: meta-analysis of 24 studies. *Int J Cancer* 2005, **114**(1):101-108.
- 595 39. Hayat MJ, Howlader N, Reichman ME, Edwards BK: Cancer statistics, trends, and multiple  
596 primary cancer analyses from the Surveillance, Epidemiology, and End Results (SEER)  
597 Program. *Oncologist* 2007, **12**(1):20-37.
- 598 40. Hollingsworth MA, Swanson BJ: Mucins in cancer: protection and control of the cell surface.  
599 *Nat Rev Cancer* 2004, **4**(1):45-60.
- 600 41. Komatsu M, Tatum L, Altman NH, Carothers Carraway CA, Carraway KL: Potentiation of  
601 metastasis by cell surface sialomucin complex (rat MUC4), a multifunctional anti-adhesive  
602 glycoprotein. *Int J Cancer* 2000, **87**(4):480-486.
- 603 42. Pino V, Ramsauer VP, Salas P, Carothers Carraway CA, Carraway KL: Membrane mucin Muc4  
604 induces density-dependent changes in ERK activation in mammary epithelial and tumor cells:  
605 role in reversal of contact inhibition. *J Biol Chem* 2006, **281**(39):29411-29420.
- 606 43. Senapati S, Chaturvedi P, Chaney WG, Chakraborty S, Gnanapragassam VS, Sasson AR *et al*:  
607 Novel INTERaction of MUC4 and galectin: potential pathobiological implications for  
608 metastasis in lethal pancreatic cancer. *Clin Cancer Res* 2011, **17**(2):267-274.
- 609 44. Pinho SS, Reis CA: Glycosylation in cancer: mechanisms and clinical implications. *Nat Rev*  
610 *Cancer* 2015, **15**(9):540-555.
- 611 45. Baldus SE, Hanisch FG: Biochemistry and pathological importance of mucin-associated  
612 antigens in gastrointestinal neoplasia. *Adv Cancer Res* 2000, **79**:201-248.
- 613 46. Hanson RL, Hollingsworth MA: Functional Consequences of Differential O-glycosylation of  
614 MUC1, MUC4, and MUC16 (Downstream Effects on Signaling). *Biomolecules* 2016, **6**(3).
- 615 47. Reebye V, Frilling A, Hajitou A, Nicholls JP, Habib NA, Mintz PJ: A perspective on non-catalytic  
616 Src homology (SH) adaptor signalling proteins. *Cell Signal* 2012, **24**(2):388-392.
- 617 48. Kufe DW: Functional targeting of the MUC1 oncogene in human cancers. *Cancer Biol Ther*  
618 2009, **8**(13):1197-1203.
- 619 49. Kufe DW: Mucins in cancer: function, prognosis and therapy. *Nat Rev Cancer* 2009,  
620 **9**(12):874-885.
- 621 50. Yin BW, Lloyd KO: Molecular cloning of the CA125 ovarian cancer antigen: identification as a  
622 new mucin, MUC16. *J Biol Chem* 2001, **276**(29):27371-27375.
- 623 51. Kleeff J, Korc M, Apte M, La Vecchia C, Johnson CD, Biankin AV *et al*: Pancreatic cancer. *Nat*  
624 *Rev Dis Primers* 2016, **2**:16022.
- 625 52. Ajani JA, Lee J, Sano T, Janjigian YY, Fan D, Song S: Gastric adenocarcinoma. *Nat Rev Dis*  
626 *Primers* 2017, **3**:17036.
- 627 53. Kuipers EJ, Grady WM, Lieberman D, Seufferlein T, Sung JJ, Boelens PG *et al*: Colorectal  
628 cancer. *Nat Rev Dis Primers* 2015, **1**:15065.
- 629 54. TCGA-Network.: Integrated Genomic Characterization of Pancreatic Ductal Adenocarcinoma.  
630 *Cancer Cell* 2017, **32**(2):185-203 e113.
- 631 55. Bhasin MK, Ndebele K, Bucur O, Yee EU, Otu HH, Plati J *et al*: Meta-analysis of transcriptome  
632 data identifies a novel 5-gene pancreatic adenocarcinoma classifier. *Oncotarget* 2016,  
633 **7**(17):23263-23281.

634

635

636

## Figure legends

637

638 **Figure 1: Strategy of analysis of genes correlated with MUC4 expression in Cancer**  
639 **Cell Line Encyclopedia.** (A) Flowchart of MUC4 analysis. *MUC4* mRNA expression z-  
640 scores were extracted from Cancer Cell Line Encyclopedia using cBioportal . The list of gene  
641 correlated with *MUC4* expression was determined by using the co-expression tool. Genes  
642 presenting a Pearson's correlation higher than 0.3 or lower than -0.3 were selected.  
643 Spearman analysis was performed subsequently. Gene ontology annotation and clustering  
644 were performed using DAVID 6.8 functional annotation tool . (B) Example of *MUC4-MUC16*  
645 correlation of mRNA expression. (C) Example of *MUC4-MUC20* correlation of mRNA  
646 expression.

647

648 **Figure 2: *MUC4* expression in Cancer Cell Line Encyclopedia.** *MUC4* mRNA expression  
649 z-scores were extracted from Cancer Cell Line Encyclopedia (Novartis/Barretina Nature  
650 2012) using cBioportal . N=881 samples. Expression data were sorted depending on tumor  
651 type (A) and histology (B).

652

653 **Figure 3: *MUC4* expression in cancer samples from TCGA.** *MUC4* mRNA expression z-  
654 scores were extracted from TCGA samples using cBioportal. N=13 489 samples. Expression  
655 data were sorted depending on organs.

656

657 **Figure 4: Correlation of *MUC4* expression and methylation of genes correlated with**  
658 ***MUC4*.** The top genes were defined as genes harboring Pearson's correlation higher than  
659 0.5 with *MUC4* expression. *MUC4* mRNA expression and methylation score (Reduced  
660 Representation Bisulfite Sequencing: RRBS) of *ADGRF1*, *LCN2*, *MUC20*, *C10RF116*,

661 *STEAP4, SCEL, WFDC2, GJB3, SH2D3A, RNF39, PRSS22, HS3ST1, GPR87, TACST2,*  
662 *MUC16, FAM83A, LAMC2, B3GNT3, CLDN7* were extracted using  
663 <https://portals.broadinstitute.org/ccle>

664

665 **Figure 5: *MUC4* expression is associated with reduced overall survival of carcinoma.**

666 (A) Hazard ratio was calculated in population designated as *MUC4* high risk and low risk  
667 (higher value of *MUC4* for higher risk) by SurvExpress optimized algorithm in every cancer  
668 from TCGA datasets. (B) Overall survival values of *MUC4* high and low risk groups in  
669 bladder cancer, colon cancer, lung adenocarcinoma, lung squamous carcinoma, ovarian  
670 cancer, skin cancer, stomach cancer, available in TCGA datasets. The numbers below  
671 horizontal axis represent the number of individuals not presenting the event of *MUC4* high  
672 and low risk group along time.

673

674 **Figure 6: Hazard ratio of signatures defined by gene ontology terms and top-genes**

675 **correlated with *MUC4*.** (A) Hazard ratio was calculated in bladder cancer, colon cancer,  
676 lung adenocarcinoma, lung squamous carcinoma, ovarian cancer, pancreatic cancer, skin  
677 cancer and stomach cancer. The populations were defined according to GO term extracted  
678 from list of gene correlated with *MUC4* (GO 0031424: keratinization, GO 0007155: cell  
679 adhesion, GO 0019897: extrinsic component of plasma membrane, GO 0016323 :  
680 basolateral plasma membrane and GO 0016324: apical plasma membrane). (B) A Hazard  
681 ratio was calculated in populations designated as high risk and low risk for top genes  
682 (ADGRF1, LCN2, MUC20, C1ORF116, SCEL, STEAP4) that harbored Pearson's correlation  
683 with *MUC4* superior to 0.5.

684

685 **Figure 7: *MUC4/MUC16/MUC20* expression is associated with reduced overall survival**  
686 **of pancreatic adenocarcinoma. (A)** Overall survival of *MUC4/MUC16/MUC20* high and low  
687 risk group in pancreatic cancer available in TCGA datasets. High risk and low risk cohorts  
688 were determined by SurvExpress optimized algorithm. Log rang test and Hazard ratio were  
689 calculated to compare both cohorts. **(B)** Box plot of *MUC4*, *MUC16* and *MUC20* expression  
690 and the corresponding p value testing the differences between high risk and low risk groups.  
691 **(C)** Overall survival of *MUC4/MUC16/MUC20* high and low risk groups in ICGC, Stratford  
692 (GSE21521) and Zhang (GSE 28735) datasets available in SurvExpress.

