A comprehensive genomic pan-cancer classification using The Cancer Genome Atlas gene expression data

YuanYuan Li

BCBB retreat poster presentation

09/08/2016

Introduction

- The Cancer Genome Atlas (TCGA) makes available gene-expression profiles using RNA-seq for many human tumor types.
- These profiles provide a great opportunity to identify unique genes that can classify tumor types.
- Those genes may serve as biomarkers for tumor diagnosis and potential targets for drug development
- Gender differences in cancer susceptibility are consistent findings in cancer epidemiology
- Knowing whether the distinguishing features differ between males and females for the same tumor types might enhance their utility as biomarkers
- Our goals are
    1. To identify a set of genes whose expression levels classify pan-cancer tumor types when gender is ignored
    2. To identify analogous sets of genes for pan-cancer classification in sex non-specific tumors from men and from women separately

TCGA RNAseq data (33 tumor types)

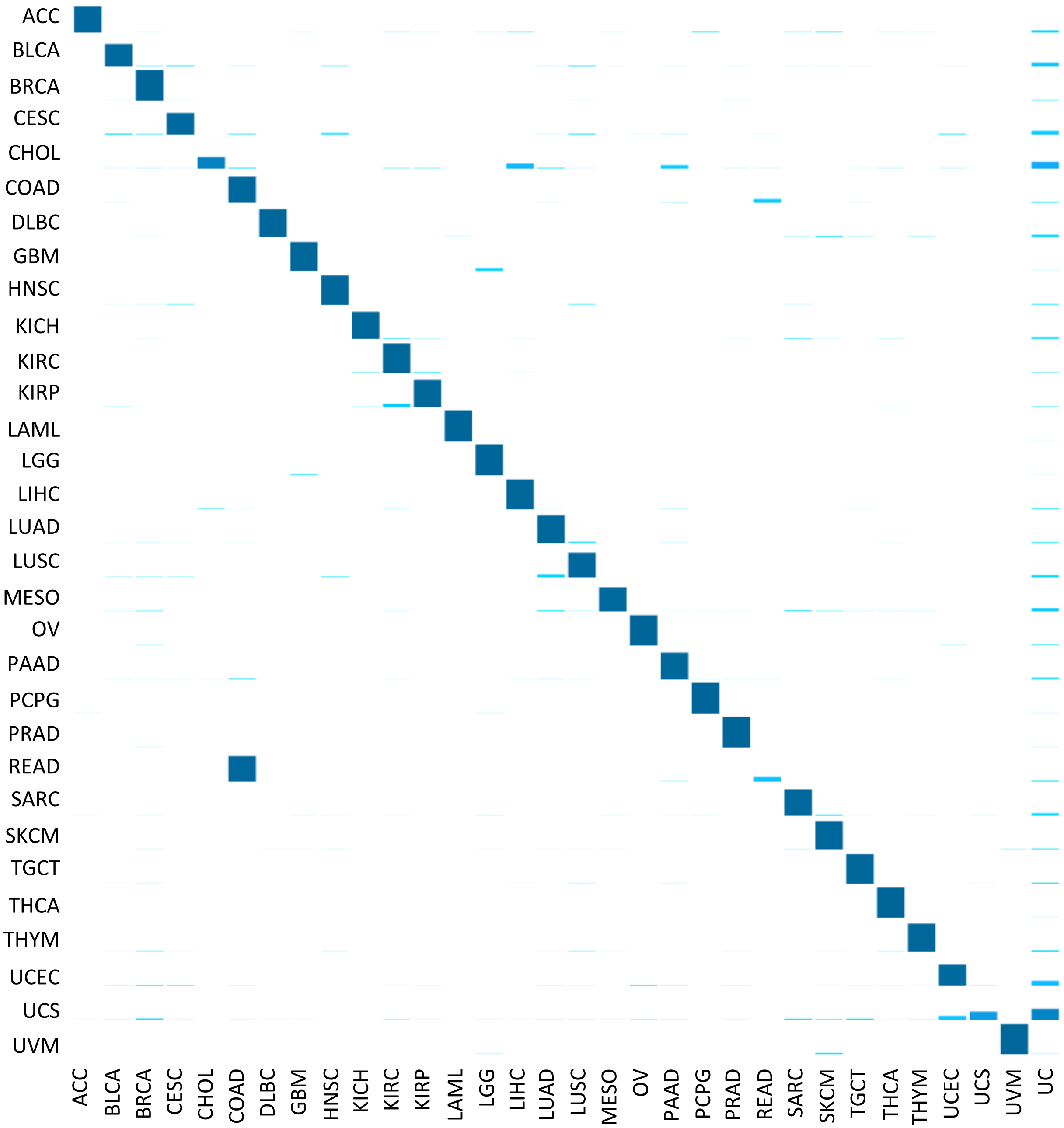
Available cancer types		**Number of Samples**		
		Pan-cancer	**Males**	**Females**
Adrenocortical carcinoma	ACC	79	31	48
Bladder urothelial carcinoma	BLCA	408	272	99
Breast invasive carcinoma	BRCA	1,102	Not used	Not used
Cervical squamous cell carcinoma and endocervical denocarcinoma	CESC	306	Not used	Not used
Cholangiocarcinoma	CHOL	36	Not used	Not used
Colon adenocarcinoma	COAD	287	156	129
Lymphoid neoplasm diffuse large B-cell lymphoma	DLBC	48	Not used	Not used
Esophageal carcinoma	ESCA	Not used	159	26
Glioblastoma multiforme	GBM	169	109	59
Head and Neck squamous cell carcinoma	HNSC	522	385	137
Kidney chromophobe	KICH	66	Not used	Not used
Kidney renal clear cell carcinoma	KIRC	534	346	188
Kidney renal papillary cell carcinoma	KIRP	291	214	77
Acute Myeloid Leukemia	LAML	173	93	80
Brain lower grade glioma	LGG	534	292	241

Liver hepatocellular carcinoma	LIHC	374	253	121
Lung adenocarcinoma	LUAD	517	240	277
Lung squamous cell carcinoma	LUSC	502	371	131
Mesothelioma	MESO	87	71	16
Rectum adenocarcinoma	READ	95	52	42
Sarcoma	SARC	263	119	144
Skin cutaneous melanoma	SKCM	473	259	156
Stomach adenocarcinoma	STAD	Not used	268	147
Testicular germ cell tumors	TGCT	156	Not used	Not used
Thyroid carcinoma	THCA	513	102	246
Thymoma	THYM	120	63	57
Uterine corpus endometrial carcinoma	UCEC	177	Not used	Not used
Uterine carcinosarcoma	UCS	57	Not used	Not used
Uveal melanoma	UVM	80	45	35
Total		9,096	4,081	2,638

