

Supplementary Information

Data Collection

The genes' coordinates used in the training phase were taken from the GRCh37/hg19 assembly (Church et al. 2011). The data was downloaded from the UCSC Genome Browser (Haeussler et al. 2019) from the following URL: <https://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/genes/hg19.refGene.gtf.gz>. Build version 37 of the assembly was used to match the genome coordinates in the methylation probes manifest file.

The methylation data we used was measured using Infinium HumanMethylation450 v1.2 BeadChip by Illumina. The genome mapping information of the methylation probes was downloaded from Illumina's website (ftp://webdata2.webdata2@usds-ftp.illumina.com/downloads/ProductFiles/HumanMethylation450/HumanMethylation450_15017482_v1-2.csv).

The TCGA data was downloaded using the TCGA-Assembler software (Wei et al. 2018; Zhu, Qiu, and Ji 2014). The DNA methylation data was downloaded using the 'DownloadMethylationData' function. The RNA-seq data was downloaded with the 'DownloadRNASeqData' function, setting the 'assayPlatform' parameter to 'gene.normalized_RNAseq'.

The additional LUAD dataset (Chen et al. 2020) was downloaded from OncoSG, the Singapore Oncology Data Portal (<https://src.gisapps.org/OncoSG/>), under 'Lung Adenocarcinoma (GIS, 2019)'.

Correspondence information between features

Some of the tested multi-omic integration algorithms require correspondence information between the features across omics. LIGER and Seurat assume that the input matrices to be integrated share the same set of features. When these methods were previously used to integrate scRNA-seq and methylation (LIGER) or scATAC-seq (Seurat) data, the input from the latter omic was converted to a matrix with gene-level features. The new features were expected to correspond to the GE features.

To summarize gene-level methylation, we used the annotations of methylation sites into six possible regions: TSS1500 (201-1500 bps upstream of the transcription start site(TSS)), TSS200 (0-200 bps upstream of the TSS), 5'UTR (untranslated region), 1stExon, Body, and 3'UTR. The HumanMethylation450 BeadChip annotations were taken from Illumina (https://support.illumina.com/downloads/humanmethylation450_15017482_v1-2_product_files.html). Of those, we used the sites in the TSS1500, TSS200, 5'UTR and 1stExon regions. We chose these regions as they showed anticorrelation with GE on the TCGA data (Supplementary Figure 4). This matched previous reports on anti-correlation between DM levels in the promoter region and the gene's expression level (Deaton and Bird 2011). The final gene-level summary was minus the average methylation signal in those regions.

Overview of the Methods Used

LIGER

LIGER (Welch et al. 2019) takes as input multiple single-cell datasets, either scRNA-seq experiments or multi-omic measurements. In the latter case, LIGER takes as input the preprocessed datasets after

conversion to a shared gene-level feature space. When LIGER is used to integrate gene expression with methylation data from mouse frontal cortical neurons, the methylation data is first converted to gene-level methylation features (non-CpG gene body methylation). The direction of the methylation signal is reversed to incorporate the assumption of general anti-correlation with gene expression in neurons (Mo et al. 2015). LIGER then employs integrative non-negative matrix factorization to create for each matrix a dataset-specific factor plus a shared factor across the datasets. The shared factor is used to jointly embed cells in a common low-dimensional space.

Seurat v3

The Seurat v3 algorithm (Stuart et al. 2019) was designed to integrate multiple scRNA-seq datasets in the SO/MD setting, but was also demonstrated to integrate scATAC-seq and scRNA-seq data in the MO/MD setting. The first step of such integration is similar to LIGER. The scATAC-seq data is converted to a gene-activity matrix, based on the accessibility of sites proximal to the gene’s transcription start site (Pliner et al. 2018). The gene activity matrix has the same feature set as the scRNA-seq matrices and it is assumed to be correlated with them. Seurat first uses canonical-correlation analysis to jointly reduce the dimension of the two datasets to a shared space. Then it identifies mutual nearest neighbors across the datasets. The pairings found are termed “anchors”. These anchor pairs are scored based on the consistency of anchors across the neighborhood structure of each dataset. The scored anchors are utilized to compute a projection mapping for each cell to embed it in the shared space.

JLMA

The joint Laplacian manifold alignment (JLMA) algorithm (C. Wang and Mahadevan 2008) learns a projection that maps datasets from two different feature spaces to a shared lower-dimensional space. This is done while simultaneously preserving the neighborhood relationships in each set and matching the local geometry of samples from the two sets. JLMA constructs a joint Laplacian matrix across the two domains, which captures the similarities within each dataset and the similarities across the datasets. The similarities across the datasets can be given as input to the algorithm (in a semi-supervised manner) or computed by the algorithm according to matching between the local geometry of the samples. In the latter case, JLMA does not require any correspondence information. The local geometry measure is computed based on the k -NN graph of each dataset. Finding the local geometry matching is computationally expensive even for small values of k , as it runs in $O(k!)$ time. After the joint Laplacian is computed, the optimal solution is found by solving a generalized eigenvalue decomposition problem.

MMD-MA

The maximum mean discrepancy-manifold alignment (MMD-MA) algorithm (Liu et al. 2019) is an unsupervised manifold alignment algorithm. It was created specifically for the task of single-cell multi-omics integration. The algorithm assumes the samples from the different omic datasets are drawn from the same initial population, but it does not require any correspondence information between the samples or the features. MMD-MA seeks an optimal alignment by minimizing an objective function with three terms. The first is the maximum mean discrepancy, which corresponds to the distance between the two mapped manifolds in the shared space. The second term, named distortion, measures relationships among data points between the original space and the shared latent space. The third is a penalty term that is intended to avoid a collapse to a trivial solution.

Benchmark Methods and Software

All experiments ran on R-4.0.1 (and python 3.8.0 for MMD-MA). For all methods, we followed the usage guidelines supplied by the creators. The preprocessing steps described in section 2.2.3 were applied to the input GE and DM data for all algorithms used in the benchmark.

LIGER

Implementation: We used the ‘`rLiger`’ R package, version 0.5.0. The methods referred to in the following subsection were supplied by this package. We followed LIGER guidelines for integrating GE and DM data (<https://welch-lab.github.io/liger/rna-methylation.html>).

Preprocessing: No further preprocessing (except the steps described in 2.2.3) was applied to the input GE and DM matrices. The aggregated gene-level methylation (described above in this supplement) was used as the methylation input for LIGER. The GE data was normalized and scaled using the `normalize` and `scaleNotCenter` methods. The DM input was not normalized and scaled as suggested by the guidelines.

Feature selection: The genes were selected using the `selectGenes` method when considering only the GE data, as suggested by the guidelines. This method selects the variable genes by comparing the variance of each gene’s expression to its mean expression.

Execution details: The default settings were used in the factorization and quantile normalizations phases of LIGER. As suggested by the guidelines, the `quantileAlignSNF` method was used with `center=T`, considering the density of the methylation data. The factorization was done with k (the number of factors) between 2 and 40, resulting in data projections in 2 to 40 dimensions.

Seurat v3

Implementation: We used the ‘`Seurat`’ R package, version 3.2.3. The methods referred to in the following subsection were supplied by this package. We followed Seurat guidelines for integration and label transfer (<https://satijalab.org/seurat/archive/v3.2/integration.html>).

Preprocessing: No further preprocessing (except the steps described in 2.2.3) was applied to the input GE and DM matrices. The aggregated gene-level methylation (described above in this supplement) was used as the methylation input for Seurat. The GE data was normalized using the `NormalizeData` method with the relative count normalization method. This step is not documented in the guidelines but empirically improved the results in all tested cases.

Feature selection: The genes were selected using the `FindVariableFeatures` method with the default parameters and selection method.

Execution details: The default settings were used. In the step of identifying anchors (using `FindIntegrationAnchors`), we used 30 neighbors when filtering the anchors (`k.filter=30`). We ran the algorithm with all possible dimensions between 2 and 40.

JLMA

Implementation: We used our implementation based on the JLMA paper, as we didn’t find an R package implementing JLMA. The implementation code is part of the INTEND project on GitHub.

Preprocessing: No further preprocessing (except the steps described in 2.2.3) was applied to the input GE and DM matrices. The aggregated gene-level methylation (described above in this supplement) was used as the methylation input for JLMA.

Feature selection: We selected the n genes with the highest variance in expression for $n = 500$ and 2000 (the algorithm ran in two variants). We scaled the inputs such that each feature (gene) had zero mean and unit variance.

Execution details: As mentioned in subsection 2.2.4, we computed the cross-omic similarity matrix for JLMA based on the aggregated gene-level methylation matrix. We used the hyper-parameter $\mu = 1$.

MMD-MA

Implementation: The algorithm's source code was downloaded from <https://noble.gs.washington.edu/proj/mmd-ma/>. We made minor changes in the source code to allow us to run the algorithm for the desired dimensions.

Preprocessing: No further preprocessing (except the steps described in 2.2.3) was applied to the input GE and DM matrices. As mentioned in subsection 2.2.4, we ran MMD-MA with both the original methylation data and the aggregated gene-level methylation (described above in this supplement) as inputs.

Feature selection: When using the original methylation data, no feature selection method was applied before computing the inter-similarity matrices for the GE and DM inputs. When running MMD-MA with the gene-level methylation data, we selected the n genes with the highest variance in expression for $n = 500$ and 2000 (the algorithm ran in two variants). In this case, we scaled the inputs such that each feature (gene) had zero mean and unit variance.

Execution details: We ran MMD-MA with dimensions 2,10,20,30, 40. We did not run it for all possible dimensions between 2 and 40 due to long running times.

INTEND algorithm

Here is a complete mathematical explanation of the two phases described in section 2.1 in the main paper.

The training phase

The preliminary training phase aims to learn connections between GE and DM using training data. Its inputs are expression and methylation profiles for the same set of n samples. E_{train} is an $|f_E| \times n$ expression matrix, where f_E is the set of genes for which the expression was measured. The methylation matrix M_{train} has dimensions $|f_M| \times n$, where f_M is the set of measured methylation sites. The goal is to determine a function $p^{(g)}$ for every gene g , that predicts the expression level of g based on the methylation levels of potentially relevant sites. Let $f_M^{(g)} \subseteq f_M$ be the set of relevant sites (its creation is described below). For a methylation profile $m^{(g)} \in \mathbb{R}^{|f_M^{(g)}|}$, we seek a function $p^{(g)}: \mathbb{R}^{|f_M^{(g)}|} \rightarrow \mathbb{R}$, s.t. $p^{(g)}(m^{(g)})$ is the predicted expression level of g .

Model

For each $g \in f_E$ we set $f_M^{(g)}$ to be all the probed methylation sites in the range $[C_{5'-end} - 10\text{kb}, C_{3'-end} + 10\text{kb}]$, where $C_{5'-end}$ and $C_{3'-end}$ are the coordinates of g 's 5'-end and 3'-end on the chromosome, respectively. The size of $f_M^{(g)}$ may vary due to the variability in gene length and the assay's coverage. Genes that had less than two measured methylation sites were removed from the model. Let $f_M = \cup f_M^{(g)}$ the union of the used methylation sites for all genes.

For each g , after obtaining $f_M^{(g)}$, INTEND uses Lasso regression model to learn the prediction function $p^{(g)}$ and select model features. Lasso was run using the glmnet R package and the optimal value of the penalty constant was chosen using 10-fold cross-validation on the training set. After calculating $p^{(g)}$ for every g in every training sample, the 2000 genes with the highest R^2 between predicted and observed gene expression are identified for use in the next stages of INTEND.

