Scalable Integration of Multiomic Single Cell Data Using Generative Adversarial Networks

Valentina Giansanti[1,2], Francesca Giannese[2], Oronza A. Botrugno[3], Giorgia Gandolfi[2], Chiara Balestrieri[2], Marco Antoniotti[1,4], Giovanni Tonon[2,3] and Davide Cittaro[2*]

[1]Department of Informatics, Systems and Communication, Università degli Studi di Milano-Bicocca, Milan, Italy
[2]Center for Omics Sciences, IRCCS San Raffaele Institute, Milan, Italy
[3] San Raffaele Scientific Institute, Division of Experimental oncology, Functional genomics of cancer
[4]Bicocca Bioinformatics Biostatistics and Bioimaging Centre -- B4, Università degli Studi di Milano-Bicocca, Milan, Italy

*To whom correspondence should be addressed.

Running Title: Multimodal data integration with MOWGAN

Character count: 62,951 (with spaces)

Abstract

Single cell profiling has become a common practice to investigate the complexity of tissues, organs and organisms. Recent technological advances are expanding our capabilities to profile various molecular layers beyond the transcriptome such as, but not limited to, the genome, the epigenome and the proteome. Depending on the experimental procedure, these data can be obtained from separate assays or from the very same cells. Despite development of computational methods for data integration is an active research field, most of the available strategies have been devised for the joint analysis of two modalities and cannot accommodate a high number of them.

To solve this problem, we here propose a multiomic data integration framework based on Wasserstein Generative Adversarial Networks (MOWGAN) suitable for the analysis of paired or unpaired data with high number of modalities (>2). At the core of our strategy is a single network trained on all modalities together, limiting the computational burden when many molecular layers are evaluated. Source code of our framework is available at https://github.com/vgiansanti/MOWGAN.

Keywords: data integration/deep learning/multimodal data/single cell analysis

Introduction

Availability of single-cell sequencing technologies made easier the understanding of complex organization of tissues and organs and the interplay among different types of cells. Here, cell properties can be characterized at different layers, in terms of transcriptome (1–4), genome (5–7) epigenome (8–10) and proteome (11,12). Depending on the experimental procedure, these data are available from the same cells or from separate assays (13). Multiomic data from the same cells are currently limited to up to 2/3 modalities (or layers) but the prospective is to have a higher rate of co-observation (14,15), as long as additional assays do not compromise profiling of other layers, keeping the cell intact.

Proper integration of multiomic data is one of the grand challenges in single-cell data analysis (16). Integration is considered a difficult task as it requires specific computational models supporting multidimensional data, it is based on generally unknown models of mutual dependence and causal relationships among modalities, and it includes all the analysis challenges that are peculiar to each single modality (*e.g.*, sampling bias, high sparsity).

Data integration links different data sources to derive a more comprehensive and biologically meaningful description of the object under analysis. This task can be addressed in three different ways depending on the type of available anchors used to combine multiple data sources (17): we refer to *horizontal integration* when integration is performed between different datasets representing the same data type (*i.e.*, multiple scRNA-seq dataset) collected from multiple samples at different locations or time points. V*ertical integration* is performed when more modalities are assayed from the same cell, for example epigenome and transcriptome (10,18,19) transcriptome and proteins (20), genome and transcriptome (21) or any other combination (22–25). Vertical integration is challenging both from an experimental and an interpretative point of view. Finally, we refer to *diagonal integration* if both cells and features are different for all datasets. This last scenario may be

considered the hardest to solve and yet it is possibly the most common, given the pace at which single-cell datasets are produced (26).
While horizontal integration can be treated as a problem of batch correction (27), vertical and diagonal concern with the multimodality integration field. To the best of our knowledge, a fully integrated and generalizable way to analyse such data is still missing. The main solutions proposed so far refer to *Manifold Alignment* (MA) applications and *Deep Learning* (DL) (28–42). Both approaches share the final goal of representing multiple feature sets in a common manifold embedding. MA methods try to find a common latent space (manifold) to describe the data, while DL methods develop networks specific for the single omic and work on their embeddings, (*i.e.*, learned low-dimensional representations of the given data).
In order to produce their results, both MA and DL methods must assume some constraints. Restrictive assumptions are generally applied on the data, like correspondences between the features (*e.g.*, converting chromatin accessibility data into gene activity scores (43)) and/or between cells, or assumptions on the data distribution. These requirements are difficult to fulfil and to generalize, making them unfit for datasets where no prior knowledge is available. Furthermore, they are typically considered to address the integration of only two molecular layers (*e.g.*, RNA and ATAC), so scaling to three or more modalities could be impractical.
In addition to technical limitations and experimental costs, many large datasets have been made available for single omics only (44–46). This urges the need for the development of an approach to integrate paired or unpaired data which is also free from assumptions and flexible with respect to the number of omics available. To address these issues, we propose MOWGAN, a DL framework for the analysis of Multi-Omics paired or unpaired data based on Wasserstein Generative Adversarial Networks. Our approach is designed to accommodate any kind and number of single-cell assays without priors on the relationships among the inputs. MOWGAN learns the structure of single assays and infers the optimal couplings between pairs of assays. In doing so, MOWGAN generates synthetic multiomic datasets that can be used to transfer information among the measured assays by bridging (47). We benchmarked MOWGAN with existing methods for multimodal single-assays data integration on various single cell datasets and show how it could be used to unveil hidden biology. We show that MOWGAN generates much more reliable results when looking at the cell type annotation shared between the layers.
A python implementation of MOWGAN can be found at https://github.com/vgiansanti/MOWGAN.

Results

Overview of MOWGAN

The core component of the framework is a Wasserstein Generative Adversarial Network (WGAN) with gradient penalty (WGAN-GP). A WGAN-GP is a generative adversarial network that uses the Wasserstein (or Earth-Mover) loss function and a gradient penalty to achieve Lipschitz continuity (48,49). Like any other GAN, the WGAN-GP is composed of two subnetworks, called *generator* and *critic*. MOWGAN's *generator* outputs a synthetic dataset where cell pairing is introduced across multiple modalities.
MOWGAN's inputs are molecular layers embedded into a feature space having the same dimensionality C (Figure 1). In principle any dimensionality reduction technique could be

used (*e.g.*, PCA, SVD…). To capture local topology within each dataset, cells in each embedding are sorted by the first component of its Laplacian Eigenmap (LE). This step maximises the likelihood of having similar cells close to each other within each layer. Training of the WGAN-GP is performed in mini-batches, iteratively sampling random subsets of cells from the whole sorted dataset. First, a mini-batch is selected from one modality (usually RNA) and a Bayesian ridge regressor *R* is trained on the mini-batch embedding and the corresponding eigenvectors from the LE. Next *n* mini-batches (default *n* = 50) are selected from the second modality. We apply *R* on all mini-batches and their eigenvectors. The resulting scores are used to sort mini-batches and select the batch in the second modality (*e.g.*, ATAC) with local properties more similar to the first, therefore maximising the probability to select cells drawn from same cell types. This procedure is repeated for all modalities, always using one layer (*e.g.*, RNA) as reference. In preliminary implementations of MOWGAN, without LE sorting and pre-matching, we noticed the WGAN-GP could only capture global topology of the data, resulting in random mixing of cell types (Figure S1).
The mini-batches are combined to form a tensor (*N*, *M*, *C*), where *N* is the number of cells in a mini-batch (default *N* = 256), *M* is the number of modalities evaluated (2 in case of just the RNA and ATAC layers), and *C* is the rank of the input embeddings. In any case, *C* is set at most to the lowest matrix rank across all modalities.
Semi-supervised training is implemented to include prior information about the dataset. This allows to model, for example, batch-specific models that are eventually merged. Lastly, MOWGAN can explicitly include paired data during training.
We implemented MOWGAN in Python using Tensorflow libraries to support of GPU-enabled hardware and scale to large datasets.

MOWGAN preserves data topology and biological information

We prototyped MOWGAN on public scRNA and scATAC data of peripheral blood mononuclear cell (PBMC) for which paired and unpaired experiments were available (10x Genomics datasets). The paired dataset (PBMC_1) is a true multimodal experiment, with transcriptome and epigenome data from the same cells; the unpaired dataset (PBMC_2) includes data from two separate experiments, one for each modality.
We evaluated MOWGAN output in terms of shared information between layers (Adjusted Mutual Information of cell clusters, AMI), integrability (Local Inverse Simpson's Index, LISI (50)) and ability to recapitulate biological information within the same modality. Experiments were performed using either PBMC_1 or MOWGAN as data bridge to transfer labels from RNA to ATAC modality.
First, we identified a set of cell clusters *T* ("truth") in each modality, we used data bridges (either the PBMC_1 dataset or synthetic pairings generated by MOWGAN) to transfer labels from one modality to another. We denote the set of transferred cell clusters as *P* ("prediction"). Labels are propagated from the first modality D_1 to MOWGAN and then from MOWGAN to the second modality D_2 using label transfer properties of a Stochastic Block Model (51). As every cell in D_1 and D_2 will be annotated by *T* and *P*, we could calculate mutual information. When measured on PBMC_1 data using direct transfer by cell identity, mutual information was computed comparing *T* annotations of RNA and ATAC and found below 1 (AMI = 0.691), indicating that different modalities encode slightly different cell properties.
We obtained lower mutual information values when cell clusters were transferred using MOWGAN, which was expected (AMI_{RNA} = 0.662, AMI_{ATAC} = 0.594 for PBMC_1, AMI_{RNA} =