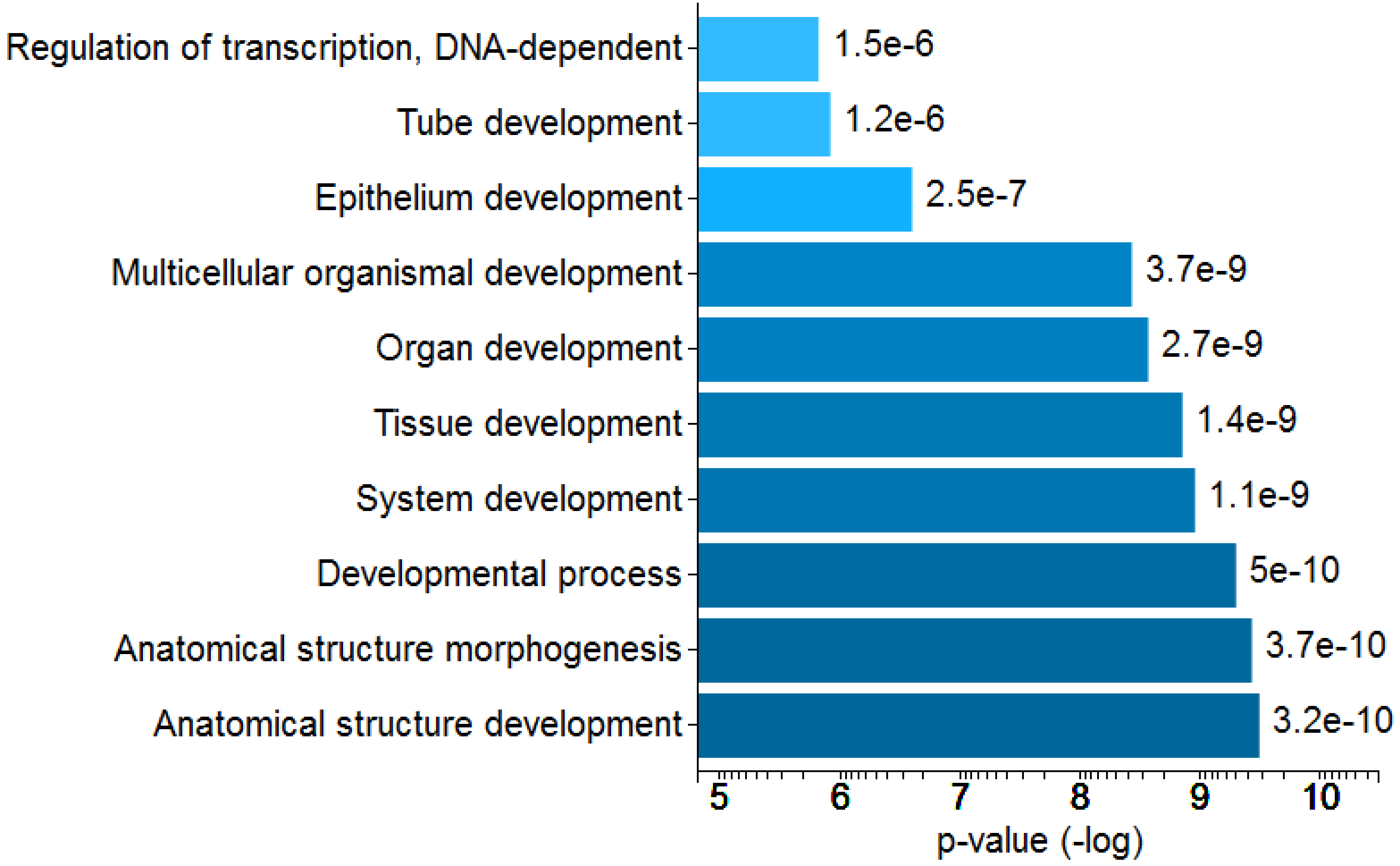
Methods

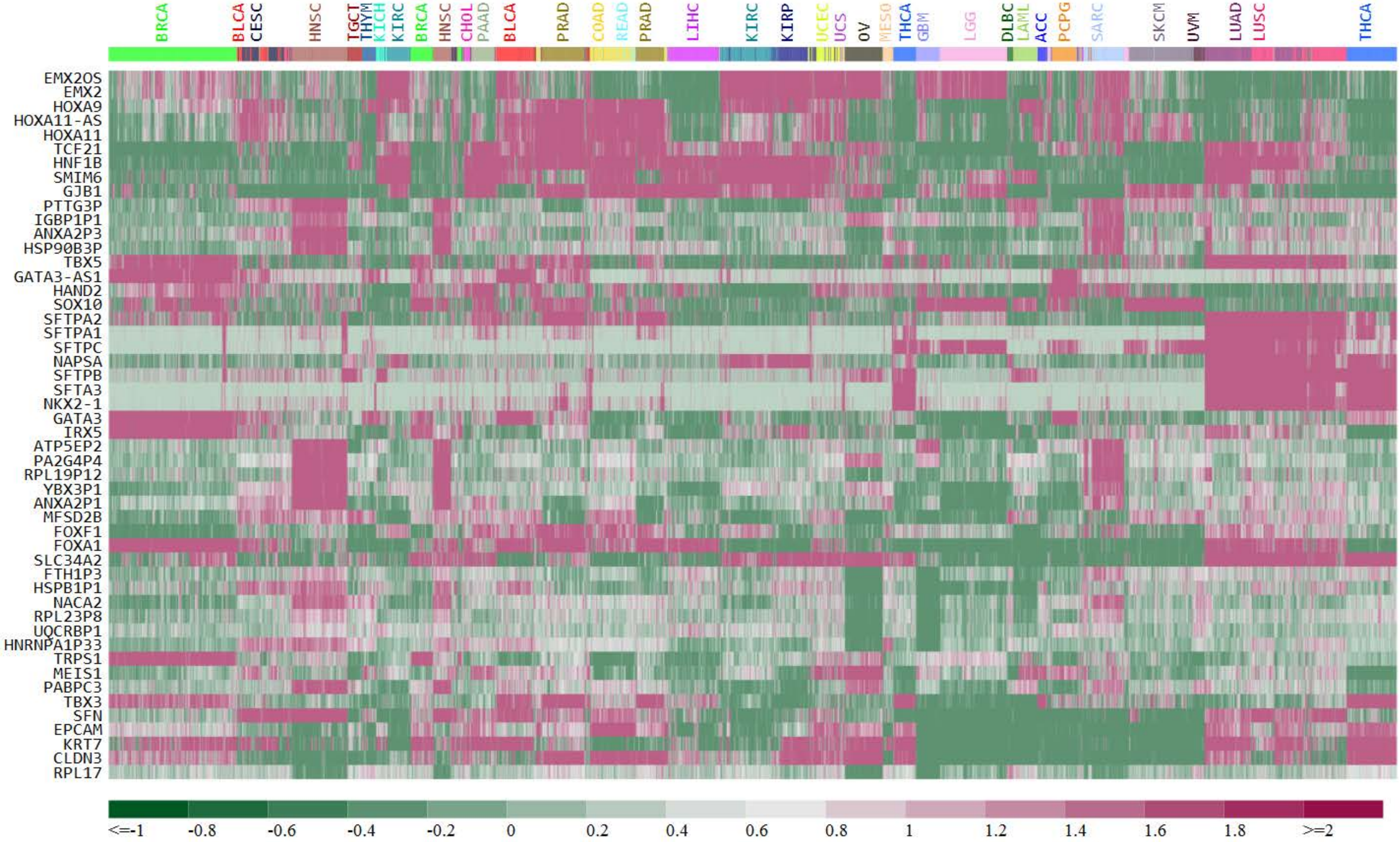
- Used our in-house GA/KNN (Genetic algorithm/K-nearest neighbors) algorithm for multi-class classification of 31 tumor types
- Randomly split samples into 75/25 training & testing sets (proportional allocation of each tumor type)
- Used the training set to obtain 2,000 near-optimal classifiers (each consisting of 20 genes) from repeated runs
- Applied each resultant near-optimal classifier to the test set
- Compared the predicted class with the true class to calculate training & testing performance
- To see if sex non-specific tumors (not related to reproductive organs) differ between males & females, we repeated the analysis for 23 tumor types separately for males & for females.

Results: pan-cancer 31 tumor types


ACC
BLCA
BRCA
CESC
CHOL
COAD
DLBC
GBM
HNSC
KICH
KIRC
KIRP
LAML
LGG
LIHC
LUAD
LUSC
MESO
OV
PAAD
PCPG
PRAD
READ
SARC
SKCM
TGCT
THCA
THYM
UCEC
UCS
UVM
UC

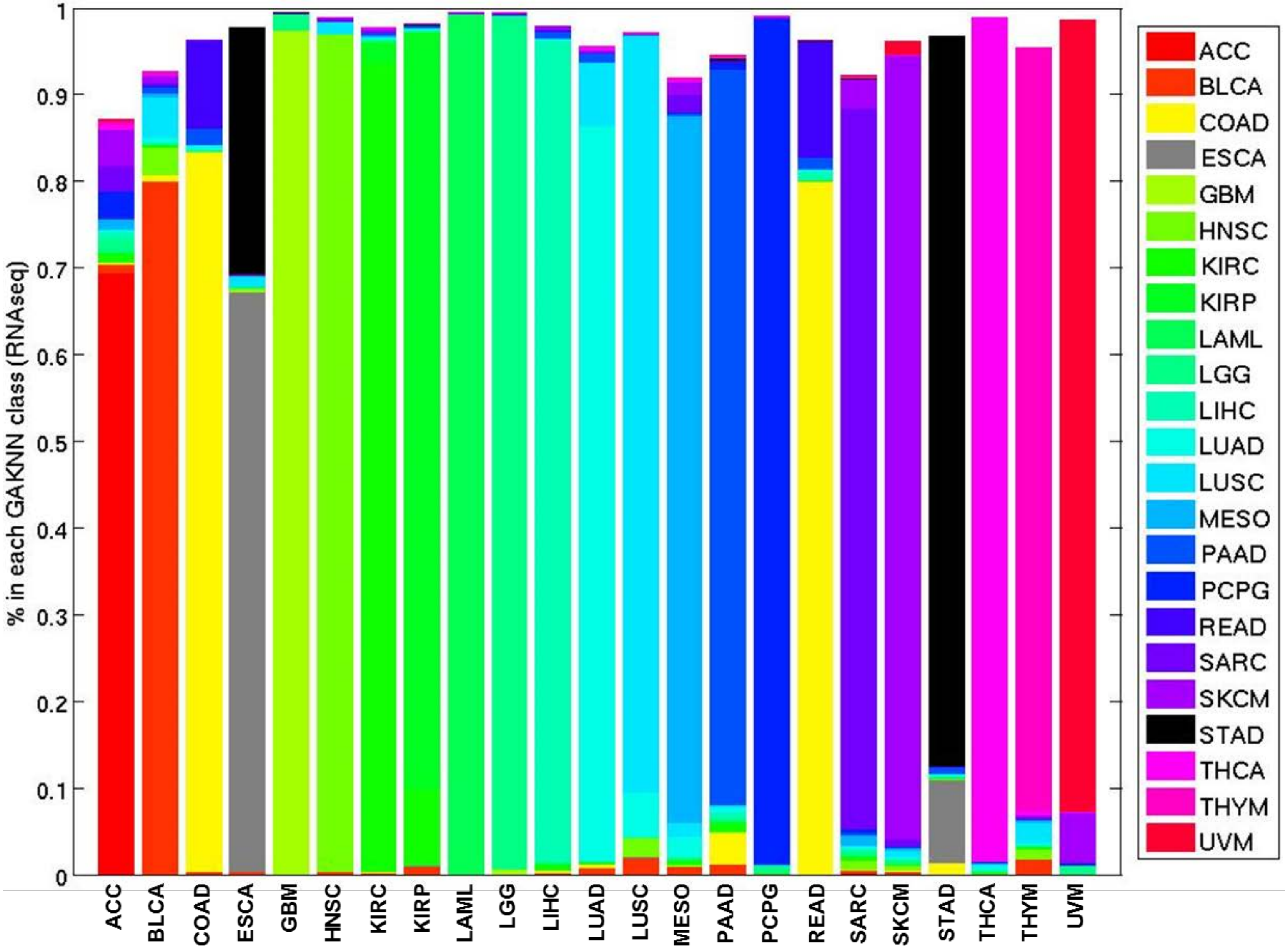
Enriched gene ontology (GO) terms for the top 200 genes from the pan-cancer



