

Supplementary Material:

Gene Expression Prediction using Low-Rank Matrix Completion

Arnav Kapur^{1,*}

Kshitij Marwah¹

Gil Alterovitz^{1,2}

¹Biomedical Cybernetics Laboratory, Harvard Medical School, Boston, MA 02115 and ²Department of Health Science and Technology/Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139.

Abstract

In this supplement, we describe the source of gene expression data used to evaluate low-rank prediction, provide additional discussion and results on additional studies.

S.1 Gene Expression Data

Our evaluation set consisted of gene expression datasets sourced from the curated expression data (GDS) repository of the Gene Expression Omnibus (NCBI GEO), with samples within a dataset referring to a common Platform, and curated high-throughput sequencing data from ArrayExpress. The expression values of each Sample within a curated dataset were transformed on a log2 scale. Failure of prediction for select datasets was indicated when successive iterates had an increasing error, implying divergence. This was indicated when the Omega error increased beyond a set threshold. The predictions were performed on a Windows® 8.0 desktop CPU with an Intel® Core™ i5-3470 processor (3.40 GHz), and 16 GB RAM. The evaluation set included studies performed on the following microarray platforms:

- Affymetrix Human Genome U95A Array
- Affymetrix Human Genome U95 Version 2 Array
- Affymetrix Human Genome U95B Array
- Affymetrix GeneChip Human Genome HG-U133 Plus 2 Array
- Affymetrix Human Genome U133 Plus 2.0 Array
- Affymetrix Human Genome U133A Array
- Affymetrix Human Genome U133A 2.0 Array
- Affymetrix Human Genome U133B Array
- Affymetrix HT HG-U133+ PM Array Plate
- Affymetrix Human Gene 1.0 ST Array
- Affymetrix Human Gene 1.1 ST Array
- ABI Human Genome Survey Microarray Version 2
- Affymetrix GeneChip Human Genome U133A 2.0 Array
- Agilent-012097 Human 1A Microarray (V2) G4110B
- Agilent-012391 Whole Human Genome Oligo Microarray G4112A
- Illumina HumanHT-12 V3.0 expression beadchip
- Illumina HumanRef-8 v2.0 expression beadchip
- Illumina HumanRef-8 v3.0 expression beadchip

- NCI_UniGEM2_HCC
- NHICU Human 19K v1.0
- NuGO array (human) NuGO_Hs1a520180
- Research Genetics (Invitrogen) - GF211 Microarray Filter
- Rosetta/Merck Human RSTA Custom Affymetrix 2.0 microarray

The protocol hardware for RNA-seq based studies were variants of the Illumina sequencing kits:

- Illumina HiSeq 1500
- Illumina HiSeq 2000
- Illumina HiSeq 2500
- Illumina Genome Analyzer II
- Illumina Genome Analyzer IIx
- Illumina HiScanSQ

S.2 Differential Expression and Bayesian Network Analysis

We present results of identifying differentially expressed genes in three additional studies, and the comparison of the analysis with low-rank predicted and original datasets. Differentially expressed genes were identified in both original and low-rank predicted profiles of acute myeloid leukemia (Stirewalt *et al.*, 2008), non-small cell lung carcinoma (Lu *et al.*, 2010) and the oral mucosa of smokers (Boyle *et al.*, 2010).

For each dataset, we performed the predictions at different observabilities, by randomly selecting the checkpoint expression values at each observability. The genes were then ranked according to log2 fold change in both the original and constructed datasets, using the Bayes moderated t-test.

This supplement includes disease classification results on individual test examples trained with the original Pancreatic Ductal Adenocarcinoma [4] and low-rank predicted dataset with 50% missing values, demonstrating the ability to derive prognostic and diagnostic signatures from low-rank predicted gene expression data.

