De novo pathway-based biomarker identification

Supplementary Material

Nicolas Alcaraz, Markus List, Richa Batra, Fabio Vandin,
Henrik J. Ditzel and Jan Baumbach

1 Supplementary Material and Methods

1.1 Meta-gene scoring

In order to produce to an activity matrix to score the gene sets representing pathways, we used the single sample Gene Set Enrichment Method (ssGSEA) [4] as implemented in the GSVA [3] R package with default parameters. We briefly present the ssGSEA procedure as described in [1]:

Given a gene set G of size N_G and a single sample S of the data set of N_G genes, the gene values in the measurement are replaced by their ranks according to their absolute expression values, from high to low $L = \{r_1, r_2, ..., r_N\}$. An enrichment score for $ES(G, S)$ is then obtained by the sum of the difference between the weighted empirical cumulative distribution function (ECDF) of the genes in the measurement

$$ES(G,S) = \sum_{i=1}^{N} |P_G^W(G,S,i) - P_{N_G}(G,S,i)|$$

where:

$$P_G^W = \sum_{r_j \in G, j \leq i} \frac{|r_j|^\alpha}{\sum_{r_j \in G} |r_j|^\alpha}$$

and

$$P_{N_G} = \sum_{r_j \in G, j \leq i} \frac{1}{N - N_G}$$

The calculation is repeated for each gene set and each in the dataset. The exponent α is set to it's default value of 1/4. For the case when a gene in the gene set is not found in the dataset, then it's set to the average expression value of all genes for that sample.

1.2 Removal of correlated features

Features were clustered based on their Spearman's correlation coefficient using TransClust [5], with a threshold value of 0.9. This produced clusters of features where the average Spearman correlation value of all pairs of features within each cluster was above 0.9. Finally, the feature with the highest average similarity within the cluster was taken as cluster representative, while all other features in the cluster were discarded from further workflow steps.

1.3 Feature selection and model building

To select a small set of predictive features we used the varSelRF R package, which performs feature selection with Random Forests using a recursive feature elimination approach. The feature selection procedure implemented in varSelRF [2] starts by building a random forest with all features and then iteratively proceeds to remove 20% of the least important features. This procedure is repeated until a model with only two features is left. The model with the lowest out-of-bag error (OOB) is reported as final solution.

1.4 Evaluation measures

The F-score is defined as

$$2 * \frac{P * R}{P + R}$$

Where P is precision, defined $P = \frac{tp}{tp+fp}$ and R is recall, defines as $P = \frac{tp}{tp+tn}$. And tp = true positives, tn = true negatives, fp = false positives.

Let A and B be two sets, the Jaccard Index is defined as following:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

1.5 Random networks

Two types of networks randomization strategies were tested: i) node label permutation and ii) degree preserving rewiring.

Given the parameter L, degree preserving rewiring consists of the following steps while $L > 0$:

1. Randomly sample two edges $e_1 = (u_1, v_1)$ and $e_2 = (u_2, v_2)$ from $E(G)$
2. If $e_{n1} = (u_1, v_2) \notin E(G)$ and $e_{n2} = (u_2, v_1) \notin E(G)$, create edges e_{n1}, e_{n2} and remove e_1, e_2 from $E(G)$, decrease $L := L - 1$, else go to 1.

Where L was set to $4 * |E|$ for each network. For each randomization strategy, 20 networks were generated and the 5-fold cross-validation scheme was executed with the full pipeline for all types of features on the TCGA genes expression (Figure S6) and DNA methylation (Figure S7) cohorts.

2 Supplementary Tables

Table S1: Number of samples by PAM50 subtypes for TCGA and ACES datasets

Dataset	Basal	Her2	LumA	LumB
TCGA	98	58	230	127
Desmedt-June07	42	23	68	43
Hatzis-Pusztai	73	20	28	26
Ivshina	14	21	36	21
Loi	4	6	20	22
Miller	9	11	43	18
Minn	21	12	13	15
Pawitan	15	18	57	45
Schmidt	21	17	74	47
Symmans	7	16	90	84
WangY	36	10	4	2
WangY-ErasmusMC	51	34	89	70
Zhang	4	3	62	47

Table S2: KEGG pathways related to the hallmarks of cancer

KEGG ID	Pathway Name	Hallmark ID
hsa03410	Base excision repair	7
hsa03420	Nucleotide excision repair	7
hsa03430	Mismatch repair	7
hsa03440	Homologous recombination	7
hsa03450	Non-homologous end-joining	7
hsa04010	MAPK signaling pathway	1
hsa04012	ErbB signaling pathway	1
hsa04070	Phosphatidylinositol signaling system	1
hsa04150	mTOR signaling pathway	1
hsa04310	Wnt signaling pathway	2
hsa04330	Notch signaling pathway	3
hsa04350	TGF-beta signaling pathway	4,8,10
hsa04370	VEGF signaling pathway	3
hsa04110	Cell cycle	1
hsa04115	p53 signaling pathway	4,5,6,7
hsa04210	Apoptosis	5
hsa04510	Focal adhesion	4,9
hsa04520	Adherens junction	9
hsa04640	Hematopoietic cell lineage	8,10
hsa04610	Complement and coagulation cascades	8,10
hsa04620	Toll-like receptor signaling pathway	8,10
hsa04621	NOD-like receptor signaling pathway	8,10
hsa04622	RIG-I-like receptor signaling pathway	8,10
hsa04623	Cytosolic DNA-sensing pathway	8,10
hsa04650	Natural killer cell mediated cytotoxicity	8,10
hsa04612	Antigen processing and presentation	8,10
hsa04660	T cell receptor signaling pathway	8,10
hsa04662	B cell receptor signaling pathway	8,10

KEGG ID	Pathway Name	Hallmark ID
hsa04664	Fc epsilon RI signaling pathway	8,10
hsa04666	Fc gamma R-mediated phagocytosis	8,10
hsa04670	Leukocyte transendothelial migration	8,10
hsa04672	Intestinal immune network for IgA production	8,10
hsa04062	Chemokine signaling pathway	8,10
hsa00030	Pentose phosphate pathway	6
hsa04512	ECM-receptor interaction	9
hsa04060	Cytokine-Cytokine receptor interaction	1,10
hsa04024	cAMP signaling pathway	1
hsa04151	PI3K-Akt signaling pathway	1,4
hsa04630	Jak-STAT signaling pathway	9,10
hsa03320	PPAR signaling pathway	1
hsa04611	Patelet Activation	8
hsa00010	Glycolysis / Gluconeogenesis	6
hsa00190	Oxidative phosphorylation	6
hsa00020	Citriate cycle (TCA cycle)	6
hsa00260	Glycine serine and threonine metabolism	6
hsa00471	D-Glutamine and D-Glutamate metabolsim	6
hsa00330	Arginine and proline metabolsim	6
hsa04066	HIF-1 signaling pathway	6
hsa00250	Alanine, aspartate and glutamate matabolism	6
hsa00564	Glycerophopholipid metabolism	4
hsa04810	Regulation of actin cytoskeleton	1,3,10
hsa05230	Central Carbon metabolism in cancer	6
hsa05231	Choline metabolism in cancer	1,3,5,10
hsa04064	NF-kappa B signaling pathway	8,10

Table S3: Hallmarks of Cancer

Hallmark ID	Hallmark Name
1	Sustaining Proliferative Signaling
2	Enabling Replicative Immortality
3	Inducing Angiogenesis
4	Evading Growth Suppressors
5	Resisting Cell Death
6	Deregulating Cellular Energetics
7	Genome Instability and Mutation
8	Avoiding Immune Destruction
9	Activating Invasion and Metastasis
10	Tumor-promoting Inflammation

Table S4: Average overlap of genes contained in the selected features for all runs within the CV loop for TCGA