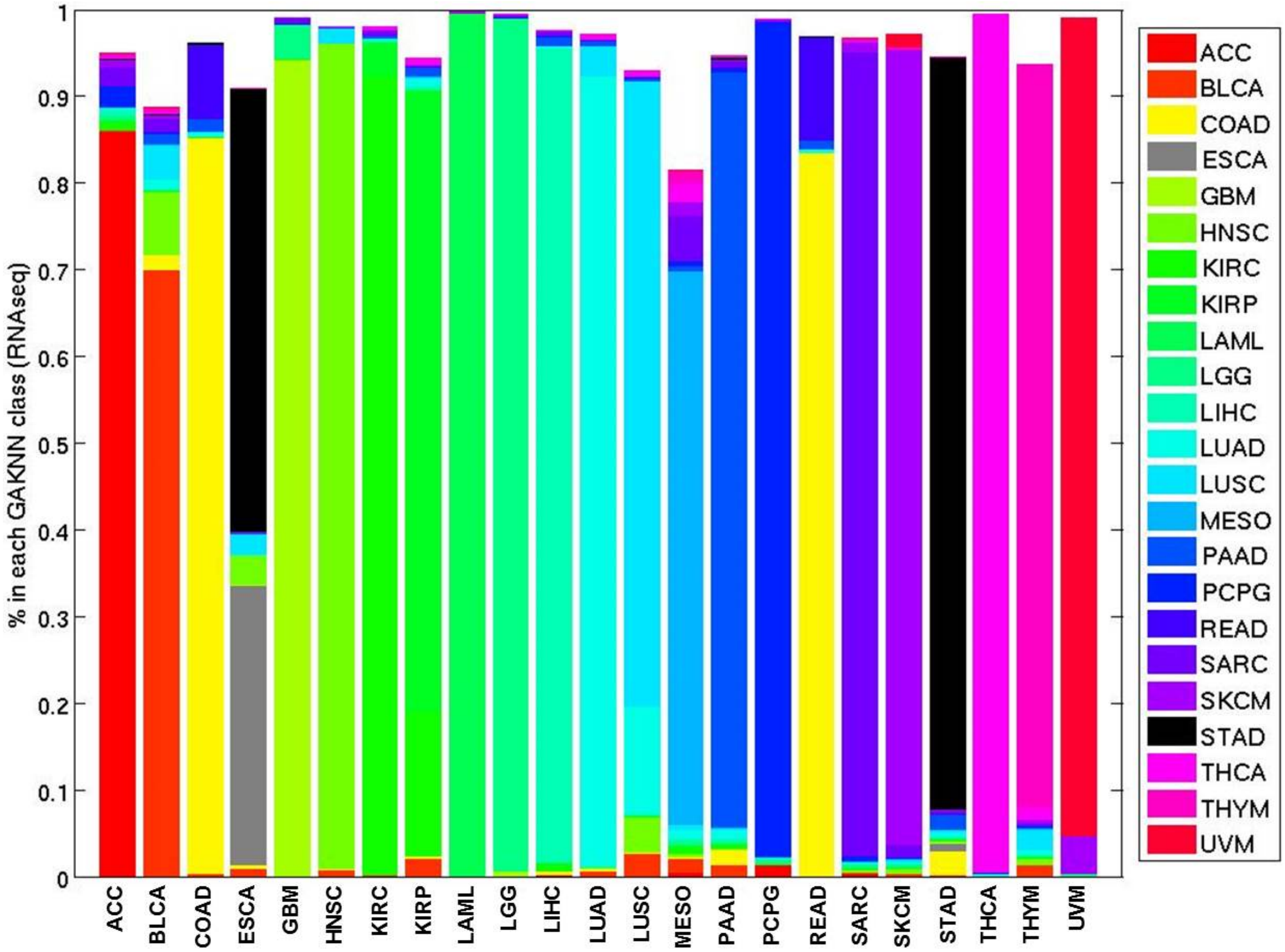

BRCA
BLCA
CESC
HNSC
TGCT
THYM
KICH
KIRC
BRCA
HNSC
CHOL
PAAD
BLCA
PRAD
COAD
READ
PRAD
LIHC
KIRC
KIRP
UCEC
UCS
OV
MESO
THCA
GBM
LGG
DLBC
LAML
ACC
PCPG
SARC
SKCM
UVM
LUAD
LUSC
THCA
EMX2OS
EMX2
HOXA9
HOXA11-AS
HOXA11
TCF21
HNF1B
SMIM6
GJB1
PTTG3P
IGBP1P1
ANXA2P3
HSP90B3P
TBX5
GATA3-AS1
HAND2
SOX10
SFTPA2
SFTPA1
SFTPC
NAPSA
SFTPB
SFTA3
NKX2-1
GATA3
IRX5
ATP5EP2
PA2G4P4
RPL19P12
YBX3P1
ANXA2P1
MFSD2B
FOXF1
FOXA1
SLC34A2
FTH1P3
HSPB1P1
NACA2
RPL23P8
UQCRBP1
HNRNPA1P33
TRPS1
MEIS1
PABPC3
TBX3
SFN
EPCAM
KRT7
CLDN3
RPL17
<=-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
>=2

