Tel-Aviv University

Raymond and Beverly Sackler Faculty of Exact Sciences

The Blavatnik School of Computer Science

Integration of Gene Expression and DNA Methylation Data Across Different Experiments

Thesis submitted in partial fulfillment of graduate requirements for

The degree "Master of Sciences" in Tel-Aviv University

School of Computer Science

By

Yonatan Itai

Under the supervision of

Prof. Ron Shamir

May 2023

Acknowledgments

I would like to express my sincere appreciation to the individuals who have supported me throughout my academic journey and have helped me reach this significant milestone. Without their unwavering support, I wouldn't have been able to achieve my academic goals.

First and foremost, I would like to extend my heartfelt thanks to my supervisor, Professor Ron Shamir. He has been an inspiring mentor throughout my research, providing me with invaluable guidance, encouragement, and professionalism. His vast knowledge and expertise have taught me a lot, and his constant support has helped me stay focused and motivated. Ron’s kind words and helpful correspondence helped me make it to the finish line.

I am also grateful to Nimrod Rappoport, who has provided me with exceptional support and guidance throughout this journey. Nimrod has been with me from the beginning and has been instrumental in advising me on my research. His support and encouragement have helped me overcome challenges and accomplish my goals.

I would also like to extend my appreciation to the members of the computational genomics lab, including Lianrong, David, Tom, Hagai, Dan C., Omer, Naama, Hadar, Dan F., Eran, and Ron. I also thank Gilit Zohar-Oren, who provided excellent administrative assistance.

Lastly, I would like to express my deepest gratitude to my family, my parents Moshe and Tamar, and my wife Yuval. They have been my pillars of strength and have provided me with endless support and understanding throughout my academic journey. Their belief in me has been the driving force behind my success, and I couldn't have achieved this milestone without their constant encouragement and motivation.

Abstract

Integrative analysis of multi-omic datasets has proven to be extremely valuable in cancer research and precision medicine. However, obtaining multimodal data from the same samples is often difficult. Integrating multiple datasets of different omics remains a challenge, with only a few available algorithms developed to solve it.

Here, we present INTEND (IntegratioN of Transcriptomic and EpigeNomic Data), a novel algorithm for integrating gene expression and DNA methylation datasets covering disjoint sets of samples. To enable integration, INTEND learns a predictive model between the two omics by training on multi-omic data measured on the same set of samples. In comprehensive testing on eleven TCGA cancer datasets spanning 4329 patients, INTEND achieves significantly superior results compared to four state-of-the-art integration algorithms. We also demonstrate INTEND's ability to uncover connections between DNA methylation and the regulation of gene expression in the joint analysis of two lung adenocarcinoma single-omic datasets from different sources. INTEND's data-driven approach makes it a valuable multi-omic data integration tool.

The code for INTEND is available at https://github.com/Shamir-Lab/INTEND.

Contents

1 Introduction

Emerging technological advances in recent years have made high throughput genome-wide sequencing a central tool for biological research. It allows the collective analysis of various types of biological data (commonly termed 'omics'), in a single tissue or even at the level of a single cell. These include genomics – covering the DNA sequence itself; transcriptomics – the expression levels of genes in the form of messenger RNAs; epigenomics –reversible modifications on the genetic data, e.g. DNA methylation and chromatin accessibility; proteomics – the levels of translated proteins; and more (**Figure 1**). Although the analysis of a single omic may generate meaningful insights, a multi-omic integrative analysis can lead to comprehensive understanding of a biological system and its complexities. For brevity, will use throughout the term *integration* for integrative analysis. Hence, integrating different omic datasets is one of the most interesting challenges in computational biology today, with the potential of opening new avenues in cancer research and precision medicine (1–3)

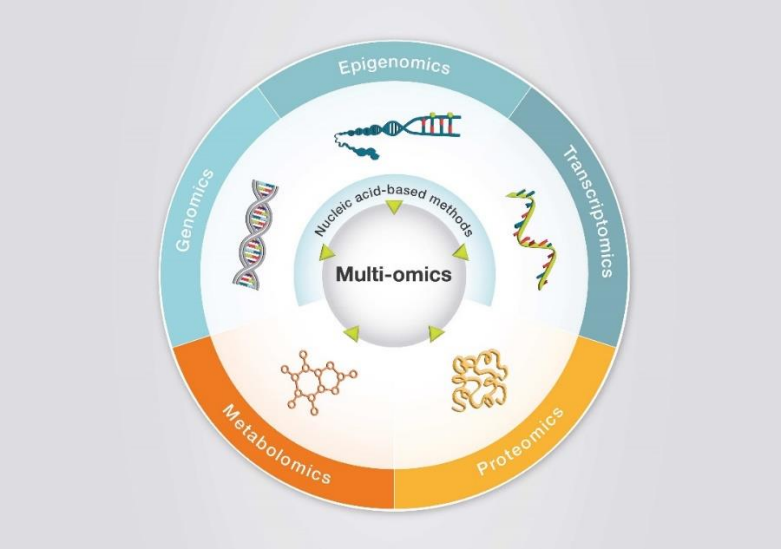


Figure 1. Multi-omics (Figure source: https://www.thermofisher.com/il/en/home/brands/thermo-scientific/molecular-biology/molecular-biology-learning-center/molecular-biology-resource-library/spotlight-articles/supporting-multi-omics-approaches.html)

1.1 Multi-omic integration – diverse problems, diverse approaches

One way to obtain multi-omics data for analysis is to simultaneously measure more than one omic from the same tissue. For example, TCGA (The Cancer Genome Atlas) (4) contains multimodal data for numerous tissues spanning dozens of cancer types. The main data types covered by TCGA are genotype, copy number variations, genome methylation, mRNA expression, and miRNA expression, along with clinical data. Multimodal data can be also obtained at the cell level by simultaneously measuring multiple types of molecules within the cell (5–7). Such technologies are relatively new and expensive, and thus so far there is much less data of multiple omics from the same cells.

Schematically, we can categorize the integration problems into three scenarios (**Figure 4A**):

a. Single omic – multiple datasets (SO/MD). Here only one omic type is used but multiple datasets (typically experiments from different labs or studies) need to be analyzed together.
b. Multiple omic – single dataset (MO/SD). Here there is one set of samples on which several omics were measured, and the feature sets of the different omics are disjoint.
c. Multiple omics – multiple datasets (MO/MD). This problem generalizes both (a) and (b).

Many algorithms were developed to handle the integration in the MO/SD setting. These include DIABLO (8), iCluster (9), and MOFA/MOFA+ (10, 11), which use latent variable analysis approach; iNMF (12), which uses non-negative matrix factorization; similarity-based methods like SNF (13), NEMO (14, 15) and MONET (16); and scAI (17), which specializes in single-cell data. Other algorithms were developed to tackle the integration in the SO/MD setting. These algorithms should balance the tradeoff between the removal of batch effects and the

conservation of biological variance (18). Relevant examples are MNN (19), Seurat v3 (20), scVI (21), Scanorama (22), LIGER (23), Conos (24) and Harmony (25).

The challenge we address in this paper is the composition of the two problems discussed above: MO/MD integration. Only a few algorithms have been developed to tackle this challenge. Both LIGER and Seurat v3 were used to integrate different omic datasets of disjoint sets of cells, specifically transcriptome and epigenome datasets. LIGER was shown to integrate scRNA-seq with genome-wide DNA methylation, and Seurat to integrate scRNA-seq with scATAC-seq (measuring chromatin accessibility).

The motivation behind integrating datasets across different experiments arises from the difficulties to obtain multimodal data from the same samples. These difficulties may be technical inabilities, as mentioned in the context of single-cell data, and economical, a significant factor also in the case of bulk sequencing data. An algorithm that can integrate two different omic datasets measured from disjoint sets of samples, could assist researchers in utilizing data that has already been collected in the past, allowing a multi-omic systemic view on the investigated subject. This could increase efficiency, both in time and in cost. Consider the situation where the methylation patterns inside tumors of a specific cancer subtype are being investigated. The multi-omics approach could suggest further inquiry of the epigenome-transcriptome connections, i.e. obtaining mRNA sequencing from every tumor and conducting an integrative analysis of the methylation and gene expression patterns together. As RNA-seq data is widely available for many cancer subtypes, it may be the case that such RNA-seq data is already available for other samples of that cancer subtype. With an algorithm that can integrate RNA-seq and DNA methylation datasets measured on disjoint samples, the researcher could conduct an integrative multi-omic analysis while measuring only the methylation patterns, thus requiring fewer resources.

The algorithms for MO/MD integration can be classified according to the correspondence information they require as input. Some methods require partial correspondence between the samples (either tissues or cells). One example is the semi-supervised correspondence approach of the MAGAN algorithm (26). This approach uses matching pairs of samples from both datasets to learn the correct alignment of the datasets. Other methods, as LIGER and Seurat, require correspondence information between the features of the different omics. Finally, some methods do not require any correspondence information and assume a common underlying structure that is maintained across technologies and omics. Such methods usually belong to the class of machine learning algorithms that solve the unsupervised manifold alignment problem. One algorithm that uses such techniques to integrate single-cell multi-omics data is the maximum mean discrepancy-manifold alignment (MMD-MA) algorithm (27) Another algorithm that can jointly embed two datasets, without any correspondence information between their features or samples, is the joint Laplacian manifold alignment algorithm (JLMA) (28). Using a method that does not require any correspondence information may sound appealing, but may not perform adequately when the assumed common underlying structure is weak.

In our study, we developed a method for the integration of transcriptomic and epigenomic data across different experiments. We focused on the integration of gene expression and DNA methylation. Specializing in two particular creates a less general method, but allows us to develop a stronger model: we can incorporate the known biological connections between gene expression and DNA methylation.

1.2 Associations between DNA methylation and gene expression

The regulation of gene expression allows cells to increase or decrease the production of proteins or RNA. Such adjustments enable response to external changes in the environment and to internal signals within cells. In complex multicellular organisms, the regulation of genes

in particular cellular contexts enables the differentiation and proliferation of cells. Epigenetic modifications mainly include DNA methylation and histone protein modifications, which alter the chromatin structure. These modifications are known to be key factors in the regulation of gene expression. In the last two decades, a strong connection has been established between epigenetic modifications and the development of cancer. Hence, the integration of transcriptomic and epigenomic data has the potential to broaden our understanding of the molecular mechanisms orchestrating the regulation of genes, in both normal and malignant tissues.

DNA methylation in mammals occurs almost exclusively in the 5' position of a Cytosine followed by a Guanine, commonly termed a CpG site. CpG dinucleotides tend to cluster in CpG islands (CGIs), regions with a high frequency of CpG sites. The majority of CpG dinucleotides (75%) throughout the mammalian genomes are methylated (29), except for CGIs, which are mostly unmethylated. About 70% of the proximal promoters of human genes contain a CGI, and reciprocally, about 50% of the CGIs are located near a gene's transcription start site (TSS). In fact, CGIs are strongly linked to the regulation of transcription (30). Although CGIs are mostly hypomethylated, there are known examples of their methylation, resulting in stable silencing of the associated promoter (**Figure 2**). However, it is believed that CGI methylation does not initiate the silencing of genes, but assists in making the silenced state permanent (30). For example, in X chromosome inactivation, the methylation process of CGIs in the X chromosome has been shown to start only after gene silencing. However, when DNA methylation is inhibited, genes in the X chromosome can be reactivated.

The connection between CGI hypermethylation and silencing of genes is not the only relationship observed between methylation and gene expression. There is evidence of both strong positive and strong negative correlations between gene-body methylation and gene expression (31). Other studies have shown that hypermethylation of CGIs in cancer tissues is

not always accompanied by a decrease in gene expression (32). These findings suggest that DNA methylation can play diverse roles in gene regulation, depending on the genomic context (33). This should be considered when using multi-omic integration algorithms like LIGER and Seurat, which require correspondence information between the features of the different omics. The methods that are currently used to link the feature spaces of DNA methylation and gene expression assume a simplistic connection between the two (see LIGER description in the next section). The complex and not fully understood relationship between DNA methylation and gene expression stresses the necessity for a more sophisticated approach.

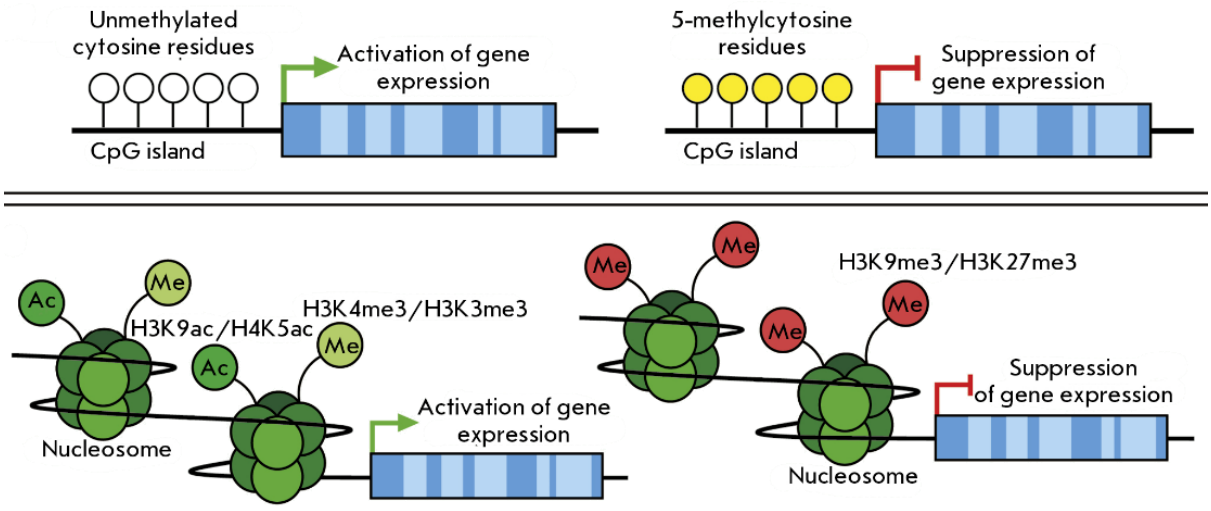


Figure 2. The major epigenetic mechanisms regulating gene expression. (A) – Methylation of cytosine residues in the CpG island located in the gene promoter region. (B) – The most common modifications of the histone proteins involved in gene expression activation and suppression. (Figure source: DNA Methylation As an Epigenetic Mechanism in the Development of Multiple Sclerosis, https://actanaturae.ru/2075-8251/article/view/11043).

1.3 Our approach

In this paper, we present a novel algorithm for the MO/MD problem. The algorithm is called INTEND (IntegratioN of Transcriptomic and EpigeNomic Data). Specifically, INTEND aims to integrate gene expression (GE) and DNA methylation (DM) datasets covering *disjoint* sets of samples. INTEND does not use any correspondence information between the samples in the

two datasets (e.g. knowing which GE and DM profiles originated from the same individual). To handle the complex connections between DM and GE, INTEND learns a predictive model between the two, by training on multi-omic data measured on the same set of samples. To the best of our knowledge, this is the first use of a predictive model in the context of the studied problem.

As a preliminary step, for each gene, INTEND learns a function that predicts its expression based on the methylation levels in sites located proximal to it. To integrate the target methylation and gene expression datasets, INTEND first predicts for each methylation profile its expression profile. Then, it identifies a set of genes that will be used for the joint embedding of the expression and predicted expression datasets. At this stage, both datasets share the same feature space. INTEND then employs canonical-correlation analysis (CCA) to jointly reduce their dimension.

We evaluated the performance of INTEND by comparing it to four state-of-the-art MO/MD integration methods: LIGER, Seurat v3, JLMA, and MMD-MA. The first two require correspondence information between the different omic features, in order to create a common feature space before the integration, whereas the last two do not require such information. We used eleven TCGA cancer datasets spanning 4329 patients for testing the algorithms in multiple integration tasks. We also showed the utility of the method in identifying SKCM cancer subtypes and in joint analysis of LUAD using two single-omic datasets obtained from different individuals.

1.3.1 Integration methods used in the benchmark

LIGER

LIGER (23) 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 (34). 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 (20) 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 (35). 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 (28) 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 (27) 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.

1.3.2 Additional computational background

Lasso Regression

INTEND uses Lasso regression (37, 38) to learn a predictive model between DM and GE. Lasso regression is a multivariate linear regression with an l_1-norm penalty. Given a dataset $\left(\left(x_{i1}, x_{i2}, \dots x_{ip}\right), y_i\right), i = 1,2, \dots, n,$ where x_{ij} are the predictor variables and y_i are the

response values, Lasso solves the l_1-regularization problem by finding $\beta = \{\beta_1, \beta_2, \ldots \beta_p\}$ that minimizes:

$$\sum_{i=1}^{n}\left(y_i - \sum_j x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| \quad (1)$$

This is equivalent to minimizing the sum of squares with the constraint $\sum|\beta_j| \leq c$. This shrinks the coefficients, some of them to zero, thus employing also variable selection. The tuning parameter λ controls the power of the l_1-penalty. We solve the minimization problem via the coordinate-descent method and choose the optimal λ using 10-fold cross-validation on the training set. This is done with the glmnet R package.

Cross validation

K-fold cross-validation uses the following approach to evaluate a model **(Figure 3)**:

- **Step 1**: Randomly divide a dataset into k groups, or "folds", of roughly equal size.
- **Step 2**: Choose one of the folds to be the holdout set. Fit the model on the remaining k-1 folds. Calculate the test mean squared error (MSE) on the observations in the fold that was held out.
- **Step 3**: Repeat this process k times, using a different set each time as the holdout set.
- **Step 4**: Calculate the overall test MSE to be the average of the k test MSE's.

•

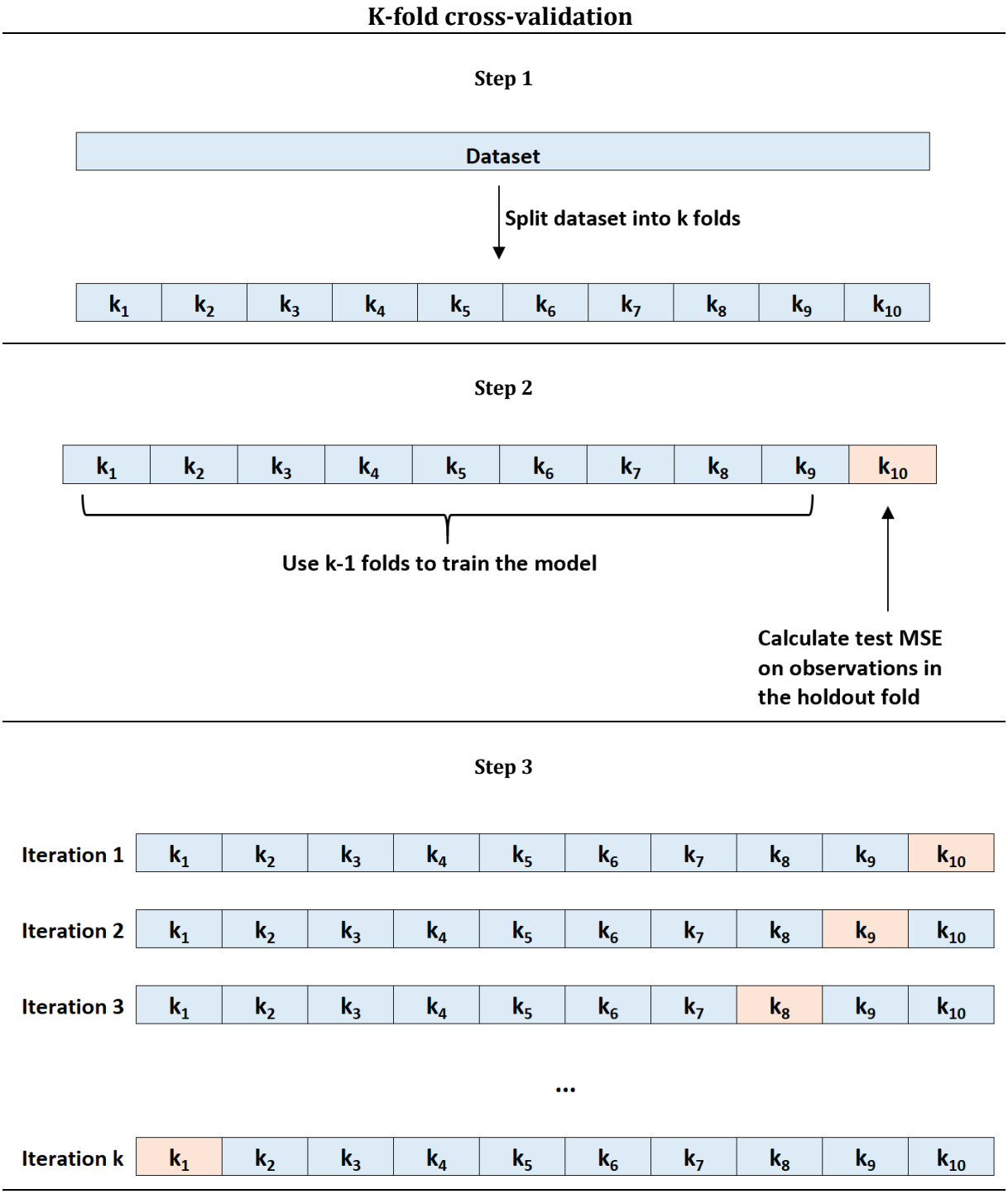


Figure 3. K-fold cross validation (source: https://www.statology.org/k-fold-cross-validation)

2 Materials and Methods

2.1 INTEND algorithm

INTEND works in two phases (**Figure 4B**). The training phase receives as input training data consisting of GE and DM profiles measured on the same set of samples. The algorithm uses this data to learn the connections between the omics. This will allow it later to make accurate predictions of expression levels of specified genes based on a given methylation profile. The training process can be executed once for any number of future integration tasks. Intuitively, the multimodal data used in the training process should be "biologically similar" to the datasets that INTEND will integrate subsequently. However, as we shall show, even when we used INTEND to integrate datasets covering tumor types that were different from the ones covered by the multimodal training data, it performed well.

