SCUOLA DI DOTTORATO
UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

Department of
Informatics, Systems and Communication

PhD program in Computer Science
Cycle XXXV

Integration of heterogeneous single cell data with Wasserstein Generative Adversarial Networks

Valentina Giansanti

N° 854371

Tutor: Prof. Raimondo Schettini

Supervisor: Prof. Marco Antoniotti

Co-supervisor: Dr. Davide Cittaro

Coordinator: Prof. Leonardo Mariani

ACADEMIC YEAR

2021 - 2022

CONTENTS

ABSTRACT

Tissues, organs and organisms are complex biological systems. They are objects of many studies aiming at characterizing their biological processes. Understanding how they work and how they interact in healthy and unhealthy samples gives the possibility to infer, correcting and preventing dysfunctions, possibly leading to diseases.

Recent advances in single-cell technologies are expanding our capabilities to profile at single-cell resolution various molecular layers, by targeting the transcriptome, the genome, the epigenome and the proteome. The number of single-cell datasets, their size and the diverse modalities they describe are continuously increasing, prompting the need to develop robust methods to integrate multiomic datasets, whether paired from the same cells or, most challenging, from unpaired separate experiments. The integration of different sources of information results in a more comprehensive description of the whole system.

Most published methods allow the integration of limited number of omics (generally two) and require assumptions about their inter-relationships. They often impose the conversion of a data modality into the other one (*e.g.*, ATAC peaks converted in a gene activity matrix). This step introduces an important level of approximation, which could affect the following analysis.

Here we propose MOWGAN (Multi Omic Wasserstein Generative Adversarial Network), a deep-learning based framework to integrate multimodal data supporting high number of modalities (more than two) which is also agnostic about their relationships (no assumption is imposed).

We prototyped our approach on public data with available paired and unpaired RNA and ATAC experiments. Each modality is embedded into feature spaces with same

dimensionality across all modalities. This step prevents any conversion between data modalities. The embeddings are used to train a Wasserstein Generative Adversarial Network to understand the coupling between multiple modalities. The output of the generative model can be integrated with the original data and used to bridge information across omics.

Particular attention is reserved to the organization of the input data to train the model in mini-batches. This allows MOWGAN to have a network architecture independent of the number of modalities evaluated. Indeed, the framework was also successfully applied to integrate three (*e.g.*, RNA, ATAC and protein) and four modalities (*e.g.*, RNA, ATAC, protein, histone modifications).

MOWGAN's performance was evaluated both in terms of both computational scalability and biological meaning, the latter being the most important to avoid erroneous conclusion. A comparison was conducted with published methods, concluding that MOWGAN performs better when looking at the ability to retrieve the correct biological identity (*e.g.*, cell types) and associations.

In conclusion, MOWGAN is a powerful tool for multi-omics data integration in single-cell, addressing most of the critical issues observed in the field.

CHAPTER 1

SINGLE CELL TECHNOLOGIES

All organisms, from bacteria to humans, are constituted of at least one cell. In the human body their number is estimated to be 3.72×10^{13} (Wang & Navin, 2015). For more than 150 years biologists have tried to classify and characterize the cell types of different organisms by looking at their functions or at their molecular components (Regev et al., 2017). This proved to be a very difficult task, made even worst by the absence of an agreement on the definition on what a "cell type" or a "cell state" is.

Technological advancement had a great impact on the ability to identify cell types. Starting with the advent of next-generation sequencing (NGS) technologies in the early 2000s, genome-wide sequencing of DNA and RNA became a *de-facto* procedure to investigate biology. More recently, high-throughput single-cell molecular profiling approaches have been introduced. They were chosen as the "technology of the year" by "Nature Methods" in 2013 (Pennisi, 2012). This is a very active field where new methods are proposed every year and existing ones are improved (Regev et al., 2017).

Single-cell profiling is not just a trend. The excitement for this technology is due to the knowledge that has been gained by investigating at such deep resolution. If the difficulties in cell characterization were previously due to analysis performed on bulk tissues samples (*i.e.*, composed of millions of cells), now the tissue heterogeneity can be easily handled. Applications previously impossible to carry out can now be addressed: microorganisms that cannot be cultured can be researched on, as well as the

characterization of earliest differentiation events in human embryogenesis and the tumour microenvironment (Shapiro et al., 2013). This prompted investments in projects aiming at the description of cells in whole organisms. The Human Cell Atlas (Wang & Navin, 2015) initiative ambition is to come out with a map showing the relationships among molecular layers in a tissue/organ. This will work as a reference to study diseases but also to investigate the mechanisms that control cell types differentiation and behaviours.

Most of the efforts were at first dedicated to the *single-assay* sequencing of the transcriptome at single-cell resolution, with the so called single-cell RNA sequencing (scRNA-seq) technologies. Nowadays, methods to profile the genome, epigenome, DNA methylation and 3D organization are also available (Kashima et al., 2020). The possibility to observe these molecular layers in detail boosted the understanding of the processes underling multiple phenomena, like development, aging and disease.

An even more important step in this direction has been achieved through the development of multi-omics approaches, where more than one molecular layer can be evaluated in parallel for a single cell with *multi-assays* experiment. With these approaches, the relationship between layers (which otherwise would remain unknown) can be investigated, revealing regulatory and functional mechanisms in healthy and unhealthy samples (Ogbeide et al., 2022). Thereby, a cell's identity can be truthfully understood.

Different methods have been developed to perform the multi-omics analysis. A recent review (Ogbeide et al., 2022) focused on their categorization based on chronological release and profiled layers. Following the structure of that review, from here on some of the most interesting *multi-assays* methods are introduced. Understanding why they are applied and which layers they profile will be useful to follow this work discussion.

Genome and Transcriptome Sequencing

The first attempts toward multi-omics approaches were dedicated to the joint analysis of the genome and the transcriptome. The interest in the combination of these two elements is motivated by the observation that modifications in the genome have

consequences only if they determine the transcriptome of a different gene. In that case, we observe a modification of the cell's phenotype, leading to cells heterogeneity.

In this category we find G&T-seq (Macaulay et al., 2015). Starting from a physical separation of the RNA from the DNA, the two molecules are first amplified and then sequenced. By analysing the obtained materials, the authors demonstrated for the first time the correlation between chromosomal copy number and gene expression. They were also able to identify the causative genomic modification of a breast cancer cell line, highlighting the potential of single-cell multi-omics analysis in tumour studies.

G&T-seq has also some limitations: the amplification step introduces sequence errors and allelic and locus dropout. Obtaining high coverage data is also expensive, limiting the applications to 100s or 1000s of cells.

Epigenome and Transcriptome Sequencing

The epigenome is the record of chemical changes in the genome determining the functions and regulation of gene expression. Linking genome regulation and gene expression in the same cell can inform on the mechanisms leading to disease development but also on the lineage determination and developmental dynamics.

Two are the approaches used to evaluate epigenetic influence in multi-omics single cell experiments: DNA methylation and chromatin accessibility measured by the assay for transposase-accessible chromatin using sequencing (ATAC-seq) (Buenrostro et al., 2013). Compared to DNA methylation, ATAC-seq is characterized by a higher throughput, that is increased even more with the implementation of ligation-based combinatorial indexing strategy enabling processing of millions of nuclei per experiment (Zhu et al., 2019). SHARE-seq (Ma et al., 2020) improved the sensibility of the combinatorial indexing to evaluate the chromatin potential to predict the gene expression in a cell. This approach was complemented by microfluidic platforms, first by SNARE-seq (S. Chen et al., 2019) and later with its adaptation to 10X Genomics Chromium platform.

Euchromatin and Heterochromatin Sequencing

Open chromatin captured with ATAC-seq is just a small fraction of the chromatin in a cell. The majority is composed of heterochromatin, the compacted part responsible for the genome stability. Single-cell genome and epigenome by transposase sequencing (scGET-seq) (Tedesco et al., 2022) is a recently introduced assay enabling the analysis of both compact and accessible chromatin at single-cell resolution (Annex I). It is built on top of scATAC-seq microfluidic platform by engineering a hybrid transposase to link both open and compacted chromatin. The evaluation of the two chromatin's stages allows to measure epigenetic plasticity in terms of chromatin dynamics (De Pretis & Cittaro, 2022).

Protein and Transcriptome Sequencing

Proteins determine cell's behaviour but due to their biochemical characteristics their measurement is dependent on antibody-based protein detection or mass spectrometry. The antibody-based approach implies that protein detection is feasible only if a specific antibody is included in the panel, otherwise the protein will remain unseen. This represents a critical limitation for the application on these techniques. Nevertheless, simultaneous profiling of proteins and other assays is possible. Cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq) (Stoeckius et al., 2017) is one of the first methods to evaluate RNA and proteins in the same cell. This result is obtained through cell's labelling with antibody-specific oligonucleotide before the amplification step.

Simultaneous profiling of more than two assays

The simultaneous profiling of two omics is only the first step for a knowledge that is omni-omics comprehensive. Even if technological limitations have already been observed (*e.g.*, proteomic profiling), several techniques to profile more than two molecular layers have already been developed. Among them, ASAP-seq (Mimitou et al., 2021) has been proposed for the concurrent analysis of ATAC, protein and mitochondrial DNA; TEA-Seq (Swanson et al., 2021) can be used for the evaluation of ATAC, RNA and

epitopes and scCUT&TAG-pro (B. Zhang et al., 2022) allows for the measurement of histone modifications and proteins abundances on whole cells.

CHAPTER 2

SINGLE CELL INTEGRATION AND LABEL TRANSFERRING

While technological advancement allows to profile multiple assays from the same cells, robust methodological analysis is needed to integrate their information.

Integration of multi-omics data was indicated as one of the grand challenges in single-cell data analysis (Lähnemann et al., 2020).

The problem is worsened by the necessity to face at the same time challenges related to each omic, for which a proper protocol may not have been established. Indeed, if a broad consensus has been reached for analysis of scRNA-seq data, this is not entirely true for modalities that can be considered younger. Analysis of scATAC-seq data is an example of this non-agreement (H. Chen et al., 2019). The community is still working to deliver efficient ways to extrapolate information from the peaks. They are generally converted into a gene activity score, a much-discussed approximation that differently weigh co-accessible elements to a gene's promoter region (Cusanovich et al., 2018). There are also suggestions to use reference set of regions instead of the whole-genome to align the data (Giansanti et al., 2020) (Annex II), allowing for a faster execution of the pipeline without losing the ability to characterize cells. This procedure, called *pseudo-alignment*, was previously introduced for bulk and single-cell RNA-seq data (Bray et al., 2016; Patro et al., 2017).

Why focusing on data integration when addressing single data analysis is already a complex problem? Data integration links different data sources to derive a more comprehensive and biologically meaningful description of the system under analysis. It

also helps to solve the interdependence and the causal relationship among modalities. Its importance is reflected by the number of tools developed and the initiatives arising to collect ideas and guidelines (*e.g.*, the Multi-modal Single-Cell Data Integration Competition, denoted NeurIPS challenge (Lance et al., 2022)).

Depending on the anchors used to combine the data sources, it is possible to distinguish three different integration tasks (Argelaguet et al., 2021) (Figure 1):

1. *Horizontal integration*: the datasets represent one modality, observed in multiple samples, locations or time points. For example, the integration of scRNA-seq data collected from two or more subjects is a horizontal integration task.
2. *Vertical integration*: multiple modalities are assayed from the same cells. The integration of data generated in a SNARE-seq experiment falls in this category.
3. *Diagonal integration*: each dataset represents a different modality for a different sample, without any correspondence between cells and features.

The diagonal integration 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 (Svensson et al., 2018). In addition, many large datasets have been made available for single omic only (Bock et al., 2021; Regev et al., 2017; Schaum et al., 2018; K. Zhang et al., 2021).

While horizontal integration can be treated as a problem of batch correction (Tran et al., 2020), vertical and diagonal settings 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.

So far, two are the most suitable approaches for data integration in single-cell so far: *Manifold Alignment* (MA) and *Deep Learning* (DL). Both methods share the final goal of representing multiple feature sets in a common manifold embedding. While MA approaches try to find a common latent space (manifold) to describe the data, DL develops omic-specific networks to learn and combine a low-dimensional representation of the data.

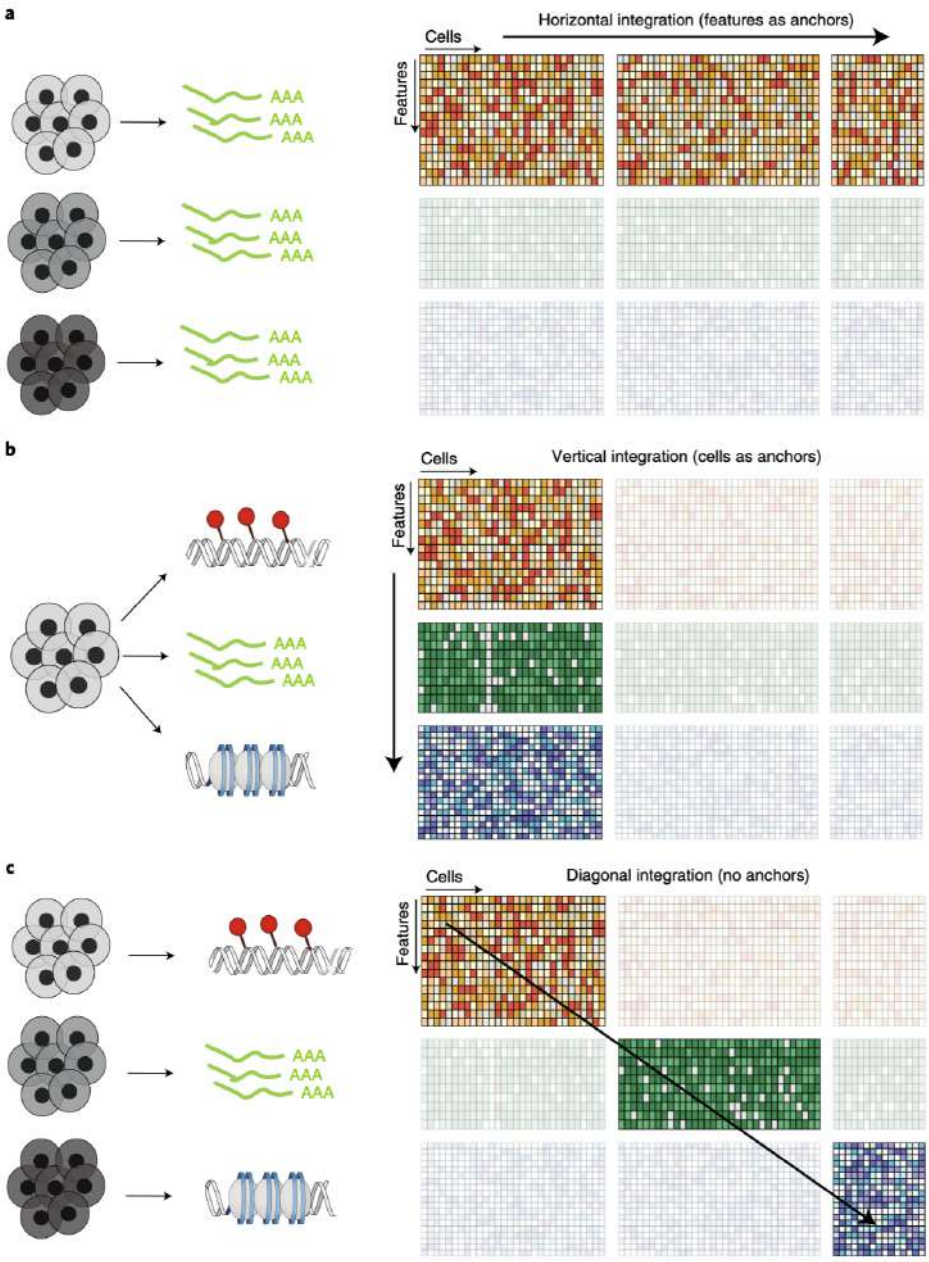


Figure 1 Data integration in single-cell. (a) Horizontal integration, where one modality is observed in multiple samples. (b) Vertical integration, for the same cell more than one modality is collected. (c) Diagonal integration, both cells and features are unmatched between the datasets. *Figure from* (Argelaguet et al., 2021).

Both MA and DL approaches depend on some constrains. Restrictive assumptions are generally applied to the data, like correspondences across the features (*e.g.*, converting ATAC peaks in gene activity scores) and/or cells, or to the data distribution. The major drawback is that these requirements are usually difficult to generalize, making them unfit for datasets where no prior knowledge is available.

Furthermore, they are generally developed to address only two molecular layers (*e.g.*, RNA and ATAC) and scaling to three or more omics could be impractical.

From here on, integration tools are formally introduced. Given the broad literature, they were selected from the ones published on journals. They are divided by methodology (MA or DL), and particular attention is given to DL applications. The integration scenario is also highlighted.

MANIFOLD ALIGNMENT INTEGRATION TOOLS

MMD-MA – MAXIMUM MEAN DISCREPANCY MANIFOLD ALIGNMENT

MMD-MA (J. Liu et al., 2019) is an unsupervised method to align single-cell data when no connection is known *a priori*. The only assumption is that the data points share a manifold structure that can be learned. The learning is made through the optimization of an objective function composed of three elements, designed to address specific tasks. The first term is the maximum mean discrepancy term that enforces the data points to have the same distribution in the latent space. A second term is involved to preserve the structure of the single data between the input and output space. The last term, a penalty term, prevents falling into a trivial solution.

The algorithm was tested on three simulated datasets and one real dataset composed of 61 cells, where both gene expression and methylation where measured. The authors underlined a problem of scalability of their method, as the algorithm was developed to store kernel matrices in memory.

MOFA+ - MULTI-OMICS FACTOR ANALYSIS

MOFA+ (Argelaguet et al., 2020) is a stochastic variational inference framework for the vertical integration of single-cell datasets. Datasets are organized in views and groups, where views are non-overlapping sets of features and groups are non-overlapping sets of samples (*e.g.*, different experimental conditions).

MOFA+ learns a low-dimensional representation of the data where *K* latent factors explain the molecular variability. The *K* factors and the dataset features are

linked by a weight matrix that reveals the importance of each feature in the embedding, even though the model is not able to capture complex non-linearities. MOFA+ output can be used for downstream analysis, such as clustering and trajectory analysis.

BINDSC – BI-ORDER INTEGRATION OF SINGLE-CELL DATA

bindSC (Dou et al., 2020) generates a co-embedding for two unpaired datasets. The framework is based on the bi-order canonical correlation analysis (bi-CCA) algorithm. The inputs are the count matrices of the two modalities, linked by a gene score matrix evaluated by the bi-CCA algorithm. The matrix is optimized to maximize the correlation between the datasets and, for RNA and ATAC integration, it can be initialized as a gene activity matrix. Bi-CCA outputs canonical correlation vectors (CCA) to project cells in a common low-dimensional space.

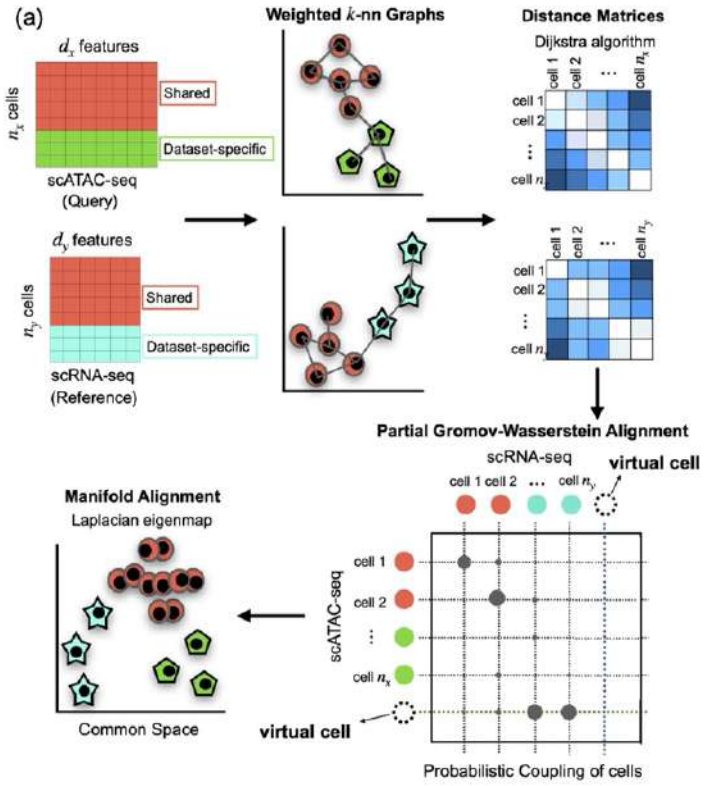


Figure 2 PAMONA overview. Starting from data matrices representing different single-cell layers (*i.e.*, gene expression, chromatin accessibility, etc.), it first constructs the weighted *k*-NN graph for each modality to later minimize the geodesic distance between cells of different datasets. *Figure adapted from* (K. Cao et al., 2021).

PAMONA – A PARTIAL GROMOV-WASSERSTEIN-BASED MANIFOLD ALIGNMENT ALGORITHM

Pamona (K. Cao et al., 2021) is a framework for the alignment of cells from different modalities that preserves the shared and dataset-specific structures. Inputs are the data matrices while outputs are the probabilistic coupling of cells and a common low-dimensional space. To achieve these results, Pamona first computes weighted *k*-NN graphs for each modality and the geodesic distances between the cells in the datasets. The probabilistic coupling matrix is obtained as the result of a partial Gromov-Wassertein optimal transport problem (Figure 2).

The framework was tested on both synthetic and real datasets. The incorporation of cell annotation was proved to increase the performances and, in addition, the application is feasible for both paired and unpaired dataset.

SCOT – SINGLE-CELL MULTI-OMICS ALIGNMENT WITH OPTIMAL TRANSPORT

SCOT (Demetci et al., 2022) is an unsupervised alignment method based on Gromov-Wasserstein optimal transport. It is proposed to address the diagonal integration problem but can be applied for horizontal integration too. SCOT's aim is to find a probabilistic matrix enforcing the coupling between the different modalities. To this end, it preserves the local geometry of the data by computing a *k*-NN graph for each molecular layer. The intra-domain distances are evaluated by constructing a graph distance matrix for each *k*-NN. A probabilistic coupling matrix is later minimized to conserve the intra-domain distances. The same coupling matrix is also used to project one data onto the other, performing the alignment (Figure 3). This is the main design difference between SCOT and MMD-MA: the former projects one modality into the other one, the latter projects both modalities into a latent space.

SCOT was tested on four synthetic datasets (three already used to test MMD-MA and one created with Splatter (Zappia et al., 2017) to simulate a scRNA dataset) and two real datasets. The latter were generated with multiple-assays techniques: SNARE-seq, to profile RNA and ATAC, and scGEM (Cheow et al., 2016), which profile both gene expression and DNA methylation. Results showed that SCOT performed better compared to MMD-MA.

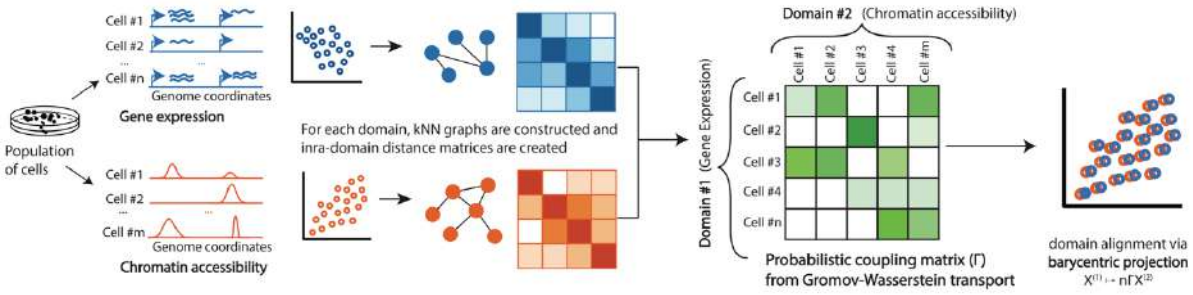


Figure 3 SCOT graphical abstract. Cells are sampled from different datasets. A *k*-NN graph is constructed for each dataset as well as an intra-domain distance matrix. The distance between the two intra-domains distance matrices is minimized. One domain is finally projected into the other one. *Figure from* (Demetci et al., 2022).

DEEP LEARNING INTEGRATION TOOLS

MAGAN: MANIFOLD ALIGNMENT GENERATIVE ADVERSARIAL NETWORK

MAGAN (Amodio & Krishnaswamy, 2018) is a manifold alignment algorithm developed for single and multi-assay experiments. It is one of the first examples of the application of Generative Adversarial Networks (GANs) for single-cell alignment task.

The model is composed of two distinctive GANs, one for each modality. During the training, a loss function with a correspondence loss term is minimized. This term enforces the manifolds to be fully aligned and must be chosen accordingly to the specific application. This implies the possibility to modify the term when some correspondence between the data is known.

The model was tested on synthetic data, on a subset of the MNIST dataset and on biological data. In details, it was tested on cells where both RNA and flow cytometry measurements were available.

SCIM – SINGLE CELL DATA INTEGRATION VIA MATCHING

SCIM (Stark et al., 2020) is a tool for integrating single cell datasets across technologies. It falls in the horizontal integration task. SCIM's assumption is that the cell's distributions remain the same, even when different technologies are used.

The model is based on autoencoders, one for each technology. They learn a technology-invariant latent space that is later aligned with a bipartite matching algorithm (a discriminator acting on the latent space).

sciCAN – Single Cell data Integration via Cycle-consistent Adversarial Network

scRNA and scATAC integration can be performed also with sciCAN (Xu et al., 2021). This tool is based on cycle adversarial networks and was tested on both paired and unpaired datasets. When paired datasets were used, the training was agnostic to the true pairing. scATAC data were converted in a gene activity matrix.

The model is proposed for both representation learning and modality alignment. An encoder projects RNA and ATAC data into a joint low dimensional embedding. The second task is addressed by two discriminators. One is linked to the encoder, and it is trained with adversarial domain adaptation loss. It learns to distinguish between the two modalities. The second discriminator links with a generator that generates chromatin accessibility data starting from RNA embedding.

scMVAE – Single Cell Multimodal Variational Autoencoder

scMVAE was proposed for the vertical integration of scRNA and scATAC data (Zuo & Chen, 2021). Here, to handle the different dimensionality of the two dataset, scATAC data are converted into a gene activity matrix. Data are modelled as a zero-inflated negative binomial (ZINB) distribution to learn a joint embedding. Cells of the two modalities are first clustered together and later passed through an omic-specific decoder. This step allows the reconstruction of the features, accounting for the different normalization processing but also returning denoised data.

Three learning strategies are proposed. The first one, called PoE, aims to estimate the joint posterior by the product of posteriors of each omic data. The second one uses a neural network to embed together the latent representation of the modalities obtained by separate networks. The last one directly concatenates the features of both omics and pass them as input to a network.

Portal – Adversarial domain translation

Portal (Zhao et al., 2021) returns a harmonized representation of cells of different experiment in a shared latent space. It was applied to horizontal integration settings, but it was also tested in the diagonal context. In the latter case, RNA and ATAC data were used, with ATAC data converted in gene activity.

Portal works on the first *p* principal components (PCs) of the data, used to characterize their embeddings. They are inputs to an adversarial domain translation framework composed of both encoders and generators for each dataset. While the encoders remove domain-specific properties of the data, the generators link one domain to the other, introducing in the data generation process the domain-specific effects characterizing the output domain.

scMM – Mixture-of-experts deep generative model for integrated analysis of single-cell multiomics data

scMM (Minoura et al., 2021) is a framework based on mixture-of-experts (MoE) multimodal deep generative models. It was developed to improve the interpretability of joint embeddings and the prediction between modalities. The input is composed of multimodal data, in paired measurement. Each modality is described by a characteristic distribution: ZINB distribution for ATAC-seq data, negative binomial (NB) for scRNA and protein data. The framework is composed by a variational autoencoder (VAE) for each modality. VAEs are trained to learn low-dimensional joint variational posterior factorized by a MoE. scMM was tested on datasets with up to two modalities (*i.e.*, RNA and ATAC data or RNA and protein data). A trained scMM model can be used with just one modality data to infer the missing one, achieving cross-modal generation.

Cobolt – Multimodal Variational Autoencoder based on hierarchical Bayesian generative model

Cobolt (Gong et al., 2021) is a framework proposed for the joint analysis of multi-assay and single-assay data (Figure 4). In the simplest formulation, it can also be used on paired cells only. The output of the model is a joint representation, in a low dimensional embedding, of the inputs. To this end, Cobolt models the data using the Latent Dirichlet

Allocation (LDA) model. A multimodal VAE is used on this representation of the data to transfer learning between the multi-assay and single-assay modalities.

Cobolt assumes that cells can be categorized in *K* types. Each category is filled with the cells whose features contributed to the characterization or activation of that category. All cells will be represented by a vector of activation for category, that lies in a *K*-dimensional space that can be used for downstream analysis.

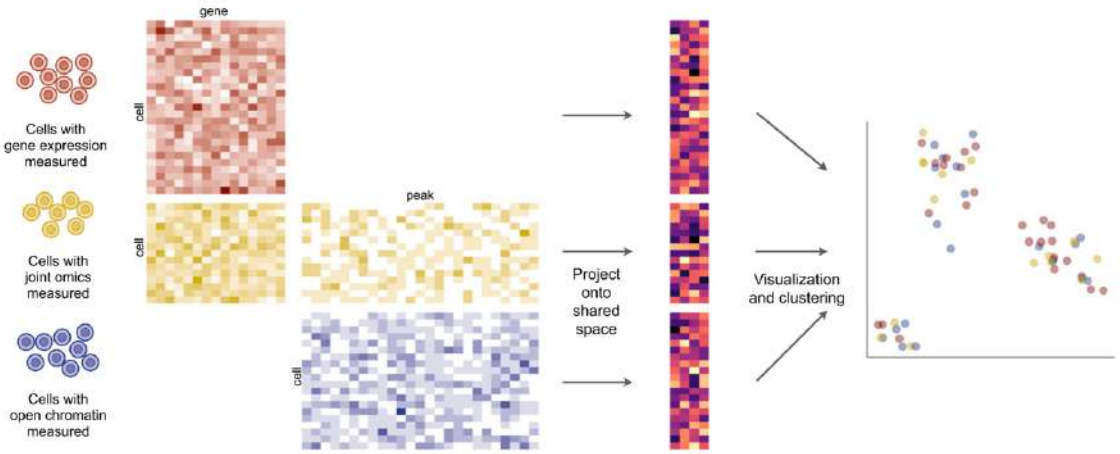


Figure 4 Cobolt workflow. Inputs are datasets from two modalities, in single- and multi-assays. Cobolt learns a low-embedding where all datasets can be represented. This embedding can be used for visualization and clustering of all data together. *Figure from* (Gong et al., 2021).

GLUE – GRAPH-LINKED UNIFIED EMBEDDING

GLUE was one of the first tools implementing graph-based neural network in single-cell (Z. J. Cao & Gao, 2022). Autoencoders, which have modality-specific structure but equal outputs dimensions, are first used to learn a low-dimensional embedding for each modality. GLUE takes advantage of the biological knowledge about the relationship between modalities, and it uses a guidance graph to build a graph variational autoencoder to link the embeddings.

The framework was also successfully tested on the alignment of three modalities: gene expression, chromatin accessibility and DNA methylation.

scMMGAN – Single-Cell Multi-Modal GAN

One of the most recent proposals for the integration of unpaired single cell datasets is scMMGAN (Amodio et al., 2022). The model architecture is based on GANs, one for each modality, with pairwise generators in each mapping direction (Figure 5). The number of generators increase quadratically with the number of modalities included in the analysis.

The main contribution of this work is the introduction of a correspondence geometry loss term in the training phase. This term improves the ability of the network to align cells of the same cell type. This caution is introduced since the network could learn a mapping function between incoherent cell states. The correspondence loss is evaluated as the minimization of the difference between the eigenvectors of the diffusion maps of the datasets.

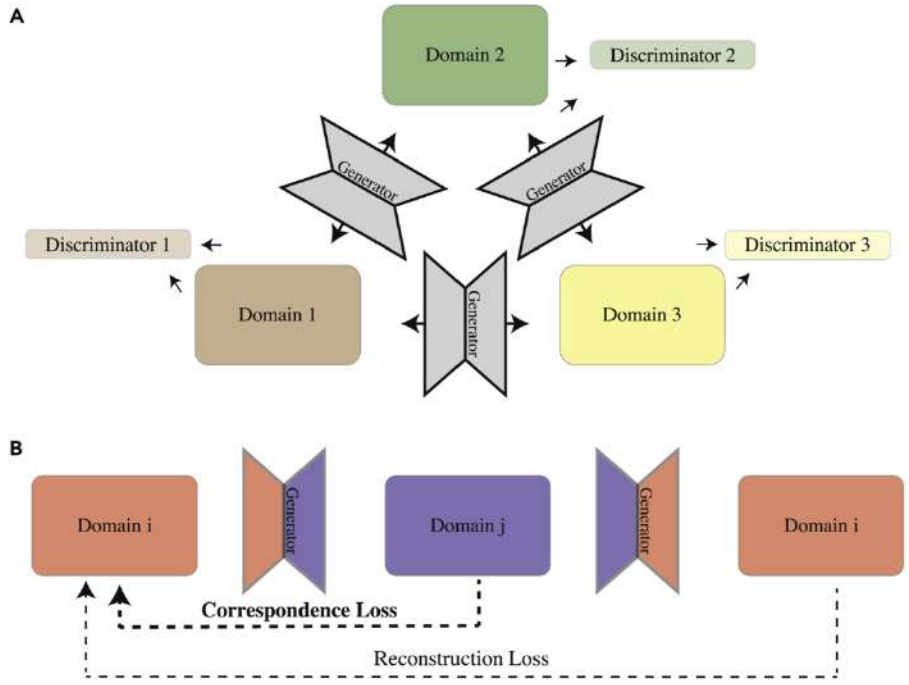


Figure 5 scMMGAN overview. (A) scMMGAN maps data from one modality to another. To this end, the architecture is composed of as many generator-discriminator networks as the number of modalities to analyse. (B) For each domain the training is performed based on a cost function composed of multiple terms. The correspondence loss term is the novelty introduced in scMMGAN. *Figure adapted from* (Amodio et al., 2022).

After this overview regarding the most recent published single cell data integration tool, the main points addressed by MA and DL can be recapitulated as:

- MA tools aim to identify a low dimensional space where the datasets are projected together. To this end, they work on the optimization of a function

composed of multiple terms. Each term takes into accounts specific properties of the datasets and of the alignment problem. Most of the tools work on the integration of two omics.

- In the DL field, autoencoders and GANs are implemented to extrapolate a common embedding for the data, or to achieve modality translation. Most of the tools require common features (*e.g.*, by converting ATAC peaks to a gene activity matrix) or prior biological knowledge. Theoretically, these tools can be easily adapted for the integration of more than two datasets, with the risk of an exponential growth of the number of networks to train.

LABEL TRANSFERRING

As previously introduced, the purpose of integration between different datasets and modalities is to achieve a more comprehensive understanding of the biology. This aspect is translated into the possibility to characterize a sample through its transcriptome, genome, etc. In a multi-assays experiment, as in the 10X Multiome platform, a specific cell is concurrently pictured by gene expression and chromatin accessibility. Every annotation obtained from one modality can be directly *transferred* to the other, as the cell identities are known.

For single-assay dataset, the cell-to-cell correspondence is not available, and *label transferring* approaches must be applied. This step is required every time the information should be passed from one dataset to another (*i.e.*, for every integration setting).

The transferring can be obtained with different approaches, and it is still an active research topic. It is out of the scope of this thesis to compare the available methods. An idea that is recently taking place is to use multi-modal data as a bridge to connect unpaired datasets (Hao et al., 2022). This is also the idea behind Cobolt, where the multiomic dataset is treated as an anchor.