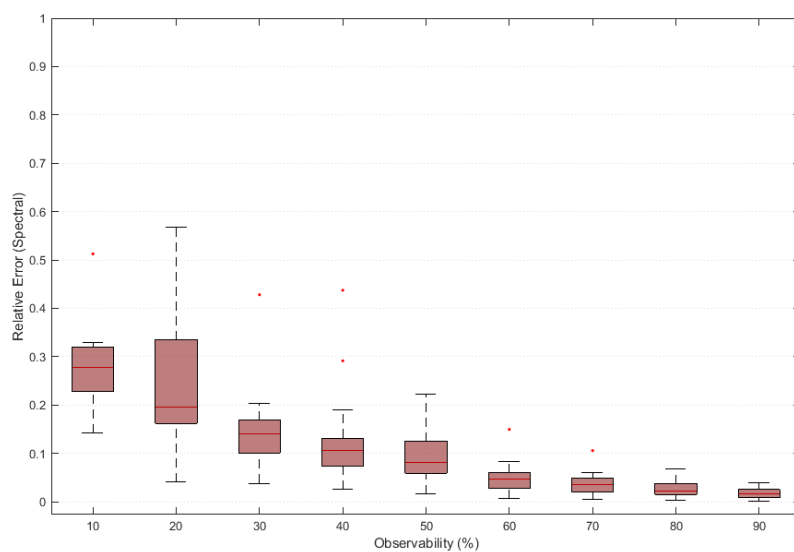


Figure S.1: Boxplot represents spectral relative error in prediction of converged RNA-seq datasets evaluated on a combined total of 897 samples, sourced from 14 gene expression studies. The observability quantifies fraction of measurements known prior to prediction. Each dataset was replicated into 9 equivalents with varying observability (10%-90%). Edges of box represent 25% and 75% coverage and the whiskers extend to 99.3% coverage.

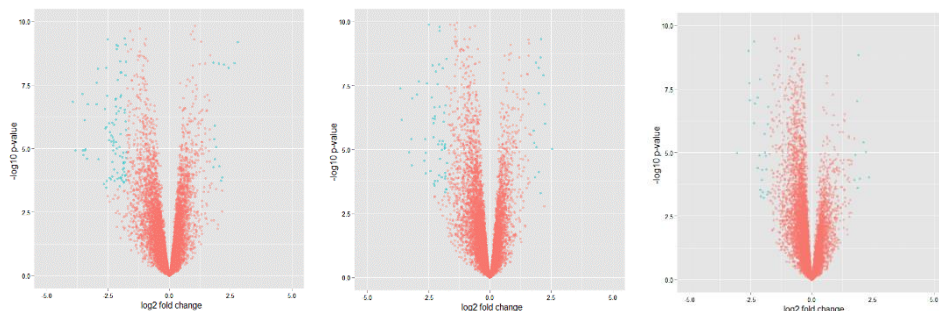


Figure S.2: Comparison of differential analysis on original and predicted acute myeloid leukemia (AML) datasets. Volcano plots represent differentially expressed in original AML dataset (leftmost), low-rank predicted dataset with 40% values unknown, and with 50% values unknown (rightmost).

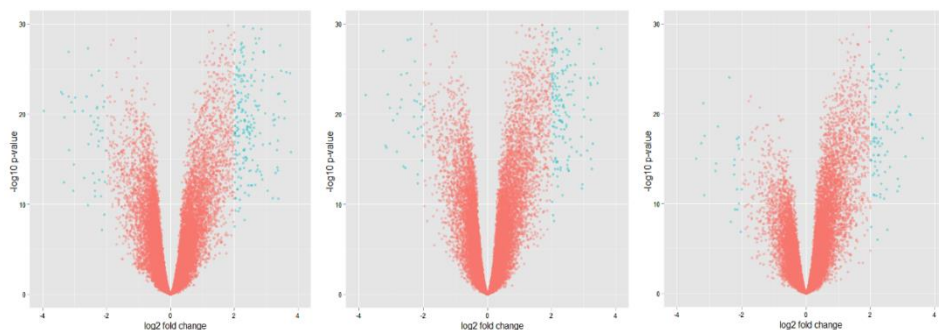


Figure S.3: Comparison of differential analysis on original and predicted non-small cell lung carcinoma (NSCLC) datasets. Volcano plots represent differentially expressed in original NSCLC dataset (leftmost), low-rank predicted dataset with 40% values unknown, and with 70% values unknown (rightmost).