693

694 **Figure 8: Expression and ROC curves of the *MUC4/MUC16/MUC20* signature in a**  
695 **pancreatic adenocarcinoma dataset. (A)** *MUC4*, *MUC16* and *MUC20* mRNA expression  
696 was evaluated in GSE28735 dataset to analyze whether the mRNA level differed between  
697 normal and tumor tissues. Statistical analyses were performed using paired t-test (\*\*\*\*  
698  $p < 0.0001$ , \*\*  $p < 0.01$ ). (B) ROC curves and Area under ROC measurement (AUROC) of  
699 *MUC4*, *MUC16*, *MUC20* and the combination in GSE28735 dataset.

700

701 **Table 1: List of mRNA positively correlated with *MUC4*.** Data were retrieved from 881  
702 samples of Cancer Cell Line Encyclopedia (Novartis/Broad, Nature 2012). Correlation  
703 analysis was performed using cBioPortal.org online tool. 178 genes presented a Pearson's  
704 correlation higher than 0.3.

705

706 **Table 2: List of mRNA negatively correlated with *MUC4*.** Data were retrieved from 881  
707 samples of Cancer Cell Line Encyclopedia (Novartis/Broad, Nature 2012). Correlation  
708 analysis was performed using cBioPortal.org online tool. 9 genes presented a Pearson's  
709 correlation lower than -0.3.

710

711 **Table 3: Gene ontology clustering on genes correlated with *MUC4* expression.** Gene  
712 list was retrieved from 881 samples of Cancer Cell Line Encyclopedia (Baretina, Nature  
713 2012). 187 genes that are positively (n=178) or negatively (n=9) correlated with *MUC4*  
714 expression were selected. Functional Annotation and gene clustering were performed using  
715 David Functional Annotation Tool.

716

717 **Table 4: Hazard-ratio and survival analysis of high and low risk in TCGA tumor**  
718 **databases.** Hazard ratio and p-value were determined using SurvExpress tool. Risk group  
719 were determined using the optimization algorithm (maximize) from the ordered prognostic  
720 index (higher values of *MUC4* expression for higher risk).

721

722

723 **Supplemental material legends**

724

725 **Supplemental Figure 1: *MUC4* Oncoprint in Cancer Cell Line Encyclopedia.** *MUC4*  
726 alterations were explored in Cancer Cell Line Encyclopedia dataset using cBioPortal webtool.  
727 The oncoprint represents the amplification, deletion, up regulation or in frame mutation.

728

729 **Supplemental Figure 2: *MUC4* expression in normal tissues.** *MUC4* expression was  
730 analyzed with <https://gtexportal.org>. Expression is shown as log10 of RKPM (read per  
731 kilobases of transcript per million map reads). Boxplot are shown as median and 25/75%  
732 percentile. Outliers are represented as points.

733

734 **Supplemental figure 3: interaction network of the proteins correlated with *MUC4***  
735 **expression.** Interacting proteins were determined by String 10 tool and are represented by  
736 nodes. Edges represent a relationship between two nodes (known interaction from curated  
737 databases or experimentally determined; predicted interaction from gene neighborhood, gene  
738 fusion or co-occurrence; textmining; co-expression; protein homology). The obtained network  
739 was divided in 3 clusters by k-means clustering.

740

741 **Supplemental Figure 4: Correlation of *MUC4* expression and copy numbers of genes**  
742 **correlated with *MUC4*.** The top genes were defined as genes harboring Pearson's  
743 correlation higher than 0.5 with *MUC4* expression. *MUC4* mRNA expression and log2 copy  
744 number of *ADGRF1*, *LCN2*, *MUC20*, *C10RF116*, *STEAP4*, *SCEL*, *MUC16* were extracted  
745 using <https://portals.broadinstitute.org/ccle>

746

747 **Supplemental Figure 5: Overall survival of MUC4/MUC16/MUC20 high and low risk**  
748 **groups in cancer datasets available in TCGA. (A)** Overall survival of  
749 *MUC4/MUC16/MUC20* high and low risk groups in bladder cancer, colon cancer, lung  
750 adenocarcinoma, lung squamous adenocarcinoma, skin cancer and stomach cancer. High  
751 risk and low risk cohorts were determined by SurvExpress optimized algorithm. Log rang test  
752 and Hazard ratio were calculated to compare both cohorts. The numbers below horizontal  
753 axis represent the number of individuals not presenting the event of *MUC4* high and low risk  
754 group along time. **(B)** Overall survival of *MUC4/MUC16/MUC20* high and low risk group in  
755 liver and acute myeloid leukemia (AML).

756

757 **Supplemental Figure 6: *MUC4-MUC16* and *MUC4-MUC20* correlation of mRNA**  
758 **expression in 45 tumor tissues of GSE28735 PDAC dataset.**

759

760 **Supplemental Figure 7: *MUC4*, *MUC16* and *MUC20* expression in bladder, colorectal,**  
761 **lung, stomach, skin and ovarian cancer datasets. *MUC4*, *MUC16* and *MUC20* mRNA**  
762 expression was evaluated in datasets to analyze whether the mRNA level differed between  
763 normal and tumor tissues. (A) GSE13507 contains 165 bladder cancer and 58 ANT samples.  
764 (B) GSE30219 contains 14 normal lung, 85 adenocarcinomas and 61 squamous cancer  
765 samples. (C) GSE40967 contains 566 colorectal cancers and 19 normal mucosae. (D)  
766 GSE27342 contains 80 tumors and 80 paired ANT tissues. (E) GSE4587 contains 2 normal,  
767 2 melanomas and 2 metastatic melanomas. (F) GSE14407 contains 12 ovarian  
768 adenocarcinomas and 12 normal ovary samples. Statistical analyses were performed using  
769 paired t-test (\*  $p < 0.05$ , \*\*  $p < 0.01$ ).

770

771 **Supplemental table 1: Ontology of genes correlated with *MUC4* expression.** Gene list  
772 was retrieved from 881 samples of Cancer Cell Line Encyclopedia (Novartis/Broad, Nature  
773 2012). 187 genes that are positively (n=178) or negatively (n=9) correlated with MUC4  
774 expression were selected. Functional Annotation was performed using David Functional  
775 Annotation Tool.

776

777 **Supplemental table 2: Hazard-ratio and survival analysis of most significant genes**  
778 **clustered in GO term associated with MUC4 expression in TCGA tumor databases.**  
779 Hazard ratio and p-value were determined using SurvExpress tool  
780 (<http://bioinformatica.mty.itesm.mx/SurvExpress>). Risk groups were sorted depending on the  
781 major GO term GO 0031424, GO 00071555, GO 0019897, GO 0016323 and GO 0016324  
782 using the optimization algorithm (maximize) from the ordered prognostic.

783

784 **Supplemental table 3: Hazard-ratio and survival analysis of top genes associated with**  
785 **MUC4 expression in TCGA tumor databases.** Hazard ratio and p-value were determined  
786 using SurvExpress tool (<http://bioinformatica.mty.itesm.mx/SurvExpress>). Risk groups were  
787 defined using the optimization algorithm (maximize) from the ordered prognostic. Selected  
788 genes (ADGRF1, LCN2, MUC20, C1ORF116, SCEL, STEAP4) harbored Pearson's  
789 correlation with MUC4 > 0.5.

**Table 1: List of mRNA positively correlated with MUC4.** Data were retrieved from 881 samples of Cancer Cell Line Encyclopedia (Novartis/Broad, Nature 2012). Correlation analysis was performed using cBioPortal.org online tool. 178 genes presented a Pearson's correlation higher than 0.3.