Four datasets are required for the bridge integration (Figure 6). Considering the transferring between RNA and ATAC data, the datasets needed are: one single-assay RNA dataset, one single-assay ATAC dataset, RNA and ATAC multimodal datasets. In this

way the problem is moved from a diagonal integration setting to a horizontal setting, which is much easier to be addressed.

Of course, this approach cannot be applied when multi-omics data are not available. Hence, our idea for MOWGAN: if we can simulate multi-omics data for every combination of modalities we are interested in, we could use the bridge integration to transfer annotations between single-assay datasets.

One more topic that should be mentioned in multi-omics data analysis is the cross-modality translation. This task concerns the forecast of a modality from another one. From a technical point of view, it resembles a standard machine translation or prediction task, allowing non-biological technician to work on it (Lance et al., 2022). Deep learning is mostly applied in this field, with generative model and autoencoders being the most applied architectures.

For more details on modality prediction, please refer to the specific literature (*e.g.*, Martinez-De-Morentin et al., 2021; Wu et al., 2021).

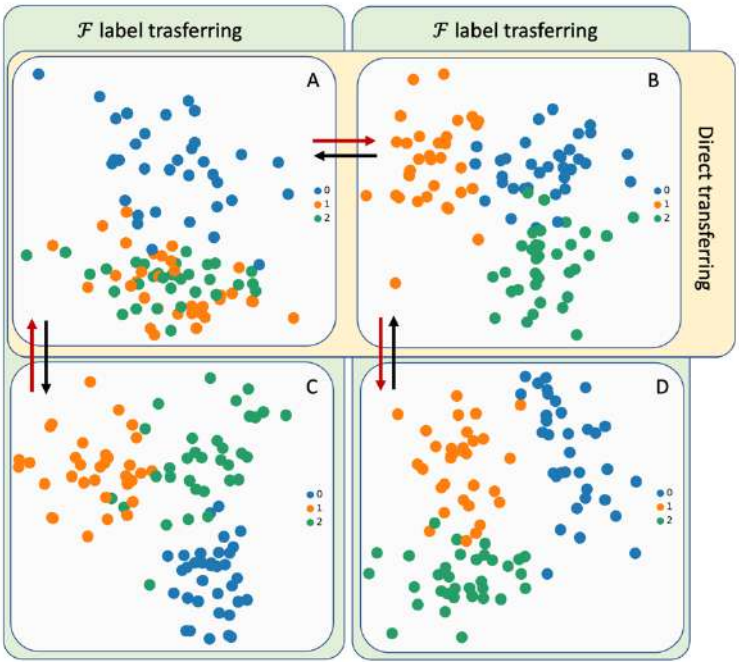


Figure 6 Bridge Integration schema. *A* and *B* are multiomic dataset (e.g., RNA and ATAC from 10x Multiome). Any label can be directly passed from *A* to *B* and vice versa as they represent the same cells (i.e., cell 1 in dataset *A* is cell 1 in dataset *B*). Dataset *C* is a single-modality dataset representing the same modality expressed by *A*. *D* is a single-modality dataset representing the same modality expressed by *B*. The bridge integration uses the multiomic dataset *A* and *B* as intermediary between *C* and *D*. Red arrows show the transferring from *C* to *D*, black arrows the path from *D* to *C*. The label transferring between *C*-*A* and *D*-*B* can be achieved with an appropriate method $\mathcal{F}$.

CHAPTER 3

DEEP LEARNING, BEFORE BIOLOGY

Most of the innovations introduced in Deep Learning are first developed and tested in applications completely different from single cell and, broadly speaking, from biology. One for all, computer vision is the field that benefits the most from Deep Learning advancements and, at the same time, it is the one that has contributed the most to its growth.

The methods and tasks described in the previous chapter have been widely studied in other fields, such as processing of audio and video signals.

Ngiam (Ngiam et al., 2011) proposed a model for audio-visual bimodal feature learning. The model, represented in Figure 7, is composed of two autoencoders sharing a hidden layer. The autoencoders input are data from different modalities (audio and video) and outputs the reconstructed data. This kind of architecture is the same that can be applied for common representation learning and modality translation in single-cell.

This example wants to demonstrate that architectures and solutions proposed for one application can be adapted to work in different contexts. The interoperability and adaptation of Deep Learning algorithms are the reasons why their application is increasing in biology and, especially, in the single-cell field, where the data availability is not an issue.

In this chapter, we will go into the details of generative adversarial networks (GANs) and their evolutions. They are among the most recent models and their use is increasing both in terms of application fields and tasks.

Given their excellent results in generating synthetic data, they were chosen as the core of our framework, MOWGAN.

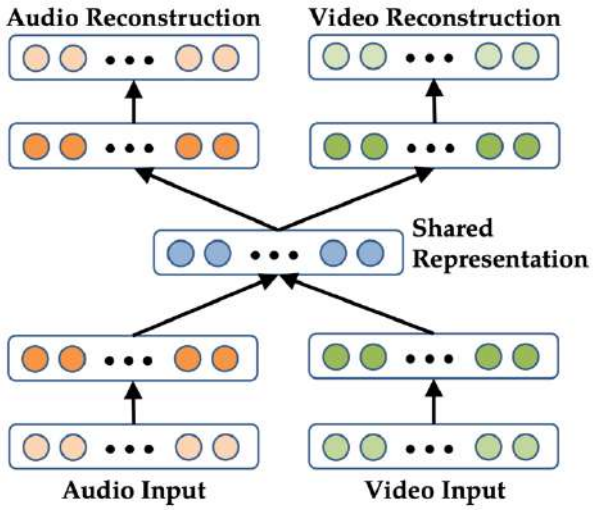


Figure 7 Bimodal Deep Autoencoder. Two autoencoders are trained together on two data modalities (audio and video data). The autoencoders share a hidden layer where the features are concatenated. *Figure from* (Ngiam et al., 2011).

GENERATIVE ADVERSARIAL NETWORKS

Generative Adversarial Networks (GANs) were first introduced in 2014 (Goodfellow et al., 2014). They are a DL-based model to generate synthetic data when annotated datasets are not available. The framework is composed of two different networks, a generator G and a discriminator D, trained together in an adversarial process. The generator aims to produce data resembling the real ones. This concept is translated into the ability of the generator to produce synthetic data with the same distribution of the real data. This distribution is not known *a priori,* but during the training step the generator learns the parameters of the distribution best fitting the data. The discriminator takes in input both the real data and the data generated by G and aims to correctly discriminate between true and fake (synthetic) data. When D is no more able to make such distinction, it means that the data coming from G are plausible examples.

Basically, the training procedure for G is to maximise the probability of D making a mistake. Based on the output of D, G updates the parameters (Figure 8).

In a more formal way, given the data x, we search for the generator's distribution p_g. $p_z(z)$ is an input noise variable and $G(z;\theta_d)$ is the generator mapping from the data z to the distribution with parameters θ_d. $D(x;\theta_d)$ is the discriminator that outputs a single scalar. $D(x)$ is the probability of the data x to come from the real data rather than p_g. D and G are simultaneously trained to minimax the function $V(G,D)$:

$$\min_G \max_D V(G,D) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}\left[\log 1 - D\big(G(z)\big)\right]$$

Goodfellow *et al.* show that, for a fixed G, there is a unique optimal D:

$$D^*{}_G(x) = \frac{p_{data}(x)}{p_{data}(x) + p_g(x)}$$

In their work, it is also shown that G is optimal when $p_g(x) = p_{data}(x)$, *i.e.*, when D is maximally confused. For optimal D, training G is also equivalent to minimizing the Jensen-Shannon divergence between $p_{data}(x)$ and $p_g(x)$.

To avoid overfitting, the authors suggest updating G only once every k steps of optimization for D, where k is a parameter to be set. Backpropagation is used to obtain the gradients.

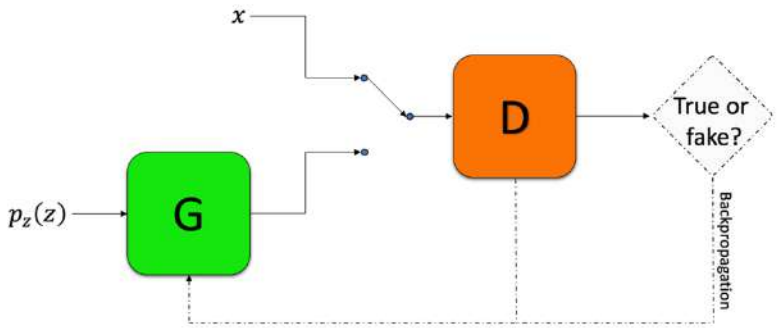


Figure 8 GAN model. $p_z(z)$ is an input vector for the generator network G. x represents the real data that, together with the output of G, is the input of a discriminative network D. D aims to understand if the input is true (coming from the real dataset) or fake (coming from G). Based on D output, the parameters of G and D are updated.

WASSERSTEIN GENERATIVE ADVERSARIAL NETWORKS

Optimizing a GAN is notoriously difficult (Arjovsky & Bottou, 2017; Radford et al., 2015; Salimans et al., 2016). One of the most common problems encountered is the model collapse, with the generator being able to represent only a fraction of the data.

To facilitate the training, it was suggested by many researchers to change the cost function, by replacing the Jensen-Shannon divergence. Arjovsky (Arjovsky et al., 2017) proposed an approximation of the Earth-Mover (EM) distance, or Wasserstein distance, developing the so-called Wasserstein GANs (WGANs).

In a WGAN, the discriminator is converted into a *critic*. The critic scores the realness or falseness of given data. The EM distance represents the cost to optimally transport mass $\gamma(x, y)$ from a distribution $\mathbb{P}_r$ to $\mathbb{P}_g$:

$$W(\mathbb{P}_r, \mathbb{P}_g) = \inf_{\gamma \in \Pi(\mathbb{P}_r, \mathbb{P}_g)} \mathbb{E}_{(x,y)\sim\gamma}[\|x - y\|]$$

where $\Pi(\mathbb{P}_r, \mathbb{P}_g)$ is the set of all joint distributions $\gamma(x, y)$ whose marginals are $\mathbb{P}_r$ and $\mathbb{P}_g$. The Kantorovich-Rubinstein duality (Villani, 2007) tells that:

$$W(\mathbb{P}_r, \mathbb{P}_g) = \sup_{\|f\|_L \leq 1} \mathbb{E}_{x\sim\mathbb{P}_r}[f(x)] - \mathbb{E}_{x\sim\mathbb{P}_\theta}[f(x)]$$

which is the 1-Lipschitz functions $f: \mathcal{X} \to \mathbb{R}$. To find the function f, we can train a neural network with weights $\omega \in \mathcal{W}$ and backpropagate through $\mathbb{E}_{z\sim p(z)}\left[\nabla_\theta f_\omega\left(g_\theta(z)\right)\right]$ as we would do in a normal GAN. To have the weights ω in a compact space $\mathcal{W}$, it is suggested to clamp the weights into a restricted range after each iteration.

Compared to the standard GANs, the EM cost function is more likely to return terms useful to update the generator. The stability during the learning phase, as well as the problem of mode collapse, are improved. WGANs are also more robust to changes in the generator architecture. This is due to the critic being a *k*-Lipschitz continuous function.

WASSERSTEIN GENERATIVE ADVERSARIAL NETWORKS WITH GRADIENT PENALTY

Problems with weight clipping were reported by Gulrajani and colleagues (Gulrajani et al., 2017b). They observed that this procedure can lead to undesired behaviours, like the generation of poor samples or the failure to converge. Analysing the problem, they found out that these behaviours were due to interactions between the weight constraint and the EM distance: if the clipping threshold is not well evaluated, it can lead to vanishing or exploding gradients.

In their paper, Gulrajani *et al*. proposed an alternative solution to weight clipping to help stabilize the training. They thought of penalizing the norm of gradient of the critic with respect to its input. This procedure is called *gradient penalty*. Their objective function is defined as:

$$L = \mathop{\mathbb{E}}_{\tilde{x}\sim\mathbb{P}_g}[D(\tilde{x})] - \mathop{\mathbb{E}}_{x\sim\mathbb{P}_r}[D(x)] + \lambda \mathop{\mathbb{E}}_{\hat{x}\sim\mathbb{P}_{\hat{x}}}[(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2]$$

where $\mathbb{P}_r$ is the data distribution, $\mathbb{P}_g$ is the model distribution and $\tilde{x} = G(z)$, meaning the output of the generator G given some input z from the distribution $p(z)$.

CHAPTER 4

MOWGAN

MOWGAN is a tool for the synthetic generation of coupled, multiomic, single-cell dataset. The development started from the evaluation of scGAN (Marouf et al., 2020) and CoGAN (M.-Y. Liu & Tuzel, 2016). The former is a WGAN for the synthetic generation of scRNA data. scGAN demonstrated the suitability of the GANs class to learn scRNA data properties. CoGAN, instead, is a network to learn the joint distribution of multi-domain images.

We hypothesized that networks in the GAN's family could be applied to other single cell modalities with similar results and, moreover, that they could generate multimodal data if properly trained.

We prototyped our approach on public scRNA and scATAC data from peripheral mononuclear cell (PBMC) for which paired and unpaired experiments exist (from here on, *PBMC_1* and *PBMC_2* dataset respectively).

We applied MOWGAN also on *PBMC_1* and *PBMC_2* combined with other modalities to test a three- and four-layers integration. In this case, PBMC CITE-seq and CUT&Tag-pro data were used.

When available, our tool can also take advantage of prior knowledge on the dataset, typically sample identity, to improve MOWGAN's performance. For demonstration, we used a public dataset of patient derived organoids of Colorectal Cancer (CRC) with batch information.

In this chapter the architectures, data, tests and validations performed during MOWGAN's development until the definition of the final framework will be discussed.

WGAN-GP: ARCHITECTURE

The core component of the framework is a 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 (Arjovsky et al., 2017, Gulrajani et al., 2017). Like all other GANs, the WGAN-GP is composed of two subnetworks, called *generator* and *critic*.

In our idea, a single WGAN-GP is trained with all molecular layers together. After training, the *generator* outputs synthetic dataset where cells are paired, even when the training has been performed with unpaired single-cell data.

The WGAN-GP training is performed in mini-batches, where cells are sampled from the whole dataset (*i.e.*, from each molecular layer). The sampled data are combined in a vector of shape (N, M, C), where N is the number of cells for modality in the mini-batch (generally 256), M the number of modalities evaluated (2, in case of just the RNA and ATAC layers), and C is the number of components in each embedding used for the analysis.

The *generator* is designed with three *convolutional 1D* layers (Conv1D) and two *batch normalization* layers (BN). The *critic* is designed with two Conv1D layers and a Dense layer with 1 unit. All Conv1D layers are characterized by a kernel size of M, stride 1 and the ReLU activation function.

Finally, different optimizers are used for each component: Adam optimizer (Kingma & Ba, 2015) for the *generator* with $\mathrm{learning\ rate} = 0.001$, $\mathrm{beta\_1} = 0.5$, $\mathrm{beta\_2} = 0.9$, $\mathrm{epsilon} = 1e-07$ and the AMSgrad option (Reddi et al., 2018); RMSprop optimizer with $\mathrm{learning\ rate} = 0.0005$ for the *critic* (Hinton et al., 2012).

After the training, the generator returns a dataset still in the shape of (N, M, C). A *k*-NN regressor for each modality is used to reconstruct the count matrix.

DATASETS

Public datasets were used to test and develop MOWGAN. They are introduced in the following.

*PBMC_1 AND PBMC_2**

Public data for 10k PBMC were downloaded from 10x Genomics web site (https://www.10xgenomics.com/resources/datasets/pbmc-from-a-healthy-donor-granulocytes-removed-through-cell-sorting-10-k-1-standard-2-0-0, https://www.10xgenomics.com/resources/datasets/10k-human-pbmcs-3-ht-v3-1-chromium-x-3-1-high, https://www.10xgenomics.com/resources/datasets/10-k-peripheral-blood-mononuclear-cells-pbm-cs-from-a-healthy-donor-1-standard-1-2-0). After the preprocessing, the RNA and ATAC paired layers were further reduced to include only the common cells. Table 1 summarizes the number of cells and features per dataset after preprocessing.

Table 1 Dataset description. *PBMC_1* (paired dataset) and *PBMC_2** (unpaired dataset). In the paired dataset, 8320 cells were analysed for RNA and ATAC, with respectively 4696 genes and 31048 regions. The RNA unpaired dataset is instead composed of 9835 cells and 2657 genes. The ATAC unpaired dataset has 6989 cells for 30524 regions.

	PBMC_1		PBMC_2*	
	RNA	ATAC	RNA	ATAC
Cells	8320		9835	6989
Features	4696	31048	2657	30524

A 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. The intra-modal transition (*i.e.*, from RNA *PBMC_1* to RNA *PBMC_2**, and from ATAC *PBMC_1* to ATAC *PBMC_2**) was performed

by Schist label transfer function (Morelli et al., 2021) (Annex III). Figure 9 shows the cell types on the dataset's UMAP representation.

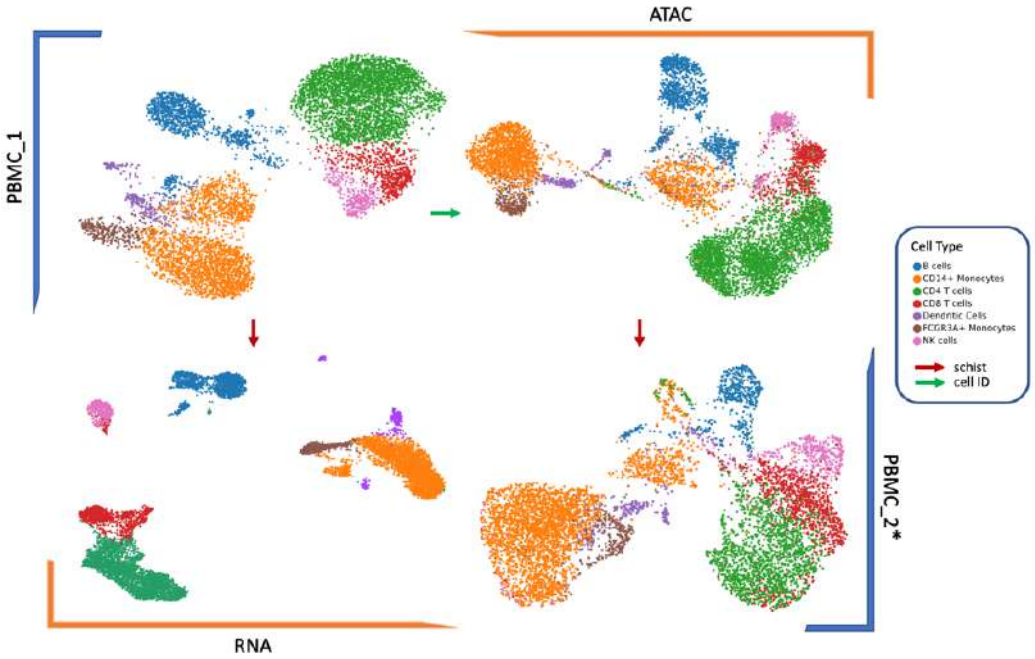


Figure 9 Cell annotation. Cell type was defined on RNA *PBMC_1* and transferred by cell ID to the ATAC *PBMC_1* layer. Schist's label transfer function was applied to transfer the annotation from RNA *PBMC_1* to RNA *PBMC_2** and from ATAC *PBMC_1* to ATAC *PBMC_2** (intra-modalities transfers).

PBMC CITE-SEQ

Public PBMC CITE-seq data were downloaded from the Scanpy tutorial (https://scanpy-tutorials.readthedocs.io/en/latest/cite-seq/pbmc5k.html).

The RNA processed dataset is made of 5076 cells for 3081 genes. In the protein layer, the same cells observed in RNA were selected, subsetting the dataset to 5076 cells for 32 antibodies.

PBMC CUT&TAG-PRO

The scCUT&Tag-pro dataset (B. Zhang et al., 2022) is available at https://zenodo.org/record/5504061. It is composed by six histone modification coupled with the abundance of 173 surface proteins (Table 2). Cell type annotation was already available. H3K27me3 and H3K4me2 were selected to test MOWGAN as markers of silencing and activation. No filtering was applied.

Table 2 scCUT&Tag dataset description.

	H3K27ac	H3K27me3	H3K4me1	H3K4me2	H3K4me3	H3K9me3
Cells	15609	8232	12770	9575	10386	8304
Regions	52981	45383	37665	24707	21758	59416

CRC DATASET

scRNA and scGET-seq data for three human-derived colorectal cancer organoids are available at E-MTAB-9659. The RNA layer is made of 6486 cells for 772 genes, while GET layer is made of 14308 cells and 57094 genomic regions. Details are available in Table 3.

Table 3 CRC dataset. Cells for batch.

	CRC_6	CRC_17	CRC_39
RNA	517	4864	1105
GET	4310	4998	5000

PREPROCESSING

To ensure the benchmarks' comparability, standard processing was applied to filter and normalized the PBMC data. For RNA, cells with $200 < \text{expressed genes} < 20.000$, less than 40% of mitochondrial genes and genes present in more than 10 cells were selected. Counts per cell were normalized and log-transformed. Genes were selected to include only the highly variable ones.

For the ATAC analysis, the genome was segmented in windows of 5000 bp. Count matrices were generated using peak_counts.py script from the scatACC repository (https://github.com/dawe/scatACC). Cells with more than 30% of captured regions and regions common to more than 80% of cells were selected. Data were normalized and log transformed. Highly variable regions were selected.

Proteins count matrices were normalized using Centered Log Ratio and log transformed. No filtering was applied, apart from the selection of cells common to the matched layers (*i.e.*, RNA in CITE-seq dataset).

Histone data were normalized and log transformed. No filtering was applied.

CRC data were pre-processed as illustrated in the original paper (Tedesco et al., 2022).

BASELINE

To estimate the shared information in real multiomic dataset, clusters were identified in *PBMC_1* using the Leiden algorithm (Traag et al., 2019) with resolution 0.5 on both RNA and ATAC layers. We observed a slight discrepancy between the annotations. They have adjusted mutual information (AMI), ranging between 0 and 1, equal to 0.68, marking the upper limit we may expect also for shared information in synthetic coupled data.

We also evaluated the bridge integration between our four datasets. To this end, annotations A and B (Leiden clusters with resolution 0.5) were defined in the *PBMC_2** RNA and ATAC layers respectively. A was transferred to ATAC *PBMC_2** through *PBMC_1*. The AMI between A and B is 0.61.

The same procedure was applied on the *PBMC_1* dividing each layer into two sub-datasets. In this case, $AMI = 0.87$.

Next, the Local Inverse Simpson's Index ($LISI$) score (Korsunsky et al., 2019) was evaluated. This index measures the integrability between multiple batches of the same modality and requires the application of Harmony (Korsunsky et al., 2019). The $LISI$ score ranges between 1 and the number of batches considered, where 1 indicates a poor integration.

Figure 10 shows first two Harmony's principal components coloured by cell types (defined in each dataset) and batches. RNA layers have a good overlap ($LISI = 1.2$), which can be appreciated from the relative plots where cell types are also in line. The same is not entirely true for the ATAC counterpart. Even if the ATAC $LISI$ score is higher ($LISI = 1.35$) compared to RNA, the algorithm is not able to correctly integrate the B cells, which in case of *PBMC_2** are mixed with other cell types, while in *PBMC_1* form a separate cluster.

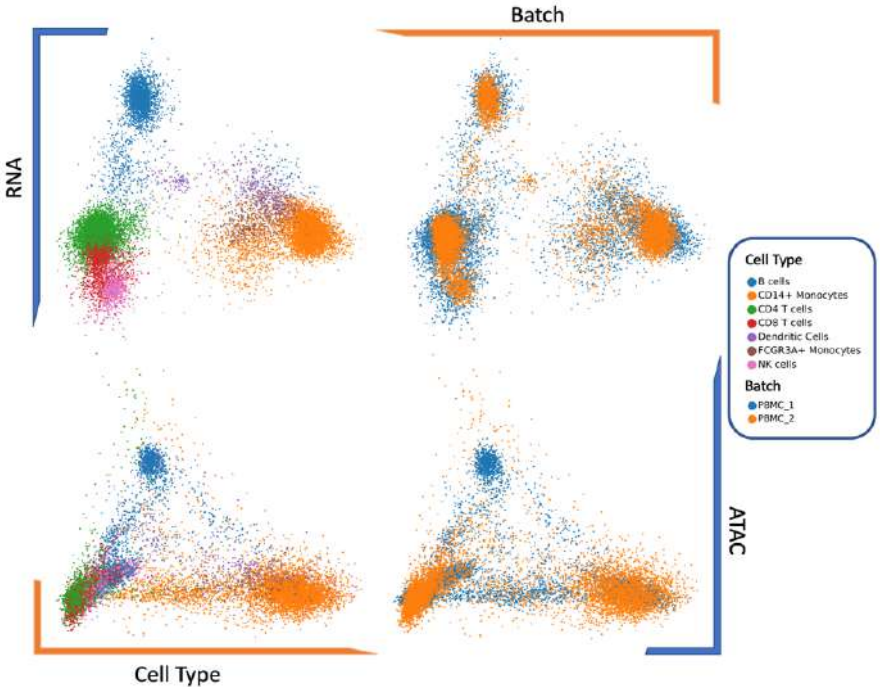


Figure 10 Real datasets integration. The first and second rows show the integration between two RNA batches and two ATAC batches respectively. Column one is coloured by the cell type annotation that was already defined in all datasets, while the second column is coloured by their batch name.

To check the biology shared between different batches, a Gene Set Enrichment Analysis ($GSEA$) (Subramanian et al., 2005) was performed. The $GSEA$ is computed on *PBMC_2**. For each cell type, the reference set of features (genes or regions) is defined in *PBMC_1*: selected features have $\text{adjusted p-value} < 0.005$ and $\text{logfoldchange} > 0$. Cell type clusters in *PBMC_2** were tested against all clusters in *PBMC_1*.

As expected, cell types in RNA *PBMC_2** are clearly enriched for the gene set characterizing the same cell type in RNA *PBMC_1* (Table 4). This is not the case for ATAC (Table 5). For example, the NK ATAC *PBMC_2** is especially enriched for regions describing the CD8 ATAC *PBMC_1* cluster.

From these observations, derived on real data, we can conclude that biological and topological properties are not entirely conserved even between batches of the same molecular layer.

Table 4 RNA GSEA analysis. The GSEA is performed on RNA *PBMC_2** while taking as a reference the genes per cell type in RNA *PBMC_1* with p-value < 0.005 and logfoldchange > 0. For each dataset/cell type comparison, the table reports the normalized enrichment score (NES), the p-value (Pval) and the dales discovery rate (FDR).

PBMC_2* / PBMC_1	NK	FCGR3A+ Monocytes	Dendritic Cells	CD8 T cells	CD4 T cells	CD14+ Monocytes	B cells
NK	NES=2.092 Pval=0.000 FDR=0.000	NES=0.529 Pval=0.970 FDR=0.983	NES=0.903 Pval=0.646 FDR=0.792	NES=1.627 Pval=0.000 FDR=0.004	NES=inf Pval=Nan FDR=0.000	NES=inf Pval=Nan FDR=0.000	NES=-1.184 Pval=0.167 FDR=0.125
FCGR3A+ Monocytes	NES=1.391 Pval=0.084 FDR=0.171	NES=1.651 Pval=0.000 FDR=0.018	NES=1.117 Pval=0.270 FDR=0.447	NES=0.855 Pval=0.708 FDR=0.737	NES=-1.640 Pval=0.000 FDR=0.000	NES=1.183 Pval=0.150 FDR=0.373	NES=0.889 Pval=0.596 FDR=0.823
Dendritic Cells	NES=1.157 Pval=0.313 FDR=0.397	NES=0.982 Pval=0.510 FDR=0.634	NES=1.675 Pval=0.000 FDR=0.000	NES=1.061 Pval=0.404 FDR=0.538	NES=-0.810 Pval=0.750 FDR=0.850	NES=0.625 Pval=1.000 FDR=0.957	NES=1.284 Pval=0.040 FDR=0.260
CD8 T cells	NES=1.714 Pval=0.000 FDR=0.000	NES=-0.937 Pval=1.000 FDR=0.625	NES=-1.346 Pval=0.000 FDR=0.250	NES=1.879 Pval=0.000 FDR=0.000	NES=0.479 Pval=1.000 FDR=0.996	NES=-1.165 Pval=0.200 FDR=0.500	NES=-1.144 Pval=0.250 FDR=0.333
CD4 T cells	NES=-1.562 Pval=0.000 FDR=0.062	NES=-1.456 Pval=0.000 FDR=0.047	NES=-1.639 Pval=0.077 FDR=0.047	NES=1.007 Pval=0.510 FDR=0.492	NES=1.997 Pval=0.000 FDR=0.000	NES=-1.868 Pval=0.000 FDR=0.000	NES=inf Pval=Nan FDR=0.000
CD14+ Monocytes	NES=0.839 Pval=0.780 FDR=0.830	NES=0.969 Pval=0.550 FDR=0.922	NES=0.858 Pval=0.850 FDR=1.000	NES=-1.970 Pval=0.000 FDR=0.000	NES=-1.508 Pval=0.000 FDR=0.000	NES=1.145 Pval=0.110 FDR=0.506	NES=inf Pval=Nan FDR=0.000
B cells	NES=-0.692 Pval=0.929 FDR=0.909	NES=1.031 Pval=0.440 FDR=0.566	NES=1.262 Pval=0.030 FDR=0.230	NES=-1.513 Pval=0.143 FDR=0.136	NES=-1.809 Pval=0.000 FDR=0.000	NES=0.600 Pval=0.960 FDR=0.960	NES=1.532 Pval=0.000 FDR=0.041

Table 5 ATAC GSEA analysis. The GSEA is performed on ATAC *PBMC_2** while taking as a reference the regions per cell type in ATAC *PBMC_1* with p-value < 0.005 and logfoldchange > 0. For each dataset/cell type comparison, the table reports the normalized enrichment score (NES), the p-value (Pval) and the dales discovery rate (FDR).

PBMC_2* / PBMC_1	NK	FCGR3A+ Monocytes	Dendritic Cells	CD8 T cells	CD4 T cells	CD14+ Monocytes	B cells
NK	NES=1.638 Pval=0.000 FDR=0.000	NES=Nan Pval=Nan FDR=Nan	NES=-1.408 Pval=0.133 FDR=0.133	NES=0.853 Pval=0.820 FDR=0.820	NES=inf Pval=Nan FDR=0.000	NES=0.793 Pval=0.729 FDR=0.729	NES=inf Pval=Nan FDR=0.000
FCGR3A+ Monocytes	NES=-0.987 Pval=0.333 FDR=0.333	NES=2.096 Pval=0.000 FDR=0.000	NES=1.329 Pval=0.000 FDR=0.000	NES=-2.248 Pval=0.000 FDR=0.000	NES=inf Pval=Nan FDR=0.000	NES=1.835 Pval=0.000 FDR=0.000	NES=inf Pval=Nan FDR=0.000
Dendritic Cells	NES=Nan Pval=Nan FDR=Nan	NES=Nan Pval=Nan FDR=Nan	NES=1.203 Pval=0.241 FDR=0.241	NES=Nan Pval=Nan FDR=Nan	NES=Nan Pval=Nan FDR=Nan	NES=Nan Pval=Nan FDR=Nan	NES=1.203 Pval=0.241 FDR=0.241
CD8 T cells	NES=1.992 Pval=0.000 FDR=0.000	NES=Nan Pval=Nan FDR=Nan	NES=-1.410 Pval=0.143 FDR=0.143	NES=1.196 Pval=0.070 FDR=0.070	NES=0.433 Pval=1.000 FDR=1.000	NES=-1.206 Pval=0.211 FDR=0.211	NES=inf Pval=Nan FDR=0.000
CD4 T cells	NES=1.229 Pval=0.179 FDR=0.179	NES=-1.306 Pval=0.000 FDR=0.000	NES=-2.268 Pval=0.000 FDR=0.000	NES=0.737 Pval=0.980 FDR=0.980	NES=1.052 Pval=0.250 FDR=0.250	NES=-1.803 Pval=0.100 FDR=0.100	NES=0.340 Pval=1.000 FDR=1.000
CD14+ Monocytes	NES=-1.485 Pval=0.075 FDR=0.075	NES=1.877 Pval=0.000 FDR=0.000	NES=1.254 Pval=0.204 FDR=0.204	NES=-1.334 Pval=0.021 FDR=0.021	NES=Nan Pval=Nan FDR=Nan	NES=1.626 Pval=0.030 FDR=0.030	NES=-3.199 Pval=0.000 FDR=0.000
B cells	NES=0.774 Pval=0.774 FDR=0.774	NES=1.141 Pval=0.312 FDR=0.312	NES=1.489 Pval=0.000 FDR=0.000	NES=inf Pval=Nan FDR=0.000	NES=inf Pval=Nan FDR=0.000	NES=1.230 Pval=0.080 FDR=0.080	NES=1.194 Pval=0.000 FDR=0.000

TRAINING AND VALIDATION

NAÏVE TRAINING

MOWGAN was first applied to *PBMC_2**. Given unpaired molecular dataset (*e.g.*, RNA and ATAC single assays dataset), a naïve training is not sufficient for the WGAN-GP to induce the coupling. The WGAN-GP learns to reproduce the data in the order they are given in input. Therefore, if the mini-batch is composed of cells randomly sampled from the whole dataset, the WGAN-GP will learn the internal structure of the single layer but not how to match them. Indeed, if the Leiden clusters are transferred between data generated with this training strategy, the annotation is completely mixed in the receiving embedding (Figure 11). This is confirmed by $AMI = 0.05$.

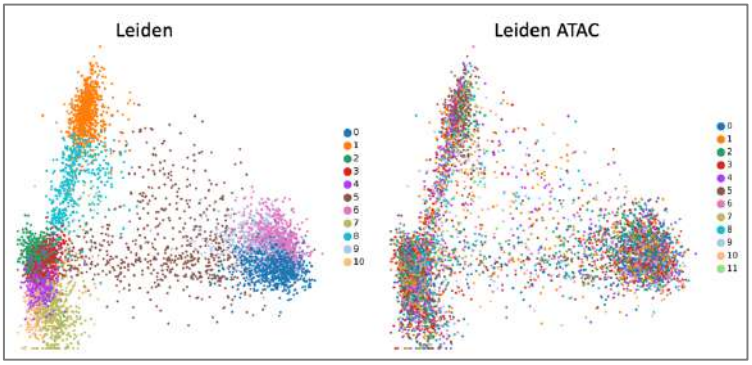


Figure 11 MOWGAN naïve training. When the training is performed on data randomly sampled from the whole dataset, the network can learn only the structure of the molecular layers, but it does not induce the pairing. This is confirmed by the extremally low *AMI* score (*AMI*=0.05) between the Leiden clusters. In the picture, the PCs generated by MOWGAN for the RNA layer are coloured by the Leiden clusters identified on the RNA itself (first column), and by the clusters transferred from ATAC (Leiden ATAC). On RNA's PCs, the transferred clusters are completely mixed, meaning the cells in the synthetic dataset don't have the correct association.

INFORMED TRAINING

From these observations, we understood the mini-batches composition to be a critical aspect. Concerning this, our intuition was that the data, within and between datasets, should be organized to have cells representing the same biology in similar positions.

Our first attempt was to sort the data based on hierarchical clustering, obtained through the Ward's linkage. During the checks on the generated data, a problem was noted with the cell type annotation on the ATAC layer. Indeed, the cell type transferred

from the original data and the one passed through the bridge integration (*i.e.*, from RNA) were inverted (Figure 12). This means the hierarchical clustering returns a reverse order for the RNA and ATAC layers.