Results: male & female 23 sex non-specific tumor types

Male testing performance



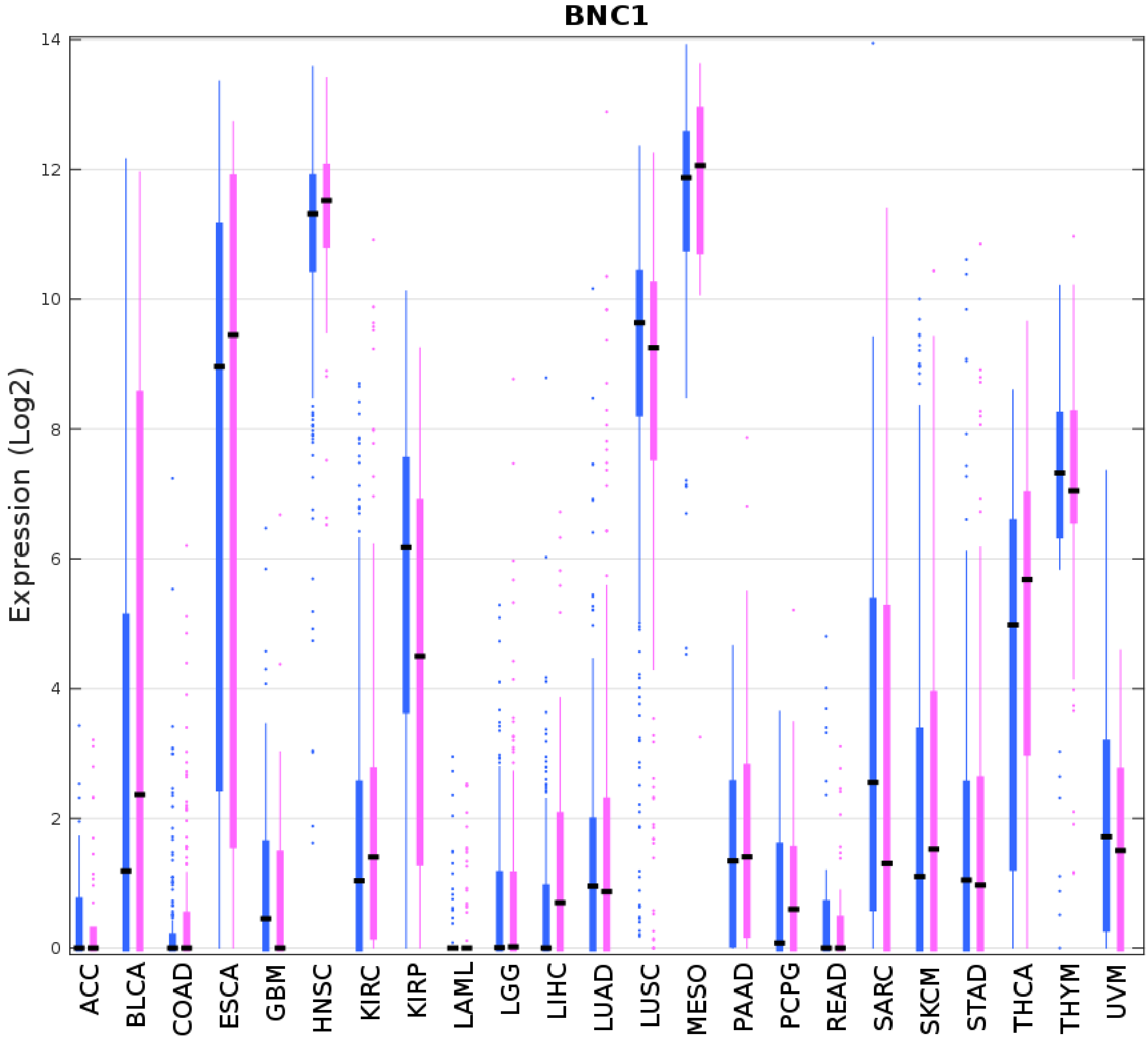
Female testing performance



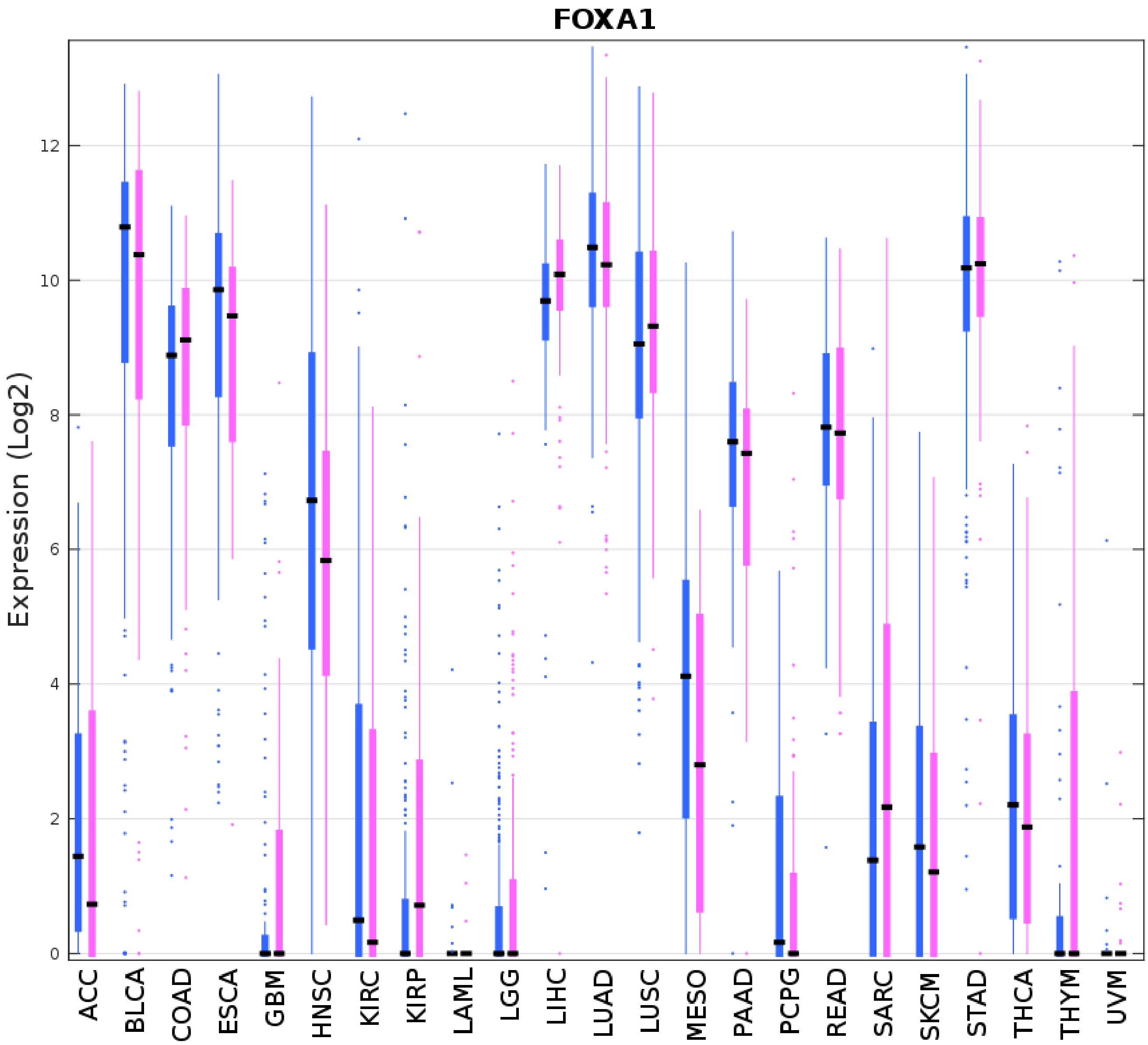
Gene ranks male vs female

	Gene	Rank from full female dataset	Rank from full male dataset	Difference (F-M)	Mean (SD) rank from 8 matched male datasets	Difference (F-meanM)
Genes ranked higher using male samples than female samples	**BNC1**	**931**	**44**	**887**	**54(16)**	**877**
	FAT2	391	89	302	143(23)	248
	KRT5	327	46	281	165(57)	162
	RNF43	298	93	205	81(14)	217
	S1PR5	280	99	181	98(38)	182
	ANKS4B	244	96	148	115(20)	129
	CSTA	217	92	125	129(33)	88
	ANXA8	160	47	113	121(36)	39
	KRT8	174	64	110	94(22)	80
	CLRN3	203	97	106	86(15)	117
Genes ranked higher using male samples than female samples	**FOXA1**	**81**	**416**	**-335**	**237(92)**	**-156**
	AMY1A	99	369	-270	386(162)	-287
	HPN	73	335	-262	256(94)	-183
	LAD1	44	268	-224	129(40)	-85
	PDZK1	82	292	-210	228(79)	-146
	TMC5	54	240	-186	139(50)	-85
	KIF12	88	248	-160	324(135)	-236
	STK32A	78	225	-147	123(28)	-45
	CFAP221	80	186	-106	94(21)	-14
	TRIM29	85	187	-102	143(25)	-58
	HOXA11	83	183	-100	291(77)	-208

BNC1 in 23 sex non-specific tumor types



FOXA1 in 23 sex non-specific tumor types



Summary & Conclusion

- Pan-cancer classification (ignoring gender) of 9,096 samples into 31 tumor types using RNA-seq gene expression alone was remarkably accurate.
  - > 90% testing accuracy
  - Classification accuracies were high for except 3 tumors
  - Genes whose expression best discriminated all tumors
    - ~1/3 were pseudogenes
    - ~1/3 were transcription factors
    - and ~1/3 were encoded proteins involved in cell adhesion, ion and small molecular transport, protein synthesis and folding, and lung function

Summary & Conclusion (Cont'd)

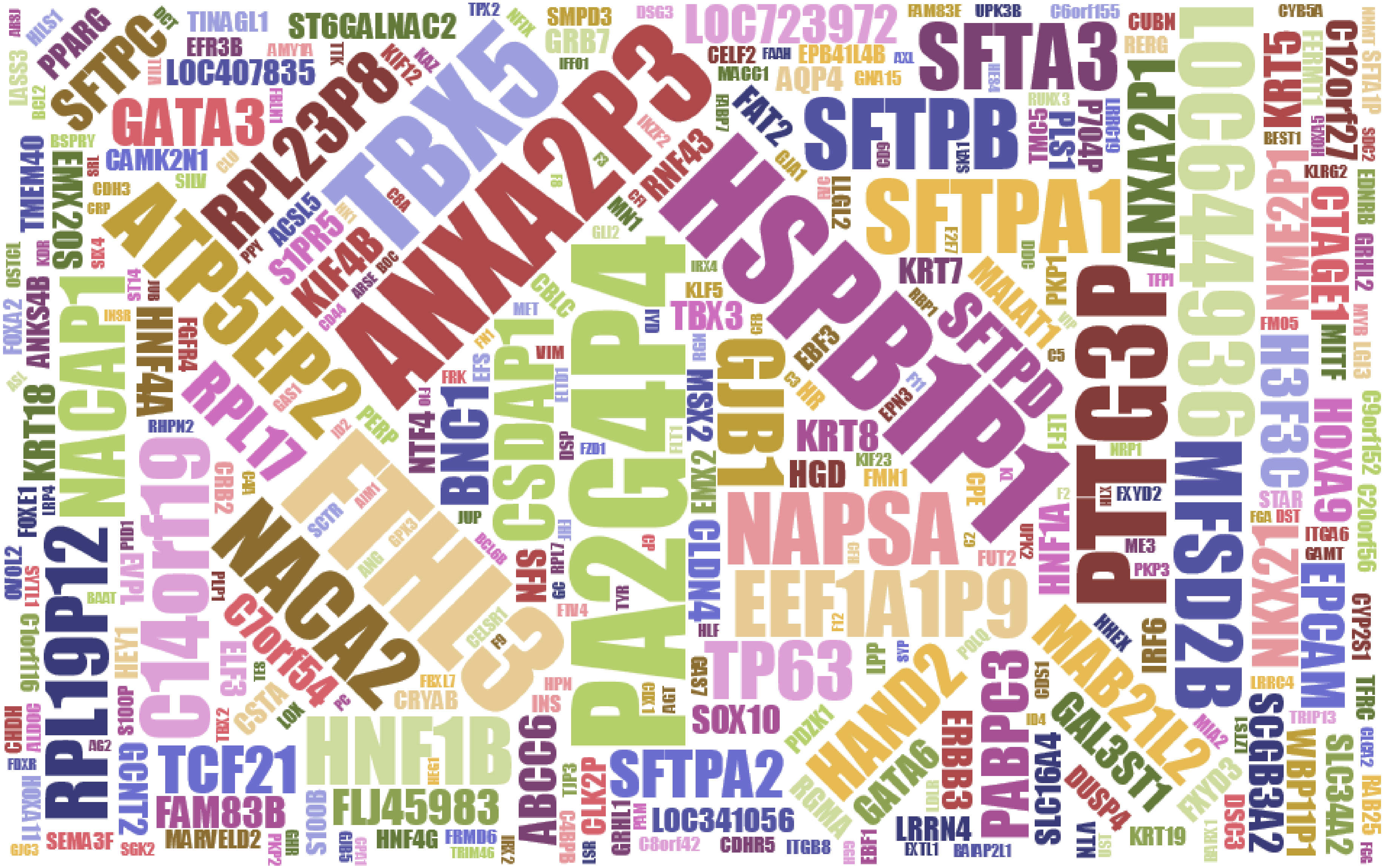
- Repeating the analysis for 23 sex non-specific tumor types in males & females separately led to similarly accurate classification in each gender and overlapping but distinct gene lists
  - > 80% of the 100 most discriminative genes were common between males & females
  - Genes that were differentially expressed between male & female included: *BNC1*, *FAT2*, *FOXA1*, & *HOXA11*
- The few top-ranked discriminative genes that differed between males and females might be related to sex differences in tumor incidence and prognosis

Thank you!