Correlated gene	cytoband	Pearson's correlation	Spearman's correlation
ADGRF1	6p12.3	<b>0.56</b>	0.40
LCN2	9q34	<b>0.54</b>	0.41
MUC20	3q29	<b>0.54</b>	0.42
C1ORF116	1q32.1	<b>0.52</b>	0.47
SCEL	13q22	<b>0.52</b>	0.43
STEAP4	7q21.12	<b>0.51</b>	0.35
WFDC2	20q13.12	<b>0.48</b>	0.31
GJB3	1p34	<b>0.48</b>	0.35
SH2D3A	19p13.3	<b>0.48</b>	0.45
RNF39	6p21.3	<b>0.47</b>	0.35
PRSS22	16p13.3	<b>0.47</b>	0.41
HS3ST1	4p16	<b>0.46</b>	0.35
GPR87	3q24	<b>0.46</b>	0.35
TACSTD2	1p32	<b>0.46</b>	0.41
MUC16	19p13.2	<b>0.46</b>	0.37
FAM83A	8q24.13	<b>0.45</b>	0.34
LAMC2	1q25-q31	<b>0.45</b>	0.32
B3GNT3	19p13.1	<b>0.45</b>	0.40
CLDN7	17p13.1	<b>0.45</b>	0.44
ELF3	1q32.2	<b>0.44</b>	0.44
MIR205HG	1q32.2	<b>0.44</b>	0.37
PPL	16p13.3	<b>0.44</b>	0.40
MPZL2	11q24	<b>0.44</b>	0.43
TMPRSS4	11q23.3	<b>0.44</b>	0.46
C6ORF132	6p21.1	<b>0.43</b>	0.36
FGFBP1	4p15.32	<b>0.43</b>	0.38
IRF6	1q32.3-q41	<b>0.43</b>	0.44
LAMB3	1q32	<b>0.43</b>	0.31
CDH3	16q22.1	<b>0.43</b>	0.41
SPINT1	15q15.1	<b>0.43</b>	0.42
EHF	11p12	<b>0.43</b>	0.41
CYSRT1	9q34.3	<b>0.42</b>	0.33
MACC1	7p21.1	<b>0.42</b>	0.38
MST1R	3p21.3	<b>0.42</b>	0.41
SERPINB5	18q21.33	<b>0.42</b>	0.39
TMEM30B	14q23.1	<b>0.42</b>	0.40
CLDN4	7q11.23	<b>0.41</b>	0.37
LIPH	3q27	<b>0.41</b>	0.36
ALS2CL	3p21.31	<b>0.41</b>	0.37
ITGB6	2q24.2	<b>0.41</b>	0.37

RAB25	1q22	<b>0.41</b>	0.41
CNKSR1	1p36.11	<b>0.41</b>	0.43
TSPAN1	1p34.1	<b>0.41</b>	0.36
CEACAM6	19q13.2	<b>0.41</b>	0.37
KLK10	19q13	<b>0.41</b>	0.37
UCA1	19p13.12	<b>0.41</b>	0.32
CXCL16	17p13	<b>0.41</b>	0.35
ELMO3	16q22.1	<b>0.41</b>	0.44
PRSS8	16p11.2	<b>0.41</b>	0.42
ST14	11q24-q25	<b>0.41</b>	0.40
TRIM29	11q23.3	<b>0.41</b>	0.37
GRHL2	8q22.3	<b>0.40</b>	0.40
PTK6	20q13.3	<b>0.40</b>	0.34
FLJ23867	1q25.2	<b>0.40</b>	0.31
TMC4	19q13.42	<b>0.40</b>	0.38
CDH1	16q22.1	<b>0.40</b>	0.39
SDR16C5	8q12.1	<b>0.39</b>	0.35
S100A14	1q21.3	<b>0.39</b>	0.38
GJB5	1p35.1	<b>0.39</b>	0.33
JUP	17q21	<b>0.39</b>	0.40
TMC5	16p12.3	<b>0.39</b>	0.42
SCGB1A1	11q12.3	<b>0.39</b>	0.34
MROH6	8q24.3	<b>0.38</b>	0.39
MAL2	8q23	<b>0.38</b>	0.41
ESRP1	8q22.1	<b>0.38</b>	0.42
GALNT3	2q24-q31	<b>0.38</b>	0.38
CBLC	19q13.2	<b>0.38</b>	0.40
FUT3	19p13.3	<b>0.38</b>	0.42
PKP3	11p15	<b>0.38</b>	0.39
EPHA1	7q34	<b>0.37</b>	0.39
AGR2	7p21.3	<b>0.37</b>	0.33
CDS1	4q21.23	<b>0.37</b>	0.37
S100P	4p16	<b>0.37</b>	0.36
ARL14	3q25.33	<b>0.37</b>	0.33
KRTCAP3	2p23.3	<b>0.37</b>	0.41
BIK	22q13.31	<b>0.37</b>	0.38
SFN	1p36.11	<b>0.37</b>	0.41
TMEM125	1p34.2	<b>0.37</b>	0.44
C19ORF33	19q13.2	<b>0.37</b>	0.35
LSR	19q13.12	<b>0.37</b>	0.41
MISP	19p13.3	<b>0.37</b>	0.39
ESRP2	16q22.1	<b>0.37</b>	0.39
PAK6	15q14	<b>0.37</b>	0.37
KRT4	12q13.13	<b>0.37</b>	0.32

ANKRD22	10q23.31	<b>0.37</b>	0.40
MARVELD2	5q13.2	<b>0.36</b>	0.38
LAD1	1q25.1- q32.3	<b>0.36</b>	0.38
F11R	1q21.2- q21.3	<b>0.36</b>	0.44
CGN	1q21	<b>0.36</b>	0.42
ARHGEF16	1p36.3	<b>0.36</b>	0.43
KIAA1522	1p35.1	<b>0.36</b>	0.33
DMKN	19q13.12	<b>0.36</b>	0.34
STAP2	19p13.3	<b>0.36</b>	0.34
EVPL	17q25.1	<b>0.36</b>	0.38
ITGB4	17q25	<b>0.36</b>	0.36
MARVELD3	16q22.2	<b>0.36</b>	0.42
CCDC64B	16p13.3	<b>0.36</b>	0.38
KLF5	13q22.1	<b>0.36</b>	0.35
KRT6A	12q13.13	<b>0.36</b>	0.33
EXPH5	11q22.3	<b>0.36</b>	0.37
PLEKHA7	11p15.1	<b>0.36</b>	0.33
PRRG4	11p13	<b>0.36</b>	0.33
ADAP1	7p22.3	<b>0.35</b>	0.35
IL1RN	2q14.2	<b>0.35</b>	0.36
EPCAM	2p21	<b>0.35</b>	0.38
PVRL4	1q23.3	<b>0.35</b>	0.31
EPS8L1	19q13.42	<b>0.35</b>	0.39
PRRG2	19q13.33	<b>0.35</b>	0.43
FXYD3	19q13.12	<b>0.35</b>	0.37
CRB3	19p13.3	<b>0.35</b>	0.40
MYO5C	15q21	<b>0.35</b>	0.37
TC2N	14q32.12	<b>0.35</b>	0.38
PLEKHG3	14q23.3	<b>0.35</b>	0.35
FAM83H	8q24.3	<b>0.34</b>	0.39
FRK	6q21- q22.3	<b>0.34</b>	0.31
FAM110C	2p25.3	<b>0.34</b>	0.35
KDF1	1p36.11	<b>0.34</b>	0.40
KLK6	19q13.3	<b>0.34</b>	0.38
SPINT2	19q13.1	<b>0.34</b>	0.39
TTC9	14q24.2	<b>0.34</b>	0.32
FOXA1	14q21.1	<b>0.34</b>	0.36
TJP2	9q13-q21	<b>0.33</b>	0.31
ARHGEF5	7q35	<b>0.33</b>	0.33
MAPK13	6p21.31	<b>0.33</b>	0.32
ZNF165	6p21.3	<b>0.33</b>	0.41
ANXA3	4q21.21	<b>0.33</b>	0.30
B3GNT5	3q28	<b>0.33</b>	0.32

ZBED2	3q13.2	<b>0.33</b>	0.31
GRHL1	2p25.1	<b>0.33</b>	0.38
FERMT1	20p12.3	<b>0.33</b>	0.31
SPRR1A	1q21-q22	<b>0.33</b>	0.31
S100A9	1q21	<b>0.33</b>	0.33
PCSK9	1p32.3	<b>0.33</b>	0.34
CEACAM5	19q13.1- q13.2	<b>0.33</b>	0.33
KLK8	19q13	<b>0.33</b>	0.36
GNA15	19p13.3	<b>0.33</b>	0.32
KRT19	17q21.2	<b>0.33</b>	0.32
TNS4	17q21.2	<b>0.33</b>	0.41
PLEK2	14q23.3	<b>0.33</b>	0.32
DTX4	11q12.1	<b>0.33</b>	0.31
TSPAN15	10q22.1	<b>0.33</b>	0.34
CHMP4C	8q21.13	<b>0.32</b>	0.38
DAPP1	4q25-q27	<b>0.32</b>	0.32
PROM2	2q11.1	<b>0.32</b>	0.37
AIM1L	1p36.11	<b>0.32</b>	0.42
GRHL3	1p36.11	<b>0.32</b>	0.34
MYH14	19q13.33	<b>0.32</b>	0.41
TJP3	19p13.3	<b>0.32</b>	0.40
DSC2	18q12.1	<b>0.32</b>	0.32
LLGL2	17q25.1	<b>0.32</b>	0.40
IL18	11q23.1	<b>0.32</b>	0.32
OVOL1	11q13	<b>0.32</b>	0.40
CORO2A	9q22.3	<b>0.31</b>	0.34
TMEM184A	7p22.3	<b>0.31</b>	0.40
MAP7	6q23.3	<b>0.31</b>	0.33
IL20RA	6q23.3	<b>0.31</b>	0.37
DDR1	6p21.3	<b>0.31</b>	0.32
FAM83B	6p12.1	<b>0.31</b>	0.37
LAMP3	3q26.3- q27	<b>0.31</b>	0.36
OVOL2	20p11.23	<b>0.31</b>	0.41
KCNK1	1q42-q43	<b>0.31</b>	0.35
PTAFR	1p35- p34.3	<b>0.31</b>	0.34
FUT2	19q13.3	<b>0.31</b>	0.38
LRG1	19p13.3	<b>0.31</b>	0.32
ST6GALNAC1	17q25.1	<b>0.31</b>	0.43
GRB7	17q12	<b>0.31</b>	0.38
ATP2C2	16q24.1	<b>0.31</b>	0.42
PLA2G10	16p13.1- p12	<b>0.31</b>	0.39
SCNN1A	12p13	<b>0.31</b>	0.40
TMEM45B	11q24.3	<b>0.31</b>	0.38