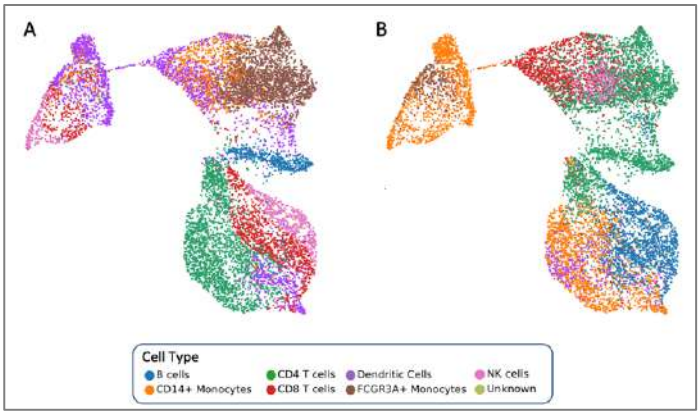


Figure 12 Cell type inversion. (A) Cell type transferred with Schist from the true ATAC dataset to the synthetic ATAC. (B) Cell type transferred from the true RNA layer to the synthetic ATAC with the bridge integration. The two cell types are inverted, meaning the association induced between RNA and ATAC in the generated dataset is incorrectly done.

We moved on to test the Laplacian Eigenmaps (LE), a non-linear dimensionality reduction technique that preserves the local geometry of the data. We hypothesized that the local neighbourhood of each cell is roughly conserved across modalities. This property should be reflected in the structural properties of the underlying *k*-NN graphs, hence in the eigenvalues of the graph Laplacian. LE secures that cells close in the original space will be close also in the reduced space. By sorting each dataset by the first component of its LE, we aim to maximise the probability to have cells representing the same biology (*e.g.*, cell type) in a similar order. Still, the ordering is not guaranteed to be conserved between datasets. Therefore, the selection of the (N, M) cells for the WGAN-GP's training is done in an iterative way that include an additional control.

First, a mini-batch from one modality (generally RNA) is selected, and a Bayesian ridge regressor is trained on the mini-batch embedding and the corresponding eigenvectors. Following this step, n mini-batches ($n = 50$) are selected from the second modality. The trained Bayesian ridge regressor is applied on all n mini-batches and their eigenvectors. Based on the returned score, the n_i (with $1 < i < n$) mini-batch with the best score is selected, which represent the batch in the second modality (*e.g.*, ATAC) with local characteristics more akin to what was already selected in the RNA mini-batch.

This should further maximise the probability of selecting cells from RNA and ATAC representing the same cell types.

Finally, the RNA mini-batch and the best-scored ATAC mini-batch are combined to form the vector of shape (N, M, C) used to train the WGAN-GP.

AWARE TRAINING

Prior information on the dataset can be used to guide the training. For example, if batches are shared (even partially) between the layers, batch-specific models can be trained. The final, synthetic, coupled dataset would be the concatenation of the data generated by all trained models.

In addition, when the datasets are already annotated (by cell type or other characteristics), the initial sorting can be done directly on such annotation. In this way, we will skip all the iterative, inferring phase, also speeding up the training.

HYPERPARAMETER TUNING

A grid search policy was applied to test the dependency of the model from i) the filters f in the Conv1D layers and ii) the number of components C characterizing the embeddings (here, PCs).

To simplify our evaluation of the results, we decided to discard the ATAC *PBMC_2** layer as it was difficult to integrate with ATAC *PBMC_1*. We defined another dataset, *PBMC_2*, made of RNA *PBMC_2** and ATAC *PBMC_1*. Therefore, *PBMC_2* is still an unpaired dataset.

The grid search was performed on *PBMC_1* and *PBMC_2*. We tested for filters $f = \{8,32,64,128,256,512\}$ and components $C = \{5,10,15\}$. A total of 216 models were trained with the informed training policy. The training time for single model was $\sim 3h$ every 100.000 epochs.

Within the trained models, 51 did not produced data due to *NANs* in the *generative* and *critic* loss function and 1 produced just a point in the embedding space. We believe the *NANs* to be the result of a bad initialization of the network's weights and that they could disappear (at least in some models) if the training is re-initialized. We decided to not proceed with the re-training as we already collected enough models.

Therefore, qualitative metrics were evaluated for a total of 164 models. For them, the $LISI$ and AMI scores were calculated.

Figure 13 and Figure 14 show the $LISI$ score for all the models in *PBMC_1* and *PBMC_2* respectively. Results are divided for molecular layer (ATAC or RNA). In each panel, scores are grouped based on the filters f used in the first and second network's layers. The colour identifies the number of components C evaluated in the embedding, that is also the number of filters in the generator's most internal layer.

From these results, it appears that the training failures are most common for networks characterized by few filters in the first and/or second layer. Instead, there is not a clear dependency between the performance and the embedding's components. This suggests that networks with limited number of filters are not powerful enough to learn the structure of the data. There is also no dependency of the $LISI$ score to the addressed molecular layer, implying that the model is not biased towards one data.

As for the evaluation of AMI, we had to consider that the score could be high even for models where there is no resemblance with the original data. For this reason, Figure 15 shows the AMI for trained models in *PBMC_1* and *PBMC_2* sorted by the ATAC $LISI$ and with $AMI \geq 0.4$. We decided to evaluate ATAC' and not RNA's $LISI$, as ATAC seemed to be slightly more difficult to reproduce. On average, the best models have $AMI = 0.45$.

We also performed visual inspection of the embeddings and loss functions for all trained models. This was useful to confirm the relationship between metrics, returned embeddings and loss trend.

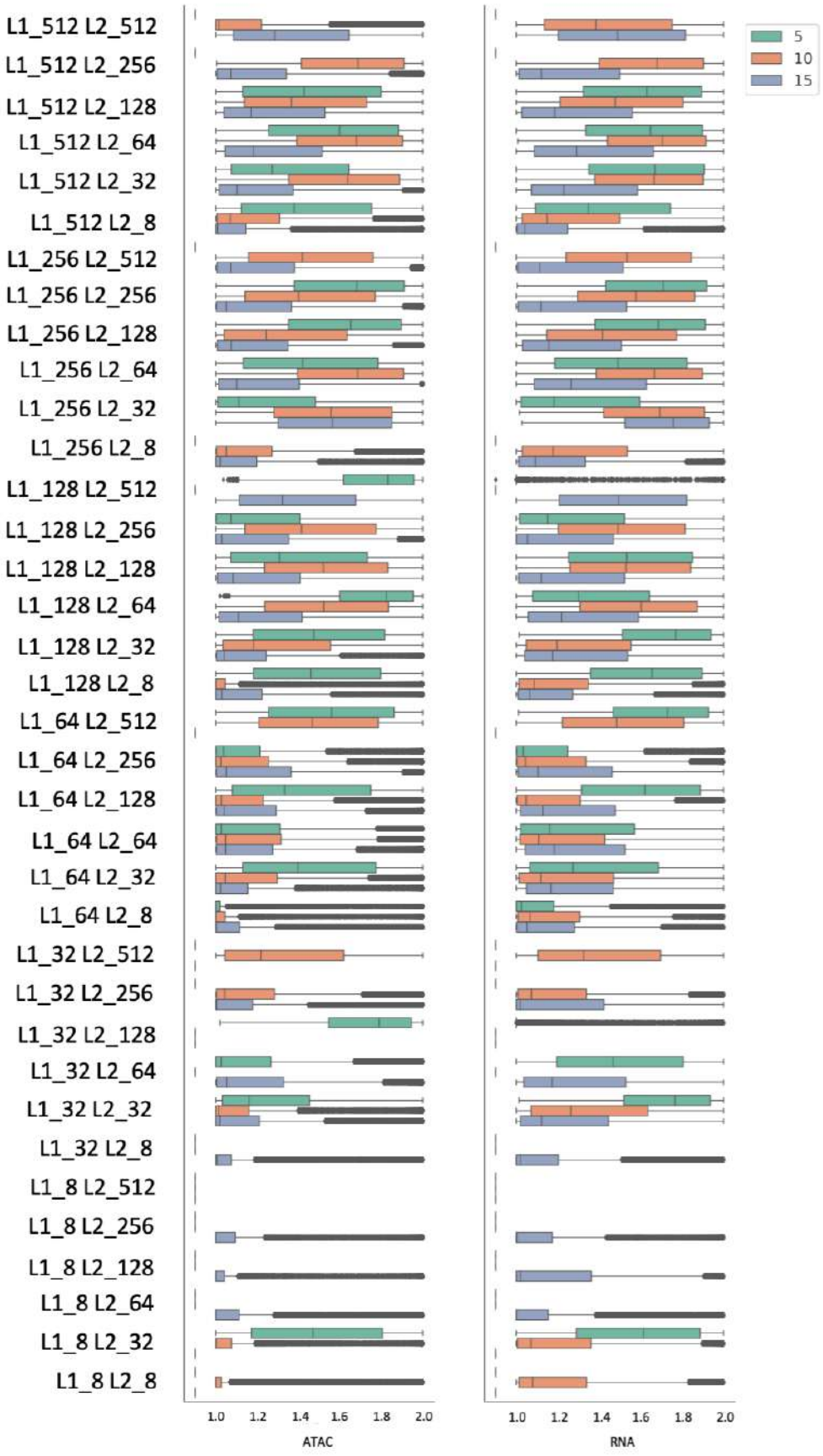


Figure 13 LISI score for *PBMC_1*. Results are divided between layers (ATAC and RNA). In each panel, scores are grouped for filters f used in the first two layers of the WGAN-GP. Colours represents the components C.

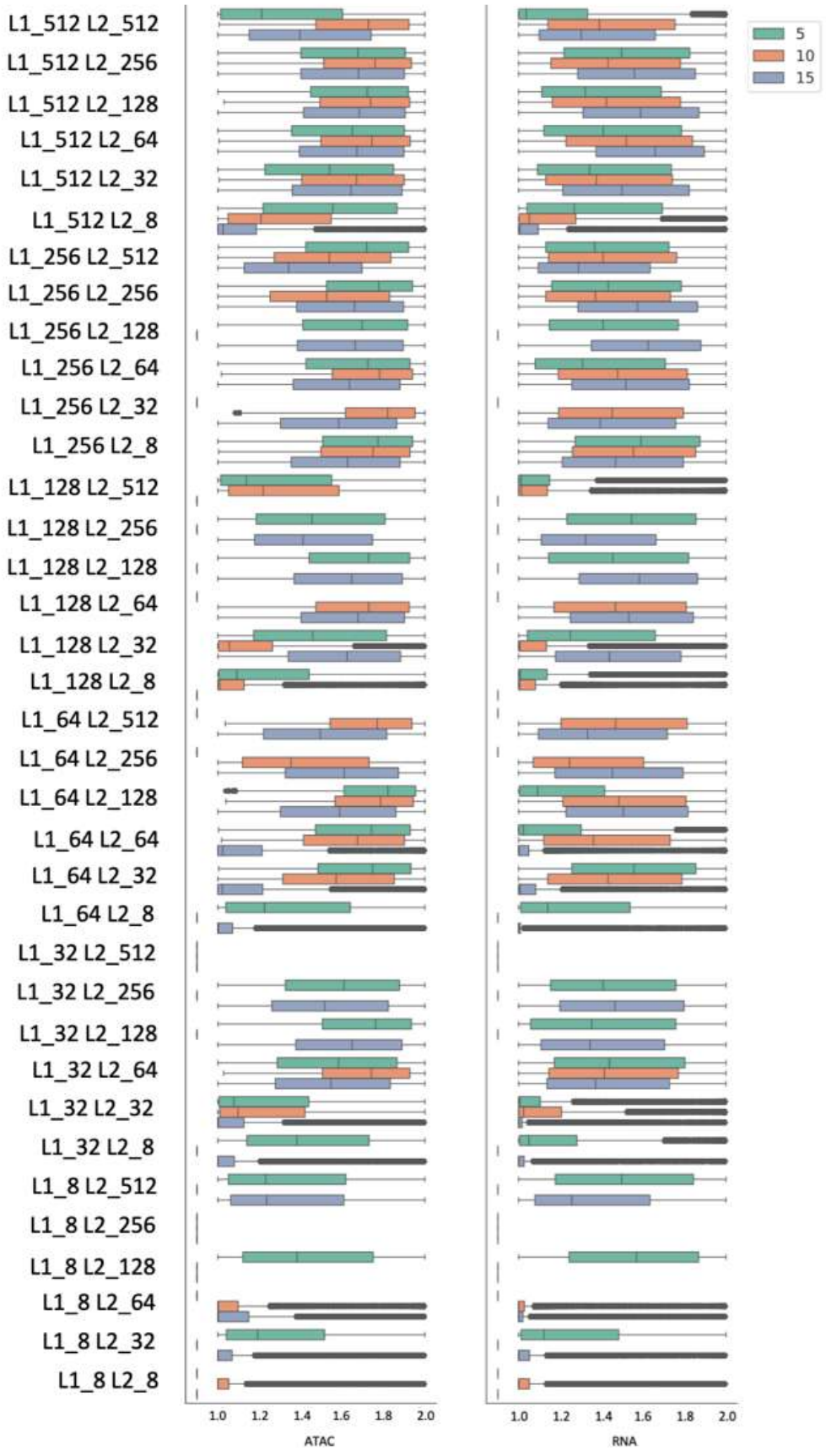


Figure 14 LISI score for *PBMC_2*. Results are divided between layers (ATAC and RNA). In each panel, scores are grouped for filters f used in the first two layers of the WGAN-GP. Colours represents the components C.

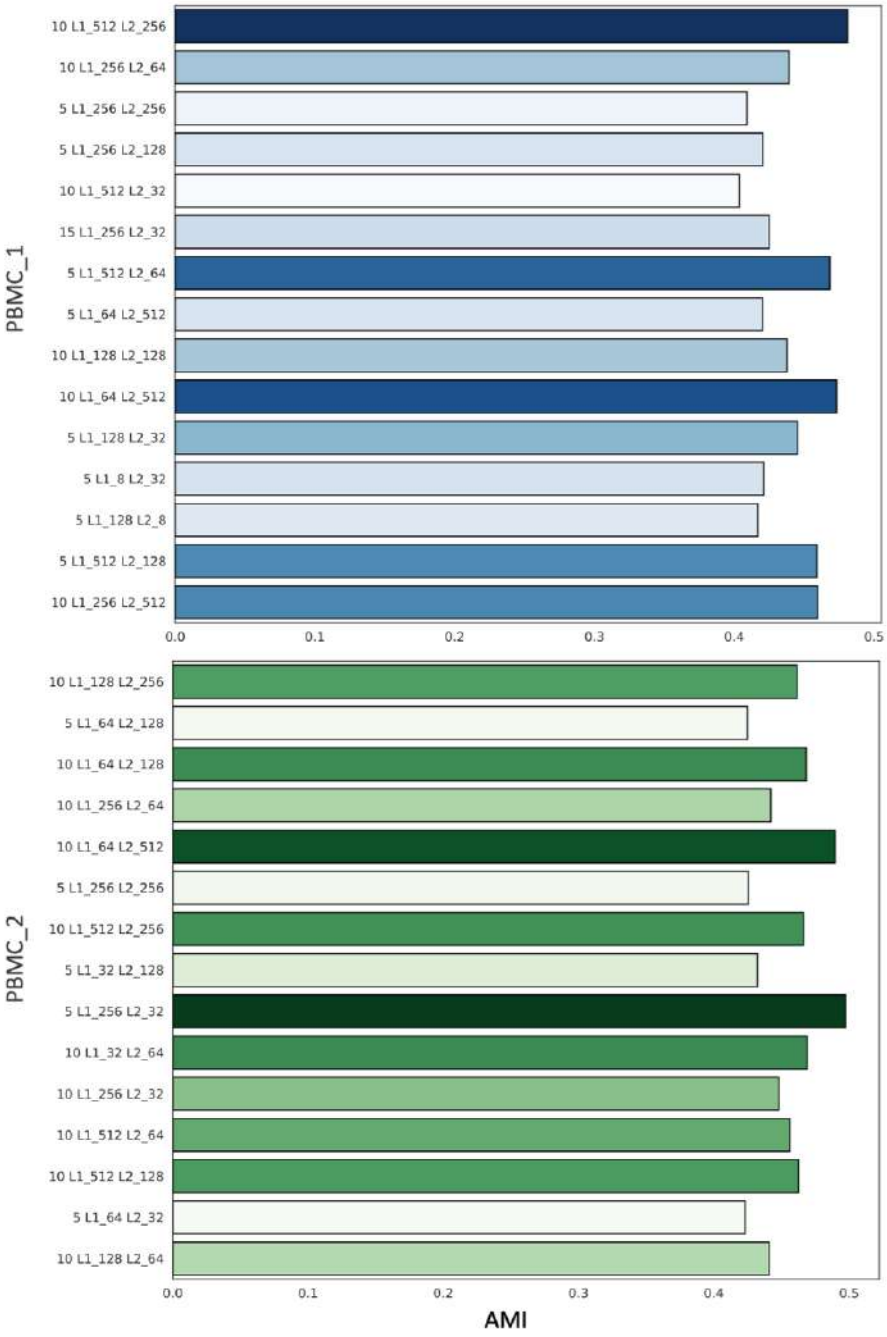


Figure 15 AMI score. For all trained models in *PBMC_1* and *PBMC_2*, the scores were sorted for the LISI calculated on the ATAC layer. Here, the AMI is reported for the top 15 models per ATAC LISI. Darker colours indicate higher AMI. The results suggest that the models should not include convolutional layers with to limited number of filters.

VALIDATION

Based on AMI and $LISI$ scores, the best trained models for *PBMC_1* and *PBMC_2* are:

- *PBMC_1*: $f = (512,256), C = 10$
- *PBMC_2*: $f = (128,256), C = 10$

From here on, they will be called respectively "*PBMC_1* MOWGAN" and "*PBMC_2* MOWGAN". We further analysed these data in terms of UMAP, loss functions and clusters concordance (Figure 16 and Figure 17), verifying the good quality of the data. We bridged the synthetic data with the original ones to transfer the cell types and confirmed the results with the GSEA (from Table 6 to Table 9). We found that RNA MOWGANs layers have much higher similarity with the original data compared to ATAC layers. However, this was expected also considering the difficulties in the association we encountered in the baseline analysis. Nevertheless, in ATAC MOWGANs we don't have inverted association between the main cell types (e.g., CD14 and CD4).

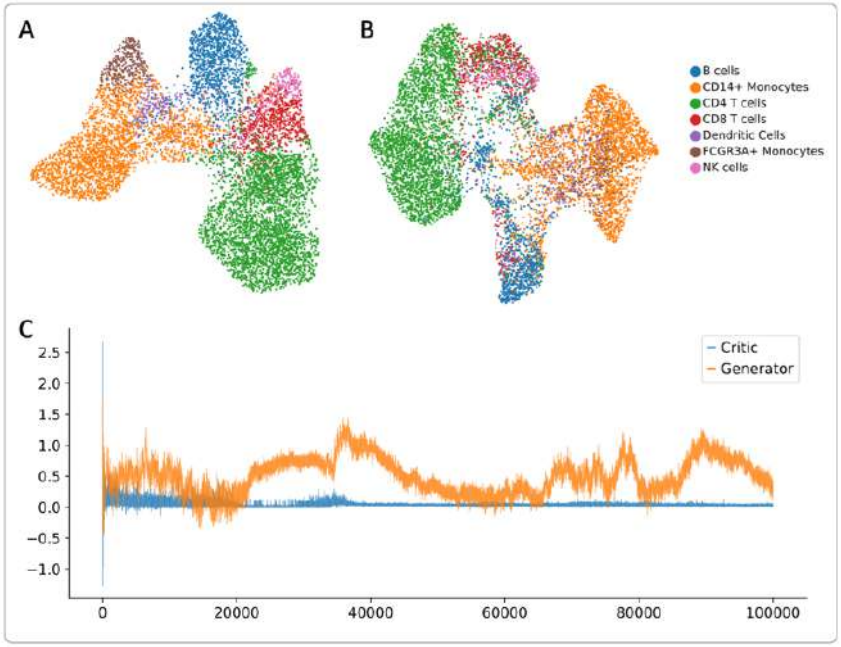


Figure 16 *PBMC_1* MOWGAN. (A) RNA UMAP coloured by the cell type transferred from the original RNA data. (B) ATAC UMAP coloured by the cell type transferred from the synthetic RNA. (C) Critic and generator loss function during the training (100.000 epochs).

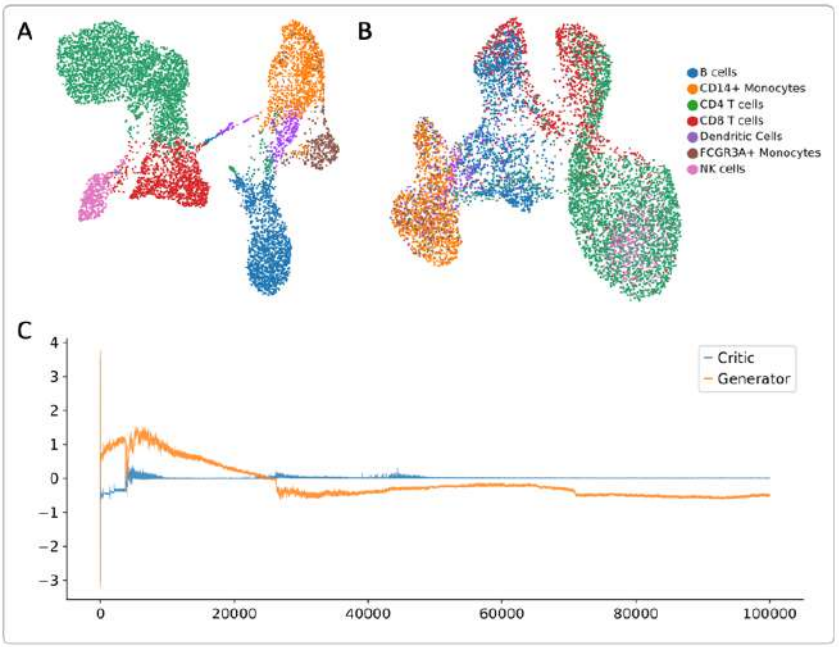


Figure 17 *PBMC_2* MOWGAN. (A) RNA UMAP coloured by the cell type transferred from the original RNA data. (B) ATAC UMAP coloured by the cell type transferred from the synthetic RNA. (C) Critic and generator loss function during the training (100.000 epochs).

Table 6 GSEA of RNA *PBMC_1* MOWGAN.

MOWGAN / TRUE	NK	FCGR3A+ Monocytes	Dendritic Cells	CD8 T cells	CD4 T cells	CD14+ Monocytes	B cells
NK	NES=2.220 Pval=0.000 FDR=0.000	NES=1.821 Pval=0.000 FDR=0.000	NES=1.092 Pval=0.370 FDR=0.495	NES=1.663 Pval=0.000 FDR=0.004	NES=-1.195 Pval=0.000 FDR=0.312	NES=0.823 Pval=0.770 FDR=1.000	NES=-0.869 Pval=1.000 FDR=1.000
FCGR3A+ Monocytes	NES=-0.808 Pval=1.000 FDR=0.700	NES=2.501 Pval=0.000 FDR=0.000	NES=1.195 Pval=0.130 FDR=0.318	NES=-1564 Pval=0.000 FDR=0.000	NES=-1.342 Pval=0.192 FDR=0.307	NES=1.790 Pval=0.000 FDR=0.000	NES=1.235 Pval=0.160 FDR=0.168
Dendritic Cells	NES=inf Pval=Nan FDR=0.000	NES=1.460 Pval=0.000 FDR=0.045	NES=2.144 Pval=0.000 FDR=0.000	NES=inf Pval=Nan FDR=0.000	NES=-3.090 Pval=0.000 FDR=0.000	NES=1.571 Pval=0.000 FDR=0.006	NES=1.257 Pval=0.040 FDR=0.188
CD8 T cells	NES=2.150 Pval=0.000 FDR=0.000	NES=1.088 Pval=0.417 FDR=0.355	NES=0.458 Pval=0.979 FDR=0.994	NES=1.859 Pval=0.000 FDR=0.000	NES=0.495 Pval=1.000 FDR=0.992	NES=0.677 Pval=0.902 FDR=0.959	NES=-0.541 Pval=1.000 FDR=1.000
CD4 T cells	NES=-1.501 Pval=0.000 FDR=0.000	NES=-1.201 Pval=0.273 FDR=0.259	NES=-1.710 Pval=0.000 FDR=0.000	NES=inf Pval=Nan FDR=0.000	NES=1.665 Pval=0.000 FDR=0.010	NES=-1.610 Pval=0.000 FDR=0.000	NES=inf Pval=Nan FDR=0.000
CD14+ Monocytes	NES=-1.888 Pval=0.000 FDR=0.000	NES=1.855 Pval=0.000 FDR=0.000	NES=0.944 Pval=0.630 FDR=0.738	NES=-1.956 Pval=0.000 FDR=0.000	NES=-0.657 Pval=0.895 FDR=0.895	NES=1.934 Pval=0.000 FDR=0.000	NES=0.682 Pval=0.930 FDR=0.928
B cells	NES=-1.076 Pval=0.500 FDR=0.533	NES=1.330 Pval=0.080 FDR=0.073	NES=2.061 Pval=0.000 FDR=0.000	NES=-0.788 Pval=1.000 FDR=0.826	NES=-1.961 Pval=0.000 FDR=0.000	NES=0.819 Pval=0.840 FDR=1.000	NES=1.938 Pval=0.000 FDR=0.000

Table 7 GSEA of ATAC *PBMC_1* MOWGAN.

MOWGAN / TRUE	NK	FCGR3A+ Monocytes	Dendritic Cells	CD8 T cells	CD4 T cells	CD14+ Monocytes	B cells
NK	NES=1.881 Pval=0.000 FDR=0.000	NES=-3.389 Pval=0.000 FDR=0.000	NES=inf Pval=Nan FDR=0.000	NES=1.888 Pval=0.000 FDR=0.000	NES=inf Pval=Nan FDR=0.000	NES=inf Pval=Nan FDR=0.000	NES=0.378 Pval=1.000 FDR=1.000
FCGR3A+ Monocytes	NES=-1.807 Pval=0.000 FDR=0.000	NES=0.695 Pval=1.000 FDR=1.000	NES=inf Pval=Nan FDR=0.000	NES=-1.584 Pval=0.000 FDR=0.000	NES=-1.106 Pval=0.354 FDR=0.354	NES= Pval= FDR=	NES=inf Pval=Nan FDR=0.000
Dendritic Cells	NES=-0.954 Pval=1.000 FDR=1.000	NES=0.667 Pval=1.000 FDR=1.000	NES=1.159 Pval=0.060 FDR=0.060	NES=-1.061 Pval=0.000 FDR=0.000	NES=-1.502 Pval=0.050 FDR=0.050	NES=0.489 Pval=1.000 FDR=1.000	NES=0.851 Pval=0.990 FDR=0.990
CD8 T cells	NES=1.416 Pval=0.000 FDR=0.000	NES=-5.629 Pval=0.000 FDR=0.000	NES=-3.758 Pval=0.000 FDR=0.000	NES=1.888 Pval=0.000 FDR=0.000	NES=0.550 Pval=1.000 FDR=1.000	NES=inf Pval=Nan FDR=0.000	NES=inf Pval=Nan FDR=0.000
CD4 T cells	NES=inf Pval=Nan FDR=0.000	NES=-1.717 Pval=0.000 FDR=0.000	NES=-1.273 Pval=0.064 FDR=0.064	NES=inf Pval=Nan FDR=0.000	NES=NA Pval=NA FDR=NA	NES=-2.909 Pval=0.000 FDR=0.000	NES=inf Pval=Nan FDR=0.000
CD14+ Monocytes	NES=inf Pval=Nan FDR=0.000	NES=NA Pval= NA FDR= NA	NES=NA Pval=NA FDR=NA	NES=inf Pval=Nan FDR=0.000	NES=-4.059 Pval=0.000 FDR=0.000	NES=NA Pval=NA FDR=NA	NES=inf Pval=Nan FDR=0.000
B cells	NES=inf Pval=Nan FDR=0.000	NES=inf Pval=Nan FDR=0.000	NES=inf Pval=Nan FDR=0.000	NES=inf Pval=Nan FDR=0.000	NES=inf Pval=Nan FDR=0.000	NES=inf Pval=Nan FDR=0.000	NES=NA Pval=NA FDR=NA

Table 8 GSEA of RNA *PBMC_2* MOWGAN.

MOWGAN / TRUE	NK	FCGR3A+ Monocytes	Dendritic Cells	CD8 T cells	CD4 T cells	CD14+ Monocytes	B cells
NK	NES=1.904 Pval=0.000 FDR=0.000	NES=0.428 Pval=1.000 FDR=0.999	NES=0.626 Pval=0.990 FDR=0.994	NES=1.122 Pval=0.140 FDR=0.214	NES=inf Pval=Nan FDR=0.000	NES=0.541 Pval=1.000 FDR=1.000	NES=inf Pval=Nan FDR=0.000
FCGR3A+ Monocytes	NES=0.653 Pval=0.990 FDR=1.000	NES=2.460 Pval=0.000 FDR=0.000	NES=1.242 Pval=0.000 FDR=0.048	NES=inf Pval=Nan FDR=0.000	NES=-1.808 Pval=0.000 FDR=0.000	NES=1.603 Pval=0.000 FDR=0.000	NES=0.554 Pval=0.990 FDR=0.996
Dendritic Cells	NES=0.394 Pval=1.000 FDR=0.999	NES=1.297 Pval=0.000 FDR=0.064	NES=2.024 Pval=0.000 FDR=0.000	NES=inf Pval=Nan FDR=0.000	NES=inf Pval=Nan FDR=0.000	NES=1.434 Pval=0.000 FDR=0.003	NES=0.872 Pval=0.870 FDR=1.000
CD8 T cells	NES=1.235 Pval=0.030 FDR=0.128	NES=inf Pval=Nan FDR=0.000	NES=inf Pval=Nan FDR=0.000	NES=1.611 Pval=0.000 FDR=0.000	NES=0.606 Pval=1.000 FDR=1.000	NES=inf Pval=Nan FDR=0.000	NES=inf Pval=Nan FDR=0.000
CD4 T cells	NES=inf Pval=Nan FDR=0.000	NES=-3.585 Pval=0.000 FDR=0.000	NES=inf Pval=Nan FDR=0.000	NES=inf Pval=Nan FDR=0.000	NES=1.340 Pval=0.010 FDR=0.083	NES=-2.210 Pval=0.000 FDR=0.000	NES=inf Pval=Nan FDR=0.000
CD14+ Monocytes	NES=0.520 Pval=1.000 FDR=1.000	NES=1.690 Pval=0.000 FDR=0.000	NES=1.360 Pval=0.000 FDR=0.005	NES=inf Pval=Nan FDR=0.000	NES=0.562 Pval=0.947 FDR=0.978	NES=2.438 Pval=0.000 FDR=0.000	NES=0.624 Pval=1.000 FDR=1.000
B cells	NES=inf Pval=Nan FDR=0.000	NES=inf Pval=Nan FDR=0.000	NES=0.971 Pval=0.600 FDR=0.697	NES=inf Pval=Nan FDR=0.000	NES=inf Pval=Nan FDR=0.000	NES=inf Pval=Nan FDR=0.000	NES=1.804 Pval=0.000 FDR=0.000

Table 9 GSEA of ATAC *PBMC_2* MOWGAN.

MOWGAN / TRUE	NK	FCGR3A+ Monocytes	Dendritic Cells	CD8 T cells	CD4 T cells	CD14+ Monocytes	B cells
NK	NES=inf Pval=Nan FDR=0.000	NES=inf Pval=Nan FDR=0.000	NES=inf Pval=Nan FDR=0.000	NES=1.700 Pval=0.000 FDR=0.000	NES=0.727 Pval=1.000 FDR=1.000	NES=inf Pval=Nan FDR=0.000	NES=0.377 Pval=1.000 FDR=1.000
FCGR3A+ Monocytes	NES=-0.982 Pval=0.400 FDR=0.400	NES=NA Pval=NA FDR=NA	NES=0.860 Pval=0.960 FDR=0.960	NES=inf Pval=Nan FDR=0.000	NES=-0.837 Pval=0.840 FDR=0.840	NES=NA Pval=NA FDR=NA	NES=0.375 Pval=1.000 FDR=1.000
Dendritic Cells	NES=-1.007 Pval=0.500 FDR=0.500	NES=inf Pval=Nan FDR=0.000	NES=0.798 Pval=0.970 FDR=0.970	NES=0.622 Pval=1.000 FDR=1.000	NES=-1.424 Pval=0.000 FDR=0.000	NES=inf Pval=Nan FDR=0.000	NES=0.885 Pval=1.000 FDR=1.000
CD8 T cells	NES=inf Pval=Nan FDR=0.000	NES=-5.368 Pval=0.000 FDR=0.000	NES=-4.666 Pval=0.000 FDR=0.000	NES=1.238 Pval=0.000 FDR=0.000	NES=0.809 Pval=1.000 FDR=1.000	NES=inf Pval=Nan FDR=0.000	NES=inf Pval=Nan FDR=0.000
CD4 T cells	NES=1.253 Pval=0.070 FDR=0.070	NES=-1.610 Pval=0.000 FDR=0.000	NES=-1.312 Pval=0.064 FDR=0.085	NES=inf Pval=Nan FDR=0.000	NES=1.637 Pval=0.000 FDR=0.000	NES=-2.469 Pval=0.000 FDR=0.000	NES=inf Pval=Nan FDR=0.000
CD14+ Monocytes	NES=-1.699 Pval=0.000 FDR=0.000	NES=NA Pval=NA FDR=NA	NES=NA Pval=NA FDR=NA	NES=inf Pval=Nan FDR=0.000	NES=-2.379 Pval=0.000 FDR=0.000	NES=NA Pval=NA FDR=NA	NES=NA Pval=NA FDR=NA
B cells	NES=inf Pval=Nan FDR=0.000	NES=inf Pval=Nan FDR=0.000	NES=inf Pval=Nan FDR=0.000	NES=0.898 Pval=0.990 FDR=0.990	NES=inf Pval=Nan FDR=0.000	NES=inf Pval=Nan FDR=0.000	NES=NA Pval=NA FDR=NA

EMBEDDING AND INVERSE TRANSFORMATION

MOWGAN's inputs are molecular layers embedded into a feature space having the same dimensionality C. By transforming the data in an embedding, the problem of the different number of features per dataset is solved. In principle any dimensionality reduction technique for which an inverse transformation is defined could be used (*e.g.*, PCA, UMAP). The inverse transformation is required to reconstruct the matrix $N \times F$, where N is the number of cells and F the number of features (genes or regions).

Indeed, after training, MOWGAN's first output is a new (coupled) dataset where the information is still expressed by a matrix $N \times C$. To this matrix the inverse transformation can be applied. However, the definition of the inverse transformation limits MOWGAN's applicability to only a bunch of embeddings. Moreover, the transformation introduces an error E in the reconstructed matrix and could be computationally expensive to perform. For these reasons, we decided to quantify E and implement an alternative solution for the matrix reconstruction.

First, we applied the reduction techniques (*i.e.*, PCA, NMF) on the count matrix X to compute C components for all our data. Then, we used the already implemented

inverse transformation to reconstruct the count matrix X'. We defined E as the mean squared error between X and X'.

The alternative solution proposed here is a *k*-NN regressor with *k*=2. The regressor is fit to learn the relationship between X and C in the selected embedding. The *k*-NN regressor is later applied on MOWGAN's output to predict X'.

Table 10 shows the errors introduced by the inverse transformations and the *k*-NN regressor. For the *k*-NN regressor, E is lower.

Table 10 Mean squared error E. Dimensionality reductions techniques introduce an error E in the reconstruction phase. If the reconstruction is performed by the defined inverse transformation (IT) of the specific technique, E is higher compared to the reconstruction done by a *k*-NN regressor.

		PBMC_1		PBMC_2*	
		RNA	ATAC	RNA	ATAC
PCA	IT	0.15	0.03	0.09	0.03
	k-NN	0.07	0.01	0.04	0.01
NMF	IT	0.15	0.03	0.09	0.03
	k-NN	0.07	0.01	0.04	0.01

BENCHMARKING

We tested the performances of four tools on our reference datasets: Pamona, SCOT, COBOLT and scMMGAN (see Chapter 2).