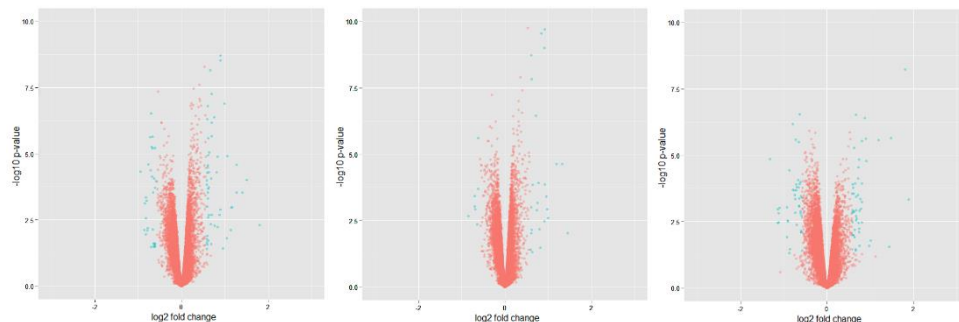


Figure S.4: Comparison of differential analysis on original and predicted oral mucosa datasets. Volcano plots represent differentially expressed in the original oral mucosa of smokers when compared with that of non-smokers (leftmost), low-rank predicted dataset with 40% values unknown, and with 70% values unknown (rightmost).

Table S.1.A: Top unique differentially expressed genes in non-small cell lung carcinoma when ranked according to log2-fold-change in **(a)** original dataset, **(b)** predicted dataset on known checkpoint values with 60% observability, and **(c)** data with known checkpoint values without prediction at 60% observability. Significance is demonstrated by adjusted P-values for fold change in every gene by using eBayes with Benjamini–Hochberg correction.

Gene Ranking	Original Dataset				Predicted Dataset (60%)				Checkpoint Dataset (60%)			
	Probe ID	Symbol	log FC	Adj. P-Value	Probe ID	Symbol	log FC	Adj. P-Value	Probe ID	Symbol	log FC	Adj. P-Value
1.	210081_at	<i>AGER</i>	4.397615	$< 1 \times 10^{-10}$	210081_at	<i>AGER</i>	4.281678	$< 1 \times 10^{-10}$	228698_at	<i>SOX7</i>	3.653878	0.0577
2.	209470_s_at	<i>GPM6A</i>	4.198425	$< 1 \times 10^{-10}$	232578_at	<i>CLDN18</i>	4.097738	$< 1 \times 10^{-10}$	217414_x_at	<i>NA</i>	3.554961	0.3019
3.	232578_at	<i>CLDN18</i>	4.160882	$< 1 \times 10^{-10}$	209470_s_at	<i>GPM6A</i>	3.789067	$< 1 \times 10^{-10}$	219866_at	<i>CLIC5</i>	3.511534	0.0577
4.	206209_s_at	<i>CA4</i>	3.869421	$< 1 \times 10^{-10}$	203980_at	<i>FABP4</i>	3.541133	$< 1 \times 10^{-10}$	225274_at	<i>PCYOX1</i>	3.452415	0.2424
5.	204712_at	<i>WIF1</i>	3.773603	$< 1 \times 10^{-10}$	206209_s_at	<i>CA4</i>	3.513062	$< 1 \times 10^{-10}$	228766_at	<i>CD36</i>	3.418187	0.0891
6.	203980_at	<i>FABP4</i>	3.747368	$< 1 \times 10^{-10}$	209469_at	<i>GPM6A</i>	3.435685	$< 1 \times 10^{-10}$	230469_at	<i>RTKN2</i>	3.375043	0.0003
7.	230560_at	<i>STXBP6</i>	3.627475	$< 1 \times 10^{-10}$	230560_at	<i>STXBP6</i>	3.424782	$< 1 \times 10^{-10}$	205200_at	<i>NA</i>	3.359106	0.2563
8.	209469_at	<i>GPM6A</i>	3.615812	$< 1 \times 10^{-10}$	219230_at	<i>TMEM100</i>	3.419413	$< 1 \times 10^{-10}$	200902_at	<i>SEP15</i>	3.320577	0.2563

