Supplementary Figures

Figures S1-S3. Distribution of Classes for each Dataset. Y-axis is the number of samples. The X-axis lists the classes used for binary and multiclass (MC) classifications. Colors group classes of similar type. Each figure depicts a different dataset for RBM (S9), NCBI (S10), TCGA (S11)

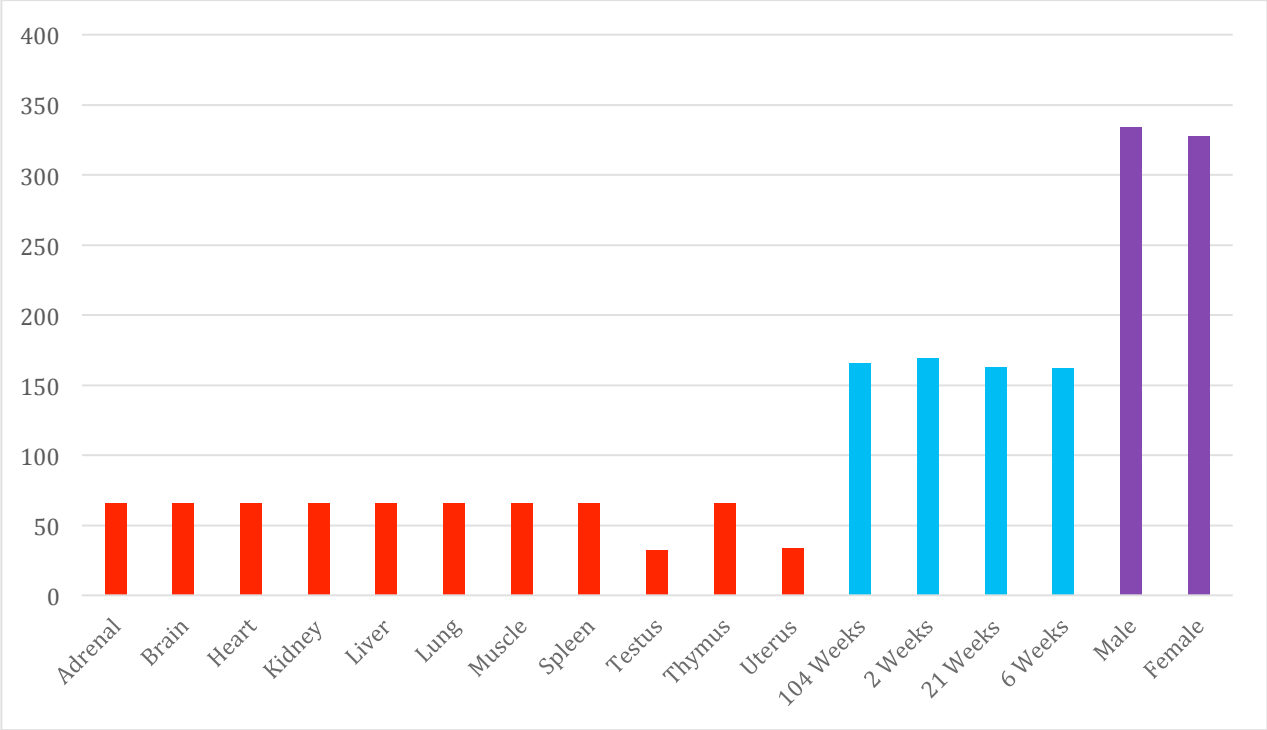

Figure S1- RBM

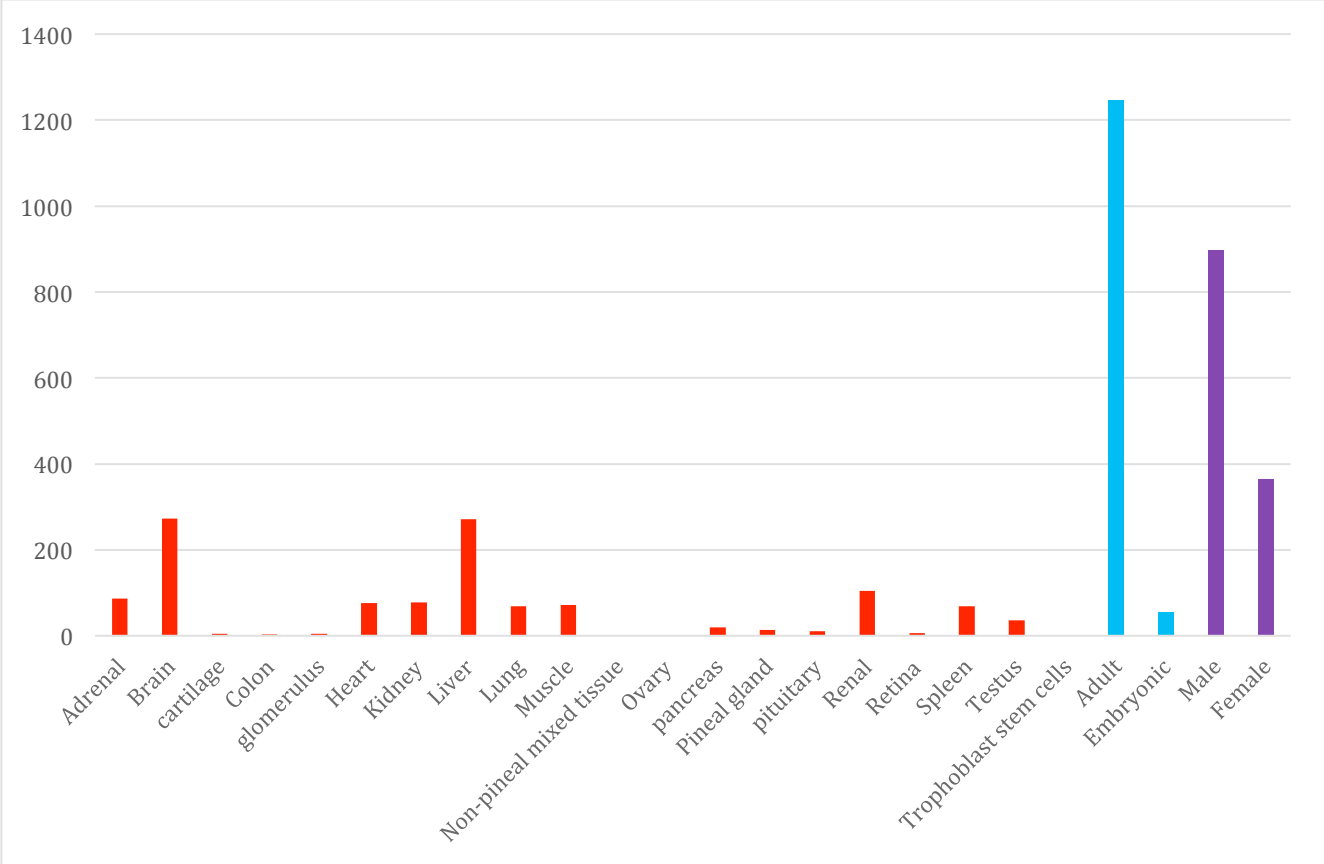

Figure S2-NCBI