The embedding phase

The inputs for the main phase of the algorithm are:

1. A DM matrix M , for one target set of samples (T1), of dimensions $|f_M| \times n_M$
2. A GE matrix E for a second, disjoint target set of samples (T2), of dimension $|f_E| \times n_E$

3. A desired dimension d for the shared space

Additionally, the prediction functions $p^{(g)}$ for each g from the preliminary step are used. The requested output is a $d \times (n_M + n_E)$ matrix denoted S , which contains the projections of the input and predicted expression profiles into the shared d -dimensional space. The phase has three steps:

Step 1: Gene expression prediction using methylation data

Let $p^{(g)}$ be the learned prediction function for gene g and let m_1, m_2, \dots, m_{n_M} be the methylation profiles in M . Recall that $m_i^{(g)}$ describes the methylation levels of m_i in $f_M^{(g)}$ (possibly with some coefficients zeroed by the Lasso process). We apply $p^{(g)}$ on $m_i^{(g)}$ and get the predicted expression $e_i^{(g)}$. We denote the predicted expression profile for m_i as $e_i = \{e_i^{(g)} \mid g \in f_E\}$. This step results in the predicted expression matrix $P = (e_1, e_2, \dots, e_{n_M})$.

Step 2: Selecting genes

Denote the 2000 genes selected in the training phase by G_R . Let G_E denote the 2000 genes with the highest expression variability in E . Let G_P denote the 2000 genes with the highest variance in the predicted expression P . We select the following genes from E and P :

$$G_s = G_R \cap G_E \cap G_P \quad (1)$$

The resulting matrices are E_{G_s} and P_{G_s} , with dimensions $|G_s| \times n_E$ and $|G_s| \times n_M$ respectively. The size of G_s varies depending on the training and target datasets. Finally, each row of E_{G_s} and P_{G_s} is centered and scaled separately so that each feature has zero mean expression level and unit variance.

Step 3: Embedding

The last step applies CCA to E_{G_s} and P_{G_s} , and produces the integrated matrix S . CCA is a dimension reduction method that finds linear combinations of features across datasets such that these combinations have maximum correlation (37).

Let us denote $X = E_{G_s} \in \mathbb{R}^{|G_s| \times n_E}$ and $Y = P_{G_s} \in \mathbb{R}^{|G_s| \times n_M}$. Let $d \leq \min(n_E, n_M)$. CCA aims to find canonical correlation vectors $u_1, \dots, u_d, v_1, \dots, v_d$ such that the correlations between the projections Xu_i and Yv_i are maximized, under the constraint that Xu_i is uncorrelated with Xu_j for $j < i$ and the same for Yv_i and Yv_j . To get the first pair of canonical correlation vectors, the following optimization problem should be solved:

$$(u_1, v_1) = \underset{u \in \mathbb{R}^{n_E}, v \in \mathbb{R}^{n_M}}{\operatorname{argmax}} u^T X^T Y v \quad s. t. \begin{cases} u^T X^T X u = 1 \\ v^T Y^T Y v = 1 \end{cases} \quad (2)$$

When $|G_s|$ is smaller than the number of samples n_E and/or n_M , the solution for u_1, v_1 is not unique. To overcome this, as proposed in Butler et al., the covariance matrix within each dataset is treated as if it were diagonal, resulting in the following problem:

$$(u_1, v_1) = \underset{u, v}{\operatorname{argmax}} u^T X^T Y v \quad s. t. \begin{cases} \|u\|_2^2 = 1 \\ \|v\|_2^2 = 1 \end{cases} \quad (3)$$

We scale and center the columns of X and Y to have a mean of 0 and variance of 1 (in the previous step the same process was applied to the rows). The problem can be solved using Lagrange multipliers, as described in the following subsection.

CCA optimization problem solution

To solve the optimization problem in equation (3), we use the Lagrange multipliers method. We denote $K = X^T Y$. Let:

$$L = u^T K v - \frac{\lambda_1}{2} (u^T u - 1) - \frac{\lambda_2}{2} (v^T v - 1) \quad (5)$$

Differentiating L with respect to u and v gives:

$$\frac{\delta L}{\delta u} = K v - \lambda_1 u = 0 \rightarrow K v = \lambda_1 u \quad (6)$$

$$\frac{\delta L}{\delta v} = K^T u - \lambda_2 v = 0 \rightarrow K^T u = \lambda_2 v \quad (7)$$

Left-multiplying (6) and (7) by u^T and v^T respectively, and using the constraints $\|u\|_2^2 = 1$ and $\|v\|_2^2 = 1$:

$$\lambda_1 = u^T K v = v^T K^T u = \lambda_2 \quad (8)$$

Thus u and v are the left and right unit singular vectors of K with singular value $\lambda = \lambda_1 = \lambda_2$. Since the objective is to maximize $u^T K v$, then u_1, v_1 are the left and right unit singular vectors of K with the greatest singular value. We claim that $\forall i \in \{1, \dots, d\}$, u_i and v_i are the left and right unit singular vectors of K with the i^{th} greatest singular value. Let u_i and v_i be the i^{th} unit singular vectors of K . Then we showed that (u_i, v_i) maximizes over all $u \in \mathbb{R}^{n_E}, v \in \mathbb{R}^{n_M}$, the correlation between Xu and Yv . As $u_i^T u_j = v_i^T v_j = 0$ for $j < i$, and we assumed that $X^T X$ and $Y^T Y$ are diagonal, then $\text{Cor}(Xu_i, Xu_j) = \text{Cor}(Yv_i, Yv_j) = 0$ for $j < i$. Hence the optimal canonical-correlation vectors can be obtained by SVD of $K = X^T Y$. We denote $U = (u_1, u_2, \dots, u_d) \in \mathbb{R}^{n_E \times d}$ and $V = (v_1, v_2, \dots, v_d) \in \mathbb{R}^{n_M \times d}$ where u_i and v_i are the i -th left and right singular vectors, respectively. The output of this step is the matrix $S = [U^T \ V^T]$, of dimensions $d \times (n_E + n_M)$, containing the embeddings of samples from both target sets in the shared d -dimensional space.

Supplementary Tables

Supplementary Table 1. Number of features and samples in each TCGA dataset before and after the handling of missing values

Dataset	Number of samples				Number of features			
	Gene expression		DNA methylation		Gene expression		DNA methylation	
	Before	After	Before	After	Before	After	Before	After
AML	173	173	194	194	20530	20530	526729	432429
BLCA	427	427	440	440	20530	20530	526729	431716
COAD	328	328	353	353	20530	20530	526729	431308
LGG	534	534	534	534	20530	20530	526729	431991
LIHC	424	424	430	430	20530	20530	526729	430791
LUAD	576	576	507	507	20530	20530	526729	431486
PAAD	183	183	195	195	20530	20530	526729	428806
PRAD	550	550	553	553	20530	20530	526729	432201
SARC	265	265	269	269	20530	20530	526729	428486
SKCM	473	473	475	475	20530	20530	526729	430579
THCA	572	572	571	571	20530	20530	526729	432307

Supplementary Table 2. Average FOSCTTM score for INTEND with and without applying CCA at the end of the embedding phase. When running with CCA the requested shared space dimension d ranges from 2 to 40, and the presented score is the best across all values of d . The optimal d is written in parentheses. When running without CCA the dimension is the size of the selected gene set. The set size is written in parentheses.

Dataset	INTEND – with CCA	INTEND – without CCA
AML	2.416 (25)	5.184 (191)
BLCA	0.040 (39)	0.857 (362)
COAD	0.025 (37)	1.361 (297)
LGG	6.815 (22)	10.072 (222)
LIHC	0.139 (36)	1.088 (339)
LUAD	0.062 (32)	0.892 (359)
PAAD	0.546 (30)	4.781 (362)
PRAD	0.374 (38)	1.843 (295)
SARC	0.052 (35)	0.616 (382)
SKCM	0.027 (39)	1.043 (379)
THCA	3.073 (11)	5.849 (264)

Supplementary Table 3. Comparison of correlation extraction from LUAD dataset integration results. The procedure described in section 3.4.2 was repeated with the integration results of INTEND, LIGER, Seurat and MMD-MA. The number of mutual nearest neighbors used was 270, 142, 61, and 231, respectively. The analysis presented in the table included approx. 2.5 million gene-site pairs that had significant TCGA-observed correlation (p -value <0.01). We tested the percentage of these pairs that were detected with significant estimated correlation (p -value <0.01), the percentage of these pairs with the same correlation sign of estimated and TCGA-observed correlations, and the R^2 for the correlation between estimated and TCGA-observed correlations.

Algorithm	INTEND	LIGER	Seurat	MMD-MA
% of gene-site pairs with estimated significant correlation (p -value <0.01)	20.17	3.78	5.12	17.45
% of gene-site pairs with estimated and TCGA-observed correlation with same correlation sign	74.51	47.29	52.08	56.73
R^2 for the correlation between estimated and TCGA-observed correlation	0.374	0.015	0.003	0.061

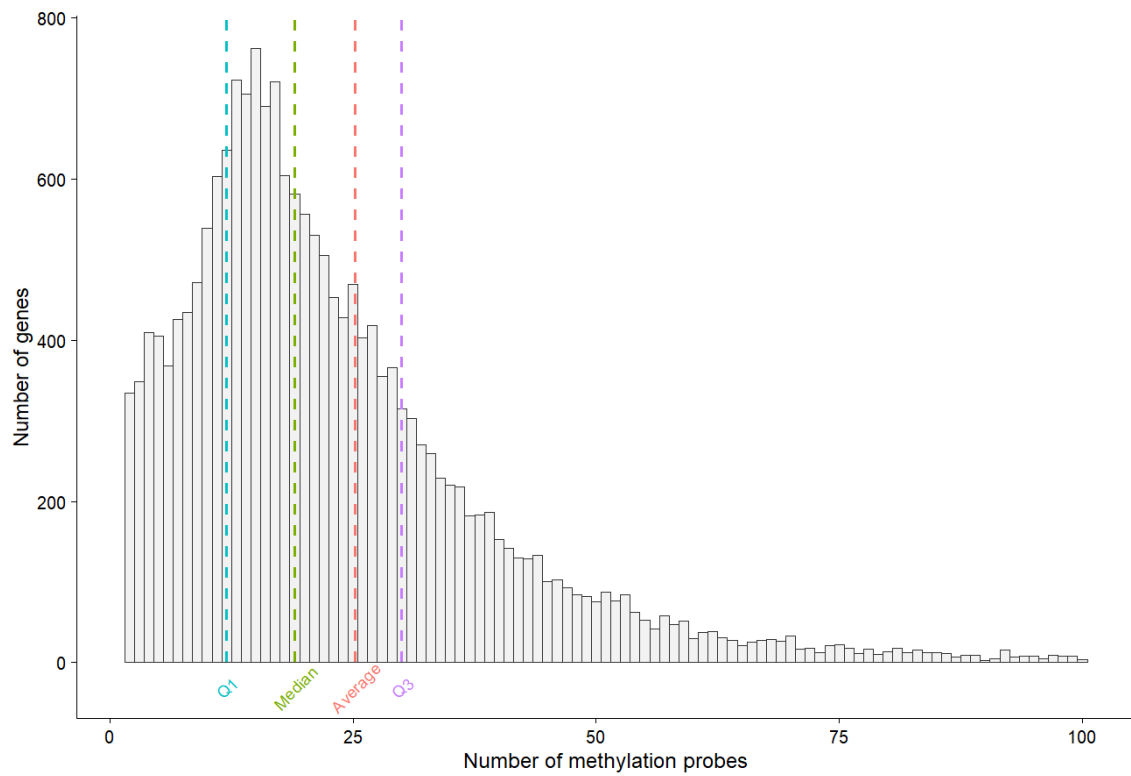
Supplementary Table 4. Top ten methylation sites with the strongest negative estimated correlations out of the 964 sites in 1Mb range from TK1

Methylation site	Location on chromosome 17 (build GRCh37/hg19)	Correlation coefficient estimation
cg11868461	75830800	-0.5181234
cg06643271	76128170	-0.5016554
cg24988684	76128556	-0.4887382
cg10460946	76247467	-0.4631925
cg11493223	76128522	-0.4516062
cg02911077	76128621	-0.4396759
cg18901278	76128531	-0.4280529
cg04947157	76128481	-0.4135805
cg03742808	76128634	-0.4130168
cg05110391	76588634	-0.4063449

Supplementary Table 5. Effect of the size and diversity of the training set data on INTEND performance. The table shows average FOSCTTM score (percent) on four target cancer datasets. Training dataset #1 includes **seven** cancer datasets: AML, BLCA, LGG, LUAD, PAAD, PRAD, and THCA. Training dataset #2 is different for each target dataset and includes **ten** cancer datasets, all eleven used in the study (AML, BLCA, COAD, LIHC, LGG, LUAD, PAAD, PRAD, SARC, SKCM and THCA), excluding the target dataset.