All selected tools return embeddings where cells are aligned between molecular layers. Like MOWGAN, Pamona, SCOT and scMMGAN inputs are single assay data. Pamona and SCOT implement a Gromov-Wasserstein optimal transport solution, while scMMGAN uses GANs. COBOLT is also a deep learning tool, but the inputs are both paired and unpaired dataset. For this reason, only in this case *PBMC_1* and *PBMC_2** were used.

The tools were used with standard parameters, following the tutorials. Cell type annotations were evaluated on the aligned embeddings. This step did not require any

label transferring as the tools do not generate new datasets. In Figure 18 the originals, MOWGANs and the four tools' embeddings are represented together, coloured by cell types. Panel A and B show the outcomes on *PBMC_1* and *PBMC_2* respectively. We can observe how PAMONA's embeddings are entirely different from the others. Pamona stress the linear relationship between the cells returning an embedding that is hard to trust. SCOT and scMMGAN outputs are much more similar to the reference. Nevertheless, they introduce an incorrect association between cells of different layers.

COBOLT outputs three embeddings: one for the unpaired RNA, one for the unpaired ATAC and one for the paired data. They are topologically similar between each other, but they do not relate with the reference. Moreover, the cell type annotation is completely mixed in the single-assays ATAC embedding, suggesting something went wrong during the training.

For this benchmark, the $LISI$ score can be used as a measure of the distortion introduced in the dataset due to the alignment. For all four tools, the $LISI \cong 1$, meaning the returned embeddings are not integrable with the originals (*i.e.*, they are distorted).

Performances were evaluated also in terms of homogeneity, completeness and V-measure between cell types (ground truth) and Leiden clusters calculated on the different embeddings (Table 11 and Table 12).

Pamona has high scores, but we remark the unreliability of its outputs, at least on the tested datasets. COBOLT has good performances relatively to the paired and RNA unpaired layers. Its performances drastically drop on the unpaired ATAC layer. MOWGAN, SCOT and scMMGAN are in line, with SCOT slightly overtaking.

Our benchmark indicates that the true discriminant between MOWGAN and the other tools is the ability to generate data with 1) minor distortion and 2) less mislabelling.

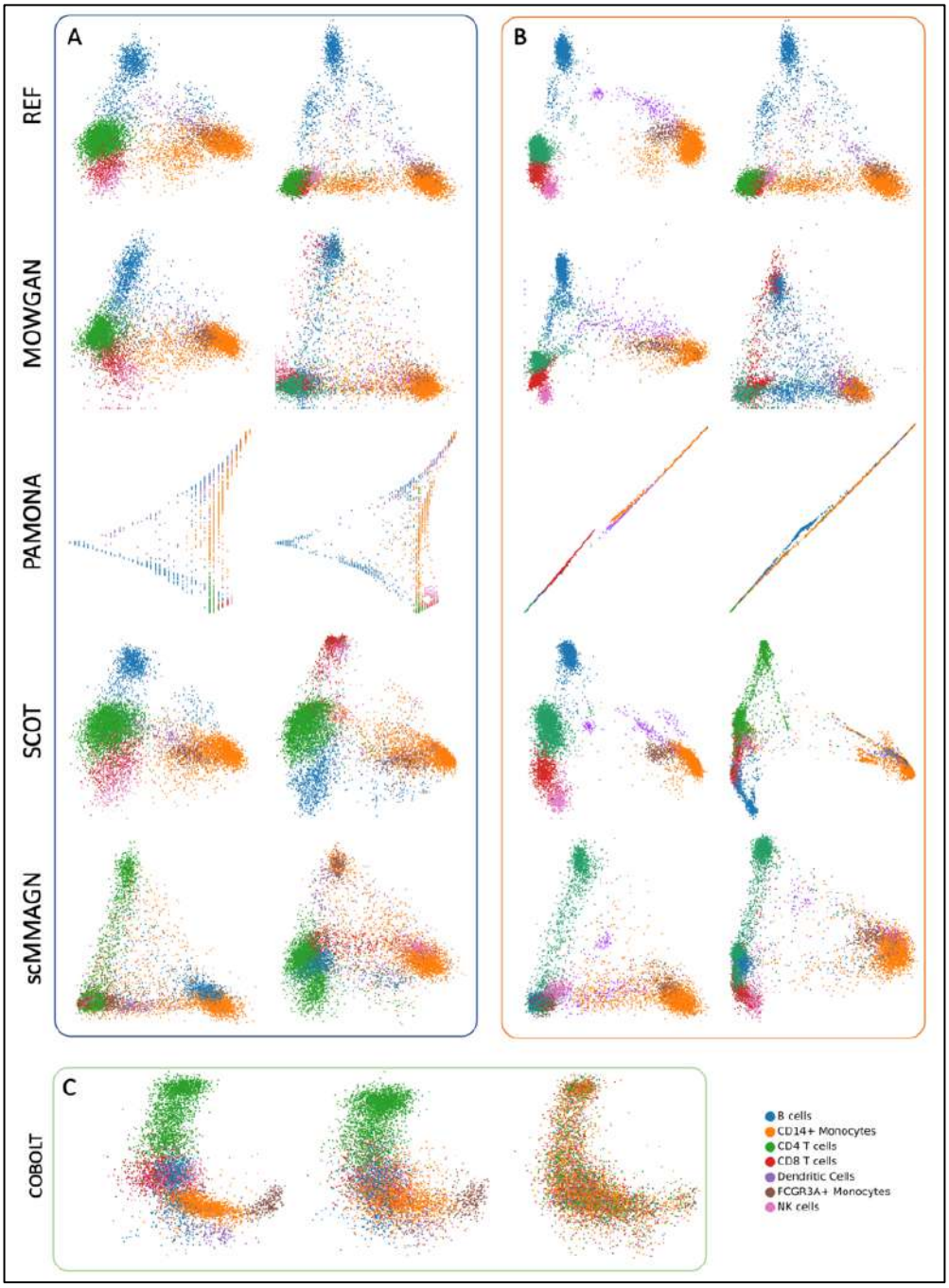


Figure 18 Tools benchmark on *PBMC_1* (A), *PBMC_2* (B) and *PBMC_2** (C). In (A) and (B), the first column represents the embeddings for the RNA layer. The second column is the embeddings for ATAC. In (C) we have in order the embeddings for the unpaired RNA, for the paired dataset and for the unpaired ATAC. All embeddings are coloured for cell types.

Table 11 MOWGAN, PAMONA, SCOT and scMMGAN performances. All metrics are in line, with SCOT slightly surpassing the others.

		PBMC_1		PBMC_2*	
		RNA	ATAC	RNA	ATAC
Ref	Homogeneity	1.00	0.81	0.94	0.81
	Completeness	0.62	0.54	0.62	0.54
	V measure	0.76	0.65	0.74	0.65
MOWGAN	Homogeneity	0.71	0.53	0.85	0.52
	Completeness	0.52	0.36	0.56	0.37
	V measure	0.60	0.43	0.68	0.44
PAMONA	Homogeneity	0.89	0.82	0.94	0.79
	Completeness	0.38	0.35	0.40	0.33
	V measure	0.53	0.49	0.56	0.46
SCOT	Homogeneity	0.86	0.68	0.94	0.71
	Completeness	0.54	0.34	0.58	0.33
	V measure	0.66	0.46	0.72	0.45
scMMGAN	Homogeneity	0.74	0.70	0.89	0.70
	Completeness	0.43	0.42	0.53	0.40
	V measure	0.54	0.53	0.67	0.51

Table 12 COBOLT performances. COBOLT was run on combination of paired and unpaired data. It returns three different embeddings: for the unpaired RNA and ATAC layers and one for the paired dataset. Performances on the unpaired ATAC layer are drastically low.

		PBMC_1		PBMC_2*	
		RNA	ATAC	RNA	ATAC
Ref	Homogeneity	1.00	0.81	0.94	0.74
	Completeness	0.62	0.54	0.62	0.41
	V measure	0.76	0.65	0.74	0.53
Cobolt	Homogeneity	0.87	0.87	0.92	0.005
	Completeness	0.49	0.49	0.56	0.003
	V measure	0.63	0.63	0.70	0.003

CHAPTER 5

ADVANCED APPLICATIONS

MOWGAN allows to work on challenging settings. It accepts dataset with complex experimental design (*e.g.*, multiple batches and/or conditions). It also allows the integration of more than two molecular layers.

BATCH-INFORMED TRAINING

The experimental design can be used to inform the training. In Chapter 4, this was called "aware training". To summarize, the dataset is segmented with respect to a property (*e.g.*, sample identity) shared across layers. Thus, subset-specific models are trained. A complete dataset is later reconstructed by merging all models' outputs.

The batch-aware training improves the quality of the synthetic data. In this setting, the proportion of each subset in the merged dataset can also be chosen.

To demonstrate the benefits of an aware training, MOWGAN was applied on a human-derived colorectal cancer organoids (CRC) dataset. The dataset is composed of three organoids/batches (CRC_6, CRC_17, CRC_39). scRNA and scGET data were collected.

MOWGAN was applied 1) on the whole dataset (not batch-informed training) and 2) on organoid-specific subsets (batch-informed training). In both scenarios, PCs and tensor train decomposition (TTD) components were used for RNA and GET data

respectively. For the sake of simplicity, we will call the generated dataset CRC_NB and CRC_B respectively.

In the former case, the analysis we performed is the same as the PBMC datasets. Therefore, we trained a single WGAN-GP model and used Schist to transfer the batch annotation on the synthetic data (CRC_NB) (Figure 19).

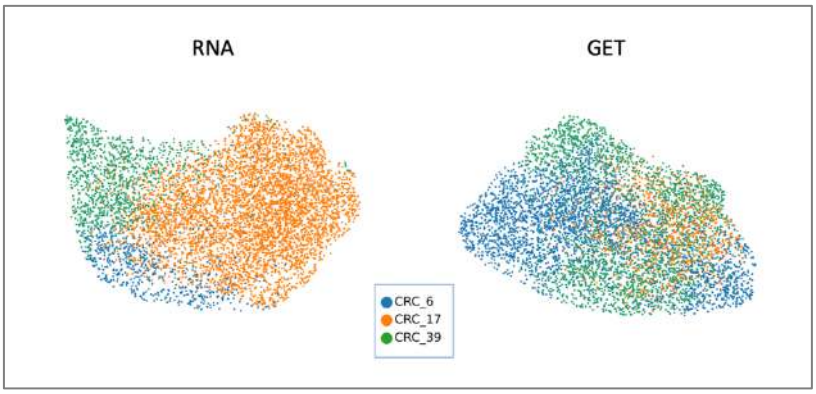


Figure 19 CRC_NB RNA and GET UMAP colored by batch transferred by Schist.

In the latter case, three models were trained. On CRC_B the batch is derived directly from the trained model. This knowledge can be used to test the reliability of the label transfer system. To this end, we applied Schist to CRC_B, and we compared Schist annotation with the real one (Table 13 and Figure 20).

Table 13 CRC _B performance metrics. The metrics were evaluated between the true batch name and the one transferred thought Schist.

		RNA	ATAC
CRC_B	Homogeneity	0.90	0.28
	Completeness	0.87	0.26
	V measure	0.88	0.27
	Accuracy	0.97	0.77

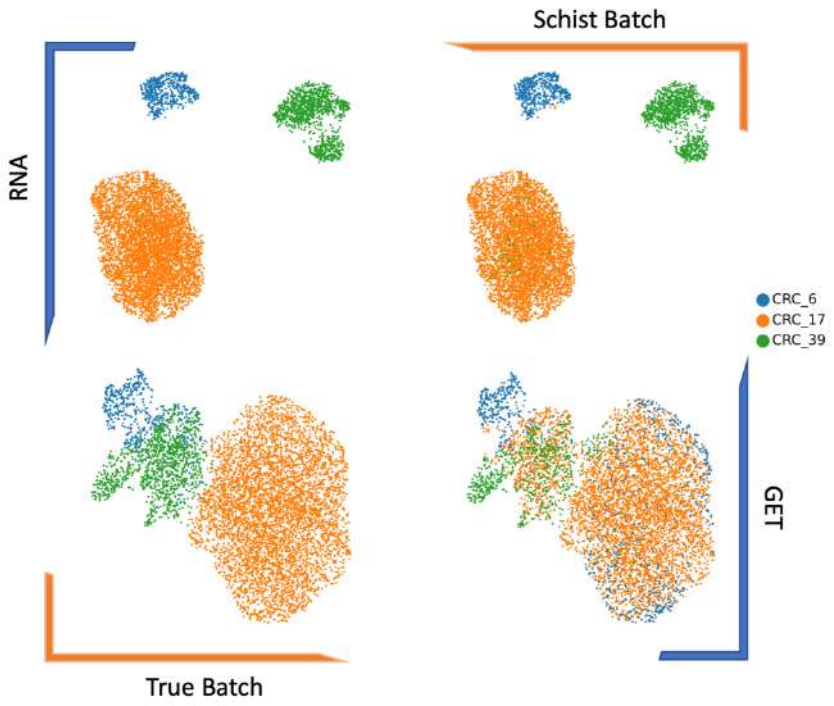


Figure 20 CRC_B dataset. UMAPs for the RNA and GET synthetic data colour by the true batch name, and the batch transferred through Schist from the original data. The two labels agree in the RNA layer. In GET, the label transfer in a bit more uncertain.

On the RNA layer the two annotations are mostly in agreement. The performance decreases on the GET layer, even if the accuracy is still high. In this setting, the accuracy is an index of the quality of the transferring system. This means that in the CRC_B dataset we would trust more a transferring direction from RNA→GET compared to GET→RNA.

To refine the dataset, we can subset CRC_B to include only cells with the same annotations. Moreover, given the model error:

$$E = 1 - accuracy$$

if N are the observations we want in the synthetic dataset, the model could generate S samples, with:

$$S = (1 + E)N$$

To demonstrate the improved performances in CRC_B compared to CRC_NB, the two datasets were integrated, and Leiden clusters were evaluated. We computed the quality metrics between the batch transferred by Schist and the clusters computed in

the integrated objects (Table 14). CRC_B outperforms CRC_NB in the RNA layers. Between the GET layers no real difference is appreciated.

Table 14 CRC_B vs CRC_NB. Metrics were calculated between Leiden clusters derived on the integrated object and the batch annotation transferred with Schist.

	CRC_B		CRC_NB	
	RNA	ATAC	RNA	ATAC
Homogeneity	0.82	0.27	0.45	0.25
Completeness	0.34	0.08	0.19	0.12
V measure	0.48	0.14	0.26	0.16
AMI	0.30		0.22	

CRC_B and CRC_NB are not integrable. Globally the $LISI$ scores are in line with what was already observed in the baseline analysis ($LISI_{rna} = 1.37, LISI_{get} = 1.26$). Nevertheless, when the $LISI$ is calculated on batch-subsets the performances drop especially for CRC_6 and CRC_17 (Table 15). Moreover, we observed a better integration between CRC_B ($LISI_{rna} = 1.55, LISI_{get} = 1.33$) and original data compared to CRC_NB ($LISI_{rna} = 1.42, LISI_{get} = 1.25$) and the originals. This is true also for $LISI$s evaluated on single batches (Table 16), where lower performances are registered especially in the RNA CRC_NB layer compared to the RNA CRC_B layer.

Table 15 CRC_NB and CRC_B integration. The LISI scores evaluated for single batch.

	CRC_6	CRC_17	CRC_39
RNA	1.14	1.13	1.24
GET	1.15	1.21	1.31

Table 16 CRC_B and CRC_NB integration with the original data. The LISI scores are evaluated for single batch.

		CRC_6	CRC_17	CRC_39
Original + CRC_B	RNA	1.86	1.47	1.74
	GET	1.19	1.40	1.26
Original + CRC_NB	RNA	1.19	1.41	1.32
	GET	1.25	1.27	1.30

THREE- AND FOUR-LAYERS INTEGRATION

MOWGAN can integrate more than two molecular layers. For demonstration, we present four different case studies. PCs were used as embeddings. The WGAN-GP architecture was adjusted to match the kernel-size of the Conv1D layers to the number of modalities evaluated.

CASE STUDY I

First, the integration of RNA, ATAC and proteins abundance data was tested. The RNA and proteins layers derived from the CITE-Seq dataset. ATAC data were the same as in *PBMC_1*. MOWGAN was not informed of the coupling between RNA and proteins.

The results reported here were noteworthy. MOWGAN's embeddings were extremely integrable with the originals ($LISI_{RNA} = 1.7, LISI_{ATAC} = 1.6, LISI_{protein} = 1.7$). The Leiden clusters evaluated on each synthetic dataset had $AMI \geq 0.46$ with clusters evaluated in the other layers.

Figure 21 shows the cell types transferred from the originals to MOWGAN's data by Schist. Moreover, RNA annotation was also transferred to ATAC and protein layers by cell ID. The RNA annotation mostly agrees with the cell type defined by Schist on the ATAC and protein layers.

Table 17 summarizes the quality metrics between cell types and Leiden clusters for the reference (real datasets) and MOWGAN's data (cell type transferred by Schist). We observed a decrease in the performance compared to the baseline, especially in the RNA layer. Nevertheless, the overall performance is high.

Table 17 Case study I: RNA, ATAC and protein abundance quality scores. Metrics are evaluated between cell types and Leiden clusters.

		RNA	ATAC	Protein
Ref	Homogeneity	0.85	0.82	0.74
	Completeness	0.73	0.55	0.61
	V measure	0.79	0. 66	0.67
MOWGAN	Homogeneity	0.65	0.64	0.67
	Completeness	0.51	0.58	0.58
	V measure	0.57	0.61	0.62

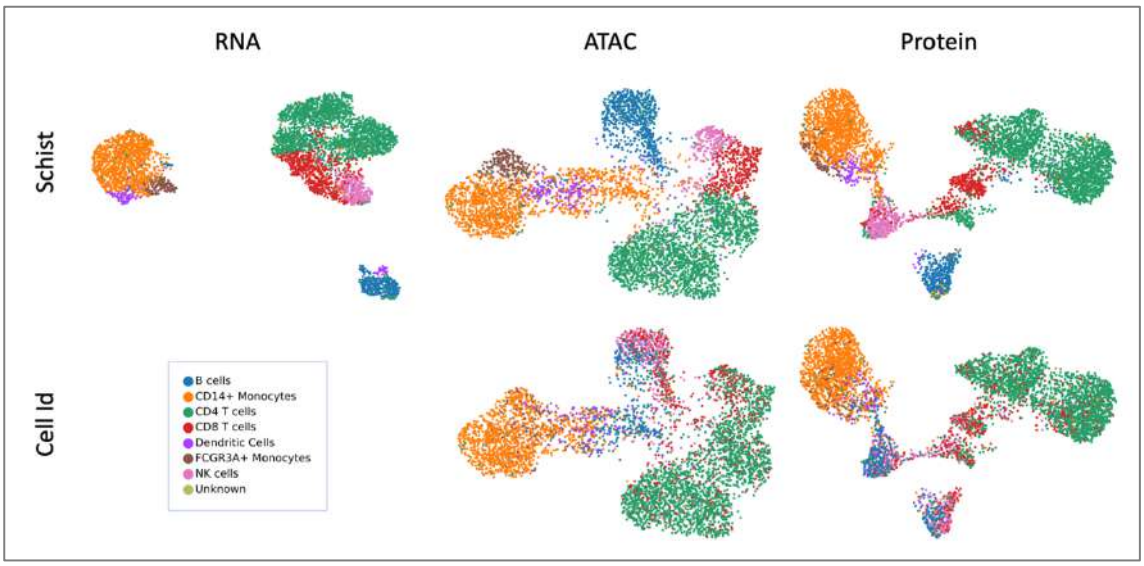


Figure 21 Case study I: integration of RNA, ATAC and protein abundance. In the first row, MOWGAN's embeddings are coloured by the cell type transferred by Schist from the original data. In the second row, the RNA annotation is transferred to ATAC and protein data by cell identity. Annotations are consistent with each other.

CASE STUDY II

The second case study was the integration of RNA and ATAC with a histone modification dataset. In details, we integrated *PBMC_2* and H3K27me3 from the public scCUT&Tag-pro dataset. Cell type annotation provided with scCUT&Tag-pro data is slightly different from the one used in PBMC. Therefore, we modified it accordingly (Table 18).

Table 18 Cell type annotation. Cell types in 10x and scCUT&Tag-pro were renamed to "Common Annotation" to match them between the two datasets.

10x	scCUT&Tag-pro	Common Annotation
CD4 T cells CD8 T cells	CD4 T CD8 T other T	T cells
CD14+ Monocytes FCGR3A+ Monocytes	Mono other	Monocytes
B cells	B	B cells
NK cells	NK	NK cells
Dendritic Cells	DC	Dendritic Cells

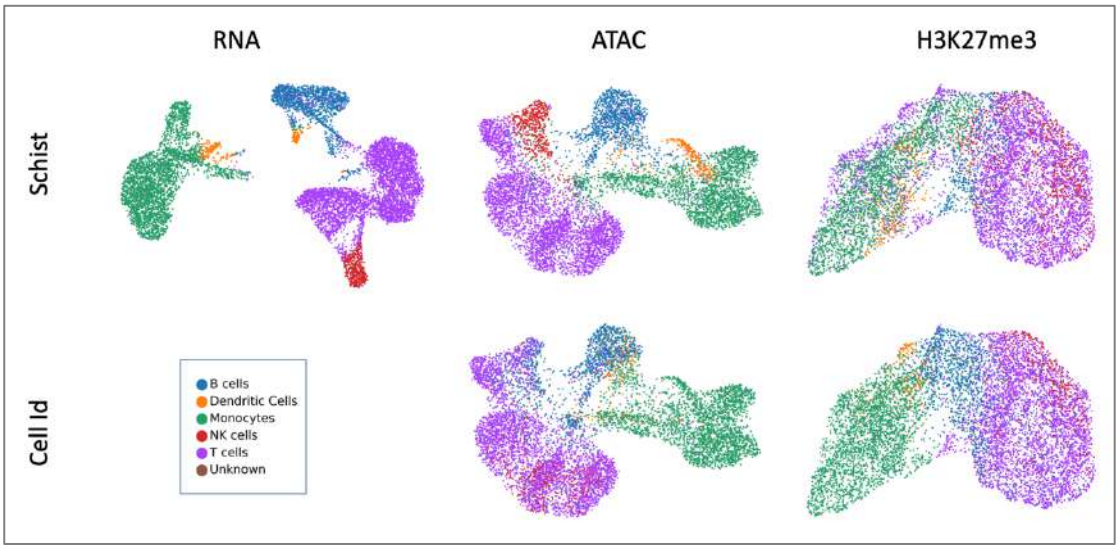


Figure 22 Case study II: integration of RNA, ATAC and H3K27me3 histone modification. In the first row, MOWGAN's embeddings are coloured by the cell type transferred by Schist from the original data. In the second row, the RNA annotation is transferred to ATAC and H3K27me3 data by cell identity. Annotations are consistent with each other.

Compared to the first case study, MOWGAN's results were here less integrable with the original data ($LISI_{RNA} = 1.30, LISI_{ATAC} = 1.66, LISI_{H3K27me3} = 1.37$). Nevertheless, the cell type association between different layers was higher. Indeed, the accuracy between the cell type transferred by Schist and the one passed through cell identity from RNA to ATAC/H3K27me3 was higher than 60% (Figure 22). Moreover, the quality metrics in

Table 19 show that MOWGANs data retain the properties of the reference. In the reference itself, H3K27me3 appears to be of poor quality. Indeed, clusters identified

in this layer are not representative of biological properties (Table 19). This behaviour can also be observed in the synthetic H3K27me3 data.

Table 19 Case study II: RNA, ATAC and H3K27me3 quality scores. Metrics are evaluated between cell types and Leiden clusters.

		RNA	ATAC	H3K27me3
Ref	Homogeneity	0.95	0.84	0.55
	Completeness	0.49	0.45	0.27
	V measure	0.64	0.59	0.37
MOWGAN	Homogeneity	0.83	0.70	0.34
	Completeness	0.49	0.46	0.21
	V measure	0.61	0.56	0.26

CASE STUDY III

Here, the integration of four molecular layers was tested for the first time. *PBMC_1* was used in combination with the H3K27me3 histone modification and its protein abundance data. Also in this case, we used the redefined cell type annotation presented in Table 18.

We observed high integrability between MOWGAN's data and the originals, with the lowest performance registered once again by the histone layer ($LISI_{RNA} = 1.67, LISI_{ATAC} = 1.67, LISI_{protein} = 1.75, LISI_{H3K27me3} = 1.38$). The same trend was observed in the other quality metrics (Table 20).

The cell type annotations (Figure 23) mostly agree between layers. We registered the highest accuracy between RNA and ATAC cell types (80%) and the lowest between RNA and the histone (55%).

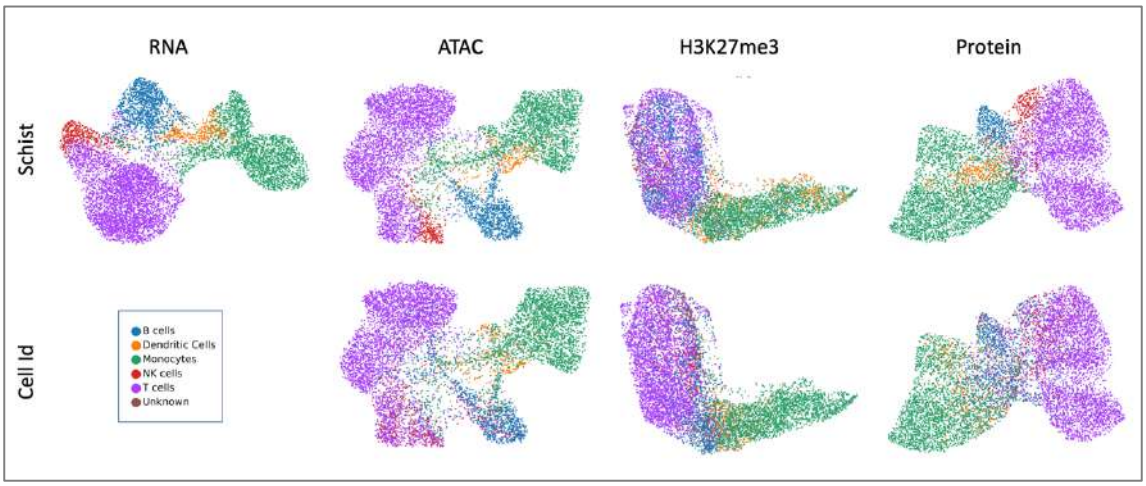


Figure 23 Case study III: integration of RNA, ATAC, H3K27me3 histone modification and protein abundance. In the first row, MOWGAN's embeddings are coloured by the cell type transferred by Schist from the original data. In the second row, the RNA annotation is transferred to ATAC, H3K27me3 and protein data by cell identity. Annotations are consistent with each other.

Table 20 Case study III: RNA, ATAC, H3K27me3 histone modification and protein abundance quality scores. Metrics are evaluated between cell types and Leiden clusters.

		RNA	ATAC	Protein	H3K27me3
Ref	Homogeneity	0.99	0.84	0.81	0.55
	Completeness	0.49	0.45	0.41	0.29
	V measure	0.67	0.59	0.54	0.38
MOWGAN	Homogeneity	0.75	0.73	0.64	0.38
	Completeness	0.61	0.48	0.44	0.32
	V measure	0.67	0.58	0.53	0.35

CASE STUDY IV

The last scenario we addressed is the integration of *PBMC_1* with two histone modifications: H3K27me3 and H3K4me2. Integrability ($LISI_{RNA} = 1.73, LISI_{ATAC} = 1.66, LISI_{H3K4me2} = 1.65, LISI_{H3K27me3} = 1.41$), cell type agreement between each molecular layer and RNA's cell types ($0.65 \leq \text{accuracy} \leq 0.77$) (Figure 24), and quality metrics (Table 21) were all good. In the reference, H3K4me2 has higher association between clusters and cell types compared to H3K27me3 (Table 21). This is also verified in the synthetic data.

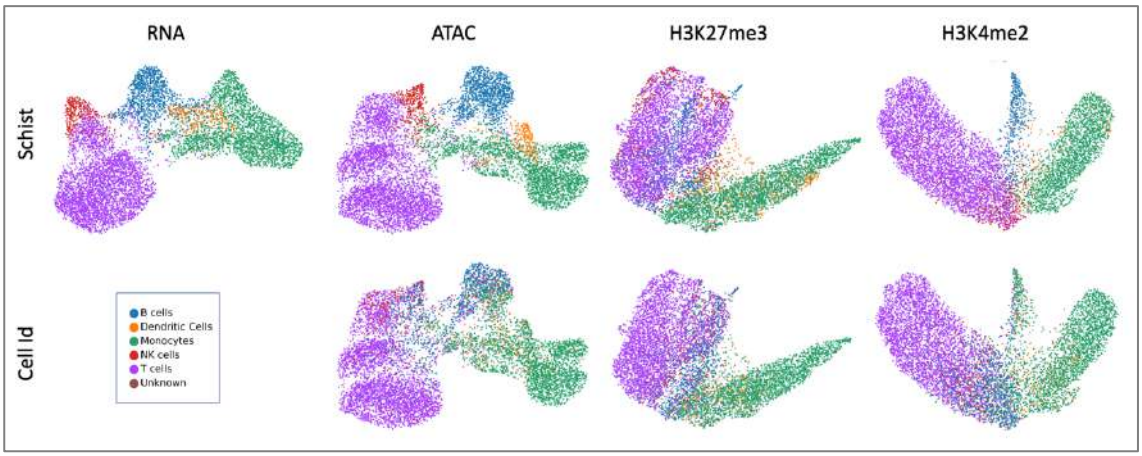


Figure 24 Case study IV: integration of RNA, ATAC, H3K27me3 and H3K4me2 histone modifications integration. In the first row, MOWGAN's embeddings are coloured by the cell type transferred by Schist from the original data. In the second row, the RNA annotation is transferred to ATAC, H3K27me3 and H3K4me2 data by cell identity. The annotations are consistent with each other.

Table 21 Case study IV: RNA, ATAC, H3K27me3 and H3K4me2 histone modifications quality scores. Metrics are evaluated between cell types and Leiden clusters.

		RNA	ATAC	H3K27me3	H3K4me2
Ref	Homogeneity	0.99	0.84	0.55	0.65
	Completeness	0.49	0.45	0.27	0.45
	V measure	0.66	0.59	0.37	0.54
MOWGAN	Homogeneity	0.70	0.76	0.46	0.65
	Completeness	0.58	0.52	0.32	0.56
	V measure	0.64	0.62	0.38	0.60

DISCUSSION

The characterization of biological entities (*e.g.*, cell type) requires studying the processes underlying cell ecosystem. To this end, single-cell sequencing technologies were introduced. However, they cannot guarantee the simultaneous observation of more than one molecular layer, essential to comprehensively depict cells.

The need of tools to integrate multiple single cell data, corroborated by the vast amount of software tools developed, prompted the development of MOWGAN, a machine learning framework for the generation of synthetic paired multi-omics single-cell datasets. It should be underlined that integration methods could be extended to other, non-single cells, contexts such as the study of large bulk assays collections (*e.g.*, The Cancer Genome Atlas data).

To summarize, MOWGAN's inputs are low-dimensional representation of unpaired multi-modal data. Cells in each dataset are sorted based on the first component of their Laplacian Eigenmap and later used to train a single WGAN-GP. The training is performed in mini-batches. A Bayesian ridge regressor is iteratively applied to guide the mini-batch construction.

MOWGAN's first output is a synthetic embedding for each input molecular layer. The embedding is converted into a matrix $N \times F$ (with N the number of generated cells and F the number of features available for each layer in the original dataset) by a *k*-NN regressor.

MOWGAN's outputs verify two main properties:

1. Integrability between synthetic and original data
2. Induced coupling between molecular layers in the synthetic data

The first property assess that the generated data maintain the structural characteristics of the original ones. The second measures the pairing introduced in the data. Moreover, it was demonstrated that the association established between the layers is respectful of the biology.

Therefore, MOWGAN can be used to bridge real unpaired dataset. Annotations defined i) in the true unpaired data or ii) in the integrated object with MOWGAN's data, can be transferred to every other modality for which coupled data were generated.

To prototype MOWGAN, public data were used. First, PBMC dataset annotated with the same cell types were evaluated. AMI was calculated between Leiden clusters in RNA and ATAC paired data to measure the share information between different modalities in the same cells. That is the upper limit of shared information that could be expected in simulated data.

GSEA was performed between cell types in two RNA datasets and two ATAC datasets (*i.e.*, intra-modality evaluation). These analyses demonstrated how the biological information between RNA datasets is more conserved compared to what can be seen between ATAC data.

AMI and GSEA were performed also on MOWGAN synthetic data. As expected, the AMI introduced in the data was lower to the one observed in the reference. The features expressed by RNA MOWGAN data for all cell types were consistent with the characterization of the cell type in the reference. For the ATAC counterpart, as in the reference, results were not so straightforward. Indeed, enrichments for the same regions of the reference were found only for few cell types (*e.g.*, NK cells, CD8 cells, CD4 cells).

We reasoned that the different performance of the GSEA in RNA and ATAC, both real and synthetic data, could be due to the high number of features in ATAC, and the small contribution they have in the characterization of the cell type.

MOWGAN was used on the CRC dataset to demonstrate the improved performance when the training is modified to include prior knowledge on the dataset, as for the batch's origin. Indeed, MOWGAN was trained on CRC with and without the batch information. When the batch information was not used, the synthetic dataset was less integrable with the original one, affecting the ability to transfer annotation between the data.

Due to the model architecture, MOWGAN can be trained on more than two datasets without need to modify the networks. This is an important accomplishment. Many tools support the integration of more than two modalities, but generally require a drastic change in the model architecture. Moreover, MOWGAN is consistent to changes in the hyperparameters, meaning that results are generalizable, and fine-tuning is not strongly required.

MOWGAN was first applied to integrate RNA, ATAC and protein data (case study I). RNA and protein were from the CITE-seq dataset, while ATAC was from 10x. MOWGAN was agnostic to the pairing in the CITE-seq part. Nevertheless, it was able to create association between RNA and proteins and, most importantly, with ATAC, a dataset generated using a different platform and derived from a different blood sample.

To demonstrate again the ability of MOWGAN to integrate three datasets of different origins (*i.e.*, platforms, aliquots and laboratories), RNA and ATAC unpaired data from 10x were integrated to the H3K27me3 histone modification from the scCUT&Tag-pro dataset (case study II). The histone layer was not particularly clean, meaning that cell types were not clearly identified. Indeed, the association between clusters and cell types was lower compared to the other layers, as reported in Table 19.

To conclude, a four layers integration was tested first between RNA and ATAC paired data with H3K27me3 histone modification and its protein abundance (case study III), and later with H3K27me3 and H3K4me2 histone modifications (case study IV). In both cases, MOWGAN was uninformed of the correct pairing between cells of different modalities. Results for H3K27me3 were always poorer compared to the other molecular layers it was integrated with, whereas H3K4me2 data had better performances. This was true also for the reference, and therefore it is reflected in MOWGAN performance.

The main problem emerging from the literature about multi-omics data integration tools is the introduction of incorrect associations during the inference process. Indeed, four tools were here tested for the integration of RNA and ATAC PBMC data: Pamona, SCOT, COBOLT and scMMGAN. They were selected for benchmarking due to the similarity shared with MOWGAN in the data processing, allowing a direct comparison between the results. Tools as Seurat (Stuart et al., 2019) and scVI (Lopez et al., 2018), usually applied for data integration, were not evaluated as they require data

transformation (*i.e.*, calculation of gene activity matrix) or they work with paired assays only.

Performances were evaluated in terms of homogeneity, completeness and V-measure between cell types and Leiden clusters. The LISI score was also considered as a measure of distortion. Other metrics were investigated as potentially useful but later discarded, due to inapplicability and redundance of information. Among them, the Preserve Paired Jaccard Index (PPJI) introduced in (Martinez-De-Morentin et al., 2021). In the original paper, PPJI is calculated between labels (*e.g.*, Leiden clusters or cell types) observed in the original embedding versus the ones in the embedding post alignment, to evaluate the conservation and/or improvement in the (cell types) definition in the integrated space. This imply a one-to-one relationship between old and new annotations, that is not present in MOWGAN, making the score not applicable in this context.