For the embedding phase, INTEND's inputs are from two disjoint cohorts, denoted T1 and T2. They include a DM matrix for T1 and a GE matrix for T2. It proceeds in three steps: (1) Creation of predicted GE matrix for T1 based on the DM data. (2) Selection of a subset of the genes based on the predicted GE for T1, the GE for T2, and the trained model from the preliminary step. (3) Reducing jointly the dimension of the two GE datasets on the selected gene set.

2.1.1 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 $p^{(g)}: \mathbb{R}^{|f_M^{(g)}|} \longrightarrow \mathbb{R}$, s.t. $p^{(g)}\left(m^{(g)}\right)$ is the predicted expression level of g.

Model

We hypothesized that accurately predicting the expression levels of even a small number of genes, from an input methylation matrix, will enable successful integration. To achieve this goal, we developed a prediction model considering the known connections between methylation in promoter CGIs and gene expression (30), as well as gene-body methylation (31). Furthermore, to capture the variation in the correlation between methylation and expression across the CGI, its shores and shelves, and also outside CGIs (32), the model uses the methylation levels in each probe separately.

For each $g \in f_E$ we set $f_M^{(g)}$ to be all the probed methylation sites in the range $[C_{5'\text{-end}} - 10\text{kb}, C_{3'\text{-end}} + 10\text{kb}]$, where $C_{5'\text{-end}}$ and $C_{3'\text{-end}}$ are the coordinates of g's $5'$-end and $3'$-end on the chromosome, respectively. While in certain cases more distal methylation sites were reported to affect gene expression (36), the main effect is usually due to proximal sites (30). We limited the range in order to have modest-size gene models. As we will show, such models provide a good basis for the integration task.

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 $f_M = \bigcup f_M^{(g)}$ the union of the used methylation sites for all genes.

For each g, after obtaining $f_M^{(g)}$, INTEND uses Lasso regression model (37, 38) 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. Using Lasso allows the preliminary step to handle genes with a large number of methylation sites, by ignoring sites that have little relevance for the gene expression prediction. For example, in a TCGA training set that we used, spanning 10 cancer subtypes (the datasets listed in **Table 1**, excluding LUAD) and spanning 3852 tumor samples, $f_M^{(g)}$ contained 25 sites on average, with a maximum of 1055 sites per gene (**Supplementary Figure 1**). However, the maximal number of probes for which the regression resulted with non-zero coefficients for a specifed gene was only 424, with an average of 21 sites per gene (**Supplementary Figure 2**).

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. For example, using the above training set, the average R^2 of all 19143 genes considered was 0.30, and the average R^2 of the top 2000 genes was 0.68 (**Supplementary Figure 3**).

Note that when applying the preliminary step to certain cancer subtypes, the subsequent algorithmic steps use only data from other subtypes, in order to avoid overfitting.

2.1.2 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 $m_1, m_2, \ldots, m_{n_M}$ 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 $e_i = \left\{ e_i^{(g)} \,\middle|\, g \in f_E \right\}$. This step results in the predicted expression matrix $P = (e_1, e_2, \ldots, e_{n_M})$.

Step 2: Selecting genes

Denote the 2000 genes selected in the training phase by G_R. The expression of these genes has the highest likelihood to be predicted accurately by the methylation profile, at least in the tissue types and states included in the training set. However, the target datasets may originate from a different tissue type or state. Hence, an additional heuristic for feature selection is employed. Genes may be regulated by mechanisms other than DNA methylation. Thus we assumed that the genes that are most likely to be regulated by the methylation profile are the ones with high variance in both methylation and expression levels. 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:

$$G_s = G_R \cap G_E \cap G_P \tag{2}$$

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 (39). It was used in computational genomics to project datasets that share the same samples but have different features (the MO/SD setting) to a common low-dimensional feature space. CCA has been used in this way, for example, in multi-omic clustering (15, 40). In contrast, here we apply CCA to E_{G_s} and P_{G_s}, which covers samples from different datasets but share the same set of selected genes (similar to the SO/MD setting). This approach for utilizing CCA was introduced in Seurat v2 (41).

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:

$$(u_1, v_1) = \underset{u \in \mathbb{R}^{n_E}, v \in \mathbb{R}^{n_M}}{\operatorname{argmax}} u^T X^T Y v \quad s.t \begin{cases} u^T X^T X u = 1 \\ v^T Y^T Y v = 1 \end{cases} \tag{3}$$

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:

$$(u_1, v_1) = \underset{u,v}{\operatorname{argmax}} u^T X^T Y v \quad s.t \begin{cases} \|u\|_2^2 = 1 \\ \|v\|_2^2 = 1 \end{cases} \tag{4}$$

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. See the Supplement for details.

The code for INTEND is available at https://github.com/Shamir-Lab/INTEND.

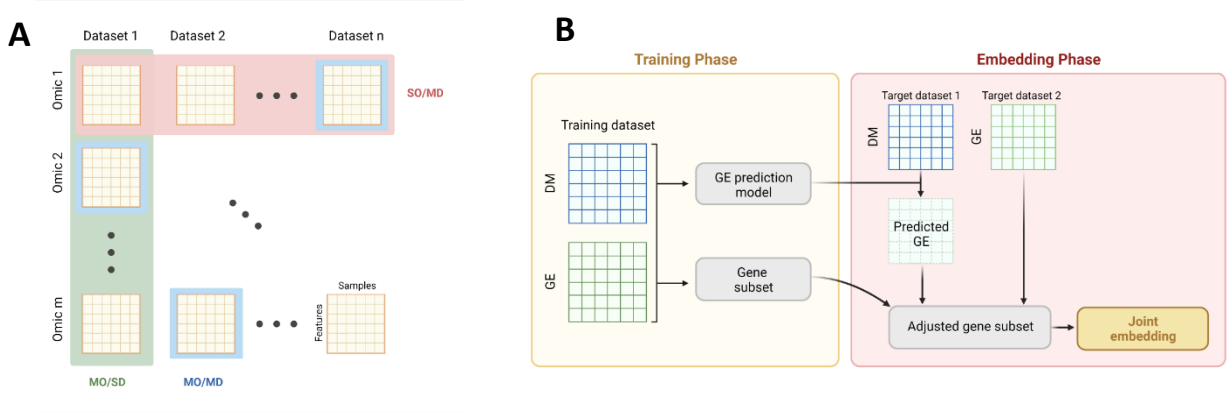


Figure 4. (A) Three scenarios of integration problems: Green: single omic – multiple datasets (SO/MD); red: multiple omic – single dataset (MO/SD); blue: multiple omics – multiple datasets (MO/MD). (B) An overview of the two phases of INTEND: the training phase and the embedding phase.

2.2 Data

2.2.1 TCGA data

To assess performance, we used RNA-seq and DM data from TCGA (4) covering 11 different cancer types. See **Table 1** for cancer types, their abbreviations and statistics. The data was downloaded using the TCGA-Assembler software (42, 43). We used only 4329 samples for which both omics were measured.

The DM data we used was gathered with Illumina's Infinium HumanMethylation450 BeadChip assay. The levels of $>$ 450,000 methylation sites were reported as β-values. The RNA-seq data was gathered with Illumina HiSeq assay, and quantified using RSEM (44). In each GE and DM sample the zero counts were removed, then the raw count values were divided by the 75th percentile of the counts, and then multiplied by 1000. In both omics, we downloaded the data after these transformations from the TCGA website.

Table 1. Summary information of TCGA cancer datasets used

Cancer type	Abbreviation	Number of patient samples		
		Gene expression	DNA methylation	Both
Acute Myeloid Leukemia	**AML**	173	194	**170**
Bladder Urothelial Carcinoma	**BLCA**	427	440	**425**
Colon Adenocarcinoma	**COAD**	328	353	**298**
Brain Lower-Grade Glioma	**LGG**	534	534	**530**
Liver Hepatocellular Carcinoma	**LIHC**	424	430	**414**
Lung Adenocarcinoma	**LUAD**	576	507	**477**
Pancreatic Adenocarcinoma	**PAAD**	183	195	**183**
Prostate Adenocarcinoma	**PRAD**	550	553	**533**
Sarcoma	**SARC**	265	269	**263**
Skin Cutaneous Melanoma	**SKCM**	473	475	**473**
Thyroid Carcinoma	**THCA**	572	571	**563**

2.2.2 An additional LUAD gene expression dataset

In addition to the TCGA LUAD data, we used RNA-seq profiles from 172 tumors of LUAD patients from Singapore (45). GE was quantified with RSEM and normalized as done for the TCGA data.

2.2.3 Data preprocessing

To handle missing values, for each dataset, features with $> 5\%$ missing values were removed, and then samples with $> 5\%$ missing values were removed. Subsequently, the missing values per each feature were imputed to the mean of this feature across all samples. The number of features and samples in each dataset we used, before and after the handling of missing values,

are described in **Supplementary Table 1.** Finally, for GE data from all sources and for all purposes, we added 1 pseudo-count to each value and log-transformed the result.

2.2.4 Running other algorithms

We evaluated the performance of INTEND by comparing it to four state-of-the-art MO/MD integration methods: LIGER, Seurat v3, JLMA, and MMD-MA. The methods are briefly described in the Supplement. To use LIGER and Seurat, we supplied the algorithms with an aggregated gene-level methylation matrix as input, as they require correspondence information between features across omics. The aggregated matrix computation process is described in the Supplement. JLMA and MMD-MA algorithms do not require correspondence information between the features. However, empirical results from (27) showed that JLMA failed to integrate GE and DM using the local geometry metric as a measure for cross-omic similarity. Hence, we computed the cross-omic similarity matrix for JLMA based on the aggregated gene-level methylation matrix. For MMD-MA we used both the original methylation data and gene-level methylation matrix as inputs. We denoted the runs of JLMA and MMD-MA with the gene-level methylation matrix as JLMA WFCI (with features correspondence information) and MMD-MA WFCI. We ran all the algorithms with their default recommended hyper-parameters, and whenever applicable, we used the algorithm's pipeline for feature selection and normalization. Since MMD-MA and JLMA do not include a method for feature selection, when running them in the WFCI mode, we selected the n genes with the highest variance in expression, for $n = 500$ and 2000. Further details regarding how each of the algorithms was applied, including hyper-parameters and additional necessary preprocessing steps, are described in the Supplement.

2.3 Evaluating the quality of the results

For the TCGA data, we have the true pairing of samples that represent different omic measurements of the same patient. This pairing is not given as input to the integration

algorithms and can therefore be used to evaluate their results. We use the metric defined in Liu et al. to evaluate the algorithms. For GE and DM input datasets covering n_E and n_M samples respectively, each algorithm produces a d-long vector of the projected expression e_i for each sample i and a d-long vector of the projected estimated expression m_j based on the methylation for each sample j. For patient i, let f_i be the fraction of samples j with projections m_j closer to e_i than m_i. We call it the "fraction of samples closer than the true match" (FOSCTTM). FOSCTTM ranges from 0 to 1, where 0 means that the true match of a sample i is the closest to i in the projected space. We calculate the FOSCTTM for every sample in the GE and DM datasets, and average these values. A perfect integration will have a score of 0. For a random projection, the expected FOSCTTM is 0.5.

2.4 Clustering

For clustering (subsection 3.3), we used the k-means algorithm of Hartigan and Wong (1979), with maximum number of 100 iterations and 100 different starting solutions. We selected the desired number of clusters using the "elbow method" as described in Rappoport and Shamir (2018). Let $v(i)$ be the total within-cluster sum of squares for a solution with i clusters, then we chose i for which the point $v(i)$ had the maximum curvature. Specifically, we chose the i that maximized the following approximation of the second derivative of v:

$$v[i+1] + v[i-1] - 2v[i] \quad (5)$$

3 Results

We applied INTEND in several settings. In the first part, we applied INTEND and four other algorithms in several integration tasks of GE and DM data, using eleven cancer datasets from TCGA. We also demonstrated the utility of the method in identifying SKCM cancer subtypes.

In the second part, we used INTEND for the integration of datasets from two different sources, covering two populations of LUAD patients.

Our first set of analyses compared five algorithms: INTEND, LIGER, Seurat v3 (hereafter: Seurat), MMD-MA, and JLMA. We used eleven datasets of different cancer types from TCGA. First, we integrated GE and DM data of the same cancer type, for each of the eleven types. Next, we integrated data of four cancer types simultaneously.

3.1 Single cancer dataset integration task

We first ran the algorithms with input datasets of a single cancer subtype. We used the eleven datasets listed in **Table 1**. For each dataset, we considered only the subset of samples measured in both omics. The total number of samples used in these integration tasks was 4329, where dataset sizes ranged from 170 to 563. For each cancer dataset, we trained a new regression model in INTEND's preliminary phase, using the samples of the remaining ten cancer datasets as the training set. To evaluate the results, we used the pairing information between samples from the two omics measured on the same tissue to calculate the FOSCTTM score.

We ran the algorithms using projected space dimension d ranging from 2 to 40, and recorded the best integration scores (average FOSCTTM). The results are summarized in

Table 2 and **Supplementary Figure 5**. INTEND performed best across all datasets and all d values, and substantially better than the rest, with MMD-MA the second performer. In fact, INTEND results were often 1-2 orders of magnitude better than those of all the other methods.

Table 2. Average FOSCTTM of algorithms for integrating GE and DM data

Cancer/Alg	INTEND	LIGER	Seurat v3	MMD-MA	MMD-MA WFCI (500)	MMD-MA WFCI (2000)	JLMA WFCI (500)	JLMA WFCI (2000)
AML	**2.42** (25)	29.83 (7)	17.05 (36)	23.63 (40)	19.08 (40)	22.35 (40)	24.01 (8)	28.38 (7)
BLCA	**0.04** (39)	39.62 (9)	13.86 (40)	11.20 (40)	16.34 (40)	14.58 (40)	34.80 (40)	37.11 (40)
COAD	**0.02** (37)	26.84 (19)	19.14 (40)	12.59 (40)	12.19 (40)	12.92 (40)	32.98 (5)	34.73 (4)
LGG	**6.82** (22)	41.97 (8)	32.06 (26)	8.88 (40)	15.50 (40)	12.08 (40)	37.41 (14)	32.38 (12)
LIHC	**0.14** (36)	42.34 (3)	19.23 (38)	16.04 (30)	11.02 (30)	12.94 (30)	32.68 (21)	36.03 (12)
LUAD	**0.06** (32)	36.72 (4)	16.36 (39)	8.71 (40)	14.11 (40)	13.89 (40)	29.60 (9)	32.16 (8)
PAAD	**0.55** (30)	36.68 (15)	24.18 (35)	11.07 (40)	23.42 (40)	16.27 (40)	29.83 (3)	27.44 (2)
PRAD	**0.37** (38)	35.96 (8)	16.32 (17)	10.88 (40)	11.15 (40)	10.99 (40)	27.14 (2)	29.53 (2)
SARC	**0.05** (35)	42.06 (15)	12.86 (36)	8.86 (40)	20.97 (40)	17.42 (40)	34.47 (7)	34.73 (5)
SKCM	**0.03** (39)	42.20 (17)	18.97 (37)	16.02 (40)	20.53 (40)	16.62 (40)	32.11 (15)	34.71 (3)
THCA	**3.07** (11)	32.58 (7)	15.96 (36)	6.71 (40)	7.78 (40)	6.65 (40)	30.95 (2)	27.52 (5)
Average (all datasets)	**1.23** (31)	36.98 (10)	18.73 (34)	12.24 (39)	15.64 (39)	14.25 (39)	31.45 (11)	32.25 (9)

Average FOSCTTM score (percent) for each algorithm on each of the eleven cancer datasets. The optimal score is 0%, and the expected score for a random projection is 50%. The requested shared space dimension d ranges from 2 to 40 for each algorithm. The score shown is the best across all values of d, and the optimal d is written in parenthesis. The numbers 500 and 2000 for MMD-MA and JLMA denote the number of selected genes in the WFCI runs.

We also analyzed the contribution of the last step in INTEND – applying CCA for dimension reduction – to its performance. We measured the average FOSCTTM when using the original GE data and the imputed GE computed by INTEND, for the selected gene set (see subsection 2.1.2). Excluding the CCA step resulted in poorer FOSCTTM scores. Notably, these scores were better than all other tested algorithms in all datasets, with only one exception **(Supplementary Table 2)**.

In later analyses, we preferred to use the same space dimension d for all algorithms. MMD-MA and JLMA do not recommend a method for determining d. For Seurat, the authors originally suggested approaches to select d (41) but later noted that the identification of this

value remains a challenge (20). After running all methods for $d \in [2,40]$ for all datasets, we observed that most algorithms reach a plateau in the FOSCTTM score around $d = 40$ (**Supplementary Figure 5**). Hence, in subsequent runs we set $d = 40$ for all algorithms, with one exception: LIGER failed to run on the AML dataset with $d = 40$ or $d = 39$, so in that case we used $d = 38$.