Table S.1.B Top unique differentially expressed genes in non-small cell lung carcinoma when ranked according to log2-fold-change in **(a)** original dataset, **(b)** predicted dataset with 30% observability, and **(c)** data with known checkpoint values without prediction at 30% observability.

Gene Ranking	Original Dataset				Predicted Dataset (30%)				Checkpoint Dataset (30%)			
	Probe ID	Symbol	log FC	Adj. P-Value	Probe ID	Symbol	log FC	Adj. P-Value	Probe ID	Symbol	log FC	Adj. P-Value
1.	210081_at	<i>AGER</i>	4.397615	$< 1 \times 10^{-10}$	214387_x_at	<i>SFTPC</i>	3.615717	$< 1 \times 10^{-10}$	211813_x_at	<i>DCN</i>	3.528874	0.000616
2.	209470_s_at	<i>GPM6A</i>	4.198425	$< 1 \times 10^{-10}$	230560_at	<i>STXBP6</i>	3.239639	$< 1 \times 10^{-10}$	200716_x_at	<i>NA</i>	3.259112	0.003868
3.	232578_at	<i>CLDN18</i>	4.160882	$< 1 \times 10^{-10}$	210081_at	<i>AGER</i>	3.213539	$< 1 \times 10^{-10}$	213048_s_at	<i>SET</i>	3.014932	0.002372
4.	206209_s_at	<i>CA4</i>	3.869421	$< 1 \times 10^{-10}$	209613_s_at	<i>ADH1B</i>	3.086986	$< 1 \times 10^{-10}$	201716_at	<i>SNXI</i>	2.922509	0.000319
5.	204712_at	<i>WIF1</i>	3.773603	$< 1 \times 10^{-10}$	209074_s_at	<i>FAM107A</i>	3.020343	$< 1 \times 10^{-10}$	212537_x_at	<i>NA</i>	2.890575	0.008841
6.	203980_at	<i>FABP4</i>	3.747368	$< 1 \times 10^{-10}$	217546_at	<i>MTIM</i>	2.985399	$< 1 \times 10^{-10}$	202029_x_at	<i>RPL38</i>	2.874477	0.008984
7.	230560_at	<i>STXBP6</i>	3.627475	$< 1 \times 10^{-10}$	209470_s_at	<i>GPM6A</i>	2.943318	$< 1 \times 10^{-10}$	206621_s_at	<i>EIF4H</i>	2.853773	0.003253
8.	209469_at	<i>GPM6A</i>	3.615812	$< 1 \times 10^{-10}$	242009_at	<i>SLC644</i>	2.93128	$< 1 \times 10^{-10}$	212294_at	<i>GNG12</i>	2.837949	0.000915

Table S.2.A Top unique differentially expressed genes in oral mucosa of smokers when ranked according to log2-fold-change in **(a)** original dataset, **(b)** predicted dataset with 60% observability, and **(c)** data with known checkpoint values without prediction at 60% observability.