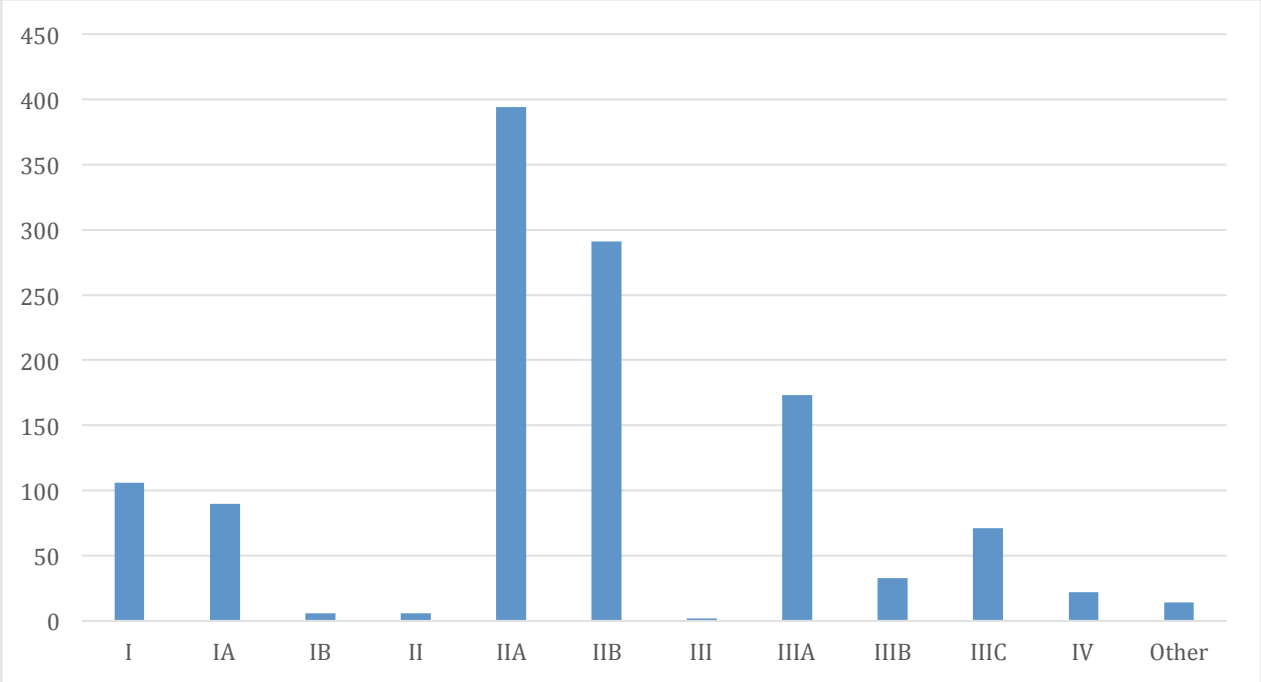

Figure S3- TCGA

Figures S4-S7 Comparing the Number of Features Selected During Feature Selection Protocol for Gene vs Isoform Based Classification. The Y-axis is the number of features. The X-axis lists the classes used for binary and multiclass (MC) classifications. Each figure depicts an individual dataset: RBM (S4), NCBI (S5), TCGA-$\log_2$ Normalized (S6), and TCGA-Raw Count (S7).