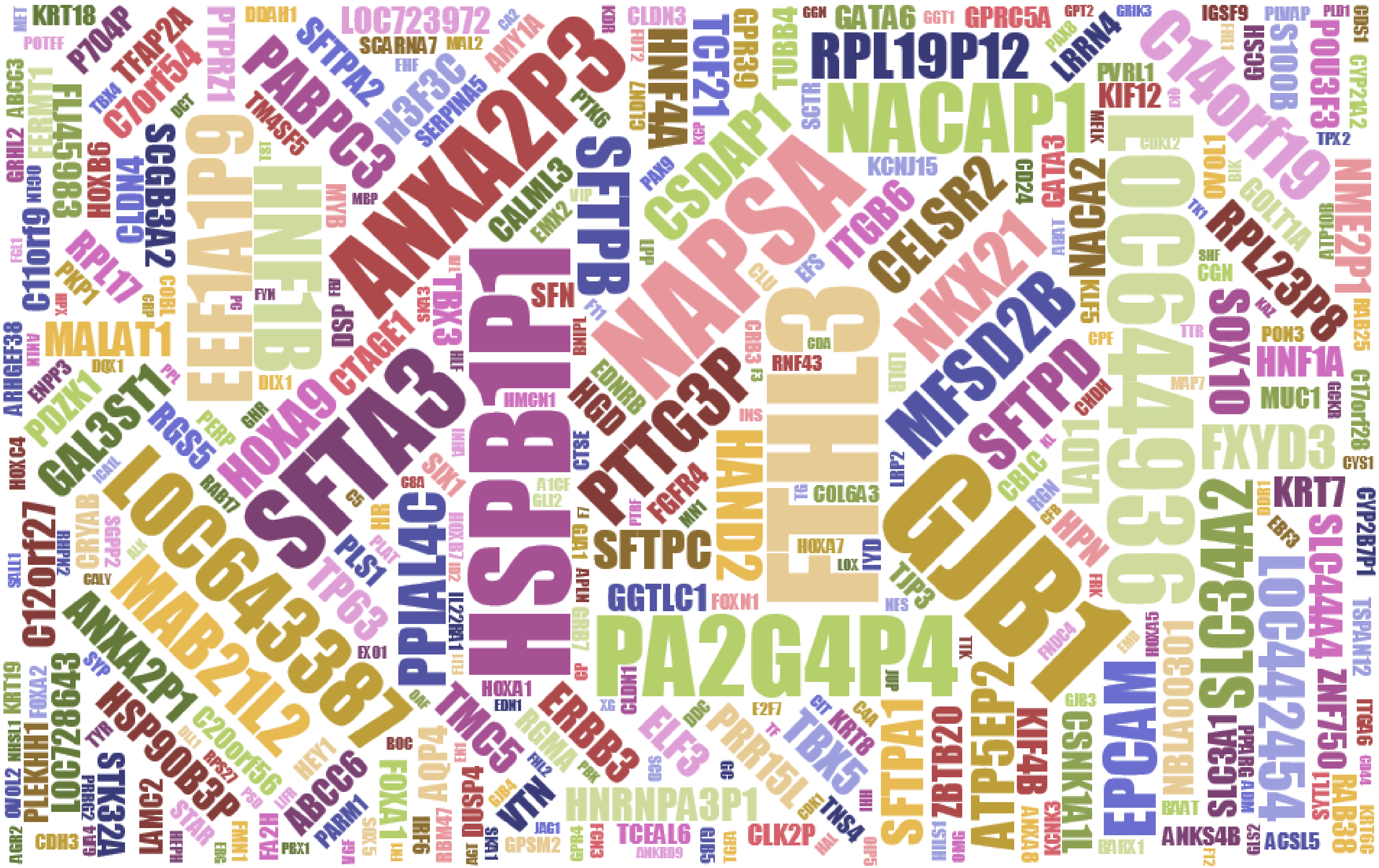
Pan-cancer word cloud

TCF21
TBX5
EMX2OS
HNF1B
EMX2
PA2G4P4
ATP5EP2
SFTA3
SFTPA1
GATA3
ETHL3
NAPSA
PTTG3P
NACA2
ANXA2P3
C14orf19
HOXA9
HOXA11AS
HOXA11
ANXA2P1
RPL19P12
HSP90B3P
HSPB1P1
SFTPA2
SFTPC
SFTPB
MFSD2B
FLJ45983
LOC644936
LOC728643
LOC407835
LOC100130933
LOC100130331
LOC642587
LOC643387
LOC649330
LOC442454
LOC723972
CELSR2
KRT7
TMEM125
C15orf21
RPL23P8
PABPC3
SLC34A2
GPM6B
SFTPD
FOXF1
NKX2-1
CSDAP1
IRX5
HAND2
NBLA00301
CTAGE1
ESR1
SFN
EEF1A1P9
NME2P1
CLDN4
SYCP2
S100A1
OVOL1
PPARG
MECOM
RPL13AP3
DSP
C1orf135
GCNT2
S1PR5
HNRNPA3P1
TRPS1
CYP2S1
EPCAM
MEIS1
ARHGEF38
BNC1
IGSF11
TMEM108
SCGB3A2
ARL4D
CLDN3
LYPD3
C17orf28
HNF4A
AQP4
TRIM29
AZGP1
CDCA7
MITF
SOX17
TJP3
KIF5C
NACAP1
GJB1
AOC3
REC8
CSNK1A1L
MAB21L2
ST6GALNAC2
KRT5
CLDN1
ESRP1
KRT19
WBP11P1
ANXA8L2
GCSH
HNF1A
C20orf56
ERBB3
SCG2
FAM110B
MUC1
FOXA2
FOXA1
TFAP2A
SOX10
FOXE1
TBX3
HGD
C11orf9
LRRN4
KRT18
IRX3
ITGA3
TNS4
LAD1
PKP1
KIF4B
ELF3
IRX2
RPL17
EFHD1
SALL1
CGN
GPR87
VTN
TP63
SPDEF
EPN3
S100A2
KHDC1
UBE2NL
SH3TC2
GRHL2
GRHL1
ST6GALNAC2
HOXB6
LLGL2
GGET2
FXYD3
PLEK2
CRB2
FUNDC2P2
BSPRY
ELF3
KIF6
DST
COBL
KLRG2
RHOV
PVRL1
IRX1
GHRH
FOXN1
P704P
ANXA8
ITGB6
CD44
POLQ
IRF6
RERG
EBF3
PRSS8
TPX2
GJA1
BIM
TRIL
LIF
PBK
GGNE2
EGFR
KLF5
AIM1
H3F3C
MEOX2
PLS1
NTF4
BNIP3
WNT7B
TMPRSS2
PLCH1
GJB5
DHCR24
ARHGAP23
FERMT1
TACSTD2
RNF43
GPA33
DUOXA2
CAPN5
MLPH
EYA2
GBP6
GALNT3
GAL3ST1
PAK8
MIA2
EDARADD
CHDH
CTXN1
PRAC
LRP4
HOXD8
HOXB5
EXTL1
ACCN2
GNG7
SYK
KRT5
SDC1
KCNJ15
MRO
EFS
SIX1
SNAI2
GLI3
GHR
F2R
DSC3
CLDN1
ASRGL1
LRIG1
SFTPD
NID2
GRB7
MF1
EBF1
ATP1B2
ASF1B
PTTG2
VIM
IVD
LSR
TIAM1
TFPI
NES
TTK
KIF12
FTO
ID1
CP
ID2
FN1
GA2
KAZ
ICA1
TEF
SIPL
GJB3
PON3
DTL
MYB
PTK6
FYN
HOXB2
CDK6
RIN1
CDK1
C4A
AFP
ADAM8
MET
JAG1
KIT
MLF1
PKP1
NEXN
TNN
DSG3
SIX4
SOX17
MAB21L1
ETV4
ROC
CLU
FMB
GMPR
HEPH
KCTD1
AG2
AXL
CJTS
TF
KDR
GJC1
LGI3
HDN
MNX1
RBP1
FGF1
CTSL2
ID4
TLR3
FLNC
PGR
TFY
TYMS
DUSP1
LOX
DHX1
DLX1
FZD7
GCNT1
ODZ2
LBR
HHEX
ADH6
LIFR
JUP
TUB
GLI2
EYA2
MUC1
HEY1
HOXA9
GD9
PHOX2A