0.411, AMI_{ATAC} = 0.448 for PBMC_2). Lower values for PBMC_2 were also expected, as the dataset is composed by two unpaired experiments which were acquired, analyzed and released over a long extent.
Data integrability was measured by LISI, a measure of the amount of data mixing in the cell neighborhood. If two datasets (*A,B*) are well integrated, we may expect the neighbors of every cell in *A* being well represented in *A* and *B* datasets, and *vice versa*. When integration fails, most of the neighbors of every cell in *A* (*B*) will be from dataset *A* (*B*) only. LISI is calculated within the same modality (D_1 or D_2 with MOWGAN) and it should be intended as a global measure of similarity of two datasets. It varies from 1 (perfect separation) to the number of modalities *M* (perfect mixing).
Baseline measures of integrability, evaluated on PBMC_1 and PBMC_2, were dependent on the modality ($LISI_{RNA}$ = 1.2, $LISI_{ATAC}$=1.35). Data generated by MOWGAN had higher LISI values ($LISI_{RNA}$ = 1.636, $LISI_{ATAC}$ 1.641 for PBMC_1 and $LISI_{RNA}$ = 1.354, $LISI_{ATAC}$ = 1.438 for PBMC_2), indicating that it is able to compute synthetic data that resemble the input.
Lastly, we evaluated the ability to transfer biological information in terms of power of identification of cell types. We identified sets of cell type markers in each modality of PBMC_1. We then calculated, for each cell type, the fold change of each feature in MOWGAN data and PBMC_2, the latter to evaluate baseline performances, and performed preranked GSEA (52) against the markers. While the WGAN-GP itself does not generate feature values, MOWGAN imputes them using kNN-regression (see Methods). We evaluated the performance as average rank *K* of NES values on the diagonal in the confusion matrix, so that *K* = 1 if all enrichments maximally match the correct cell type and *K* = 0 if the correct cell type has the lowest possible enrichments (Figure 2).
The baseline performance, evaluated with PBMC_2, indicates that different datasets, while derived with different protocols, share a large portion of explained biology (K_{RNA} = 0.964, K_{ATAC} = 0.686). Synthetic data generated with MOWGAN share much of the biological information with the respective input data, performance was high both for PBMC_1 (K_{RNA} = 0.800, K_{ATAC} = 1.000) and PBMC_2 (K_{RNA} = 0.869, K_{ATAC} = 0.952). We noticed that K_{ATAC} for MOWGAN was higher than baseline, we ascribe this result to the intrinsic denoising properties of the generative model and the kNN regressor.
In all, these results indicate that MOWGAN can induce appropriate pairings in synthetic data which, in turn, can be used to perform bridge integration of multiple modalities.

Comparison with other tools

We benchmarked MOWGAN against four tools that share similar design principles: Pamona (30), SCOT (32), COBOLT (35) and scMMGAN (29). Pamona and SCOT use Optimal Transport theory to project each modality on a single latent embedding. The new embeddings can be processed by standard workflow to identify cell clusters. scMMGAN projects data of each modality to the embedding defined on the other modality, so that two modalities can be aligned in a single space. Similarly to Pamona and SCOT, the resulting embedding can be used to identify cell clusters. COBOLT, instead, projects unpaired data on the embedding obtained from paired data, to perform bridge integration.
For each tool, we evaluated the consistency between cell types and cell clusters derived from the generated embeddings, both in PBMC_1 and PBMC_2 datasets. Cell types were annotated once on both datasets, to be consistent across all experiments. In addition to AMI between *P* and *T* cell clusters, we measured Completeness (*i.e.*, all cells of any cell type are clustered together), Homogeneity (*i.e.*, all cells of any cluster belong to the same cell

type), and their harmonic mean (V-measure). In general, we observed better performances in PBMC_1 compared to PBMC_2 and on the RNA layer compared to the ATAC layer, without any tool outperforming the others (Figure 3A, Table S1).
More in detail: SCOT had the best V-measures for RNA (V_{PBMC_1} = 0.562, V_{PBMC_2} = 0.616) and the worst for ATAC (V_{PBMC_1} = 0.378, V_{PBMC_2} = 0.335), whereas MOWGAN displayed a similar performance both on RNA (V_{PBMC_1} = 0.482, V_{PBMC_2} = 0.383) and ATAC (V_{PBMC_1} = 0.505, V_{PBMC_2} = 0.450). MOWGAN also performed similar to scMMGAN, with which shares some design principles as both rely on deep generative models. Of note, COBOLT failed in integrating ATAC data in PBMC_2 (V_{PBMC_2} = 0.009) and its high performance in PBMC_1 was probably due to the fact it explicitly uses cell identities to integrate two modalities.
Measures of mutual information reveal that all tools share similar performance on PBMC_1 dataset (average AMI_{PBMC_1} = 0.364) but MOWGAN was the only tool sustaining comparable performance on PBMC_2 (AMI_{PBMC_2} = 0.395), while the remaining tools had AMI values close to 0 (Figure 3B, Table S1).
Integrability measured by LISI showed comparable performance of MOWGAN and scMMGAN, again possibly due to the fact both are based on GAN. We report low LISI values for SCOT and Pamona, however they should not be considered in the comparison as both tools generate an embedding in a new reference system that is necessarily different from the input embeddings.
Last, we noticed some tools induce spurious cell type associations (Figure S2). For example, SCOT switched B cells with NK/CD8 T cells in PBMC_1 and Monocytes with T cells in PMBC_2. Similarly, scMMGAN suffers of several cell type switches in all the tested datasets. As previously noticed, COBOLT shuffled cell types in the ATAC layer in PBMC_2.
In all, these results indicate that all tools but COBOLT preserve the local topology of the input data as clustering properties are conserved in the data transformations. In this regard, no tool outperforms the others, suggesting an upper bound for integration tasks. However, most tools, except for MOWGAN, did not guarantee the correct cell type associations, a further evidence of the ability of MOWGAN to generate good quality and meaningful data pairings.

MOWGAN extends the integration to high number of modalities

The major advantage of MOWGAN is that it can integrate more than two molecular layers without modifying the model architecture. To test this feature we took advantage of the availability of different single cell profiles for PBMC; in particular we attempted the integration of PBMC_1 dataset with H3K27me3 and Antibody Derived Tags (ADT) data obtained with scCUT&TAG-pro (53) on PBMC, to achieve a four-layers integration. It's worth underline that the pairings between ADT and H3K27me3 and between RNA and ATAC were not exploited in this experiment. We observed high integrability between original and synthetic data generated with MOWGAN, with the lowest performance detected for H3K27me3 ($LISI_{RNA}$ = 1.67, $LISI_{ATAC}$ = 1.67, $LISI_{ADT}$ = 1.75, $LISI_{H3K27me3}$ = 1.38).
We used MOWGAN to bridge cell type annotations from RNA to all the remaining modalities, then compared with the given cell type (Figure 4). We found similar results for ATAC and ADT, both in terms of accuracy (> 0.8) and V-measure (> 0.6), whereas H3K27me3 performed worst, suggesting RNA, ATAC and ADT share similar resolution power with respect to cell types. On the other hand, H3K27me3 clearly separated myeloid and lymphoid lineages with lower resolution power in the latter. An analogous behavior of H3K27me3 has been previously described (54).

Of note, we found the performance of inverse labeling, from any modality to RNA, equally high, if not better (Table S2). In all, these results show how MOWGAN can effectively compute synthetic pairings for multiple modalities, overcoming a limitation that is common to many available tools.

Semisupervised training improves results and unveils hidden cell populations in colorectal cancer organoids

An ideal tool to integrate unpaired data would be able to work in unsupervised manner, that is it would learn the whole data structure from data themselves. That requires the algorithm to model several confounders including technical noise due to different technologies (13,55), biological noise (56,57) and batch effects (27). For this reason, we tested the ability of MOWGAN to work with multiple samples of unpaired data with or without specifying the sample identity in the training step.

We profiled three patient-derived organoids (PDO) of liver metastatic colorectal cancer (CRC) by scRNA-seq and scGET-seq (25) (Table S3). Cells for each PDO were split in two aliquots before single cell preparation, so that data are paired at sample level (*i.e.*, single cells represent the same populations) but not at modality level. Analysis of copy number alterations at single cell level revealed that all PDOs are genetically uniform and essentially monoclonal (Figure S3A). These findings were previously confirmed for CRC6 and CRC17 by bulk exome sequencing (25).

To verify the effectiveness of semisupervised training, we trained one MOWGAN model without specifying the organoid of origin (NB) and a model that instead included such information (B). In practice, to obtain model B, we trained three separated models, one per organoid, sampling batches from the whole dataset, and eventually concatenating the outputs.

We checked the average concordance between the actual organoid labels and the one we obtained transferring it from scRNA-seq to scGET-seq and *vice versa* using MOWGAN as bridge. In model NB the performance was low in terms of accuracy (ACC_{RNA} = 0.111, ACC_{GET} = 0.353) and increased in model B (ACC_{RNA} = 0.982, ACC_{GET} = 0.668).

Synthetic data were generally integrable with the original ones, with minor differences between model B ($LISI_{RNA}$ = 1.55, $LISI_{GET}$ = 1.33) and NB ($LISI_{RNA}$ = 1.42, $LISI_{GET}$ = 1.25), in line with previous observations on PBMC datasets. In all, we conclude that semisupervised learning improves the quality of the generated data and their usability for the integration.

Having set the model B is better, we performed further analysis to unveil the biological features of PDOs. We identified cell states using the multimodal capabilities of schist (51), which relies on a Stochastic Block Model with overlap. Two kNN graphs, including MOWGAN generated cells, are computed for each modality and merged specifying an edge label depending on the modality, then a Nested Stochastic Block Model (NSBM) with edge covariates is inferred. Of note, MOWGAN output is represented in the graph as a set of nodes having both types of edges, hence providing the link between modalities. The resulting groups are consistent across modalities and can be used to analyze phenotypes from the transcriptomic and the epigenetic standpoint. We computed a unified embedding using optimal transport framework to visualize data annotations in consistent way (Figure 5A-C).

Since the rate of transcription depends, among other factors, on genome accessibility, we calculated the CytoTrace score (58) from scRNA-seq and a chromtain accessibility score (59) from scGET-seq. The correlation between the two scores, averaged on the deepest level of

NSBM hierarchy, was positive (Figure 5D, r = 0.35, p = 1.94e-22). In addition, we found a positive correlation between accessibility and CRIS-B signature for CRC (Figure S3B, r = 0.18, p = 4.054e-7). We expected positive correlations in case of appropriate pairing, since CRIS-B subtype has been described associated to EMT phenotype (60), as well as higher CytoTrace scores and chromatin accessibility (61,62) from scGET-seq.