Compared to MOWGAN, the results of benchmarking tools were severely affected by untrue associations between groups of different layers, like the coupling between CD4 cells and CD14 cells. While MOWGAN is also affected by similar problems, the iterative step performed during the mini-batch selection can mitigate the issue. Nevertheless, in the results showed so far it was evident that the pairing introduced in the generated data by MOWGAN was not fully accurate. The performance depends for the most part from the quality of the input data. Indeed, if the embeddings are not representative enough of the biological information, or cell entities (*e.g.*, cell types) are not distinguishable, MOWGAN will be affected. This was already observed with the H3K27me3 analysis: from a poor-quality layer will be generated poor-quality synthetic data. Hence, the importance of preprocessing to bring out the information from the data. Clean, processed data are recommended as inputs for MOWGAN. Knowing these limitations, the usefulness of the data should be considered before the integration. On the other hand, MOWGAN is not affected by unbalanced dataset, as demonstrated with the CRC. Differential proportions within batches are overtaken by training the network in mini-batches. Nevertheless, a problem of resolution is observed: small cell types could be lost or incorrectly associated with cell types more represented.

Finally, we believe MOWGAN helps our understanding of the biological processes that take place inside cells. It is a strong framework, which output should be critically studied. In the end, a result should always be experimentally validated.

We also think that there's room for further improvement. First, the mini-batch selection strategy could be enhanced to include genetic information. To this end, if information about the genetic identities of cells is known, it could be used to tune the selection of mini-batches. Second, the introduction of weights to penalize the importance, in the training phase, of molecular layers we are not confident in is also planned. Last, model interpretability should be properly investigated: deeply looking into the training process of the WGAN-GP, such as the features that contribute the most to the characterization of the biological entities, could help unravelling the underlying biology side.

Although these improvements can be made, we believe that the development of a tool suitable for all applications/datasets is not feasible. This is an outcome expected also considering the results of a Multi-modal Single-Cell Data Integration Competition denoted NeurIPS challenge (Lance et al., 2022): "no method works best for all". This was also observed in relation to the datasets/omics already available, where tools successful in some settings don't perform as well on others.

Single cell data integration will still be an active field in the years to come, as long as new assays will be developed, but we anticipate that the new tools will lose versatility to be more case-specific.

For the time being, the challenge would be also related to the number of omics a tool is able to manage. We believe that after a few attempts, the tough interpretability of the results and the redundance of information will lead people to care more about the quality of the data compared to their abundance.

BIBLIOGRAPHY

Amodio, M., & Krishnaswamy, S. (2018). MAGAN: Aligning biological manifolds. *35th International Conference on Machine Learning, ICML 2018*, *1*.

Amodio, M., Youlten, S. E., Venkat, A., San Juan, B. P., Chaffer, C. L., & Krishnaswamy, S. (2022). Single-cell multi-modal GAN reveals spatial patterns in single-cell data from triple-negative breast cancer. *Patterns*, 100577. https://doi.org/10.1016/j.patter.2022.100577

Argelaguet, R., Arnol, D., Bredikhin, D., Deloro, Y., Velten, B., Marioni, J. C., & Stegle, O. (2020). MOFA+: A statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biology*, *21*(1). https://doi.org/10.1186/s13059-020-02015-1

Argelaguet, R., Cuomo, A. S. E., Stegle, O., & Marioni, J. C. (2021). Computational principles and challenges in single-cell data integration. In *Nature Biotechnology*. Nature Research. https://doi.org/10.1038/s41587-021-00895-7

Arjovsky, M., & Bottou, L. (2017). Towards principled methods for training generative adversarial networks. *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*.

Arjovsky, M., Chintala, S., & Bottou, L. (2017). *Wasserstein Generative Adversarial Networks*.

Bock, C., Boutros, M., Camp, J. G., Clarke, L., Clevers, H., Knoblich, J. A., Liberali, P., Regev, A., Rios, A. C., Stegle, O., Stunnenberg, H. G., Teichmann, S. A., Treutlein, B., & Vries, R. G. J. (2021). The Organoid Cell Atlas. In *Nature Biotechnology* (Vol. 39, Issue 1). https://doi.org/10.1038/s41587-020-00762-x

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, *34*(5), 525–527. https://doi.org/10.1038/nbt.3519

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., & Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, *10*(12). https://doi.org/10.1038/nmeth.2688

Cao, K., Hong, Y., & Wan, L. (2021). Manifold alignment for heterogeneous single-cell multi-omics data integration using Pamona. *Bioinformatics*, *38*(1), 211–219. https://doi.org/10.1093/bioinformatics/btab594

Cao, Z. J., & Gao, G. (2022). Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nature Biotechnology*. https://doi.org/10.1038/s41587-022-01284-4

Chen, H., Lareau, C., Andreani, T., Vinyard, M. E., Garcia, S. P., Clement, K., Andrade-Navarro, M. A., Buenrostro, J. D., & Pinello, L. (2019). Assessment of computational methods for the analysis of single-cell ATAC-seq data. *Genome Biology*, *20*(1). https://doi.org/10.1186/s13059-019-1854-5

Chen, S., Lake, B. B., & Zhang, K. (2019). High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nature Biotechnology*, *37*(12), 1452–1457. https://doi.org/10.1038/s41587-019-0290-0

Cheow, L. F., Courtois, E. T., Tan, Y., Viswanathan, R., Xing, Q., Tan, R. Z., Tan, D. S. W., Robson, P., Loh, Y. H., Quake, S. R., & Burkholder, W. F. (2016). Single-cell multimodal profiling reveals cellular epigenetic heterogeneity. *Nature Methods*, *13*(10), 833–836. https://doi.org/10.1038/nmeth.3961

Cusanovich, D. A., Hill, A. J., Aghamirzaie, D., Daza, R. M., Pliner, H. A., Berletch, J. B., Filippova, G. N., Huang, X., Christiansen, L., DeWitt, W. S., Lee, C., Regalado, S. G., Read, D. F., Steemers, F. J., Disteche, C. M., Trapnell, C., & Shendure, J. (2018). A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. *Cell*, *174*(5), 1309-1324.e18. https://doi.org/10.1016/j.cell.2018.06.052

de Pretis, S., & Cittaro, D. (2022). *Dimensionality reduction and statistical modeling of scGET-seq data*. https://doi.org/10.1101/2022.06.29.498092

Demetci, P., Santorella, R., Sandstede, B., Noble, W. S., & Singh, R. (2022). SCOT: Single-Cell Multi-Omics Alignment with Optimal Transport. *Journal of Computational Biology : A Journal of Computational Molecular Cell Biology*, *29*(1). https://doi.org/10.1089/cmb.2021.0446

Dou, J., Liang, S., Mohanty, V., Cheng, X., Kim, S., Choi, J., Li, Y., Rezvani, K., Chen, R., Chen, K., & Org, K. (2020). *Unbiased integration of single cell multi-omics data*. https://doi.org/10.1101/2020.12.11.422014

Giansanti, V., Tang, M., & Cittaro, D. (2020). Fast analysis of scATAC-seq data using a predefined set of genomic regions. *F1000Research*, *9*, 199. https://doi.org/10.12688/f1000research.22731.1

Gong, B., Zhou, Y., & Purdom, E. (2021). Cobolt: integrative analysis of multimodal single-cell sequencing data. *Genome Biology*, *22*(1). https://doi.org/10.1186/s13059-021-02556-z

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). *Generative Adversarial Networks*. http://arxiv.org/abs/1406.2661

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. (2017a). Improved training of wasserstein GANs. *Advances in Neural Information Processing Systems*, *2017-December*.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. (2017b). *Improved Training of Wasserstein GANs*. http://arxiv.org/abs/1704.00028

Hao, Y., Stuart, T., Kowalski, M., Choudhary, S., Hoffman, P., Hartman, A., Srivastava, A., Molla, G., Madad, S., Fernandez-Granda, C., & Satija, R. (2022). *Dictionary learning for integrative, multimodal, and scalable single-cell analysis*. https://doi.org/10.1101/2022.02.24.481684

Hinton, G. E., Srivastava, N., & Swersky, K. (2012). *Neural Networks for Machine Learning Lecture 6a Overview of mini-batch gradient descent.* COURSERA: Neural Networks for Machine Learning.

Kashima, Y., Sakamoto, Y., Kaneko, K., Seki, M., Suzuki, Y., & Suzuki, A. (2020). Single-cell sequencing techniques from individual to multiomics analyses. In *Experimental and Molecular Medicine* (Vol. 52, Issue 9, pp. 1419–1427). Springer Nature. https://doi.org/10.1038/s12276-020-00499-2

Kingma, D. P., & Ba, J. L. (2015). Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*.

Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P. ru, & Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods*, *16*(12). https://doi.org/10.1038/s41592-019-0619-0

Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D. J., Hicks, S. C., Robinson, M. D., Vallejos, C. A., Campbell, K. R., Beerenwinkel, N., Mahfouz, A., Pinello, L., Skums, P., Stamatakis, A., Attolini, C. S. O., Aparicio, S., Baaijens, J., Balvert, M., Barbanson, B.

de, Cappuccio, A., … Schönhuth, A. (2020). Eleven grand challenges in single-cell data science. In *Genome Biology* (Vol. 21, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s13059-020-1926-6

Lance, C., Luecken, M. D., Burkhardt, D. B., Cannoodt, R., Rautenstrauch, P., Laddach, A., Ubingazhibov, A., Cao, Z.-J., Deng, K., Khan, S., Liu, Q., Russkikh, N., Ryazantsev, G., Ohler, U., Oliveira Pisco, A., Bloom, J., Krishnaswamy, S., & Theis, F. J. (2022). *Multimodal single cell data integration challenge: results and lessons learned CZ Biohub*. https://doi.org/10.1101/2022.04.11.487796

Liu, J., Huang, Y., Singh, R., Vert, J. P., & Noble, W. S. (2019). Jointly embedding multiple single-cell omics measurements. *Leibniz International Proceedings in Informatics, LIPIcs*, *143*. https://doi.org/10.4230/LIPIcs.WABI.2019.10

Liu, M.-Y., & Tuzel, O. (2016). *Coupled Generative Adversarial Networks*. http://arxiv.org/abs/1606.07536

Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., & Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. *Nature Methods*, *15*(12). https://doi.org/10.1038/s41592-018-0229-2

Ma, S., Zhang, B., LaFave, L. M., Earl, A. S., Chiang, Z., Hu, Y., Ding, J., Brack, A., Kartha, V. K., Tay, T., Law, T., Lareau, C., Hsu, Y. C., Regev, A., & Buenrostro, J. D. (2020). Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. *Cell*, *183*(4), 1103-1116.e20. https://doi.org/10.1016/j.cell.2020.09.056

Macaulay, I. C., Haerty, W., Kumar, P., Li, Y. I., Hu, T. X., Teng, M. J., Goolam, M., Saurat, N., Coupland, P., Shirley, L. M., Smith, M., van der Aa, N., Banerjee, R., Ellis, P. D., Quail, M. A., Swerdlow, H. P., Zernicka-Goetz, M., Livesey, F. J., Ponting, C. P., & Voet, T. (2015). G&T-seq: Parallel sequencing of single-cell genomes and transcriptomes. *Nature Methods*, *12*(6), 519–522. https://doi.org/10.1038/nmeth.3370

Marouf, M., Machart, P., Bansal, V., Kilian, C., Magruder, D. S., Krebs, C. F., & Bonn, S. (2020). Realistic in silico generation and augmentation of single-cell RNA-seq data using generative adversarial networks. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-019-14018-z

Martinez-De-Morentin, X., Khan, S. A., Lehmann, R., Qu, S., Maillo, A., Kiani, N. A., Prosper, F., Tegner, J., & Gomez-Cabrero, D. (2021). *Adaptative Machine Translation between paired Single-Cell Multi-Omics Data*. https://doi.org/10.1101/2021.01.27.428400

Mimitou, E. P., Lareau, C. A., Chen, K. Y., Zorzetto-Fernandes, A. L., Hao, Y., Takeshima, Y., Luo, W., Huang, T. S., Yeung, B. Z., Papalexi, E., Thakore, P. I., Kibayashi, T., Wing, J. B., Hata, M., Satija, R., Nazor, K. L., Sakaguchi, S., Ludwig, L. S., Sankaran, V. G., … Smibert, P. (2021). Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. *Nature Biotechnology*, *39*(10), 1246–1258. https://doi.org/10.1038/s41587-021-00927-2

Minoura, K., Abe, K., Nam, H., Nishikawa, H., & Shimamura, T. (2021). A mixture-of-experts deep generative model for integrated analysis of single-cell multiomics data. *Cell Reports Methods*, *1*(5). https://doi.org/10.1016/j.crmeth.2021.100071

Morelli, L., Giansanti, V., & Cittaro, D. (2021). Nested Stochastic Block Models applied to the analysis of single cell data. *BMC Bioinformatics*, *22*(1). https://doi.org/10.1186/s12859-021-04489-7

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011). *Multimodal Deep Learning*.

Ogbeide, S., Giannese, F., Mincarelli, L., & Macaulay, I. C. (2022). Into the multiverse: advances in single-cell multiomic profiling. In *Trends in Genetics*. Elsevier Ltd. https://doi.org/10.1016/j.tig.2022.03.015

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, *14*(4), 417–419. https://doi.org/10.1038/nmeth.4197

Pennisi E. (2012). The biology of genomes. Single-cell sequencing tackles basic and biomedical questions. In *Science* (Vol. 336, Issue 6084, pp. 976–977). American Association for the Advancement of Science. https://doi.org/10.1126/science.336.6084.976

Radford, A., Metz, L., & Chintala, S. (2015). *Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks*. http://arxiv.org/abs/1511.06434

Reddi, S. J., Kale, S., & Kumar, S. (2018). On the convergence of Adam and beyond. *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*.

Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., Clevers, H., Deplancke, B., Dunham, I., Eberwine, J., Eils, R., Enard, W., Farmer, A., Fugger, L., Göttgens, B., … Yosef, N. (2017). The human cell atlas. *ELife*, *6*. https://doi.org/10.7554/eLife.27041

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). Improved techniques for training GANs. *Advances in Neural Information Processing Systems*.

Schaum, N., Karkanias, J., Neff, N. F., May, A. P., Quake, S. R., Wyss-Coray, T., Darmanis, S., Batson, J., Botvinnik, O., Chen, M. B., Chen, S., Green, F., Jones, R. C., Maynard, A., Penland, L., Pisco, A. O., Sit, R. v., Stanley, G. M., Webber, J. T., … Weissman, I. L. (2018). Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature*, *562*(7727). https://doi.org/10.1038/s41586-018-0590-4

Shapiro, E., Biezuner, T., & Linnarsson, S. (2013). Single-cell sequencing-based technologies will revolutionize whole-organism science. In *Nature Reviews Genetics* (Vol. 14, Issue 9, pp. 618–630). https://doi.org/10.1038/nrg3542

Stark, S. G., Ficek, J., Locatello, F., Bonilla, X., Chevrier, S., Singer, F., Rätsch, G., Lehmann, K. van, Rudolf, A., Al-Quaddoomi Faisal, S., Jonas, A., Ilaria, A., Sonali, A., Per-Olof, A., Marina, B., Daniel, B., Beatrice, B. S., Niko, B., Christian, B., … Gregor, Z. (2020). SCIM: Universal single-cell matching with unpaired feature sets. *Bioinformatics*, *36*. https://doi.org/10.1093/bioinformatics/btaa843

Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P. K., Swerdlow, H., Satija, R., & Smibert, P. (2017). Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, *14*(9), 865–868. https://doi.org/10.1038/nmeth.4380

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., Hao, Y., Stoeckius, M., Smibert, P., & Satija, R. (2019). Comprehensive Integration of Single-Cell Data. *Cell*, *177*(7), 1888-1902.e21. https://doi.org/10.1016/j.cell.2019.05.031

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., & Mesirov, J. P. (2005). *Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles*. www.pnas.orgcgidoi10.1073pnas.0506580102

Svensson, V., Vento-Tormo, R., & Teichmann, S. A. (2018). Exponential scaling of single-cell RNA-seq in the past decade. In *Nature Protocols* (Vol. 13, Issue 4, pp. 599–604). Nature Publishing Group. https://doi.org/10.1038/nprot.2017.149

Swanson, E., Lord, C., Reading, J., Heubeck, A. T., Genge, P. C., Thomson, Z., Weiss, M. D. A., Li, X. J., Savage, A. K., Green, R. R., Torgerson, T. R., Bumol, T. F., Graybuck, L. T., & Skene, P. J. (2021). Simultaneous trimodal single-cell measurement of transcripts, epitopes, and chromatin accessibility using tea-seq. *ELife*, *10*. https://doi.org/10.7554/ELIFE.63632

Tedesco, M., Giannese, F., Lazarević, D., Giansanti, V., Rosano, D., Monzani, S., Catalano, I., Grassi, E., Zanella, E. R., Botrugno, O. A., Morelli, L., Panina Bordignon, P., Caravagna, G., Bertotti, A., Martino, G., Aldrighetti, L., Pasqualato, S., Trusolino, L., Cittaro, D., & Tonon, G. (2022). Chromatin Velocity reveals epigenetic dynamics by single-cell profiling of heterochromatin and euchromatin. *Nature Biotechnology*, *40*(2), 235–244. https://doi.org/10.1038/s41587-021-01031-1

Traag, V. A., Waltman, L., & van Eck, N. J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-41695-z

Tran, H. T. N., Ang, K. S., Chevrier, M., Zhang, X., Lee, N. Y. S., Goh, M., & Chen, J. (2020). A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biology*, *21*(1). https://doi.org/10.1186/s13059-019-1850-9

Villani, C. (2007). Optimal Transport Old and New. *Media*, *338*.

Wang, Y., & Navin, N. E. (2015). Advances and Applications of Single-Cell Sequencing Technologies. In *Molecular Cell* (Vol. 58, Issue 4, pp. 598–609). Cell Press. https://doi.org/10.1016/j.molcel.2015.05.005

Wu, K. E., Yost, K. E., Chang, H. Y., & Zou, J. (2021). BABEL enables cross-modality translation between multiomic profiles at single-cell resolution. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(15). https://doi.org/10.1073/pnas.2023070118

Xu, Y., Begoli, E., & McCord, R. P. (2021). sciCAN: Single-cell chromatin accessibility and gene expression data integration via Cycle-consistent Adversarial Network. *BioRxiv*.

Zappia, L., Phipson, B., & Oshlack, A. (2017). Splatter: Simulation of single-cell RNA sequencing data. *Genome Biology*, *18*(1). https://doi.org/10.1186/s13059-017-1305-0

Zhang, B., Srivastava, A., Mimitou, E., Stuart, T., Raimondi, I., Hao, Y., Smibert, P., & Satija, R. (2022). Characterizing cellular heterogeneity in chromatin state with scCUT&Tag-pro. *Nature Biotechnology*, *40*(8), 1220–1230. https://doi.org/10.1038/s41587-022-01250-0

Zhang, K., Hocker, J. D., Miller, M., Hou, X., Chiou, J., Poirion, O. B., Qiu, Y., Li, Y. E., Gaulton, K. J., Wang, A., Preissl, S., & Ren, B. (2021). A single-cell atlas of chromatin accessibility in the human genome. *Cell*, *184*(24). https://doi.org/10.1016/j.cell.2021.10.024

Zhao, J., Wang, G., Ming, J., Lin, Z., Wang, Y., Consortium, T. M., Wu, A. R., & Yang, C. (2021). Adversarial domain translation networks enable fast and accurate large-scale atlas-level single-cell data integration. *BioRxiv*.

Zhu, C., Yu, M., Huang, H., Juric, I., Abnousi, A., Hu, R., Lucero, J., Behrens, M. M., Hu, M., & Ren, B. (2019). An ultra high-throughput method for single-cell joint analysis of open chromatin and transcriptome. *Nature Structural and Molecular Biology*, *26*(11), 1063–1070. https://doi.org/10.1038/s41594-019-0323-x

Zuo, C., & Chen, L. (2021). Deep-joint-learning analysis model of single cell transcriptome and open chromatin accessibility data. *Briefings in Bioinformatics*, *22*(4). https://doi.org/10.1093/bib/bbaa287

ANNEXES

In the following, the original papers published during the PhD are attached:

1. Chromatin Velocity reveals epigenetic dynamics by single-cell profiling of heterochromatin and euchromatin (Tedesco et al., 2022).
2. Fast analysis of scATAC-seq data using a predefined set of genomic regions (Giansanti et al., 2020).
3. Nested Stochastic Block Models applied to the analysis of single cell data (Morelli et al., 2021).

ANNEX I: CHROMATIN VELOCITY REVEALS EPIGENETIC DYNAMICS BY SINGLE-CELL PROFILING OF HETEROCHROMATIN AND EUCHROMATIN.

nature biotechnology

ARTICLES

https://doi.org/10.1038/s41587-021-01031-1

Check for updates

Chromatin Velocity reveals epigenetic dynamics by single-cell profiling of heterochromatin and euchromatin

Martina Tedesco[1,2,12], Francesca Giannese[3,12], Dejan Lazarević[3], Valentina Giansanti[3,4], Dalia Rosano[2,11], Silvia Monzani[5], Irene Catalano[6,7], Elena Grassi[6,7], Eugenia R. Zanella[7], Oronza A. Botrugno[2], Leonardo Morelli[3], Paola Panina Bordignon[1,8], Giulio Caravagna[9], Andrea Bertotti[6,7], Gianvito Martino[1,8], Luca Aldrighetti[10], Sebastiano Pasqualato[5], Livio Trusolino[6,7], Davide Cittaro[3] ✉ and Giovanni Tonon[2,3] ✉

Recent efforts have succeeded in surveying open chromatin at the single-cell level, but high-throughput, single-cell assessment of heterochromatin and its underlying genomic determinants remains challenging. We engineered a hybrid transposase including the chromodomain (CD) of the heterochromatin protein-1α (HP-1α), which is involved in heterochromatin assembly and maintenance through its binding to trimethylation of the lysine 9 on histone 3 (H3K9me3), and developed a single-cell method, single-cell genome and epigenome by transposases sequencing (scGET-seq), that, unlike single-cell assay for transposase-accessible chromatin with sequencing (scATAC-seq), comprehensively probes both open and closed chromatin and concomitantly records the underlying genomic sequences. We tested scGET-seq in cancer-derived organoids and human-derived xenograft (PDX) models and identified genetic events and plasticity-driven mechanisms contributing to cancer drug resistance. Next, building upon the differential enrichment of closed and open chromatin, we devised a method, Chromatin Velocity, that identifies the trajectories of epigenetic modifications at the single-cell level. Chromatin Velocity uncovered paths of epigenetic reorganization during stem cell reprogramming and identified key transcription factors driving these developmental processes. scGET-seq reveals the dynamics of genomic and epigenetic landscapes underlying any cellular processes.

Cancers are characterized by extensive interindividual and intratumor heterogeneity down to the single-cell level[1]. This fuels clonal evolution and treatment resistance[2], the leading cause of death for individuals with cancer. The mechanisms underlying such resistance are still largely unknown, especially for standard chemotherapeutic and immunotherapeutic regimens. Increasingly detailed analyses of cancer genomes before and after treatment have so far failed to identify genetic causes that could explain the ensuing refractoriness to therapy. Recently, epigenetic changes have emerged as key contributors of drug resistance in cancer[3–8], suggesting that only a comprehensive assessment of the genetic changes of the cancer genome, including somatic mutations and copy number changes, alongside a detailed description of the concomitant chromatin remodeling events that ensue after treatment could provide the insights required to tackle this pressing unmet clinical need.

As for single-cell epigenetics, the recent introduction of transposases such as Tn5, which allow for the fragmenting and sequencing of native accessible chromatin in bulk (ATAC-seq[9]) as well as at the single-cell level (scATAC-seq[10]), is providing key insights into the cellular status of open chromatin. However, the epigenetic modifications of large portions of the genome that have essential roles in cellular physiology are excluded from this analysis. For instance, to our knowledge, there are no single-cell methods able to probe compacted chromatin, that is, heterochromatin, which encompasses up to half of the entire genome[11] and harbors and regulates a large array of transposable elements and non-coding RNAs (ncRNAs)[11–13]. Heterochromatin is assembled and maintained through H3K9me3 (refs. [12,14]), and its accurate regulation is essential for cells, for example, contributing toward the definition of cell identity[12,13] and the maintenance of genomic integrity[15].

While single-cell transcriptomic analysis has fostered ground-breaking insights into the biology of healthy and diseased tissues, including cancer[16,17], to our knowledge, a tool that comprehensively audits at the single-cell level both the genomic and the epigenetic landscape has not been reported.

Results

Tn5 is able to tagment compacted chromatin featuring H3K9me3. We first determined whether Tn5 is able to tagment compacted chromatin if properly redirected. To this end, we exploited a transposase-assisted chromatin multiplex immunoprecipitation (TAM-ChIP) approach, which combines the

[1]Università Vita-Salute San Raffaele, Milano, Italy. [2]Functional Genomics of Cancer Unit, Division of Experimental Oncology, IRCCS San Raffaele Scientific Institute, Milano, Italy. [3]Center for Omics Sciences, IRCCS San Raffaele Institute, Milano, Italy. [4]Department of Informatics, Systems and Communication, University of Milano-Bicocca, Milano, Italy. [5]Biochemistry and Structural Biology Unit, Department of Experimental Oncology, IEO, IRCCS European Institute of Oncology, Milano, Italy. [6]Department of Oncology, University of Torino School of Medicine, Candiolo, Torino, Italy. [7]Candiolo Cancer Institute FPO- IRCCS, Candiolo, Torino, Italy. [8]Neuroimmunology Unit, Institute of Experimental Neurology, Division of Neuroscience, IRCCS San Raffaele Hospital, Milano, Italy. [9]Department of Mathematics and Geosciences, University of Trieste, Trieste, Italy. [10]Hepatobiliary Surgery Division, IRCCS San Raffaele Hospital, Milano, Italy. [11]Present address: Department of Surgery and Cancer, Imperial College London, London, UK. [12]These authors contributed equally: Martina Tedesco, Francesca Giannese. ✉e-mail: cittaro.davide@hsr.it; tonon.giovanni@hsr.it

antibody-mediated targeting of chromatin immunoprecipitation with the ability of Tn5 to tagment DNA, leading to chromatin fragmentation and barcoding of the chromatin surrounding the antibody binding site (Extended Data Fig. 1a). We choose a primary antibody that recognizes the histone mark H3K9me3 (or H3K4me3 used as a control), in line with a recent report[18], that was then bound by a secondary antibody conjugated to Tn5. H3K4me3 TAM-ChIP–seq profiles mirrored the corresponding H3K4me3 chromatin immunoprecipitation sequencing (ChIP–seq) profiles. Instead, when a Tn5–secondary antibody complex that recognizes H3K9me3-specific primary antibody was used, Tn5 tagmented H3K9me3-enriched compacted chromatin regions (Extended Data Fig. 1b), which was confirmed by real-time quantitative PCR (RT–qPCR) (Extended Data Fig. 1c).

Together, these experiments demonstrate that Tn5, if properly redirected, is able to sever and tag H3K9me3-compacted chromatin.

Hybrid CD HP-1α–Tn5 targets H3K9me3 chromatin regions. TAM-ChIP towards H3K9me3 was only partially effective in guiding Tn5 transposase toward closed chromatin. Additionally, this approach relies on immunoprecipitation, which poses technical challenges. We hence reasoned that the most straightforward approach to target compacted chromatin would entail the modification of the natural tropism of Tn5. To this end, we extensively reviewed proteins and domains targeting H3K9me3. We then selected HP-1α, one of the hallmark proteins involved in heterochromatin assembly and maintenance that specifically binds H3K9me3 through its CD[19–21].

We generated a hybrid protein whereby the HP-1α CD was cloned alongside Tn5 (Extended Data Fig. 2a). To link the CD with Tn5 transposase, we took advantage of the natural linker that connects the CD and the chromoshadow domain of HP-1α, which we extended with two artificial linkers of different length (TnH 1–TnH 4; Extended Data Fig. 2a). All four hybrid constructs were as efficient as the native Tn5 (either the commercial Nextera enzyme or in-house produced enzyme (hereafter, Tn5)) to fragment and insert oligos into genomic DNA (gDNA; Extended Data Fig. 2b).

We then determined whether TnH 1–TnH 4 were able to target chromatin harboring H3K9me3 histone modifications by tagmenting native chromatin on permeabilized nuclei (Extended Data Fig. 2c). Unlike Nextera and Tn5 enzymes, hybrid Tn5 constructs indeed cut and inserted oligos in regions enriched for H3K9me3 while retaining affinity toward accessible sequences (Fig. 1a,b and Extended Data Fig. 2d,e). We identified the construct TnH 3 (hereafter referred to as TnH) as the most efficient (Fig. 1b and Extended Data Fig. 2d,e).

We next reasoned that combining Tn5 and TnH in a single experiment could provide a comprehensive perspective of both accessible and compacted chromatin (Fig. 1c). We thus loaded each of the two transposases with a set of specific barcoded oligos to discriminate Tn5 from TnH tagmentation products (Fig. 1c). We then tested the effect of varying the Tn5-to-TnH ratio (Extended Data Fig. 3a) or adding the two enzymes sequentially (Extended Data Fig. 3b) on the transposition reaction. The sequential use of native Tn5 followed by TnH provided the most comprehensive mapping of the two chromatin profiles.

Together, these results demonstrate that a sequential combination of Tn5 and TnH is able to differentiate accessible versus compacted chromatin, thus defining the whole-genome epigenetic distribution of euchromatin and heterochromatin. We call this method GET-seq (genome and epigenome by transposases sequencing).

GET-seq at the single-cell level (scGET-seq). We then attempted to implement this method to single-cell analysis. To obtain droplet-based scGET-seq, we modified the Chromium Single Cell ATAC v1 protocol (10x Genomics) and replaced the provided ATAC transposition enzyme (10x Tn5, 10x Genomics) with Tn5 and TnH in appropriate enzyme proportions.

We first assessed the distribution of reads assigned to unique cell barcodes by using 10x Tn5, TnH, Tn5 or a combination of TnH and Tn5 (scGET-seq) in Caki-1 cells and found that the four profiles were overlapping (Extended Data Fig. 4a). We next explored the portion of the genome that was captured by each transposase. TnH had the higher mean distribution of coverage per cell with a smaller standard deviation than either Tn5 or 10x Tn5 (Extended Data Fig. 4b), suggesting that, even at the single-cell level, TnH captures genome areas that are not targeted by conventional transposases. Indeed, when single-cell Tn5 and TnH data were each combined in pseudobulks and compared to the ChIP–seq data obtained in the same cells using H3K9me3 and H3K4me3 antibodies, TnH was able to target regions positive for H3K9me3 as well as H3K4me3 (Extended Data Fig. 4d), in line with the bulk TnH results (Fig. 1a).

We then determined whether scGET-seq was able to capture cell identity. To this end, we sequenced a mixture of HeLa (20%) and Caki-1 (80%) cancer cell lines, which originate from different tissues (cervix and kidney, respectively). Cells were clearly separated in two clusters sized with the expected proportions (Fig. 2a).

To further confirm the identity of the clusters, we used available bulk ATAC-seq data for both cell lines and generated a score for each cell line. The respective scores clearly distinguished each cell line cluster (Fig. 2a), in accordance with standard scATAC-seq results (Fig. 2b).

Together, these data confirm that GET-seq can be applied to droplet-based single-cell approaches and is able to easily differentiate cells derived from different genetic backgrounds.

Genomic copy number variants (CNVs) at the single-cell level. The definition of genomic CNVs using scATAC-seq remains imprecise because only accessible chromatin regions are surveyed by this approach, and the remaining genomic sequences can only be imputed from adjacent regions[22].

As TnH also targets H3K9me3-enriched chromatin regions, we tested whether it could also be harnessed to define CNVs. Whole-genome sequencing (WGS) revealed several CNVs in both cell lines (fraction of genome altered, Caki-1 = 0.475 and HeLa = 0.508). The correlation between the genomic profiles obtained with WGS and the average pseudobulk profile obtained from single-cell data was much higher for the TnH signal than for the 10x Tn5 signal at various resolutions (Fig. 2c and Extended Data Fig. 5).

A closer inspection of the segmentation profiles at the single-cell level revealed that scATAC-seq is able to define CNVs at a coarse resolution (10 Mb), as previously determined[22]. Even at this resolution, scGET-seq showed a much higher consistency for both cell lines than 10x Tn5 (Extended Data Fig. 5c). After increasing the resolution up to 500 kb, scGET-seq remained reliable while the ability of scATAC-seq to identify CNVs degraded, and large swaths of the genome were excluded from the analysis (Extended Data Fig. 5a,b). In fact, the signal emerging from scATAC-seq correlated closely with the location of regulatory elements throughout the genome, unlike scGET-seq (Fig. 2d).

We tested the ability of scGET and 10x to call CNV events using a machine learning approach. To this end, we called CNVs from bulk WGS data of Caki-1 and HeLa cells. We then split scGET-seq and scATAC-seq genomic bins into training and test sets (proportion 70:30) and trained a logistic regression classifier and a support vector machine with linear kernel (SVM). We calculated their accuracy and F1 scores on the test set. scGET-seq performed better than scATAC-seq regardless of the classifier and the resolution, with the performance depending on the number of cells included in the analysis (Fig. 2e).

