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

Yuanyuan Li[1], Juno M. Krahn[2], Nicole Croutwater[1], Kevin Lee[1], David M. Umbach[1], and Leping Li[1]

[1]Biostatistics and Computational Biology Branch, [2]Genome Integrity & Structural Biology Laboratory, National Institute of Environmental Health Sciences, NIH, DHHS, RTP, NC, USA

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 & 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

Main findings

- Pan-cancer classification (ignoring gender) of 9,096 samples into 31 tumor types using RNA-seq gene expression alone was remarkably accurate.
  - Our algorithm found gene sets that correctly classified >90% of the tumor samples on average.
  - Classification accuracies were high for except 3 tumors
  - Among 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, & lung function
- Repeating the analysis for 23 sex non-specific tumor types in males and females separately gave 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 in at least one of the 23 tumor types included: BNC1, FAT2, FOXA1, & HOXA11
- The few top-ranked discriminative genes that differed between males and females might play a role in differential cancer susceptibility between genders
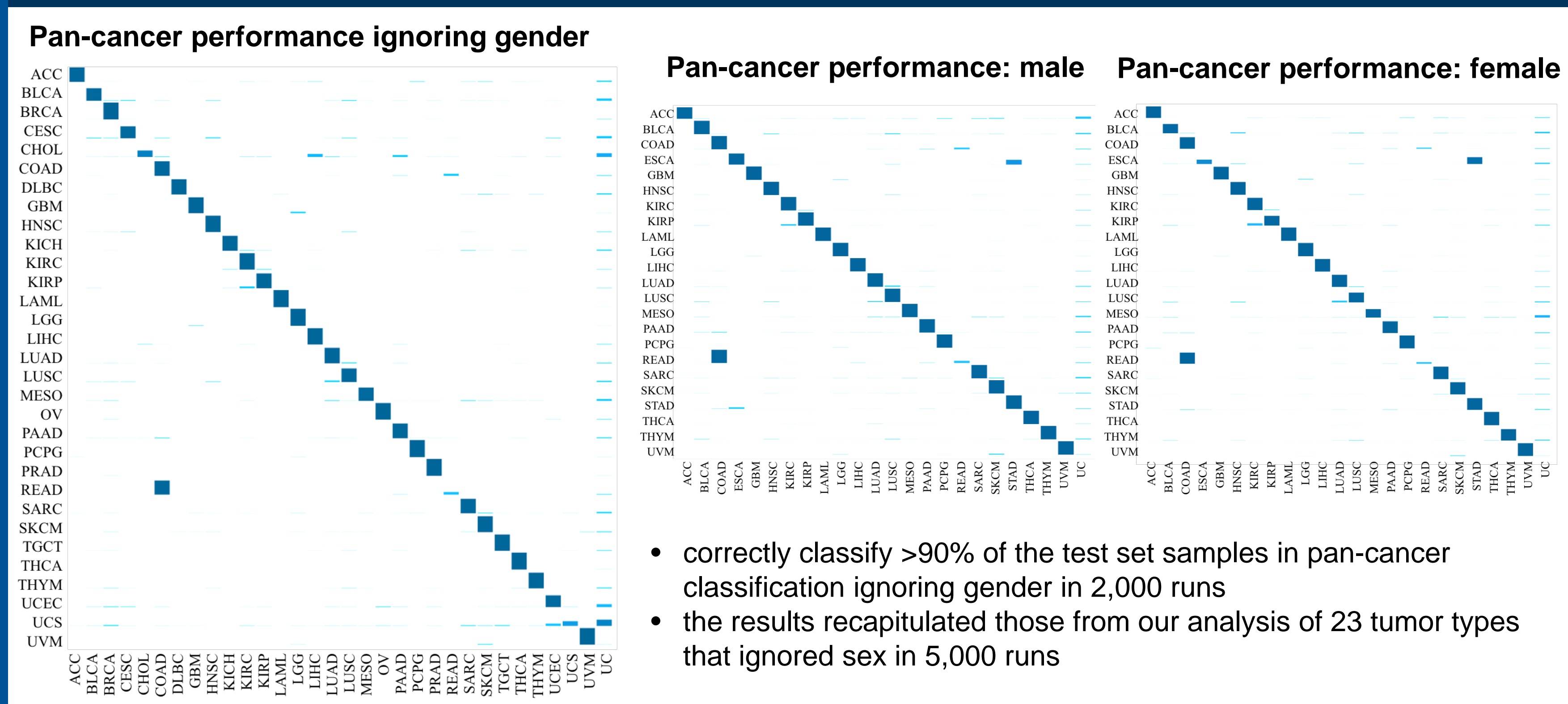
Data