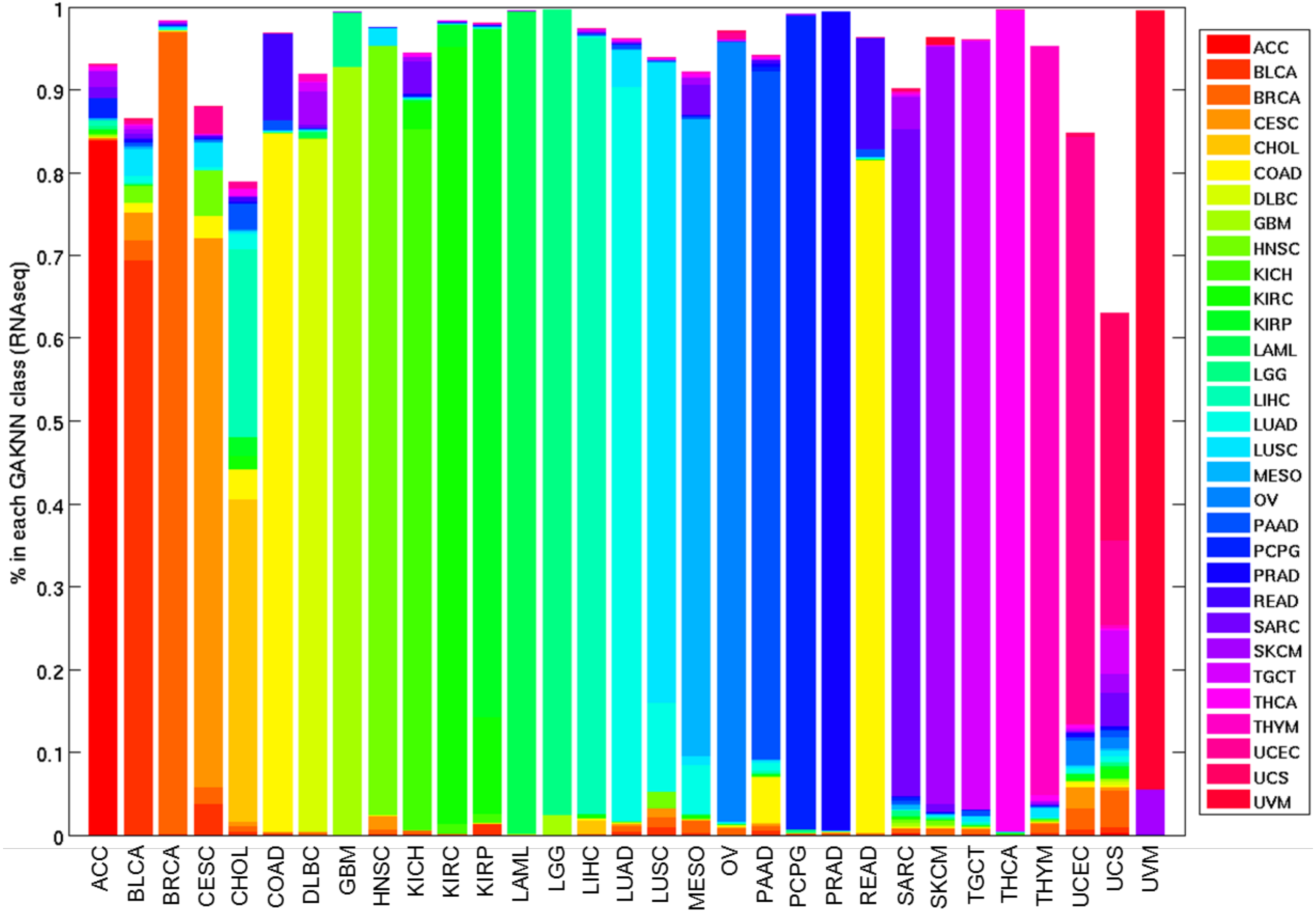
Male word cloud



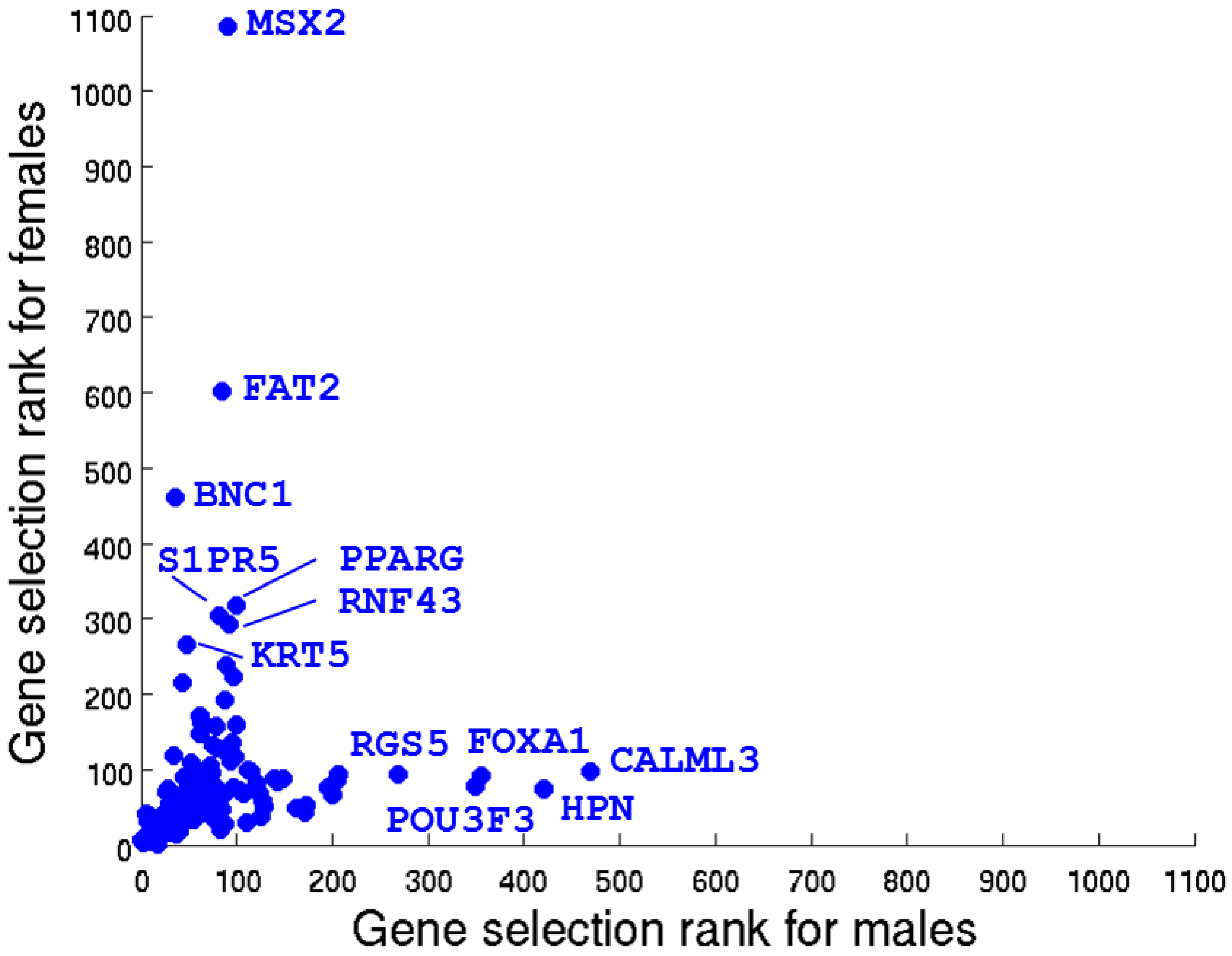
Female word cloud