We then evaluated the activity of various transcription factors from gene expression data, using decoupler (63) and a cell-level measure derived from Total Binding Affinity (TBA, (64)), computed from chromatin accessibility (see Methods). Correlation coefficients of the two measures were skewed to positive values (Figure 5E, skeweness = 0.251), again supporting the good quality of the integration. Among other transcription factors (Table S4), we found high correlation for NFE2L2 (r = 0.39, p = 1.742e-28), which regulates stress response (65) and has been recently found involved in CRC resistance to therapy (66). Interestingly, NFE2L2 is strongly associated with only one subpopulation in CRC39 (Figure 5G), indicating the PDOs, while being genetically uniform, display phenotypically diverse cell states.

We focused our attention at level 4 of the NSBM hierarchy as it provided a good trade-off between modularity and detail on the subpopulations within each PDO. We identified 32 clusters, seven of which include more than 200 cells in both modalities. All PDOs, apart from CRC6, were composed of different subpopulations with distinct phenotypes (Figure 5F). We scored multiple expression signatures studied for their relevance in drug resistance (67) and confirmed that both CRC17 and CRC39 included subpopulations which could support a differential response to treatments (Figure 5H, Figure S3C).

Importantly, these cell populations could be hardly identified by using scRNA-seq or scGET-seq data alone. First, we noticed that the NSBM hierarchy of cell clusters identified in single assays did not strictly follow the PDO identities and, in fact, some clusters are shared across different organoids, particularly in the scGET-seq data. Cell groups identified on the integrated data appear to be more consistent with the organoid identity than groups identified on the single modalities (Figure S3D); this behavior is true at any level of the NSBM hierarchies. Consequently, some phenotypes we observed above could not be clearly assigned to specific subpopulations.

In all, these data confirm that MOWGAN is able to produce good quality pairings to perform bridge integration across modalities. Moreover, we show that semisupervised training is an effective approach to operate when multiple samples/batches are available.

Discussion

Methods to probe molecular phenotypes of single cells at high throughput have been available for more than a decade now. Such methods have been rapidly adopted and are now standard techniques in many fields of biological studies. The largest part rely on analysis of gene expression, for which many technologies are available (68). Methods to investigate single cell epigenomes has also increased in number overtaking scRNA-seq (69). The development of unimodal approaches, including methods to investigate genome sequences (70) or proteomes (71,72), is paralleled by the introduction of multimodal techniques, whose number has also exploded in recent years (14,15,73). Unless innovative chemistry/microfluidic strategies will be introduced, multimodal investigation will be inherently limited to methods that preserve cell or molecule integrity. Moreover, despite the availability of multimodal techniques, their analysis remains challenging (74), moreover

they may be less informative of single-assay methods (75). For all these reasons, the development of computational strategies to pair unimodal assays is flourishing.
Along this path, we introduced MOWGAN, a DL-based framework that makes use of Wasserstein Generative Adversarial Networks (WGAN) to generate synthetic multimodal data for bridge integration of different single-cell assays. Given the heterogeneity of available technologies, we designed MOWGAN to work without the need of "biologically principled anchors" across modalities, such as Gene Activity Score for scATAC-seq data (43) or Chromatin Silencing Score for repressive histone marks profiled in scCUT&Tag (54).
Comparison with other tools on the integration of two unpaired modalities (scRNA-seq and scATAC-seq) shows MOWGAN performs similar to the top-level methods, without any tool outperforming the others, suggesting an intrinsic upper bound for this task. The major improvement of MOWGAN over existing tools is instead the possibility to work with more than two modalities. To do so, it aligns all input data to one base modality of choice, typically the transcriptome. Consequently, MOWGAN does not generate a data barycenter, nor it computes an embedding in a latent space. In this regard, MOWGAN could be used in a modular workflow, allowing the adoption of the method of choice to bridge information. We are aware that other tools can operate on high number of modalities, however they require partial overlaps among data (76) or explicit code refactoring (29).
Another important feature of MOWGAN is the possibility to specify experimental pairings, being those cell or sample identities. Semisupervised training of MOWGAN, when possible, reduces the possibility to hallucinate, a well-known issue that affects deep generative models (77), and allows the computation of good-quality and biologically meaningful data.
While it is established that "no method works best for all", we believe that MOWGAN has sufficient power to generalize over various technologies. We showed that integration of multiple modalities, notwithstanding that data are generated from separate experiments, provides information in a synergistic way, hence revealing hidden patterns. Our findings are particularly relevant in the context of intratumor heterogeneity. In our PDO dataset, all but one samples have been derived from tumors previously exposed to chemotherapy, therefore it's expected they contain (epigenetic) subclones selected for drug resistance; CRC39, on the other hand, has been derived from a naïve tumor and it would be tempting to speculate on the existence of subclones conferring drug resistance that could eventually expand upon therapy.

Materials and methods

WGAN-GP architecture

MOWGAN core component is a WGAN-GP composed of two networks, a *generator* and a *critic*. The *generator* is designed with 3 *convolutional 1D* layers (Conv1D) and 2 *batch normalization* layers (BN). The *critic* is designed with 2 Conv1D layers and a single Dense layer with 1 unit. All Conv1D layers have common parameters: *kernel size* = 2, *stride* = 1, *activation function* = ReLu. The number of *filters* for Conv1D layer changes with the network depth: default number of filters for the *generator* are 512, 128 and 20, for the *critic* 128 and 512.
Subnetwork uses different optimizers: Adam optimizer for the *generator* with *learning rate* = 0.001, *beta_1* = 0.5, *beta_2* = 0.9, *epsilon* = 1e-07; RMSprop optimizer with *learning rate* = 0.0005 in the *critic*.

Grid search

We applied a grid search policy to test the dependency of MOWGAN to hyperparameters. We tested multiple filters f = {8,32,64,128,256,512} in the Conv1D layers, and number of components selected from the embeddings C = {5,10,15}. We also set C as the number of filters in the *generator*'s hidden layer. A total of 216 models were trained, with training time for single model of approximately 3h every 100,000 epochs. 51 models returned NaNs in the loss function, 1 model produced data collapsed in a single point.

Feature reconstruction

We fit a kNN-regressor with k = 2 for each modality to learn the relationship between the count matrix and the embedding in the input data. The regressor was then applied on the WGAN-GP output to predict the count matrix of the synthetic data.

PBMC datasets

Public data for PBMC were downloaded from 10x Genomics resources site. PBMC_1 dataset includes Single Cell Multiome ATAC + Gene Expression (https://www.10xgenomics.com/resources/datasets/pbmc-from-a-healthy-donor-granulocytes-removed-through-cell-sorting-10-k-1-standard-2-0-0). PBMC_2 includes Single Cell Gene Gene Expression (https://www.10xgenomics.com/resources/datasets/10k-human-pbmcs-3-ht-v3-1-chromium-x-3-1-high) and Single Cell ATAC (https://www.10xgenomics.com/resources/datasets/10-k-peripheral-blood-mononuclear-cells-pbm-cs-from-a-healthy-donor-1-standard-1-2-0).
Transcriptome data were aligned to hg38 reference genome using STARSolo (78), providing gencode v36 as gene model (79). ATAC data were aligned to hg38 reference genome using bwa (80), we then extracted the counts over 5kb windows spanning the entire genome. For all datasets we calculated PCA, which was used as input for MOWGAN. Cell clusters were identified using the Planted Partition Block Model, implemented in schist.

PBMC scCUT&TAG-pro

H3K27me3 and ADT data for PBMC profiled with scCUT&Tag-pro (53) were downloaded from https://zenodo.org/record/5504061 (81). R objects were converted to hdf5 and then loaded into scanpy objects. We retained cell type annotation provided with the datasets. Data were normalized and log transformed before computation of the PCA embedding.

Patient-derived colorectal cancer organoids (PDOs)

Samples from three individuals with liver metastatic gastrointestinal cancers were obtained following written informed consent, in line with protocols approved by the San Raffaele Hospital Institutional Review Board and following procedures in accordance with the Declaration of Helsinki of 1975, as revised in 2000. PDO cultures were established as previously reported (25,82).
scGET-seq was performed as previously described (25,83) on a Chromium platform (10x Genomics) using 'Chromium Single Cell ATAC Reagent Kit' V1 chemistry (manual version CG000168 Rev C) and 'Nuclei Isolation for Single Cell ATAC Sequencing' (manual version CG000169 Rev B) protocols. The provided ATAC transposition enzyme (10x Tn5; 10x Genomics) was replaced with a sequential combination of Tn5 and TnH functional

transposons in the transposition mix assembly step. Specifically, a transposition mix containing 1.5 µl of 1.39 µM Tn5 was incubated for 30 min at 37 °C, then 1.5 µl of 1.39 µM TnH was added for a 1-h incubation. Nuclei suspensions were prepared to get 5,000 nuclei as target nuclei recovery. Final libraries were loaded on a Novaseq6000 platform (Illumina) to obtain 100,000 reads per nucleus, and a custom read 1 primer was added to the standard Illumina mixture (5'-TCGTCGGCAGCGTCTCCGATCT-3').
scRNA-seq was performed on a Chromium platform (10x Genomics) using 'Chromium Single Cell 3' Reagent Kits v3' kit manual version CG000183 Rev C (10x Genomics). Final libraries were loaded on a Novaseq6000 platform (Illumina) to obtain 50,000 reads per cell.