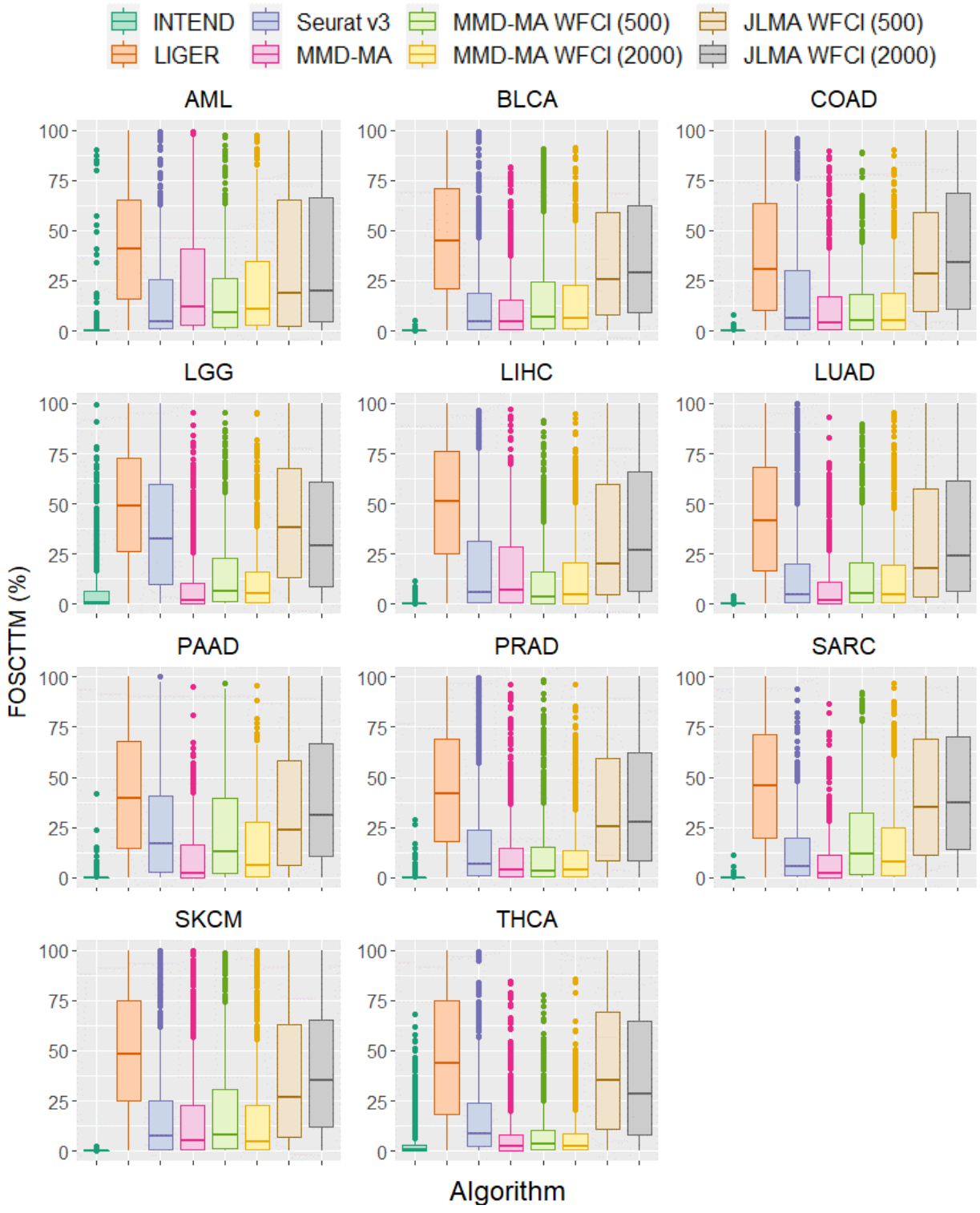


Figure 5. Distribution of FOSCTTM (%) scores in INTEND results on each cancer type.

Next, we analyzed the FOSCTTM per sample across all methods and datasets. **Figure 5** shows boxplots of the FOSCTTM per sample for each algorithm and cancer dataset using $\boldsymbol{d = 40}$.

INTEND's advantage was prominent, with the entire FOSCTTM interquartile range ($\boldsymbol{IQR}$) at zero for eight of the 11 datasets tested. In six of the 11 datasets, the FOSCTTM was perfect (zero) for $> \mathbf{90}\%$ of the samples.

We analyzed in more detail the results for the COAD dataset. We used UMAP (47) for the 2D projection of the samples from the original omic feature spaces and from the integration shared space. **Figure 6** shows the results for INTEND, LIGER, Seurat, and MMD-MA algorithms. The results for JLMA WFCI and MMD-MA WFCI versions are presented in **Supplementary Figure 6**.

Figure 6A-B show the projections from the original feature spaces. One can appreciate that pairwise distances are not preserved between the omics. **Figure 6C-F** show for each algorithm the projections from the shared feature space. It is evident that the level of mixing between the two omics is highest for INTEND, intermediate for MMD-MA and lower for Seurat and LIGER. **Figures 6G-J** show the same projections as in **Figures 6C-F** with the 10 samples of **Figure 6B** marked. Evidently, INTEND does a much better job in projecting omics from the same sample to close positions. For example, the two points labeled 3 belong to distinct clusters of samples in both the DM and the GE spaces. INTEND was the only method to succeed in projecting the points from both omics into the same cluster in the shared space. A similar advantage of INTEND was obtained for all other cancer types, even when the average FOSCTTM was higher (**Supplementary Figures 7-16**).

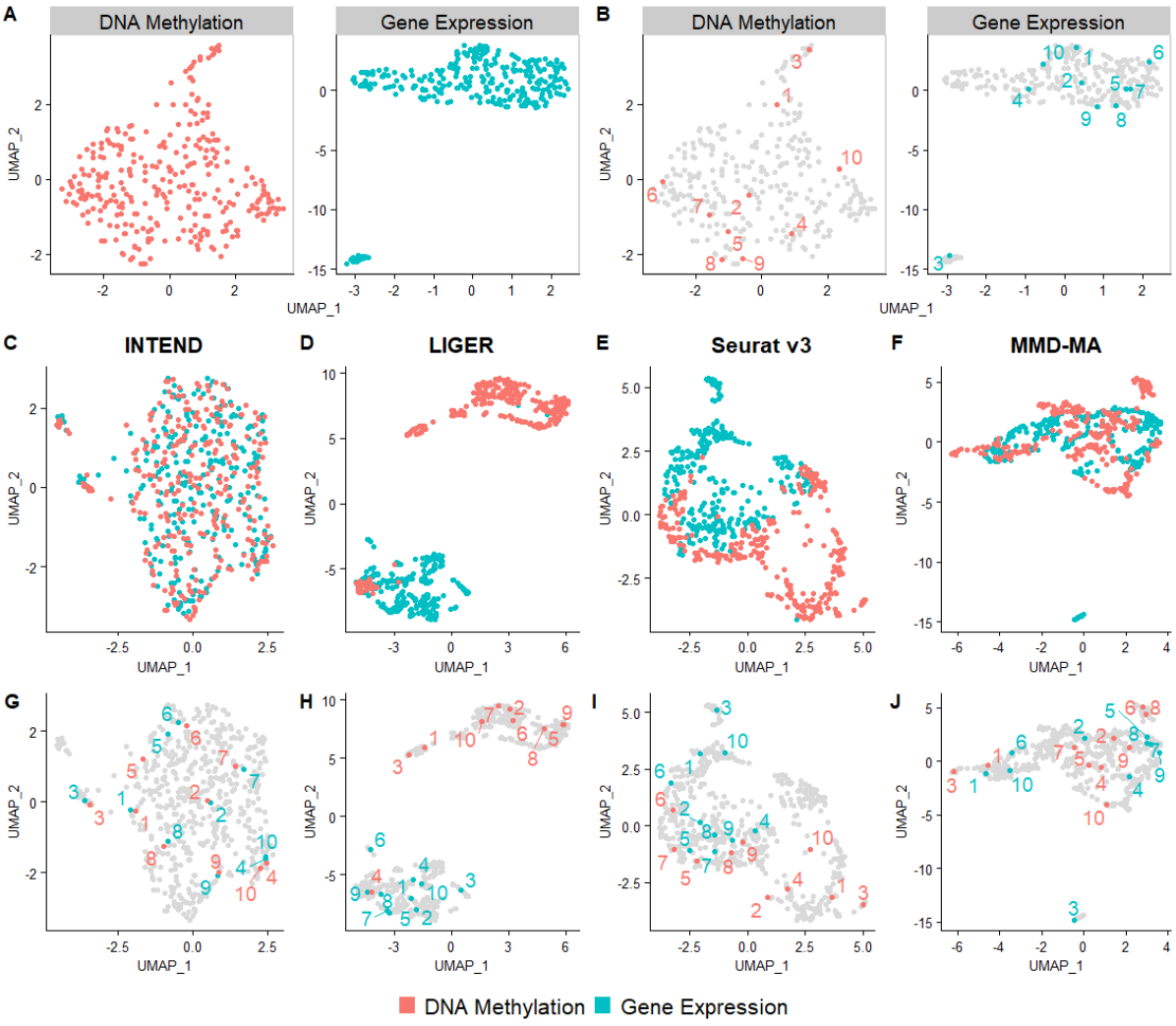


Figure 6. Results of integration of GE and DM samples from the colon adenocarcinoma dataset by different algorithms. (A) UMAP plots of the original data. (B) The same plots as in A. To appreciate concordance between omics, ten samples were randomly selected, and their matching points in both omics were 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.

3.2 Joint integration of multiple cancer types

In a second test, we applied the algorithms on four cancer datasets simultaneously. We used the datasets of COAD, LIHC, SARC and SKCM, covering 1448 GE and DM profiles. We did not supply the cancer type of each sample to the algorithms. We used the remaining seven TCGA datasets as the training set in INTEND's training phase. INTEND performed this task with the best FOSCTTM integration score (**Supplementary Figure 17**), with perfect

FOSCTTM for > 65% of the samples, and 1-2 orders of magnitudes better than the other methods: The mean scores were 0.37% for INTEND, 41.59% for LIGER, 9.33% for Seurat, and 4.01% for MMD-MA.

Figure 7 shows 2D projections of the mapping by each of the algorithms. INTEND, Seurat and MMD-MA projected the samples from the different cancer datasets into separate clusters in the shared space (**Figure 7G-J**). In contrast, LIGER failed to preserve the biological variance among the tissue types, mapping samples of different types to the same clusters (**Figure 7I**). While INTEND mixed the samples from both omics in each cancer type cluster, Seurat and MMD-MA created clusters with substantial separation between the samples from each omic (**Figure 7C-F**).

To further evaluate the results, we tested the quality of classifying the DM samples to specific cancer types based on the types of their neighboring GE samples in the shared space, as follows. Each DM sample was assigned by majority voting to the cancer type most represented among its five closest GE samples in the shared space. The confusion matrices between the inferred and true assignments are shown in **Figure 7K-N**. INTEND performed best, with > 97% of DM samples in each cancer type correctly classified. MMD-MA performed slightly worse: three cancer types had high accuracy classification, but the SARC cancer type had > 9% of the samples misclassified as SKCM. For Seurat, three cancer types had high accuracy classification, but the SKCM cancer type had >26% of the samples misclassified as SARC. The LIGER projections led to the lowest accuracy classification.

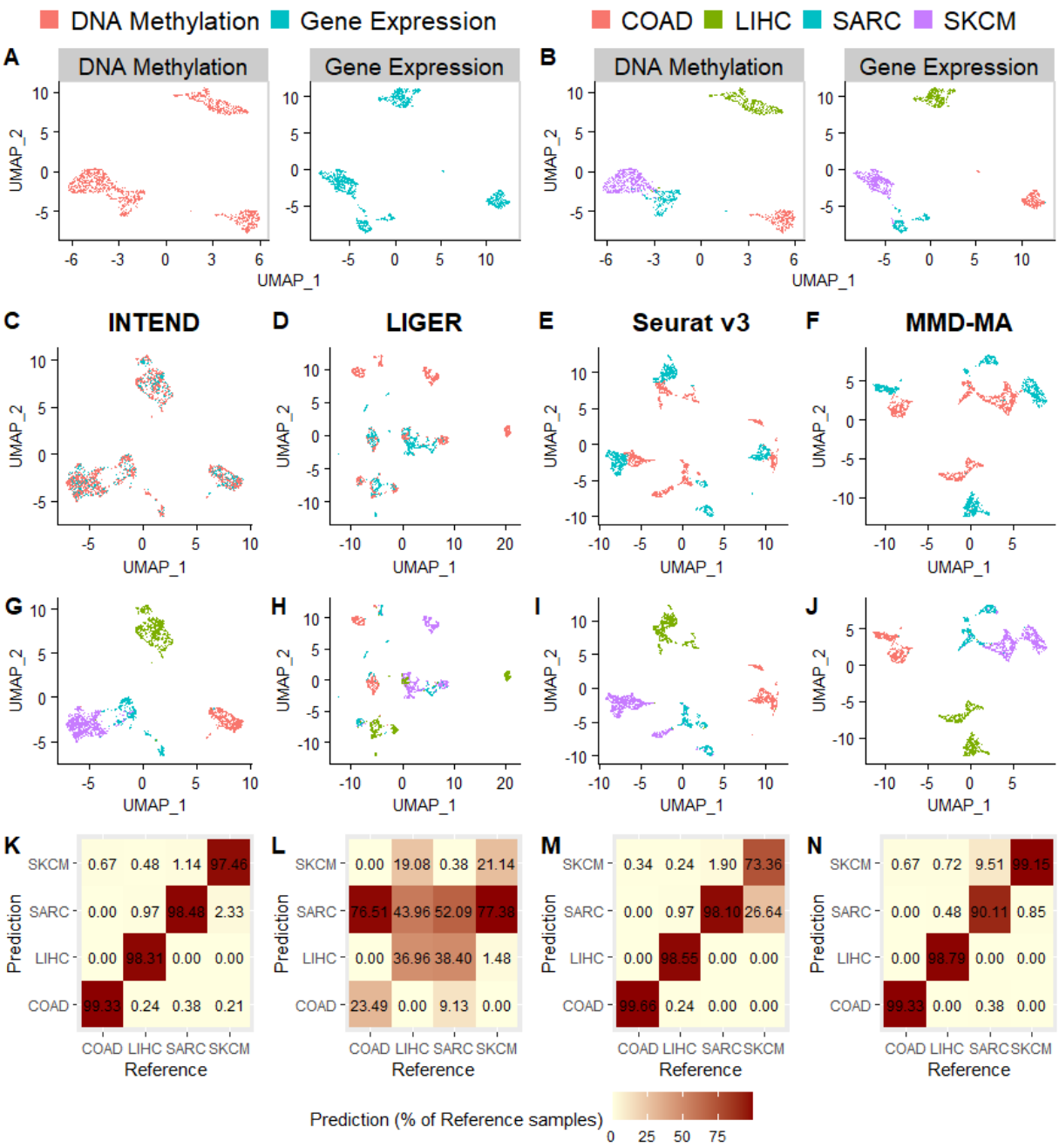


Figure 7. Results of joint integration of GE and DM samples of four cancer datasets: COAD, LIHC, SARC, and SKCM. (A-B) UMAP plots of the original data colored by omic (A) and by cancer type (B). (C-J) UMAP plots of the sample projections into the shared space by INTEND, LIGER, Seurat v3, and MMD-MA, colored by omic (C-F) and by cancer type (G-J). (K-N) Confusion matrices for the classification of the DM sample projections into cancer types based on majority vote among the five nearest GE samples in the shared space.

3.3 Using INTEND to identify subtypes in skin cutaneous melanoma

Clustering of single-omic cancer data is commonly used to identify subtypes. The quality of the clustering solution can be evaluated by the significance of separation in survival among subtypes. It has been observed that for certain cancer types, one omic may produce much better clustering than another. For example, Rappoport and Shamir (2018) benchmarked eight clustering algorithms on the TCGA SKCM data, and observed that GE profile clustering produced clusters with significant difference in survival in all algorithms, while in DM profile clustering only one algorithm showed such result. We hypothesized that in such cases, we could use INTEND to obtain GE predictions from the DM data, then jointly embed in the shared space the predictions and a set of GE profiles from the same cancer subtype, and achieve higher significance of separation in survival between clusters of the embedded predictions.

We used a dataset of 473 SKCM samples from TCGA that had both GE and DM profiles. We created 30 random partitions of this set into two equal disjoint groups, and for each partition, we used the first group's DM profiles and the second's GE profiles. We used INTEND to obtain a predicted GE matrix (P) from the DM samples and then embed P jointly with the GE profiles. Call the embedded P data EP. For the training phase of INTEND model, we used samples from all TCGA datasets listed in **Table 1** but excluded the SKCM dataset.

We first clustered separately the original partitioned DM and GE data. We performed each clustering task using k-means (see Methods) after selecting the 2000 features with the highest variance and normalizing the features to have zero mean and a standard deviation of one (as in Rappoport and Shamir (2018)). We ran the algorithm for k between 2 and 15, and selected the desired number of clusters using the "elbow method" (see Methods). We measured differential survival between clusters by computing the p-value for the log-rank test. We estimated the p-values using permutation tests (48). As we hypothesized, in most cases, the clustering of the

GE data obtained more significant differential survival between clusters than the clustering of the DM data, with the log-rank p-value of the first being lower in 27 of the 30 partitions.

Next, for each of the 30 partitions, we used INTEND's joint embedding of the DM and GE samples to classify the DM samples based on the k-means clustering of the GE samples. Each DM sample was assigned by majority voting (with ties broken at random) to the cluster most represented among the five GE embeddings closest to its matching EP representation in the shared space. In 23 of the 30 splits, clustering the DM samples using this method obtained more significant differential survival than using the k-means clustering of the DM samples. The average log-rank p-values for the clusterings for all 30 random splits were: 0.07 for the GE k-means clustering, 0.56 for the DM k-means clustering, and 0.21 for the integration-based DM clustering, as described above.

We further investigated one of the 23 partitions for which the integration-based DM clustering achieved more significant differential survival than the DM clustering. For that partition, the DM clustering resulted in two clusters with insignificant differential survival (p-value=0.978, **Figure 8A**), whereas the GE clustering resulted in two clusters with significant differential survival (p-value=0.018, **Figure 8B**). The integration-based DM clustering also obtained significant differential survival between clusters (p-value=0.036, **Figure 8C**).

Next, we tested whether the subtypes obtained by the integration-based DM clustering were biologically or clinically more similar to those obtained by the GE k-means clustering. We found that primary tumor and metastases samples were represented in each of the DM k-means clusters exactly in their portion of all DM samples (18.26% of primary tumor samples in both clusters). By contrast, when looking at the GE clusters, the primary tumor samples were overrepresented in one cluster and underrepresented in the other (28.21% of primary tumor samples in the first cluster, 5.94% in the second, 17.89% in all GE samples). We observed a similar pattern in the integration-based DM clustering: 23.77% of primary tumor samples in

one cluster and 11.34% in the other (and 18.26% in all DM samples). This example shows the potential of transferring biological information between GE and DM samples measured on different populations, using INTEND's integration.

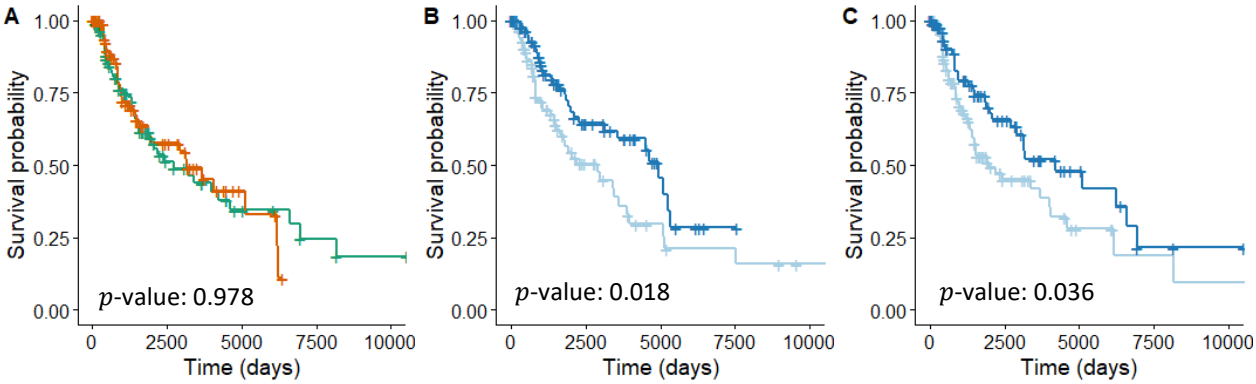


Figure 8. Kaplan-Meier plots of clusters of SKCM patients obtained using DM profiles, GE profiles, and their INTEND embeddings. (A) Plot for clusters of the original DM profiles. (B) Plot for clusters of the original GE profiles. (C) Plot for clusters of the DM profiles obtained by the integration-based clustering. See **Supplementary Figure 18A-E** for the UMAP plots and the clusters.

We also compared our results to iCluster, a widely used algorithm for multi-omic subtype identification (9). Since iCluster requires multi-omic measurements from each sample, in order to make a fair comparison, we used the entire multi-omic SKCM TCGA dataset, which comprises GE and DM profiles from 473 samples. We used the same feature selection and normalization as we used for the k-means clustering. To determine the lower dimension of the data in iCluster, we used the dimension with maximal deviance ratio as defined by the authors. We ran iCluster for dimensions between 1 and 14, corresponding to the number of clusters between 2 and 15. We also ran that same procedure with k-means and INTEND, on the full set of 473 samples. Specifically, we clustered the 473 GE profiles using k-means and then obtained a clustering of the DM profiles based on the GE clustering, by assigning each DM profile to the cluster most represented amongst the five GE embeddings closest to its DM embedding. It is important to note that in that INTEND did not use the correspondence information between

the GE and DM profiles, but only predicted the GE profiles from the DM profiles. Surprisingly, using INTEND's joint embedding of the DM and GE samples to classify the DM samples based on the k-means clustering of the GE samples, achieved a significantly better separation of survival between clusters compared to the multi-omic clustering provided by iCluster. The log-rank *p*-values for the clusterings were 0.39 for the DM k-means clustering, 0.0014 for the GE k-means clustering, 0.0062 for the integration-based DM clustering, and 0.14 for the iCluster multi-omic clustering. Therefore, our results suggest that our method outperforms iCluster in multi-omic subtype identification.