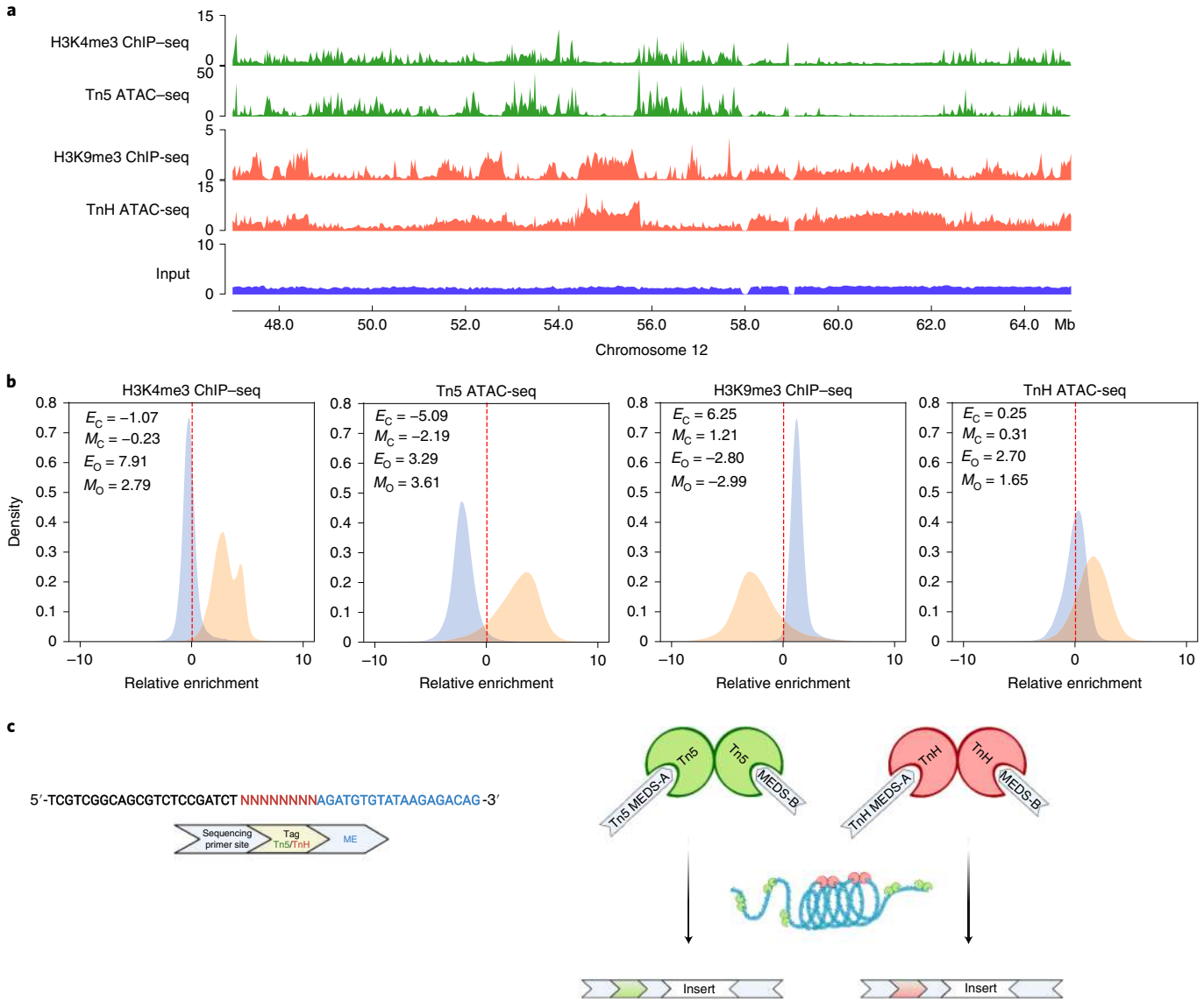


Fig. 1 | The Tn5 transposon is able to target H3K9me3-enriched regions. a, Enrichment profile of H3K4me3-associated (green) and H3K9me3-associated (red) regions obtained by ChIP-seq compared to Tn5 (green) and TnH (red) tagmentation profile obtained by ATAC-seq. The ChIP-seq input track is shown as a control (violet); Mb, megabases. **b**, Distribution of the enrichment of Tn5 and TnH transposons relative to genomic background in regions enriched for H3K4me3 (orange) or H3K9me3 (blue) expressed as $\log_2$ (ratio) of the signal over the genomic input. Enrichment over the same regions for H3K4me3 and H3K9me3 ChIP-seq are reported as reference; E_c, global enrichment over H3K9me3-marked regions; E_o, global enrichment over H3K4me3-marked regions; M_c, modal enrichment over H3K9me3-marked regions; M_o, modal enrichment over H3K4me3-marked regions. **c**, General schematic of the GET-seq transposon structure. Standard Tn5ME-A oligo was replaced by 49-nucleotide (nt) oligos composed of 22 nt for read 1 sequencing primer binding, 8-nt tags to discriminate Tn5 from TnH tagmentation products and standard 19-bp mosaic end (ME) sequence for transposase binding (created with BioRender.com). The data shown refer to experiments performed on Caki-1 cells; MEDS, mosaic end double-stranded oligonucleotides.

Together, these data show the feasibility of single-cell profiling by GET-seq, which allows for a more precise description of genomic features than scATAC-seq.

scGET-seq identifies clonality in human-derived organoids. To ascertain the ability of GET-seq to define clonality, we decided to rely on a more physiological experimental setting than cell lines, human-derived organoids (PDOs). We thus used a tumor–normal matched design to generate whole-exome data derived from two hepatic metastases of primary colorectal tumors. The analysis of somatic single-nucleotide variants (SNVs) and allele-specific copy numbers showed high levels of aneuploidy for both samples (CRC6, triploid; CRC17, tetraploid). From the analysis of allele frequency spectra and cancer cell fractions, we found no evidence of ongoing subclonal expansions, concluding that CRC6 and CRC17 are monoclonal, a common characteristic of late-stage colorectal cancer[23,24] (Extended Data Fig. 6a). From these samples, we generated PDOs (Extended Data Fig. 6b), which we then profiled with scGET-seq. The CNV analysis confirmed the existence of two main cellular populations with defining genomic features, closely mimicking the two CRC6 and CRC17 cancer populations (Fig. 3a and Extended Data Fig. 6c). To provide quantitative support to this observation, we also calculated the posterior marginal probability distribution of the number of observable clones. This analysis confirmed that scGET-seq could correctly identify two clusters, corresponding to CRC6 and CRC17. Notably, only a minority of the cells assessed were misclassi-

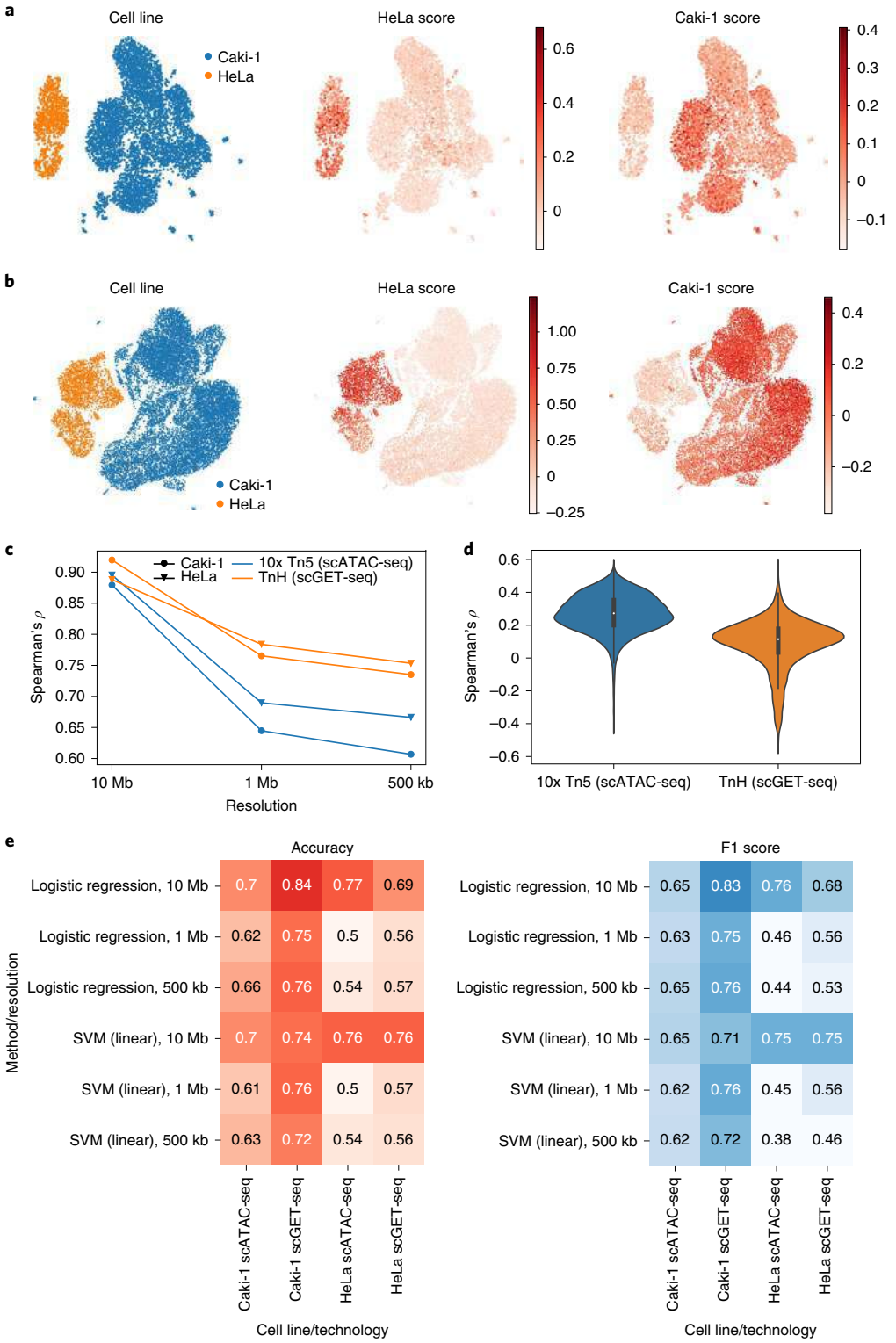


Fig. 2 | Assessment of scGET-seq strategy and genomic copy number at the single-cell level. a, Uniform manifold approximation and projection (UMAP) embedding showing individual cells in a mixture of Caki-1 and HeLa cells at known proportions (80:20) profiled by scGET-seq. Cells are identified according to a signature calculated on specific DNase I hypersensitive sites (DHS) identified from bulk studies. **b**, UMAP embedding showing individual cells in a mixture of Caki-1 and HeLa cells at known proportions (80:20) profiled by standard scATAC-seq. Cells are identified according to a signature calculated on specific DHS identified from bulk studies**. c**, Spearman's correlation values between the segmentation profile of Caki-1 and HeLa cells at increasing resolution. The signal from bulk sequencing was compared to the average cell signal obtained in single-cell profiling. scGET-seq (orange) shows consistently higher correlation than standard scATAC-seq (blue); kb, kilobases. **d**, Spearman's correlation values between the segmentation profiles and the density of regulatory elements in the GeneHancer catalog. White dots in the box plots represent the median, boxes span between the 25th and 75th percentiles and whiskers extend 1.5× the interquartile range; $n = 323$ regions. **e**, Heat map showing the performance of two different classifiers on genomic alterations (amplifications, deletions and normal segments) in HeLa and Caki-1 cells. Each classifier has been trained at increasing resolution on scGET-seq and scATAC-seq data separately. Both classifiers perform worse on HeLa cells than in Caki-1 cells given the lower numerosity.

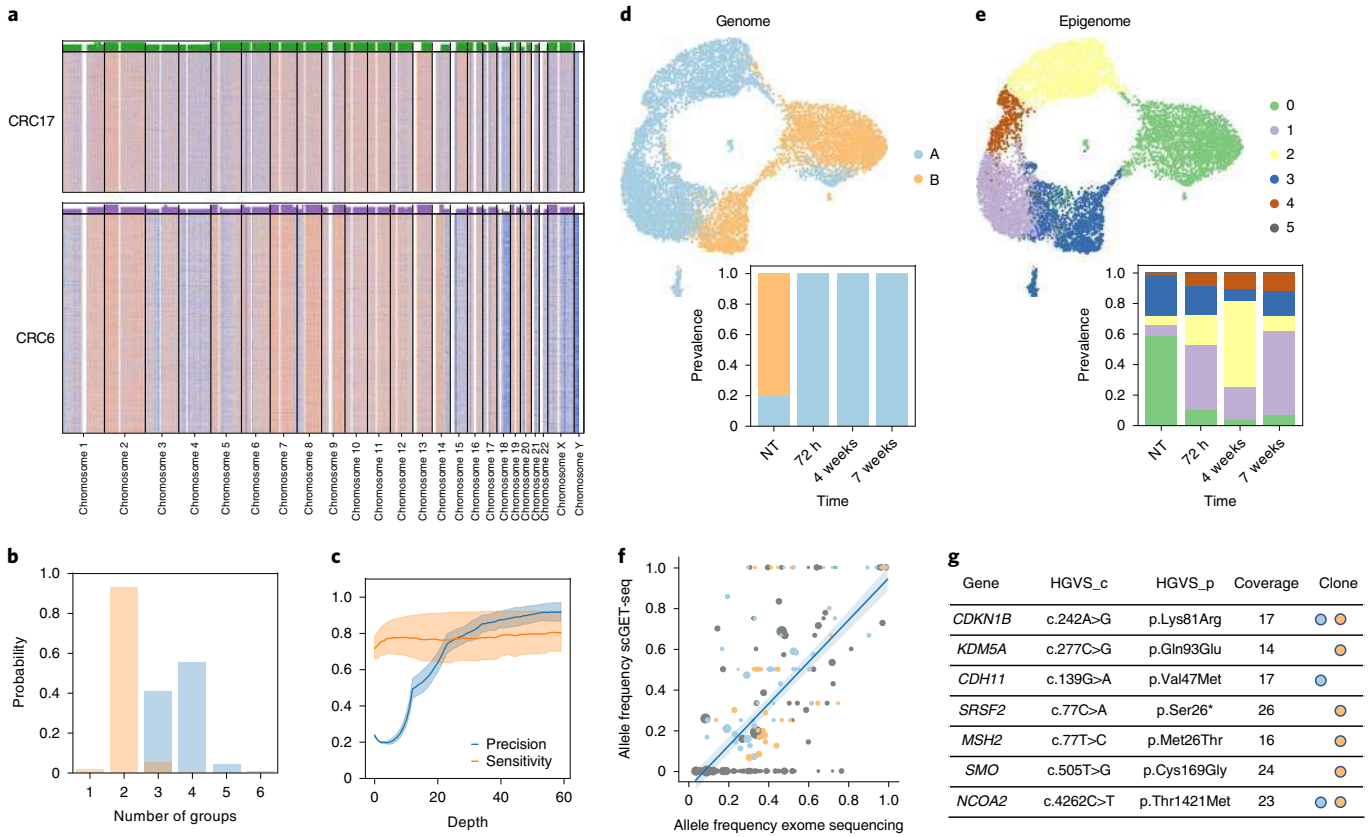


Fig. 3 | Analysis of human-derived samples by scGET-seq. a, Segmentation profile in individual cells profiled by scGET-seq of two PDOs (CRC6 and CRC17). The heat maps show the genomic landscape of two discovered clones assigned to each organoid. scGET-seq data are expressed as normalized $\log_2$ (ratio) of the signal in 1-Mb windows with respect to the average per cell coverage. Centromeric regions and genome gaps were excluded from the analysis and are colored in white. Bar plots on the top of each heat map represent the absolute copy number evaluated from whole-exome sequencing. **b**, Distribution of the marginal posterior probability of the number of cell clusters identified using TnH-derived reads (orange) or Tn5-derived reads (blue). Analysis of clonal structure with Tn5-derived reads, as in scATAC-seq, may lead to overclustering. **c**, Analysis of the performance of variant calling in PDO samples as a function of coverage on the profiled variants. The shaded interval represents the range of values for two samples, and the solid line represents the geometric mean. Sensitivity is calculated as TP/(TP + FN), and precision is calculated as TP/(TP + FP); TP, correctly identified alleles; FP, alleles identified by scGET-seq and not by exome sequencing; FN, alleles identified by exome sequencing and not by scGET-seq. Depth threshold is applied on variants profiled by scGET-seq. **d**,**e**, UMAP embeddings of scGET-seq profiles of individual cells derived from PDX samples. Cells are colored according to the clones derived from segmentation data (**d**) or epigenome analysis (**e**). Below each UMAP embedding, a bar plot represents the abundance of subpopulations over time; NT, not treated (time zero). **f**, Scatter plot of allele frequency of somatic mutations identified by whole-exome sequencing of the primary tumor in relation to the allele frequency detected by genotyping scGET-seq data. Dot size is proportional to coverage in scGET-seq, while color matches the clones in **d**; gray dots are mutations shared by two clones (Pearson's $r = 0.712$, $P = 7.93 \times 10^{-38}$, $n = 389$). **g**, Representative mutations of COSMIC hallmark genes found in scGET-seq data that were not present in the primary tumor. Each mutation is associated with the corresponding genetic clone using the appropriate color code.

fied (Supplementary Table 1). A similar analysis on Tn5-derived reads showed a tendency for overclustering and cell misclassification (Fig. 3b and Supplementary Table 1). We finally explored the accuracy of variant calling (that is, presence/absence of a variant) by comparing genotyped clones with known variants profiled in the bulk samples. We found that the dependency of precision and sensitivity at different depth thresholds were in line with previous observations[25], although values were slightly smaller and sample dependent (Fig. 3c).

Together, these results suggest that scGET-seq can be successfully used to concomitantly obtain detailed information on the single-cell epigenetic landscape as well on the underlying genomic structure.

Genomic and epigenetic landscape of resistant cancer clones. To exploit the ability of scGET-seq to capture the genomic and epigenetic landscape of single cells, we used PDX models of colon carcinoma where we have shown that resistance to therapy may arise from the selection of clones endowed with specific genetic lesions along with features of plasticity that are not driven by genomic modifications but most likely by chromatin reshaping[26,27]. We therefore followed cancer evolution in one PDX model throughout several weeks of treatment with the clinically approved epidermal growth factor receptor (EGFR) antibody cetuximab (Extended Data Fig. 7a). Analysis of genomic segmentation by scGET-seq revealed two major clones in the absence of treatment (Fig. 3d and Extended Data Fig. 7b). Conversely, cells were separated into six different clones when assessing the pretreatment epigenetic landscape (Fig. 3e). When the impact of treatment was assayed, clone A was predominant, while clone B was present at very low frequency (Fig. 3d). By contrast, the epigenetic landscape of cetuximab-treated PDX samples was more heterogenous, with epigenetic subclones embedded within genetic clones (Fig. 3e).

We next sought to identify processes that might provide biological insights into epigenetic mechanisms of resistance to EGFR

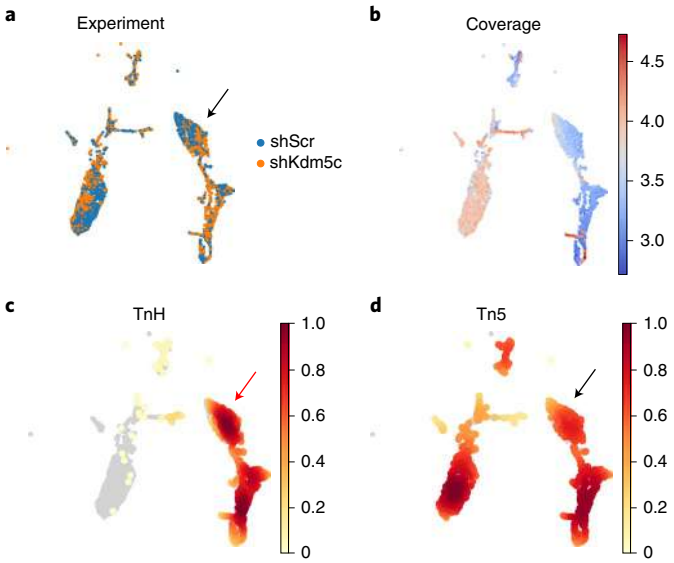


Fig. 4 | scGET-seq profiling of NIH-3T3 cells after *Kdm5c* knockdown. **a**, UMAP embedding showing the location of cells transfected with shKdm5c or shScr. **b**, UMAP embedding of individual cells colored by read coverage. Two main clusters appear depending on the coverage. **c**,**d**, UMAP embedding highlighting the density of cells with high signal over pericentromeric heterochromatin marked by the major primer (see text), as recovered by TnH (**c**) or Tn5 (**d**). The two signals are unevenly distributed and tend to localize where there are higher amounts of shScr cells. All data refer to experiments performed in NIH-3T3 cells.

blockade. To this end, we performed functional enrichment analysis using the genes associated with the regions that were differentially affected in the various clones (Supplementary Table 2). In the epigenetic clones most associated with resistance, there was a significant enrichment of pathways linked to refractoriness to EGFR inhibitors, including the phospholipase C pathway[28], transforming growth factor-β (TGF-β) signaling[29] and the WNT pathway[30] (Extended Data Fig. 7c). These results are in line with our previous observations that cancer cells exposed to targeted therapies do show resistance patterns related to genomic plasticity phenotypes, most likely driven by chromatin remodeling phenomena[26,27].

As scGET-seq includes sequences for portions of the genome that are eluded by conventional ATAC-seq, we next sought to determine whether we could also define SNVs within single cells. Not all exome SNVs were captured by scGET-seq; nonetheless, there was a highly significant correlation between the mutations identified by bulk exome sequencing conducted on the primary tumor and the scGET-seq results (Fig. 3f). Notably, by virtue of the single-cell analysis, it was possible to ascribe the mutations to specific clones.

scGET-seq was also able to identify mutations not present in the initial bulk exome sequencing in the starting sample and mutations that affected established cancer genes (tier 1, COSMIC Cancer Gene Census, version 92 (ref. [31]); Supplementary Table 3), including *CDKN1B*, *KDM5A*, *CDH11*, *SRSF2*, *MSH2*, *SMO* and *NCOA2* (Fig. 3g) (the enrichment for COSMIC mutations was significant for variants profiled at high depth, that is, higher than 15; odds ratio = 1.55; $P = 3.57 \times 10^{-3}$, Fisher's exact test). At this stage, it remains to be ascertained whether the mutations that were found by single-cell analysis but not by bulk sequencing were developed de novo by the PDX or were already present in the original population at frequencies too low to be detected by the limited coverage of exome sequencing.

Together, these results suggest that scGET-seq could be used to comprehensively assess the tumor genome (including both CNVs and SNVs) and the epigenome, illuminating paths of cancer evolution, clonality and drug resistance.

scGET-seq captures chromatin status at the single-cell level. We next determined whether scGET-seq might capture the dynamics between accessible and compacted chromatin at the single-cell level. We have recently demonstrated that ablation of the histone demethylase Kdm5c hampers H3K9me3 deposition, impairing heterochromatin assembly and maintenance in NIH-3T3 cells[32]. We performed scGET-seq in cells before and after *Kdm5c* knockdown. We identified two neatly distinguished cell groups, including short hairpin scramble (shScr) and shKdm5c cells, respectively (Fig. 4a). Seeking to find an explanation for this pattern, we discovered that this distinction was driven by the total number of reads per cell (Fig. 4b). We surmised that this pattern might be driven by the cell cycle status, namely, high coverage associated with cells in the S and G2/M cycle phases during or after DNA replication and low coverage linked to cells in the G1 cycle phase before the replication of DNA. To test our hypothesis, we applied a strategy derived from ref. [10] where we analyzed the distribution of Repli-seq[33–35] signal over differentially enriched DHS regions between high- and low-coverage cells. We found that high-coverage cells are characterized by a higher, less variable fraction of early replicating regions (Extended Data Fig. 8a) in contrast to the highly variable values characterizing the low-coverage cells. This pattern suggests that cells with high coverage are indeed in mitosis, as confirmed by the scores calculated on lamin B1-associated domain data[33] (Extended Data Fig. 8b).

To decode the relationship between accessible and compacted chromatin as captured by scGET-seq, we focused our analysis on major repeats, regions of the genome that undergo compaction during the cell cycle through the acquisition of H3K9me3 residues. As Kdm5c acts and heterochromatin assembly occurs during middle/late S phase, we focused on the G1/S phase of the cell cycle[32,36]. The signal emerging from Tn5 was weaker in G1/S cells where Kdm5c expression was not knocked down (Fig. 4a,d, black arrow, compared to TnH in Fig. 4c, red arrow), likely because these cells present a normal assembly of H3K9me3 and heterochromatin, and therefore Tn5 would be unable to tag compacted DNA. Conversely, the signal from TnH showed a more even distribution in G1/S cells, irrespective of Kdm5c status, as TnH targets both accessible and compacted chromatin (Fig. 4c).

We tested whether our observation was statistically significant fitting a linear model that considers the enrichment over TnH and Tn5 as an interaction term when looking for groupwise specific markers. We found that TnH enrichment was significantly higher than Tn5 in groups 3 and 6 (Extended Data Fig. 8c,d), where indeed shScr cells are present at a higher percentage, suggesting that TnH is able to selectively capture regions of the genome, such as chromatin decorated with H3K9me3, which Tn5 is unable to reach.

Together, these data suggest that GET-seq pinpoints quantitative differences between the two enzymes arising from the local chromatin status.

scGET-seq defines cell identity and developmental paths. The modulation of H3K9 methylation and chromatin compaction are pivotal mechanisms underlying organismal development and cellular reprogramming. We thus explored the potential role of scGET-seq in illuminating these processes. To this end, we explored the single-cell profiles of cultured fibroblasts (FIBs) undergoing reprogramming into induced pluripotent stem cells (iPSCs) that were obtained from two unrelated healthy individuals and of iPSCs undergoing differentiation into neural progenitor cells (NPCs). In parallel, we performed single-cell RNA sequencing (scRNA-seq) analysis on cells from the same samples.

Low-dimensional representation of single-cell data from scGET-seq and scRNA-seq separated FIBs, iPSCs and NPCs into

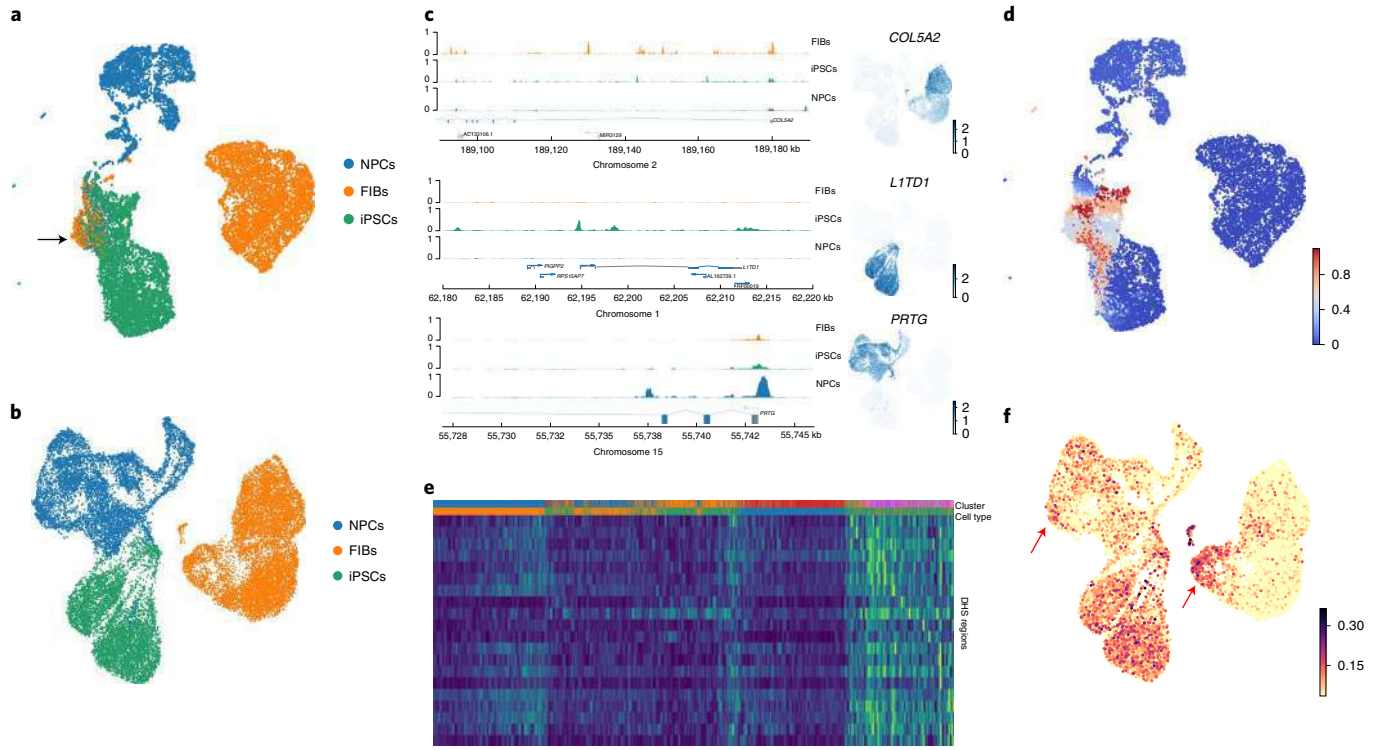


Fig. 5 | scGET-seq defines cell identity and developmental trajectories of FIBs, iPSCs and NPCs. a, UMAP embedding showing scGET-seq profiling of human FIBs, iPSCs and NPCs. The black arrow shows a small subset of FIBs and NPCs clustering alongside iPSCs. **b**, UMAP embedding showing scRNA-seq profiling of the same cell populations derived from the same samples as in **a**. **c**, Profiles show the pseudobulk Tn5 signal for three selected regions among the top differentially enriched in the three cell types; tracks are colored according to cell type as in **a** and **b**. A UMAP embedding colored by the level of expression of the corresponding gene is reported on the right of each profile. **d**, UMAP embedding of cells profiled by scGET-seq and colored by entropy (differentiation potential) as estimated by Palantir. **e**, Heat map showing the enrichment of Tn5 over the top 20 regions associated with a high entropy as result of a generalized linear model. The first annotation row is colored by cell cluster, and the second annotation row is colored by the cell type. **f**, UMAP embedding of cells profiled by scRNA-seq and colored by the expression signature derived from genes associated with regions depicted in **e**. The red arrows show the subsets of NPCs and FIBs that share similar features with iPSCs.

three distinct populations (Fig. 5a,b). Notably, UMAP representations of both scGET-seq and scRNA-seq data showed that iPSCs and NPCs were in close proximity, while FIBs were isolated from the other two populations, with the exception of a small subset of FIBs and to a lesser extent NPCs clustering alongside iPSCs exclusively in the scGET-seq data (Fig. 5a, black arrow).

We next explored the genomic regions more closely defining each population. Notably, the GET-seq sequences most significantly enriched in each cell type were in proximity of genes that are crucial for the biology of each population, such as *COL5A2* for FIBs, *L1TD1* for iPSCs[37] and *PRTG* for NPCs[38] (Fig. 5c and Supplementary Table 4), with concomitant expression in the corresponding populations.

We next sought to determine whether the epigenetic landscapes depicted by scGET-seq could be exploited to capture cell fate probabilities. Indeed, it has been recently proposed that cell fate choices are driven by a continuum of epigenetic choices more than a series of discrete bifurcations alongside developmental paths[39]. To this end, a tool has been recently devised, Palantir[39], that is able to capture these dynamics from scRNA-seq data. When we applied Palantir to the GET-seq dataset, we found three main fate branches (Extended Data Fig. 9a) defining a group of cells endowed with an intense differentiation potential (Fig. 5d), which included iPSCs and the subset of FIBs and NPCs clustering alongside iPSCs (Fig. 5a).

Intrigued by these results, we then explored the regions defining these cellular populations endowed with the highest differentiation potential (Fig. 5e). We found that these regions resided, for the most part, in pericentromeric regions (Supplementary Table 5), in line with recent reports supporting a crucial role for these genomic areas as drivers of pluripotency[40–43]. We hence used the genes associated with these regions to generate a differentiation signature, which we then applied to scRNA-seq data. This signature highlighted a subset of NPCs as well as FIBs sharing similar features in the scRNA-seq data (Fig. 5f, red arrows).

Together, these results suggest that GET-seq is able to capture the epigenetic diversity arising during developmental processes and identify key factors engaged in the process. Additionally, this approach may uncover epigenetic events arising before the appearance of the concomitant transcriptomic events.

Chromatin Velocity to define epigenetic vectors. Prompted by the quantitative properties of scGET-seq highlighted in the shKdm5c experiment, we sought to investigate developmental dynamics in terms of differential unfolding of chromatin. RNA velocity is a tool recently introduced that uses scRNA-seq data to capture not only the overall developmental direction of each cell but also its kinetics, that is, the differential displacement by which various cells travel through states[44]. We hence explored whether it is feasible to obtain single-cell trajectories using scGET-seq data. Instead of using the ratio between unspliced and spliced mRNA, as in RNA velocity, we exploited the ratio between Tn5 and TnH signals at any given location under the assumption that an increase in this value points to a dynamic process leading to more relaxed chromatin, while the opposite is indicative of chromatin compaction (Extended Data Fig. 9b). We found that this approach, which we named Chromatin

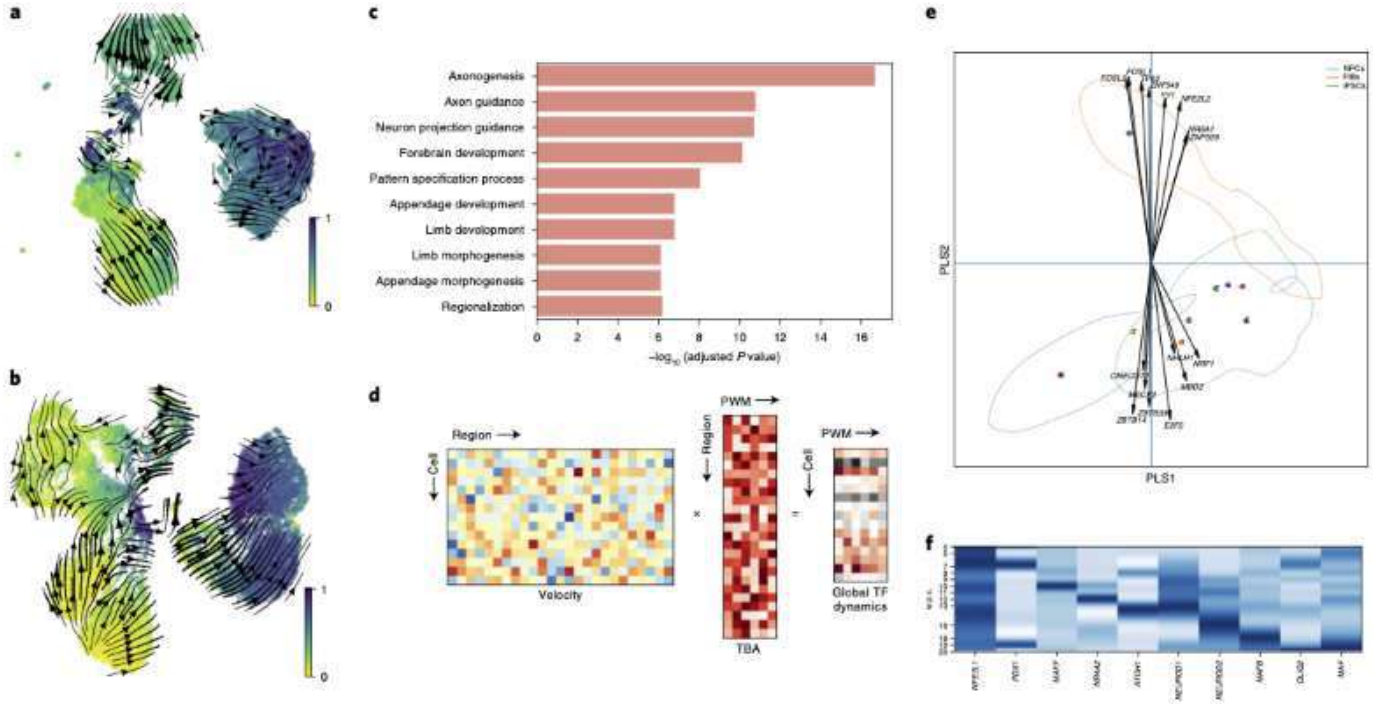


Fig. 6 | Chromatin Velocity. a, UMAP embedding of differentiating single cells profiled by scGET-seq. Cells are colored by velocity pseudotime, and arrow streams indicate the Chromatin Velocity extracted using scvelo. **b**, UMAP embedding of differentiating single cells profiled by scRNA-seq. Cells are colored by velocity pseudotime, and arrow streams indicate the RNA velocity extracted using scvelo. **c**, Selected terms enriched for genes associated with the top dynamic regions. **d**, Schematic representation of the TF analysis. The matrix of velocities calculated over the top dynamic regions is multiplied by the matrix of total binding affinity (TBA) calculated for all position weight matrices (PWMs) in *Homo sapiens* comprehensive model collection (HOCOMOCO) v11 over the same regions. The final matrix contains a single value for each cell for each PWM representing the relevance of a specific TF in the dynamic process happening over that cell. **e**, PLS plot of cell TF analysis matrix. Each dot represents the centroid of all cells belonging to a specific cell group; dots are colored according to cell groups in Extended Data Fig. 8c. Arrows indicate the loading of the top four PWMs in each quadrant. The colored contours indicate the density estimates of the three cell types. **f**, Heat map showing the average expression profiles of TFs with the top ten most negative on PLS1 during early brain development. Darker color indicates higher expression; w.p.c., weeks postconception.

Velocity, is indeed able to capture not only the overall direction but also the velocity of chromatin remodeling (Fig. 6a), with a pattern similar to RNA velocity (Fig. 6b). Of note, the overall pattern of chromatin velocity recapitulates Palantir results in highlighting a group of cells, including iPSCs, NPCs and FIBs, from which most differentiation processes appeared to arise (Figs. 5d and 6a). Also, RNA velocity revealed that the subset of FIBs enriched for the differentiation signature represented the origin from which the FIB population arose (Fig. 6b).