EZR	6q25.3	<b>0.30</b>	0.31
ARAP2	4p14	<b>0.30</b>	0.31
CDCP1	3p21.31	<b>0.30</b>	0.30
PTPRU	1p35.3	<b>0.30</b>	0.30
KLC3	19q13	<b>0.30</b>	0.36
EPN3	17q21.33	<b>0.30</b>	0.39
ARHGAP27	17q21.31	<b>0.30</b>	0.35
FA2H	16q23	<b>0.30</b>	0.40

**Table 2: List of mRNA negatively correlated with MUC4.** Data were retrieved from 881 samples of Cancer Cell Line Encyclopedia (Novartis/Broad, Nature 2012). Correlation analysis was performed using cBioPortal.org online tool. 9 genes presented a Pearson's correlation lower than -0.3.

Correlated gene	cytoband	Pearson's correlation	Spearman's correlation
SLC35B4	7q33	<b>-0.30</b>	-0.32
IFFO1	12p13.3	<b>-0.30</b>	-0.36
TTC28	22q12.1	<b>-0.31</b>	-0.33
VKORC1	16p11.2	<b>-0.31</b>	-0.35
DIXDC1	11q23.1	<b>-0.31</b>	-0.31
ATP8B2	1q21.3	<b>-0.32</b>	-0.33
ST3GAL2	16q22.1	<b>-0.32</b>	-0.31
ZEB1	10p11.2	<b>-0.33</b>	-0.35
MTFR1L	1p36.11	<b>-0.34</b>	-0.35

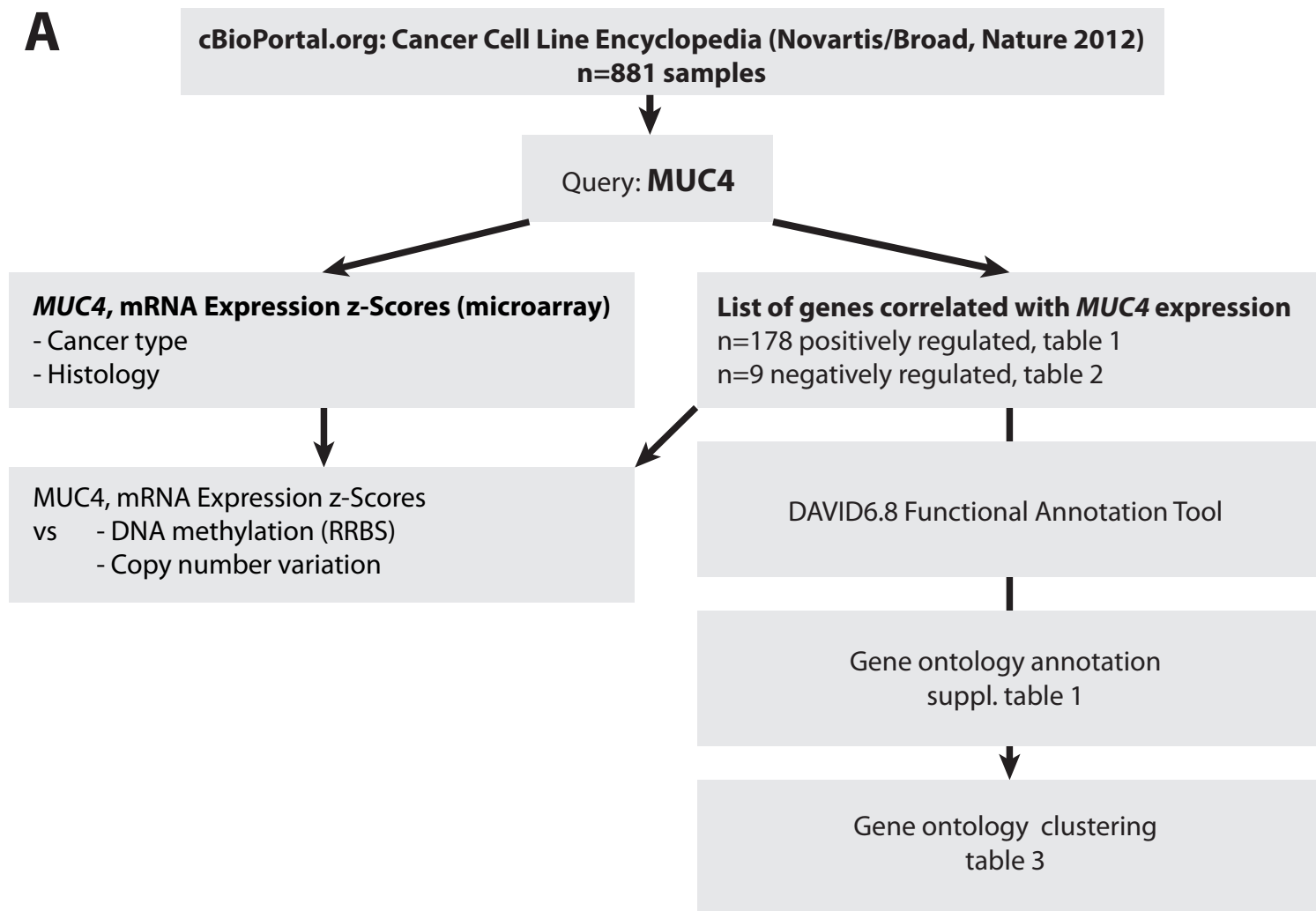
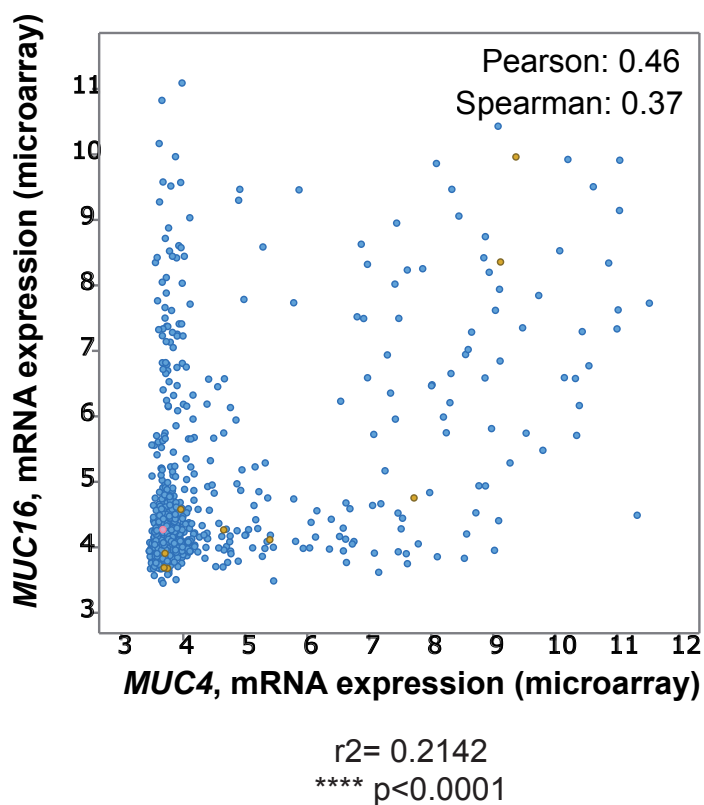
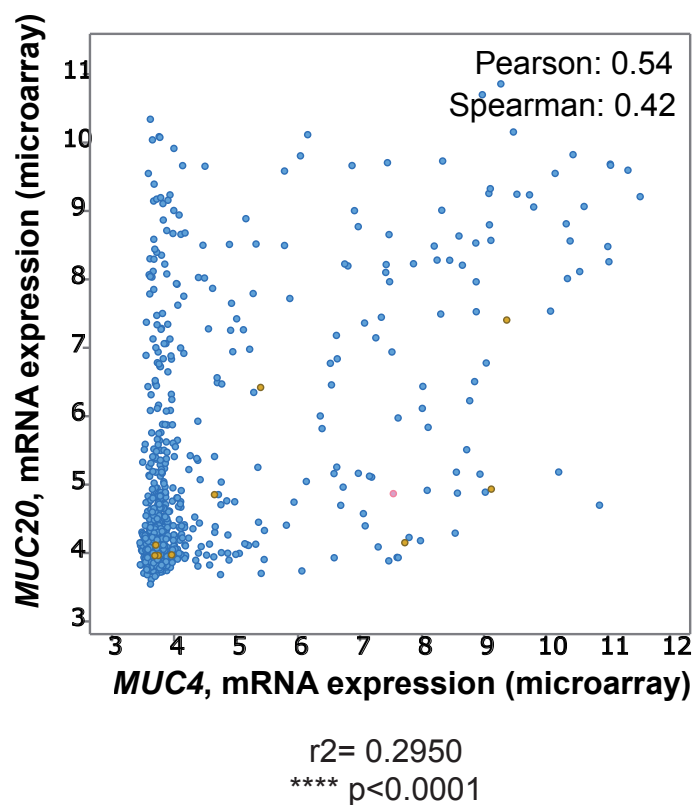
**Table 3: Gene ontology clustering on genes correlated with MUC4 expression.** Gene list was retrieved from 881 samples of Cancer Cell Line Encyclopedia (Baretina, Nature 2012). 187 genes that are positively (n=178) or negatively (n=9) correlated with MUC4 expression were selected. Functional Annotation and gene clustering were performed using David Functional Annotation Tool (<https://david.ncifcrf.gov/>).

