Electronic Supplementary Material (ESI) for Molecular Omics.

Table 1: Autoencoder architecture for independent datasets

Dataset	# genes	# non-zero genes	# perceptrons in each layer of Autoencoder
TCGA-5	60478	58462	58462 - 720 - 60 - 720 - 58462
Melanoma-B	23686	22831	22831 - 250 - 21 - 250 - 22831
Breast-B	57820	35148	35148 - 480 - 35 - 480 - 35148
Colorectal-B	34127	34127	34127 - 480 - 35 - 480 - 34127
BrPr-CTC	58347	53008	53008 - 720 - 60 - 720 - 53008
Breast-M	22656	22656	22656 - 250 - 21 - 250 - 22656

Table 2: Potential biomarkers for independent datasets

Dataset	Biomarker
TCGA-5	ENSG00000160862.11, ENSG00000189377.7, ENSG00000176153.11, ENSG00000159182.4, ENSG00000171747.7, ENSG00000157765.10, ENSG00000111341.8, ENSG00000146674.13, ENSG00000164825.3, ENSG00000130176.6, ENSG00000026025.12, ENSG00000094755.15, ENSG00000120885.18, ENSG00000158715.5, ENSG00000109846.6, ENSG00000143387.11, ENSG00000170323.7, ENSG00000158710.13, ENSG00000087086.12, ENSG00000105388.13, ENSG00000140988.14, ENSG00000084207.14, ENSG00000143320.7, ENSG00000272398.4, ENSG00000133392.15, ENSG00000168542.11, ENSG00000113140.9, ENSG00000086548.8, ENSG00000130513.6, ENSG00000134352.18, ENSG00000244468.1, ENSG00000142515.13, ENSG00000211895.4, ENSG00000197747.7, ENSG00000163191.5, ENSG00000096384.18, ENSG00000171858.16, ENSG00000187837.3, ENSG00000112306.7, ENSG00000111716.11, ENSG00000075624.12, ENSG00000118785.12, ENSG00000122026.9, ENSG00000115414.17, ENSG00000100097.10
Melanoma-B	TMSB4X, CD8A, S100B, RPS12, MTRNR2L8, MTRNR2L2, SERPINE2, EEF1A1, RPS5, RPLP1, TYR, ACTB, SERPINA3, GNB2L1, PMEL
Breast-B	'ENSG00000019582.10', 'ENSG00000075624.9', 'ENSG00000108298.5', 'ENSG00000111341.5', 'ENSG00000111669.10', 'ENSG00000112096.12', 'ENSG00000113140.6', 'ENSG00000133048.8', 'ENSG00000133110.10', 'ENSG00000134333.9', 'ENSG00000137673.4', 'ENSG00000140105.13', 'ENSG00000143185.3', 'ENSG00000156234.7', 'ENSG00000157601.9', 'ENSG00000158869.6', 'ENSG00000161016.11', 'ENSG00000164919.6', 'ENSG00000166710.13', 'ENSG00000167996.11', 'ENSG00000189403.10', 'ENSG00000197956.5', 'ENSG00000198886.2', 'ENSG00000198899.2', 'ENSG00000198938.2', 'ENSG00000205542.6', 'ENSG00000213639.5', 'ENSG00000215066.3', 'ENSG00000222414.1', 'ENSG00000229117.4', 'ENSG00000251562.3', 'ENSG00000106819.7', 'ENSG00000110848.4', 'ENSG00000212907.2', 'ENSG00000102265.7', 'ENSG00000198727.2', 'ENSG00000175898.4', 'ENSG00000143933.12', 'ENSG00000272379.1', 'ENSG00000173812.6', 'ENSG00000228253.1', 'ENSG00000198888.2'
Colorectal-B	A_23_P120660, A_33_P3234641, A_33_P3241582, A_23_P251593, A_23_P26713, A_33_P3375668, A_33_P3370461, A_23_P116694, A_33_P3587376, A_33_P3270852, A_23_P402751, A_33_P3296852, A_33_P3231005, A_33_P3336696, A_23_P356484, A_33_P3396434, A_33_P3315763, A_33_P3388491, A_32_P24581, A_24_P763243, A_33_P3293164, A_23_P26294
BrPr-CTC	'ENSG00000067066_SP100', 'ENSG00000075624_ACTB', 'ENSG00000080824_HSP90AA1', 'ENSG00000087086_FTL', 'ENSG00000087460_GNAS', 'ENSG00000092841_MYL6', 'ENSG00000106541_AGR2', 'ENSG00000110484_SCGB2A2', 'ENSG00000111341_MGP', 'ENSG00000112306_RPS12', 'ENSG00000115648_MLPH', 'ENSG00000120885_CLU', 'ENSG00000130066_SAT1', 'ENSG00000132475_H3F3B', 'ENSG00000140988_RPS2', 'ENSG00000142515_KLK3', 'ENSG00000142676_RPL11', 'ENSG00000145425_RPS3A', 'ENSG00000149273_RPS3', 'ENSG00000156508_EEF1A1', 'ENSG00000163359_COL6A3', 'ENSG00000164266_SPINK1', 'ENSG00000167751_KLK2', 'ENSG00000167996_FTH1', 'ENSG00000168028_RPSA', 'ENSG00000171345_KRT19', 'ENSG00000182774_RPS17', 'ENSG00000196531_NACA', 'ENSG00000197956_S100A6', 'ENSG00000198763_MT-ND2', 'ENSG00000198938_MT-CO3', 'ENSG00000205542_TMSB4X', 'ENSG00000209082_MT-TL1', 'ENSG00000234741_GAS5', 'ENSG00000280614_FP236383', 'ENSG00000281383_FP671120'
Breast-M	19203, 1297, 20762, 18587, 12576, 20811, 11865, 5728, 9056, 16112, 3618, 2633, 17035, 6645, 21755, 14128, 9378, 11377, 16909, 5254, 987, 6450, 5506, 12742, 18540, 11308, 17056, 6817, 13312, 1470, 1806