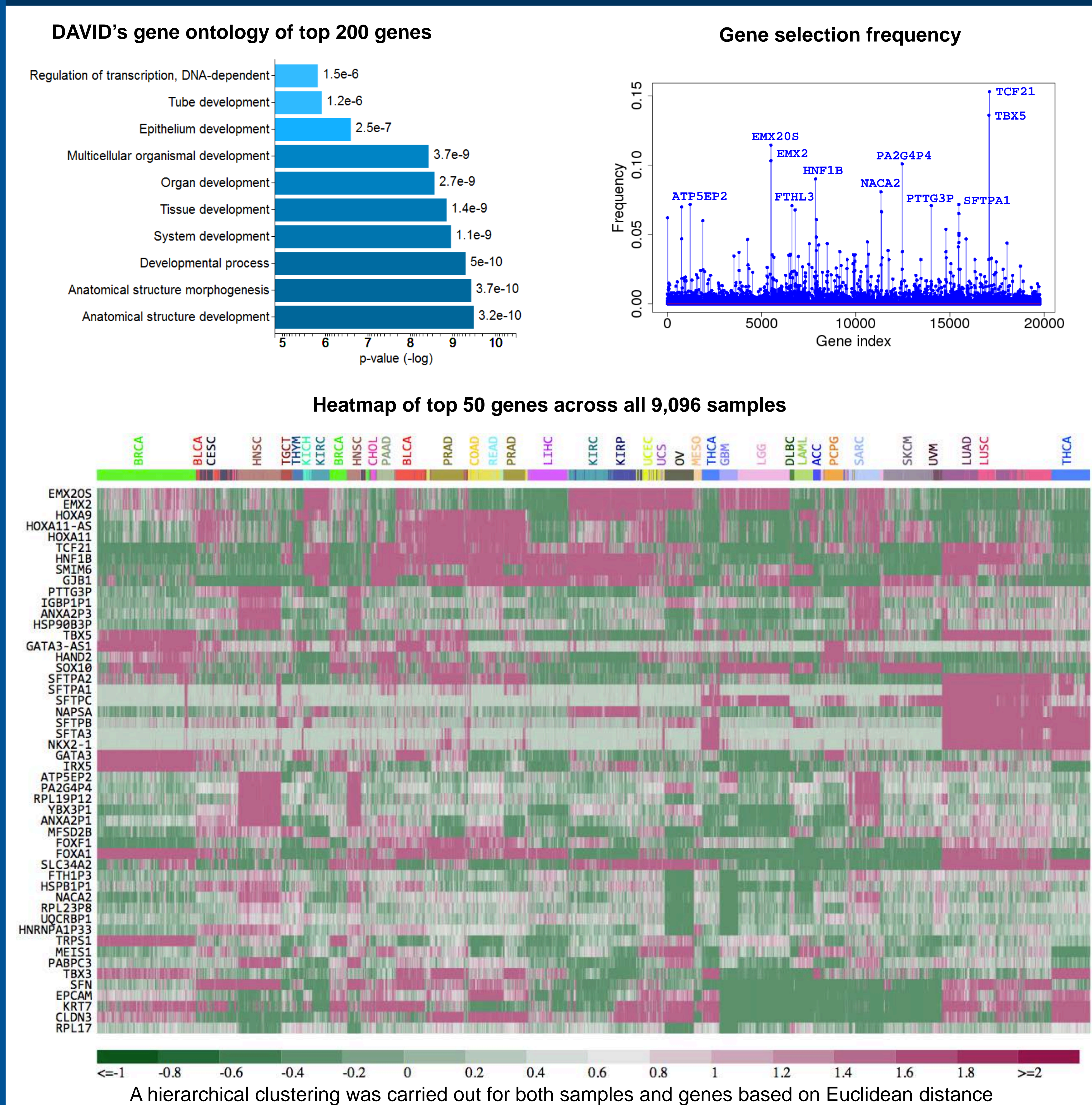| Available cancer types		Number of Samples		
		Pan-cancer	Males	Females
Adrenocortical carcinoma	ACC	79	31	48
Bladder urothelial carcinoma	BLCA	408	272	99
Breast invasive carcinoma	BRCA	1,102	Not used	Not used
Cervical squamous cell carcinoma and endocervical adenocarcinoma	CESC	306	Not used	Not used
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