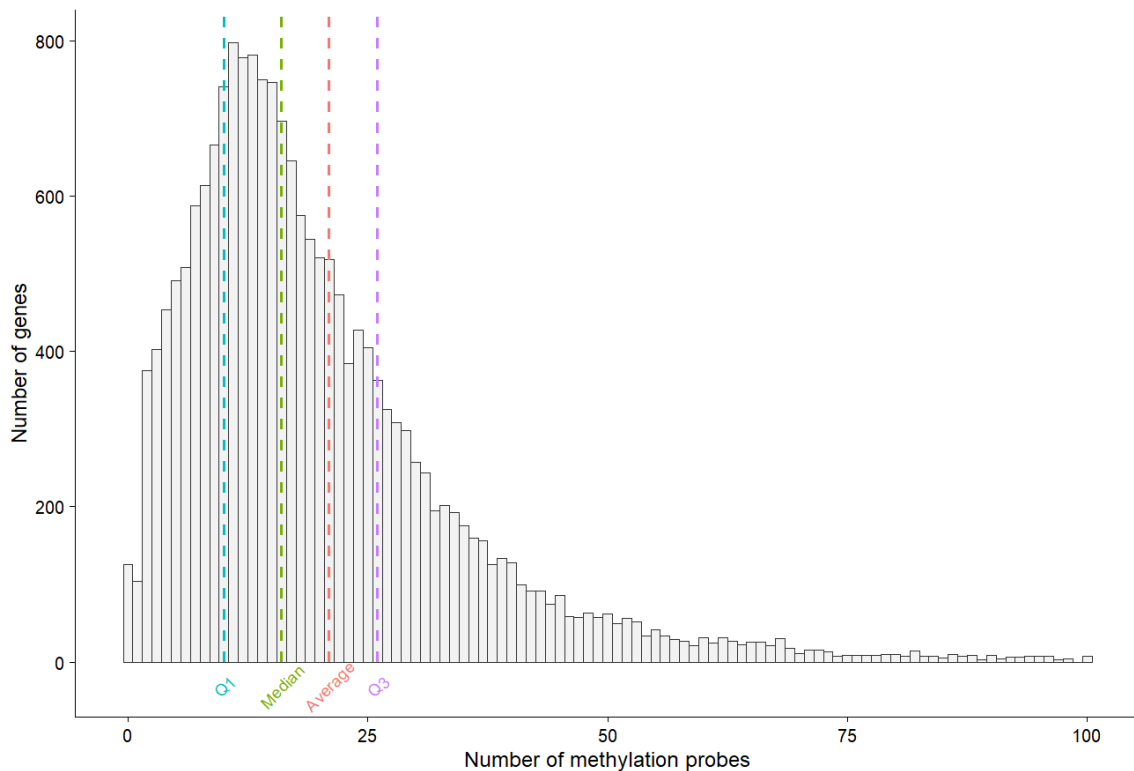
Target Dataset	FOSCTTM (%) Training dataset #1	Training dataset #1 size (7 datasets)	FOSCTTM (%) Training dataset #2	Training dataset #2 size (10 datasets)	Ratio between FOSCTTM when trained on dataset #1 and when trained on dataset #2
COAD	0.116	2881	0.043	4031	2.7
LIHC	0.325	2881	0.152	3915	2.14
SARC	0.144	2881	0.068	4066	2.11
SKCM	0.150	2881	0.027	3856	5.43

Supplementary Figures



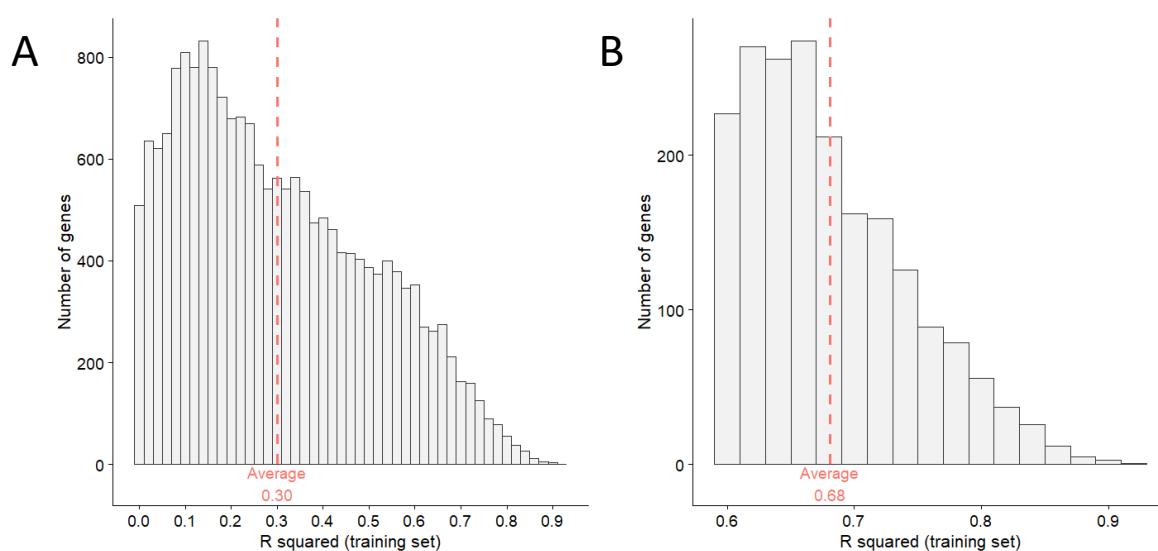
Supplementary Figure 1

Histogram of the number of methylation sites per gene. Average: 25.22, median: 19, interquartile range (*IQR*): 12-30. The maximum number of methylation sites per gene was 1055 (outside the plot axis limits).



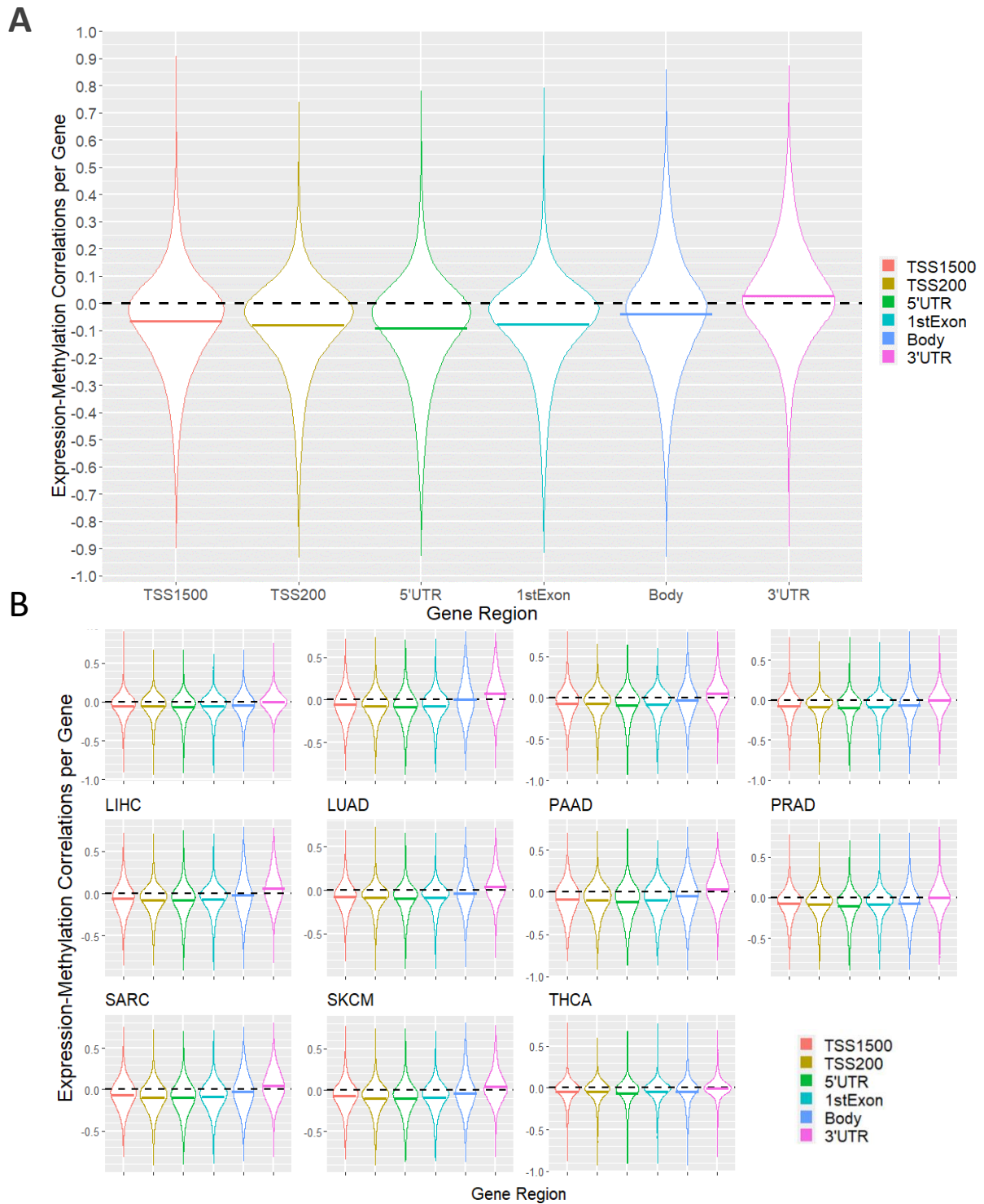
Supplementary Figure 2

Histogram of the number of methylation sites per gene in the model after Lasso shrinkage on the TCGA data. The model was trained on ten cancer subtypes data from TCGA: AML, BLCA, COAD, LGG, LIHC, PAAD, PRAD, SARC, SKCM, and THCA. Average: 20.93, median: 16, interquartile range (*IQR*): 10-26. The maximum number of methylation sites per gene was 424 (outside the plot axis limits).