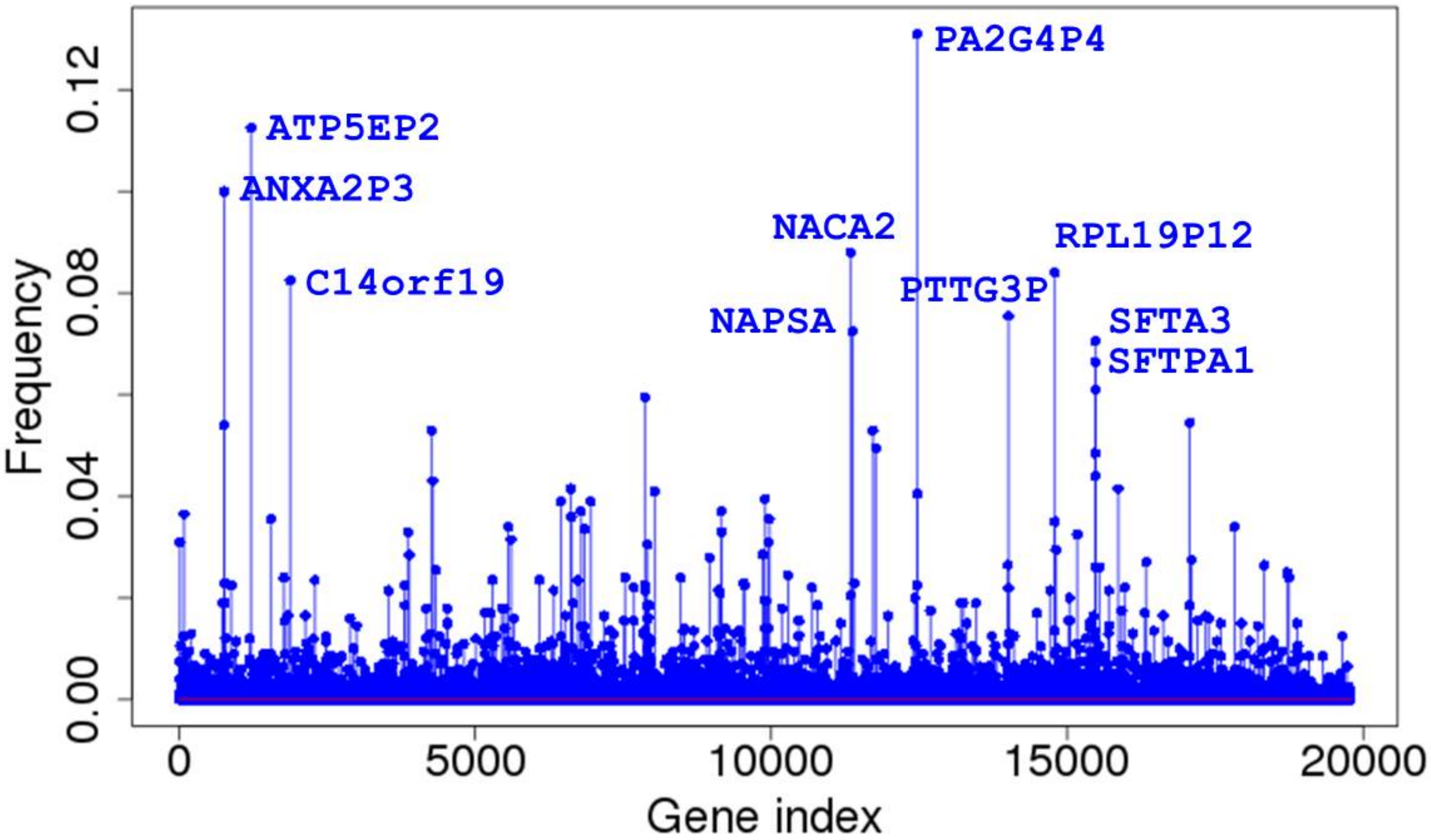
Pan-cancer testing performance



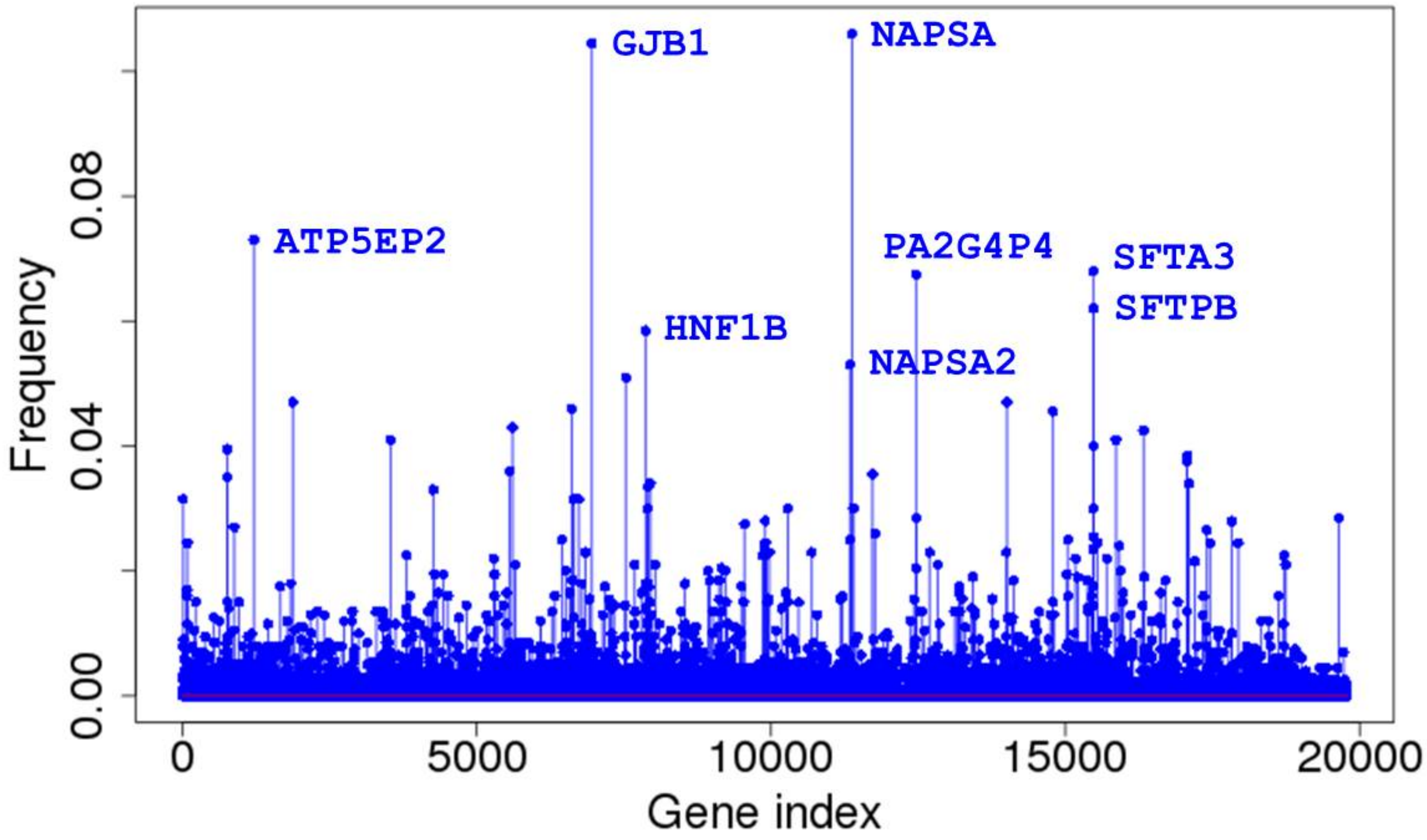
Top 100 ranked genes in males vs females