	KPM_HTRIdb	KPM_HNET	KPM_HPRD	KPM_I2D	CPDB	MSIG	SG	PAM50
KPM_HTRIdb	516.80	149.10	101.28	182.36	119.64	210.28	45.40	25.24
KPM_HNET		369.56	116.36	186.44	127.92	196.62	41.82	27.78
KPM_HPRD			243.66	141.90	91.36	129.82	26.52	17.98
KPM_I2D				475.92	131.22	220.54	43.30	23.82
CPDB					1942.42	574.50	32.34	26.96
MSIG						2832.48	68.48	42.66
SG							82.54	15.42
PAM50								50

Table S5: Overlap of genes contained in the selected features for the final models build from the full TCGA data set

	KPM_HTRIdb	KPM_HNET	KPM_HPRD	KPM_I2D	CPDB	MSIG	SG	PAM50
KPM_HTRIdb	736	210	106	308	162	309	89	32
KPM_HNET		293	86	158	121	156	56	26
KPM_HPRD			169	96	81	97	29	15
KPM_I2D				458	166	186	57	25
CPDB					2570	574	46	29
MSIG						2254	87	41
SG							107	24
PAM50								50

3 Supplementary Figures

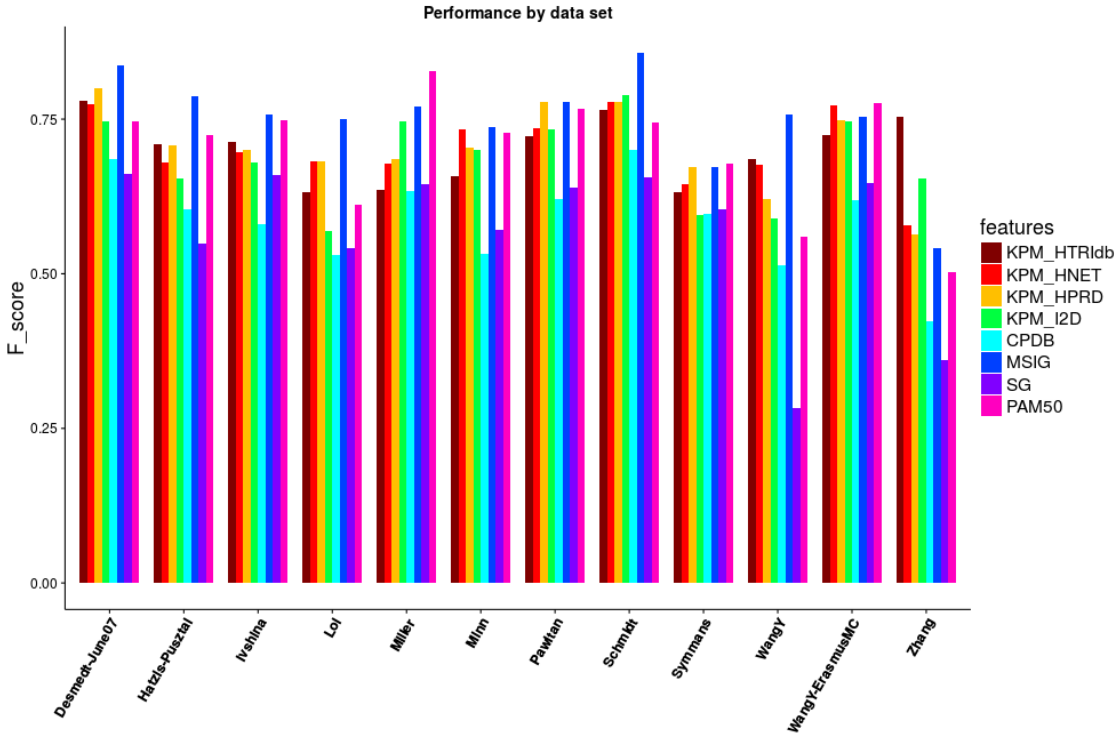


Figure S1: Performance (F-score) of all features in each of the 12 ACES validation datasets.

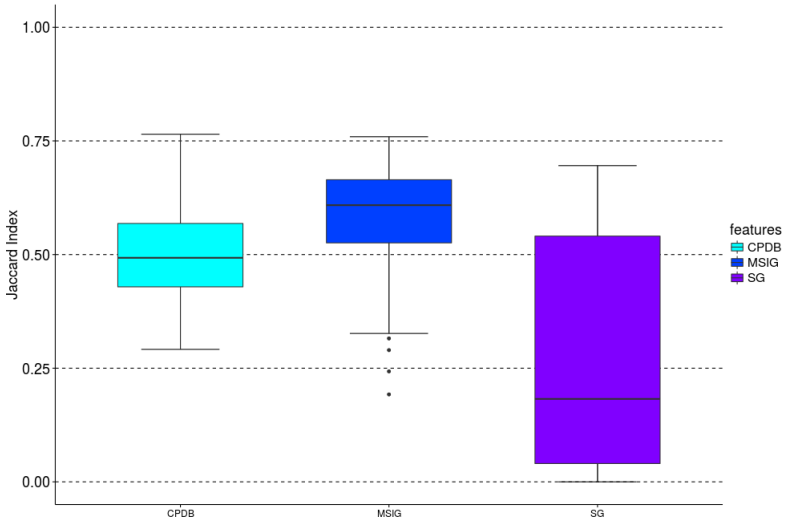


Figure S2: Stability of MG features from *a priori* pathways inside the cross-validation evaluation scheme for the TCGA DNA methylation cohort. The Jaccard Index (a) was computed for the selected features of each pair of folds.