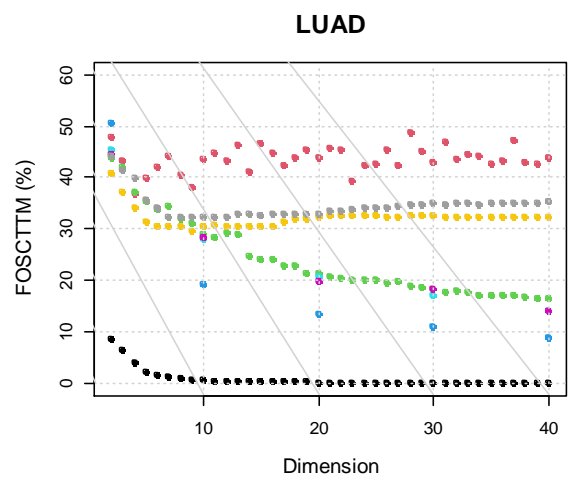
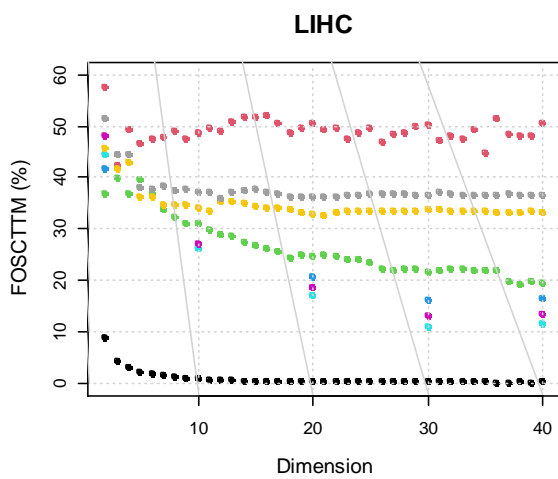
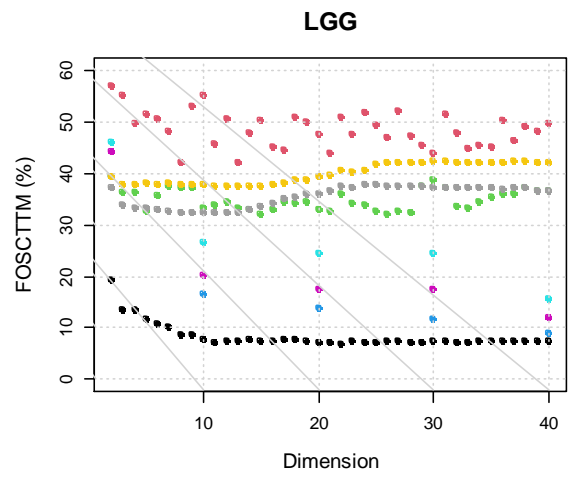
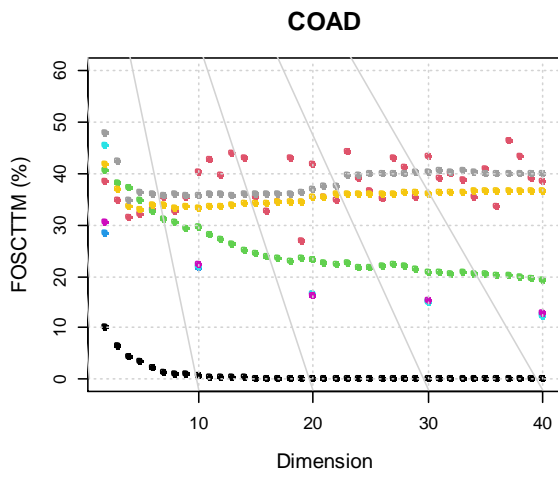
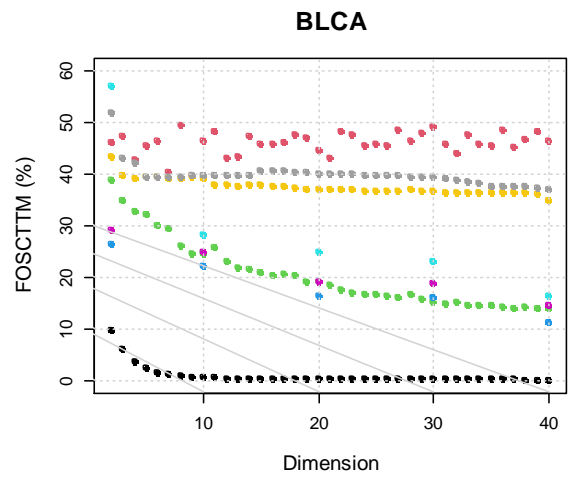
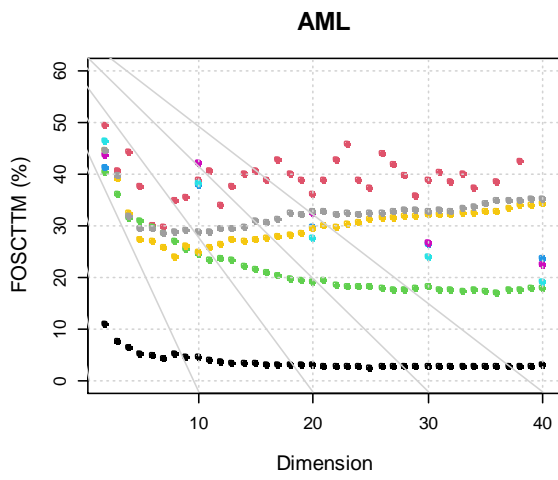
Supplementary Figure 3

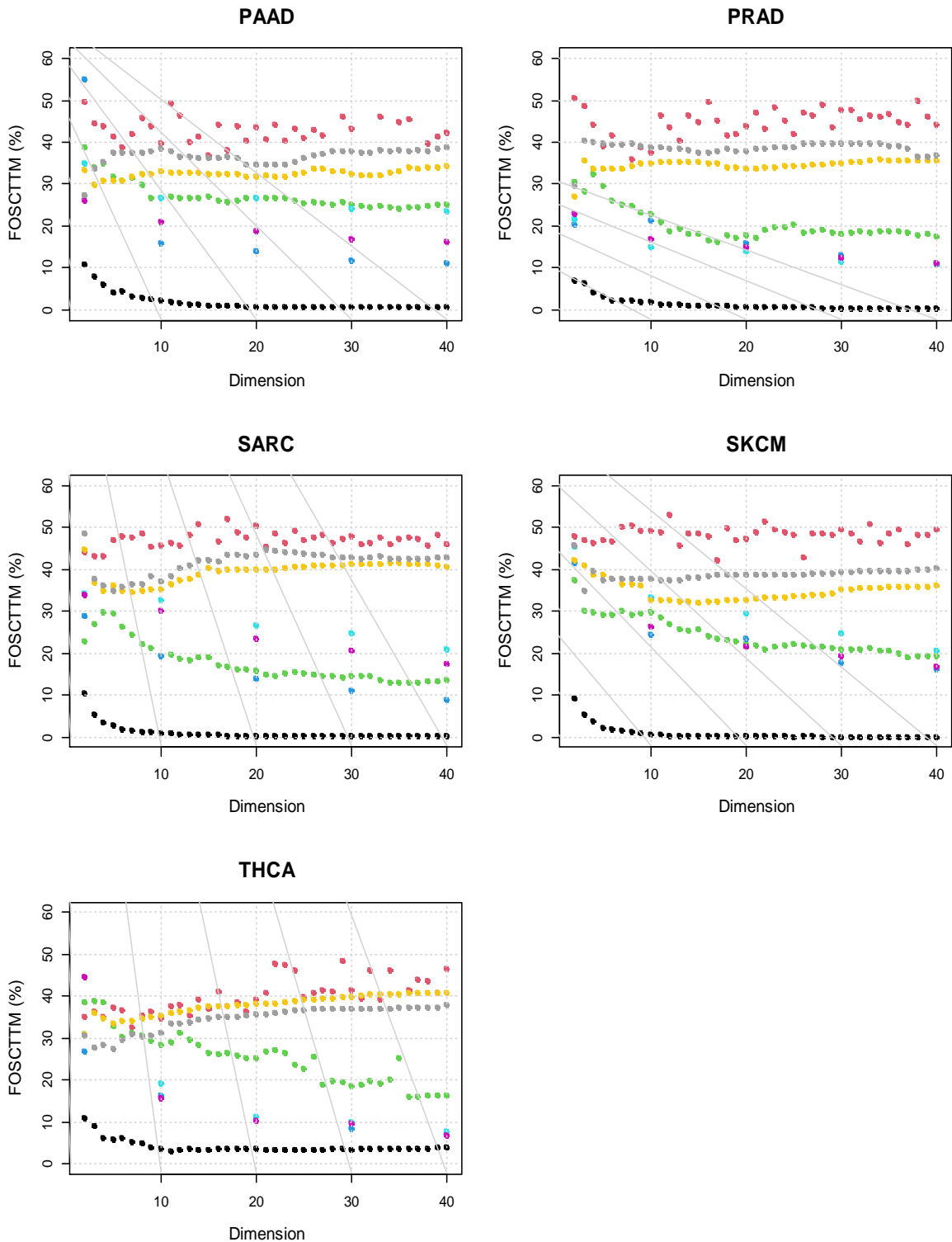
Histograms of R^2 values between predicted and observed gene expression, when training on GE and DM data of 10 cancer subtypes from TCGA (the datasets listed in **Table 1**, excluding LUAD), covering 3852 tumor samples. (A) All 19143 genes, (B) The 2000 genes with the highest R^2 .



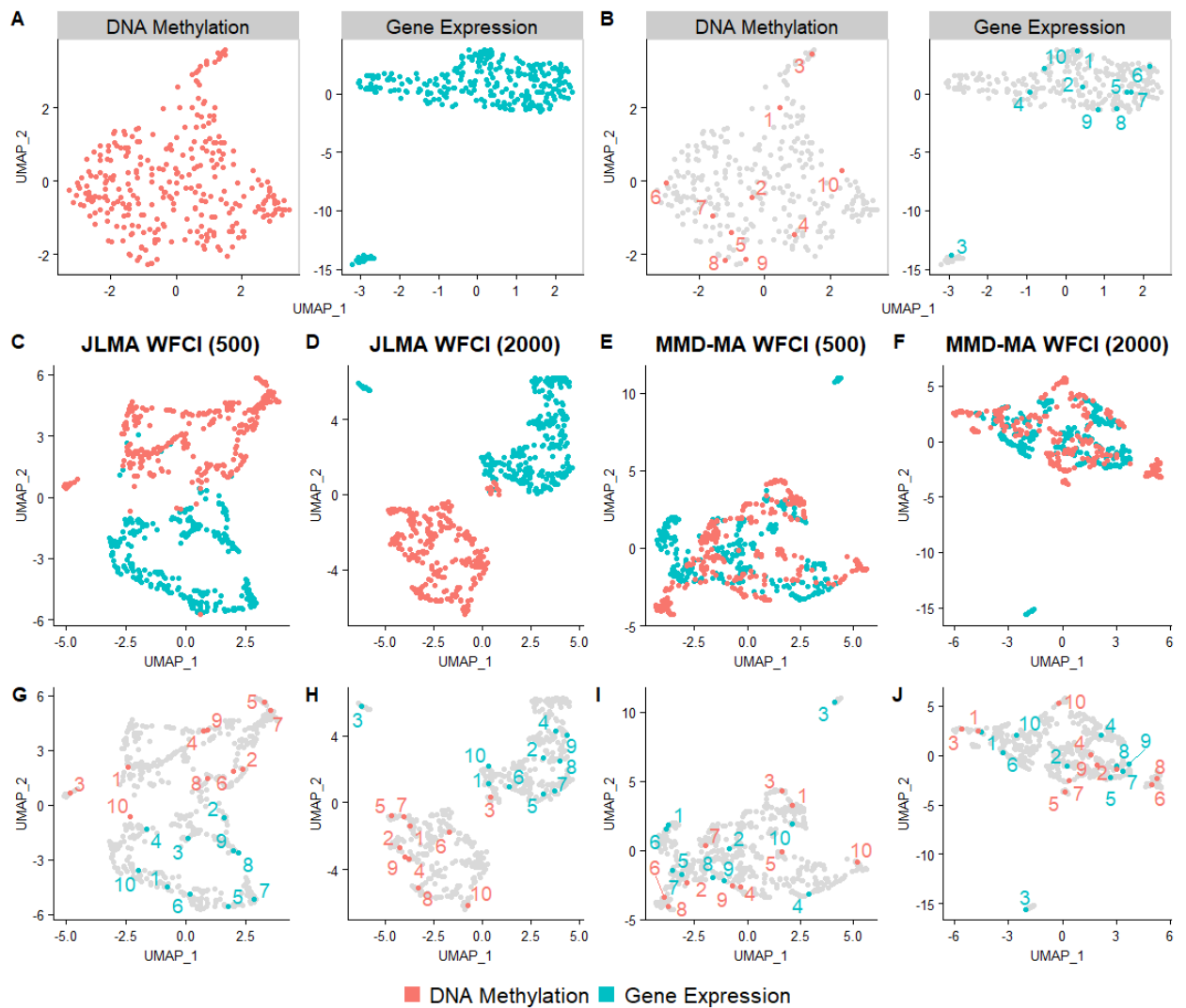
Supplementary Figure 4: Correlation between gene expression and DNA methylation levels in different genomic regions

Correlations were measured for genes that had both expression and methylation data in the specified region of the gene. Data included samples from eleven cancer types from the TCGA database (**Table 1**). The horizontal line in each violin plot is the mean correlation for the region, and the black dashed line shows a correlation of zero. (A) Summary over all subtypes, (B) Results for each subtype separately. The mean correlation for TSS1500, TSS200, 5'UTR, and 1stExon is < -0.04 for every subtype, and < -0.065 on average across subtypes. The Body regions exhibit a positive mean correlation for one subtype (BLCA), and 3'UTR for seven.