Analysis of PBMC datasets

Standard processing was applied to filter and normalized the data: for RNA, cells with less than 200 expressed genes and genes present in less than 10 cells were discarded; for ATAC, cells with more than 30% of captured regions and regions common to more than 80% of cells were selected. For PBMC_1, we further removed all cells that were not shared across the modalities.
We used PCA as embedding scheme for all datasets. Cell clusters were identified using a Planted Partition Block Model provided by schist (51).
Cell type annotation was defined on RNA PBMC_1 by evaluating cell markers, as illustrated in the Scanpy tutorial (https://scanpy-tutorials.readthedocs.io/en/latest/pbmc3k.html). The annotation was directly transferred to ATAC PBMC_1 by cell identity. Annotations were transferred to PBMC_2 using label transfer function provided by schist.
When we tested integration of four modalities, known cell types were converted to the ones provided in (53) according to the following schema:

(CD4 T cells, CD8 T cells)$_{10x}$, (CD4 T, CD8 T, Other T)$_{scCUT\&Tag\text{-}pro}$: T cells
(CD14+ Monocytes, FCGR3A+ Monocytes)$_{10x}$, (Mono, Other) $_{scCUT\&Tag\text{-}pro}$: Monocytes
(B cells)$_{10x}$, (B)$_{scCUT\&Tag\text{-}pro}$: B cells
(NK cells)$_{10x}$, (NK)$_{scCUT\&Tag\text{-}pro}$: NK cells
(Dendritic cells)$_{10x}$, (DC)$_{scCUT\&Tag\text{-}pro}$: Dendritic cells

Analysis of PDOs

Sequencing reads for scGET-seq were processed as previously described (83). Briefly, we counted the number of aligned reads in 5kb windows spanning the genome, both for tn5 and tnH enzymes. We used raw counts to fit Zero Inflated Poisson distributions at cell level, which were in turn used to estimate the excess of accessible and compact chromatin in each bin. We removed bins with coverage lower than the 90th percentile and cells with less than 2000 counts. The count matrices were subjected to tensor train decomposition (TTD), used as low-dimensionality representation of the data.
Sequencing reads for scRNA-seq were aligned to hg38 reference genome using STARSolo (78), using GENCODE v36 as gene model (79). We removed cells that a) were identified as doublets by scrublet (84), b) had more than 25% of mitochondrial reads and c) having less than 2000 genes identified. After normalization, identification of highly variable genes, regression of cell cycle and scaling, we obtained low dimensional representation by PCA.
We used 20 components of TTD and PCA as input for MOWGAN with default parameters, computing three different models, one for each organoid, which were then concatenated so that output data have a model label. We used schist to transfer batch identity from input

data to MOWGAN output and retained only synthetic data having model label equal to the transferred label.

We used schist (51) to infer multimodal cell clusters (*scs.inference.nested_nsbm_multi*) on the integrated dataset. Once cell clusters have been identified, we used the cell marginals (*i.e.*, the probability of a cell to be assigned to any cluster) as lineage specifications in CellRank (85), so that driver features could be identified in any dataset (*e.g.*, gene drivers for scRNA-seq, region drivers for scGET-seq).

In order to produce a unified UMAP embedding, we first computed separate UMAP for RNA and GET layers, initializing with PAGA positions (86) calculated on RNA using level 4 of the NSBM hierarchy. We then performed Domain Adaptation by unbalanced Optimal Transport based on Sinkhorn algorithm (87) implemented in PythonOT (88), using level 4 of the NSBM hierarchy as anchors for semisupervised fit.

Transcription factor activity was evaluated on scRNA-seq data, excluding MOWGAN synthetic data), using decoupler (63) with default parameters. To calculate the activity of transcription factors in scGET-seq data we first calculated Total Binding Affinity (64) of HOCOMOCO v11 motifs (89) in each genomic bin, using bionumpy (90). We selected the top 200 region drivers for each cluster, having positive correlation and q-value lower than 1E-20, for a total of 1253 regions. We then calculated the motif score *M* as

$$S = X_m \cdot T, \qquad M = \sqrt{|(S)|}\, sgn\, S$$

where X_m is the accessibility value estimated on scGET-seq data and *T* is the TBA value for each genomic bin. Once we obtained *M*, we regressed out coverage and scaled the resulting matrix. CRC Intrinsic Subtypes (CRIS) signatures have been collected from CMSCaller package (91).

Acknowledgements

We would like to thank all members of COSR for the support and fruitful discussions. This work has been supported by the Accelerator Award: A26815 entitled: ``Single-cell cancer evolution in the clinic'' funded through a partnership between Cancer Research UK and Fondazione Italiana per la Ricerca Sul Cancro (AIRC).

Author contributions

V.G. and D.C. devised MOWGAN, performed the analysis and wrote the manuscript. V.G. wrote the software. O.A.B. performed the experiments on PDOs. G.G. and C.B. characterized PDOs. F.G. performed single-cell experiments. M.A. and G.T. supervised the work. All authors read the manuscript.

Disclosure and competing interests statement

The authors declare no competing interest.

Data availability

scRNA-seq data for three PDOs have been deposited to ArrayExpress with the following ID: E-MTAB-13123. scGET-seq data for CRC39 have been deposited to ArrayExpress with the following ID: E-MTAB-13126. scGET-seq data for CRC6 and CRC17 were downloaded from ArrayExpress (E-MTAB-10219)

References

1. Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell. 2015 May 21;161(5):1187–201.

2. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell. 2015 May 21;161(5):1202–14.

3. Rosenberg AB, Roco CM, Muscat RA, Kuchina A, Sample P, Yao Z, et al. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. Science. 2018 Apr 13;360(6385):176–82.

4. Clark IC, Fontanez KM, Meltzer RH, Xue Y, Hayford C, May-Zhang A, et al. Microfluidics-free single-cell genomics with templated emulsification. Nat Biotechnol. 2023 Mar 6;

5. Chen C, Xing D, Tan L, Li H, Zhou G, Huang L, et al. Single-cell whole-genome analyses by Linear Amplification via Transposon Insertion (LIANTI). Science. 2017 Apr 14;356(6334):189–94.

6. Wang J, Fan HC, Behr B, Quake SR. Genome-wide single-cell analysis of recombination activity and de novo mutation rates in human sperm. Cell. 2012 Jul 20;150(2):402–12.

7. Zong C, Lu S, Chapman AR, Xie XS. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. Science. 2012 Dec 21;338(6114):1622–6.

8. Bartosovic M, Kabbe M, Castelo-Branco G. Single-cell CUT&Tag profiles histone modifications and transcription factors in complex tissues. Nat Biotechnol. 2021 Jul;39(7):825–35.

9. Rotem A, Ram O, Shoresh N, Sperling RA, Goren A, Weitz DA, et al. Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. Nat Biotechnol. 2015 Nov;33(11):1165–72.

10. Clark SJ, Argelaguet R, Kapourani C-A, Stubbs TM, Lee HJ, Alda-Catalinas C, et al. scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. Nat Commun. 2018 Feb 22;9(1):781.

11. Gebreyesus ST, Siyal AA, Kitata RB, Chen ES-W, Enkhbayar B, Angata T, et al. Streamlined single-cell proteomics by an integrated microfluidic chip and data-independent acquisition mass spectrometry. Nat Commun. 2022 Jan 10;13(1):37.

12. Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay PK, Swerdlow H, et al. Simultaneous epitope and transcriptome measurement in single cells. Nat Methods. 2017 Sep;14(9):865–8.

13. Zhu C, Preissl S, Ren B. Single-cell multimodal omics: the power of many. Nat Methods. 2020 Jan;17(1):11–4.

14. Ogbeide S, Giannese F, Mincarelli L, Macaulay IC. Into the multiverse: advances in single-cell multiomic profiling. Trends Genet. 2022 Aug;38(8):831–43.

15. Vandereyken K, Sifrim A, Thienpont B, Voet T. Methods and applications for single-cell and spatial multi-omics. Nat Rev Genet. 2023 Mar 2;1–22.

16. Lähnemann D, Köster J, Szczurek E, McCarthy DJ, Hicks SC, Robinson MD, et al. Eleven grand challenges in single-cell data science. Genome Biol. 2020 Feb 7;21(1):31.

17. Argelaguet R, Cuomo ASE, Stegle O, Marioni JC. Computational principles and challenges in single-cell data integration. Nature Biotechnology. 2021 May 3;

18. Chen S, Lake BB, Zhang K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. Nat Biotechnol. 2019 Dec;37(12):1452–7.

19. Ma S, Zhang B, LaFave LM, Earl AS, Chiang Z, Hu Y, et al. Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. Cell. 2020 Nov 12;183(4):1103-1116.e20.

20. Stoeckius M, Zheng S, Houck-Loomis B, Hao S, Yeung BZ, Mauck WM, et al. Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. Genome Biol. 2018 Dec 19;19(1):224.

21. Macaulay IC, Haerty W, Kumar P, Li YI, Hu TX, Teng MJ, et al. G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. Nat Methods. 2015 Jun;12(6):519–22.

22. Hou Y, Guo H, Cao C, Li X, Hu B, Zhu P, et al. Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. Cell Res. 2016 Mar;26(3):304–19.

23. Meers MP, Llagas G, Janssens DH, Codomo CA, Henikoff S. Multifactorial profiling of epigenetic landscapes at single-cell resolution using MulTI-Tag. Nat Biotechnol. 2023 May;41(5):708–16.

24. Swanson E, Lord C, Reading J, Heubeck AT, Genge PC, Thomson Z, et al. Simultaneous trimodal single-cell measurement of transcripts, epitopes, and chromatin accessibility using TEA-seq. eLife. 2021 Apr 9;10.

25. Tedesco M, Giannese F, Lazarević D, Giansanti V, Rosano D, Monzani S, et al. Chromatin Velocity reveals epigenetic dynamics by single-cell profiling of heterochromatin and euchromatin. Nat Biotechnol. 2022 Feb;40(2):235–44.

26. Svensson V, Vento-Tormo R, Teichmann SA. Exponential scaling of single-cell RNA-seq in the past decade. Nat Protoc. 2018 Apr;13(4):599–604.

27. Tran HTN, Ang KS, Chevrier M, Zhang X, Lee NYS, Goh M, et al. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. Genome Biol. 2020 Jan 16;21(1):12.

28. Ahmed KT, Sun J, Cheng S, Yong J, Zhang W. Multi-omics Data Integration by Generative Adversarial Network. Bioinformatics. 2021 Aug 20;

29. Amodio M, Youlten SE, Venkat A, San Juan BP, Chaffer CL, Krishnaswamy S. Single-cell multi-modal GAN reveals spatial patterns in single-cell data from triple-negative breast cancer. Patterns. 2022 Sep;100577.