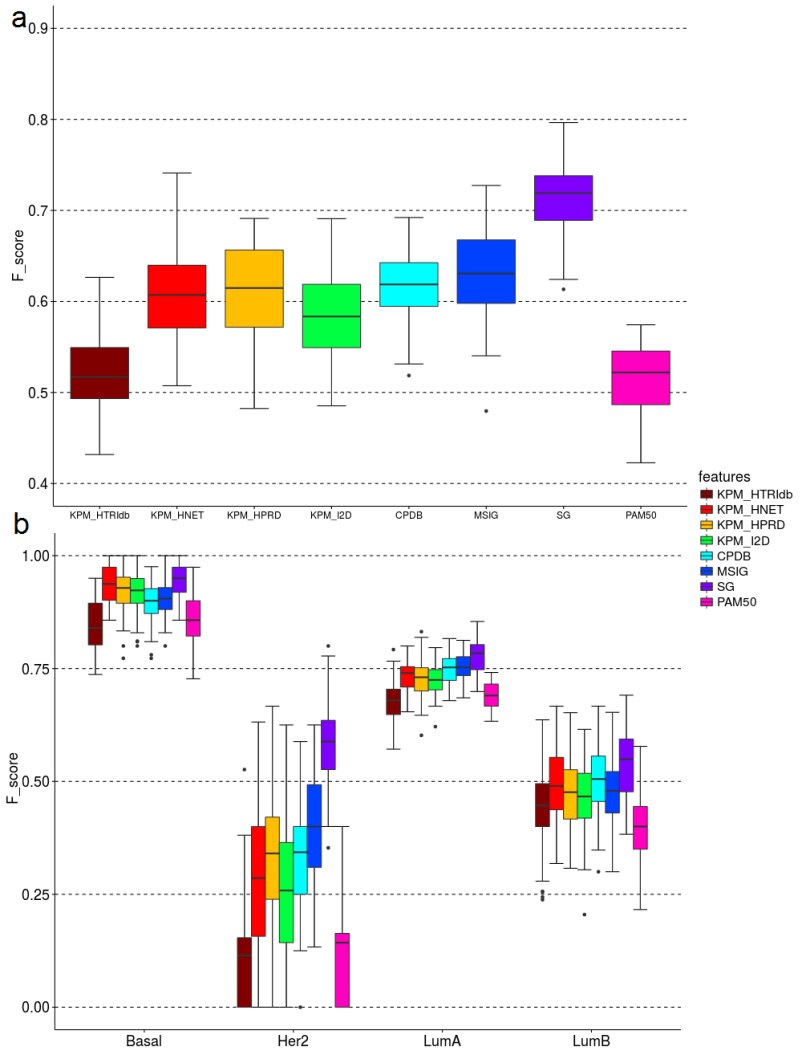
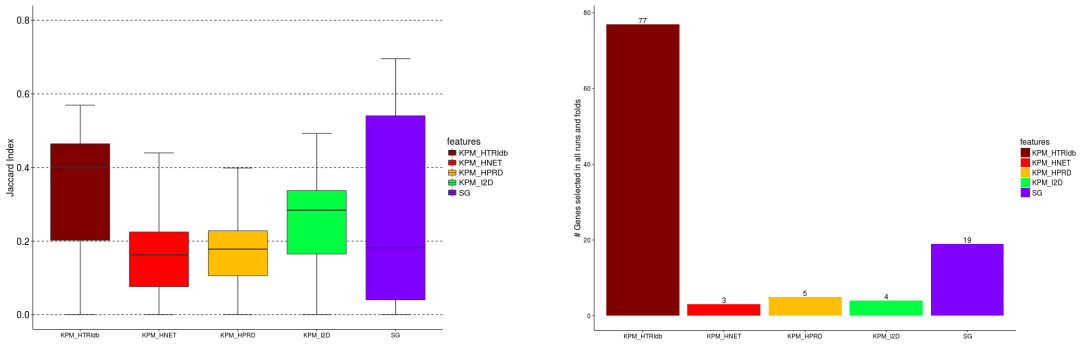
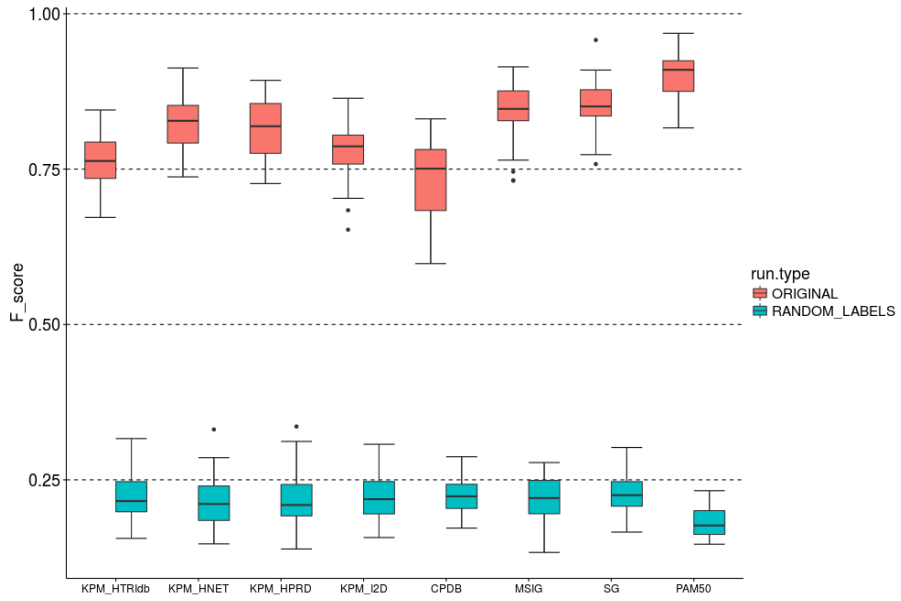


Figure S3: Prediction performance (F-score) for the different models in the TCGA DNA methylation. Top figure (a) corresponds to the overall performance, bottom figure (b) to performance by subtype.

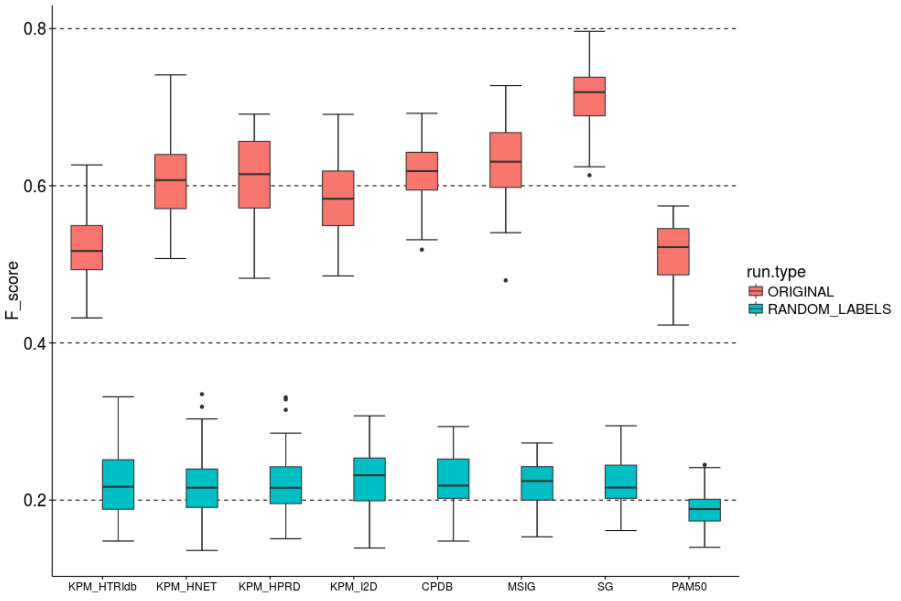


(a) (b)

Figure S4: Stability of gene markers from *de novo* pathways inside the cross-validation evaluation scheme for the TCGA DNA methylation cohort. The Jaccard Index (a) was computed for the genes within the selected features for each pair of folds. In (b) the number of genes that were selected for all runs.

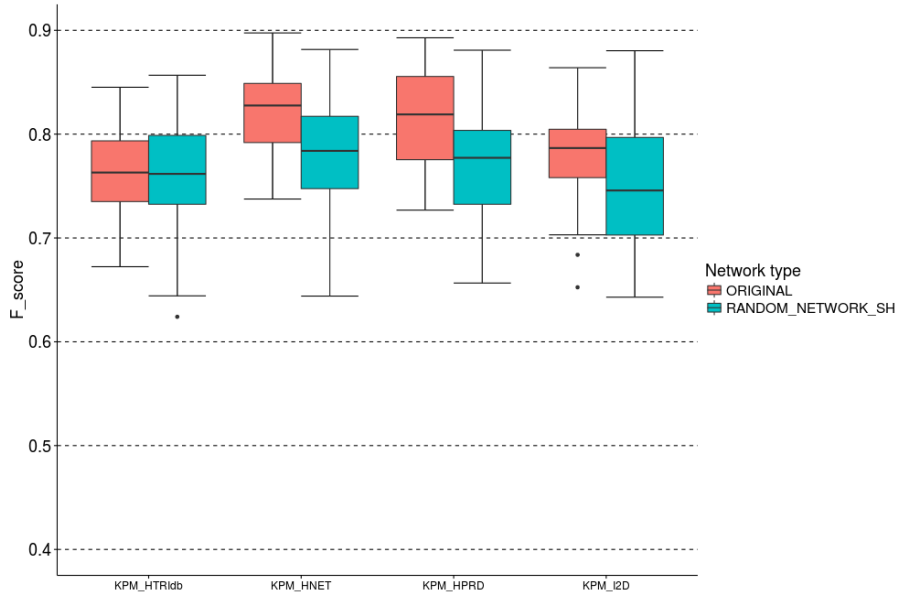


(a)