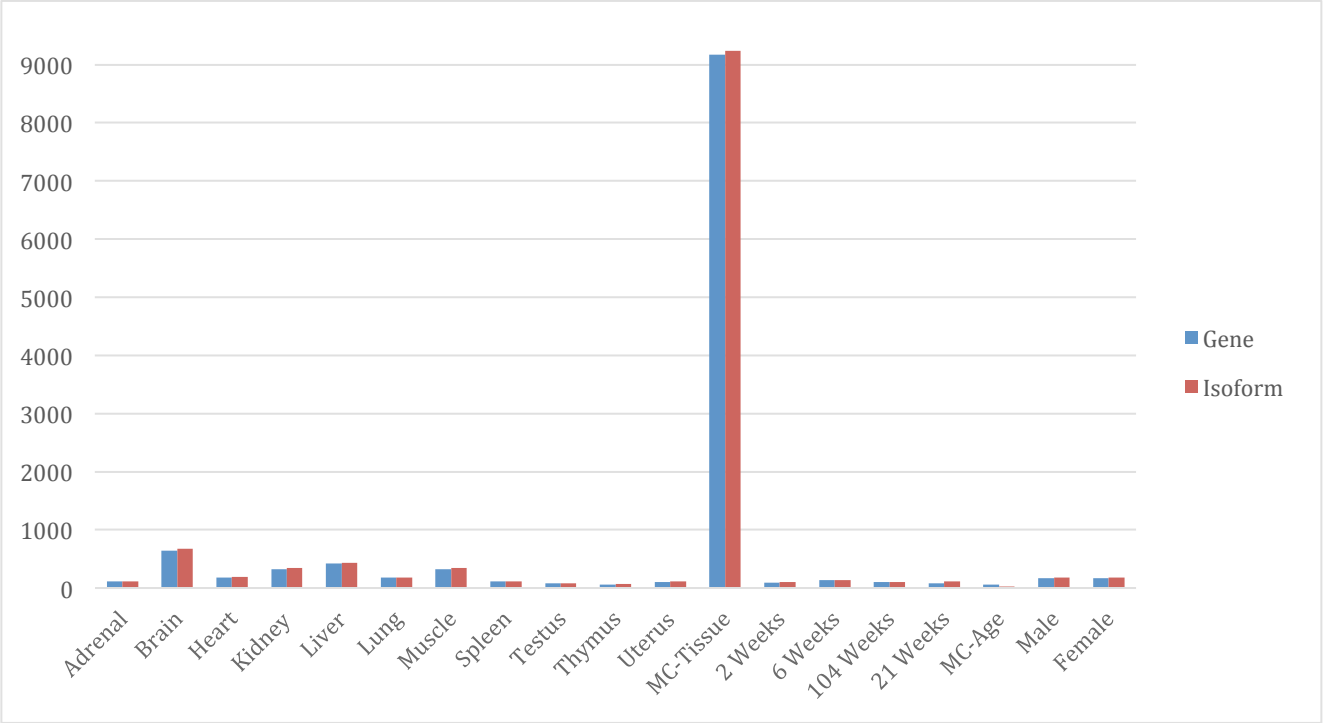

Figure S4-RBM

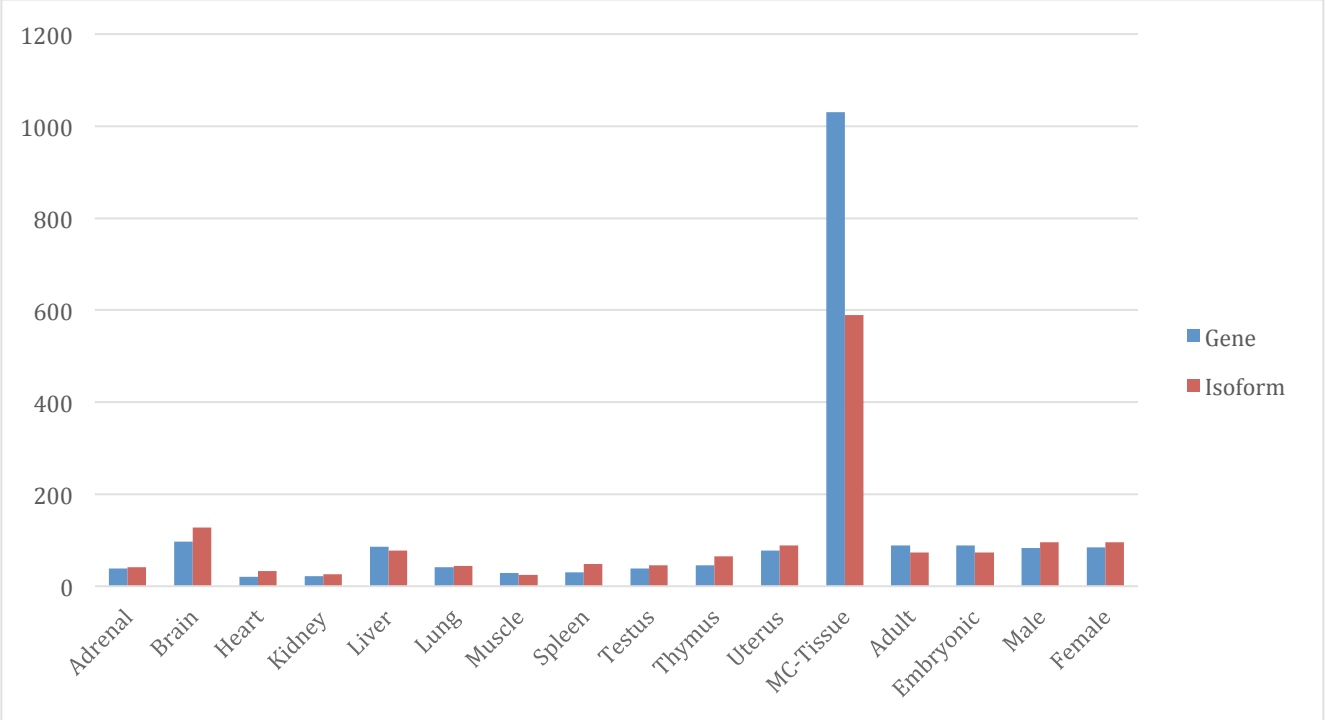

Figure S5-NCBI

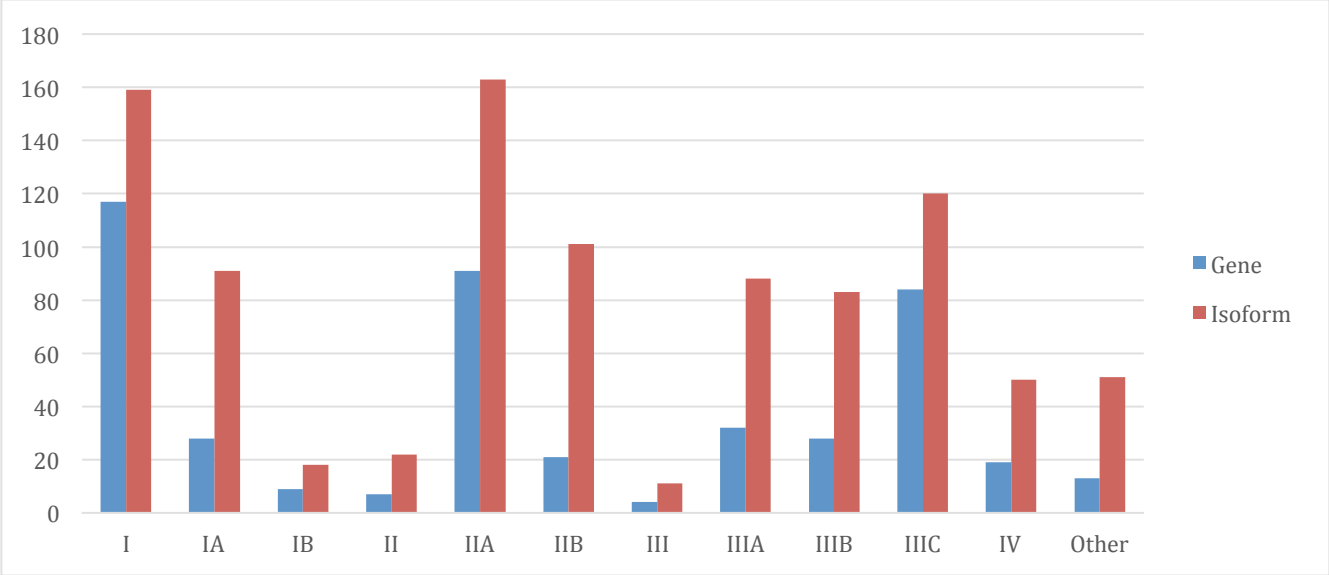

Figure S6- TCGA Log_2 Normalized

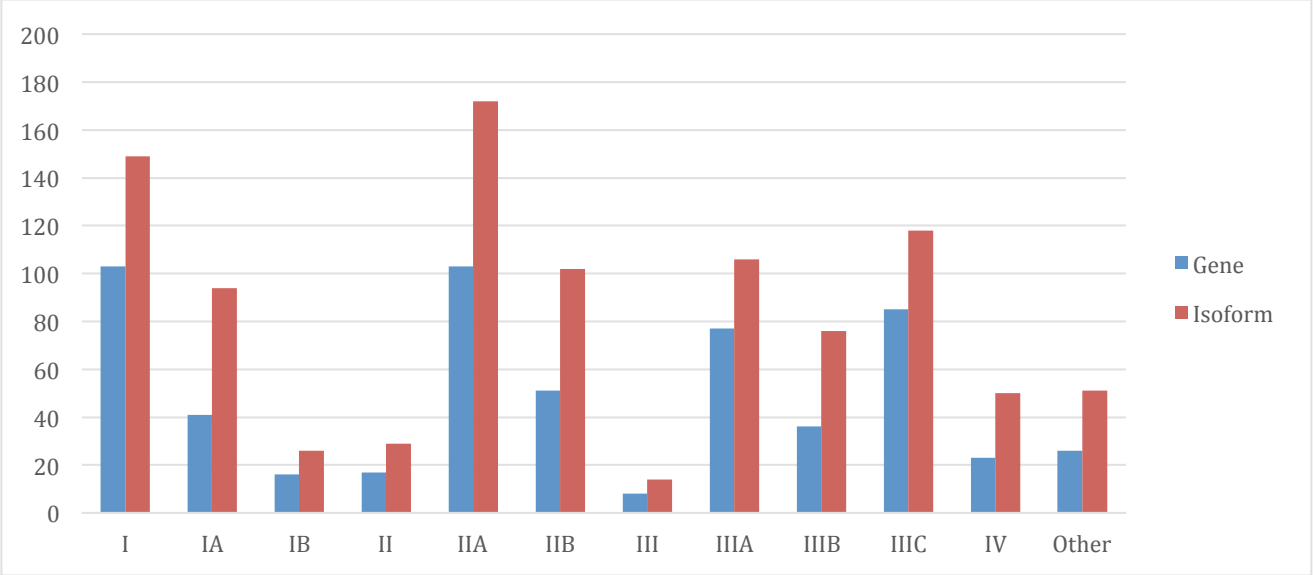

Figure S7 – TCGA Raw Counts

Figures S8–S11 Comparing the Number of Features Selected Post-Feature Selection across Normalization Techniques for Gene Features. The Y-axis is the number of features. The X-axis lists the classes used for binary and multiclass (MC) classifications. Each figure depicts a different dataset for RBM (S8), NCBI (S9), TCGA-$\log_2$ Normalized (S10), and TCGA-Raw Count (S11).