Enrichment score	Gene Ontology terms and annotations	Count	P value
7.08	Cell-cell adherens junction	18	1.4E-8
	cadherin binding involved in cell-cell adhesion	17	2.0E-8
	cell-cell adhesion	14	2.2E-6
5.44	Tight junction	10	6.6E-8
	bicellular tight junction	10	1.4E-6
	Tight junction	9	8.1E-6
	bicellular tight junction assembly	5	2.4E-4
4.67	Pleckstrin homology-like domain	17	2.6E-6
	Pleckstrin homology domain	13	9.3E-6
	domain:PH	11	8.0E-5
	PH	12	1.1E-4
3.35	SH2 domain	8	9.1E-5
	domain:SH2	7	2.3E-4
	SH2 domain	7	3.9E-4
	SH2	6	4.8E-3
3.34	Glycoprotein	64	6.0E-5
	glycosylation site:N-linked (GlcNAc...)	61	1.1E-4
	disulfide bond	44	6.4E-4
	signal peptide	48	9.7E-4
	Disulfide bond	48	9.8E-4
	Signal	54	2.2E-3
2.76	topological domain:Cytoplasmic	53	8.1E-5
	Membrane	91	1.6E-4
	transmembrane region	66	8.5E-4
	topological domain:Extracellular	42	9.2E-4
	Transmembrane helix	66	7.2E-3
	Transmembrane	66	7.7E-3
	integral component of membrane	59	8.4E-2
2.6	domain:SH3	9	1.9E-4
	SH3 domain	9	6.5E-4
	Src homology-3 domain	8	4.4E-3
	SH3	6	6.9E-2
2.48	signal peptide	48	9.7E-4
	Secreted	31	2.0E-3
	extracellular region	25	1.9E-2
2.43	establishment of protein localization to plasma	6	4.9E-5
	membrane	5	3.0E-3
	cell adhesion molecule binding	4	3.5E-1
	actin cytoskeleton		
2.32	extracellular matrix organization	10	1.2E-4
	Epidermolysis bullosa, junctional, non-Herlitz type	3	2.8E-4
	Epidermolysis bullosa	4	2.8E-4
	hemidesmosome assembly	3	5.7E-3

	ECM-receptor interaction	4	2.9E-2
	Focal adhesion	5	7.2E-2
	PI3K-Akt signaling pathway	4	5.0E-1
2.19	Serine protease	8	2.5E-4
	Peptidase S1, trypsin family, active site	7	3.9E-4
	domain:Peptidase S1	7	4.7E-4
	active site:Charge relay system	9	5.3E-4
	Peptidase S1	7	9.1E-4
	Trypsin-like cysteine/serine peptidase domain	7	1.3E-3
	Tryp_SPc	7	1.6E-3
	extrinsic component of plasma membrane	4	1.7E-3
	Peptidase S1A, chymotrypsin-type	6	4.1E-3
	serine-type endopeptidase activity	8	1.2E-2
	serine-type peptidase activity	4	2.3E-2
	Protease	8	2.0E-1
	Zymogen	4	2.9E-1
	proteolysis	7	3.5E-1
	Hydrolase	13	8.1E-1
1.74	CP2 transcription factor	3	1.3E-3
	region of interest:Transcription activation	3	3.5E-3
	chromatin DNA binding	3	1.1E-1
	sequence-specific DNA binding	8	2.3E-1
1.69	O-glycan processing	6	2.7E-4
	Glycosphingolipid biosynthesis - lacto and neolacto series	4	9.8E-4
	protein glycosylation	6	4.7E-3
	Glycosyltransferase	7	1.8E-2
	topological domain:Luminal	10	2.1E-2
	Golgi cisterna membrane	4	3.6E-2
	Signal-anchor	9	4.8E-2
	Golgi apparatus	12	1.0E-1
	Golgi membrane	9	2.0E-1
	Metabolic pathways	9	7.5E-1
1.51	Rho guanyl-nucleotide exchange factor activity	5	6.4E-3
	regulation of Rho protein signal transduction	5	7.6E-3
	Dbl homology (DH) domain	4	2.9E-2
	domain:DH	3	1.3E-1
	RhoGEF	3	1.6E-1

**Table 4: Hazard-ratio and survival analysis of high and low risk in TCGA tumor databases.** Hazard ratio and p-value were determined using SurvExpress tool (<http://bioinformatica.mty.itesm.mx/SurvExpress>). Risk groups were determined using the optimization algorithm (maximize) from the ordered prognostic index (higher values of MUC4 expression for higher risk).

Database	N; low vs risk group	Hazard ratio [95% CI]	P value
Bladder – BLCA–TCGA–Bladder Urothelial Carcinoma–July 2016	N=388; 251 vs 137	1.48 [1.09 ; 2]	<b>p=0.01191</b>
Breast – BRCA–TCGA Breast invasive carcinoma – July 2016	N=962; 831 vs 131	1.06 [0.67 ; 1.67]	p=0.8038
Cervical – CESC–TCGA Cervical squamous cell carcinoma and endocervical adenocarcinoma July 2016	N=191; 147 vs 44	1.55 [0.76 ; 3.17]	p=0.2275
Colon – COADREAD – TCGA Colon and Rectum adenocarcinoma June 2016	N=466; 417 vs 49	2.1 [1.19 ; 3.71]	<b>p=0.01061</b>
Esophagus – ESCA – TCGA Esophageal carcinoma June 2016	N=184; 137 vs 47	0.68 [0.4 ; 1.15]	p=0.1468
Head–Neck – HNSC – TCGA Head and Neck squamous cell carcinoma June 2016	N=502; 107 vs 395	1.26 [0.88 ; 1.78]	p=0.204
Hematologic – Acute Myeloid Leukemia TCGA	N=168; 146 vs 22	1.59 [0.97 ; 2.62],	p=0.06818
Kidney – KIPAN – TCGA Kidney PAN cancer TCGA June 2016	N=792; 555 vs 237	0.94 [0.7 ; 1.26]	p=0.6711
Kidney – KIRC – TCGA – Kidney renal clear cell carcinoma	N=415; 256 vs 159	0.98 [0.7 ; 1.37]	p=0.9115
Kidney – KIRP – TCGA Kidney renal papillary cell carcinoma June 2016	N=278; 248 vs 30	1.24 [0.52 ; 2.94]	p=0.6322
Liver – TCGA–Liver–Cancer	N=304; 137 vs 167	1.4 [0.97 ; 2.03]	p=0.07012
Lung ADK– LUAD – TCGA – Lung adenocarcinoma June 2016	N=475; 410 vs 65	1.7 [1.14 ; 2.52]	<b>p=0.008963</b>
Lung Squamous– LUSC – TCGA – Lung squamous cell carcinoma June 2016	N=175; 59 vs 116	1.69 [1.03 ; 2.78],	<b>p=0.03798</b>
Ovarian – Ovarian serous cystadenocarcinoma TCGA	N=578; 390 vs 188	1.33 [1.05 ; 1.69]	<b>p=0.01908</b>
Pancreatic – PAAD – TCGA – Pancreatic adenocarcinoma	N=176; 27 vs 149	3.94 [1.81 ; 8.61]	<b>p=0.0005756</b>
Prostate – PRAD – TCGA – Prostate adenocarcinoma June 2016	N=497; 328 vs 169	1.99 [0.57 ; 6.88],	p=0.2793
Skin – SKCM–TCGA Skin Cutaneous Melanoma July 2016	N=334; 312 vs 23	1.87 [1.08 ; 3.23]	<b>p=0.0262</b>
Stomach – STAD – TCGA – Stomach adenocarcinoma June 2016	N=352; 306 vs 46	1.58 [1 ; 2.51],	<b>p=0.04958</b>
Testis – TGCT – TCGA – Testicular Germ Cell Tumors	N=133; 93 vs 40	5.56 [0.57 ; 54.52]	p=0.1407
Thymus – THYM – TCGA – Thymoma June 2016	N=118; 90 vs 28	1.92 [0.48 ; 7.77]	P=0.3588
Thyroid – THCA – TCGA – Thyroid carcinoma – June 2016	N=247; 45 vs 202	1.98 [0.69 ; 5.64],	p=0.2019