Ovarian serous cystadenocarcinoma	OV	266	Not used	Not used
Pancreatic agenocarcinoma	PAAD	179	99	80
Pheochromocytoma and Paraganglioma	PCPG	184	82	102
Prostate adenocarcinoma	PRAD	498	Not used	Not used
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 50/50 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.
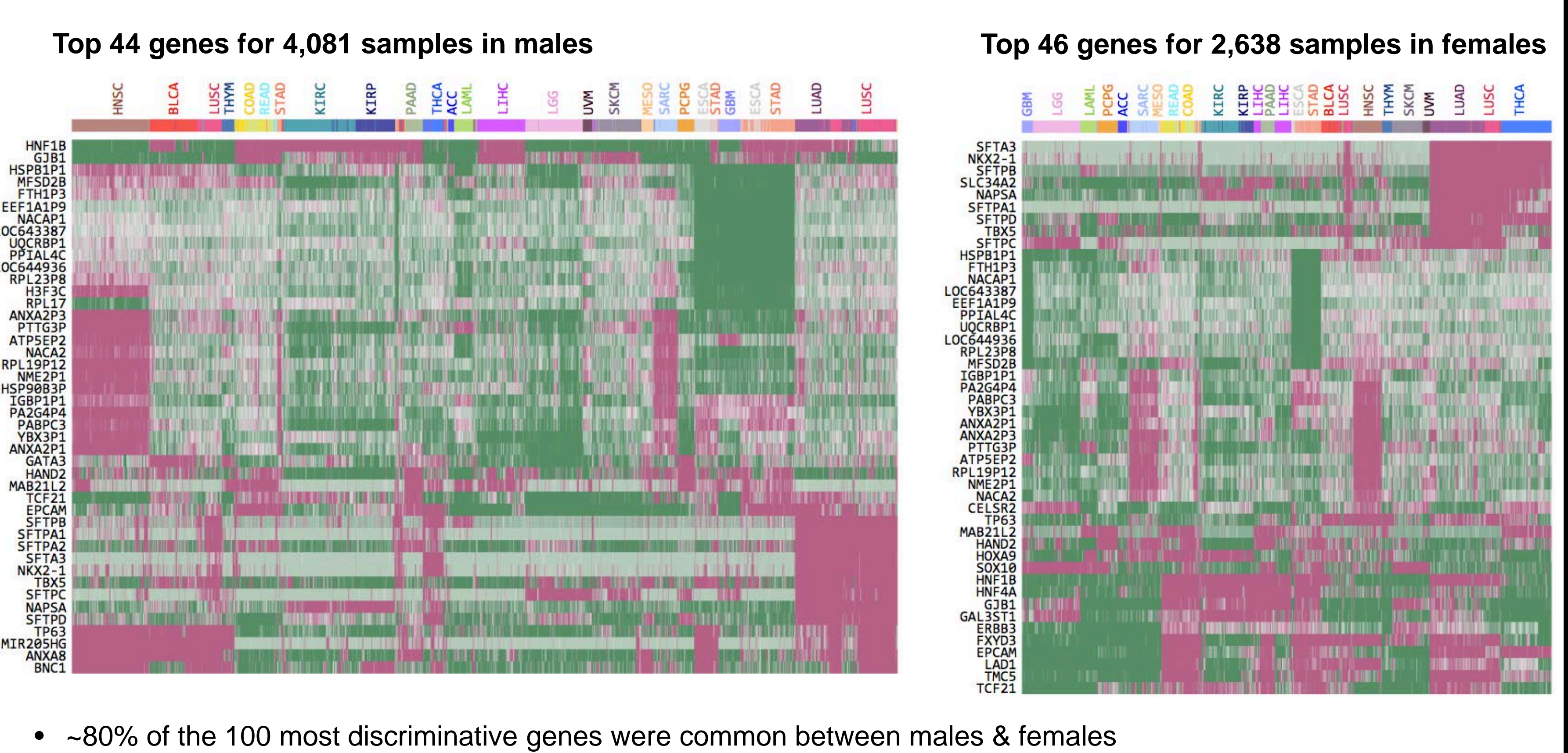
Results