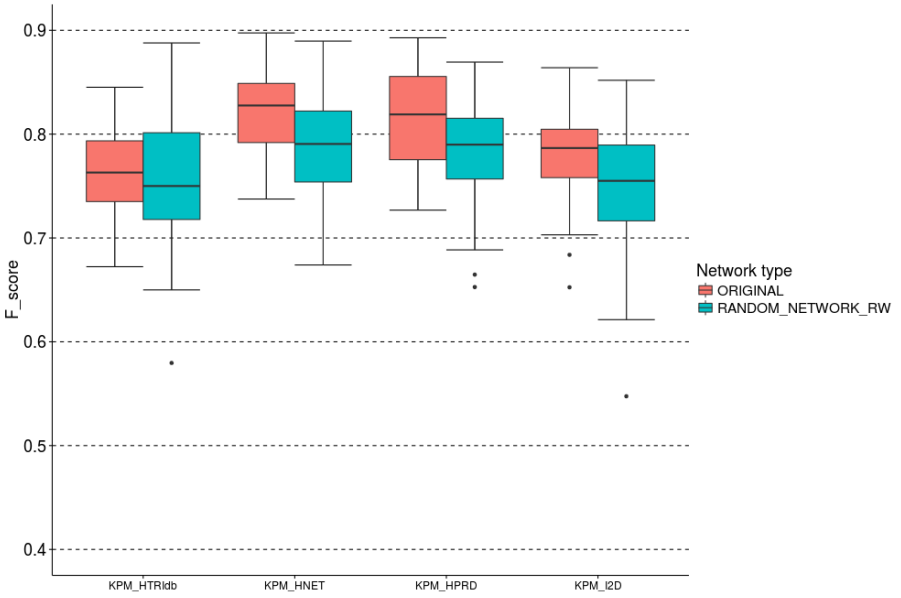


(b)

Figure S5: Evaluation on randomized sample labels for the TCGA gene expression (a) and DNA methylation (b) cohort.

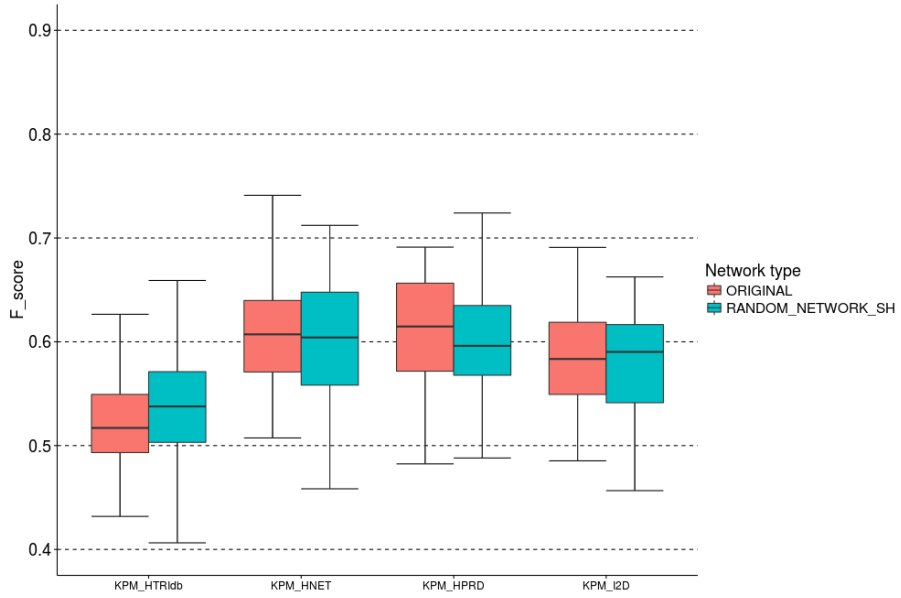


(a)

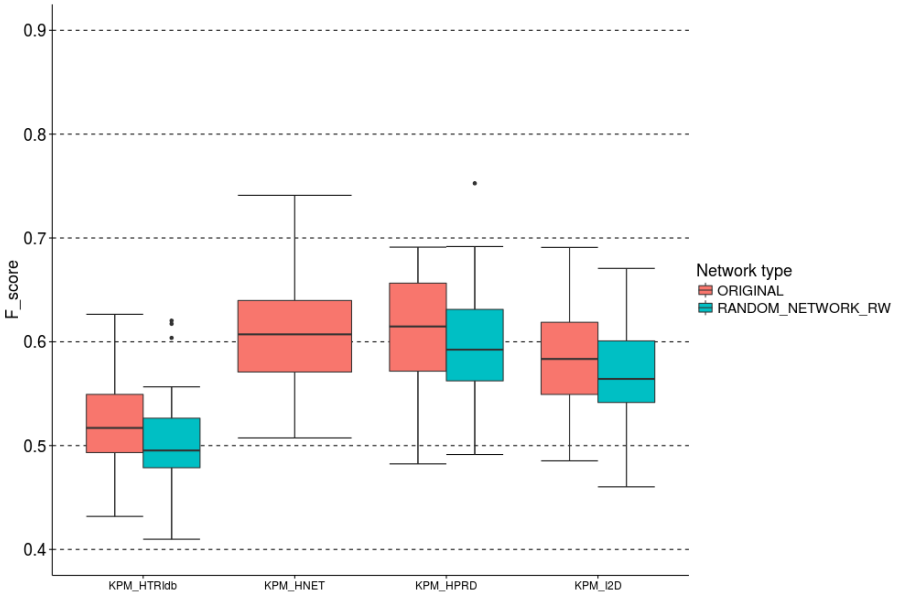


(b)

Figure S6: Evaluation on randomized networks for the TCGA gene expression cohort. Two randomization strategies were used: node label permutation (SH) and degree-preserving edge rewiring (RW)



(a)



(b)

Figure S7: Evaluation on randomized networks for the TCGA DNA methylation cohort. Two randomization strategies were used: node label permutation (SH) and degree-preserving edge rewiring (RW)

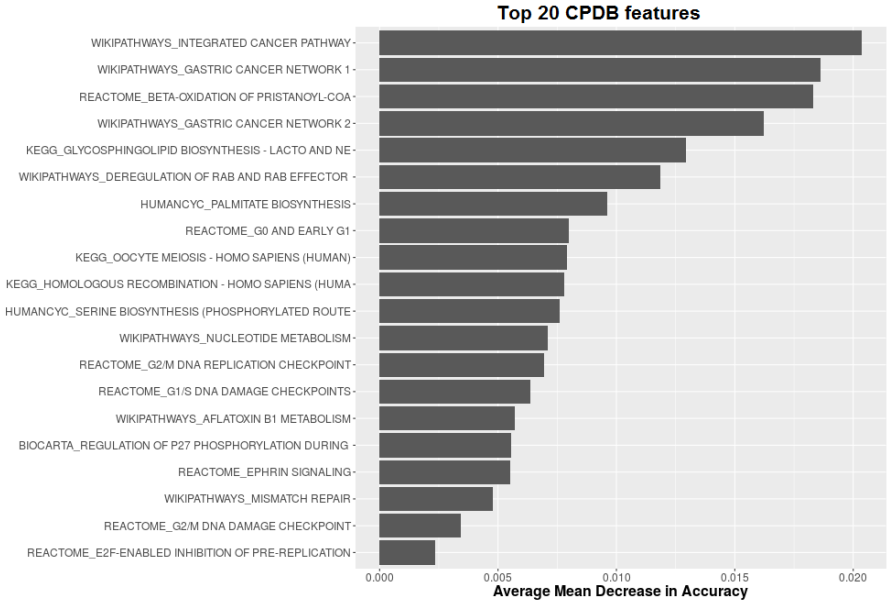


Figure S8: Top 20 most frequently selected CPDB features after the feature selection step in each cross-validation run from the TCGA gene expression data sets. Features were sorted afterwards by their averaged mean decrease in accuracy given by the random forest model.