Table 3: Cancer-specific potential biomarkers for TCGA-5 dataset with binary classification performance

Class	# BM	Biomarkers	Acc. (%)
BRCA	20	'ENSG00000110484.6', 'ENSG00000172551.9', 'ENSG00000153002.10', 'ENSG00000181617.5', 'ENSG00000124935.3', 'ENSG00000143556.7', 'ENSG00000111341.8', 'ENSG00000189058.7', 'ENSG00000164692.16', 'ENSG00000211895.4', 'ENSG00000163220.10', 'ENSG00000108821.12', 'ENSG00000163191.5', 'ENSG00000075624.12', 'ENSG00000198840.2', 'ENSG00000198763.3', 'ENSG00000184009.8', 'ENSG00000160862.11', 'ENSG00000142541.15', 'ENSG00000087086.12'	95.46
COAD	17	'ENSG00000169344.14', 'ENSG00000212907.2', 'ENSG00000198886.2', 'ENSG00000198712.1', 'ENSG00000198727.2', 'ENSG00000118785.12', 'ENSG00000198888.2', 'ENSG00000198786.2', 'ENSG00000198695.2', 'ENSG00000019582.13', 'ENSG00000234745.8', 'ENSG00000228253.1', 'ENSG00000198763.3', 'ENSG00000198840.2', 'ENSG00000211895.4', 'ENSG00000211592.5', 'ENSG00000034510.5'	92.52
KIRC	20	'ENSG00000122852.13', 'ENSG00000168878.15', 'ENSG00000163220.10', 'ENSG00000164265.7', 'ENSG00000118785.12', 'ENSG00000185303.14', 'ENSG00000168542.11', 'ENSG00000198183.10', 'ENSG00000161055.3', 'ENSG00000184009.8', 'ENSG00000211598.2', 'ENSG00000019582.13', 'ENSG00000096088.15', 'ENSG00000163191.5', 'ENSG00000111640.13', 'ENSG00000198804.2', 'ENSG00000198727.2', 'ENSG00000198886.2', 'ENSG00000198712.1', 'ENSG00000198938.2'	97.91
LUAD	18	'ENSG00000163017.12', 'ENSG00000188257.9', 'ENSG00000087086.12', 'ENSG00000198840.2', 'ENSG00000142515.13', 'ENSG00000263639.4', 'ENSG00000014257.14', 'ENSG00000198938.2', 'ENSG00000198727.2', 'ENSG00000198804.2', 'ENSG00000198886.2', 'ENSG00000122585.6', 'ENSG00000210082.2', 'ENSG00000158715.5', 'ENSG00000198695.2', 'ENSG00000198786.2', 'ENSG00000142541.15', 'ENSG00000198899.2'	88.92
PRAD	20	'ENSG00000197956.8', 'ENSG00000211890.3', 'ENSG00000111640.13', 'ENSG00000142541.15', 'ENSG00000034510.5', 'ENSG00000274012.1', 'ENSG00000198763.3', 'ENSG00000198727.2', 'ENSG00000228253.1', 'ENSG00000198886.2', 'ENSG00000198804.2', 'ENSG00000198840.2', 'ENSG00000184009.8', 'ENSG00000198899.2', 'ENSG00000198938.2', 'ENSG00000198786.2', 'ENSG00000210082.2', 'ENSG00000211459.2', 'ENSG00000102837.6', 'ENSG00000205542.9'	97.56