Gene Ranking	Original Dataset				Predicted Dataset (60%)				Checkpoint Dataset (60%)			
	Probe ID	Symbol	log FC	Adj. P-Value	Probe ID	Symbol	log FC	Adj. P-Value	Probe ID	Symbol	log FC	Adj. P-Value
1.	202437_s_at	<i>CYP1B1</i>	2.18598	$< 1 \times 10^{-10}$	202437_s_at	<i>CYP1B1</i>	1.753346	2.06×10^{-10}	226188_at	<i>LGALS1</i>	3.595662	0.9215
2.	231628_s_at	<i>SERPINB6</i>	1.801147	9.96×10^{-2}	231628_s_at	<i>SERPINB6</i>	1.447066	1.26×10^{-1}	219554_at	<i>RHCG</i>	3.582851	0.9428
3.	202436_s_at	<i>CYP1B1</i>	1.786575	$< 1 \times 10^{-10}$	202436_s_at	<i>CYP1B1</i>	1.380508	5.81×10^{-7}	203303_at	<i>DYNLT3</i>	3.533785	0.9215
4.	1553454_at	<i>RPTN</i>	1.501974	1.44×10^{-2}	1553454_at	<i>RPTN</i>	1.329623	8.73×10^{-3}	203534_at	<i>LSMI</i>	3.528252	0.6176
5.	205916_at	<i>SI00A7</i>	1.409739	2.65×10^{-2}	206198_s_at	<i>CEACAM7</i>	1.188474	8.73×10^{-3}	212284_x_at	<i>TPT1</i>	3.450075	0.9428
6.	202435_s_at	<i>CYP1B1</i>	1.333991	5.57×10^{-9}	203290_at	<i>HLA-DQA1</i>	0.999711	7.10×10^{-2}	200741_s_at	<i>RPS27</i>	3.387568	0.9428
7.	206198_s_at	<i>CEACAM7</i>	1.27266	2.68×10^{-2}	207878_at	<i>KRT76</i>	0.978685	5.15×10^{-2}	222466_s_at	<i>MRPL42</i>	3.383394	0.5249
8.	207720_at	<i>LOR</i>	1.269648	6.74×10^{-3}	205916_at	<i>SI00A7</i>	0.959343	3.08×10^{-2}	209351_at	<i>KRT14</i>	3.38324	0.9428

Table S.2.B: Top unique differentially expressed genes in oral mucosa of smokers when ranked according to log2-fold-change in **(a)** original dataset, **(b)** predicted dataset with 40% observability, and **(c)** data with known checkpoint values without prediction at 40% observability.

Gene	Original Dataset				Predicted Dataset (40%)				Checkpoint Dataset (40%)			
	Probe ID	Symbol	log FC	Adj. P-Value	Probe ID	Symbol	log FC	Adj. P-Value	Probe ID	Symbol	log FC	Adj. P-Value
Ranking												
1.	202437_s_at	<i>CYP1B1</i>	2.18598	$< 1 \times 10^{-10}$	1553454_at	<i>RPTN</i>	1.887365	4.46×10^{-2}	1553538_s_at	<i>COX1</i>	4.100917	0.923525
2.	231628_s_at	<i>SERPINB6</i>	1.801147	9.96×10^{-2}	202437_s_at	<i>CYP1B1</i>	1.807883	1.61×10^{-4}	221700_s_at	<i>UBA52</i>	3.968607	0.70759
3.	202436_s_at	<i>CYP1B1</i>	1.786575	$< 1 \times 10^{-10}$	202436_s_at	<i>CYP1B1</i>	1.484648	5.17×10^{-3}	210927_x_at	<i>JTB</i>	3.898795	0.644659
4.	1553454_at	<i>RPTN</i>	1.501974	1.44×10^{-2}	231628_s_at	<i>SERPINB6</i>	1.439264	2.30×10^{-1}	1553588_at	<i>NA</i>	3.764332	0.978179
5.	205916_at	<i>SI00A7</i>	1.409739	2.65×10^{-2}	202435_s_at	<i>CYP1B1</i>	1.191601	5.17×10^{-3}	218990_s_at	<i>SPRR3</i>	3.547704	0.981083
6.	202435_s_at	<i>CYP1B1</i>	1.333991	5.57×10^{-9}	204409_s_at	<i>EIF1AY</i>	1.12443	3.28×10^{-1}	200024_at	<i>RPS5</i>	3.529034	0.973445
7.	206198_s_at	<i>CEACAM7</i>	1.27266	2.68×10^{-2}	206198_s_at	<i>CEACAM7</i>	1.008266	1.87×10^{-1}	217732_s_at	<i>ITM2B</i>	3.484147	0.978179
8.	207720_at	<i>LOR</i>	1.269648	6.74×10^{-3}	207356_at	<i>NA</i>	0.972234	2.23×10^{-1}	202403_s_at	<i>COL1A2</i>	3.42695	0.923525