3.4 Joint analysis of lung adenocarcinoma datasets from different sources

Our next goal was to test the utility of INTEND in joint analysis of two datasets, one of DM profiles and one of GE profiles, coming from different sources. We used data from two studies of LUAD: GE of 172 tumor samples from Chen et al. (2020), and DM profiles of 477 samples from TCGA. The datasets were collected in different studies covering disjoint groups of LUAD patients.

3.4.1 Integration

For the training phase of the model, we used samples from all TCGA datasets listed in **Table 1** but excluded the LUAD dataset. The integration results are summarized in **Figure 9A-B.** As the two target datasets here are disjoint we cannot use FOSCTTM to evaluate their mixing in the embedding phase. As a sanity check, we considered for each sample its closest 32 neighbors (5% of the samples) in the shared space. We expected that if the local neighborhood of a sample is well mixed, the number of samples from each omic in the neighborhood would reflect the relative sizes of the target datasets. For each sample we measured the ratio between the observed and expected number of samples from the other omic in its neighborhood. If the omics are fully separated we would expect this ratio to be near zero, whereas for perfectly mixed samples we would expect it to be close to 1. The mean computed ratio for all samples in the

shared space was 1.003 ($SD = \pm 0.258$), and the IQR was $0.82 - 1.15$, indicating well-mixed samples across omics.

3.4.2 Correlations between methylation at specific sites and expression

Next, we wished to test if INTEND application on the two datasets can be used to reveal connections between specific distal DM sites and the regulation of GE in LUAD tumors, even though the GE profiles and DM profiles used here were collected from disjoint sets of patients. For this task, we extracted the estimated correlations between methylation levels at specific CpG sites and the expression levels of specified genes as follows.

We considered for every gene g, the methylation sites located within ±1Mb of g (including sites in g). There was a total of approximately 10.14 million such gene-site pairs, for which the expression and methylation levels were measured, covering 18,553 different genes. Recall that INTEND model was trained using proximal sites located only within ±10Kb from each gene, while here we explore mostly distal methylation sites. To estimate the correlation between the methylation level at site s and the expression level of gene g, we used INTEND projections to get matchings between GE and DM profiles from different patients. First, to match GE and DM profiles, we found the mutual nearest neighbors between the projections of all DM and GE samples in the shared space, using the *batchelor* R package (19). A pair of a GE profile e and a DM profile m was considered a match if the projection of m was among the k-nearest neighbors of the projection of e and vice versa (i.e. the projections of e and m are mutual k-nearest neighbors). For $k = 5$ we obtained 270 matches between GE and DM profiles (out of $172 \cdot 477 = 82{,}044$ possible matches). The matches provided an expression vector of length 270 for each gene g, and a corresponding vector of length 270 for each methylation site s, allowing the examination of the relationship between any gene and methylation site. Next, using the 270 matches, we computed the Pearson‘s correlation coefficient and tested the

statistical significance of the association between the expression level and the methylation level of each considered gene-site pair.

We wished to assess the validity of the estimated correlations, based on the created 270 matchings of GE and DM samples from the two LUAD datasets (from here on: "estimated correlations"). We compared the estimations to the correlations obtained from 477 pairs of GE and DM profiles measured from the same tissue, from the multi-omic LUAD TCGA dataset. For each of the ~10M gene-site pairs previously described, we also computed the correlation between the expression of the gene and the methylation level of the relevant site, based on the multi-omic TCGA dataset (from here on: "TCGA-observed correlations"). **Figure 9C** shows for each gene-site pair the estimated correlation versus the TCGA-observed correlation. Approximately 5.08% of the considered gene-site pairs were detected with significant correlation (p-value< 0.01), either positive or negative, according to both methods. For 95.63% of these significant pairs, the estimated correlation coefficient had the same sign as the TCGA-observed correlation. We also tested for each of the considered genes, the correlation between the estimated correlation and the TCGA-observed correlation, for all sites relevant for that gene. Out of the 18,553 considered genes, there was a significant positive correlation between the estimated and TCGA-observed correlations (p-value< 0.05) for 14,693 of the genes. The correlation between the estimated and TCGA-observed correlations was above 0.8 for 1,041 of the genes, and above 0.9 for 180 of them (**Figure 9D**). This demonstrates the potential of INTEND integration method to uncover connections between DNA methylation and the regulation of gene expression, both for proximal and distal methylation sites. Repeating the same procedure with the integration results of LIGER, Seurat, and MMD-MA for the target LUAD datasets gave inferior results **(Supplementary Table 3)**.

3.4.2.1 An in-depth look at the regulation of Thymidine Kinase 1

We chose to look in detail at the distal methylation sites of the gene Thymidine Kinase 1 (TK1). High expression of TK1 was recorded in many solid tumors, and was associated specifically with poor prognosis of patients with LUAD (49–51). We computed the correlation between the methylation levels in 964 sites within ±1Mb from TK1, and its expression level. The estimated correlations based on the matching of GE and DM profiles from INTEND projections were highly concordant with the correlations computed using the multi-omic TCGA dataset ($R^2 = 0.824$, **Figure 9E**).

Methylation patterns in enhancer regions are known to be altered in cancer and are closely linked to changes in expression of cancer-related genes (36). Therefore, we checked if strong expression-methylation correlations extracted from INTEND projections can indicate potential distal enhancer regions. We used the GeneHancer database of enhancers and their inferred target genes (52) for information on TK1 enhancers. There were eight enhancer regions supported by at least four GH sources, seven of them within a 100Kb range from TK1. **Figure 9F** shows the enhancer regions located ±100Kb from TK1, and the correlations between methylation and TK1 expression, for sites located in this range. 14 out of the 15 sites in this range with strong negative correlation (p-value$<$ 1e-5), are located in one of the documented enhancer regions. Note that all but two of them fall outside the ±10Kb used for the training phase.

Out of the 964 sites in 1Mb range from TK1, we investigated the ten sites with the strongest negative estimated correlations (full details in **Supplementary Table 4**). Eight of them are located in two of the enhancer regions shown in **Figure 9F** (seven of them in a short interval of less than 500 bases). The other two sites, cg11868461 and cg05110391, are located approximately 350Kb downstream and 400kB upstream the TSS, respectively. They were not in one of the regions marked by GeneHancer as TK1 enhancers. Nevertheless, both

cg11868461 and cg05110391 were identified as “enhancer probes” (not specifically related to TK1) by Mullen et al. (2020), using H3K27ac ChIP-seq data from normal and tumor lung tissue samples to identify lung-relevant enhancer regions.

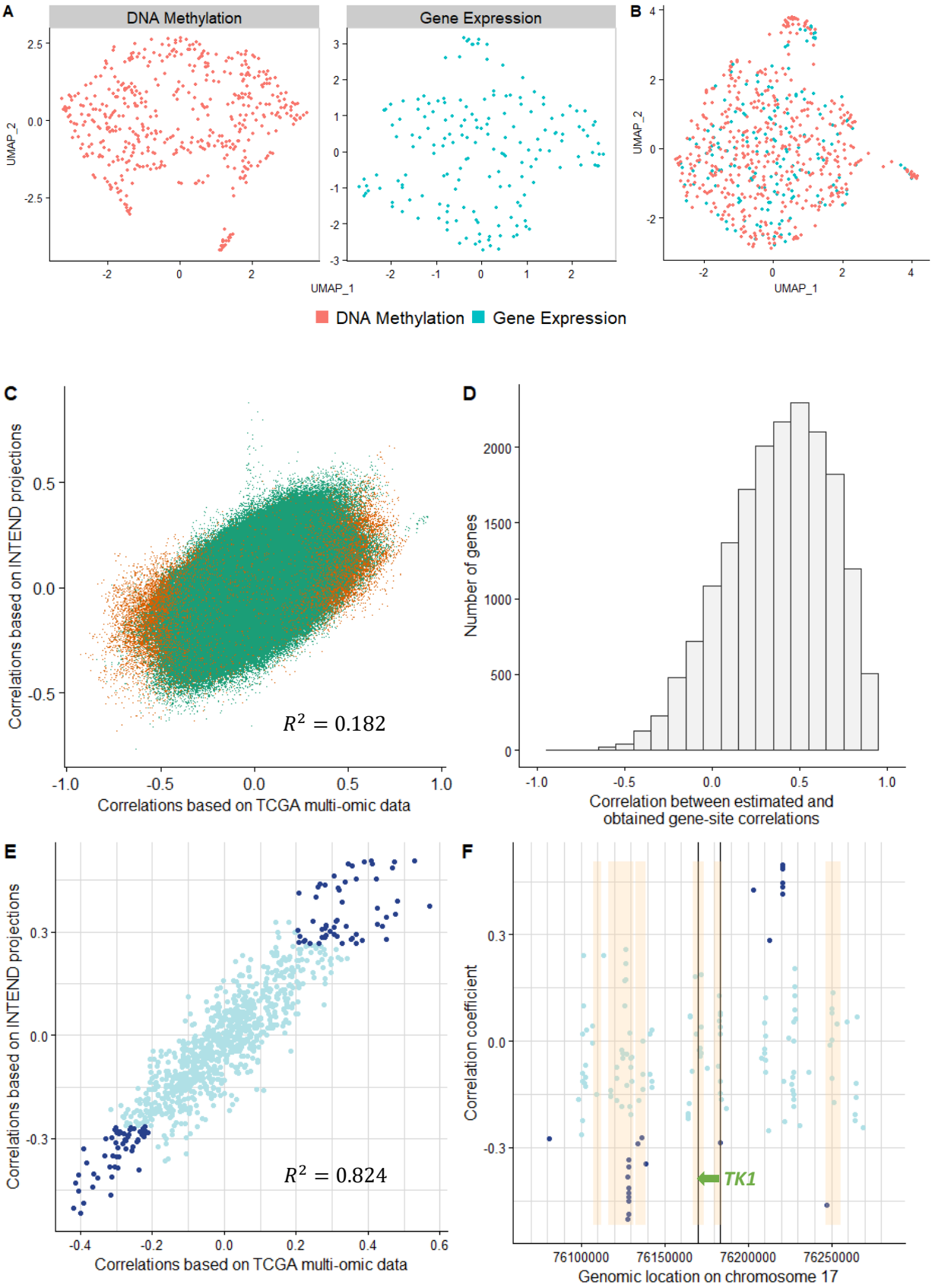

A
DNA Methylation
Gene Expression
UMAP_2
UMAP_1
B
UMAP_2
UMAP_1
DNA Methylation
Gene Expression
C
Correlations based on INTEND projections
$R^2 = 0.182$
Correlations based on TCGA multi-omic data
D
Number of genes
Correlation between estimated and obtained gene-site correlations
E
Correlations based on INTEND projections
$R^2 = 0.824$
Correlations based on TCGA multi-omic data
F
Correlation coefficient
TK1
Genomic location on chromosome 17

Figure 9. INTEND results on LUAD GE profiles from Chen et al. (2020) and DM profiles from TCGA. (A-B) UMAP plots of the original data (A) and of the projections into the shared space (B), colored by omic. (C) Scatterplot of the estimated correlations based on the matching of INTEND projections versus the observed correlations from the multi-omic TCGA dataset, for each of the considered 10.14 million gene-site pairs. The pairs for which the site is within 10Kb from the gene are colored in orange. These gene-site pairs were considered in INTEND training phase on the TCGA datasets (excluding LUAD). (D) Histogram of the correlation between the estimated and TCGA-observed gene-site correlations, per gene. (E) Correlation coefficients between TK1 expression and methylation levels, at 964 sites located ±1Mb from TK1. Y axis: correlations when TK1 expression is based on INTEND projections; x axis: correlations when both the GE and the paired DM data were taken from TCGA. Correlations with $\boldsymbol{p}$-value$< \mathbf{1}$e-5 based on both methods are colored in dark blue (F) Estimated correlation coefficients based on INTEND projections in sites located ±100Kb from TK1. The x axis shows their genomic location (build GRCh37/hg19). Correlations with $\boldsymbol{p}$-value$< \mathbf{1}$e-5 are colored in dark blue, TK1 location is marked by the green arrow. The highlighted yellow regions indicate enhancer regions supported by at least four GeneHancer sources.

4 Discussion

We presented the INTEND algorithm for integrating gene expression and DNA methylation from different datasets. We tested it on multiple multi-omic cancer datasets and compared it with extant multi-omic integration algorithms. INTEND showed significantly superior results on all tested datasets when integrating data from single and multiple cancer types, both in terms of FOSCTTM score and in classification to cancer types according to the integration results. We demonstrated the potential of INTEND to transfer biological information between GE and DM samples measured on non-overlapping populations of skin cutaneous melanoma patients. Clustering DM samples achieved higher significance of separation in survival between clusters when using the integration results of the DM and GE data, than using the original DM data only. In another typical use case, we tested INTEND in joint analysis of two lung adenocarcinoma datasets from different sources. Here INTEND demonstrated its potential to uncover connections between DNA methylation and the regulation of gene expression.

INTEND's novelty mainly resides in the incorporation of the prediction of a GE profile from a DM profile of a sample, into the MO/MD integration problem. Unlike algorithms such as LIGER and Seurat, which were developed mainly to solve the SO/MD problem and then were extended to solve the MO/MD problem, INTEND suggests another method to generate the correspondence information between features – a paramount part for the integration. INTEND presents a data-driven approach to generate a predicted GE matrix, thus effectively reducing the MO/MD problem of integrating GE and DM profiles to the simpler SO/MD problem of integrating multiple GE datasets. Importantly, the data necessary for the training phase of INTEND can represent different populations than the data used for the embedding phase. In all cases presented in this paper, the used training data originated from samples from other cancer types than represented in the target datasets for integration. It is important to note that the goal of INTEND is not to predict expression from methylation for individual genes, but rather to enable integrated analysis. In the embedding phase, INTEND uses prediction data for a selected set of genes. Although only a small portion of the genes is selected, the integrated analysis allows the examination of the relationship between any gene and methylation site, as we demonstrated in the lung cancer analysis.

INTEND has several limitations. First, the training phase requires multi-omic data measured on the same set of samples, which is not required for the other algorithms we tested. While the training data is not required to be from a similar population to the target data, it is necessary that the omics will be measured in the same method on the train and target datasets. Obtaining multi-omic measurements may be harder in several scenarios, e.g. single-cell multi-omic data. Due to the lack of appropriate single-cell training data, we applied INTEND only on bulk data, which may bias the comparison against single-cell integration methods. Further testing would be needed as such data emerges. Second, the final step in the embedding, applying CCA, may be less effective when the target datasets contain non-overlapping sample populations (e.g.

when one of the target datasets contains a group of samples from a cancer type which is not present in the second). Stuart et al. 2019 addressed this limitation of using CCA as a final step and introduced a method to overcome it, using the concept of mutual nearest neighbors to identify anchors between the target datasets.

Lastly, we note two possible directions of extending this work. The first is the integration of other pairs of omics, in addition to GE and DM, in a similar method. Here we used an established, simple biological observation, namely the relation between the state of proximal methylation sites to the gene's expression, to build a model and uncover the connections between GE and DM based on available multi-omic data. This concept may be extended to other pairs of omics with available data measuring both on the same set of samples. Another future research direction is the incorporation of methods from algorithms tackling the SO/MD integration problem, after the first step in INTEND's embedding phase, which results in the predicted GE matrix.

References

1. Chakraborty,S., Hosen,M.I., Ahmed,M. and Shekhar,H.U. (2018) Onco-Multi-OMICS Approach: A New Frontier in Cancer Research. *Biomed Res. Int.*, 10.1155/2018/9836256.

2. Efremova,M. and Teichmann,S.A. (2020) Computational methods for single-cell omics across modalities. *Nat. Methods*, 10.1038/s41592-019-0692-4.

3. Method of the Year 2019: Single-cell multimodal omics (2020) *Nat. Methods*, **17**, 1.

4. McLendon,R., Friedman,A., Bigner,D., Van Meir,E.G., Brat,D.J., Mastrogianakis,G.M., Olson,J.J., Mikkelsen,T., Lehman,N., Aldape,K., *et al.* (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 10.1038/nature07385.

5. Angermueller,C., Clark,S.J., Lee,H.J., Macaulay,I.C., Teng,M.J., Hu,T.X., Krueger,F., Smallwood,S.A., Ponting,C.P., Voet,T., *et al.* (2016) Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat. Methods*, 10.1038/nmeth.3728.

6. Clark,S.J., Argelaguet,R., Kapourani,C.A., Stubbs,T.M., Lee,H.J., Alda-Catalinas,C., Krueger,F., Sanguinetti,G., Kelsey,G., Marioni,J.C., *et al.* (2018) ScNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells e. *Nat. Commun.*, 10.1038/s41467-018-03149-4.

7. Argelaguet,R., Clark,S.J., Mohammed,H., Stapel,L.C., Krueger,C., Kapourani,C.A., Imaz-Rosshandler,I., Lohoff,T., Xiang,Y., Hanna,C.W., *et al.* (2019) Multi-omics profiling of mouse gastrulation at single-cell resolution. *Nature*, 10.1038/s41586-019-1825-8.

8. Singh,A., Shannon,C.P., Gautier,B., Rohart,F., Vacher,M., Tebbutt,S.J. and Cao,K.A.L. (2019) DIABLO: An integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics*, 10.1093/bioinformatics/bty1054.

9. Shen,R., Olshen,A.B. and Ladanyi,M. (2009) Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, 10.1093/bioinformatics/btp543.

10. Argelaguet,R., Velten,B., Arnol,D., Dietrich,S., Zenz,T., Marioni,J.C., Buettner,F., Huber,W. and Stegle,O. (2018) Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Mol. Syst. Biol.*, 10.15252/msb.20178124.

11. Argelaguet,R., Arnol,D., Bredikhin,D., Deloro,Y., Velten,B., Marioni,J.C. and Stegle,O.

(2020) MOFA+: A statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol.*, 10.1186/s13059-020-02015-1.

12. Yang,Z. and Michailidis,G. (2016) A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. *Bioinformatics*, 10.1093/bioinformatics/btv544.
13. Wang,B., Mezlini,A.M., Demir,F., Fiume,M., Tu,Z., Brudno,M., Haibe-Kains,B. and Goldenberg,A. (2014) Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods*, 10.1038/nmeth.2810.
14. Rappoport,N. and Shamir,R. (2019) NEMO: Cancer subtyping by integration of partial multi-omic data. *Bioinformatics*, 10.1093/bioinformatics/btz058.
15. Rappoport,N. and Shamir,R. (2018) Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic Acids Res.*, **46**, 10546–10562.
16. Rappoport,N., Safra,R. and Shamir,R. (2020) MONET: Multi-omic module discovery by omic selection. *PLoS Comput. Biol.*, 10.1371/journal.pcbi.1008182.
17. Jin,S., Zhang,L. and Nie,Q. (2020) scAI: an unsupervised approach for the integrative analysis of parallel single-cell transcriptomic and epigenomic profiles. *Genome Biol.*, 10.1186/s13059-020-1932-8.
18. Luecken,M.D., Büttner,M., Chaichoompu,K., Danese,A., Interlandi,M., Mueller,M.F., Strobl,D.C., Zappia,L., Dugas,M., Colomé-Tatché,M., *et al.* (2020) Benchmarking atlas-level data integration in single-cell genomics. *bioRxiv*, 10.1101/2020.05.22.111161.
19. Haghverdi,L., Lun,A.T.L., Morgan,M.D. and Marioni,J.C. (2018) Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.*, 10.1038/nbt.4091.
20. Stuart,T., Butler,A., Hoffman,P., Hafemeister,C., Papalexi,E., Mauck,W.M., Hao,Y., Stoeckius,M., Smibert,P. and Satija,R. (2019) Comprehensive Integration of Single-Cell Data. *Cell*, **177**, 1888-1902.e21.
21. Lopez,R., Regier,J., Cole,M.B., Jordan,M.I. and Yosef,N. (2018) Deep generative modeling for single-cell transcriptomics. *Nat. Methods*, 10.1038/s41592-018-0229-2.
22. Hie,B., Bryson,B. and Berger,B. (2019) Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat. Biotechnol.*, 10.1038/s41587-019-0113-3.
23. Welch,J.D., Kozareva,V., Ferreira,A., Vanderburg,C., Martin,C. and Macosko,E.Z. (2019) Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain

Cell Identity. *Cell*, **177**, 1873-1887.e17.