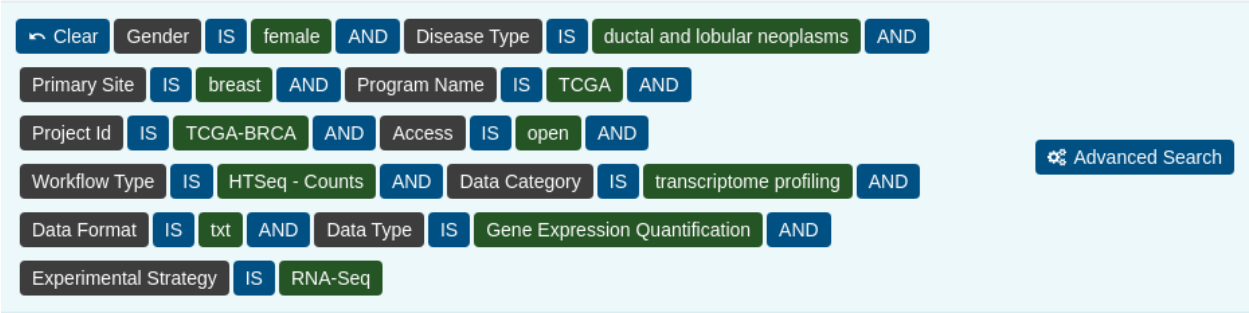
Here, # BM and Acc. stands for No. of BIomarkers and Accuracy, respectively.

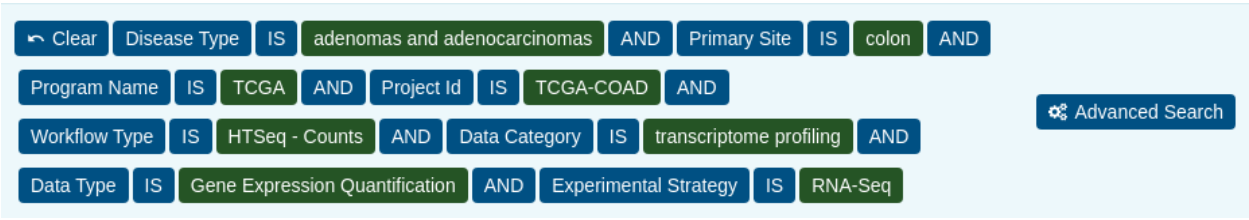
Table 4: Sub-optimal feature set for Primary UCI-5 dataset (17 features in each set)