Table S.3.A: Top unique differentially expressed genes in acute myeloid leukemia when ranked according to log2-fold-change in **(a)** original dataset, **(b)** predicted dataset with 60% observability, and **(c)** data with known checkpoint values without prediction at 60% observability.

Gene	Original Dataset				Predicted Dataset (60%)				Checkpoint Dataset (60%)			
	Probe ID	Symbol	log FC	Adj. P-Value	Probe ID	Symbol	log FC	Adj. P-Value	Probe ID	Symbol	log FC	Adj. P-Value
Ranking												
1.	215489_x_at	<i>HOMER3</i>	2.806378	2.57×10^{-8}	214575_s_at	<i>AZU1</i>	2.528969	0.000242	200674_s_at	<i>RPL32</i>	4.830232	0.655973
2.	206674_at	<i>FLT3</i>	2.79343	2.44×10^{-7}	205131_x_at	<i>CLEC11A</i>	2.246066	4.36×10^{-5}	208313_s_at	<i>SF1</i>	4.328534	0.603705
3.	210365_at	<i>NA</i>	2.704193	4.40×10^{-8}	206674_at	<i>FLT3</i>	2.235996	1.19×10^{-5}	203633_at	<i>CPT1A</i>	4.317194	0.603705
4.	204647_at	<i>HOMER3</i>	2.667401	9.41×10^{-7}	214651_s_at	<i>NA</i>	2.208055	0.011649	207783_x_at	<i>HUWE1</i>	4.244694	0.769227
5.	205899_at	<i>CCNA1</i>	2.415505	1.25×10^{-6}	215489_x_at	<i>HOMER3</i>	2.173949	1.60×10^{-6}	208645_s_at	<i>RPS14</i>	3.845344	0.783248
6.	206067_s_at	<i>WT1</i>	2.337608	6.39×10^{-8}	210365_at	<i>NA</i>	2.078888	1.66×10^{-7}	200926_at	<i>RPS23</i>	3.731364	0.835709
7.	201105_at	<i>LGALS1</i>	2.236192	1.01×10^{-6}	206067_s_at	<i>WT1</i>	2.056385	4.78×10^{-7}	211530_x_at	<i>HLA-G</i>	3.724911	0.655973
8.	205131_x_at	<i>CLEC11A</i>	2.150491	2.11×10^{-3}	205382_s_at	<i>CFD</i>	2.053737	0.004836	200775_s_at	<i>NA</i>	3.562129	0.769227

Table S.3.B: Top unique differentially expressed genes in acute myeloid leukemia when ranked according to log2-fold-change in **(a)** original dataset, **(b)** predicted dataset with 50% observability, and **(c)** data with known checkpoint values without prediction at 50% observability.