24. Barkas,N., Petukhov,V., Nikolaeva,D., Lozinsky,Y., Demharter,S., Khodosevich,K. and Kharchenko,P. V. (2019) Joint analysis of heterogeneous single-cell RNA-seq dataset collections. *Nat. Methods*, 10.1038/s41592-019-0466-z.

25. Korsunsky,I., Millard,N., Fan,J., Slowikowski,K., Zhang,F., Wei,K., Baglaenko,Y., Brenner,M., Loh,P. ru and Raychaudhuri,S. (2019) Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods*, 10.1038/s41592-019-0619-0.

26. Amodio,M. and Krishnaswamy,S. (2018) MAGAN: Aligning biological manifolds. In *35th International Conference on Machine Learning, ICML 2018*.

27. Liu,J., Huang,Y., Singh,R., Vert,J.P. and Noble,W.S. (2019) Jointly embedding multiple single-cell omics measurements. *Leibniz Int. Proc. Informatics, LIPIcs*, **143**, 1–13.

28. Wang,C. and Mahadevan,S. (2008) Manifold Alignment without Correspondence. *Ijcai.Org*.

29. Tost,J. (2010) DNA methylation: An introduction to the biology and the disease-associated changes of a promising biomarker. *Mol. Biotechnol.*, 10.1007/s12033-009-9216-2.

30. Deaton,A.M. and Bird,A. (2011) CpG islands and the regulation of transcription. *Genes Dev.*, 10.1101/gad.2037511.

31. Jjingo,D., Conley,A.B., Yi,S. V., Lunyak,V. V. and King Jordan,I. (2012) On the presence and role of human gene-body DNA methylation. *Oncotarget*, 10.18632/oncotarget.497.

32. Moarii,M., Boeva,V., Vert,J.P. and Reyal,F. (2015) Changes in correlation between promoter methylation and gene expression in cancer. *BMC Genomics*, 10.1186/s12864-015-1994-2.

33. Bhasin,J.M.M., Lee,B.H.H., Matkin,L., Taylor,M.G.G., Hu,B., Xu,Y., Magi-Galluzzi,C., Klein,E.A.A. and Ting,A.H.H. (2015) Methylome-wide Sequencing Detects DNA Hypermethylation Distinguishing Indolent from Aggressive Prostate Cancer. *Cell Rep.*, 10.1016/j.celrep.2015.10.078.

34. Mo,A., Mukamel,E.A., Davis,F.P., Luo,C., Henry,G.L., Picard,S., Urich,M.A., Nery,J.R., Sejnowski,T.J., Lister,R., *et al.* (2015) Epigenomic Signatures of Neuronal Diversity in the Mammalian Brain. *Neuron*, 10.1016/j.neuron.2015.05.018.

35. Pliner,H.A., Packer,J.S., McFaline-Figueroa,J.L., Cusanovich,D.A., Daza,R.M., Aghamirzaie,D., Srivatsan,S., Qiu,X., Jackson,D., Minkina,A., *et al.* (2018) Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. *Mol. Cell*, 10.1016/j.molcel.2018.06.044.

36. Aran,D., Sabato,S. and Hellman,A. (2013) DNA methylation of distal regulatory sites characterizes dysregulation of cancer genes. *Genome Biol.*, 10.1186/gb-2013-14-3-r21.

37. Tibshirani,R. (1996) Regression Shrinkage and Selection Via the Lasso. *J. R. Stat. Soc. Ser. B*, 10.1111/j.2517-6161.1996.tb02080.x.

38. Friedman,J., Hastie,T. and Tibshirani,R. (2010) Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.*, 10.18637/jss.v033.i01.

39. Hotelling,H. (1936) Relations Between Two Sets of Variates. *Biometrika*, 10.2307/2333955.

40. Witten,D.M. and Tibshirani,R.J. (2009) Extensions of sparse canonical correlation analysis with applications to genomic data. *Stat. Appl. Genet. Mol. Biol.*, 10.2202/1544-6115.1470.

41. Butler,A., Hoffman,P., Smibert,P., Papalexi,E. and Satija,R. (2018) Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.*, 10.1038/nbt.4096.

42. Wei,L., Jin,Z., Yang,S., Xu,Y., Zhu,Y. and Ji,Y. (2018) TCGA-assembler 2: Software pipeline for retrieval and processing of TCGA/CPTAC data. *Bioinformatics*, 10.1093/bioinformatics/btx812.

43. Zhu,Y., Qiu,P. and Ji,Y. (2014) TCGA-assembler: Open-source software for retrieving and processing TCGA data. *Nat. Methods*, 10.1038/nmeth.2956.

44. Li,B. and Dewey,C.N. (2011) RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 10.1186/1471-2105-12-323.

45. Chen,J., Yang,H., Teo,A.S.M., Amer,L.B., Sherbaf,F.G., Tan,C.Q., Alvarez,J.J.S., Lu,B., Lim,J.Q., Takano,A., *et al.* (2020) Genomic landscape of lung adenocarcinoma in East Asians. *Nat. Genet.*, 10.1038/s41588-019-0569-6.

46. Hartigan,J.A. and Wong,M.A. (1979) Algorithm AS 136: A K-Means Clustering Algorithm. *Appl. Stat.*, 10.2307/2346830.

47. McInnes,L., Healy,J. and Melville,J. (2018) UMAP: Uniform manifold approximation

and projection for dimension reduction. *arXiv*.

48. Rappoport,N. and Shamir,R. (2019) Inaccuracy of the log-rank approximation in cancer data analysis. *Mol. Syst. Biol.*, **15**, 2017–2019.
49. Malvi,P., Janostiak,R., Nagarajan,A., Cai,G. and Wajapeyee,N. (2019) Loss of thymidine kinase 1 inhibits lung cancer growth and metastatic attributes by reducing GDF15 expression. *PLoS Genet.*, 10.1371/journal.pgen.1008439.
50. Jagarlamudi,K.K. and Shaw,M. (2018) Thymidine kinase 1 as a tumor biomarker: Technical advances offer new potential to an old biomarker. *Biomark. Med.*, 10.2217/bmm-2018-0157.
51. He,E., Xu,X.H., Guan,H., Chen,Y., Chen,Z.H., Pan,Z.L., Tang,L.L., Hu,G.Z., Li,Y., Zhang,M., *et al.* (2010) Thymidine kinase 1 is a potential marker for prognosis and monitoring the response to treatment of patients with breast, lung, and esophageal cancer and non-Hodgkin's lymphoma. *Nucleosides, Nucleotides and Nucleic Acids*, 10.1080/15257771003738535.
52. Fishilevich,S., Nudel,R., Rappaport,N., Hadar,R., Plaschkes,I., Iny Stein,T., Rosen,N., Kohn,A., Twik,M., Safran,M., *et al.* (2017) GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database (Oxford).*, 10.1093/database/bax028.
53. Mullen,D.J., Yan,C., Kang,D.S., Zhou,B., Borok,Z., Marconett,C.N., Farnham,P.J., Offringa,I.A. and Rhie,S.K. (2020) TENET 2.0: Identification of key transcriptional regulators and enhancers in lung adenocarcinoma. *PLoS Genet.*, 10.1371/journal.pgen.1009023.
54. Church,D.M., Schneider,V.A., Graves,T., Auger,K., Cunningham,F., Bouk,N., Chen,H.C., Agarwala,R., McLaren,W.M., Ritchie,G.R.S., *et al.* (2011) Modernizing reference genome assemblies. *PLoS Biol.*, 10.1371/journal.pbio.1001091.
55. Haeussler,M., Zweig,A.S., Tyner,C., Speir,M.L., Rosenbloom,K.R., Raney,B.J., Lee,C.M., Lee,B.T., Hinrichs,A.S., Gonzalez,J.N., *et al.* (2019) The UCSC Genome Browser database: 2019 update. *Nucleic Acids Res.*, 10.1093/nar/gky1095.

Supplementary Information

Data Collection

The genes' coordinates used in the training phase were taken from the GRCh37/hg19 assembly (54). The data was downloaded from the UCSC Genome Browser (55) 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@ussd-ftp.illumina.com/downloads/ProductFiles/HumanMethylation450/HumanMethylation450_15017482_v1-2.csv).

The TCGA data was downloaded using the TCGA-Assembler software (42, 43). 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 (45) 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 Ilumina (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 (30). The final gene-level summary was minus the average methylation signal in those regions.

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.

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:

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

Differentiating L with respect to u and v gives:

$$\frac{\delta L}{\delta u} = Kv - \lambda_1 u = 0 \rightarrow \ Kv = \lambda_1 u \tag{2}$$

$$\frac{\delta L}{\delta v} = K^T u - \lambda_2 v = 0 \rightarrow K^T u = \lambda_2 v \tag{3}$$

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

$$\lambda_1 = u^T K v = v^T K^T u = \lambda_2 \tag{4}$$

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 $Cor(Xu_i, Xu_j) = Cor(Yv_i, Yv_j) = 0$ 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 $S = [U^T \quad V^T]$, 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)

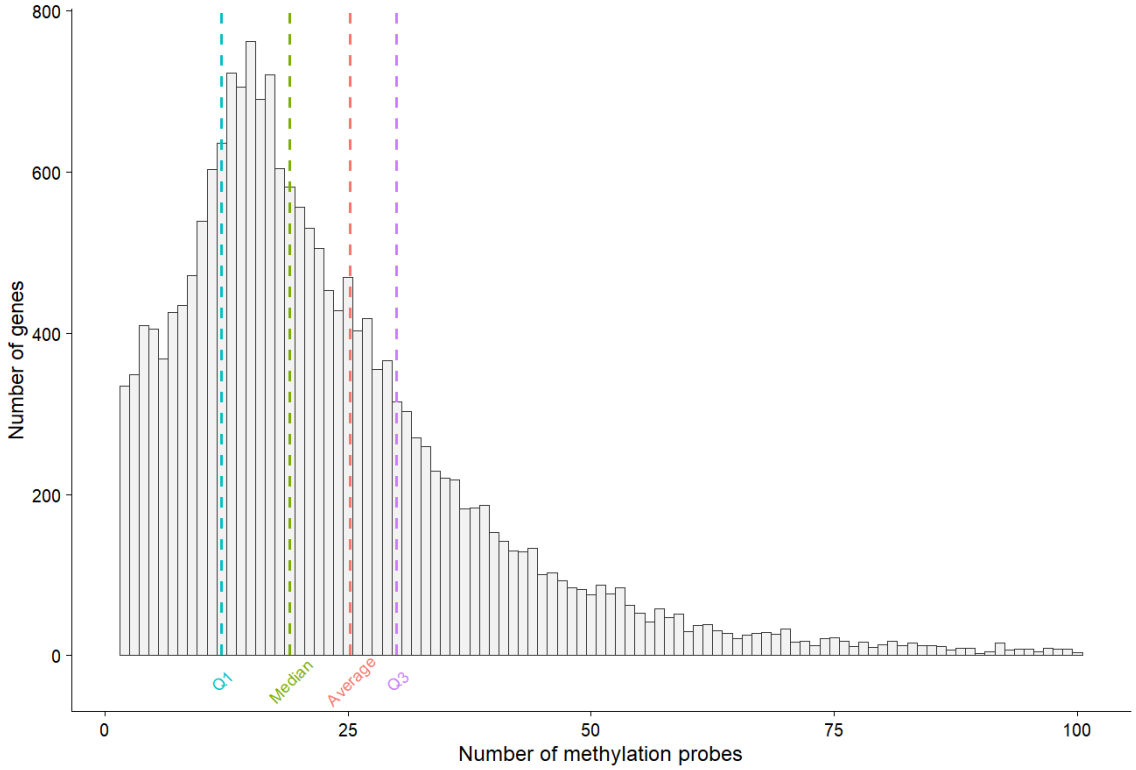
Supplementary Table 3. Comparison of correlation extraction from LUAD dataset integration results. The procedure described in section 3.4.2 was repeated with the integration results of INTEND, LIGER, Seurat and MMD-MA. The number of mutual nearest neighbors used was 270, 142, 61, and 231, respectively. 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 $\boldsymbol{R^2}$ 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

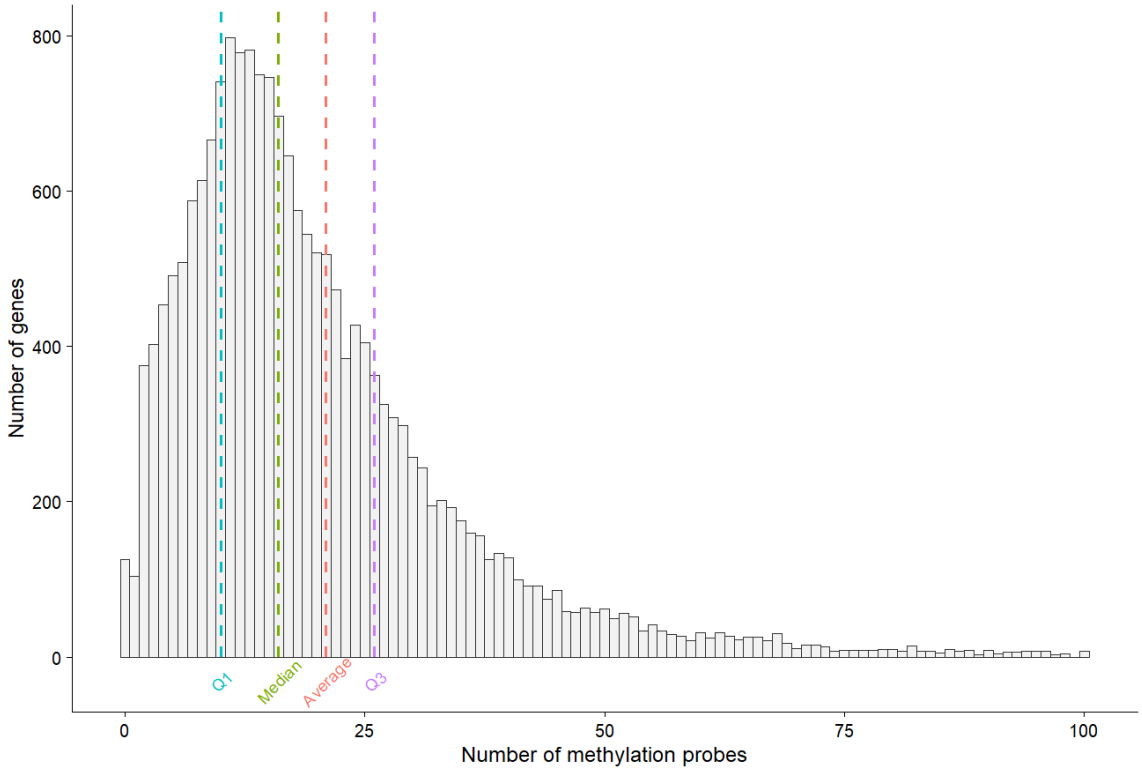
Methylation site	Location on chromosome 17 (build GRCh37/hg19)	Correlation coefficient estimation
cg11868461	75830800	-0.5181234
cg06643271	76128170	-0.5016554
cg24988684	76128556	-0.4887382
cg10460946	76247467	-0.4631925
cg11493223	76128522	-0.4516062
cg02911077	76128621	-0.4396759
cg18901278	76128531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

Supplementary Figures



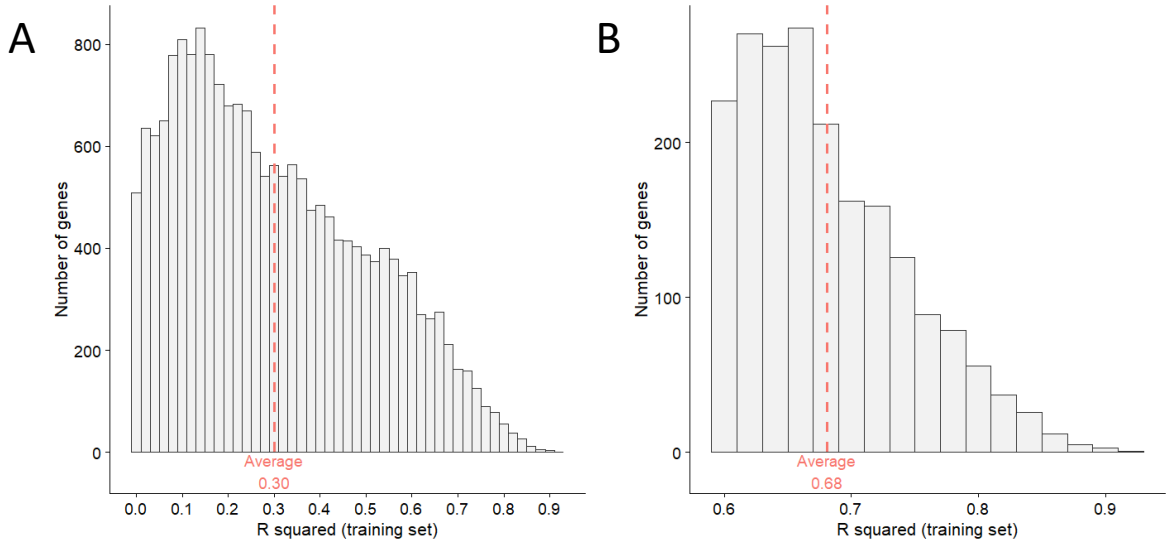
Supplementary Figure 1

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



Supplementary Figure 2

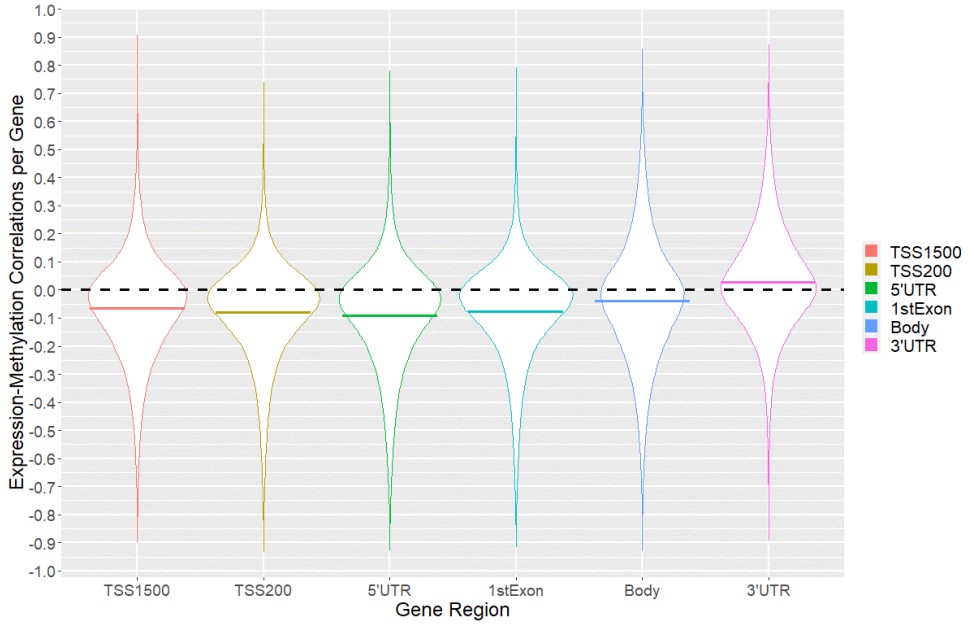
Histogram of the number of methylation sites per gene in the model after Lasso shrinkage on the TCGA data. The model was trained on ten cancer subtypes data from TCGA: AML, BLCA, COAD, LGG, LIHC, PAAD, PRAD, SARC, SKCM, and THCA. Average: 20.93, median: 16, interquantile range (IQR): 10-26. The maximum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



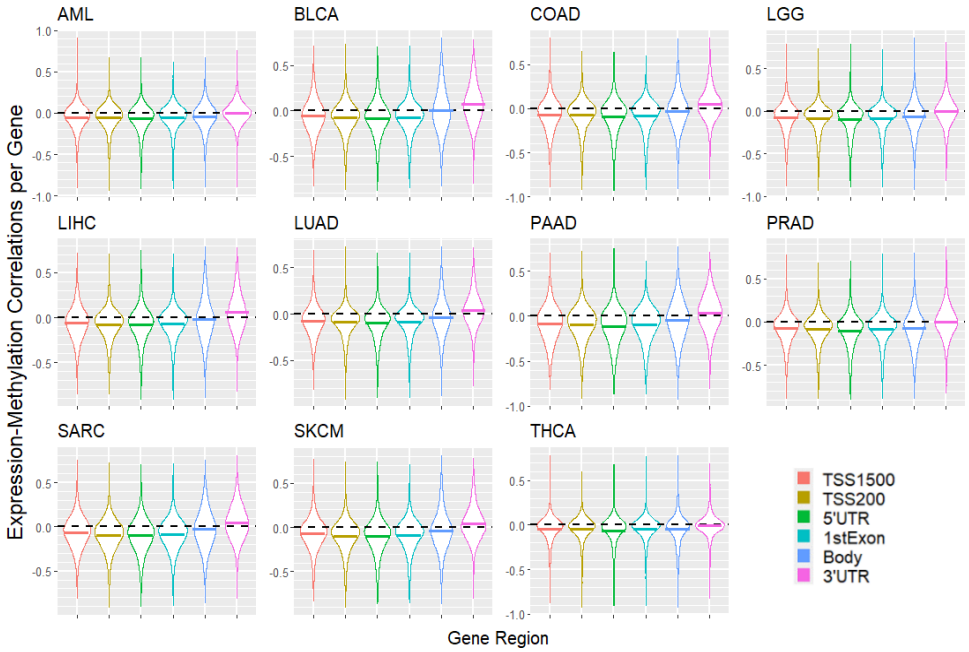
Supplementary Figure 3

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.