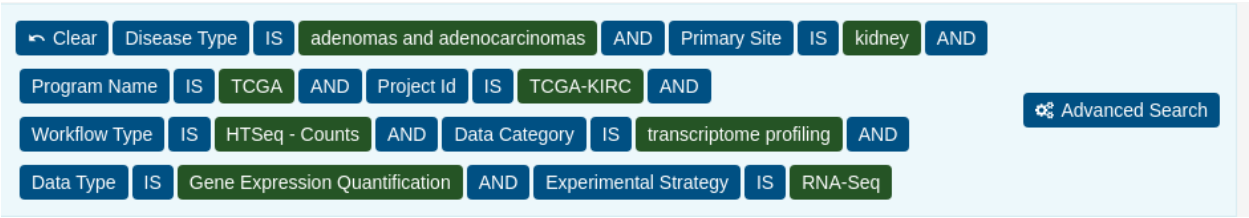
Sub-optimal Feature set	Features	Acc.±st.dev (%)
#1	'gene_17643', 'gene_8326', 'gene_10916', 'gene_6698', 'gene_18381', 'gene_16358', 'gene_18570', 'gene_15896', 'gene_228', 'gene_19375', 'gene_1322', 'gene_13801', 'gene_11910', 'gene_17173', 'gene_17170', 'gene_12851', 'gene_18392'	99.67±0.05
#2	'gene_19739', 'gene_16338', 'gene_9483', 'gene_19035', 'gene_13190', 'gene_7112', 'gene_7148', 'gene_17077', 'gene_4066', 'gene_5590', 'gene_9229', 'gene_4041', 'gene_1200', 'gene_8131', 'gene_17645', 'gene_4419', 'gene_18039'	95.03±0.12
#3	'gene_6857', 'gene_15229', 'gene_15236', 'gene_8128', 'gene_15444', 'gene_14218', 'gene_459', 'gene_11422', 'gene_15314', 'gene_15316', 'gene_552', 'gene_18388', 'gene_34', 'gene_203', 'gene_4042', 'gene_8137', 'gene_7898'	94.62±0.11
#4	'gene_232', 'gene_15202', 'gene_15272', 'gene_15242', 'gene_15300', 'gene_4421', 'gene_15197', 'gene_7218', 'gene_15281', 'gene_5380', 'gene_8127', 'gene_8146', 'gene_15254', 'gene_15253', 'gene_6694', 'gene_230', 'gene_15250'	82.99±0.17

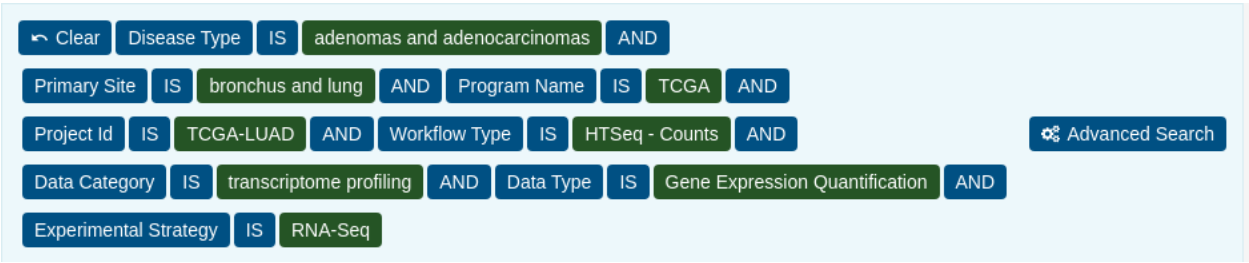
Here, Acc. and st.dev stands for Accuracy and Standard Deviation, respectively.

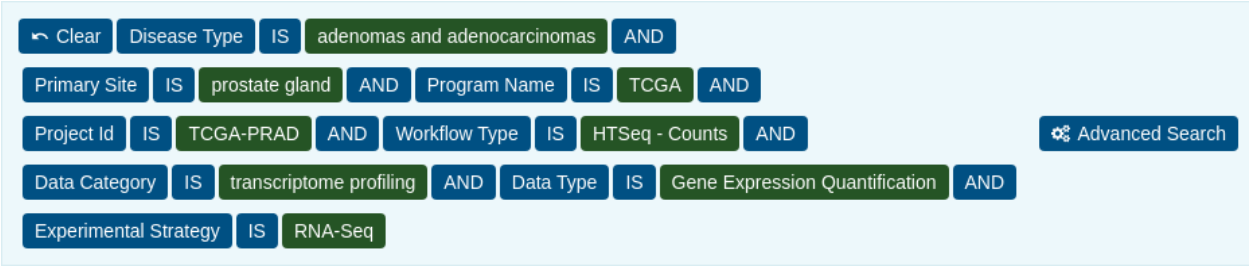
Queries to create TCGA-5 dataset from https://portal.gdc.cancer.gov/repository.











Query Date: 18 February 2022.

Remark: Though we have created three datasets using *HTSeq-Counts, HTSeq-FPKM* and *HTSeq-FPKM-UQ*, replacing *HTSeq-Counts* from the queries above, we have only used *HTSeq-FPKM* dataset in our analysis.

Example Query (for BRCA): LINK (Created 18 April 2022)