Supplementary Figure 5: Performance of the algorithms as a function of the projected dimension on eleven TCGA cancer datasets. Average FOSCTTM score versus the shared space dimension. The numbers 500 and 2000 in parenthesis denote the number of selected genes in the WFCI runs of MMD-MA and JLMA. The results of MMD-MA include only $d = 2, 10, 20, 30, 40$, due to the long runtime of the algorithm.



Supplementary Figure 6. The integration of gene expression and DNA methylation samples from the COAD dataset – results for JLMA and MMD-MA algorithms

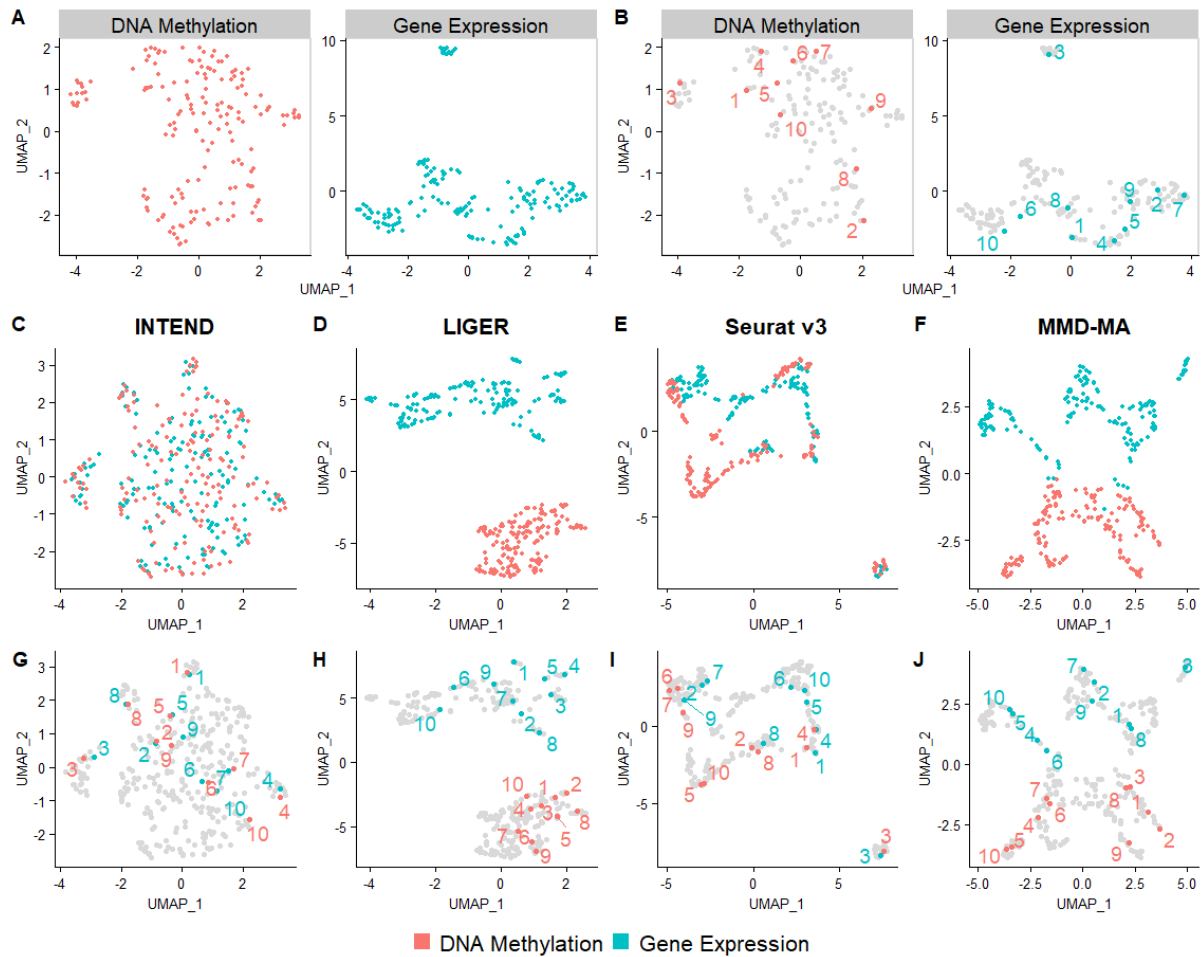
(A) UMAP plots of the original data colored by omic.

(B) UMAP plots of the original data. To appreciate concordance between omics, ten samples were randomly chosen, and their matching points in both omics are labeled and colored by omic.

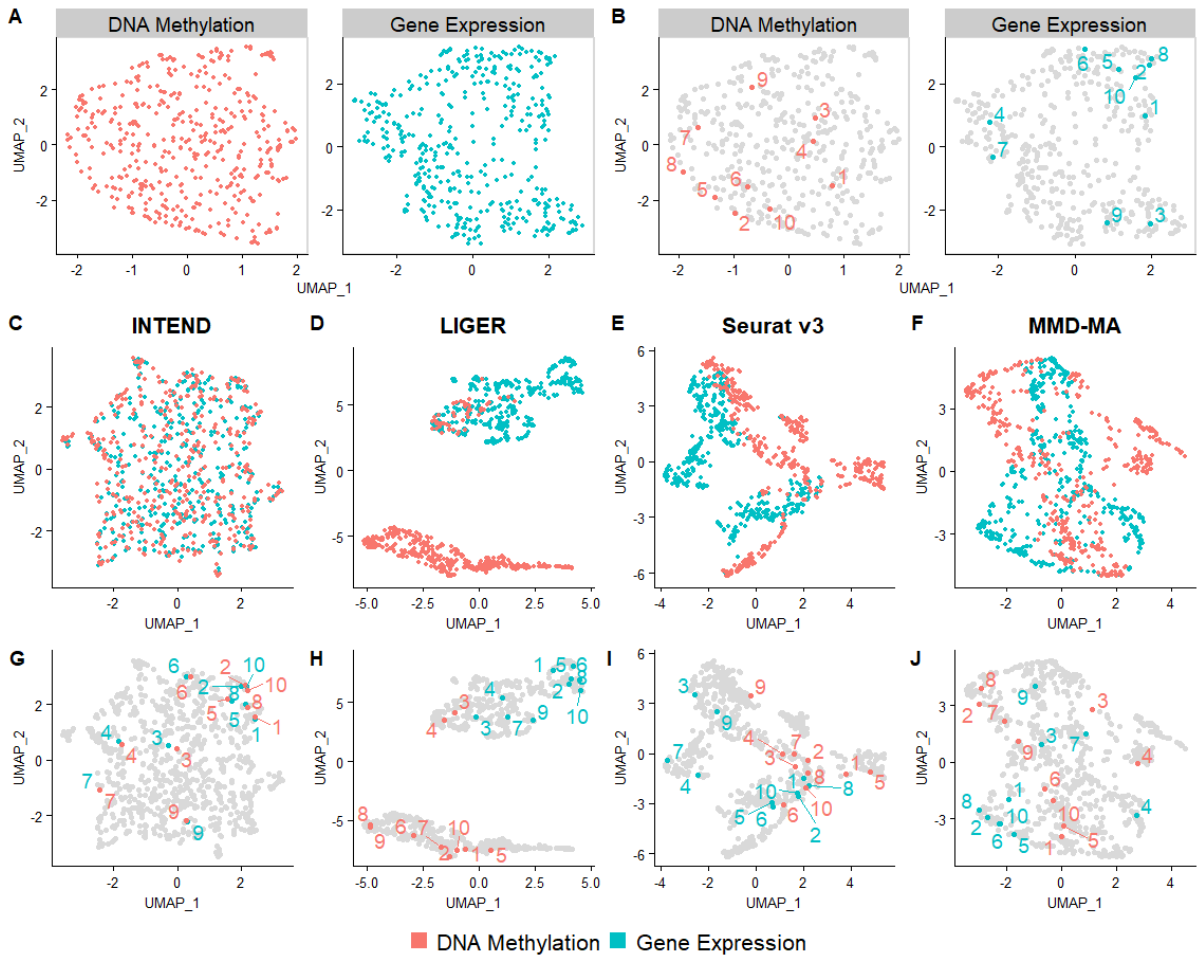
(C-J) UMAP plots of the samples after they were projected to a shared space by each algorithm, with a set of selected genes of size 500 and 2000. The samples are colored by omic (C-F) and the projection of the points from (B) are labeled in (G-J).

Supplementary Figures 7-16. Results of integration of GE and DM samples from all TCGA datasets listed in **Table 1**, excluding COAD. (A) UMAP plots of the original data. (B) The same plots as in A. To appreciate concordance between omics, ten samples were randomly selected, and their matching points in both omics were labeled. (C-F) UMAP plots of the samples after they were projected to a shared space by each algorithm. (G-J) The same plots as in C-F with the selected points labeled. In all plots colors correspond to omics.