30. Cao K, Hong Y, Wan L. Manifold alignment for heterogeneous single-cell multi-omics data integration using Pamona. Bioinformatics. 2021 Dec 22;38(1):211–9.

31. Cao Y, Fu L, Wu J, Peng Q, Nie Q, Zhang J, et al. Integrated analysis of multimodal single-cell data with structural similarity. Nucleic Acids Res. 2022 Nov 28;50(21):e121.

32. Demetci P, Santorella R, Sandstede B, Noble WS, Singh R. SCOT: Single-Cell Multi-Omics Alignment with Optimal Transport. J Comput Biol. 2022 Jan;29(1):3–18.

33. Demetci P, Santorella R, Chakravarthy M, Sandstede B, Singh R. SCOTv2: Single-Cell Multiomic Alignment with Disproportionate Cell-Type Representation. J Comput Biol. 2022 Oct 12;

34. Dou J, Liang S, Mohanty V, Cheng X, Kim S, Choi J, et al. Unbiased integration of single cell multi-omics data. BioRxiv. 2020 Dec 11;

35. Gong B, Zhou Y, Purdom E. Cobolt: integrative analysis of multimodal single-cell sequencing data. Genome Biol. 2021 Dec 28;22(1):351.

36. Liu J, Huang Y, Singh R, Vert J-P, Noble WS. Jointly Embedding Multiple Single-Cell Omics Measurements. Algorithms Bioinform. 2019 Sep 3;143.

37. Minoura K, Abe K, Nam H, Nishikawa H, Shimamura T. A mixture-of-experts deep generative model for integrated analysis of single-cell multiomics data. Cell Rep Methods. 2021 Sep 27;1(5):100071.

38. Stark SG, Ficek J, Locatello F, Bonilla X, Chevrier S, Singer F, et al. SCIM: universal single-cell matching with unpaired feature sets. Bioinformatics. 2020 Dec 30;36(Suppl_2):i919–27.

39. Wangwu J, Sun Z, Lin Z. scAMACE: Model-based approach to the joint analysis of single-cell data on chromatin accessibility, gene expression and methylation. Bioinformatics. 2021 Jun 4;

40. Xu Y, Begoli E, McCord RP. sciCAN: single-cell chromatin accessibility and gene expression data integration via cycle-consistent adversarial network. NPJ Syst Biol Appl. 2022 Sep 12;8(1):33.

41. Zhao J, Wang G, Ming J, Lin Z, Wang Y, The Tabula Microcebus Consortium, et al. Adversarial domain translation networks for integrating large-scale atlas-level single-cell datasets. Nat Comput Sci. 2022 May;2(5):317–30.

42. Zuo C, Chen L. Deep-joint-learning analysis model of single cell transcriptome and open chromatin accessibility data. Brief Bioinformatics. 2021 Jul 20;22(4).

43. Granja JM, Corces MR, Pierce SE, Bagdatli ST, Choudhry H, Chang HY, et al. ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. Nat Genet. 2021 Mar;53(3):403–11.

44. Bock C, Boutros M, Camp JG, Clarke L, Clevers H, Knoblich JA, et al. The organoid cell atlas. Nat Biotechnol. 2021 Jan;39(1):13–7.

45. Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, et al. The human cell atlas. eLife. 2017 Dec 5;6.

46. Zhang K, Hocker JD, Miller M, Hou X, Chiou J, Poirion OB, et al. A single-cell atlas of chromatin accessibility in the human genome. Cell. 2021 Nov 24;184(24):5985-6001.e19.

47. Hao Y, Stuart T, Kowalski MH, Choudhary S, Hoffman P, Hartman A, et al. Dictionary learning for integrative, multimodal and scalable single-cell analysis. Nat Biotechnol. 2023 May 25;

48. Arjovsky M, Chintala S, Bottou L. Wasserstein Generative Adversarial Networks. In: Precup D, Teh YW, editors. Proceedings of the 34th International Conference on Machine Learning. PMLR; 2017. p. 214–23.

49. Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville A. Improved training of wasserstein GANs . Red Hook, NY, USA: Curran Associates Inc.; 2017.

50. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat Methods. 2019 Dec;16(12):1289–96.

51. Morelli L, Giansanti V, Cittaro D. Nested Stochastic Block Models applied to the analysis of single cell data. BMC Bioinformatics. 2021 Nov 30;22(1):576.

52. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA. 2005 Oct 25;102(43):15545–50.

53. Zhang B, Srivastava A, Mimitou E, Stuart T, Raimondi I, Hao Y, et al. Characterizing cellular heterogeneity in chromatin state with scCUT&Tag-pro. Nat Biotechnol. 2022 Aug;40(8):1220–30.

54. Wu SJ, Furlan SN, Mihalas AB, Kaya-Okur HS, Feroze AH, Emerson SN, et al. Single-cell CUT&Tag analysis of chromatin modifications in differentiation and tumor progression. Nat Biotechnol. 2021 Jul;39(7):819–24.

55. Mereu E, Lafzi A, Moutinho C, Ziegenhain C, McCarthy DJ, Álvarez-Varela A, et al. Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. Nat Biotechnol. 2020 Jun;38(6):747–55.

56. Paszek P. From measuring noise toward integrated single-cell biology. Front Genet. 2014 Nov 19;5:408.

57. Eling N, Morgan MD, Marioni JC. Challenges in measuring and understanding biological noise. Nat Rev Genet. 2019 Sep;20(9):536–48.

58. Gulati GS, Sikandar SS, Wesche DJ, Manjunath A, Bharadwaj A, Berger MJ, et al. Single-cell transcriptional diversity is a hallmark of developmental potential. Science. 2020 Jan 24;367(6476):405–11.

59. de Pretis S, Cittaro D. Dimensionality reduction and statistical modeling of scGET-seq data. BioRxiv. 2022 Jul 2;

60. Isella C, Brundu F, Bellomo SE, Galimi F, Zanella E, Porporato R, et al. Selective analysis of cancer-cell intrinsic transcriptional traits defines novel clinically relevant subtypes of colorectal cancer. Nat Commun. 2017 May 31;8:15107.

61. Arase M, Tamura Y, Kawasaki N, Isogaya K, Nakaki R, Mizutani A, et al. Dynamics of chromatin accessibility during TGF-β-induced EMT of Ras-transformed mammary gland epithelial cells. Sci Rep. 2017 Apr 26;7(1):1166.

62. Tam WL, Weinberg RA. The epigenetics of epithelial-mesenchymal plasticity in cancer. Nat Med. 2013 Nov 7;19(11):1438–49.

63. Badia-I-Mompel P, Vélez Santiago J, Braunger J, Geiss C, Dimitrov D, Müller-Dott S, et al. decoupleR: ensemble of computational methods to infer biological activities from omics data. Bioinformatics Advances. 2022 Mar 8;2(1):vbac016.

64. Molineris I, Grassi E, Ala U, Di Cunto F, Provero P. Evolution of promoter affinity for transcription factors in the human lineage. Mol Biol Evol. 2011 Aug;28(8):2173–83.

65. Sanchez-Vega F, Mina M, Armenia J, Chatila WK, Luna A, La KC, et al. Oncogenic signaling pathways in the cancer genome atlas. Cell. 2018 Apr 5;173(2):321-337.e10.

66. O'Cathail SM, Wu C-H, Thomas R, Hawkins MA, Maughan TS, Lewis A. NRF2 Mediates Therapeutic Resistance to Chemoradiation in Colorectal Cancer through a Metabolic Switch. Antioxidants. 2021 Aug 28;10(9):1380.

67. Yang M, Davis TB, Pflieger L, Nebozhyn MV, Loboda A, Wang H, et al. An integrative gene expression signature analysis identifies CMS4 KRAS-mutated colorectal cancers sensitive to combined MEK and SRC targeted therapy. BMC Cancer. 2022 Mar 10;22(1):256.

68. Chen X, Teichmann SA, Meyer KB. From Tissues to Cell Types and Back: Single-Cell Gene Expression Analysis of Tissue Architecture. Annu Rev Biomed Data Sci. 2018 Jul 20;1(1):29–51.

69. Preissl S, Gaulton KJ, Ren B. Characterizing cis-regulatory elements using single-cell epigenomics. Nat Rev Genet. 2023 Jan;24(1):21–43.

70. Gawad C, Koh W, Quake SR. Single-cell genome sequencing: current state of the science. Nat Rev Genet. 2016 Mar;17(3):175–88.

71. Brunner A-D, Thielert M, Vasilopoulou C, Ammar C, Coscia F, Mund A, et al. Ultra-high sensitivity mass spectrometry quantifies single-cell proteome changes upon perturbation. Mol Syst Biol. 2022 Mar;18(3):e10798.

72. Labib M, Kelley SO. Single-cell analysis targeting the proteome. Nat Rev Chem. 2020 Mar;4(3):143–58.

73. Baysoy A, Bai Z, Satija R, Fan R. The technological landscape and applications of single-cell multi-omics. Nat Rev Mol Cell Biol. 2023 Jun 6;

74. Lance C, Luecken MD, Burkhardt DB, Cannoodt R, Rautenstrauch P, Laddach A, et al. Multimodal single cell data integration challenge: results and lessons learned. BioRxiv. 2022 Apr 12;

75. Booeshaghi AS, Gao F, Pachter L. Assessing the multimodal tradeoff. BioRxiv. 2023 Apr 18;

76. Ghazanfar S, Guibentif C, Marioni JC. Stabilized mosaic single-cell data integration using unshared features. Nat Biotechnol. 2023 May 25;

77. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. ACM Comput Surv. 2022 Nov 17;

78. Kaminow B, Yunusov D, Dobin A. STARsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq data. BioRxiv. 2021 May 5;

79. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 2012 Sep;22(9):1760–74.

80. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv. 2013 Jan 1;

81. Lab S. scCUT&Tag-pro datasets. Zenodo. 2021;

82. Bertotti A, Migliardi G, Galimi F, Sassi F, Torti D, Isella C, et al. A molecularly annotated platform of patient-derived xenografts ("xenopatients") identifies HER2 as an effective therapeutic target in cetuximab-resistant colorectal cancer. Cancer Discov. 2011 Nov;1(6):508–23.