Gene	Original Dataset				Predicted Dataset (50%)				Checkpoint Dataset (50%)			
	Probe ID	Symbol	log FC	Adj. P-Value	Probe ID	Symbol	log FC	Adj. P-Value	Probe ID	Symbol	log FC	Adj. P-Value
Ranking												
1.	215489_x_at	<i>HOMER3</i>	2.806378	2.57×10^{-8}	214651_s_at	<i>NA</i>	2.350099	9.25×10^{-5}	200610_s_at	<i>NCL</i>	5.909409	0.110193
2.	206674_at	<i>FLT3</i>	2.79343	2.44×10^{-7}	209905_at	<i>NA</i>	2.203104	9.67×10^{-6}	209191_at	<i>TUBB6</i>	4.111017	0.24289
3.	210365_at	<i>NA</i>	2.704193	4.40×10^{-8}	210549_s_at	<i>CCL23</i>	2.115887	3.93×10^{-6}	200088_s_at	<i>NA</i>	3.990415	0.768825
4.	204647_at	<i>HOMER3</i>	2.667401	9.41×10^{-7}	209560_s_at	<i>DLK1</i>	1.957567	1.05×10^{-4}	201144_s_at	<i>EIF2S1</i>	3.921335	0.315863
5.	205899_at	<i>CCNA1</i>	2.415505	1.25×10^{-6}	215489_x_at	<i>HOMER3</i>	1.901405	1.41×10^{-9}	201406_at	<i>NA</i>	3.738085	0.797762
6.	206067_s_at	<i>WT1</i>	2.337608	6.39×10^{-8}	217996_at	<i>PHLDA1</i>	1.86339	2.5×10^{-4}	217907_at	<i>MRPL18</i>	3.687096	0.315863
7.	201105_at	<i>LGALS1</i>	2.236192	1.01×10^{-6}	204647_at	<i>HOMER3</i>	1.84943	9.62×10^{-8}	221766_s_at	<i>FAM46A</i>	3.659216	0.713662
8.	205131_x_at	<i>CLEC11A</i>	2.150491	2.11×10^{-3}	210365_at	<i>NA</i>	1.784583	1.23×10^{-5}	200961_at	<i>SEPHS2</i>	3.580836	0.315863

Table S.4: Comparison of class predictions and probability distributions of Bayesian networks built by employing an original pancreatic ductal adenocarcinoma gene expression dataset, dataset constructed using low-rank prediction with 50% observability, and dataset constructed using sampling values with a uniform distribution at 50% observability. The table lists the results of sections of test instances in a cross-validation evaluation.