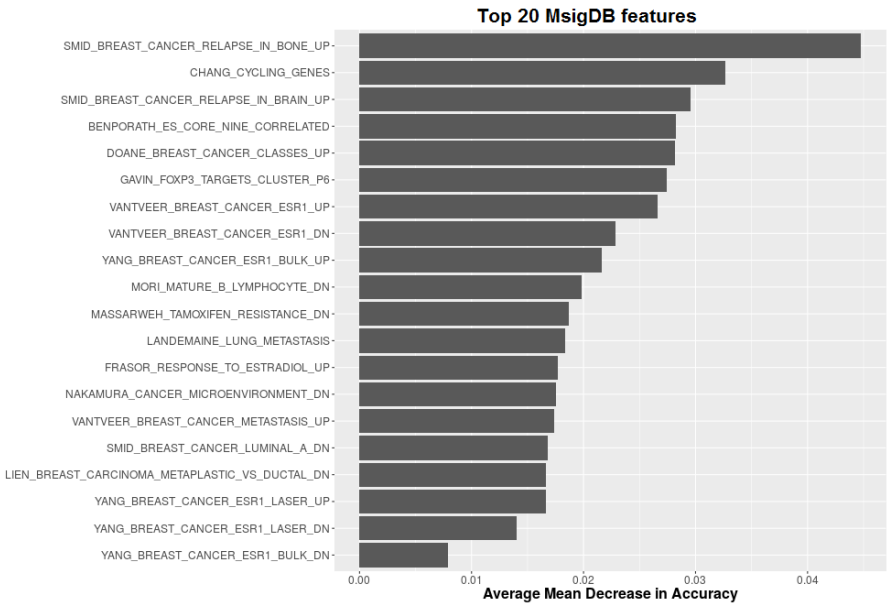


Figure S9: Top 20 most frequently selected MsigDB features after the feature selection step in each cross-validation run from the TCGA gene expression data sets. Features were sorted afterwards by their averaged mean decrease in accuracy given by the random forest

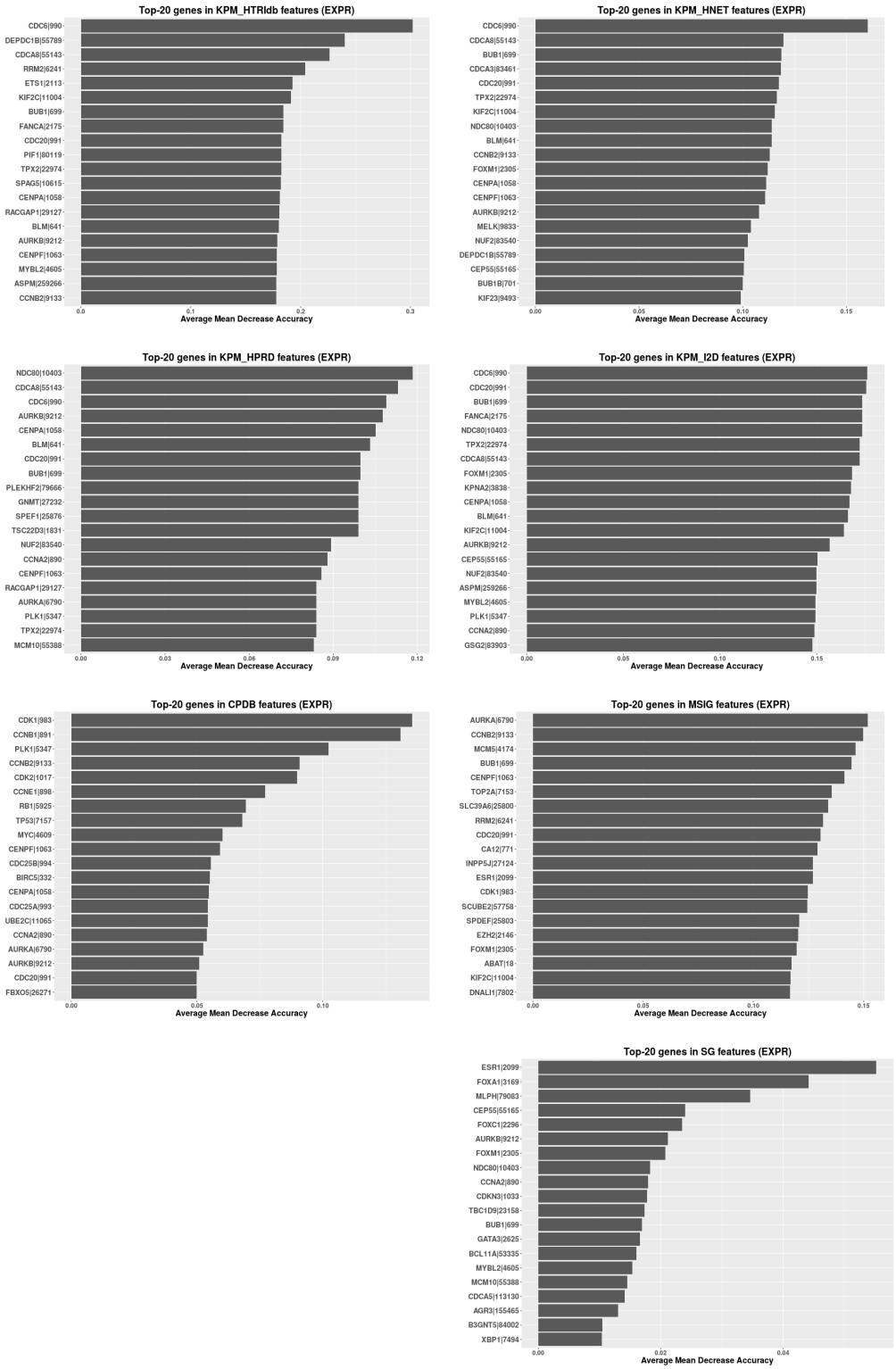


Figure S10: The top-20 most frequently selected genes (TCGA gene expression) for each feature type, sorted by their average mean decrease in accuracy from the random forest models