83. Cittaro D, Lazarević D, Tonon G, Giannese F. Analyzing genomic and epigenetic profiles in single cells by hybrid transposase (scGET-seq). STAR Protocols. 2023 Mar 29;4(2):102176.

84. Wolock SL, Lopez R, Klein AM. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. Cell Syst. 2019 Apr 24;8(4):281-291.e9.

85. Lange M, Bergen V, Klein M, Setty M, Reuter B, Bakhti M, et al. CellRank for directed single-cell fate mapping. Nat Methods. 2022 Feb;19(2):159–70.

86. Wolf FA, Hamey FK, Plass M, Solana J, Dahlin JS, Göttgens B, et al. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. Genome Biol. 2019 Mar 19;20(1):59.

87. Chizat L, Peyré G, Schmitzer B, Vialard F-X. Scaling Algorithms for Unbalanced Transport Problems. arXiv. 2016;

88. Flamary R, Courty N, Gramfort A, Alaya MZ, Boisbunon A, Chambon S, et al. POT: Python Optimal Transport. Journal of Machine Learning Research. 2021;

89. Kulakovskiy IV, Vorontsov IE, Yevshin IS, Sharipov RN, Fedorova AD, Rumynskiy EI, et al. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. Nucleic Acids Res. 2018 Jan 4;46(D1):D252–9.

90. Rand KD, Grytten I, Pavlovic M, Kanduri C, Sandve GK. BioNumPy: Fast and easy analysis of biological data with Python. BioRxiv. 2022 Dec 22;

91. Eide PW, Bruun J, Lothe RA, Sveen A. CMScaller: an R package for consensus molecular subtyping of colorectal cancer pre-clinical models. Sci Rep. 2017 Nov 30;7(1):16618.

Figure legends

Figure 1: Overview of MOWGAN. MOWGAN is a framework for the synthetic generation of multi-omics data. Given unpaired single cell data (*e.g.*, scRNA-seq and scATAC-seq from separate assays), MOWGAN learns both the specificity and the relationship between modalities, returning a new, paired, dataset. MOWGAN takes the data embeddings as input (step 1), sorts the samples according to the spectral embedding (step 2), and imputes the association between cells of different modalities through a Bayesian Ridge Regressor (step 3). Imputed pairs are used to train a WGAN-GP (step 4). The final output is a single cell multi-omics dataset including embeddings generated by the WGAN-GP and feature count matrices reconstructed using kNN regression (step 5).

Figure 2: Enrichment analysis of biologically relevant features. Within each matrix, each column represents the result of preranked GSEA of reconstructed feature values for different cell types, compared to features identified as cell type markers in the input data. In the baseline analysis (top row), any single modality (RNA or ATAC) of PBMC_2 is compared to the same modality of PBMC_1. MOWGAN outputs for PBMC_1 (center row) or PBMC_2 (bottom row) are compared to the relative modality of input data. The left column reports results for scRNA-seq, the right column reports results for scATAC-atac. Colour is proportional to NES value, dot size is proportional to the p-value of the enrichment. We report an accuracy metric (*K*) which is the average rank of NES values on the diagonal.

Figure 3: Comparison with other tools for multimodal single-cell integration. (A) The histograms display LISI and V-measure scores. V-measure is obtained comparing annotated cell types and cell clusters identified in each modality after data integration. Two groups of histograms are reported for PBMC_1 and PBMC_2 datasets separately. Both SCOT and Pamona have been marked by an asterisk since the LISI value has no meaningful interpretation, due to the fact both generate a new embedding in a coordinate system different from the respective input. (B) Histogram displaying the AMI values between cell clusters identified in scRNA-seq and scATAC-seq data after data integration.

Figure 4: Integration of multiple single cell modalities. We report the low dimensional embeddings (PCA) of multiple single cell profiles of PBMC. After data integration, we used MOWGAN to bridge cell type annotations from RNA to ATAC, ADT and H3K27me3 layers. For each modality receiving the annotation we report three embeddings: the input data colored by cell type (top left), the embedding of data generated by MOWGAN (top right) and the input data colored by the transferred cell type (bottom). We report values of accuracy and V-measure for each transfer as well.

Figure 5: Analysis of CRC PDOs. (A) Unified UMAP embeddings of the scRNA-seq (top) and scGET-seq data (bottom) where single cells have been colored by PDO of origin. (B) Unified UMAP embedding as in A, cells have been colored according to their data source, whether it is original input data or synthetic data generated by MOWGAN. (C) Same embedding as in A and B, cells have been colored by the level 4 of the NSBM hierarchy inferred on multimodal data. (D) Scatterplot representing the correlation between Chromatin Accessibility Score, calculated on scGET-seq data, and CytoTrace score, calculated on scRNA-seq data. Each point represents the median value of a score in cell groups at level 0 of the NSBM hierarchy. (E) Distribution of the correlation values between Transcription Factor activities, evaluated on scRNA-seq data, and Total Binding Affinity, evaluated on scGET-seq data. The orange line traces the smoothed profile of the histogram (F) Gene Set Enrichment Analysis of most represented cell groups after integration. The heatmap displays the Net Enrichment Score of preranked GSEA, using Hallmark gene sets. (G) NFE2L2 profiles in integrated data. We report the distribution of TF activity and TBA score of NFE2L2 in scRNA-seq (top) and scGET-seq (bottom) for the most represented cell groups. Next to each violin plot we report the relative unified UMAP embedding, colored by the same score. (H) Profile of gene signatures described for drug resistance phenotypes in CRC. The dotplot reports the score of each signature in the most represented cell groups found in integrated data.

Figure S1: Low dimensional embeddings (PCA) of PBMC_1 data generated by WGAN-GP trained with random sampling. The overall distribution of points indicates that the WGAN-GP learns data topology correctly. However, when cell types are transferred across modalities (from RNA to ATAC and vice versa) the cell group information is lost.

Figure S2: Low dimensional representation of PBMC_1 and PBMC_2 datasets generated by different tools used for benchmarking. Each row reports the embedding produced by a tool, colored by cell type. As described in the text, the analysis of these embeddings reveals spurious cell type transfers across modalities for some tools (*i.e.*, SCOT, scMMGAN and COBOLT). We report one embedding for PBMC_1 (paired data) produced by COBOLT since it projects both modalities into a single embedding. Embeddings produced by scMMGAN appear inverted compared to the reference as it projects cells of one modality (*e.g.,* RNA) to the embedding learned from the other modality (*e.g.,* ATAC) and *vice versa*.

Figure S3: (A) Segmentation profiles of PDO used in the manuscript at 1 Mb resolution. Each row represents a single cell, segmentation values are colored from red (amplification) to blue (deletion). (B) Scatterplot representing the correlation between Chromatin Accessibility Score, calculated on scGET-seq data, and CRIS-B signature score, calculated on scRNA-seq data. Each point represents the median value of a score in cell groups at level 0 of the NSBM hierarchy. (C) Profile of gene signatures described for drug resistance and phenotypes and CRC subtypes. The heatmap reports the score of each signature in cell groups found in integrated data. (D) Profile of Adjusted Mutual Information (AMI) between PDO identity and cell groups identified using the Nested Stochastic Block Models. Orange dots indicate the groups from the integrated data, blue dots indicate groups identified in single modalities. The number near each dot indicates the level of the NSBM hierarchy. The red triangle indicates the level used for the analysis described in the manuscript.