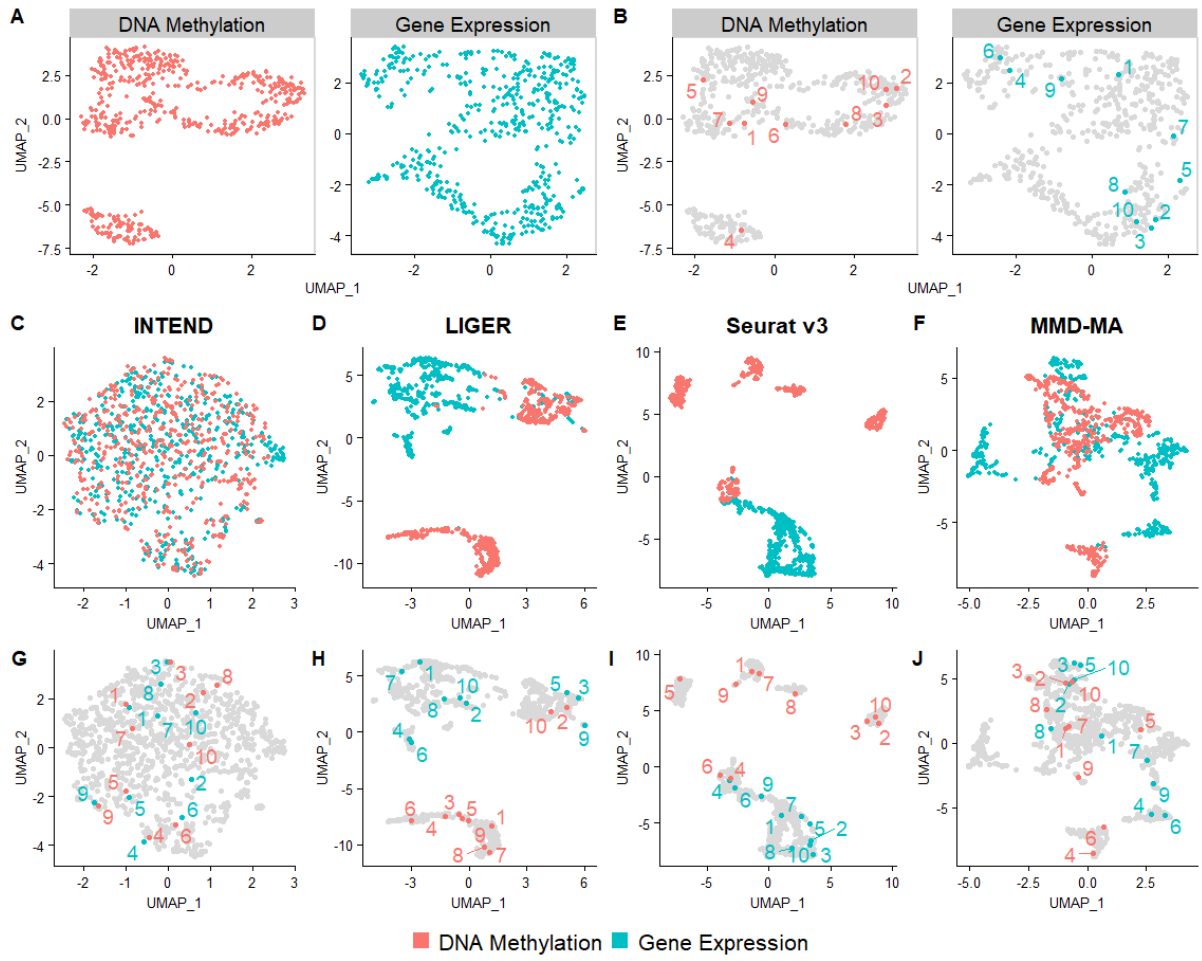
AML



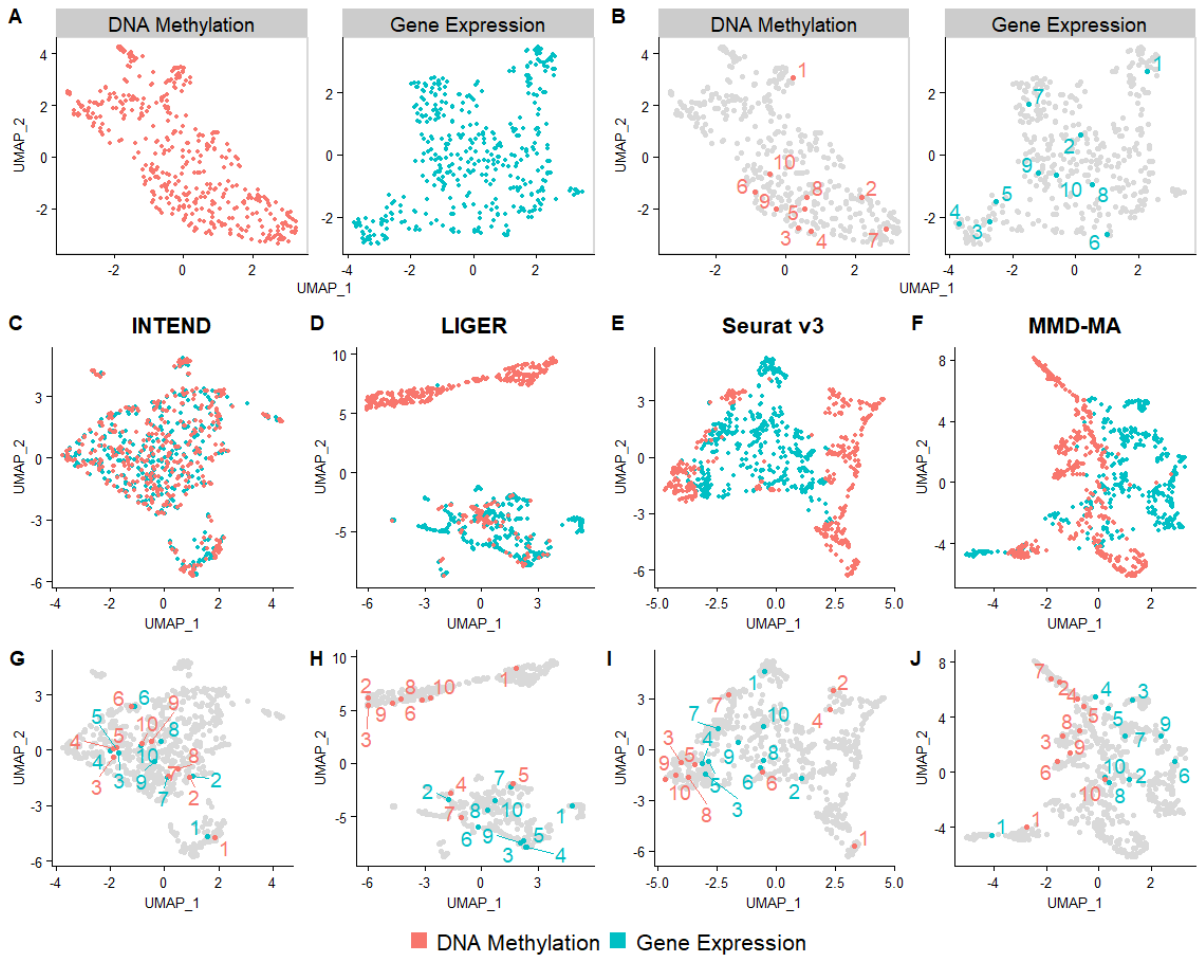
BLCA



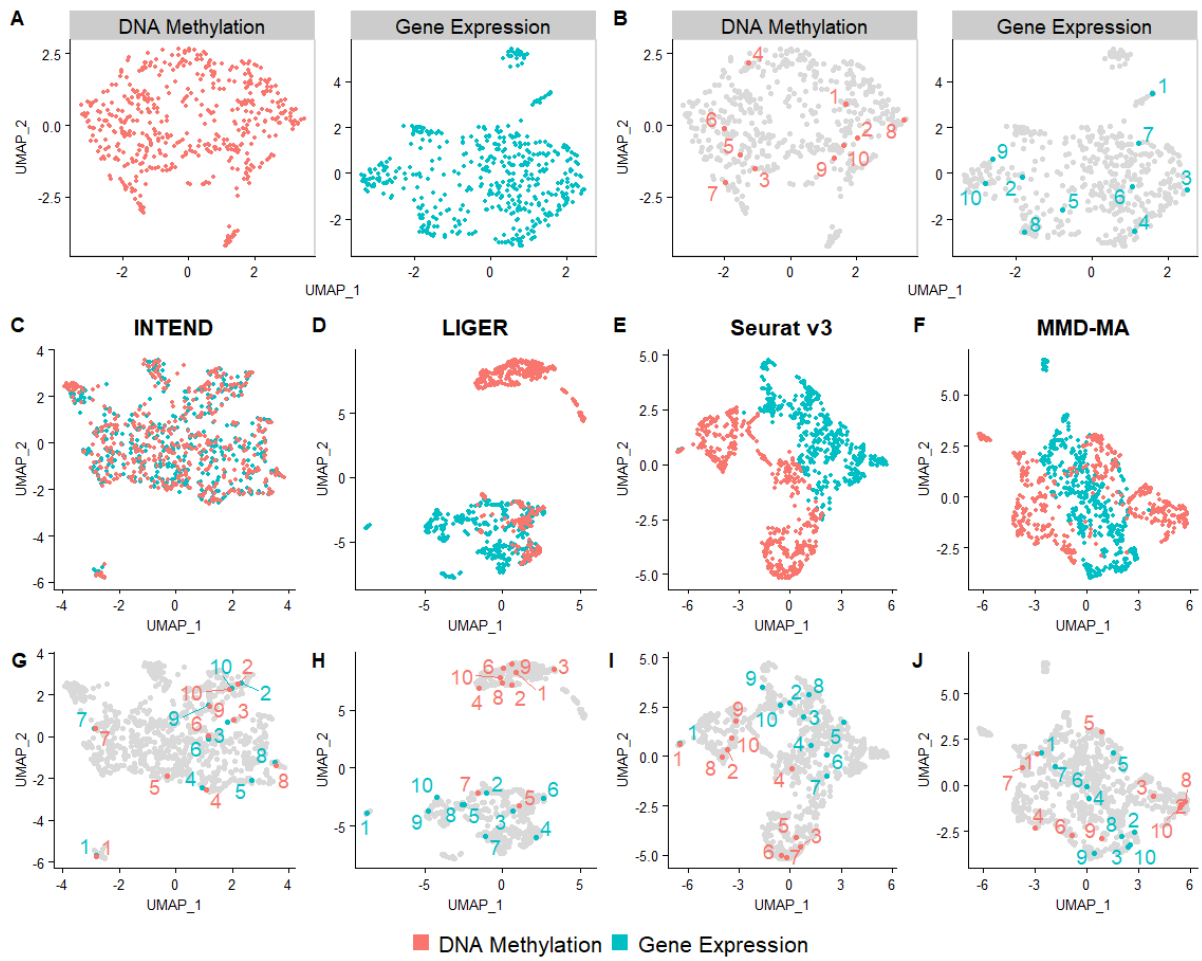
LGG



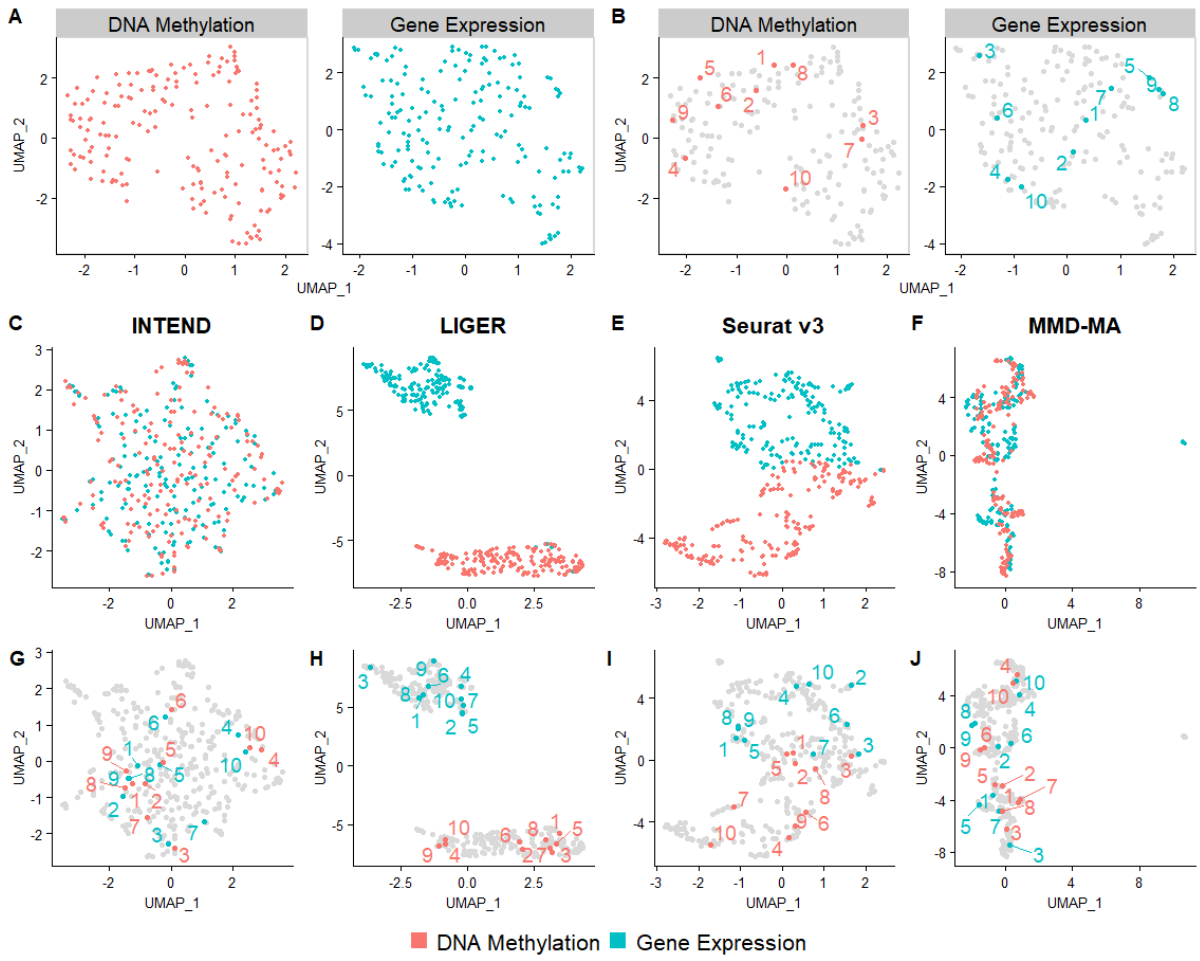
LIHC



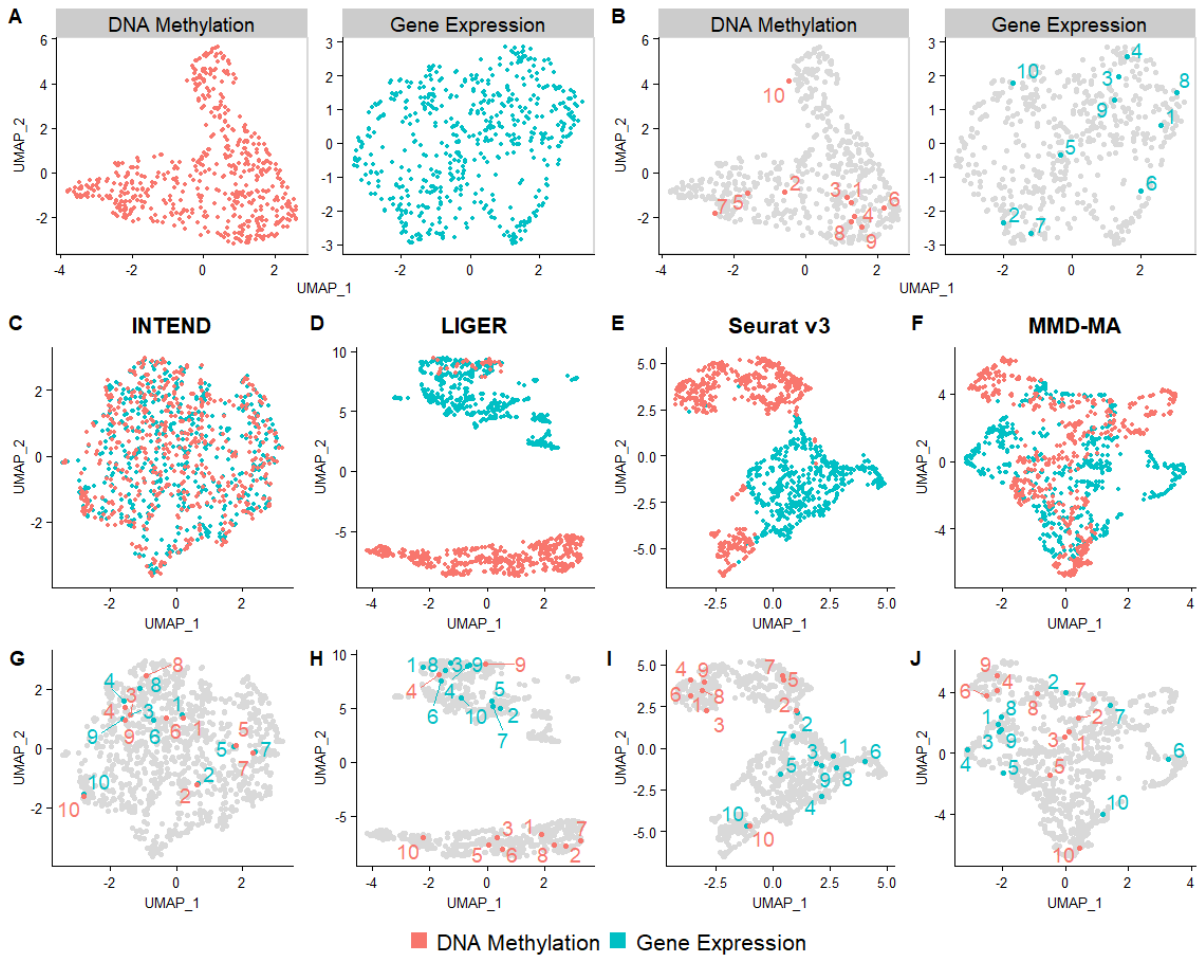
LUAD



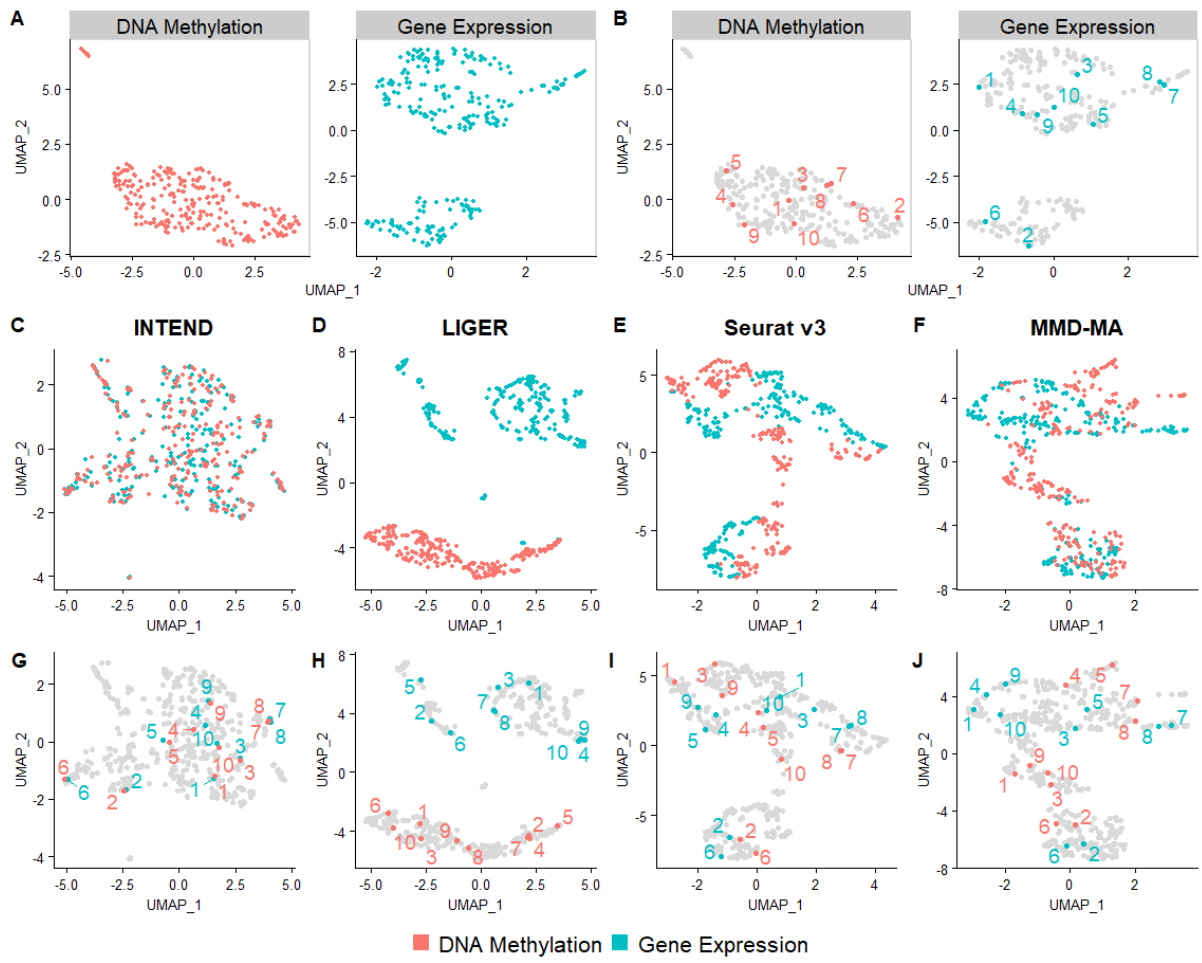
PAAD



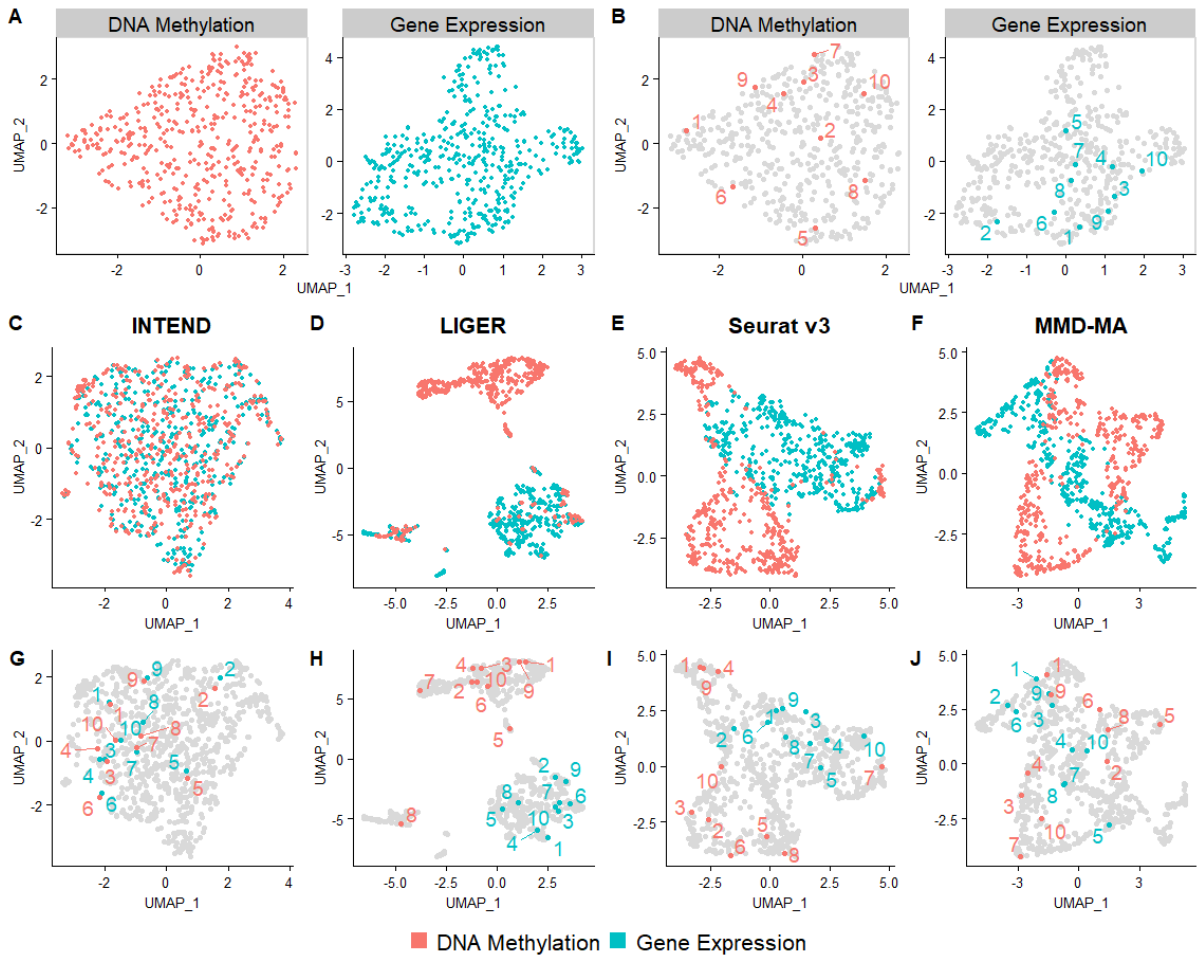
PRAD



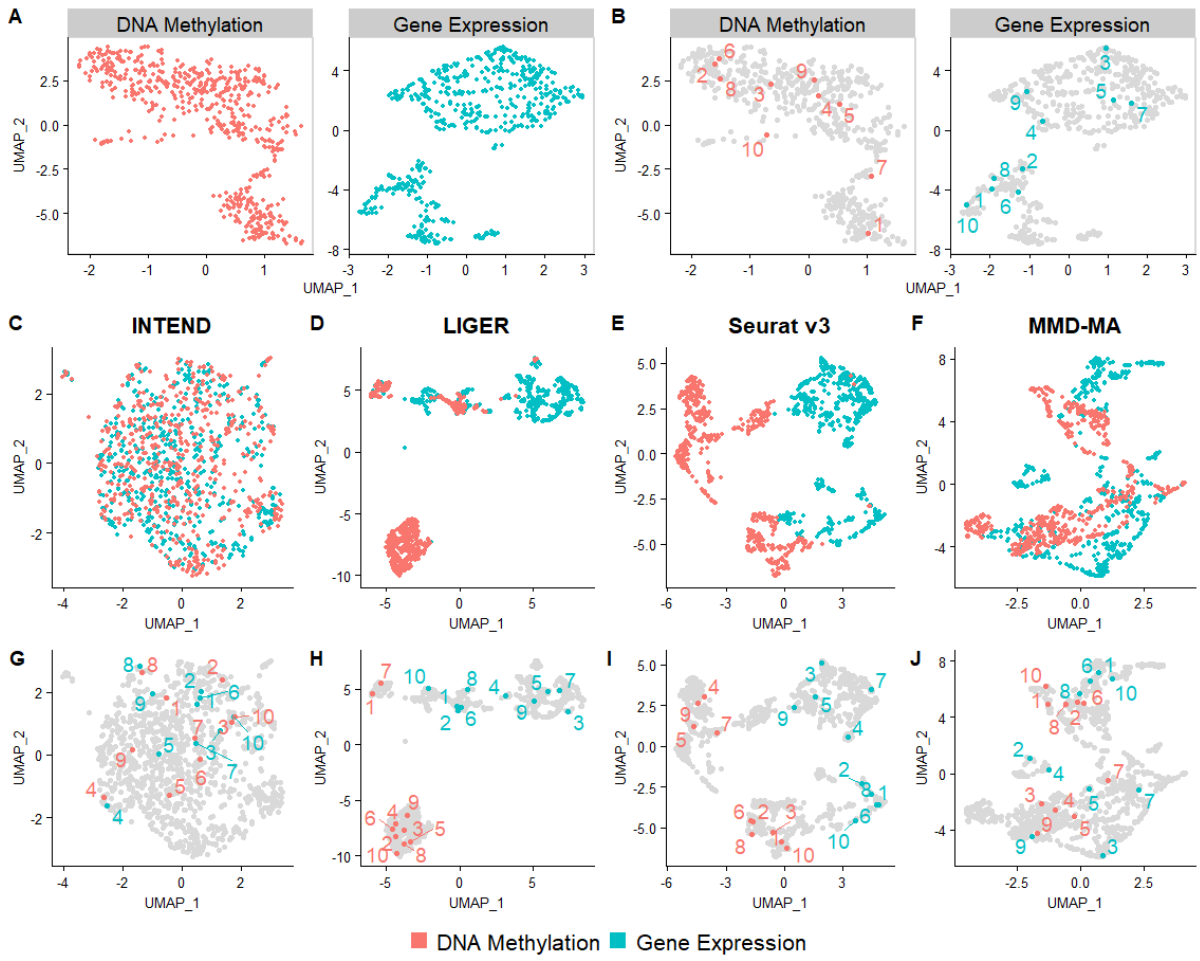
SARC

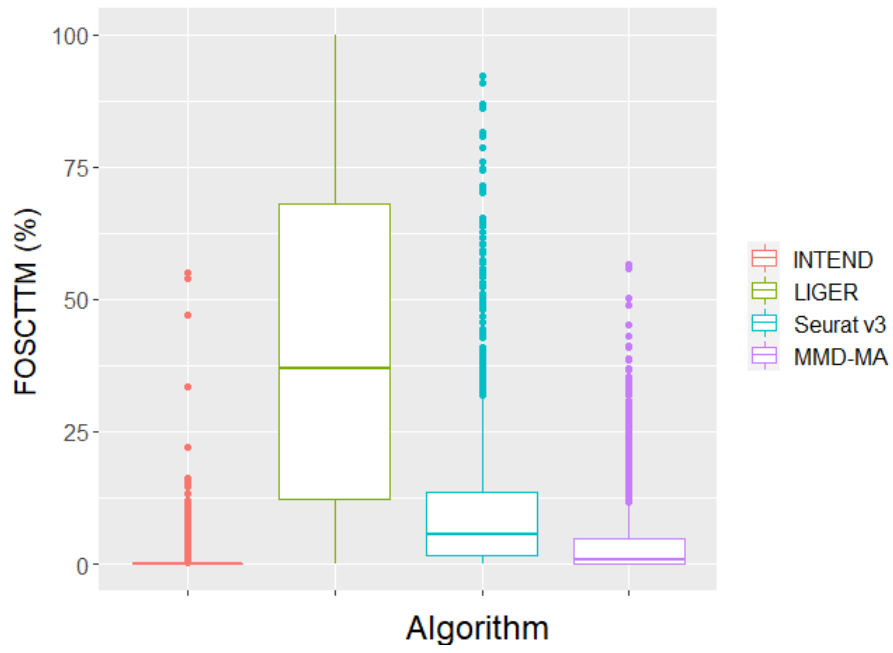


SKCM

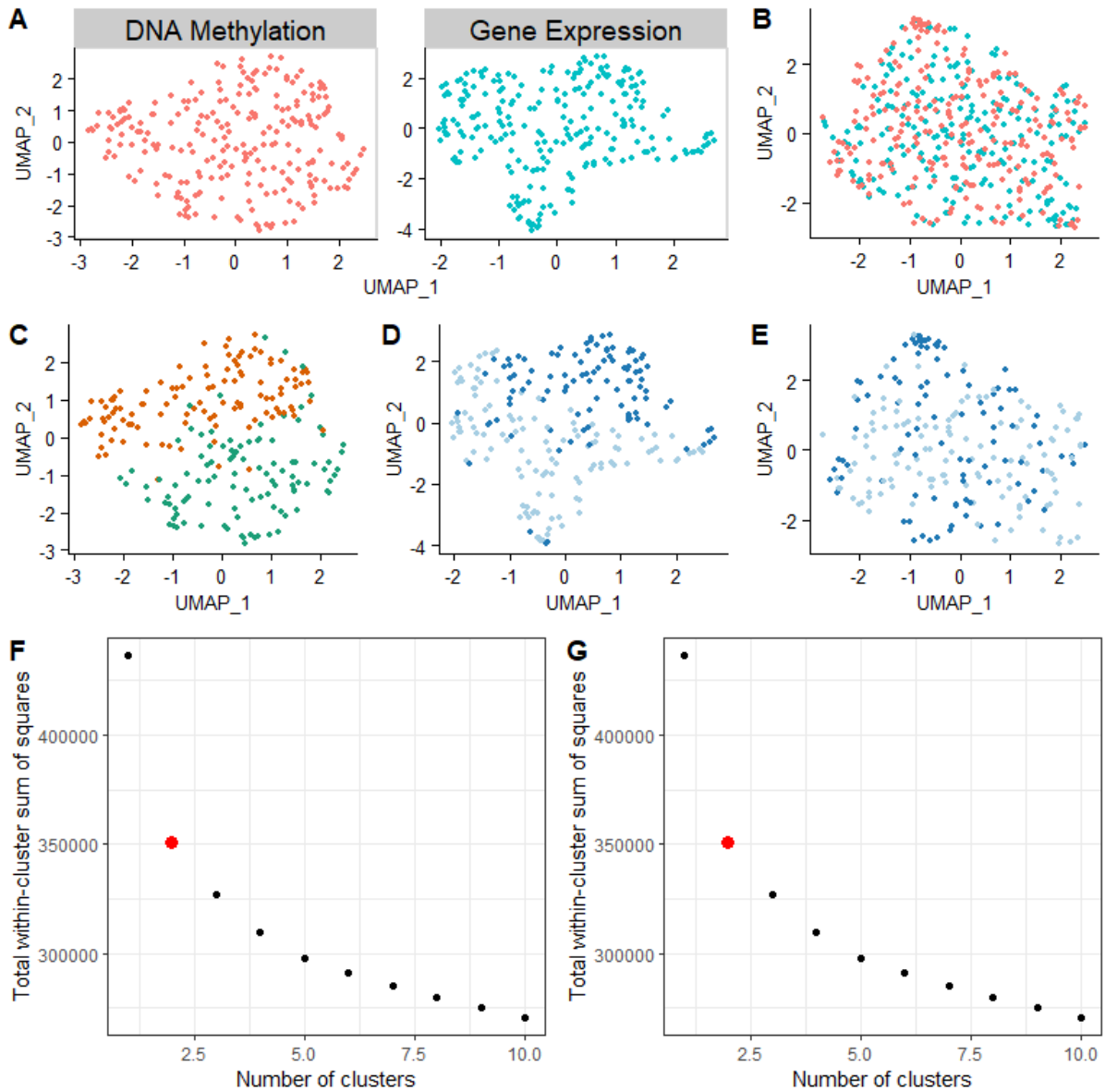


THCA





Supplementary Figure 17. FOSCTTM scores of the integration of four cancer datasets: COAD, LIHC, SARC and SKCM, simultaneously, by INTEND, LIGER, Seurat v3, and MMD-MA.



Supplementary Figure 18. SKCM clustering

(A) UMAP plots of each original omic data.

(B) UMAP plot of INTEND sample projections into the shared space colored by omic

(C) k-means clustering of the original DM samples with $k = 2$, shown on the same plot as in (A). Samples are colored according to their clusters.

(D) k-means clustering of the original GE samples with $k = 2$, shown on the same plot as in (A). Samples are colored according to their clusters.

(E) Integration-based clustering of the methylation sample embeddings into the shared space (EP, the pink points in (B)), shown on the same plot as in (B). Samples are colored according to their assigned cluster from (D). Each sample was assigned by majority voting to the cluster most represented among the five GE embeddings closest to its matching EP representation in the shared space.

(F) The total within-cluster sum of squares versus the number of clusters, for clustering DM data.

(G) The total within-cluster sum of squares versus the number of clusters, for clustering GE data.

The points with the maximum curvature are highlighted in red.