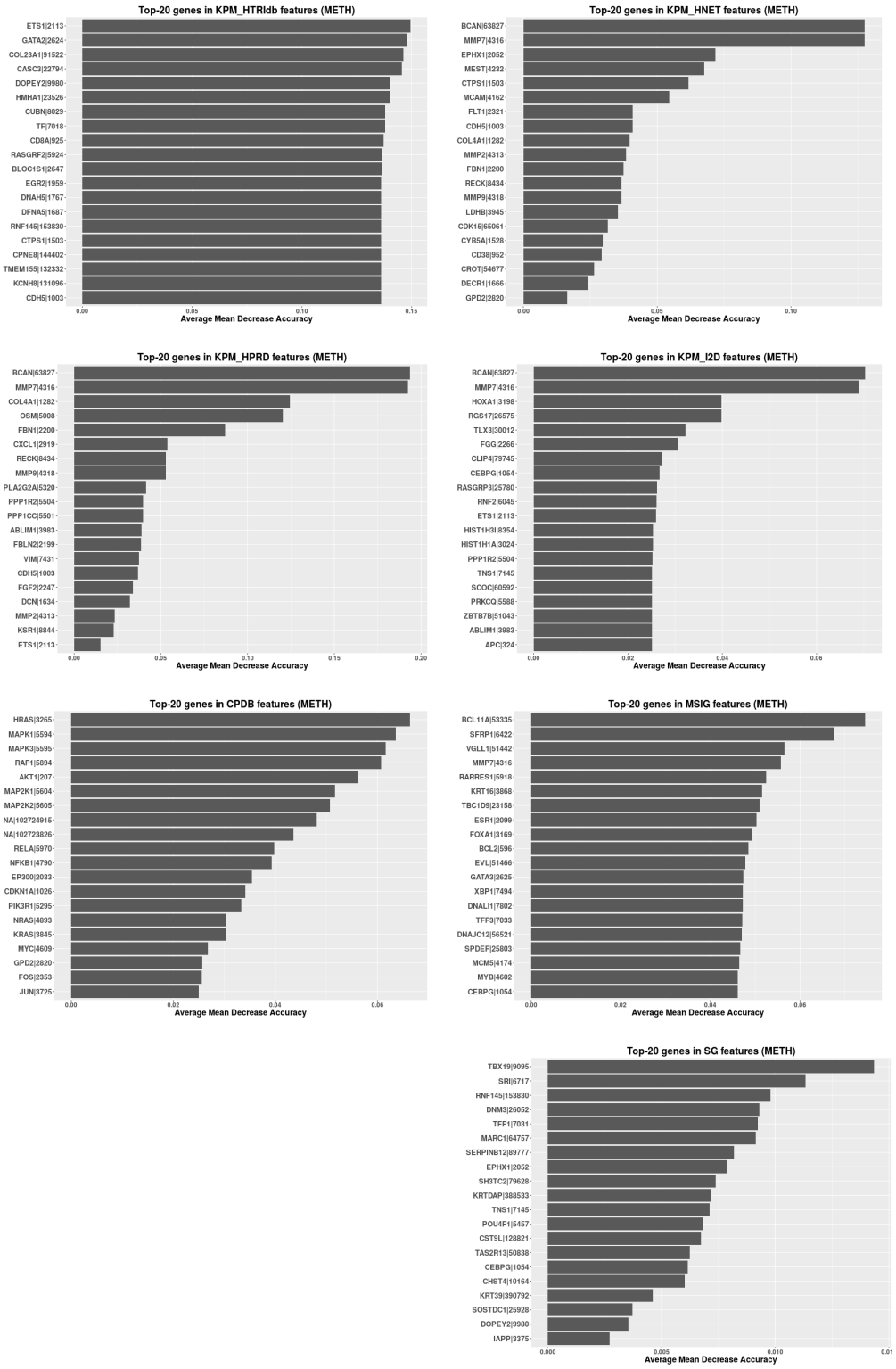


Figure S11: The top-20 most frequently selected genes (TCGA DNA methylation) for each feature type , sorted by their average mean decrease in accuracy from the random forest models.

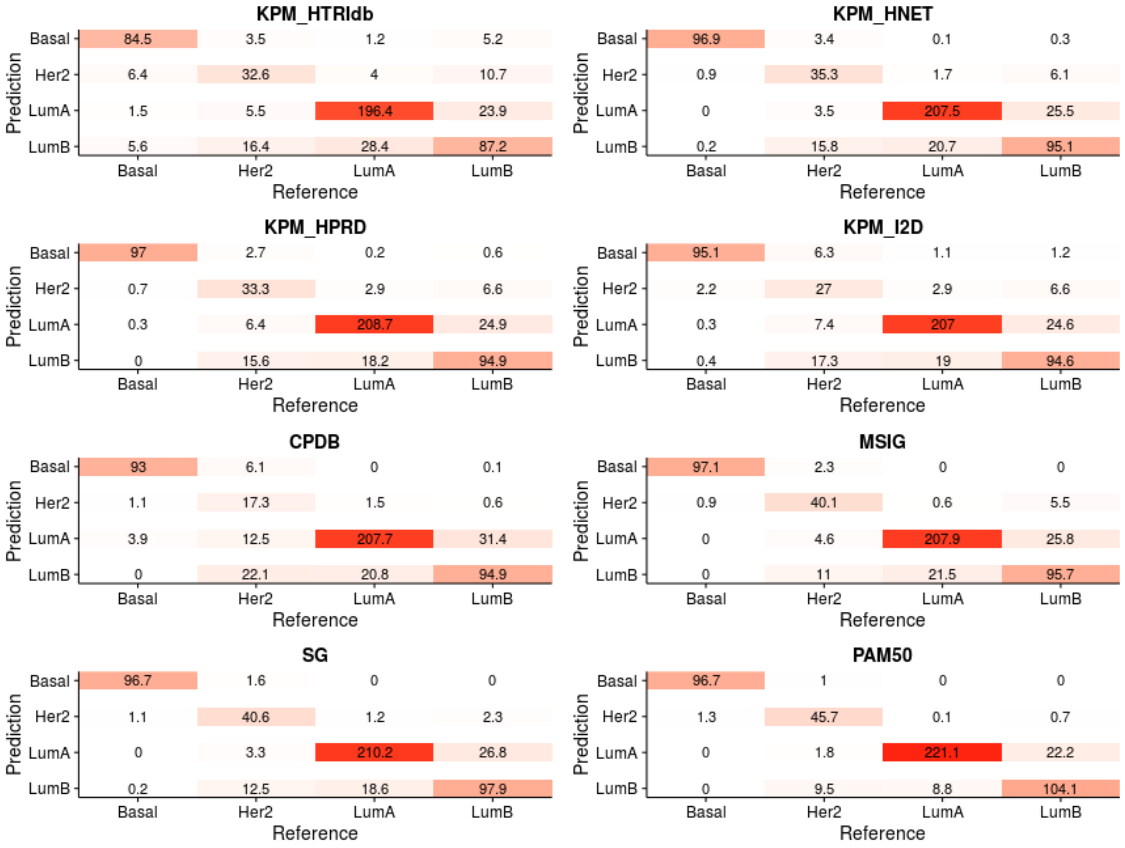


Figure S12: Confusion matrices for each of the models averaged over all cross validation repeats and folds on the TCGA expression data

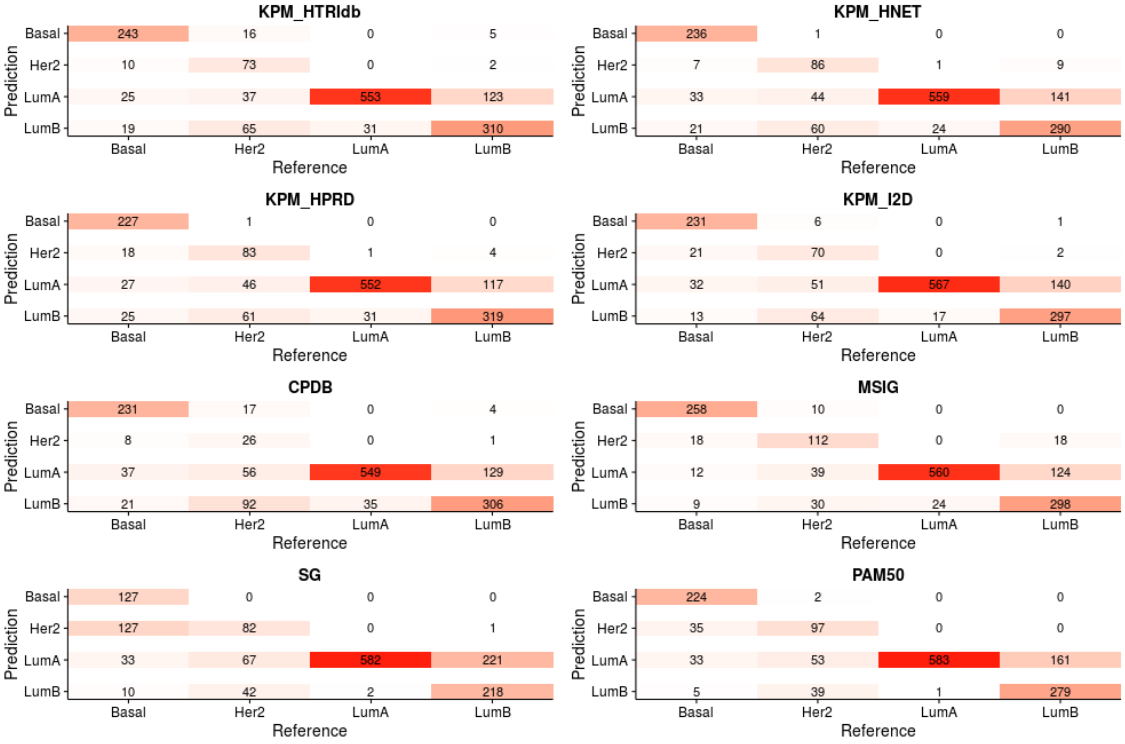


Figure S13: Confusion matrices for each of the models for the ACES validation data sets.