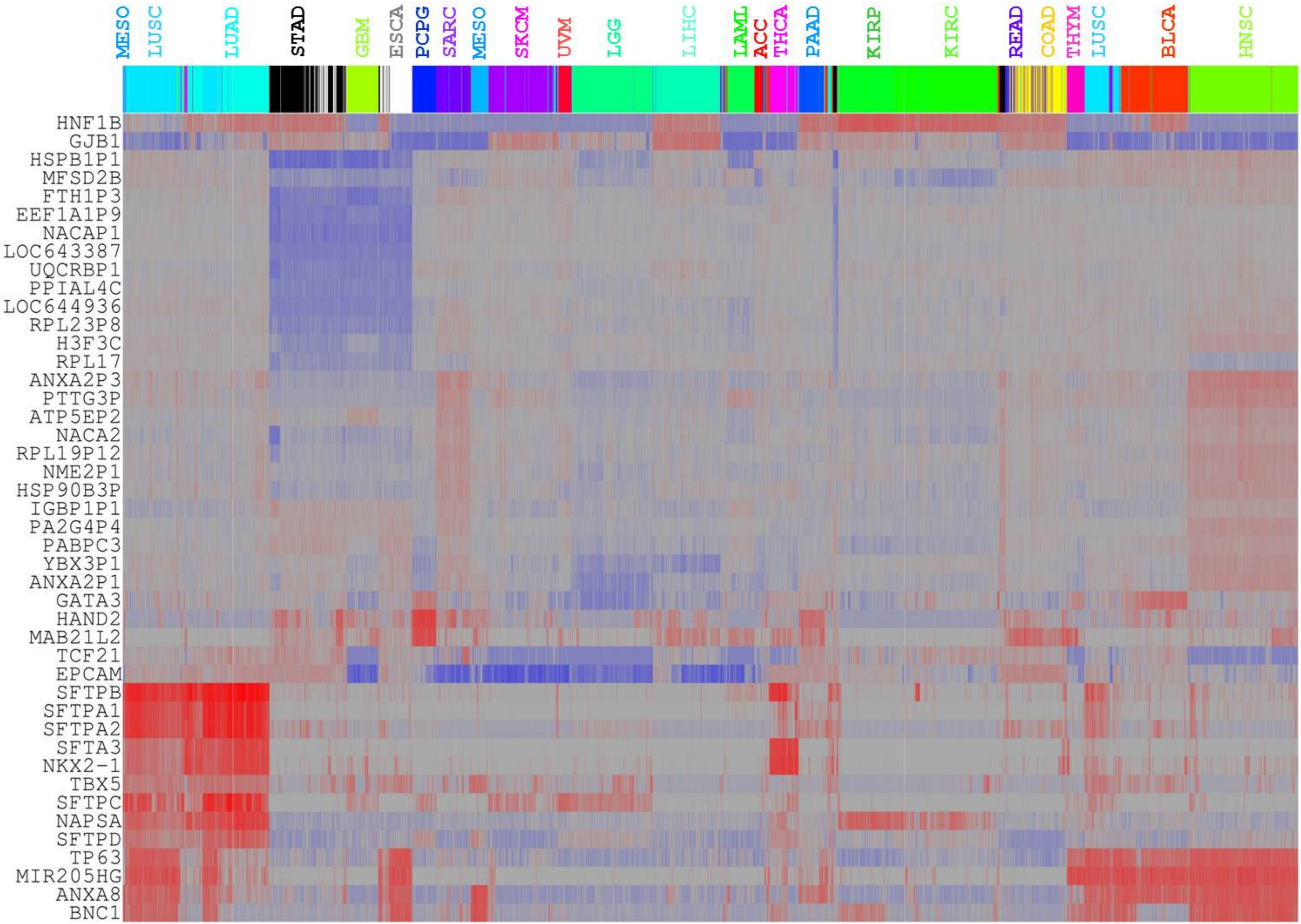
Male sex-non specific



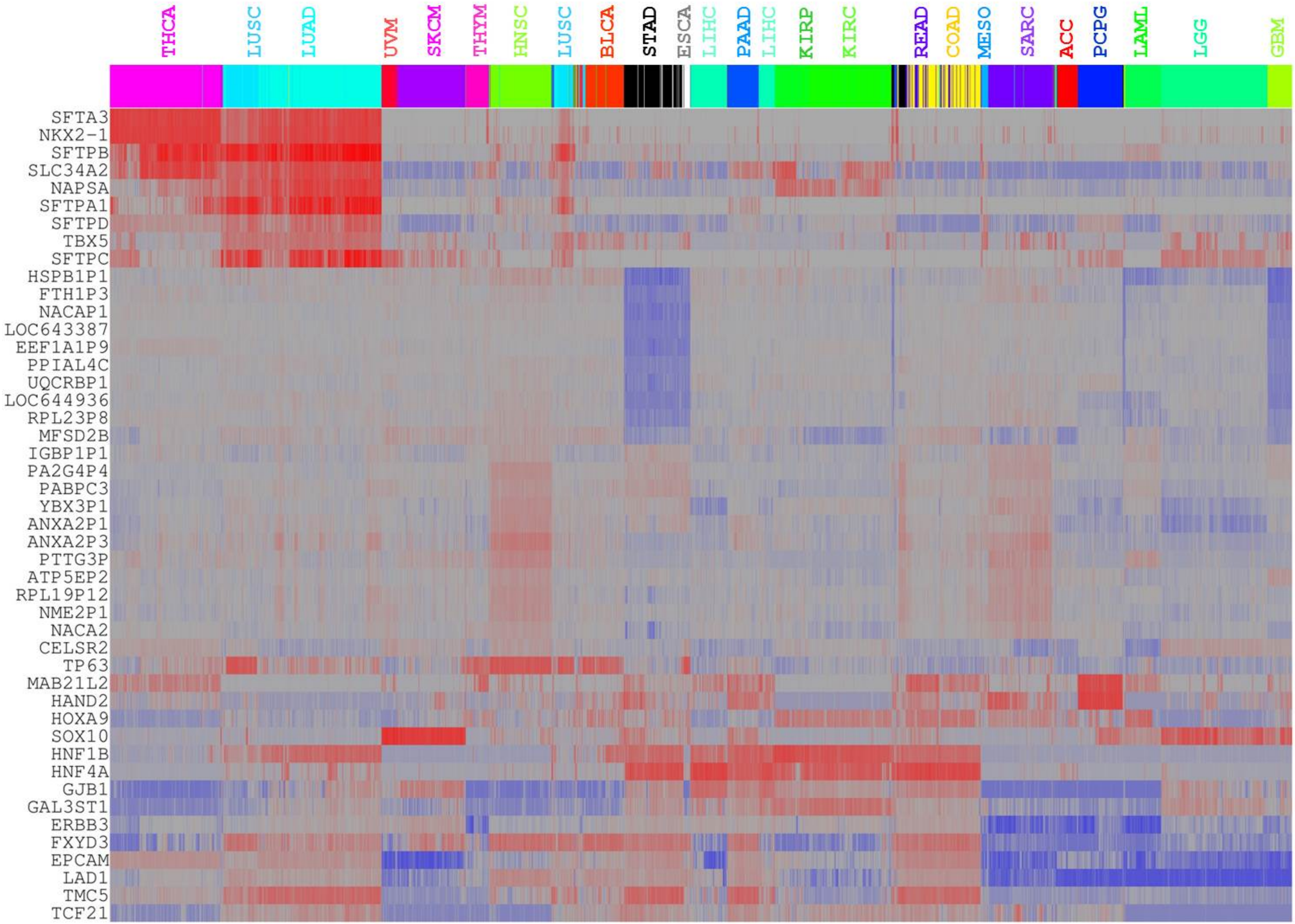
Female sex-non specific



Male sex-non specific (44 genes)



Female sex-non specific (46 genes)



Type	Sample	ACC	BLCA	BRCA	CESC	···	THYM	UCS	UVM	Unclassifiable
ACC	S1	0.911	0.000	0.002	0.000	···	0.000	0.000	0.000	0.041
	S2	0.938	0.000	0.001	0.000	···	0.000	0.000	0.000	0.028
	S3	0.885	0.000	0.000	0.000	···	0.000	0.000	0.001	0.038
	S4	0.946	0.000	0.000	0.000	···	0.000	0.000	0.000	0.019
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	S79	0.232	0.001	0.006	0.000	···	0.000	0.000	0.004	0.304
BLCA	S1	0.000	0.823	0.001	0.108	···	0.000	0.000	0.000	0.048
	S2	0.000	0.138	0.196	0.014	···	0.007	0.000	0.000	0.376
	S3	0.000	0.986	0.000	0.000	···	0.001	0.000	0.000	0.011
	S4	0.009	0.041	0.006	0.005	···	0.001	0.000	0.002	0.531
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	S408	0.000	0.983	0.000	0.002	···	0.000	0.000	0.000	0.008
BRCA	S1	0.000	0.000	1.000	0.000	···	0.000	0.000	0.000	0.000
	S2	0.000	0.000	1.000	0.000	···	0.000	0.000	0.000	0.000
	S3	0.000	0.004	0.978	0.000	···	0.001	0.000	0.000	0.007
	S4	0.000	0.001	0.986	0.000	···	0.001	0.000	0.000	0.006
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	S1102	0.000	0.004	0.979	0.000	···	0.001	0.000	0.000	0.011
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
UVM	S1	0.000	0.000	0.000	0.000	···	0.000	0.000	0.982	0.000
	S2	0.000	0.000	0.000	0.000	···	0.000	0.000	1.000	0.000
	S3	0.000	0.000	0.000	0.000	···	0.000	0.000	0.999	0.000
	S4	0.000	0.000	0.000	0.000	···	0.000	0.000	0.970	0.001
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	S80	0.000	0.000	0.000	0.000	···	0.000	0.000	0.993	0.000

Summary statistics for π_{cc} values when classifying 31 tumor types and ignoring sex of the samples. The rightmost column labeled "overall" is not based on π_{cc} but instead on a prediction using the tumor type to which each sample was assigned most often.

Type	Min.	1st Qu.	median	mean	3rd Qu.	Max.	Posterior-like prob.*
ACC	0.228	0.757	0.877	0.831	0.920	0.971	97.0%
BLCA	0.009	0.508	0.811	0.711	0.957	0.997	90.6%
CHOL	0.000	0.006	0.400	0.369	0.504	0.659	73.3%
COAD	0.181	0.769	0.851	0.829	0.906	0.984	98.6%
DLBC	0.653	0.821	0.894	0.870	0.935	0.975	100.0%
GBM	0.463	0.857	0.955	0.910	0.981	0.998	98.6%
HNSC	0.040	0.909	0.980	0.930	0.997	1.000	98.7%
KICH	0.000	0.877	0.919	0.856	0.957	0.992	96.4%
KIRC	0.003	0.980	0.998	0.933	1.000	1.000	95.7%
KIRP	0.000	0.790	0.974	0.853	0.996	1.000	92.1%
LAML	0.886	0.997	0.999	0.992	1.000	1.000	100.0%
LGG	0.562	0.989	1.000	0.972	1.000	1.000	100.0%
LHIC	0.035	0.967	0.991	0.937	0.998	1.000	97.6%
LUAD	0.002	0.877	0.960	0.884	0.990	1.000	95.9%
LUSC	0.019	0.674	0.917	0.780	0.974	1.000	88.0%
MESO	0.000	0.721	0.865	0.757	0.931	0.988	90.0%
PAAD	0.031	0.835	0.961	0.848	0.992	1.000	94.7%
PCPG	0.705	0.984	0.995	0.984	0.999	1.000	100.0%
READ	0.034	0.090	0.136	0.147	0.192	0.279	0.0%
SARC	0.026	0.776	0.908	0.830	0.962	0.998	95.7%
SKCM	0.001	0.930	0.972	0.904	0.986	1.000	96.6%
THCA	0.372	0.998	1.000	0.989	1.000	1.000	100.0%
THYM	0.079	0.904	0.988	0.893	0.998	1.000	94.4%
UCS	0.010	0.062	0.255	0.266	0.405	0.615	60.9%
UVM	0.518	0.951	0.986	0.949	0.996	1.000	100.0%
BRCA	0.010	0.975	0.994	0.967	0.999	1.000	99.4%
CESC	0.001	0.515	0.757	0.681	0.867	0.979	94.1%
OV	0.359	0.954	0.979	0.951	0.992	0.999	100.0%
PRAD	0.527	0.997	0.999	0.987	1.000	1.000	100.0%
TGCT	0.250	0.965	0.997	0.937	1.000	1.000	100.0%
UCEC	0.042	0.518	0.713	0.675	0.864	0.995	95.9%

*Porsterior-like prob. = counts of correct classified testing samples in tumor *i* / no. of samples in tumor *i*