**A****B****C**

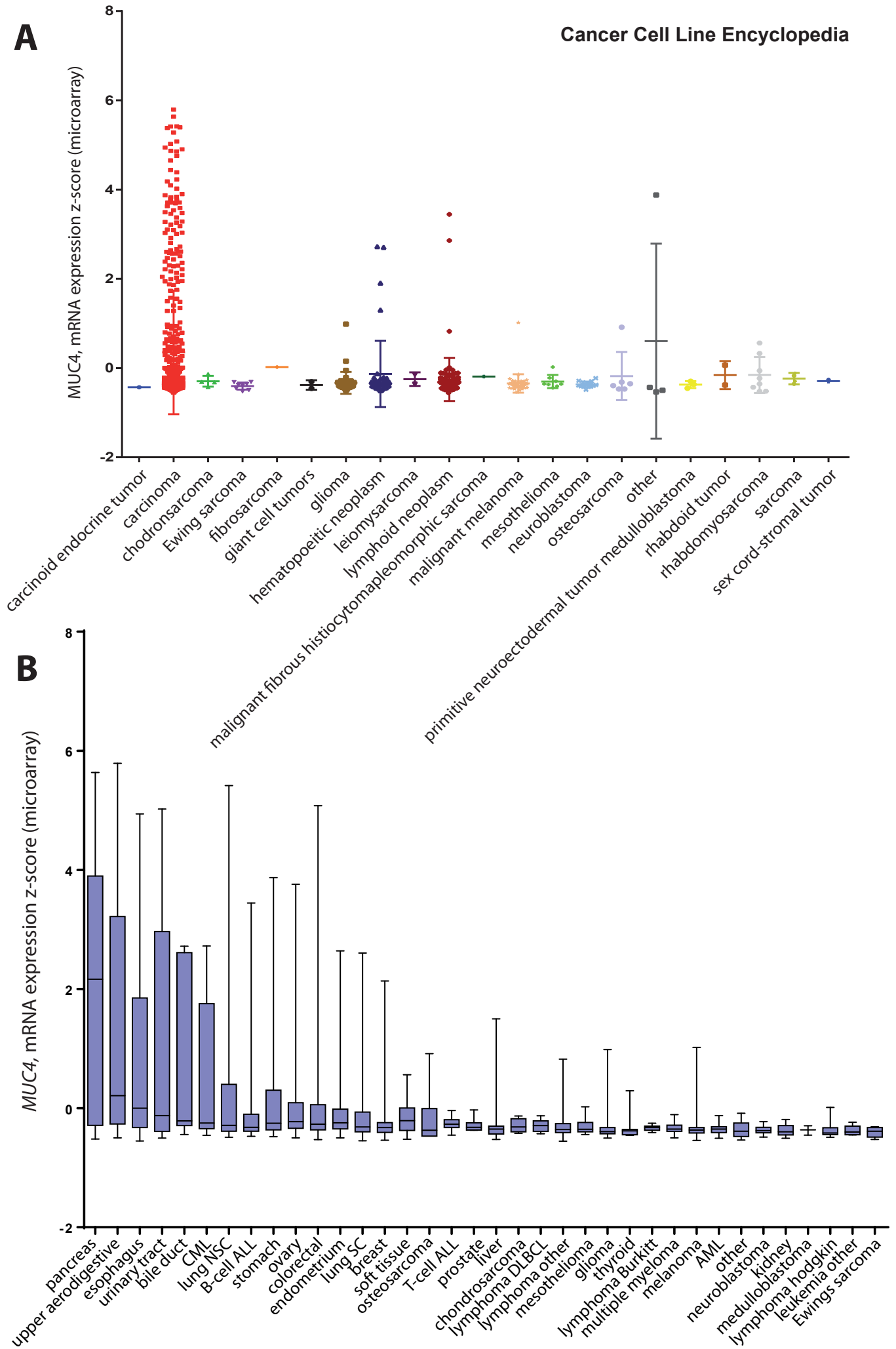


Figure 3

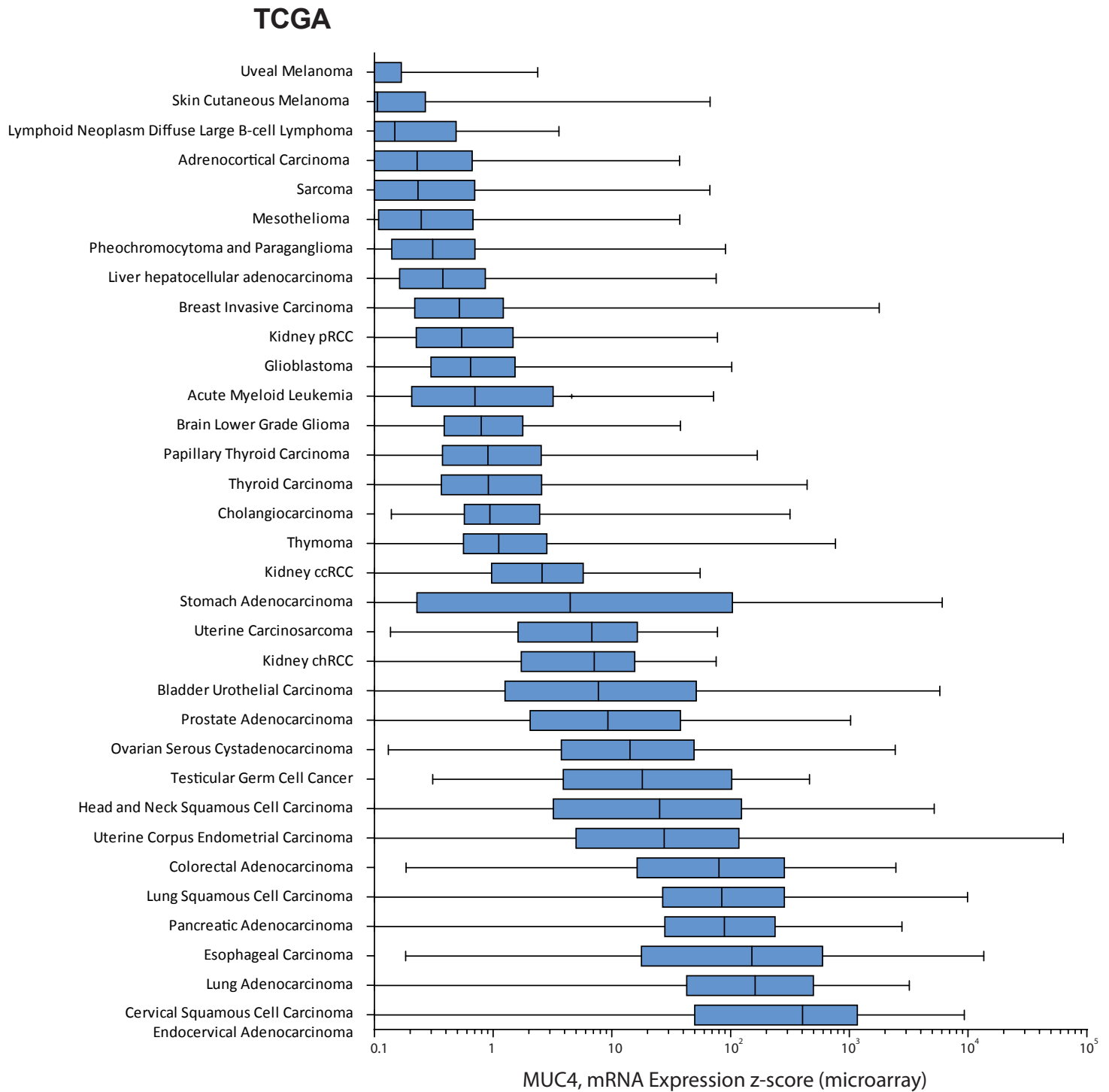
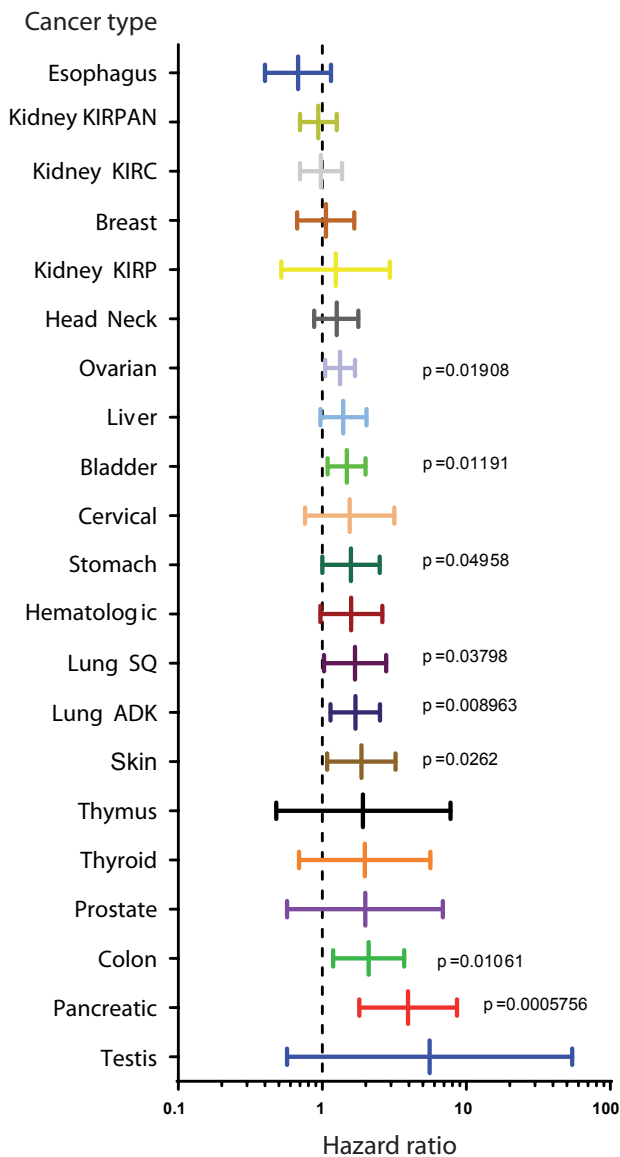