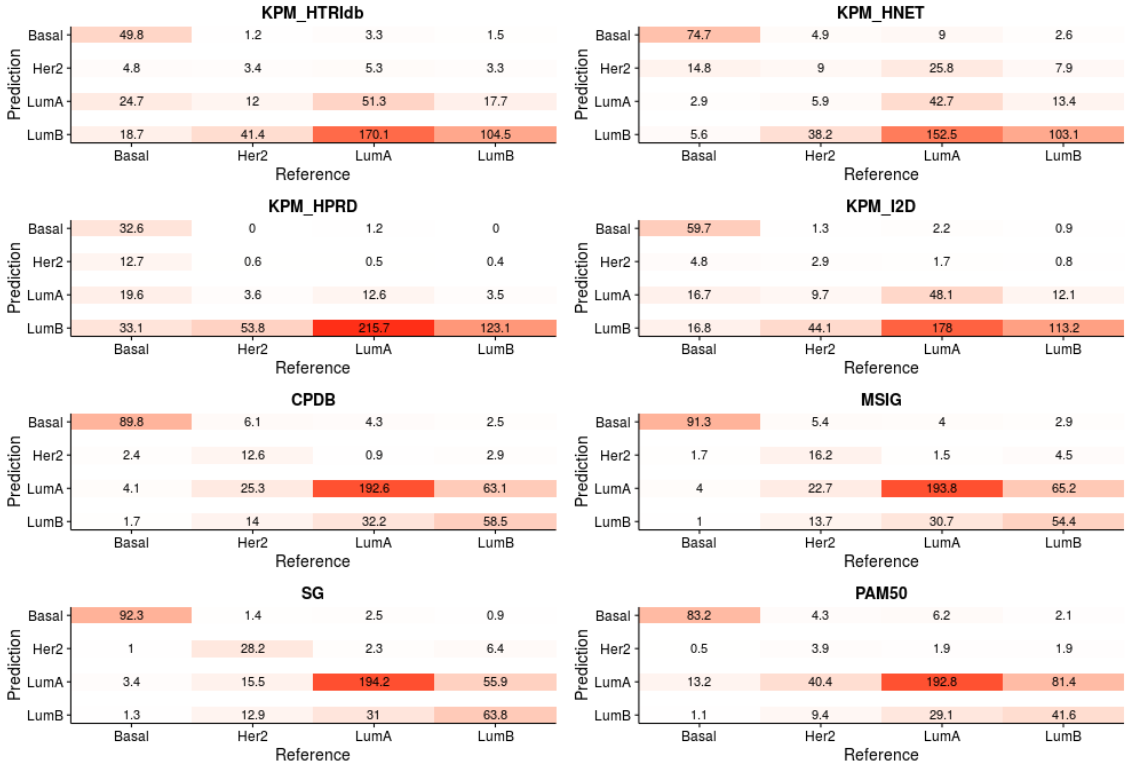


Figure S14: Confusion matrices for each of the models averaged over all cross validation repeats and folds on the TCGA DNA methylation data

