Supplemental tables of

Miriam Lohr, Birte Hellwig, Karolina Edlund, Johanna Mattsson, Johan Botling, Marcus Schmidt, Jan G. Hengstler, Patrick Micke, Jörg Rahnenführer:

Identification of sample annotation errors in gene expression datasets

Supp. Table 1: Overview of analyzed datasets. Tissue collections and gene array datasets analyzed by the male-female classifier, if available identified by their Gene Expression Omnibus (GEO) Series (GSE) number.

Type	Cohorts	Sample size (female/male)
Non-small cell lung cancer	GSE37745, Shedden, GSE31547, GSE29013, GSE14814, GSE4573, GSE31210, GSE19188, GSE31546, GSE10445	1338 (594 / 744)
Colon cancer	GSE33113, GSE12945, GSE31595, GSE4271, GSE1433, GSE17536, GSE17537	769 (358 / 411)
Other cancer	GSE5720, GSE4107, GSE42952, GSE34111, GSE31684	200 (64 / 136)
Non-cancer	GSE19027, GSE17913, GSE23343, GSE25462, GSE7821, GSE20950, GSE24427	408 (219 / 189)
Breast cancer	GSE11121, GSE2034, TRANSBIG (GSE7390/GSE6532), GSE16446, GSE20194, GSE20271, GSE22093, GSE23988	1373 (1373 / 0)
Ovarian cancer	Bild, GSE14764, GSE19829, GSE26712	426 (426 / 0)
Prostate cancer	GSE17951, GSE25136, GSE3325, GSE8218	399 (0 / 399)

Suppl. Table 2: **Detailed description of analyzed datasets.** Overview over the studied tissue collections and gene array data.

Cohort	# female	# male	# total	Type (disease or subject of study)
GSE37745	89	107	196	NSCLC
Shedden	220	223	443	NSCLC
GSE31547	36	14	50	NSCLC + controls
GSE29013	17	38	55	NSCLC
GSE14814	23	67	90	NSCLC
GSE4573	47	82	129	NSCLC
GSE31210	109	95	204	NSCLC
GSE19188	23	59	82	NSCLC
GSE31546	14	3	17	NSCLC
GSE10445	16	56	72	NSCLC
GSE4107	12	10	22	colorectal cancer
GSE33113	48	42	90	colorectal cancer
GSE31595	22	15	37	colorectal cancer
GSE12945	28	34	62	colorectal cancer
GSE14333	106	120	226	colorectal cancer
GSE17536	81	96	177	colorectal cancer
GSE17537	29	26	55	colorectal cancer
GSE4271	32	68	100	other cancer: glioma
GSE31684	25	68	93	other cancer: bladder
GSE34111	6	24	30	other cancer: gastrointestinal
GSE5720	24	30	54	other cancer: 9 different tissues
GSE42952	9	14	23	other cancer: pancreatic
GSE19027	11	48	59	bronchial epithelium of (non-) smokers with & without lung cancer
GSE17913	38	40	78	Smoking
GSE23343	7	10	17	insulin resistance/type 2 diabetes
GSE25462	28	22	50	insulin resistance/type 2 diabetes
GSE7821	28	12	40	healthy twins
GSE20950	27	12	39	insulin resistance/obesity
GSE24427	80	45	125	multiple sclerosis
GSE11121	200	0	200	breast cancer
GSE2034	286	0	286	breast cancer
TRANSBIG (GSE7390/ GSE6532)	280	0	280	breast cancer
GSE16446	114	0	114	breast cancer; chemo response
GSE20194	247	0	247	breast cancer; chemo response
GSE20271	139	0	139	breast cancer; chemo response
GSE22093	47	0	47	breast cancer; chemo response
GSE23988	60	0	60	breast cancer; chemo response
Bild	133	0	133	ovarian cancer
GSE14764	80	0	80	ovarian cancer

GSE19829	28	0	28	ovarian cancer
GSE26712	185	0	185	ovarian cancer
GSE17951	0	153	153	prostate cancer
GSE25136	0	79	79	prostate cancer
GSE3325	0	19	19	prostate cancer
GSE8218	0	148	148	prostate cancer

Suppl. Table 3: **Probe sets included in the male-female classifier**. Probe sets included into the male-female classifier, with corresponding cut points for sex evidence of samples.

Affymetrix ID	Gene	Chromosome	Cut point (99.9% quantile)	Sex evidence
221728_x_at	*XIST*	X	> 0.389	Female
214218_s_at	*XIST*	X	> 0.385	Female
201909_at	*RPS4Y1*	Y	> 0.431	Male
205000_at	*DDX3Y*	Y	> 0.276	Male

Suppl. Table 4: **Results of univariate Cox models.** Results of univariate Cox models for six NSCLC datasets. Comparison between significant genes ($p<0.01$) identified in the original cohort and significant genes identified in the reduced cohort after removal of misannotated and duplicated samples.

Dataset	No. of patients	No. of misanno-tations and duplications	No. of significant genes (original scenario)	Percentage of genes no longer significant after removal of the misannotated samples	Percentage of genes newly significant after removal of the misannotated samples
GSE37745	196	3	450	12.22	14.00
Shedden	443	14	1354	15.66	8.79
GSE29013	55	1	419	15.51	14.32
GSE4573	129	5	189	29.63	38.62
GSE31547	50	1	318	50.51	23.27
GSE19188	82	8	190	53.16	34.74