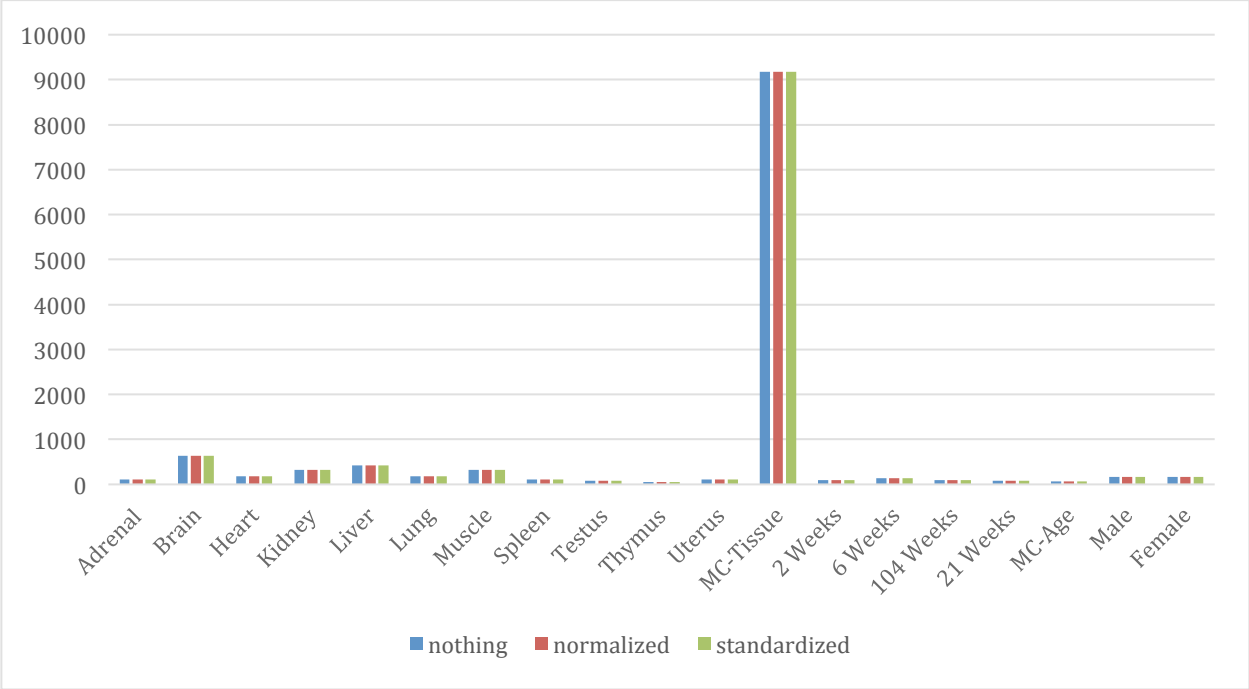

Figure S8 – RBM

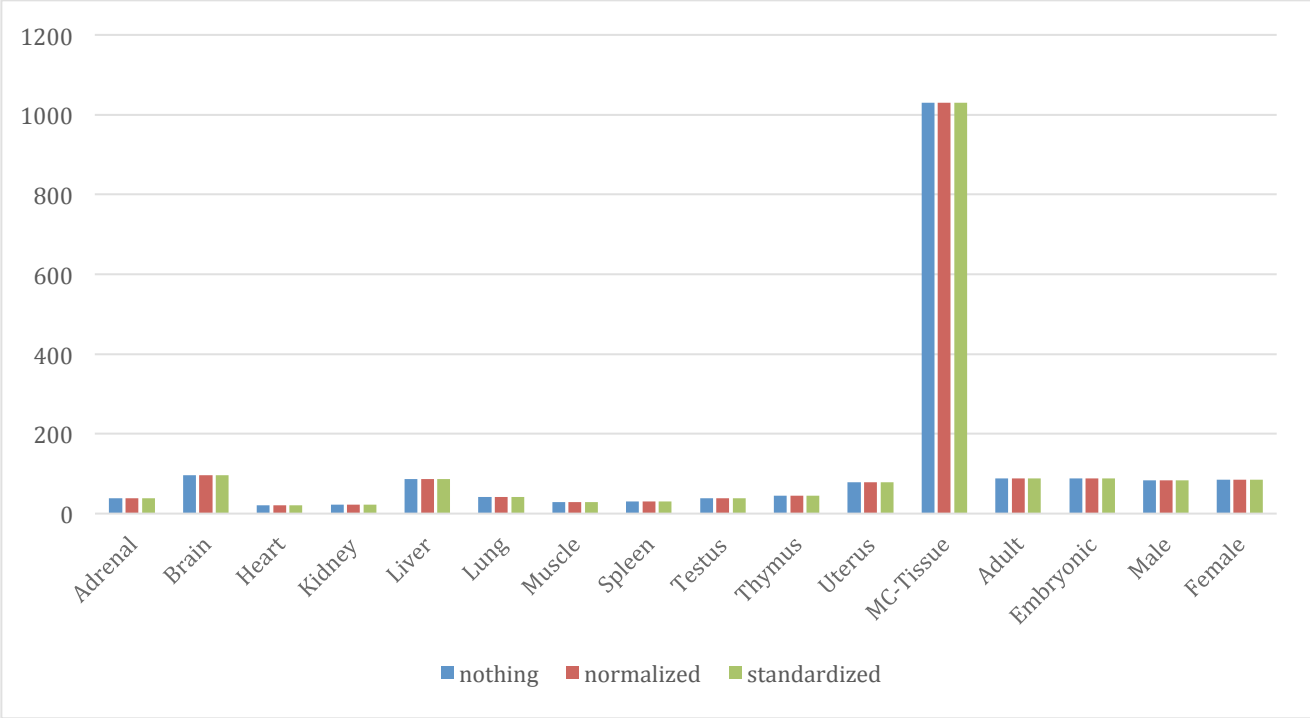

Figure S9-NCBI

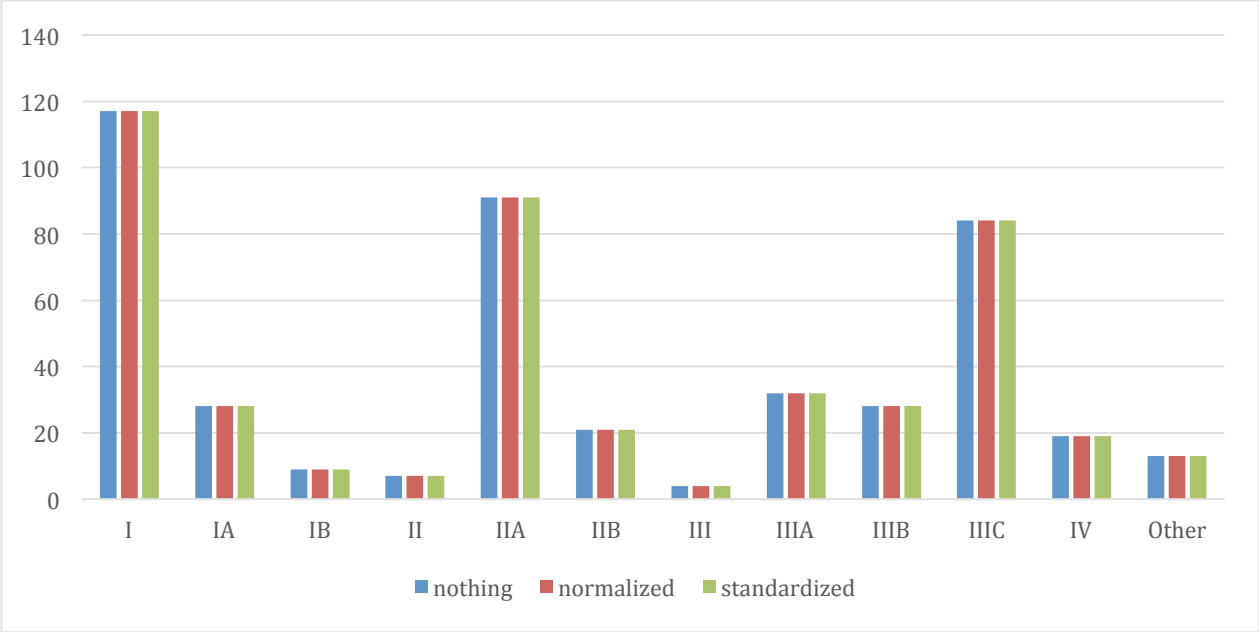

Figure S10-TCGA Log_2 Normalization

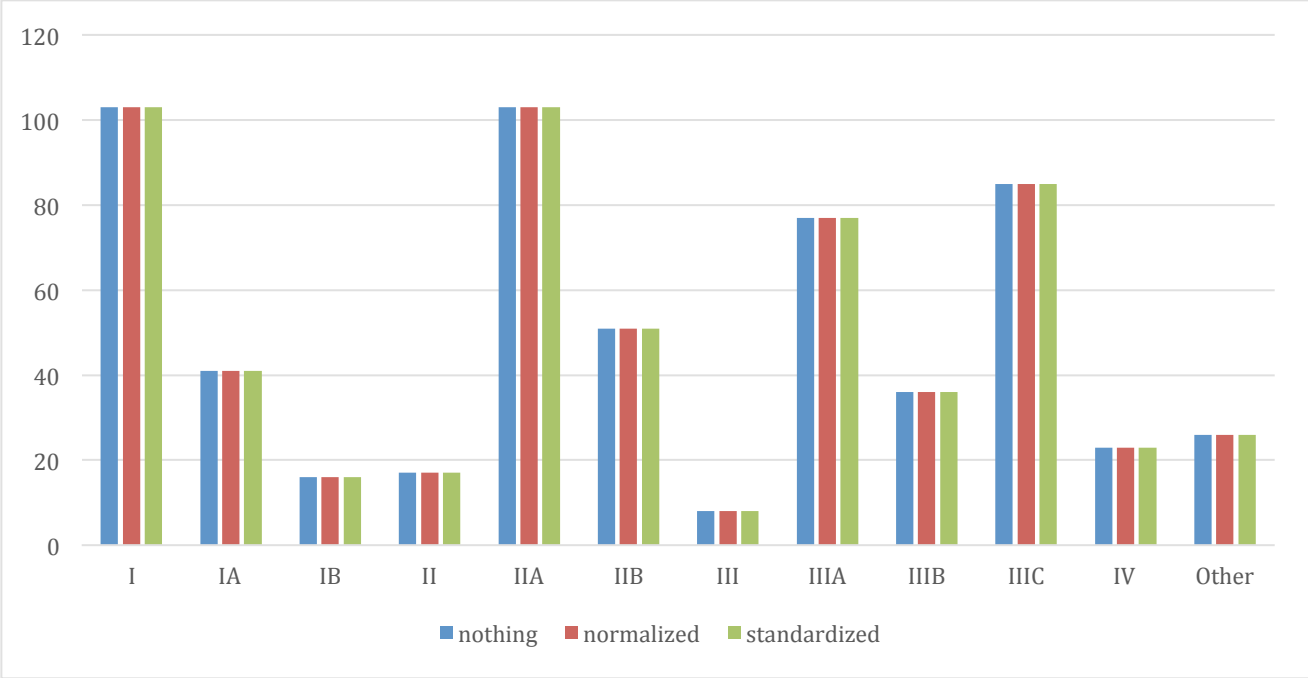

Figure S11-TCGA Raw Count

Figures S12-S13. Heat Map representation of *f* -measure standard deviation across the 10-Fold Cross Validation across machine learning methods, classes, datasets, and normalization techniques. For the majority of classification tasks, the standard deviation was less than 1% for both Gene (S12) and Transcript (S13). The top x-axis represents normalization techniques including Nothing (no normalization), Standardized, and Normalized. The bottom x-axis represents the machine learning techniques (DT = Decision Table, J48 = J48 Decision Tree, LR = Linear Regression, NB = Naïve Bayes, RF = Random Forest, SVM = Support Vector Machine). The y-axis represents the classes where MC stands for multiclass. Datasets for each panel are (A) RBM, (B) NCBI, (C) TCGA – $\log_2$ normalized counts, (D) TCGA – raw counts.