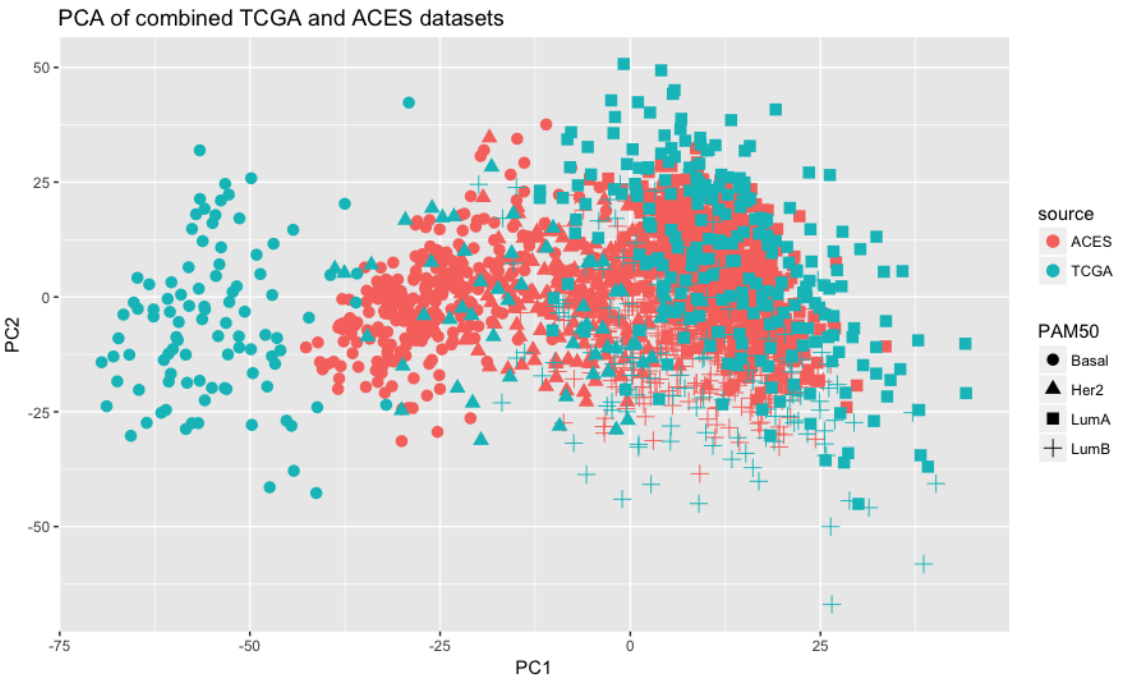


Figure S15: PCA of combined TCGA and ACES datasets

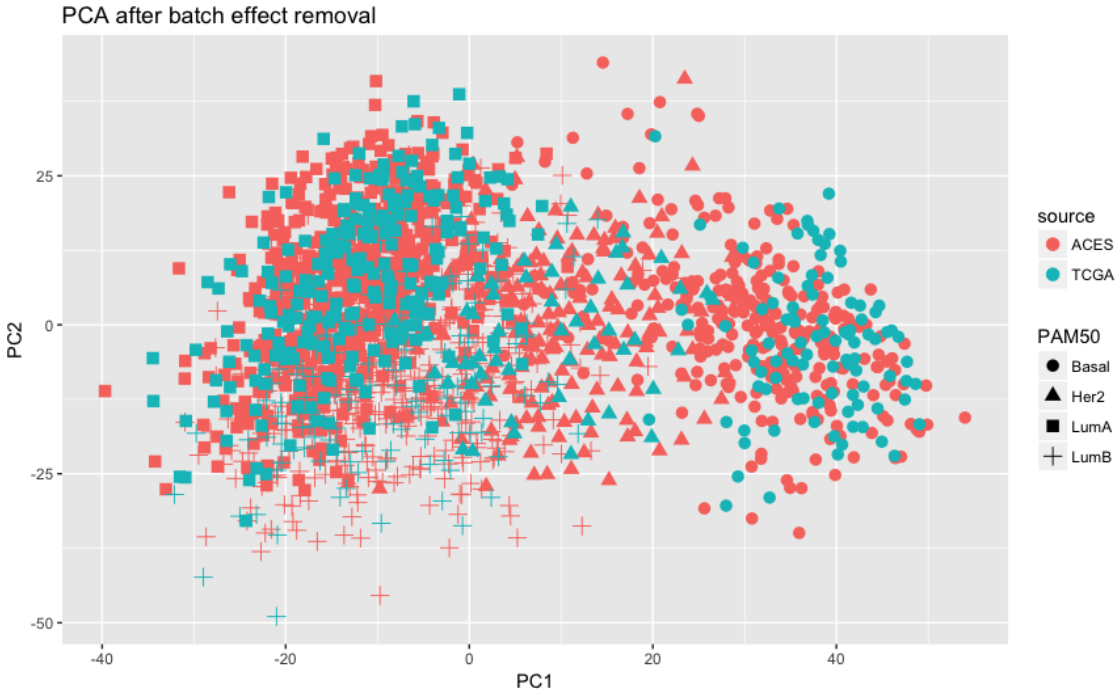


Figure S16: PCA of combined TCGA and ACES datasets after batch correction

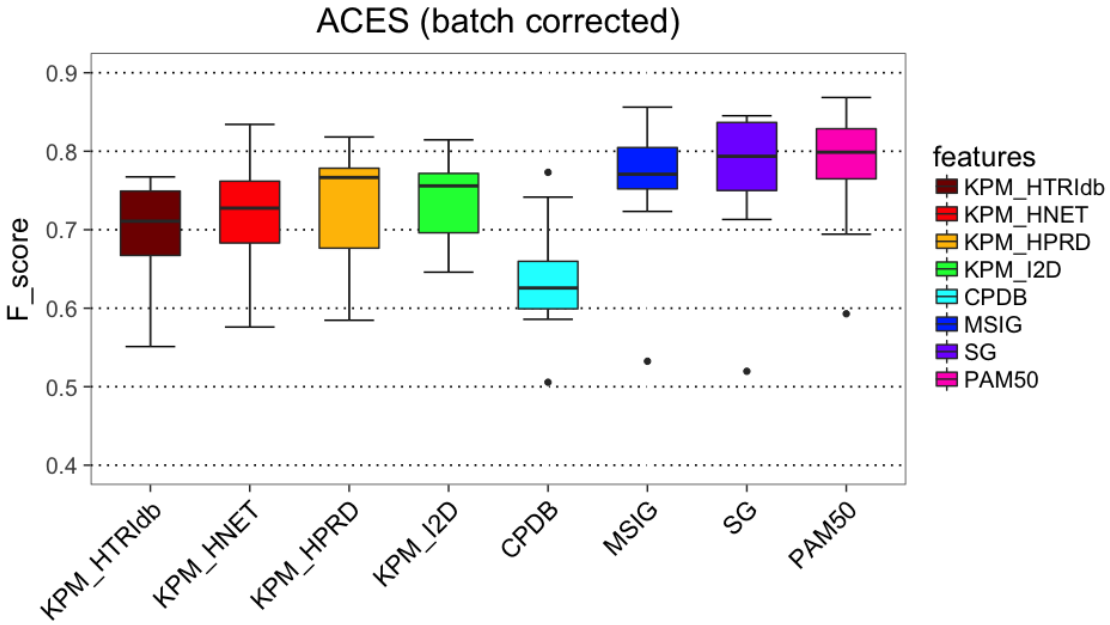


Figure S17: Performance of TCGA-trained models on ACES datasets after batch correction

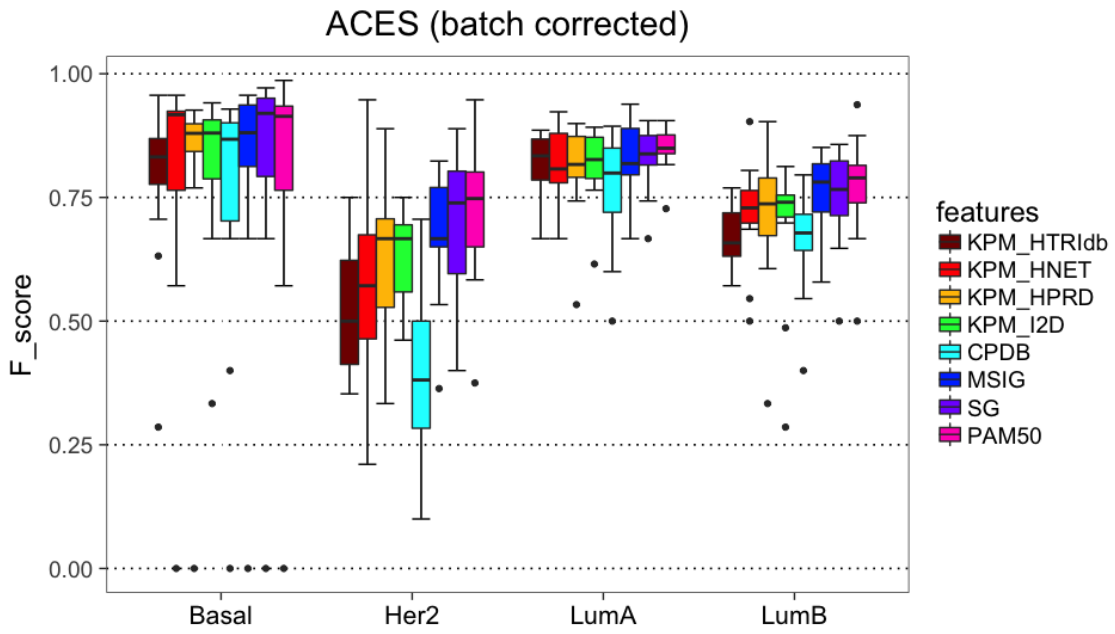


Figure S18: Performance per subtype of TCGA-trained models on ACES datasets after batch correction

References

[1] D. A. Barbie, P. Tamayo, J. S. Boehm, S. Y. Kim, S. E. Moody, I. F. Dunn, A. C. Schinzel, P. Sandy, E. Meylan, C. Scholl, S. Frohling, E. M. Chan, M. L. Sos, K. Michel, C. Mermel, S. J. Silver, B. A. Weir, J. H. Reiling, Q. Sheng, P. B. Gupta, R. C. Wadlow, H. Le, S. Hoersch, B. S. Wittner, S. Ramaswamy, D. M. Livingston, D. M. Sabatini, M. Meyerson, R. K. Thomas, E. S. Lander, J. P. Mesirov, D. E. Root, D. G. Gilliland, T. Jacks, and W. C. Hahn. Systematic rna interference reveals that oncogenic kras-driven cancers require tbk1. *Nature*, 462(7269):108–12, 2009.

[2] R. Diaz-Uriarte. Genesrf and varselrf: a web-based tool and r package for gene selection and classification using random forest. *BMC Bioinformatics*, 8:328, 2007.

[3] S. Hanzelmann, R. Castelo, and J. Guinney. Gsva: gene set variation analysis for microarray and rna-seq data. *BMC Bioinformatics*, 14:7, 2013.

[4] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 102(43):15545–50, 2005.

[5] T. Wittkop, D. Emig, S. Lange, S. Rahmann, M. Albrecht, J. H. Morris, S. Bocker, J. Stoye, and J. Baumbach. Partitioning biological data with transitivity clustering. *Nat Methods*, 7(6):419–20, 2010.