Curious to find the pathways engaged in the differentiation process, we analyzed the results of the dynamical model and identified the 1,703 DHS regions with highest likelihood of being subjected to remodeling. Functional analysis on the genes associated to these regions revealed a strong enrichment for categories related to neural morphogenesis, including axonogenesis and various pathways linked to neural development and morphogenesis, suggesting that our approach is indeed able to grasp biological processes relevant to the model (Fig. 6c and Supplementary Table 6).

As transcription factors (TF) are the key drivers of differentiation, we designed a global TF dynamic score (Fig. 6d and Methods), a cell-by-TF value that is informative of the role of specific TFs in specific cell trajectories. We applied a projection to latent structures regression analysis (PLS)[45] fitting the cell TF scores to cell clusters (Extended Data Fig. 9c and Supplementary Table 7) that clearly separated FIBs on one side and NPCs and iPSCs on the other. Several TFs already implicated in FIB development and maintenance were included, such as *FOSL2* (ref. [46]), *TP63* (ref. [47]) and *NFE2L2* (ref. [48]) (Fig. 6e). Conversely, NPCs and iPSCs were strongly enriched for TFs that are key for neural differentiation, namely *NHLH1* (ref. [49]) and *MECP2*, mutations in which lead to mental retardation[50]. MECP2, MBD2 and ZBTB33 (KAISO) exert redundant activities in neuronal development[51]. Notably, MECP2 enhances the separation of heterochromatin and euchromatin through its condensate partitioning properties[52]. Two TFs were pivotal in these cells, ONECUT1 and LHX3. It has been recently shown that ONECUT1 profoundly remodels chromatin accessibility, thus inducing a neuron-like morphology and the expression of neural genes[53]. ONECUT1 and LHX3, alongside ISLET1, tightly cooperate to dictate the transition from nascent toward maturing embryonic stem cell (ESC)-derived neurons through the engagement of stage-specific enhancers[54].

As PLS1 seems to be associated with the development stage of neural cells, we assessed whether a similar pattern is recapitulated in vivo. To this end, we analyzed expression data of developing human brain obtained from ref. [55], focusing on the early time points (4–20 weeks after conception). With the exception of *DUX4*, which was not profiled in that dataset, we found that TFs with the most negative loading on PLS1 have a single peak of expression in the early stages of brain development (Fig. 6f) and are abruptly downregulated afterwards. Similarly, TFs with the most negative loading on PLS2 include many entries that are also active in the very early stages of brain development (Extended Data Fig. 9d), such as *MBD2*, *ONECUT1* and *LHX3*.

Together, we posit that Chromatin Velocity captures epigenetic transitions underlying crucial biological processes and illuminates the hidden TF networks and wiring driving these dynamic fluxes.

Discussion

In this study, we propose a new single-cell approach, scGET-seq, based on the engineering of a Tn5 transposase targeting H3K9me3, thus providing a comprehensive epigenetic assessment of heterochromatin. Additionally, the sequencing of a much larger portion of the genome allows for the accurate and high-resolution identification of CNVs as well as the detection of SNVs at the single-cell level. We have also harnessed epigenetic data to develop a computational approach, Chromatin Velocity, that defines vectors of cellular fate and predicts future cell states based on the ratio between open and closed chromatin.

Several human diseases are the result of disrupted epigenetic processes, including cancer where the all-important relationship between genetic-driven events versus plasticity remains unclear. Indeed, the study of cancer evolution has relied on the definition of genetic lesions conferring selective advantage, such as the acquisition of somatic mutations or copy number aberrations. Yet, growing evidence points to epigenetic traits as crucially important in several cancer-related phenotypes, for instance the acquisition of drug resistance[3–8]. We envision that the engineering of additional hybrid transposases, including domains targeting other portions of the genome, could extend and integrate the information provided by TnH.

Recent enzyme-tethering strategies have been proposed for chromatin profiling, such as TAM-ChIP and most relevantly CUT&Tag[56]. Indeed, both GET-seq and CUT&Tag are applied on permeabilized live cells, exploit a streamlined Tn5-based library preparation and are suitable for low cell numbers and single cells[57]. However, CUT&Tag is based on antibody-guided tagmentation before chromatin tagmentation, while GET-seq directly targets chromatin through Tn5 tropism modification, therefore offering a more expedited procedure and removing limitations due to specific antibody availability and validation. Finally, to our knowledge, GET-seq is unique in its possibility of multiplexing analysis of different targets in the same reaction through specific barcodes in MEDS oligonucleotides.

RNA velocity adds the vector of time and direction to scRNA-seq one-dimensional data[44]. We propose here Chromatin Velocity, which provides multidimensional information at the epigenetic level. Bulk analysis has revealed that in development, cells undergo epigenetic changes, such as modulation in the opening and closing of chromatin, which precedes and prepares gene expression modifications[58–63]. Therefore, it stands to reason that RNA velocity and Chromatin Velocity are going to capture non-superimposable biological processes.

Retracing the specific engagement of TFs from scRNA-seq experiments is challenging[64]. Leveraging the detailed description of epigenome analysis provides more robust data and reduces variability, allowing for the genome-wide identification of TFs and the epigenetic dynamics of processes such as development.

In summary, we propose a new method, scGET-seq, that captures genomic and chromatin landscapes and trajectories as well as key players, which could provide important insights in fields as diverse as development, regenerative medicine and the study of human diseases, including cancer.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-021-01031-1.

Received: 5 October 2020; Accepted: 22 July 2021;
Published online: 11 October 2021

References

1. McGranahan, N. & Swanton, C. Clonal heterogeneity and tumor evolution: past, present, and the future. *Cell* **168**, 613–628 (2017).
2. Greaves, M. Evolutionary determinants of cancer. *Cancer Discov.* **5**, 806–821 (2015).
3. Liau, B. B. et al. Adaptive chromatin remodeling drives glioblastoma stem cell plasticity and drug tolerance. *Cell Stem Cell* **20**, 233–246 (2017).
4. Hangauer, M. J. et al. Drug-tolerant persister cancer cells are vulnerable to GPX4 inhibition. *Nature* **551**, 247–250 (2017).
5. Brock, A., Chang, H. & Huang, S. Non-genetic heterogeneity—a mutation-independent driving force for the somatic evolution of tumours. *Nat. Rev. Genet.* **10**, 336–342 (2009).
6. Shaffer, S. M. et al. Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. *Nature* **546**, 431–435 (2017).
7. Sharma, S. V. et al. A chromatin-mediated reversible drug-tolerant state in cancer cell subpopulations. *Cell* **141**, 69–80 (2010).
8. Flavahan, W. A., Gaskell, E. & Bernstein, B. E. Epigenetic plasticity and the hallmarks of cancer. *Science* **357**, eaal2380 (2017).
9. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
10. Buenrostro, J. D. et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).
11. Tatarakis, A., Behrouzi, R. & Moazed, D. Evolving models of heterochromatin: from foci to liquid droplets. *Mol. Cell* **67**, 725–727 (2017).
12. Ninova, M., Tóth, K. F. & Aravin, A. A. The control of gene expression and cell identity by H3K9 trimethylation. *Development* **146**, dev181180 (2019).
13. Nicetto, D. et al. H3K9me3-heterochromatin loss at protein-coding genes enables developmental lineage specification. *Science* **363**, 294–297 (2019).
14. Nakayama, J., Rice, J. C., Strahl, B. D., Allis, C. D. & Grewal, S. I. Role of histone H3 lysine 9 methylation in epigenetic control of heterochromatin assembly. *Science* **292**, 110–113 (2001).
15. Peters, A., O'Carroll, D. & Scherthan, H. Loss of the Suv39h histone methyltransferases impairs mammalian heterochromatin and genome stability. *Cell* **107**, 323–337 (2001).
16. Patel, A. P. et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396–1401 (2014).
17. Aldridge, S. & Teichmann, S. A. Single cell transcriptomics comes of age. *Nat. Commun.* **11**, 4307 (2020).
18. Henikoff, S., Henikoff, J., Kaya-Okur, H. & Ahmad, K. Efficient chromatin accessibility mapping in situ by nucleosome-tethered tagmentation. *eLife* **9**, e63274 (2020).
19. Jacobs, S. A. & Khorasanizadeh, S. Structure of HP1 chromodomain bound to a lysine 9-methylated histone H3 tail. *Science* **295**, 2080–2083 (2002).
20. Lachner, M., O'Carroll, D., Rea, S., Mechtler, K. & Jenuwein, T. Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. *Nature* **410**, 116–120 (2001).
21. Bannister, A. J. et al. Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. *Nature* **410**, 120–124 (2001).
22. Satpathy, A. T. et al. Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nat. Biotechnol.* **37**, 925–936 (2019).
23. Cross, W. et al. The evolutionary landscape of colorectal tumorigenesis. *Nat. Ecol. Evol.* **2**, 1661–1672 (2018).
24. Cross, W. et al. Stabilising selection causes grossly altered but stable karyotypes in metastatic colorectal cancer. Preprint at *bioRxiv* https://doi.org/10.1101/2020.03.26.007138 (2020).
25. Gézsi, A. et al. VariantMetaCaller: automated fusion of variant calling pipelines for quantitative, precision-based filtering. *BMC Genomics* **16**, 875 (2015).
26. Misale, S. et al. Vertical suppression of the EGFR pathway prevents onset of resistance in colorectal cancers. *Nat. Commun.* **6**, 8305 (2015).
27. Lupo, B. et al. Colorectal cancer residual disease at maximal response to EGFR blockade displays a druggable Paneth cell-like phenotype. *Sci. Transl. Med.* **12**, eaax8313 (2020).
28. Laurent-Puig, P., Lievre, A. & Blons, H. Mutations and response to epidermal growth factor receptor Inhibitors. *Clin. Cancer Res.* **15**, 1133–1139 (2009).
29. Wang, C. et al. Acquired resistance to EGFR TKIs mediated by TGFβ1/integrin β3 signaling in EGFR-mutant lung cancer. *Mol. Cancer Ther.* **18**, 2357–2367 (2019).
30. Hu, T. & Li, C. Convergence between Wnt-β-catenin and EGFR signaling in cancer. *Mol. Cancer* **9**, 236 (2010).
31. Sondka, Z. et al. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer* **18**, 696–705 (2018).
32. Rondinelli, B. et al. Histone demethylase JARID1C inactivation triggers genomic instability in sporadic renal cancer. *J. Clin. Invest.* **125**, 4625–4637 (2015).

33. Peric-Hupkes, D. et al. Molecular maps of the reorganization of genome–nuclear lamina interactions during differentiation. *Mol. Cell* **38**, 603–613 (2010).
34. Hiratani, I. et al. Global reorganization of replication domains during embryonic stem cell differentiation. *PLoS Biol.* **6**, 2220–2236 (2008).
35. Marchal, C. et al. Genome-wide analysis of replication timing by next-generation sequencing with E/L Repli-seq. *Nat. Protoc.* **13**, 819–839 (2018).
36. Rondinelli, B. et al. H3K4me3 demethylation by the histone demethylase KDM5C/JARID1C promotes DNA replication origin firing. *Nucleic Acids Res.* **43**, 2560–2574 (2015).
37. Wong, R. C. B. et al. L1TD1 is a marker for undifferentiated human embryonic stem cells. *PLoS ONE* **6**, e19355 (2011).
38. Wong, Y. H. et al. Protogenin defines a transition stage during embryonic neurogenesis and prevents precocious neuronal differentiation. *J. Neurosci.* **30**, 4428–4439 (2010).
39. Setty, M. et al. Characterization of cell fate probabilities in single-cell data with Palantir. *Nat. Biotechnol.* **37**, 451–460 (2019).
40. Wang, C. et al. Reprogramming of H3K9me3-dependent heterochromatin during mammalian embryo development. *Nat. Cell Biol.* **20**, 620–631 (2018).
41. Nicetto, D. & Zaret, K. S. Role of H3K9me3 heterochromatin in cell identity establishment and maintenance. *Curr. Opin. Genet. Dev.* **55**, 1–10 (2019).
42. Burton, A. et al. Heterochromatin establishment during early mammalian development is regulated by pericentromeric RNA and characterized by non-repressive H3K9me3. *Nat. Cell Biol.* **22**, 767–778 (2020).
43. Novo, C. L. et al. The pluripotency factor Nanog regulates pericentromeric heterochromatin organization in mouse embryonic stem cells. *Genes Dev.* **30**, 1101–1115 (2016).
44. La Manno, G. et al. RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
45. Wold, S., Sjöström, M. & Eriksson, L. PLS-regression: a basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **58**, 109–130 (2001).
46. Eferl, R. et al. Development of pulmonary fibrosis through a pathway involving the transcription factor Fra-2/AP-1. *Proc. Natl Acad. Sci. USA* **105**, 10525–10530 (2008).
47. Soares, E. & Zhou, H. Master regulatory role of p63 in epidermal development and disease. *Cell. Mol. Life Sci.* **75**, 1179–1190 (2018).
48. Zhu, M. & Zernicka-Goetz, M. Principles of self-organization of the mammalian embryo. *Cell* **183**, 1467–1478 (2020).
49. Begley, C. G. et al. Molecular characterization of NSCL, a gene encoding a helix–loop–helix protein expressed in the developing nervous system. *Proc. Natl Acad. Sci. USA* **89**, 38–42 (1992).
50. Lombardi, L. M. et al. *MECP2* disorders: from the clinic to mice and back. *J. Clin. Invest.* **125**, 2914–2923 (2015).
51. Martin Caballero, I., Hansen, J., Leaford, D., Pollard, S. & Hendrich, B. D. The methyl-CpG binding proteins Mecp2, Mbd2 and Kaiso are dispensable for mouse embryogenesis, but play a redundant function in neural differentiation. *PLoS ONE* **4**, e4315 (2009).
52. Li, C. H. et al. MeCP2 links heterochromatin condensates and neurodevelopmental disease. *Nature* **586**, 440–444 (2020).
53. Van Der Raadt, J., Van Gestel, S. H. C., Kasri, N. N. & Albers, C. A. ONECUT transcription factors induce neuronal characteristics and remodel chromatin accessibility. *Nucleic Acids Res.* **47**, 5587–5602 (2019).
54. Rhee, H. S. et al. Expression of terminal effector genes in mammalian neurons is maintained by a dynamic relay of transient enhancers. *Neuron* **92**, 1252–1265 (2016).
55. Cardoso-Moreira, M. et al. Gene expression across mammalian organ development. *Nature* **571**, 505–509 (2019).
56. Kaya-Okur, H. S. et al. CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat. Commun.* **10**, 1930 (2019).
57. Wu, S. J. et al. Single-cell analysis of chromatin silencing programs in development and tumor progression. Preprint at *bioRxiv* https://doi.org/10.1101/2020.09.04.282418 (2020).
58. Stadhouders, R. et al. Transcription factors orchestrate dynamic interplay between genome topology and gene regulation during cell reprogramming. *Nat. Genet.* **50**, 238–249 (2018).
59. Soufi, A., Donahue, G. & Zaret, K. S. Facilitators and impediments of the pluripotency reprogramming factors' initial engagement with the genome. *Cell* **151**, 994–1004 (2012).
60. Chen, J. Perspectives on somatic reprogramming: spotlighting epigenetic regulation and cellular heterogeneity. *Curr. Opin. Genet. Dev.* **64**, 21–25 (2020).
61. Li, D. et al. Chromatin accessibility dynamics during iPSC reprogramming. *Cell Stem Cell* **21**, 819–833 (2017).
62. Schwarz, B. A. et al. Prospective isolation of poised iPSC intermediates reveals principles of cellular reprogramming. *Cell Stem Cell* **23**, 289–305 (2018).
63. Zviran, A. et al. Deterministic somatic cell reprogramming involves continuous transcriptional changes governed by Myc and epigenetic-driven modules. *Cell Stem Cell* **24**, 328–341 (2019).
64. Lin, C., Ding, J. & Bar-Joseph, Z. Inferring TF activation order in time series scRNA-Seq studies. *PLoS Comput. Biol.* **16**, e1007644 (2020).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2021

Methods

Cell culture. All established cell lines were purchased from American Type Culture Collection (ATCC), except for the HEK293T cell line, which was a kind gift from L. Naldini (San Raffaele Telethon Institute for Gene Therapy). Cells were cultured in DMEM (NIH-3T3, HeLa and HEK293T) or RPMI (Caki-1) supplemented with 10% fetal bovine serum (FA30WS1810500, Carlo Erba for HEK293T cells, and 10270-106, Gibco for all the other cell lines) and 1% penicillin-streptomycin (ECB3001D, Euroclone).

TAM-ChIP. TAM-ChIP (Active Motif) was performed following manufacturer's instructions starting with 10,000,000 Caki-1 cells crosslinked with 38% formaldehyde; fixation was stopped with 0.125 M glycine. Sonication was then performed using a Covaris E220 with the following parameters: total time 6 min, 175 peak incident power, 200 cycles per burst. Sonicated chromatin (8 µg) was used as input for each experimental condition. The following antibodies were used: no antibody (No Ab), anti-H3K9me3 (ab8898, Abcam) and anti-H3K4me3 (07-473, Millipore). ChIP–seq, performed as described in ref. [32], was used as a reference for TAM-ChIP–seq (anti-H3K9me3 (ab8898, Abcam) and anti-H3K4me3 (07-473, Millipore) were used).

TAM-ChIP–RT–qPCR. TAM-ChIP was performed on two biological replicates for each condition (H3K4me3, H3K9me3 and No Ab). For each biological replicate, three technical replicates were analyzed by RT–qPCR. In TAM-ChIP–RT–qPCR one of the two H3K4me3 biological replicates was excluded because no appreciable signal was detected for any condition. For each TAM-ChIP condition, 10 ng of final library was used as input. Water was used as a negative control. RT–qPCR analysis was performed using Sybr Green Master Mix (Applied Biosystems) on the Viia 7 Real Time PCR System (Applied Biosystems). All primers used were designed on H3K9me3-enriched chromatin regions derived from reference ChIP–seq data (as previously described in ref. [32]) and used at a final concentration of 400 nM. To determine the enrichment obtained, we normalized TAM-ChIP–RT–qPCR data to No Ab samples. Primers are listed below.

Primer	Forward sequence	Reverse sequence
BRINP2	GCGCCTTCCTTACTTCCATG	AGTGGCCATCTCATTTCCCA
NTF3	AAAGGCCTTGGTCCCAGA	ATTGAAGGAACGCAGCCC
CACNA1E	GAGGGAGGAGAAAGCCGA	TTGTCCAGACCAGCCCTT

Tn5 transposase production. Tn5 transposase was produced as previously described[65] using pTXB1-Tn5 vector (Addgene, 60240). For hybrid transposases, the DNA fragment encoding human HP-1α was derived from the pET15b-HP1α (pHP1α-pre) vector[66], kindly provided by H. Kurumizaka. According to the cloning strategy, two different lengths of HP-1α polypeptide (spanning amino acids 1–93 and 1–112) were linked to Tn5, using either a three or five poly-tyrosine-glycine-serine (TGS) linker, resulting in four hybrid constructs, TnH 1–TnH 4: TnH 1, amino acids 1–93 (HP-1α)-3×TGS-Tn5; TnH 2, amino acids 1–93 (HP-1α)-5×TGS-Tn5; TnH 3, amino acids 1–112 (HP-1α)-3×TGS-Tn5; TnH 4, amino acids 1–112 (HP-1α)-5×TGS-Tn5. The 1–93 or 1–112 amino acid spanning regions of HP-1α include 1–75 amino acids of CD followed by 18 or 37 amino acids of natural linker, respectively. Construct amino acid sequences are detailed in Supplementary Data 1.

Transposon assembly. Assembly of standard and modified preannealed MEDS oligonucleotides, Tn5MEDS-A, Tn5MEDS-B and TnHMEDS-A was performed in solution following a published protocol[67]. For scGET-seq, standard ME-A oligo[65] was replaced by a combination of eight different sequences containing 8-nt tags before the 19-nt ME sequence to allow differentiation of fragments derived from either Tn5 or TnH tagmentation. Four sequences were used to replace standard Tn5ME-A (Tn5ME-A.1, Tn5ME-A.2, Tn5ME-A.7 and Tn5ME-A.8), and another four sequences were used to replace TnHME-A (TnHME-A.4, TnHME-A.5, TnHME-A.9 and TnHME-A.10). A read 1 primer binding site was reconstituted adding 8 nt (TCCGATCT) upstream of the Tn5/TnH tag. Modified Tn5ME-A sequences are reported in Supplementary Data 1.

Creation of functional transposon was performed following a previously published protocol[65].

Bulk tagmentation reaction and ATAC-seq. Bulk tagmentation was performed on Caki-1 gDNA following a published protocol[65]. Specifically, 500 ng of gDNA was incubated for 7 min at 55 °C with 1 µl of functional transposon in 1× TAPS-PEG8000 buffer in a final 20-µl volume. As a control, a parallel reaction was performed on Caki-1 gDNA but using the Nextera DNA Library Prep kit according to the manufacturer's protocol. Reactions were stopped by adding SDS at a final concentration of 0.05% and incubated for 5 min at room temperature. Then, 5 µl of this mixture was used as input for indexing PCR using standard Nextera N7xx and S5xx oligos and KAPA HiFi enzyme (Roche) using the following protocol: 3 min at 72 °C, 30 s at 98 °C followed by 13 cycles of 45 s at 98 °C, 30 s at 55 °C and 30 s at 72 °C. Libraries were then purified using 1× volume of Ampure XP beads (Beckman Coulter) and checked for fragment distribution on a TapeStation (Agilent).

ATAC-seq was performed following published protocols[9] with minor modifications. Briefly, 100,000 Caki-1 cells were pelleted and washed in 100 µl of cold 1× PBS, centrifuged for 10 min at 500*g* at 4 °C and permeabilized in 100 µl of cold lysis buffer (10 mM TrisHCl, pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% (vol/vol) Igepal CA-630) then centrifuged again for 10 min at 500*g* at 4 °C. Tagmentation was performed on cell pellets, using either Tn5 or TnH, by adding 100 µl of transposition mix (5× TAPS-PEG8000 buffer mixed with 10 µl of 1.39 µM functional transposon in a final volume of 100 µl). As a control, a parallel reaction was performed on 100,000 pelleted Caki-1 cells using the Nextera XT DNA Library Prep kit (Illumina) according to the manufacturer's protocol. Reactions were performed at 37 °C for 30 min and stopped by adding SDS at a final concentration of 0.05%. After 5 min of incubation at room temperature, reactions were purified using a QIAquick Gel Extraction kit (Qiagen) and eluted in 15 µl of Elution Buffer. Five microliters of this reaction was used as input for indexing PCR as described before. Libraries were sequenced on Illumina platforms using a 2×50-bp sequencing protocol.

scATAC-seq and scGET-seq. scATAC-seq was performed 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. Nuclei suspensions were prepared to get 10,000 nuclei as target nuclei recovery.

scGET-seq was performed as previously described, but 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.

When scGET-seq was performed using a 20:80 ratio of HeLa:Caki-1 cells, nuclei suspensions were prepared in duplicate to get 10,000 nuclei as target nuclei recovery for each replicate.

Final libraries were loaded on a Novaseq6000 platform (Illumina) to obtain 50,000 reads per nucleus with a read length of 2×50 bp. For GET-seq, the sequencing target was 100,000 reads per nucleus, and a custom read 1 primer was added to the standard Illumina mixture (5′-TCGTCGGCAGCGTCTCCGATCT-3′). Sequencing statistics for all scGET-seq experiments presented in the manuscript are reported in Supplementary Table 8.

scRNA-seq. 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.

***Kdm5c* knockdown experiment.** Lentiviral vectors were produced by transfecting HEK293T cells (a kind gift from L. Naldini, San Raffaele Telethon Institute for Gene Therapy) with pLK0.1 plasmid containing shRNAs targeting *Kdm5c* (shKd-m5c, CCGGGCAGTGTAACACACGTCCATTCTCGAGAATGGACGTGTGTTA CACTGCTTTT) or scramble (shScr)[32].

A calcium chloride method was used for transfection. Specifically, a mix containing 30 µg of transfer vector, 12.5 µg of Δr 8.74, 9 µg of Env vesicular stomatitis virus (VSV)-G, 6.25 µg of Rev and 15 µg of adenovirus (ADV) plasmid was prepared and filled up to 1,125 µl with 0.1× TE:deionized water (2:1). After 30 min of incubation with rotation, 125 µl of 2.5 M $CaCl_2$ was added to the mix and, after 15 min of incubation, the precipitate was formed by dropwise addition of 1,250 µl of 2× HBS to the mix while vortexing at full speed. Finally, 2.5 ml of precipitate was added drop by drop to 15-cm dishes with HEK293T cells at 50% confluency. After 12–14 h, the medium was replaced with 16 ml of fresh medium per dish supplemented with 16 µl of NAB per dish. After 30 h, the medium containing viral particles was collected, filtered with a 0.22-µm filter and stored at −80 °C in small aliquots to avoid freeze–thaw cycles.

NIH-3T3 cells were transduced using a six-well plate format. To this end, 2 ml of shKdm5c/shScr lentiviral vector supplemented with polybrene (final concentration, 8 µg ml^{-1}) was added to actively cycling (50% confluency) NIH-3T3 cells; one well of untransduced cells was used as a negative control. After 24 h, transduced cells were passaged in a 10-cm dish, and puromycin selection (final concentration, 4 µg ml^{-1}) was performed. Forty-eight hours after selection, half of the transduced cells were detached, washed twice with cold 1× PBS and tested for gene knockdown by RT–qPCR as described below. Following knockdown validation, 72 h after selection, all remaining cells were collected and subjected to scGET-seq as already described. Nuclei suspensions were prepared to get 10,000 nuclei as target nuclei recovery.

Gene knockdown validation by RT–qPCR. Total RNA was isolated using Trizol (Invitrogen) and purified using an RNeasy mini kit (Qiagen). cDNA was generated using a First-Strand cDNA Synthesis ImpromII A3800 kit (Promega) with random primers. RT–qPCR was performed using Sybr Green Master Mix (Applied Biosystems) on the Viia 7 Real Time PCR System (Applied Biosystems). Ten nanograms of cDNA was used as input, and water was used as a negative control.

Amplification was performed using previously validated primers[32] used at a final concentration of 400 nM except for primers for major ncRNA that were used at 200 nM. Primers for minor ncRNA were taken from ref. [68] and were used at a final concentration of 400 nM.

Human-derived colorectal cancer organoids (PDOs). Samples from two 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[69]. Briefly, fresh tissues were minced immediately after surgery, conditioned in PBS/5 mM EDTA and digested for 1 h at 37 °C in a solution composed of 2× TrypLE Select Enzyme (Thermo Fisher) in PBS/1 mM EDTA with DNAse I (Merck). Release of cells was facilitated by pipetting. Dissociated cells were collected, suspended in 120 µl of growth factor-reduced Matrigel (Corning 356231, Fisher Scientific), seeded in single domes in a 24-well flat-bottom cell culture plate (Corning) and, after dome solidification, covered with 1 ml of complete human organoid medium[69]; medium was replaced every 2–3 d. For scGET-seq analysis, after a 20-min incubation at 37 °C in a solution of 1× TrypLE Select Enzyme in PBS/1 mM EDTA, PDOs were dissociated to single cells by combining mechanical (pipetting) and enzymatic digestion, washing in 1× PBS and processing as previously described.

Human-derived colorectal cancer xenografts (PDXs). *Specimen collection and annotation.* EGFR blockade-responsive colorectal cancer and matched normal samples were obtained from one individual that underwent liver metastasectomy at the Azienda Ospedaliera Mauriziano Umberto I (Torino). The individual provided informed consent. Samples were procured, and the study was conducted under the approval of the Review Boards of the Institution.

PDX models and in vivo treatment. Tumor implantation and expansion were performed in 6-week-old male and female non-obese diabetic/severe combined immunodeficient (NOD/SCID) mice as previously described[69]. Once tumors reached an average volume of ~400 mm^3, mice were randomized into the following four treatment arms that received either placebo or cetuximab (Merck; 20 mg kg^{-1} twice weekly, intraperitoneally): (1) untreated, (2) cetuximab for 72 h, (3) cetuximab for 4 weeks and (4) cetuximab for 7 weeks. To recover enough cells from tumors that had shrunk during cetuximab treatment, multiple xenografts were minced and mixed together to obtain the individual data points of treated arms ($n=1$ in the case of untreated tumors; $n=2$ for 72 h; $n=4$ for 4 weeks; $n=5$ for 7 weeks). The whole experiment was performed twice to obtain independent biological duplicates for each experimental point. To reach the endpoint of all the experimental groups on the same day, treatments were started asynchronously. Tumor growth was monitored once weekly by caliper measurements, and approximate tumor volumes were calculated using the formula $4/3\pi\times(d/2)^2\times D/2$, where d and D are the minor tumor axis and the major tumor axis, respectively. Operators were blinded during measurements. In vivo procedures and related biobanking data were managed using the Laboratory Assistant Suite (https://doi.org/10.1007/s10916-012-9891-6). Animal procedures were approved by the Italian Ministry of Health (authorization 806/2016-PR).

scGET-seq on PDXA. At the end of treatments, mice were killed, and tumors were collected. All the tumors pertaining to each treatment arm were pooled together. The dissociation step was performed using the Human Tumor Dissociation kit (Miltenyi Biotec) with the gentleMACS Dissociator (Miltenyi Biotec) according to the manufacturer's protocol. Single cells were then subjected to scGET-seq as already described. Nuclei suspensions were prepared to get 10,000 nuclei as target nuclei recovery for each replicate.

FIB reprogramming toward iPSCs and iPSC differentiation toward NPCs. Dermal FIBs obtained from skin biopsies of two different healthy individuals (A and B) were cultured in fibroblast medium and reprogrammed with Sendai virus technology (CytoTune-iPS Sendai Reprogramming kit, Thermo Fisher) to generate human iPSC clones. iPSC clones were individually picked, expanded and maintained in mTeSR1 on human ESC (hESC)-qualified Matrigel. Human iPSC-derived NPCs were generated following the standard protocol based on dual SMAD inhibition[70]. Briefly, iPSCs were differentiated to NPCs via human embryoid bodies. Neural induction was initiated through inhibition using the dual small inhibition molecules dorsomorphin, purmorphamine and SB43152. The small molecule CHIR99021, a GSK-3β inhibitor, was added to stimulate the canonical WNT signaling pathway. The study was approved by Comitato Etico Ospedale San Raffaele (BANCA-INSPE 09/03/2017). Human FIBs, iPSCs and NPCs derived from individuals A and B were collected, counted and subjected to GET-seq and scRNA-seq, as already described, starting from the same cell suspension. Target recovery was 5,000 cells for scRNA-seq and 5,000 nuclei for scGET-seq.

Bioinformatics analysis. *Data preprocessing.* Illumina sequencing data for bulk sequencing were demultiplexed using bcl2fastq using default parameters. Sequencing data for single-cell experiments were demultiplexed using cellranger-atac (v1.0.1). Identification of cell barcodes was performed using umitools (v1.0.1)[71] using R2 as input.

Read tags for GET-seq and scGET-seq experiments, where TnH and Tn5 data are mixed, were processed with TagDust (v2.33)[72], specifying transposase-specific barcodes as first block in the hidden Markov model (HMM) model. The data preprocessing pipeline is available at https://github.com/leomorelli/scGET.

Reads for ChIP–seq, GET-seq and scGET-seq experiments were aligned to the reference genome (hg38 or mm10) using BWA-MEM v0.7.12 (ref. [73]).

Analysis of bulk sequencing data. Aligned reads were deduplicated using SAMBLASTER[74]. Genome bigwig tracks were generated using bamCoverage from the deepTools suite[75] with bins per million mapped reads (BPM) normalization. H3K4me3-enriched regions were identified using MACS v2.2.7 (ref. [76]), and H3K9me3-enriched regions were identified using SICER v2 (ref. [77]) using default parameters.

Definition of epigenome reference sets. We segmented the genome according to DHSs, as previously described[78]. Briefly, we downloaded the index of DHSs for human[79] and mouse genomes[77]; intervals closer than 500 bp were merged using bedtools[80] to create the interval set for accessible chromatin (named 'DHS'). We then took the complement of the set to create the interval set for compacted chromatin (named 'complement').

Analysis of scGET-seq data. Lists of accepted cellular barcodes were assigned to reads inside aligned BAM files using bc2rg.py script from scatACC (https://github.com/dawe/scatACC). Duplicated reads were then identified at the cell level using cbdedup.py script from the same repository. For each scGET-seq experiment, we generated four count matrices, Tn5-dhs, Tn5-complement, Tnh-dhs and TnH-complement, profiling Tn5 and TnH over accessible and compacted chromatin, respectively. Count matrices were generated using peak_count.py script from the scatACC repository. Each count matrix was processed using scanpy v1.4.6 or v1.6.0 (ref. [81]). After an initial filtering on shared regions and number of detected regions per cell, matrices were normalized and log transformed. The number of regions was used as a covariate for linear regression, and data were then scaled with a maximum value set to 10. Neighborhood was evaluated using batch-balanced KNN[82], and cell groups were identified with the Leiden algorithm[83] for cell lines or schist[84], choosing the hierarchy level that maximizes modularity. To extract a unique representation of four datasets, we applied graph fusion using scikit-fusion[85]. We first extracted a 20-component UMAP reduction of each view and built a relation graph where all views are connected to a 20-component latent space. Matrix factorization was run with 1,000 iterations five times. The resulting latent space was then added in each scanpy object as the basis for neighborhood evaluation and cell clustering.

Library saturation estimates. To estimate the library complexity, we first downsampled ten datasets (four depicted in Figs. 2a and 6, randomly chosen) at different proportions (0.1×, 0.2×, 0.5×) and calculated the number of genomic bins (5 kb) that could be found in each dataset. For each dataset, we fitted the shape parameter s of a lower incomplete gamma function. We then built a linear model fitting the number of cells and the number of duplicates to predict s (Extended Data Fig. 4c). We obtained the model $s=0.815\times N_{cells}+0.406\times(1-d)+0.2316$, where N_{cells} is the number of cells divided by 1,000, and d is the fraction of duplicated reads.

Analysis of HeLa and Caki-1 cell identity. To identify cell identity in Caki-1/HeLa mixtures, we downloaded publicly available bulk ATAC-seq data for HeLa cells (GSE106145)[86] and preprocessed as described above. We then generated a count matrix for HeLa cells and our bulk ATAC-seq for Caki-1 cells over the DHS regions using bedtools. The resulting matrix was analyzed in edgeR[87] using relative log expression (RLE) normalization and contrasting HeLa versus Caki-1 cells by a Fisher's exact test. We selected HeLa-specific regions by filtering for a false discovery rate (FDR) value of $<1\times10^{-3}$, log counts per million reads mapped (CPM) of >3 and log fold change of >0 (that is, regions enriched in HeLa cells with detectable read counts), and we took the top 200 regions that were present in scGET-seq data. We used this list to create a HeLa score using the score_genes function implemented in scanpy.

Cell cycle analysis. Identification of cell cycle phase using replication data was performed as follows. First, we identified high-coverage and low-coverage cells in each experiment by analyzing TnH-complement data. We then identified the top 500 Tn5–DHS regions characterizing each cluster.

Two-stage Repli-seq data for NIH-3T3 cells were downloaded from the 4DNucleome project (https://data.4dnucleome.org/experiment-set-replicates/4DNES7ZVDD5G/), replicated data were averaged and the $\log_2$ ratio between early stage (E) and late stage (L) was calculated. Entries in the Tn5–DHS list were assigned the average $\log_2$ (E/L) value over its interval.

Lamin B1 DamID data for NIH-3T3 cells were also downloaded from University of California Santa Cruz genome browser tables, converted to bigwig format and lifted over mm10 assembly coordinates using Crossmap[88]. The average value of lamin B1 data over Tn5–DHS regions was assigned as described above.

Differences in distribution of $\log_2$(E/L) and lamin B1 values were evaluated by Mann-Whitney *U*-test.