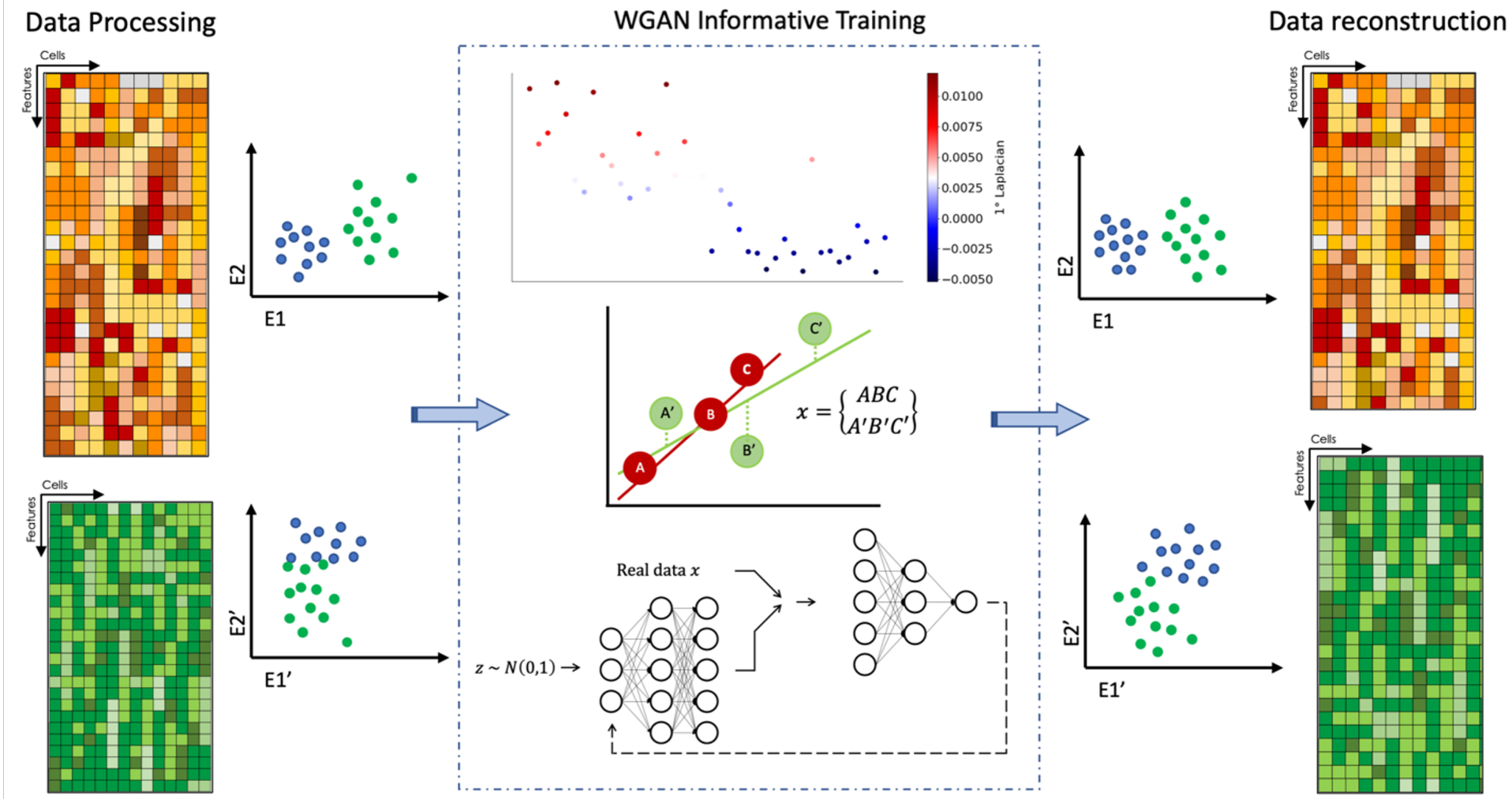
Table legends

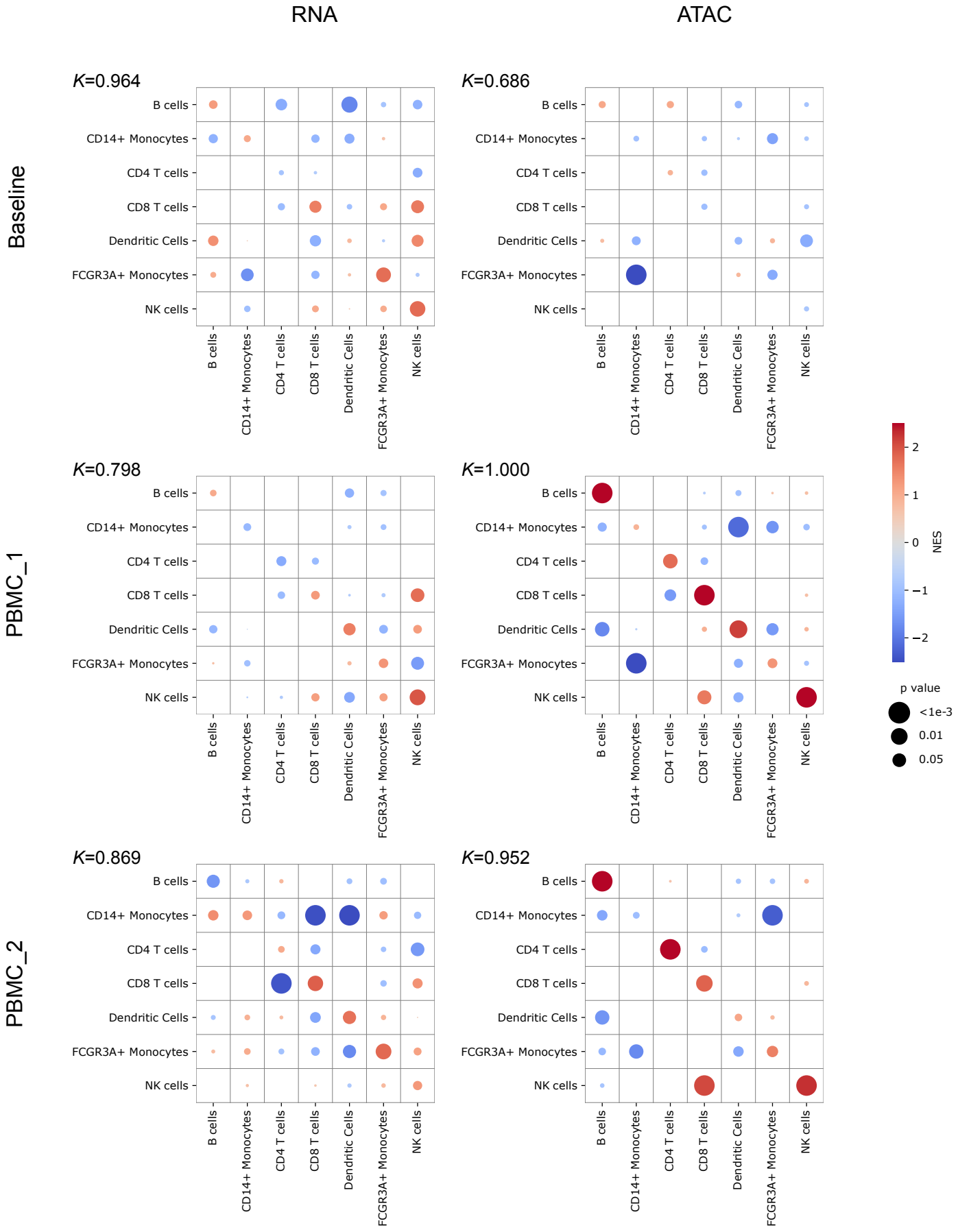
Table S1: Performance of multiple tools in integration of PBMC data. Homogeneity, completeness and V-measure were evaluated between cell types and cell clusters. We report one value for COBOLT for each metric in PBMC_1 as it produces a single embedding for paired data. LISI scores were calculated on the input embedding (PCA) and the embedding generated by the tool. Both V-measure and LISI are reported in Figure 3 in the main text.

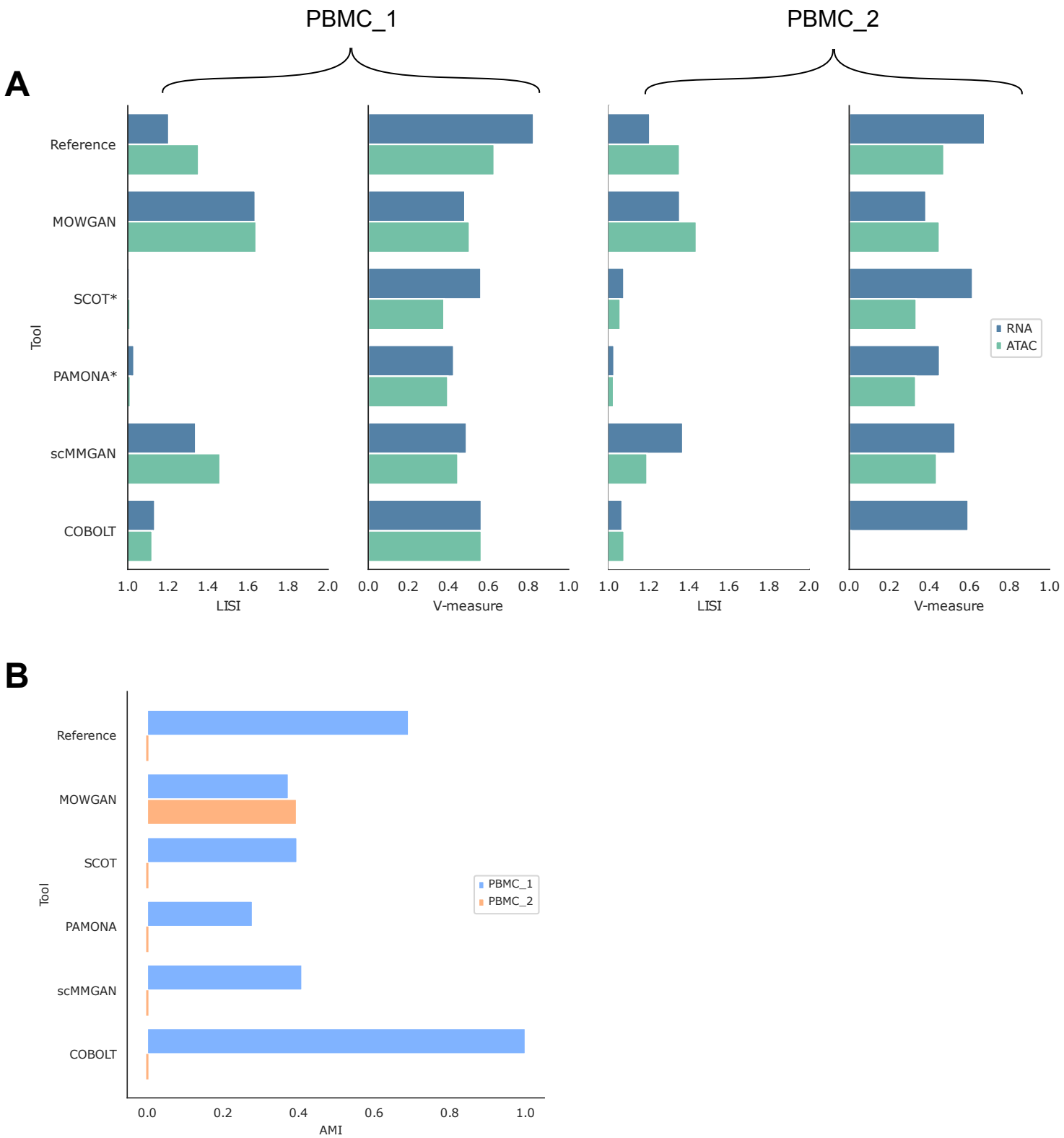
Table S2: Performance of integration of high number of modalities using MOWGAN. The table reports multiple scores calculated after four layer integration (RNA, ATAC, ADP and H3K27me3) of PBMC data, evaluated on the ability to transfer cell type annotation. The first three columns refer to label transfer from RNA to the remaining modalities. The other three columns report the values obtained by transferring labels from any other modality to RNA.