A

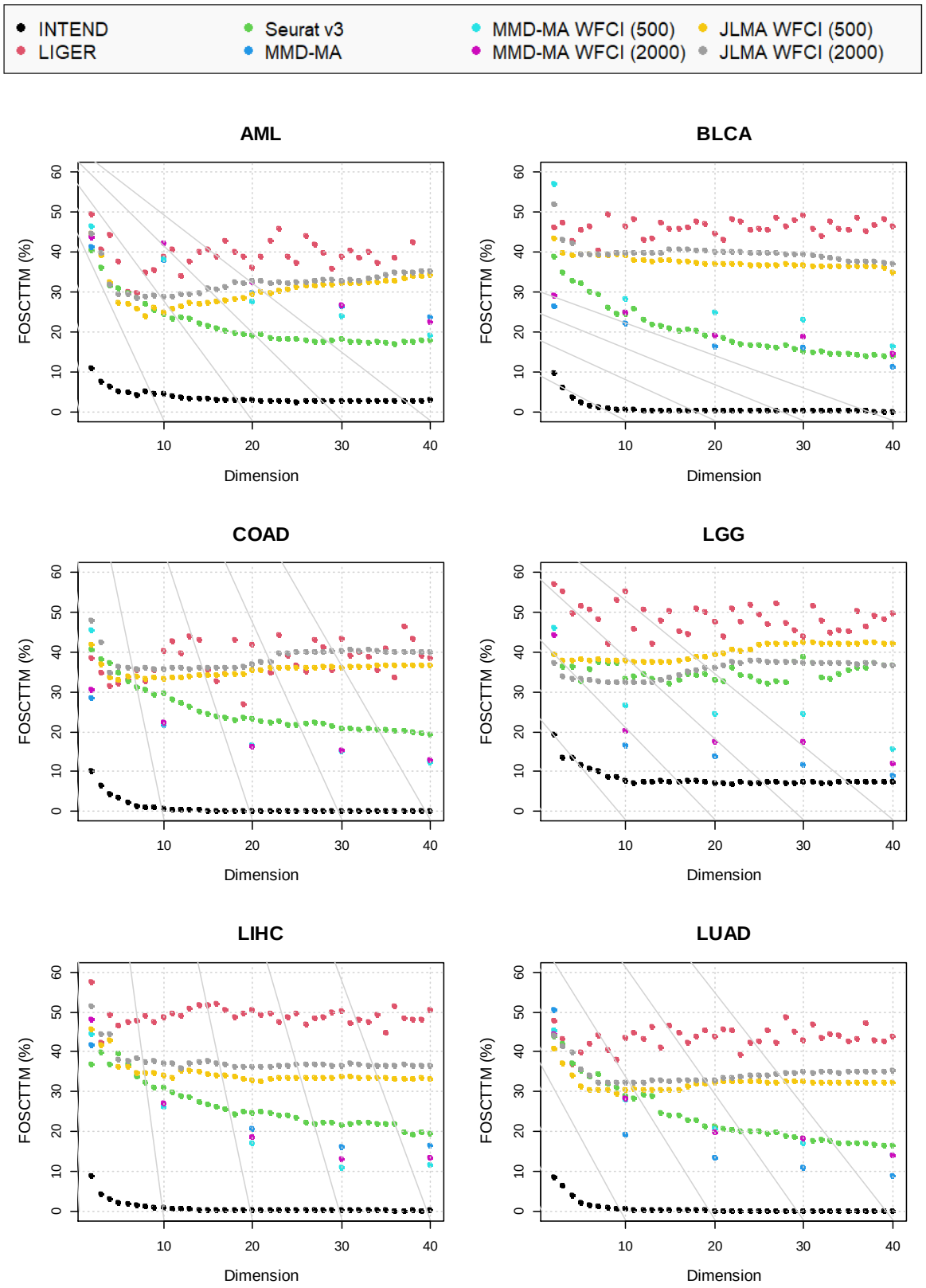


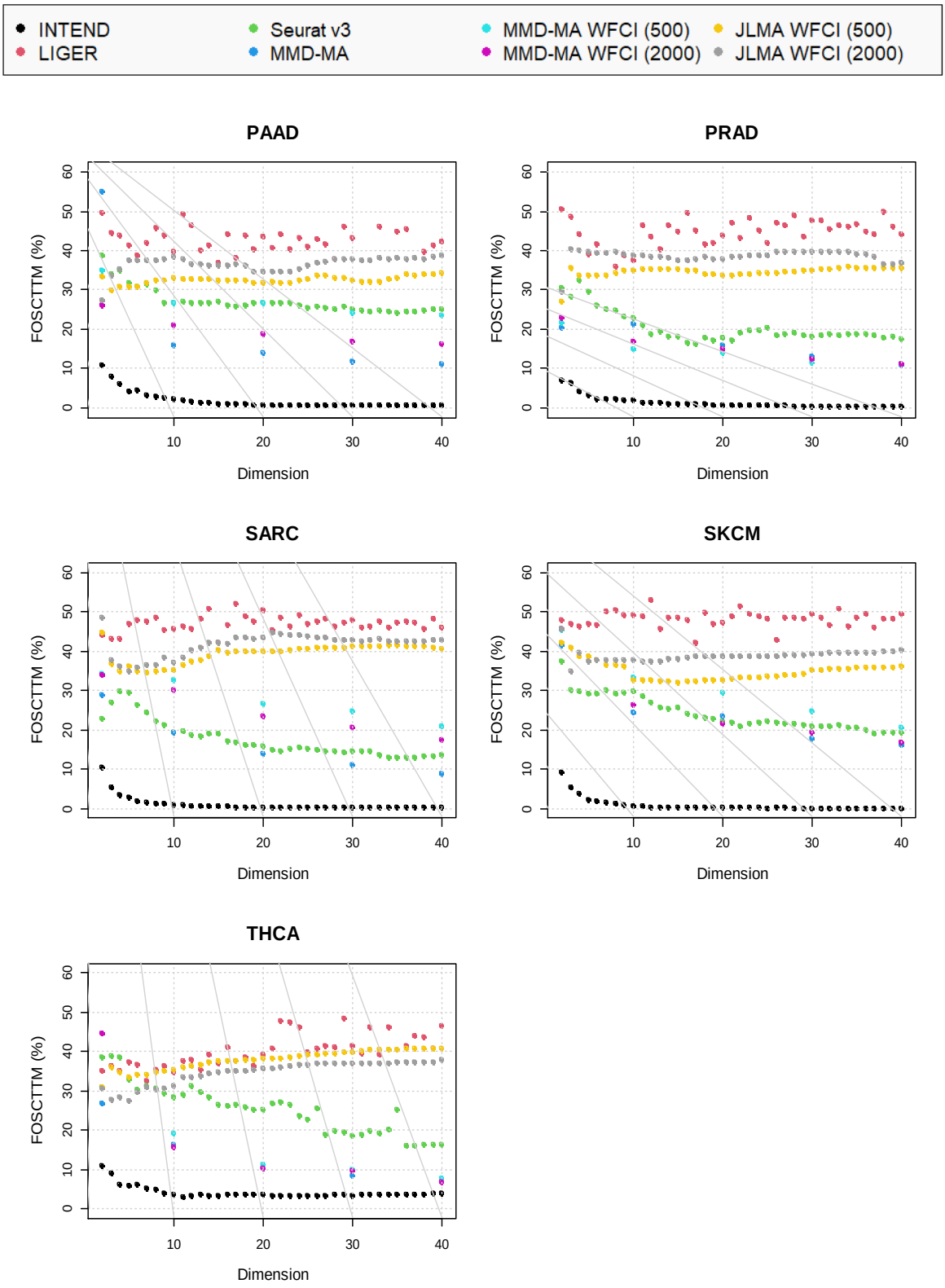
B



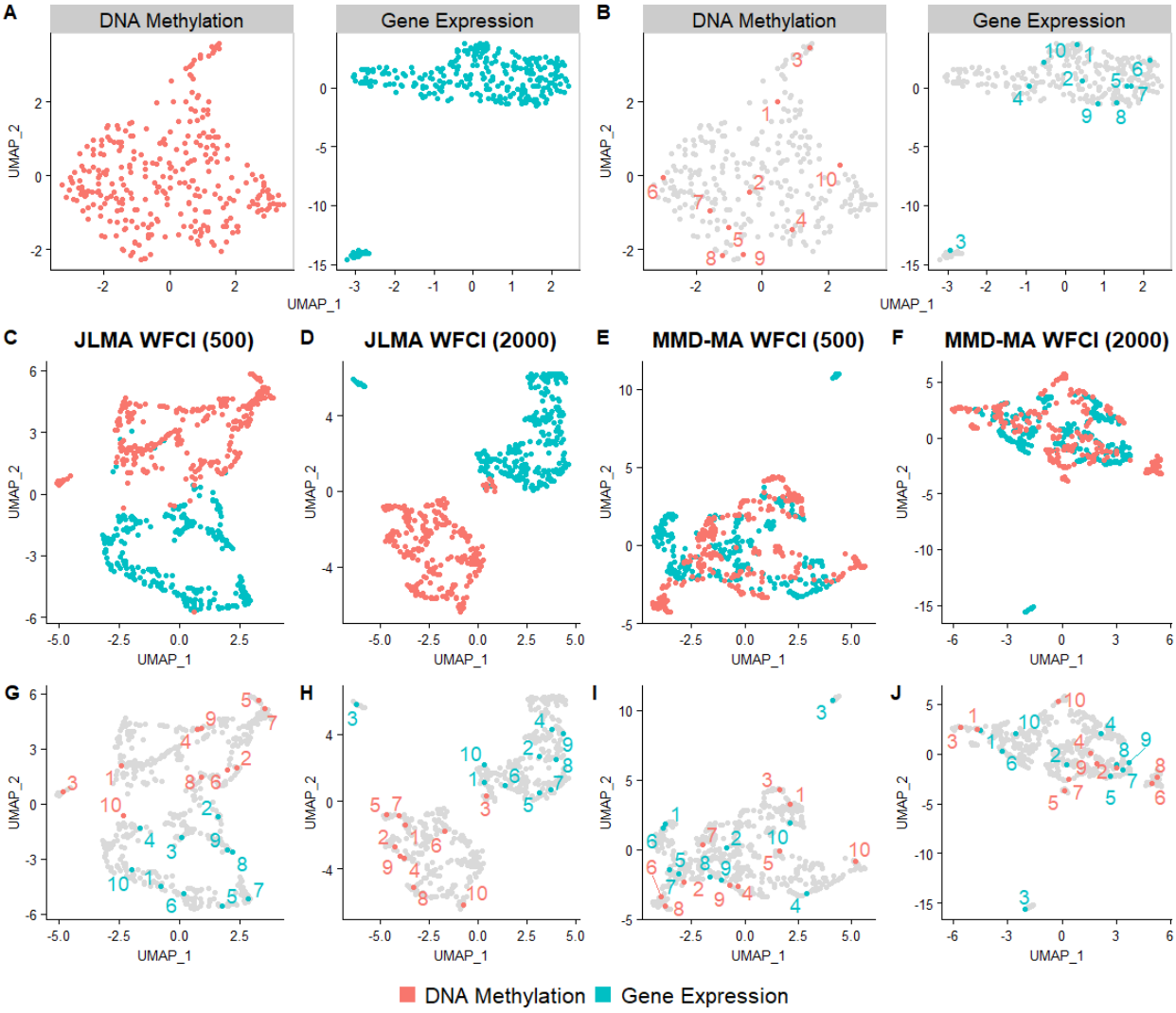
Supplementary Figure 4: Correlation between gene expression and DNA methylation levels in different genomic regions

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.


INTEND
LIGER
Seurat v3
MMD-MA
MMD-MA WFCI (500)
MMD-MA WFCI (2000)
JLMA WFCI (500)
JLMA WFCI (2000)
AML
BLCA
COAD
LGG
LIHC
LUAD
FOSCTTM (%)
Dimension
0
10
20
30
40
50
60



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 $\boldsymbol{d = 2, 10, 20, 30, 40}$, due to the long runtime of the algorithm.

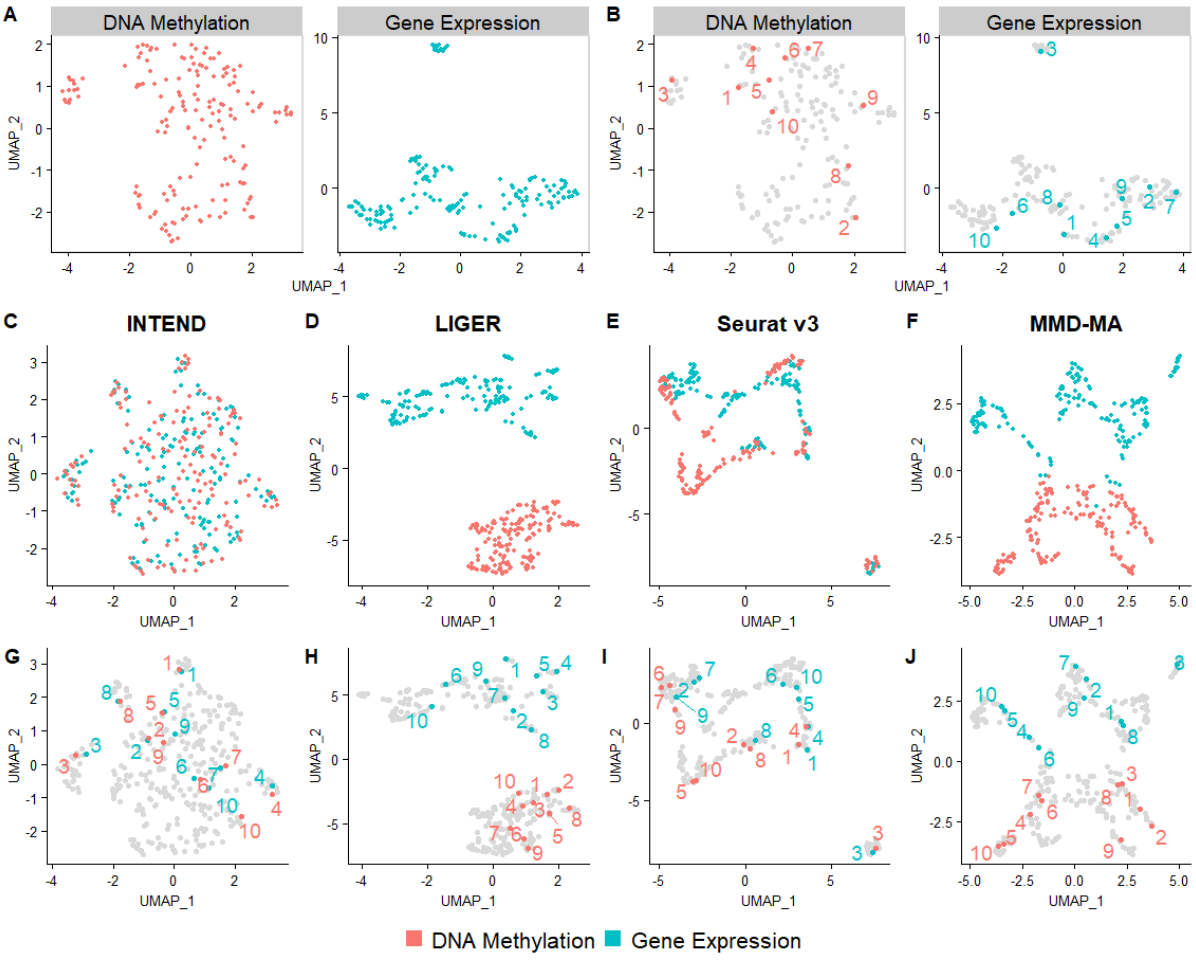


Supplementary Figure 6. The integration of gene expression and DNA methylation samples from the COAD dataset – results for JLMA and MMD-MA algorithms

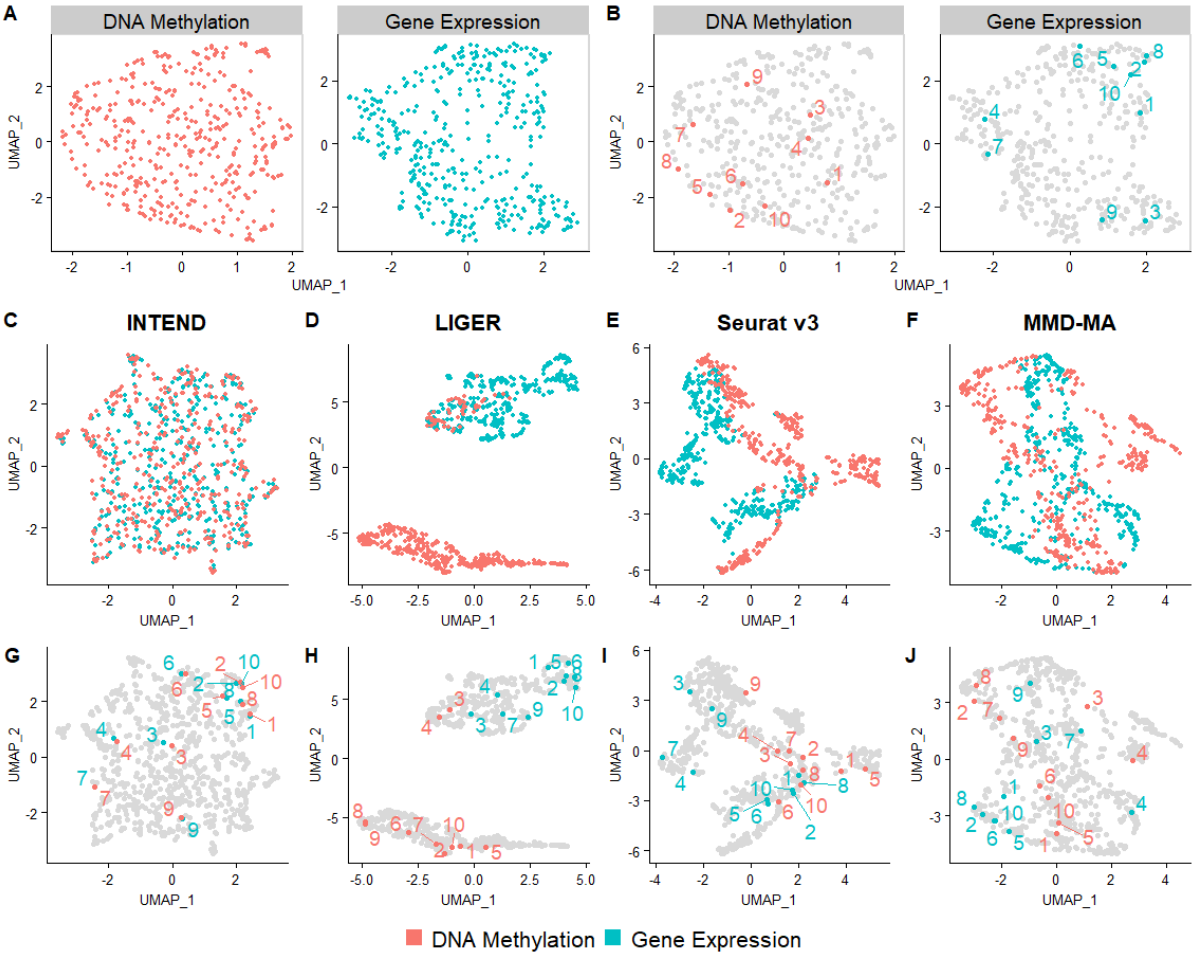
(A) UMAP plots of the original data colored by omic.
(B) UMAP plots of the original data. To appreciate concordance between omics, ten samples were randomly chosen, and their matching points in both omics are labeled and colored by omic.
(C-J) UMAP plots of the samples after they were projected to a 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.