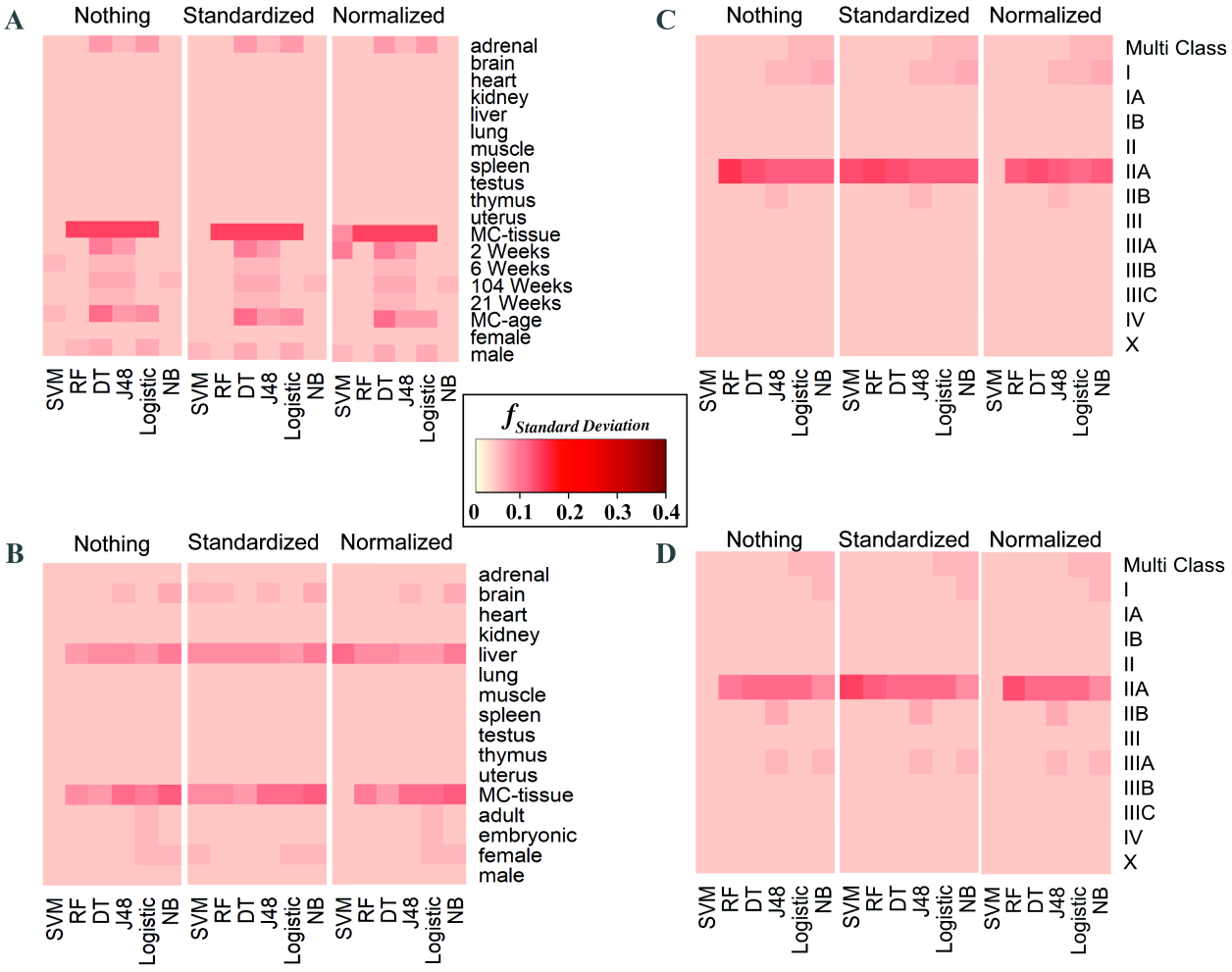

Figure S12-Gene

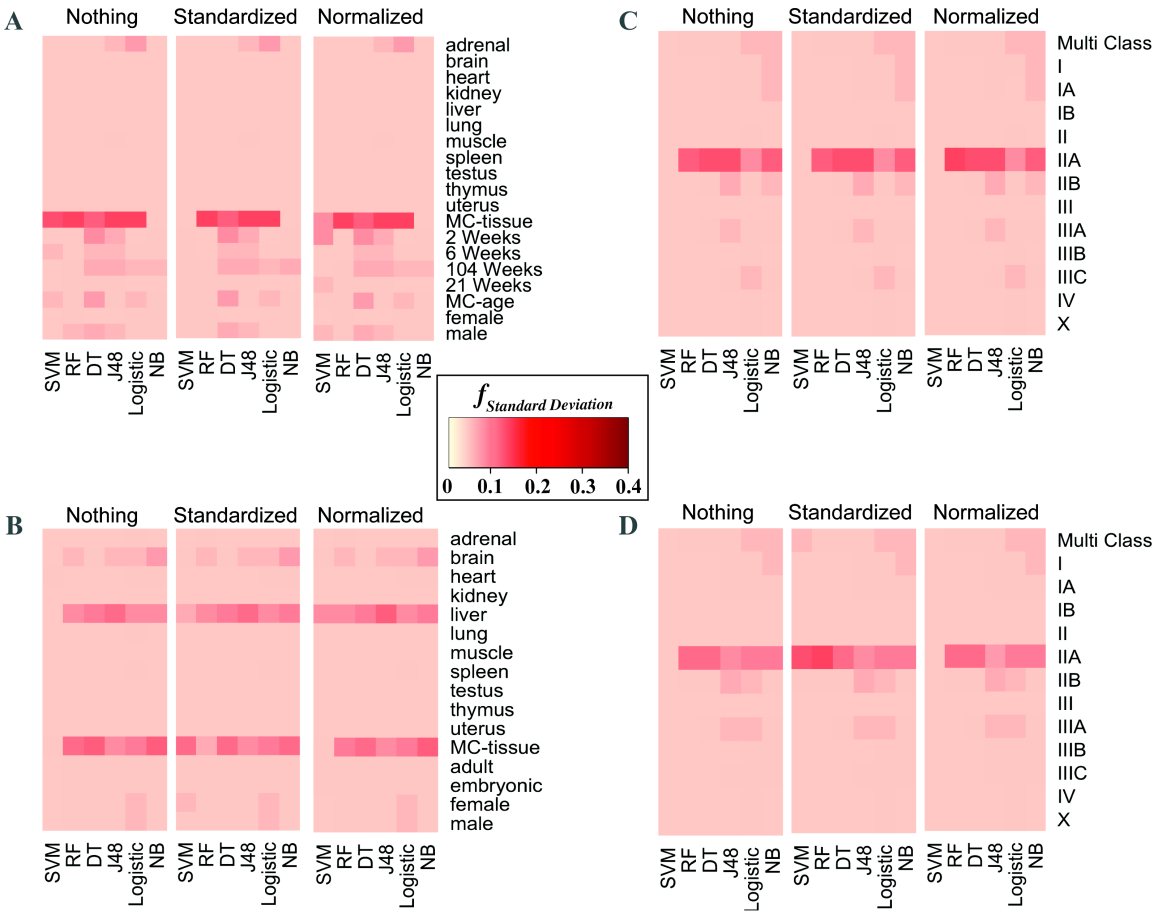

Figure S13- Transcript

Supplementary Tables

Table S1- List of all NCBI SRA Project IDs for NCBI Dataset

Accession Number	Number of Samples
SRP037986	662
SRP023266	144
SRP041131	125
SRP039021	116
SRP028932	48
SRP016501	27
SRP036442	24
SRP041119	16
SRP045777	16
SRP021090	14
SRP028515	12
SRP029760	12
SRP041920	12
SRP055430	12
SRP042370	10
SRP021119	8
SRP046247	8
SRP018407	6
SRP044684	6
SRP046248	6
SRP035358	4
SRP041741	4
SRP045117	4
SRP051483	4
SRP009272	2
SRP013262	2
SRP017140	2
SRP029980	2
SRP047494	1

Table S2. Initial number of features, and the average number of features after feature selection procedure across different classification problems

	Initial		After Feature Selection	
	Gene	**Transcript**	**Gene**	**Transcript**
RBM	25,538	29,130	659	671
NCBI	10,711	17,506	119	97
TCGA-Raw	20,524	73,592	49	82
TCGA-log_2 Normalized	20,524	73,592	38	80

Table S3- Number of Features Selected: Binary classification (average) and Multiclass classification

	Binary-Average		Multiclass	
	Gene	**Transcript**	**Gene**	**Transcript**
RBM-Tissue	230	242	9173	9237
RBM-Age	102	114	60	27
NCBI-Tissue	48	97	1030	589
TCGA-Raw-Cancer Stage	38	80	20	73
TCGA-log_2 Normalized-Cancer Stage	49	82	20	73