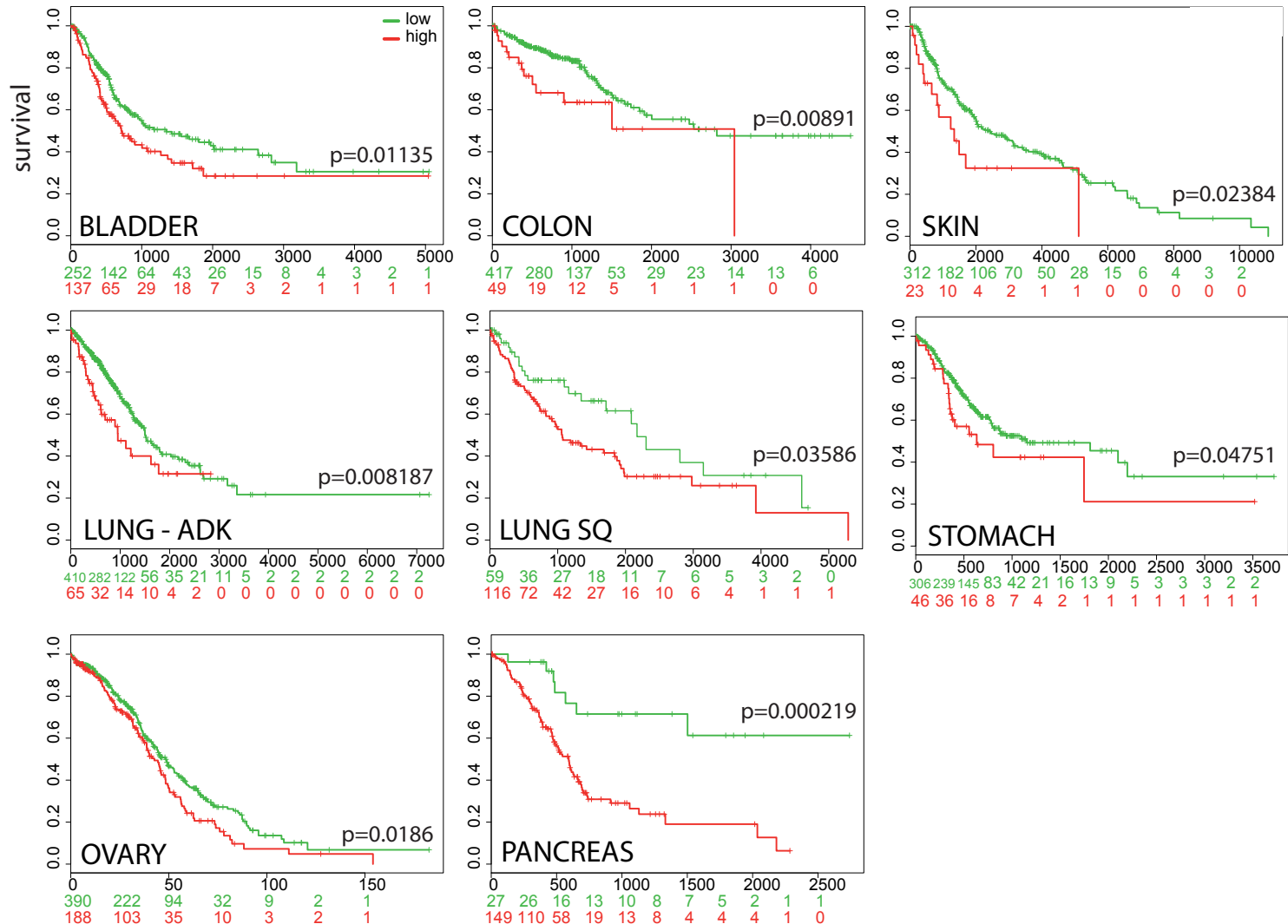


Figure 5

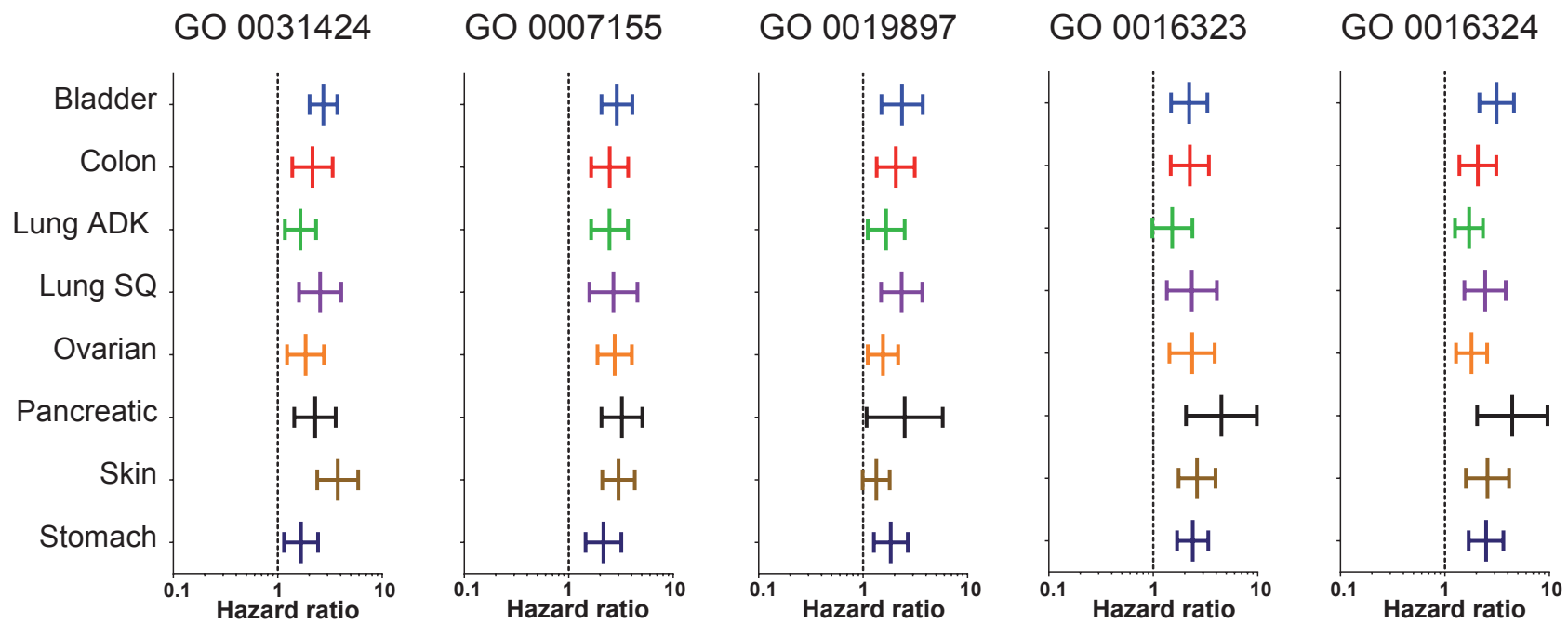
A



B

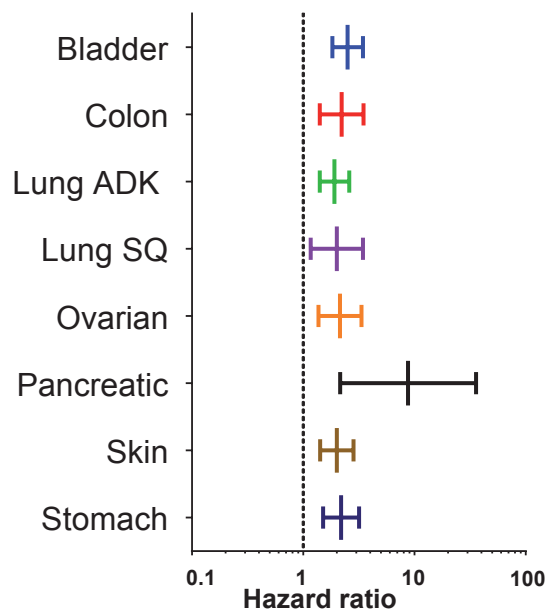


**A**

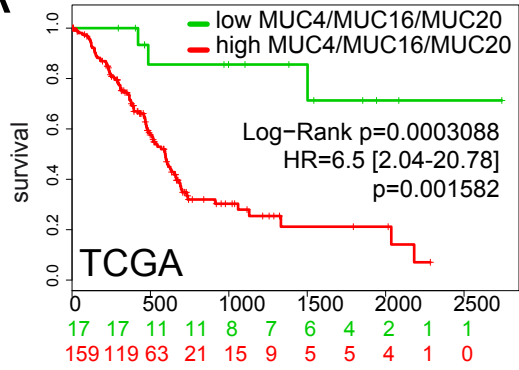


**B**

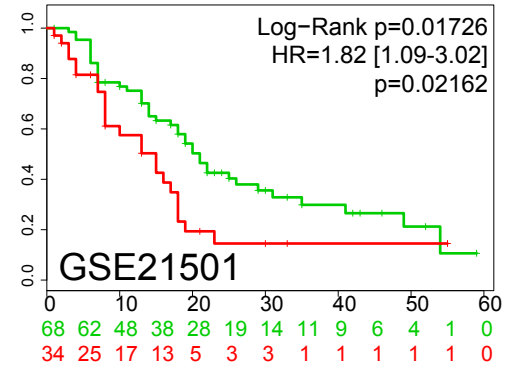
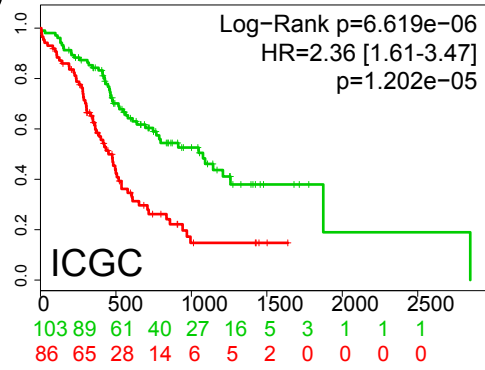