Test Instance	Original			Low-rank prediction		Uniform distribution sampled	
	Actual	Classifier predicted	Prob. dist	Classifier predicted	Prob. dist	Classifier predicted	Prob. dist
1	PDAC tumor	PDAC tumor	0.008, 0.992	PDAC tumor	0.001, 0.999	Normal pancreatic	0.978, 0.022
2	PDAC tumor	Normal pancreatic	0.555, 0.445	Normal pancreatic	0.997, 0.003	Normal pancreatic	0.62, 0.38
3	PDAC tumor	PDAC tumor	0.039, 0.961	Normal pancreatic	0.685, 0.315	PDAC tumor	0.48, 0.52
4	PDAC tumor	PDAC tumor	0.038, 0.962	PDAC tumor	0.091, 0.909	PDAC tumor	0.315, 0.685
5	Normal pancreatic	Normal pancreatic	0.962, 0.038	Normal pancreatic	1, 0	Normal pancreatic	0.663, 0.337
6	Normal pancreatic	Normal pancreatic	0.995, 0.005	Normal pancreatic	1, 0	PDAC tumor	0.164, 0.836
7	Normal pancreatic	Normal pancreatic	1, 0	Normal pancreatic	1, 0	Normal pancreatic	0.931, 0.069
8	Normal pancreatic	Normal pancreatic	1, 0	Normal pancreatic	1, 0	PDAC tumor	0.038, 0.962
9	PDAC tumor	PDAC tumor	0.001, 0.999	PDAC tumor	0, 1	Normal pancreatic	0.583, 0.417
10	PDAC tumor	Normal pancreatic	0.633, 0.367	PDAC tumor	0, 1	Normal pancreatic	0.591, 0.409
11	PDAC tumor	PDAC tumor	0, 1	PDAC tumor	0, 1	PDAC tumor	0.078, 0.922
12	PDAC tumor	Normal pancreatic	0.883, 0.117	Normal pancreatic	0.998, 0.002	PDAC tumor	0.13, 0.87
13	Normal pancreatic	Normal pancreatic	0.896, 0.104	Normal pancreatic	0.998, 0.002	Normal pancreatic	0.566, 0.434
14	Normal pancreatic	Normal pancreatic	1, 0	Normal pancreatic	0.992, 0.008	PDAC tumor	0.253, 0.747
15	Normal pancreatic	Normal pancreatic	0.955, 0.045	Normal pancreatic	1, 0	PDAC tumor	0.384, 0.616
16	Normal pancreatic	Normal pancreatic	0.871, 0.129	Normal pancreatic	0.883, 0.117	PDAC tumor	0.206, 0.794
17	PDAC tumor	PDAC tumor	0.002, 0.998	PDAC tumor	0.017, 0.983	Normal pancreatic	0.959, 0.041
18	PDAC tumor	PDAC tumor	0, 1	PDAC tumor	0.005, 0.995	PDAC tumor	0.279, 0.721
19	PDAC tumor	PDAC tumor	0.002, 0.998	Normal pancreatic	0.595, 0.405	Normal pancreatic	0.704, 0.296
20	PDAC tumor	PDAC tumor	0.02, 0.98	PDAC tumor	0.002, 0.998	PDAC tumor	0.442, 0.558
21	Normal pancreatic	Normal pancreatic	0.999, 0.001	Normal pancreatic	0.751, 0.249	PDAC tumor	0.442, 0.558
22	Normal pancreatic	Normal pancreatic	1, 0	Normal pancreatic	1, 0	PDAC tumor	0.442, 0.558
23	Normal pancreatic	PDAC tumor	0.0670.933	PDAC tumor	0.106, 0.894	PDAC tumor	0.279, 0.721
24	Normal pancreatic	Normal pancreatic	1, 0	Normal pancreatic	1, 0	PDAC tumor	0.442, 0.558
25	PDAC tumor	PDAC tumor	0, 1	PDAC tumor	0, 1	PDAC tumor	0.032, 0.968
26	PDAC tumor	PDAC tumor	0.004, 0.996	PDAC tumor	0, 1	PDAC tumor	0.104, 0.896
27	PDAC tumor	PDAC tumor	0, 1	PDAC tumor	0.028, 0.972	PDAC tumor	0.312, 0.688
28	PDAC tumor	PDAC tumor	0, 1	PDAC tumor	0.001, 0.999	Normal pancreatic	0.909, 0.091
29	Normal pancreatic	PDAC tumor	0, 1	PDAC tumor	0, 1	Normal pancreatic	0.74, 0.26
30	Normal pancreatic	Normal pancreatic	1, 0	Normal pancreatic	1, 0	Normal pancreatic	0.516, 0.484
31	Normal pancreatic	PDAC tumor	0.003, 0.997	Normal pancreatic	0.991, 0.009	Normal pancreatic	0.957, 0.043
32	Normal pancreatic	Normal pancreatic	1, 0	Normal pancreatic	0.965, 0.035	PDAC tumor	0.173, 0.827

Abbreviations: PDAC, Pancreatic ductal adenocarcinoma; Prob. dist, probability distribution.

References

- Stirewalt, D. L., *et al.* (2008). Identification of genes with abnormal expression changes in acute myeloid leukemia. *Genes, Chromosomes and Cancer*, 47(1), 8-20.
- Lu, T. P., *et al.* (2010). Identification of a novel biomarker, SEMA5A, for non-small cell lung carcinoma in nonsmoking women. *Cancer Epidemiology Biomarkers & Prevention*, 19(10), 2590-2597.
- Boyle, J.O., *et al.* (2010). Effects of cigarette smoke on the human oral mucosal transcriptome. *Cancer Prevention Research*, 3(3), 266-278.
- Badea, L., *et al.* (2008). Combined Gene Expression Analysis of Whole-Tissue and Microdissected Pancreatic Ductal Adenocarcinoma identifies Genes Specifically Overexpressed in Tumor Epithelia. *Hepato-gastroenterology*, 55(88), 2016.