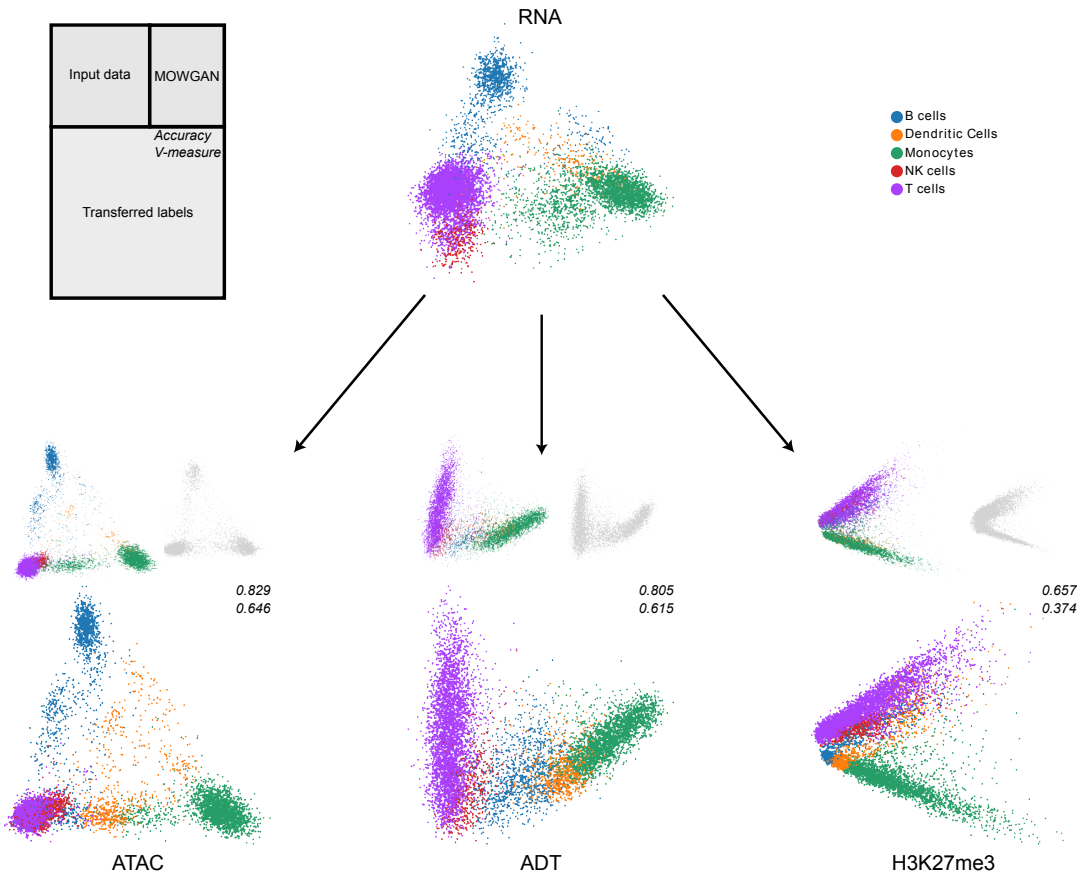
Table S3: Characterization of the PDO dataset. The table reports clinical features of three PDOs (CRC6, CRC17 and CRC39) and the mutational status of cancer genes commonly profiled in CRC. We also report the number of single cells analyzed in the corresponding scRNA-seq and scGET-seq datasets.

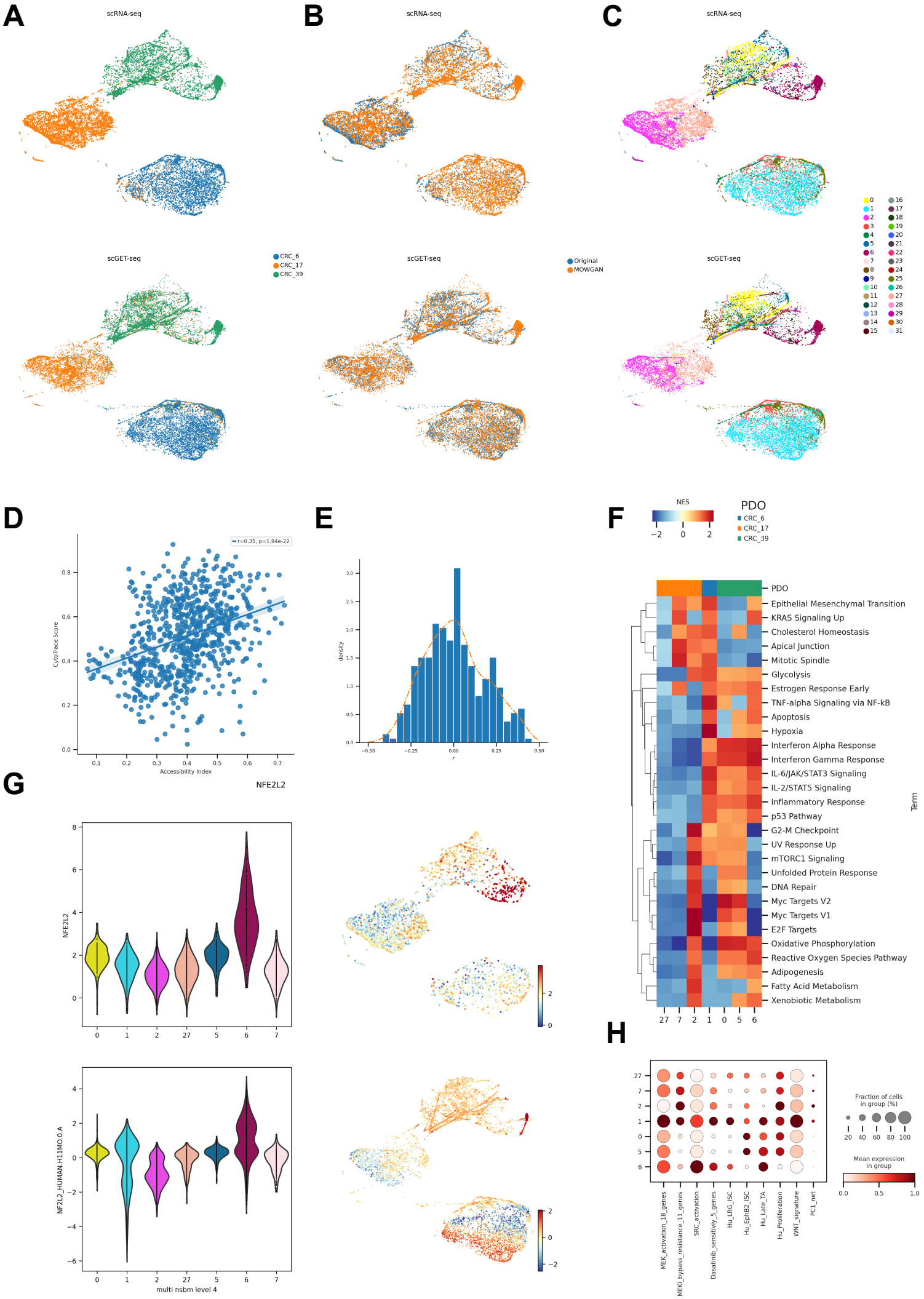
Table S4: Correlation between TF activity, computed on scRNA-seq data, and TBA scores, computed on scGET-seq data, in PDO datasets. We report Pearson's *r* value and the associated *p*-value for transcription factors that were present in both analyses. Data are sorted by decreasing correlation.


Data Processing
WGAN Informative Training
Data reconstruction
Cells
Features
E1
E2
E1'
E2'
0.0100
0.0075
0.0050
0.0025
0.0000
−0.0025
−0.0050
1° Laplacian
A
B
C
A'
B'
C'
$x = \begin{Bmatrix} ABC \\ A'B'C' \end{Bmatrix}$
Real data x
$z \sim N(0,1) \rightarrow$


RNA
ATAC
Baseline
PBMC_1
PBMC_2
K=0.964
K=0.686
K=0.798
K=1.000
K=0.869
K=0.952
B cells
CD14+ Monocytes
CD4 T cells
CD8 T cells
Dendritic Cells
FCGR3A+ Monocytes
NK cells
NES
2
1
0
−1
−2
p value
<1e-3
0.01
0.05


A
PBMC_1
PBMC_2
Reference
MOWGAN
SCOT*
PAMONA*
scMMGAN
COBOLT
Tool
RNA
ATAC
1.0 1.2 1.4 1.6 1.8 2.0
LISI
0.0 0.2 0.4 0.6 0.8 1.0
V-measure
B
Reference
MOWGAN
SCOT
PAMONA
scMMGAN
COBOLT
Tool
PBMC_1
PBMC_2
0.0 0.2 0.4 0.6 0.8 1.0
AMI


Input data
MOWGAN
Accuracy
V-measure
Transferred labels
RNA
B cells
Dendritic Cells
Monocytes
NK cells
T cells
0.829
0.646
0.805
0.615
0.657
0.374
ATAC
ADT
H3K27me3


A
scRNA-seq
scGET-seq
CRC_6
CRC_17
CRC_39
B
scRNA-seq
scGET-seq
Original
MOWGAN
C
scRNA-seq
scGET-seq
D
r=0.35, p=1.94e-22
CytoTrace Score
Accessibility Index
E
density
r
F
NES
PDO
CRC_6
CRC_17
CRC_39
PDO
Epithelial Mesenchymal Transition
KRAS Signaling Up
Cholesterol Homeostasis
Apical Junction
Mitotic Spindle
Glycolysis
Estrogen Response Early
TNF-alpha Signaling via NF-kB
Apoptosis
Hypoxia
Interferon Alpha Response
Interferon Gamma Response
IL-6/JAK/STAT3 Signaling
IL-2/STAT5 Signaling
Inflammatory Response
p53 Pathway
G2-M Checkpoint
UV Response Up
mTORC1 Signaling
Unfolded Protein Response
DNA Repair
Myc Targets V2
Myc Targets V1
E2F Targets
Oxidative Phosphorylation
Reactive Oxygen Species Pathway
Adipogenesis
Fatty Acid Metabolism
Xenobiotic Metabolism
Term
27 7 2 1 0 5 6
G
NFE2L2
NFE2L2
NF2L2_HUMAN.H11MO.0.A
multi nsbm level 4
H
Fraction of cells in group (%)
Mean expression in group
MEK_activation_18_genes
MEKi_bypass_resistance_11_genes
SRC_activation
Dasatinib_sensitivity_5_genes
Hu_LRG_ISC
Hu_EphB2_ISC
Hu_Late_TA
Hu_Proliferation
WNT_signature
PC1_net