Top genes: ADGRF1, LCN2, MUC20, C1ORF116, SCEL, STEAP4



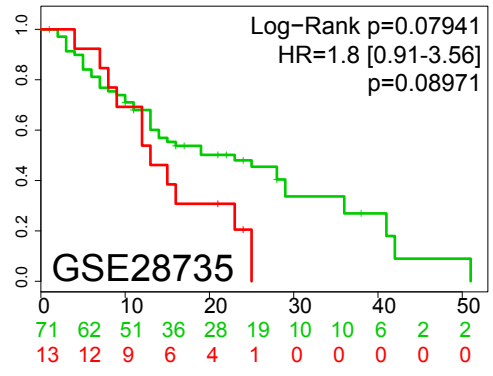
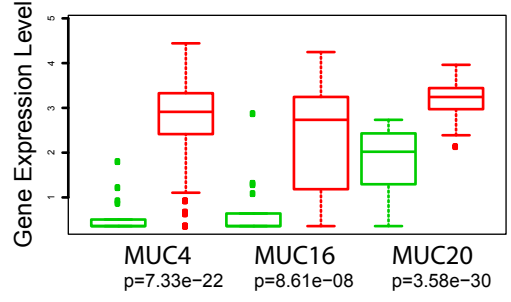
**A**

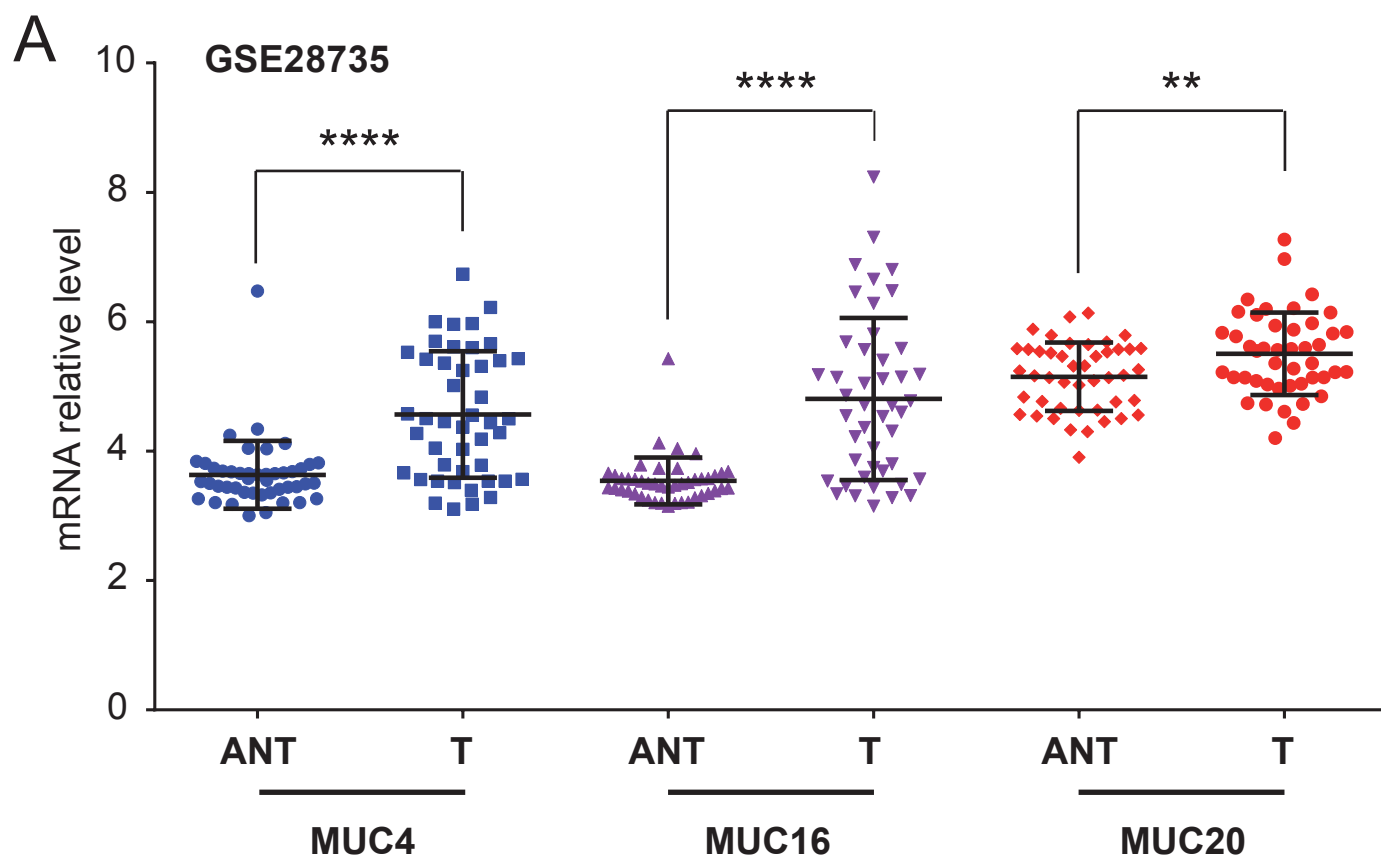


**C**



**B**



**B****ROC curve: MUC4, MUC16, MUC20**