Testing performances of GA/KNN

Pan-cancer performance ignoring gender

Pan-cancer performance: male   Pan-cancer performance: female

- correctly classify >90% of the test set samples in pan-cancer classification ignoring gender in 2,000 runs
- the results recapitulated those from our analysis of 23 tumor types that ignored sex in 5,000 runs

Top genes selected for pan-cancer (ignoring gender)

DAVID's gene ontology of top 200 genes

	p-value (-log)
Regulation of transcription, DNA-dependent	1.5e-6
Tube development	1.2e-6
Epithelium development	2.5e-7
Multicellular organismal development	3.7e-9
Organ development	2.7e-9
Tissue development	1.4e-9
System development	1.1e-9
Developmental process	5e-10
Anatomical structure morphogenesis	3.7e-10
Anatomical structure development	3.2e-10

Gene selection frequency

Heatmap of top 50 genes across all 9,096 samples

A hierarchical clustering was carried out for both samples and genes based on Euclidean distance

Heatmaps of top genes selected for sex non-specific tumors

Top 44 genes for 4,081 samples in males   Top 46 genes for 2,638 samples in females

- ~80% of the 100 most discriminative genes were common between males & females
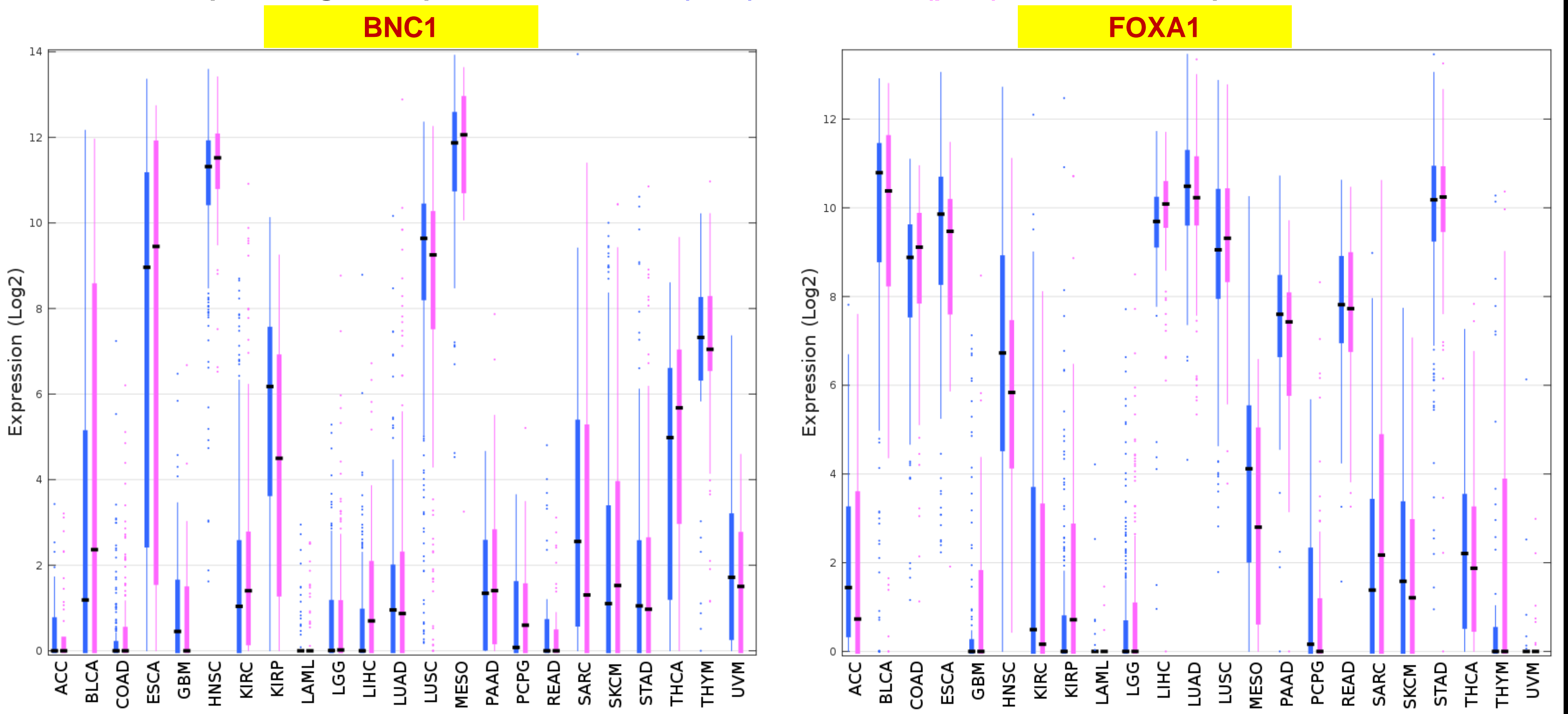
Genes that differ in rank between (males & females)

- Among the top 100 genes from each sex, 21 differed by more than 100 in rank

	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

Boxplots of gene expression for males (blue) vs females (pink) in 23 sex non-specific tumors

BNC1   FOXA1

Acknowledgments