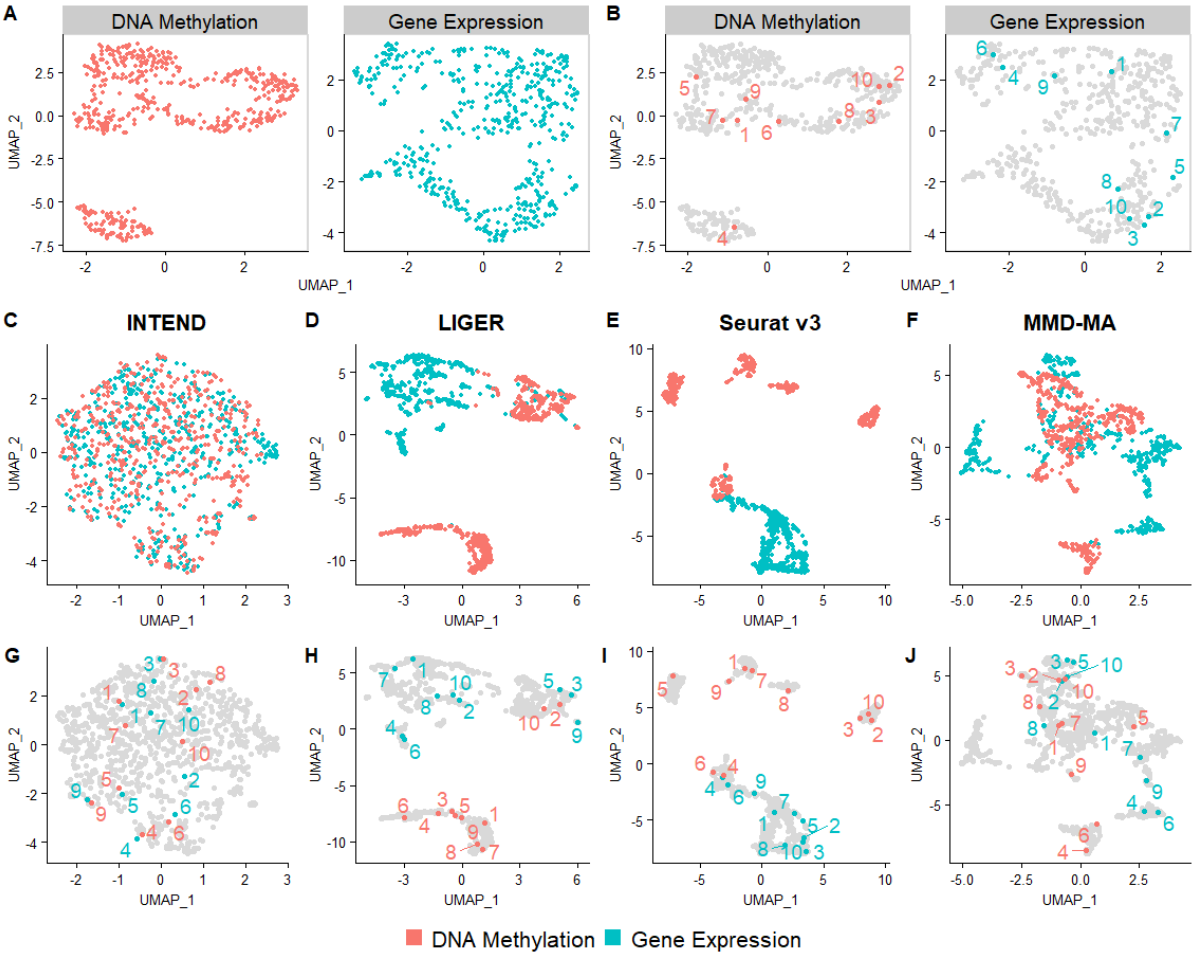
AML



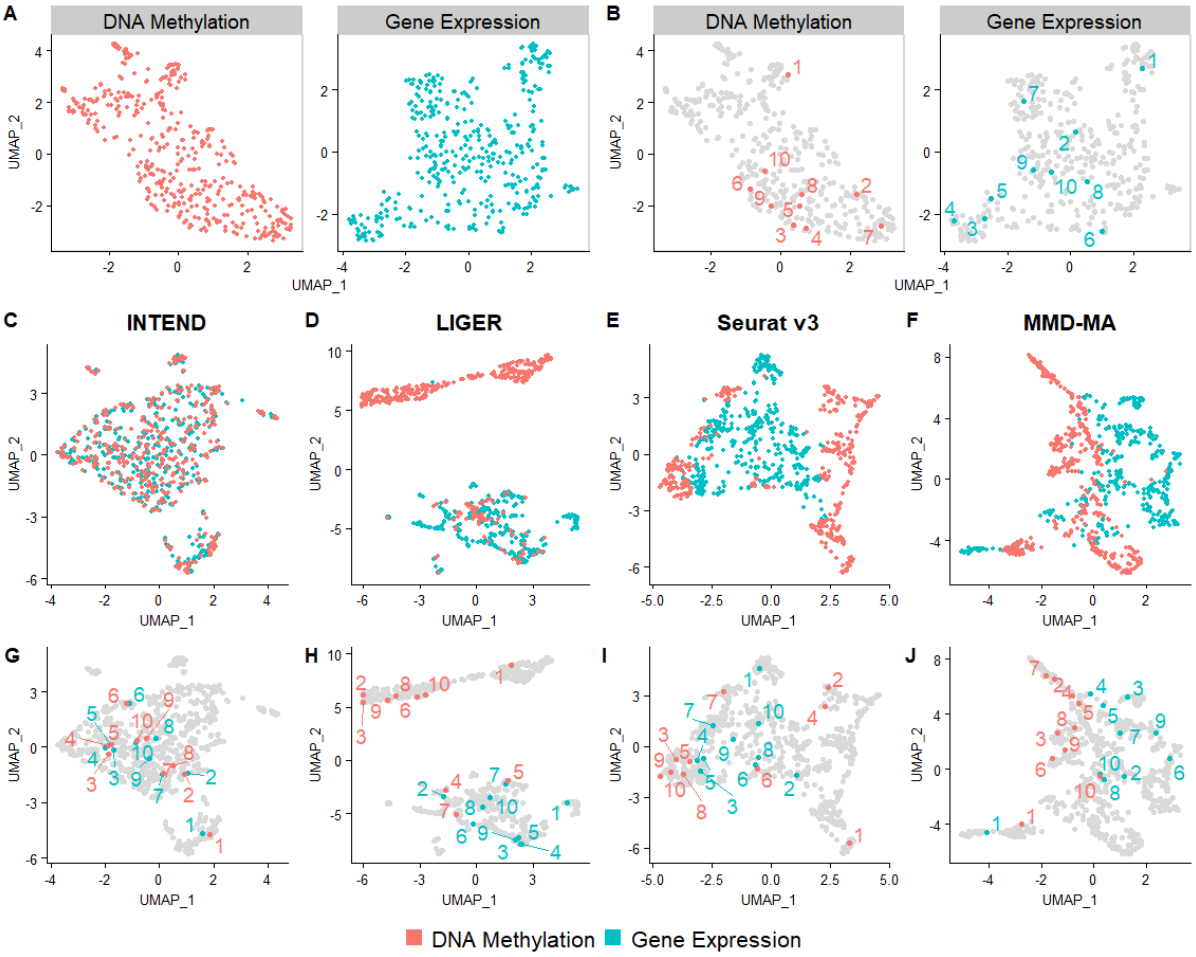
BLCA


A
DNA Methylation
Gene Expression
B
DNA Methylation
Gene Expression
UMAP_1
UMAP_2
C
INTEND
D
LIGER
E
Seurat v3
F
MMD-MA
G
H
I
J
DNA Methylation
Gene Expression

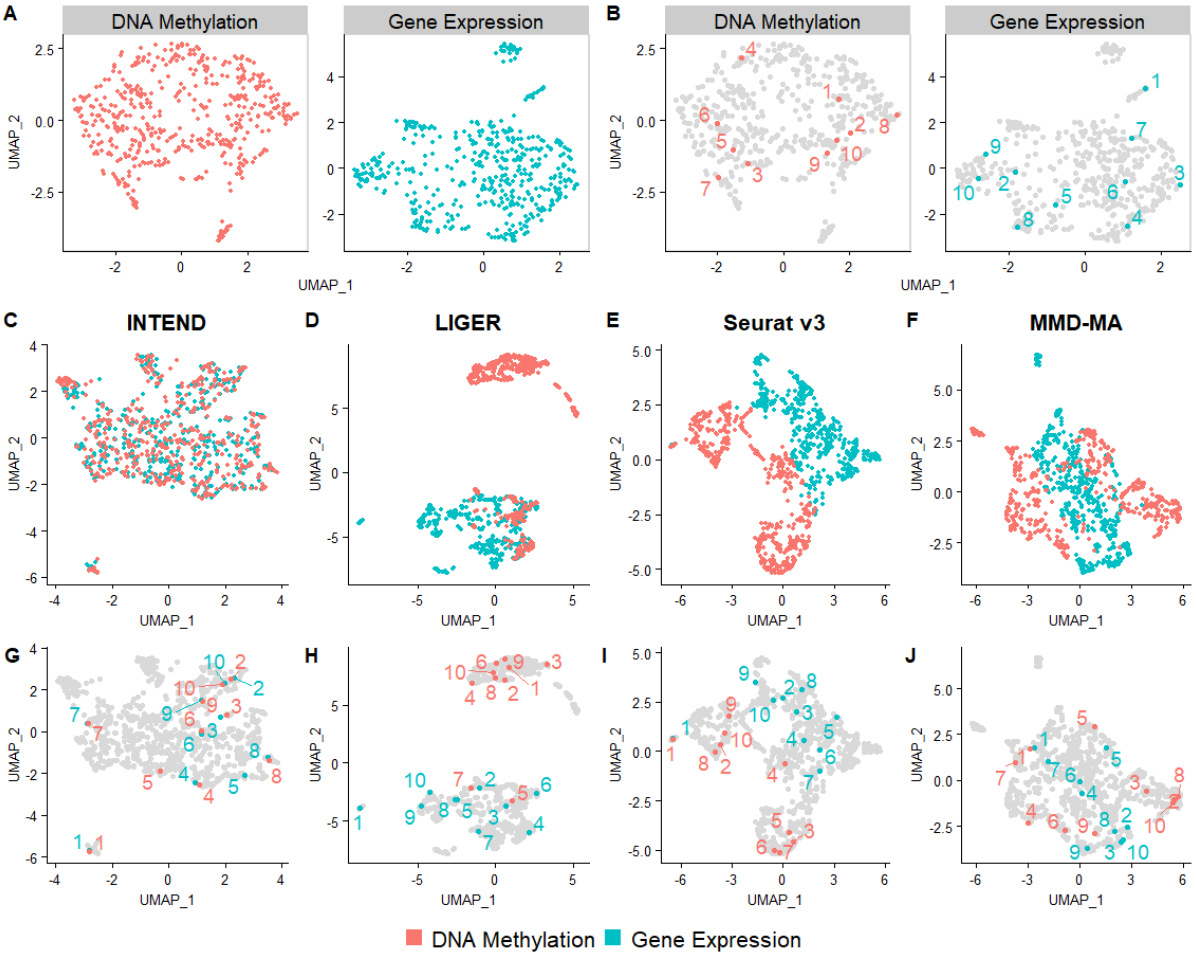
LGG


A
DNA Methylation
Gene Expression
B
DNA Methylation
Gene Expression
C
INTEND
D
LIGER
E
Seurat v3
F
MMD-MA
G
H
I
J
UMAP_1
UMAP_2
DNA Methylation
Gene Expression

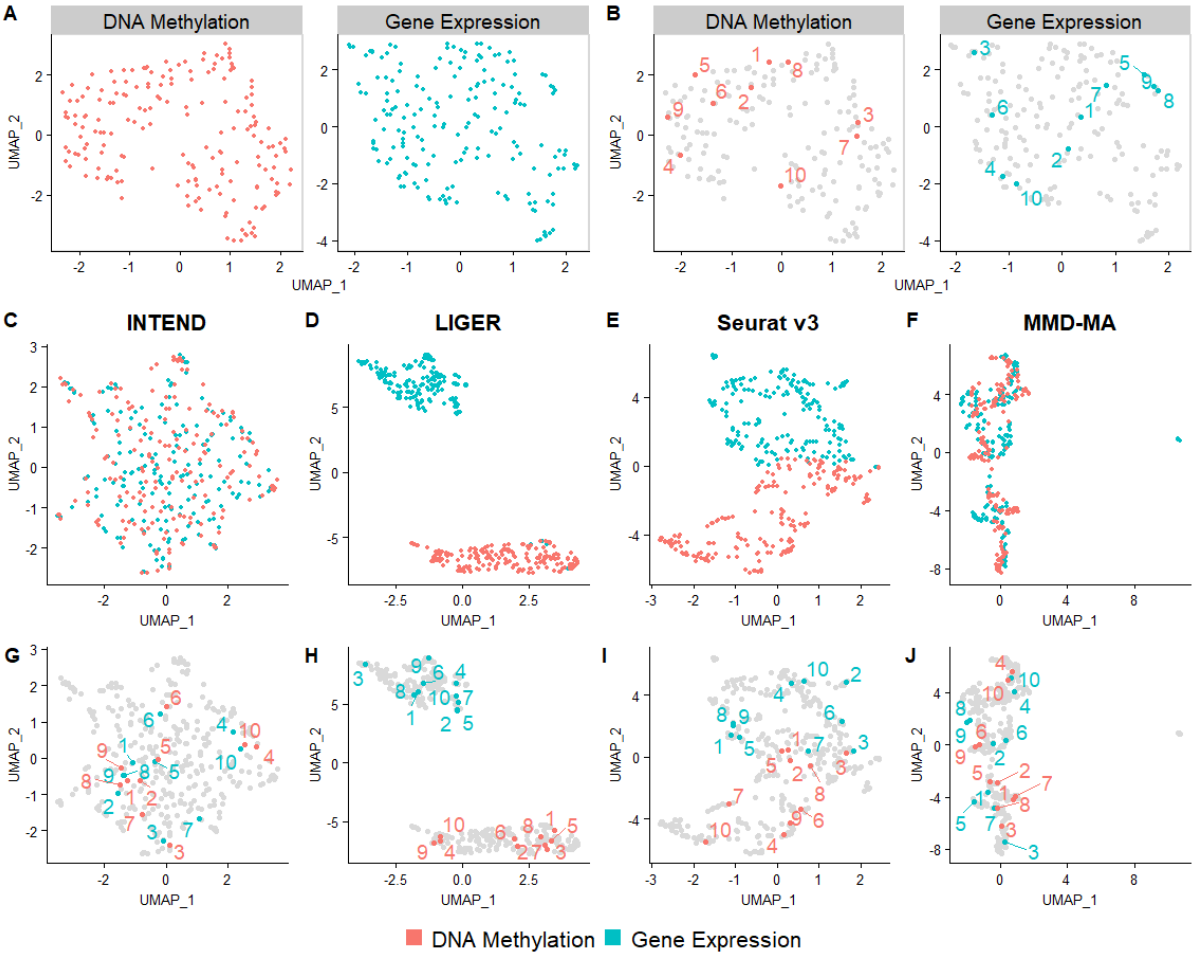
LIHC


A
DNA Methylation
Gene Expression
B
DNA Methylation
Gene Expression
C
INTEND
D
LIGER
E
Seurat v3
F
MMD-MA
G
H
I
J
UMAP_1
UMAP_2
DNA Methylation
Gene Expression

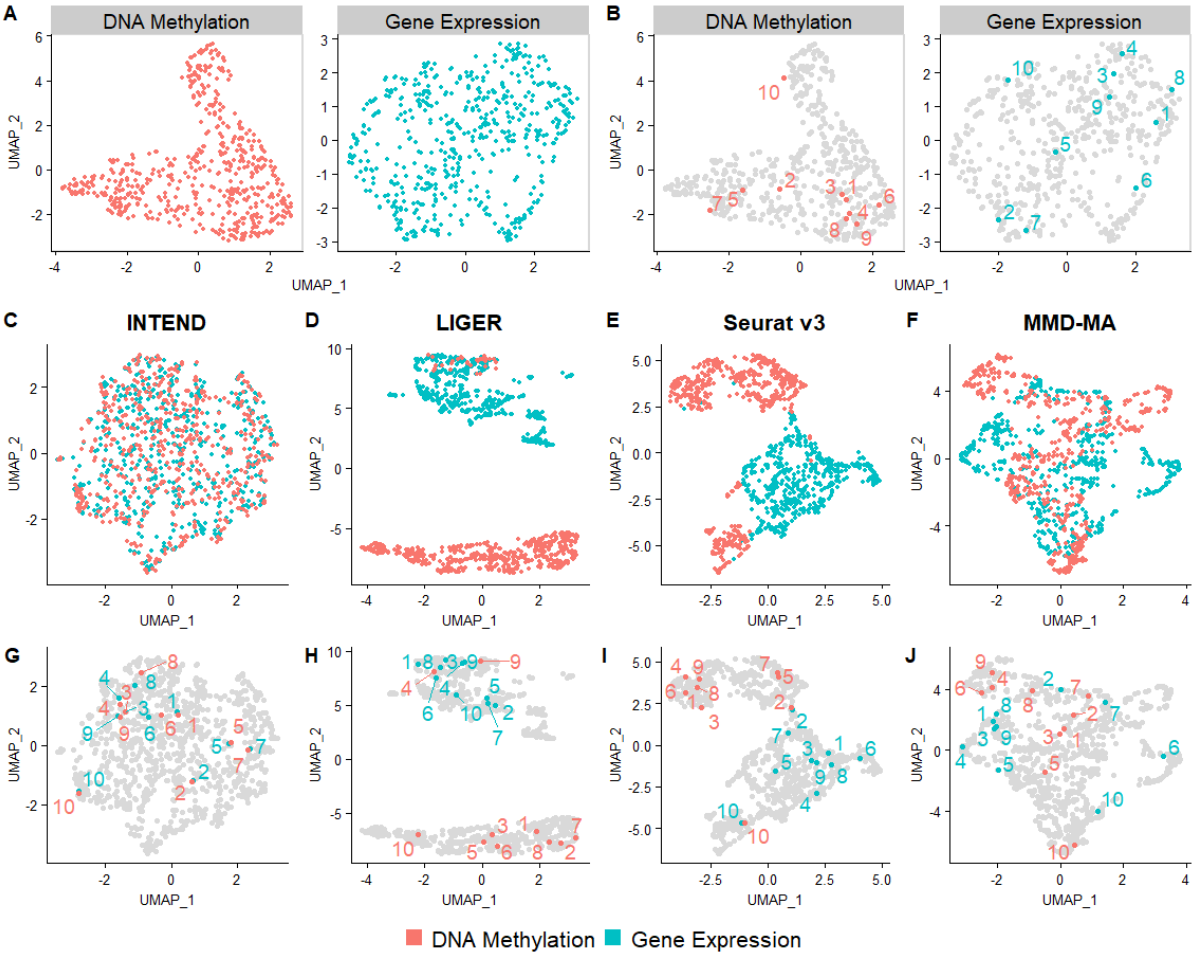
LUAD


A
DNA Methylation
Gene Expression
B
DNA Methylation
Gene Expression
C
INTEND
D
LIGER
E
Seurat v3
F
MMD-MA
G
H
I
J
UMAP_1
UMAP_2
DNA Methylation
Gene Expression

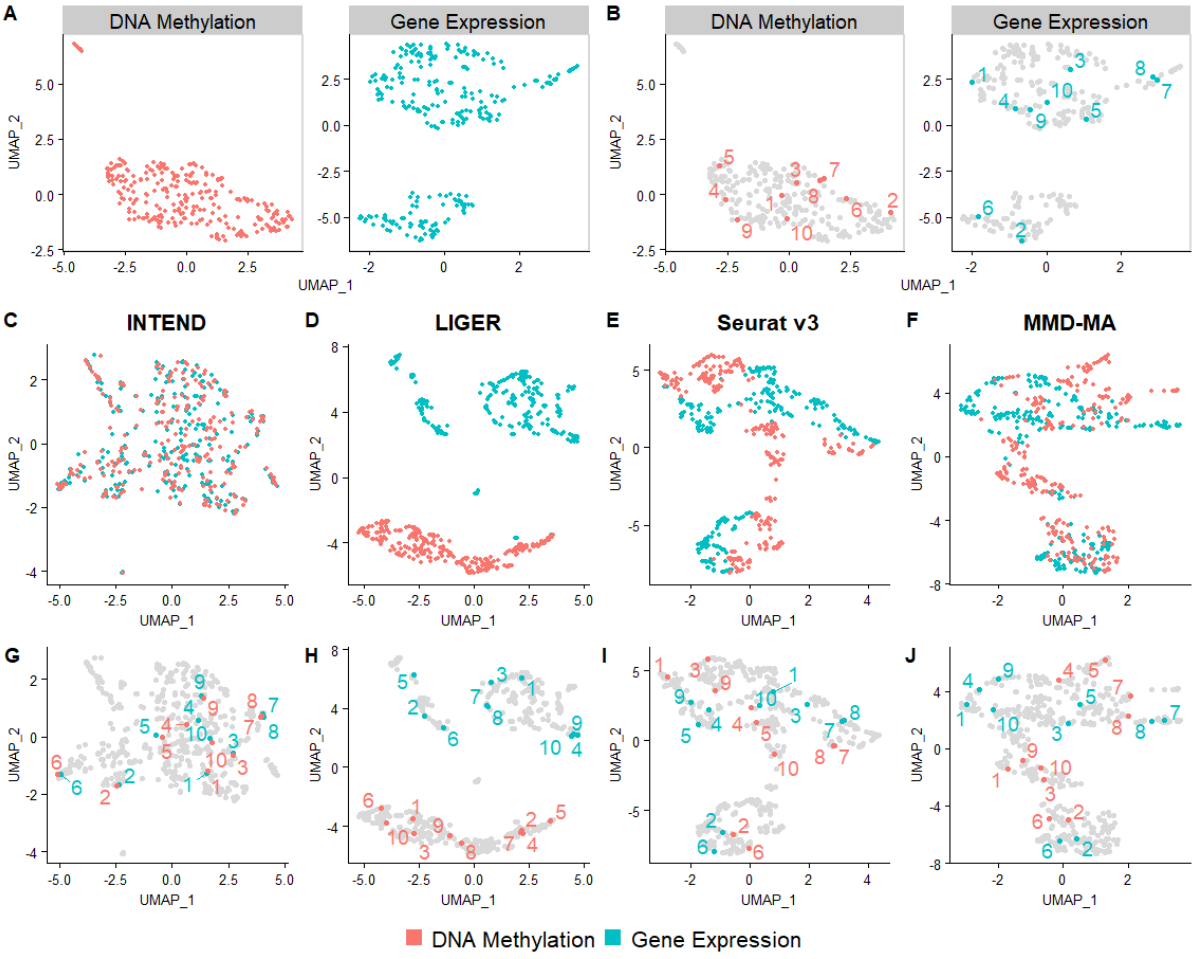
PAAD


A
DNA Methylation
Gene Expression
B
DNA Methylation
Gene Expression
C
INTEND
D
LIGER
E
Seurat v3
F
MMD-MA
G
H
I
J
UMAP_1
UMAP_2
DNA Methylation
Gene Expression

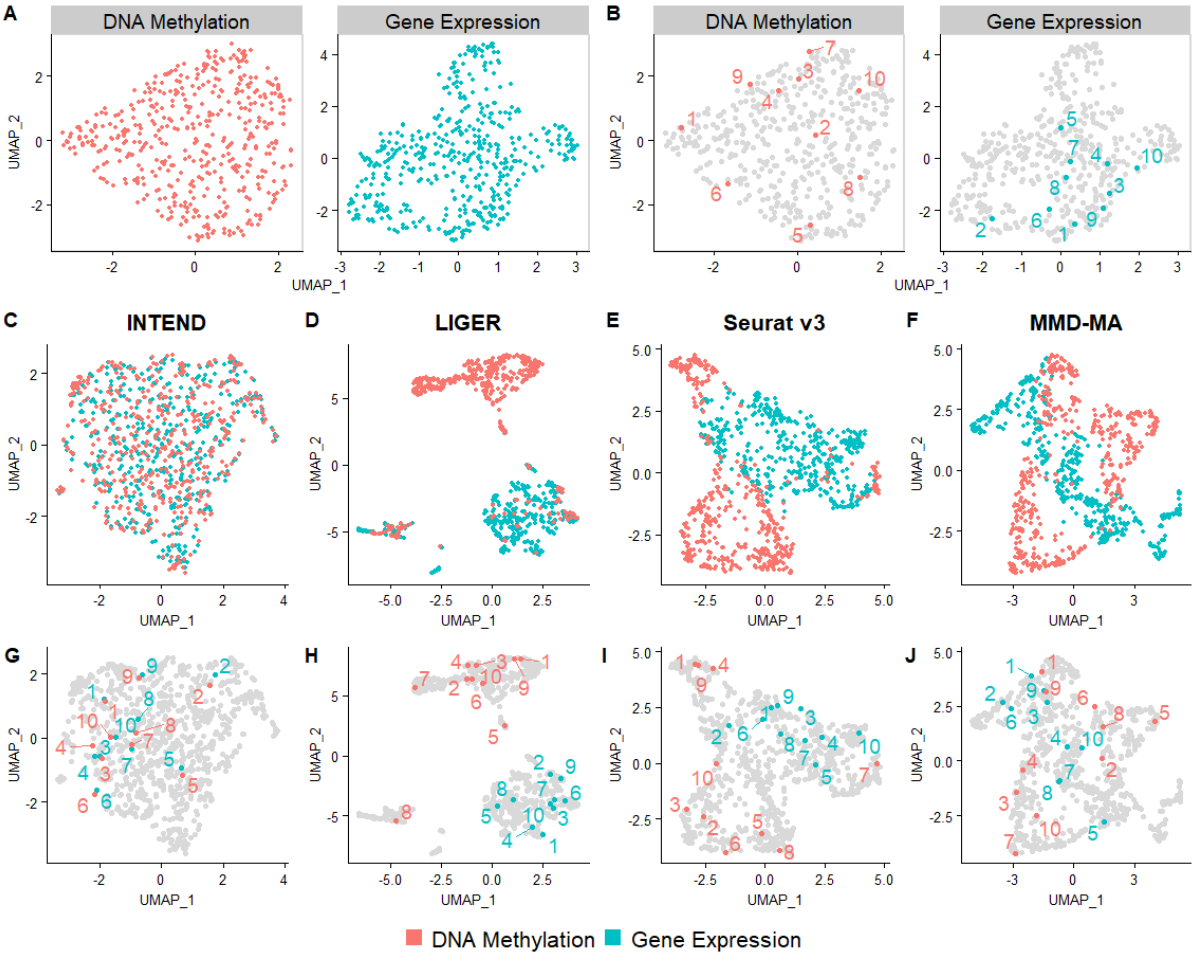
PRAD


A
DNA Methylation
Gene Expression
B
DNA Methylation
Gene Expression
C
INTEND
D
LIGER
E
Seurat v3
F
MMD-MA
G
H
I
J
UMAP_1
UMAP_2
DNA Methylation
Gene Expression

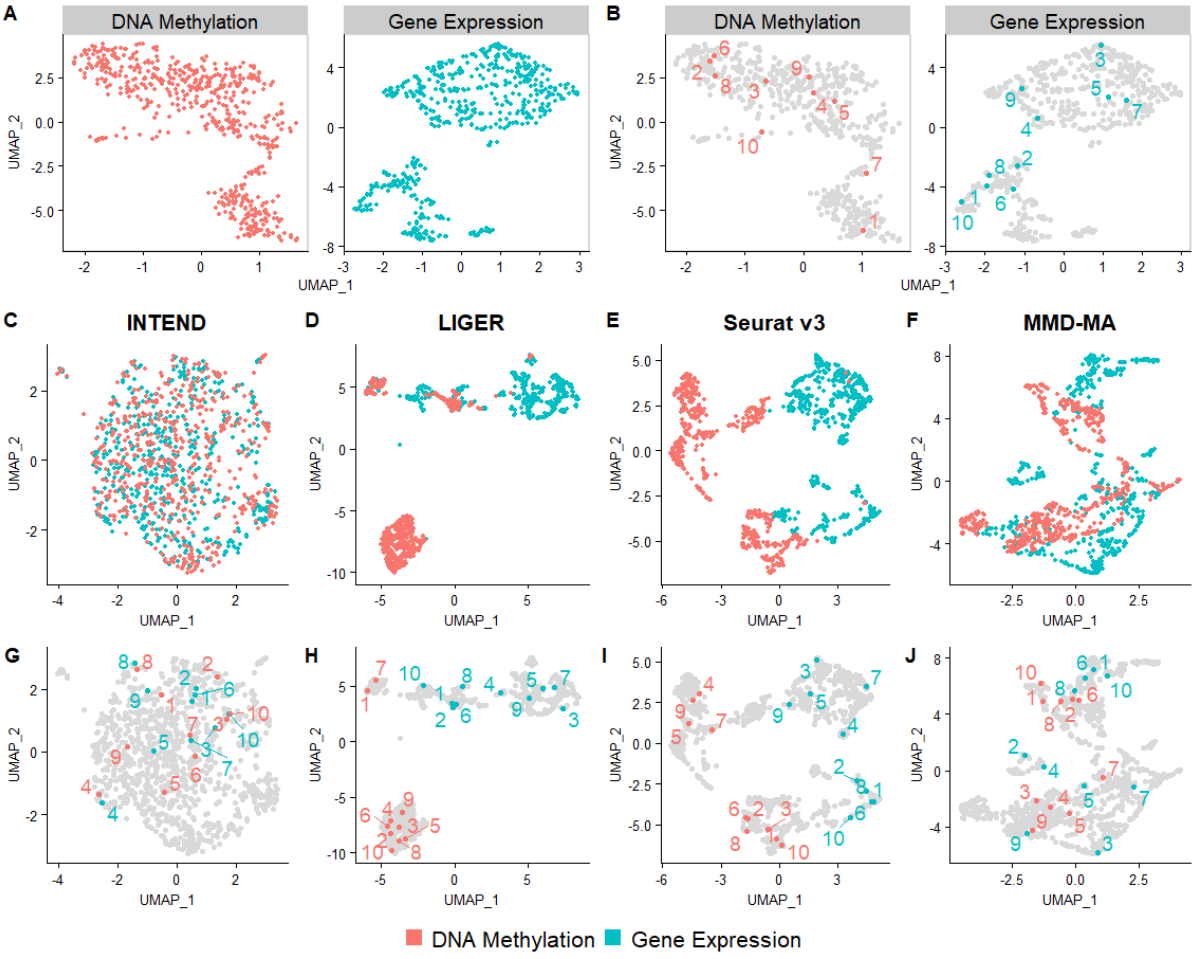
SARC

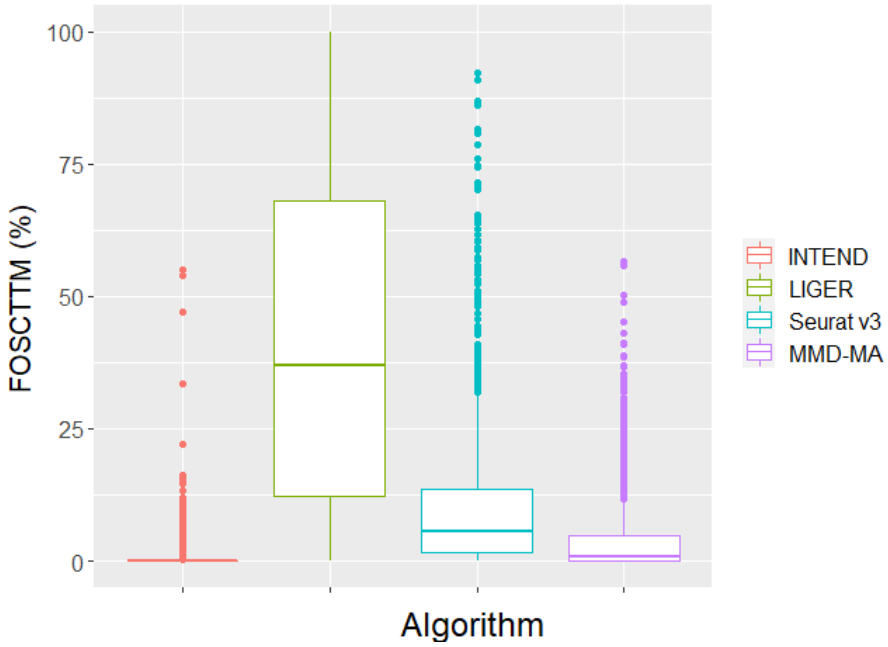

A
DNA Methylation
Gene Expression
B
DNA Methylation
Gene Expression
C
INTEND
D
LIGER
E
Seurat v3
F
MMD-MA
G
H
I
J
UMAP_1
UMAP_2
DNA Methylation
Gene Expression

SKCM

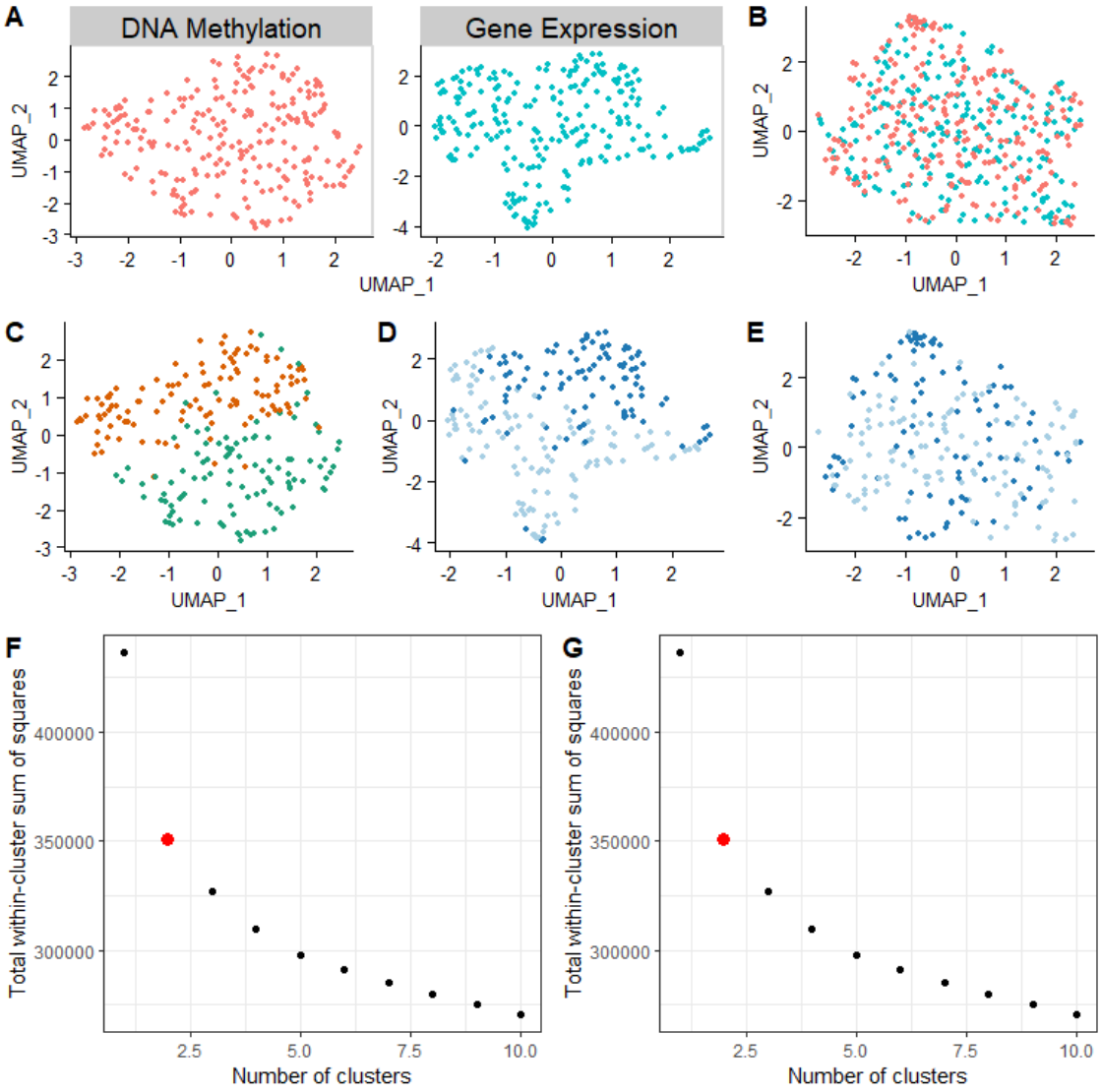

A
DNA Methylation
Gene Expression
B
DNA Methylation
Gene Expression
C
INTEND
D
LIGER
E
Seurat v3
F
MMD-MA
G
H
I
J
UMAP_1
UMAP_2
DNA Methylation
Gene Expression

THCA


A
DNA Methylation
Gene Expression
B
DNA Methylation
Gene Expression
C
INTEND
D
LIGER
E
Seurat v3
F
MMD-MA
G
H
I
J
UMAP_1
UMAP_2
DNA Methylation
Gene Expression



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.

תקציר

ניתוח אינטגרטיבי של מערכי נתונים מרובי-אומיקים (multi-omic, המכילים מידע גנומי נרחב ממספר סוגים שונים) הוכח כבעל ערך רב במיוחד בחקר הסרטן וברפואה מותאמת אישית. עם זאת, השגת נתונים מרובי-אומיקים שמקורם באותן דגימות לעיתים קרובות כרוכה בקשיים רבים. אינטגרציה של מספר מערכי נתונים ממספר אומיקים שונים עודנה מהווה אתגר, ועד כה פותח רק מספר מצומצם של אלגוריתמים כדי להתמודד עם הבעיה.

במסגרת עבודה זו, פותח INTEND (IntegratioN of Transcriptomic and EpigeNomic Data), אלגוריתם חדשני לאינטגרציה של נתוני ביטוי גנים ומתילציה של דנ"א שמקורם בקבוצות זרות של דגימות. בכדי לאפשר את האינטגרציה, INTEND לומד מודל חיזוי לביטוי גנים על בסיס מתילצית דנ"א, ע"י אימון על נתונים מרובי-אומיקים שנמדדו על אותה קבוצת דגימות. בבדיקה מקיפה על מערכי נתונים שמקורם ב-11 סוגי סרטן שונים ואשר מקיפים 4329 חולים, INTEND השיג תוצאות טובות באופן משמעותי בהשוואה לארבעה אלגוריתמי אינטגרציה מתקדמים אחרים. בעבודה זו מודגמת גם היכולת של INTEND לחשוף קשרים בין מתילצית דנ"א ובקרת ביטוי גנים בניתוח אינגרטיבי של שני מאגרי מידע של סרטן ריאות (lung adenocarcinoma) ממקורות שונים – מאגר מידע של מתילציה ומאגר מידע של ביטוי גנים. הגישה מוכוונת הנתונים של INTEND הופכת אותו לכלי רב-עוצמה לניתוח אינטגרטיבי של מידע מרובה-אומיקים.

אוניברסיטת תל אביב

הפקולטה למדעים מדויקים ע"ש ריימונד ובברלי סאקלר

בית הספר למדעי המחשב ע"ש בלווטניק

אינטגרציה של נתוני ביטוי גנים ומתילציה של דנ"א על פני ניסויים שונים

חיבור זה הוגש כעבודת גמר לתואר 'מוסמך אוניברסיטה'

בבית הספר למדעי המחשב

על ידי

יונתן איתי

בהנחיית

פרופ' רון שמיר

אייר תשפ"ג