Analysis of copy number alterations. Copy number alterations were derived from TnH data quantified over the entire genome, binned at a 5-kb resolution. Counts were extracted using peak_count.py script from the scatACC repository. Data were then processed by collapsing values into larger bins at different resolutions (10 Mb, 1 Mb and 500 kb). The value of each bin is divided by the average per cell read count. We applied linear regression of per bin GC content and mappability[89] and finally expressed values as $\log_2$ of the scaled residuals. Cell clustering was performed using schist applied on the KNN graph built with bbknn and using correlation as a distance metric. The number of clusters is defined by the highest level of the hierarchy that splits more than one group. Evaluation of the posterior distribution of number of groups is performed by equilibration of a Markov Chain Monte Carlo model with at most 1,000,000 iterations.

Classification of CNVs in Caki-1 and HeLa cells. We created a ground truth dataset by calling copy number alterations in Caki-1 and HeLa cells with Control-FREEC[89] on WGS data. We binned the resulting segments according to the desired resolution in single-cell experiments (10 Mb, 1 Mb and 500 kb), retaining three classes (loss, gain and normal).

We subsampled scATAC-seq cells and scGET-seq cells to match cell numbers and coverage distributions to avoid biases due to different data sizes. We split $\log_2$ ratio matrices into a training and a test set in a 70:30 proportion. We trained a logistic regression classifier and an SVM with the one-versus-rest strategy and increased the number of iterations to ensure convergence. We recorded accuracy and F1 score on the test sets. This process was applied on each resolution, cell type and platform.

Bulk analysis of organoid whole-exome sequencing data. Reads were aligned to the hg38 reference genome using BWA, and reads were then processed using BWA. Alignments were processed using GATK MarkDuplicates and base quality score recalibration[89]. Somatic mutations and copy number segments were identified with Sequenza[90] with default parameters. Evaluation of CNVs was performed using CNAqc[91], and clonal deconvolution was performed using MOBSTER and BMix[92] with default parameters.

Analysis of mutations. Reads for Tn5 and TnH data were separated into individual BAM files using separate_bam.py script from the scatACC repository. Known somatic mutations were genotyped using freebayes v.1.3.2 (ref. [93]) (parameters: -@ exome_somatic.vcf.gz -C 2 -F 0.01). Only variants with a depth of >1 were then considered for the analysis.

Variant calling without priors was performed using freebayes using the same thresholds. Variant call format (VCF) files were annotated using snpEff v4.3p[94] using the GRCh38.86 annotation model. Known cancer variants were annotated using COSMIC catalog[95]. Variants were then filtered for depth >10 and quality >5 if unknown and quality >1 if profiled in COSMIC.

Chromatin Velocity. Chromatin Velocity was calculated using scvelo[96]. Normalized count matrices over DHS regions for Tn5 and TnH were first filtered to include regions common to both. Then a proper object was created injecting Tn5 and TnH data in the unspliced and spliced layers, respectively. Moments were calculated on the KNN graph previously estimated. Dynamical modeling was then applied, and final velocity was calculated with regularization by latent time. Regions having a likelihood value higher than the 95th percentile were considered as marker regions.

Analysis of scRNA-seq data. Reads were demultiplexed using Cell Ranger (v4.0.0). Identification of valid cellular barcodes and unique molecular identifiers (UMIs) was performed using umitools with default parameters for 10x v3 chemistry. Reads were aligned to the hg38 reference genome using STARsolo (v2.7.7a)[97]. Quantification of spliced and unspliced reads on genes was performed by STARsolo itself on GENCODE v36 (ref. [98]). Count matrices were imported into scanpy, and doublet rate was estimated using scrublet[99]. The count matrix was filtered (min_genes = 200, min_cells = 5, pct_mito < 20) before normalization and log transformation. A KNN graph was built using bbknn. RNA velocity was estimated using scvelo dynamical modeling with latent time regularization.

TBA analysis. For each DHS region selected for likelihood, we extracted the 500-bp sequence flanking summits there included, as annotated in the DHS index. We downloaded the HOCOMOCO v11 list of PWMs[100] and calculated the TBA as defined in ref. [101] using tba_nu.py script from the scatACC repository. TBA values for multiple summits within a DHS region were summed. Final values were divided by the length of the corresponding DHS region. To obtain a cell-specific TBA value, the region-by-TBA matrix was multiplied by the cell-by-region velocity matrix.

PLS analysis was performed using the PLSCanonical function from the Python sklearn.cross_decomposition library using cell groups as targets for the matrix transformation.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Fastq files and raw count matrices have been deposited to the Array Express platform (https://www.ebi.ac.uk/arrayexpress/) with the following IDs: E-MTAB-9648, E-MTAB-10218, E-MTAB-2020, E-MTAB-10219, E-MTAB-9650, E-MTAB-9651 and E-MTAB-9659. Source data are provided with this paper.

Code availability

Code necessary to preprocess scGET-seq data is available at https://github.com/leomorelli/scGET (ref. [102]) and https://github.com/dawe/scatACC (ref. [103]). Illustrative code snippets for postprocessing are reported in Supplementary Data 2.

References

65. Picelli, S. et al. Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res.* **24**, 2033–2040 (2014).
66. Machida, S. et al. Structural basis of heterochromatin formation by human HP1. *Mol. Cell* **69**, 385–397 (2018).
67. Reznikoff, W. S. Transposon Tn5. *Annu. Rev. Genet.* **42**, 269–286 (2008).
68. Zhu, Q. et al. BRCA1 tumour suppression occurs via heterochromatin-mediated silencing. *Nature* **477**, 179–184 (2011).
69. Bertotti, A. 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.* **1**, 508–523 (2011).
70. Reinhardt, P. et al. Derivation and expansion using only small molecules of human neural progenitors for neurodegenerative disease modeling. *PLoS ONE* **8**, e59252 (2013).
71. Smith, T., Heger, A. & Sudbery, I. UMI-tools: modeling sequencing errors in unique molecular identifiers to improve quantification accuracy. *Genome Res.* **27**, 491–499 (2017).
72. Lassmann, T. TagDust2: a generic method to extract reads from sequencing data. *BMC Bioinformatics* **16**, 24 (2015).
73. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at *arXiv* https://arxiv.org/abs/1303.3997 (2013).
74. Faust, G. G. & Hall, I. M. SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics* **30**, 2503–2505 (2014).
75. Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A. & Manke, T. DeepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* **42**, 187–191 (2014).
76. Zhang, Y. et al. Model-based analysis of ChIP–seq (MACS). *Genome Biol.* **9**, R137 (2008).
77. Breeze, C. E. et al. Atlas and developmental dynamics of mouse DNase I hypersensitive sites. Preprint at *bioRxiv* https://doi.org/10.1101/2020.06.26.172718 (2020).
78. Giansanti, V., Tang, M. & Cittaro, D. Fast analysis of scATAC-seq data using a predefined set of genomic regions. *F1000Res.* **9**, 199 (2020).
79. Meuleman, W. et al. Index and biological spectrum of human DNase I hypersensitive sites. *Nature* **584**, 244–251 (2020).
80. Quinlan, A. R. BEDTools: the Swiss-Army tool for genome feature analysis. *Curr. Protoc. Bioinformatics* https://doi.org/10.1002/0471250953.bi1112s47 (2014).
81. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
82. Polański, K. et al. BBKNN: fast batch alignment of single cell transcriptomes. *Bioinformatics* **36**, 964–965 (2020).
83. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019).
84. Morelli, L., Giansanti, V. & Cittaro, D. Nested stochastic block models applied to the analysis of single cell data. Preprint at *bioRxiv* https://doi.org/10.1101/2020.06.28.176180 (2020).
85. Žitnik, M. & Zupan, B. Data fusion by matrix factorization. *IEEE Trans. Pattern Anal. Mach. Intell.* **37**, 41–53 (2015).
86. Cho, S. W. et al. Promoter of lncRNA gene *PVT1* is a tumor-suppressor DNA boundary element. *Cell* **173**, 1398–1412 (2018).
87. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2009).
88. Zhao, H. et al. CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* **30**, 1006–1007 (2014).
89. Karimzadeh, M., Ernst, C., Kundaje, A. & Hoffman, M. M. Umap and Bismap: quantifying genome and methylome mappability. *Nucleic Acids Res.* **46**, e120 (2018).
90. Favero, F. et al. Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Ann. Oncol.* **26**, 64–70 (2015).
91. Household, J., Cross, W. C. H. & Caravagna, G. A fully automated approach for quality control of cancer mutations in the era of high-resolution whole genome sequencing. Preprint at *bioRxiv* https://doi.org/10.1101/2021.02.13.429885 (2021).

92. Caravagna, G., Sanguinetti, G., Graham, T. A. & Sottoriva, A. The MOBSTER R package for tumour subclonal deconvolution from bulk DNA whole-genome sequencing data. *BMC Bioinformatics* **21**, 531 (2020).
93. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. Preprint at *arXiv* https://arxiv.org/abs/1207.3907 (2012).
94. Cingolani, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* **6**, 80–92 (2012).
95. Forbes, S. A. et al. COSMIC: mining complete cancer genomes in the catalogue of somatic mutations in cancer. *Nucleic Acids Res.* **39**, 945–950 (2011).
96. Bergen, V., Lange, M., Peidli, S., Wolf, F. A. & Theis, F. J. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat. Biotechnol.* **38**, 1408–1414 (2020).
97. Kaminow, B., Yunusov, D. & Dobin, A. STARsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq data. Preprint at *bioRxiv* https://doi.org/10.1101/2021.05.05.442755 (2021).
98. Harrow, J. et al. GENCODE: the reference human genome annotation for the ENCODE project. *Genome Res.* **22**, 1760–1774 (2012).
99. Wolock, S. L., Lopez, R. & Klein, A. M. Scrublet: computational identification of cell doublets in single-cell transcriptomic data. *Cell Syst.* **8**, 281–291 (2019).
100. Kulakovskiy, I. V. 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.* **46**, D252–D259 (2018).
101. 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.* **28**, 2173–2183 (2011).
102. Morelli, L. & Cittaro, D. scGET: pre-release of scGET repository. *Zenodo* https://doi.org/10.5281/zenodo.5095040 (2021).
103. Cittaro, D. scatACC: version 0.1. *Zenodo* https://doi.org/10.5281/zenodo.5095157 (2021).

Acknowledgements

We thank all the members of the COSR and Tonon laboratory for discussions, support and critical reading of the manuscript. We are grateful to E. Brambilla and F. Ruffini for preparation of the iPSCs and NPCs and A. Mira for assistance in the preparation of the organoids. We would like to thank S. de Pretis for the thoughtful discussions about chromatin velocity. We are grateful to G. Bucci for providing raw exome sequencing data and P. Dellabona for the coordination of the metastatic colon cancer sample collection and analysis. We also thank D. Gabellini, M. E. Bianchi, A. Agresti and S. Biffo for helpful discussions and for reviewing the manuscript. A.B. and L.T. are members of the EurOPDX Consortium. This work was partially supported by the Italian Ministry of Health with Ricerca Corrente and 5×1000 funds (S.M. and S.P.), by Associazione Italiana per la Ricerca sul Cancro (AIRC) investigator grants 20697 (to A.B.) and 22802 (to L.T.), AIRC 5×1000 grant 21091 (to A.B. and L.T.), AIRC/CRUK/FC AECC Accelerator Award 22795 (to L.T.), European Research Council Consolidator Grant 724748 BEAT (to A.B.), H2020 grant agreement 754923 COLOSSUS (to L.T.), H2020 INFRAIA grant agreement 731105 EDIReX (to A.B.), Fondazione Piemontese per la Ricerca sul Cancro-ONLUS, 5×1000 Ministero della Salute 2014, 2015 and 2016 (to L.T.), AIRC investigator grants (to G.T.) and by the Italian Ministry of Health with 5×1000 funds, Fiscal Year 2014 (to G.T.), AIRC 5×1000 ID 22737 (to G.T.) and the AIRC/CRUK/FC AECC Accelerator Award 'Single Cell Cancer Evolution in the Clinic' A26815 (AIRC number program 2279) (to G.T.).

Author contributions

M.T. performed experiments and analyzed the data. F.G. devised the methodology and experimental design, performed experiments and analyzed data. D.L. devised the methodology. V.G. performed bioinformatic analysis. D.R. performed experiments and provided experimental assistance and expertise. L.M. performed bioinformatic analysis. S.M. performed cloning and transposase production. I.C. and E.R.Z. performed in vivo experiments. O.A.B. performed experiments related to culturing and maintenance of organoids. E.G. performed bioinformatic analysis. G.C. performed analysis on whole-exome data. P.P.B. designed and supervised the FIB reprogramming and iPSC differentiation experiments. A.B. designed and supervised in vivo experiments and reviewed the data. G.M. designed and supervised the FIB reprogramming and iPSC differentiation experiments and reviewed the data. L.A. provided the primary samples used for the organoid experiments. S.P. designed and supervised transposase production and reviewed the data. L.T. designed and supervised in vivo experiments and reviewed data. D.C. designed the study, performed bioinformatic analysis and wrote the manuscript. G.T. designed the study, analyzed data and wrote the manuscript.

Competing interests

M.T., F.G., D.L., S.P., D.C. and G.T. have submitted a patent application, pending, covering TnH.

Additional information

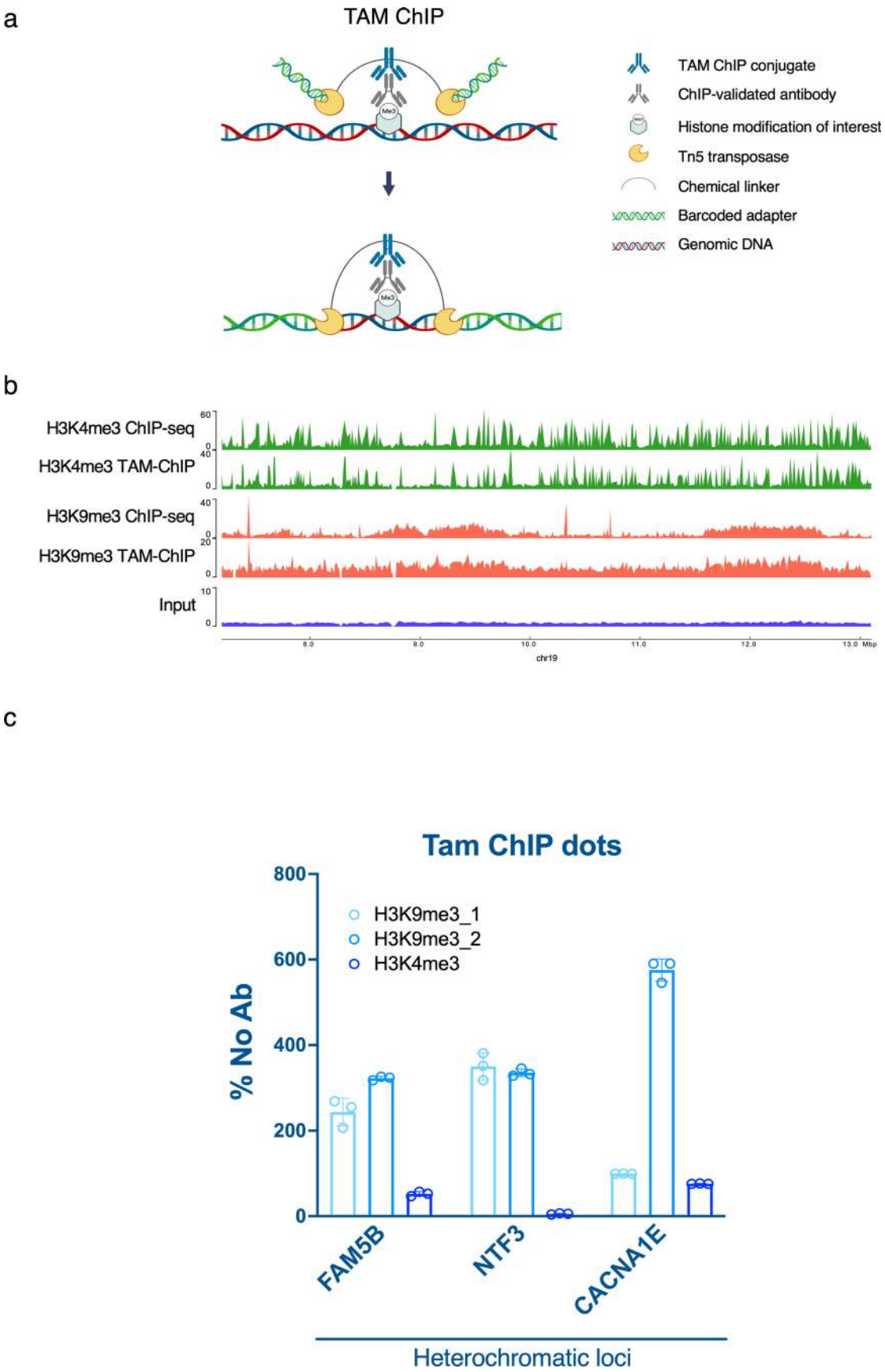
Extended data is available for this paper at https://doi.org/10.1038/s41587-021-01031-1.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-021-01031-1.

Correspondence and requests for materials should be addressed to Davide Cittaro or Giovanni Tonon.

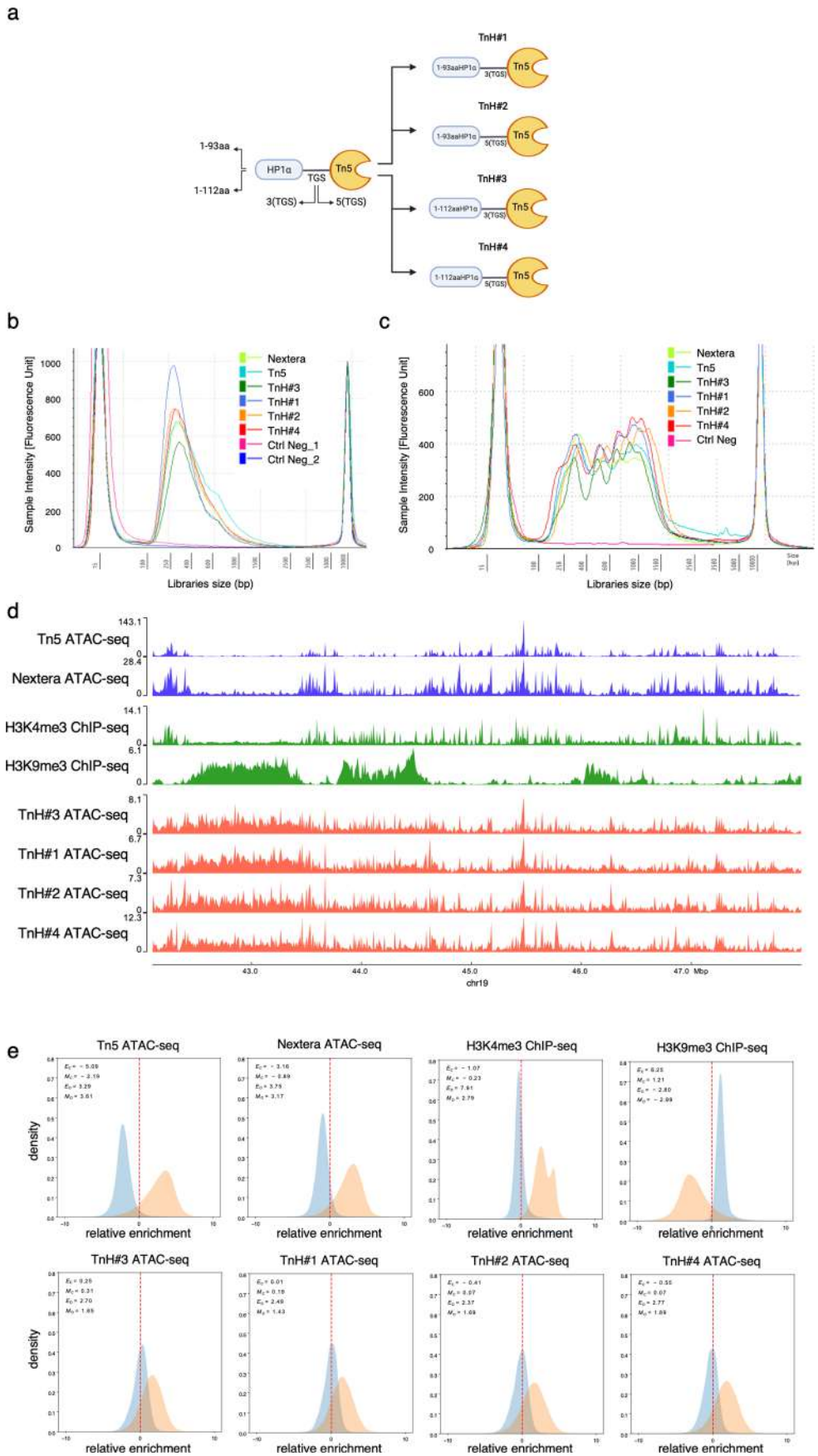
Peer review information *Nature Biotechnology* thanks Kun Zhang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work

Reprints and permissions information is available at www.nature.com/reprints.



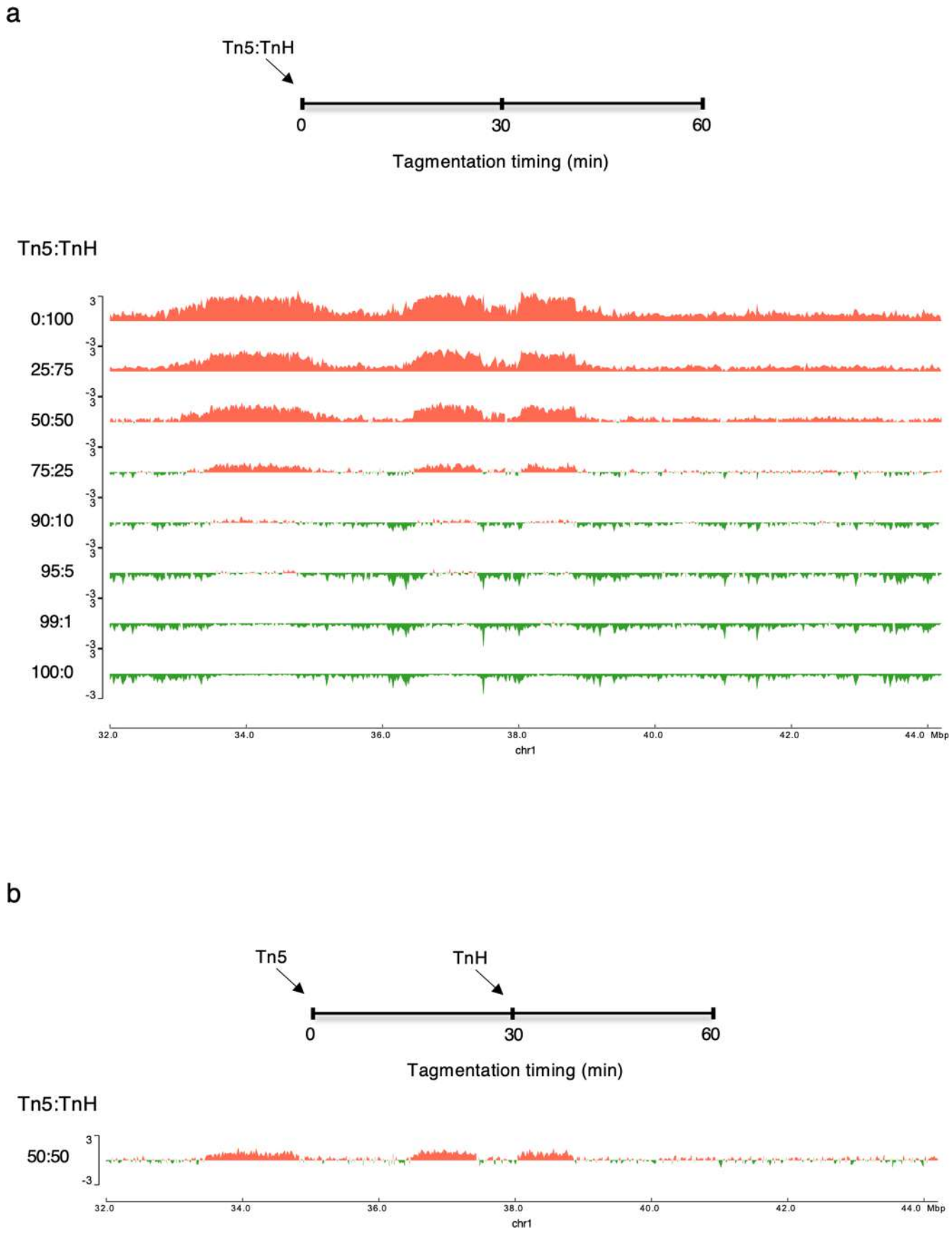
Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Tn5 transposase is able to tagment compacted chromatin featuring H3K9me3. a, General scheme of TAM-ChIP technique (created with BioRender.com). A primary antibody (ChIP-validated antibody, dark grey) binds to a specific histone modification (light grey) over the genome (blue-red). A secondary antibody (TAM-ChIP conjugate, blue) is linked to the Tn5 transposon, which is made of Tn5 transposase (yellow) and the respective barcoded adapters (green). Upon the binding of the secondary antibody to the primary antibody, the linked Tn5 transposase targets and cuts the genomic regions flanking the histone modification, adding the barcoded adapters. TAM-ChIP was performed on two biological replicates for each condition (H3K4me3, H3K9me3 and NoAb). **b**, H3K4me3 (green) and H3K9me3 (red) enrichment profiles obtained either by ChIP-seq or TAM-ChIP-seq, compared with Input ChIP control (violet). **c**, Enrichment profile of heterochromatic genes FAM5B, NTF3, CACNA1E obtained from TAM-ChIP libraries assessed by Real Time-qPCR confirms Tn5 is able to target heterochromatic loci when redirected by H3K9me3 antibody. For each biological replicate three technical replicates were analyzed by Real-Time qPCR; one of the two H3K4me3 biological replicate was excluded because no appreciable signal was detected for any condition. Whiskers represent standard deviations ($n = 3$ technical replicates). Data shown in b and c refer to experiments performed on Caki-1 cell line.



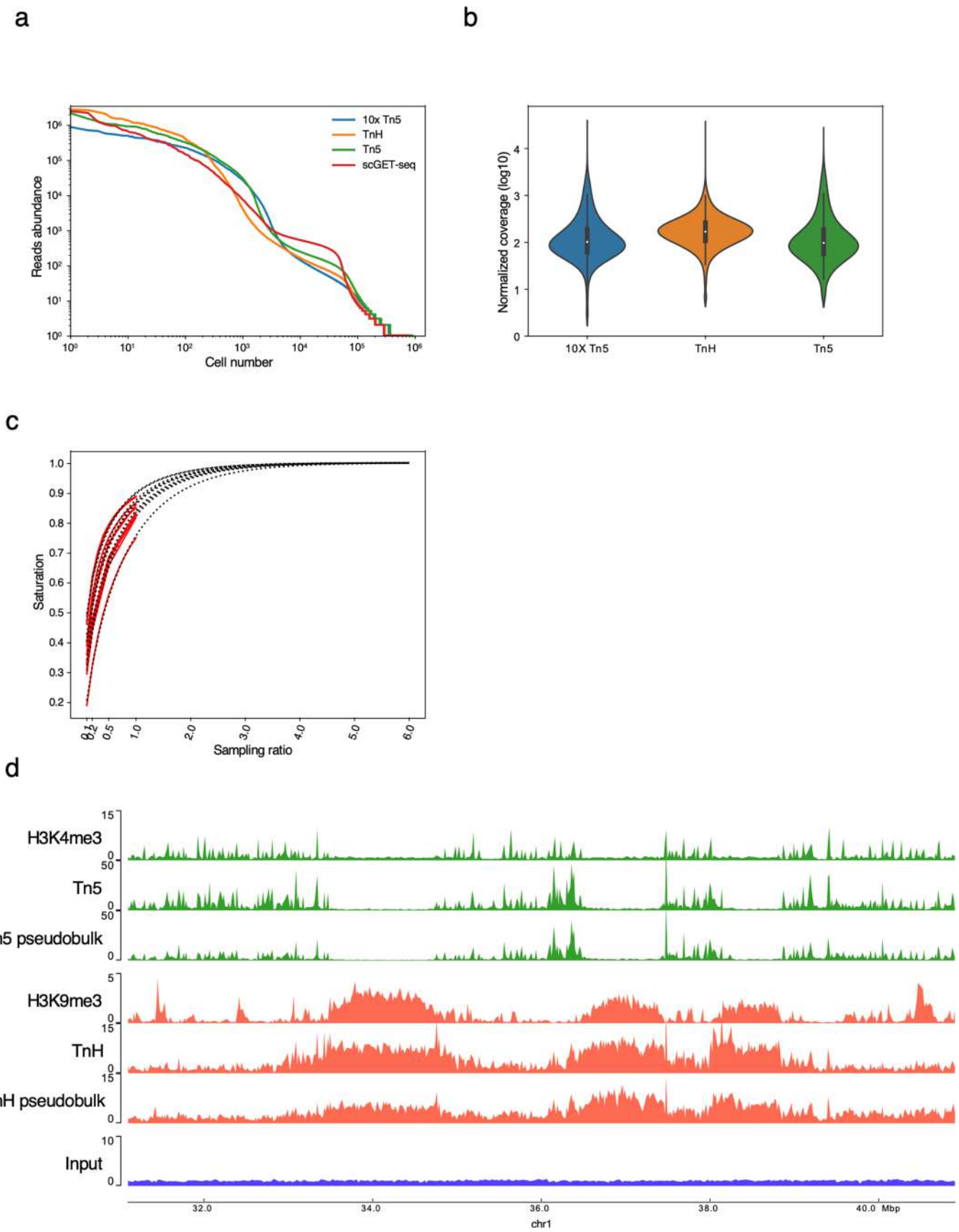
Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Hybrid CD (HP1α)-Tn5 targets H3K9me3 chromatin regions. a, Two different lengths of HP1α polypeptide (spanning amino acids 1-93 and 1-112) were linked to Tn5, using either a 3 or 5 poly-tyrosine–glycine–serine (TGS) linker, resulting in four hybrid construct, TnH#1-4. TnH#1 made of 1-93aa (HP1α) - 3x(TGS) - Tn5; TnH#2 made of 1-93aa (HP1α) - 5x(TGS) - Tn5; TnH#3 made of 1-112aa (HP1α) - 3x(TGS) - Tn5; TnH#4 made of 1-112aa (HP1α) - 5x(TGS) - Tn5. The 1-93 or 1-112aa spanning regions of HP1α include 1-75aa of CD followed by 18 or 37aa of natural linker, respectively (Created with BioRender.com). **b-c**, Tagmentation profiles relative to the four hybrid constructs (TnH#1-4) showing no difference in tagmentation efficiency relative to the native Tn5 enzyme (Nextera and Tn5 in-house produced) when targeting either genomic DNA, panel b, or native chromatin on permeabilized nuclei, panel c. **d**, Enrichment profiles relative to ATAC-seq performed with the four hybrid constructs (TnH#1-4, red) compared with native Tn5 enzyme (Nextera and Tn5 in-house produced) and with H3K4me3 and H3K9me3 ChIP-seq signals (green). **e**, Distribution of the enrichment of four TnH hybrid constructs (TnH#1-4) relative to genomic background in regions enriched for H3K4me3 (orange) or H3K9me3 (blue) expressed as log2(ratio) of the signal over the genomic Input. Enrichment over the same regions for native Tn5 enzyme (Nextera and Tn5 in-house produced), H3K4me3 and H3K9me3 ChIP-seq are reported as reference. E_c: global enrichment over H3K9me3-marked regions; E_o: global enrichment over H3K4me3-marked regions; M_c: modal enrichment over H3K9me3-marked regions; M_o: modal enrichment over H3K4me3-marked regions. Data shown in b, c and d refer to experiments performed on Caki-1 cell line.



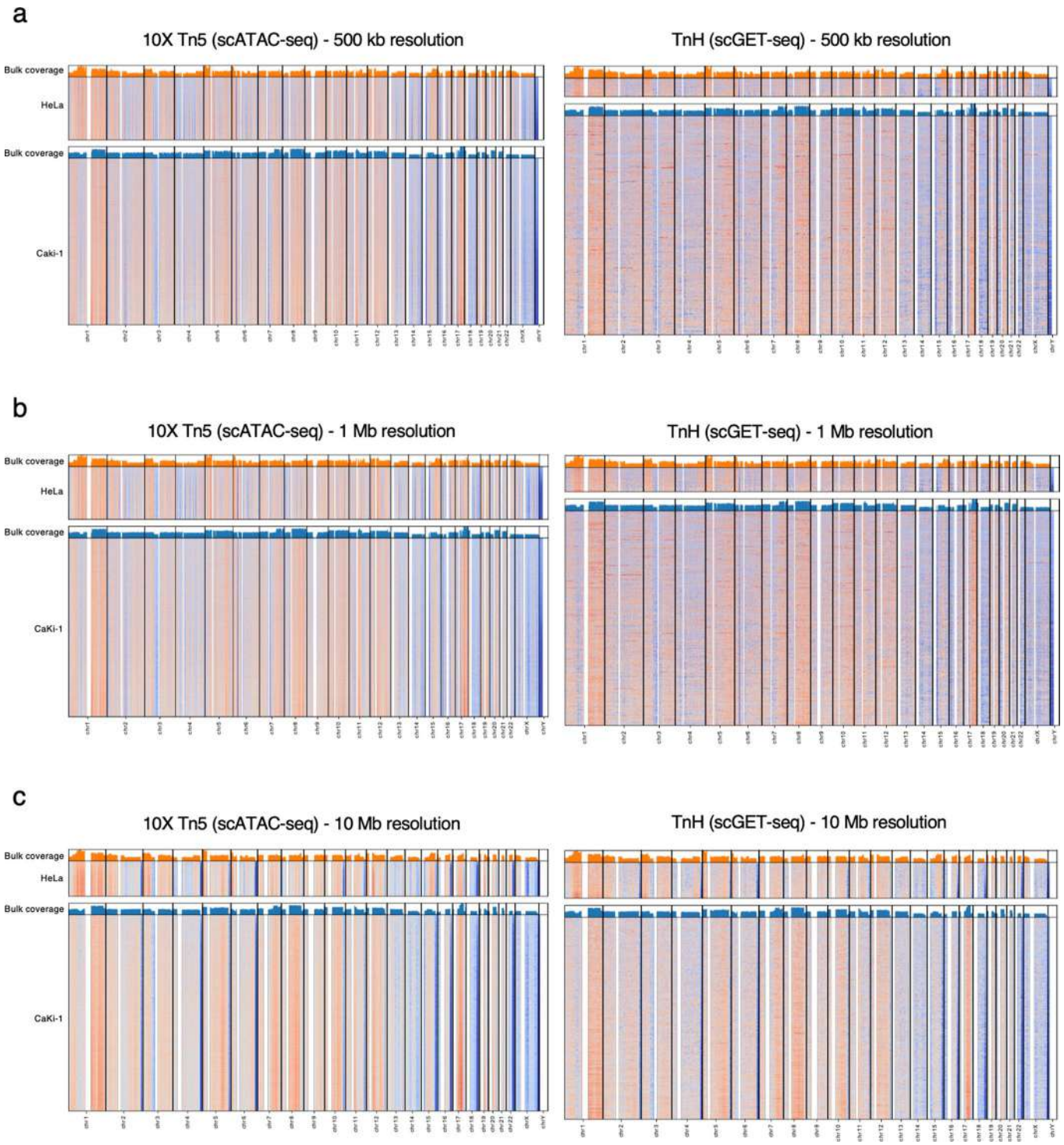
Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Optimization of ATAC-seq protocol introducing a combination of Tn5 and TnH transposases. a, Effect of altering Tn5 (green) to TnH (red) ratio on tagmentation profiles when adding both enzymes simultaneously at the beginning of the 60 minutes of the transposition reaction. **b**, Sequential addition of the same quantity of Tn5 and then TnH enzyme after 30 minutes of the transposition reaction results in a balanced distribution of enrichment signals between the two enzymes. Experiments performed on Caki-1 cell line.

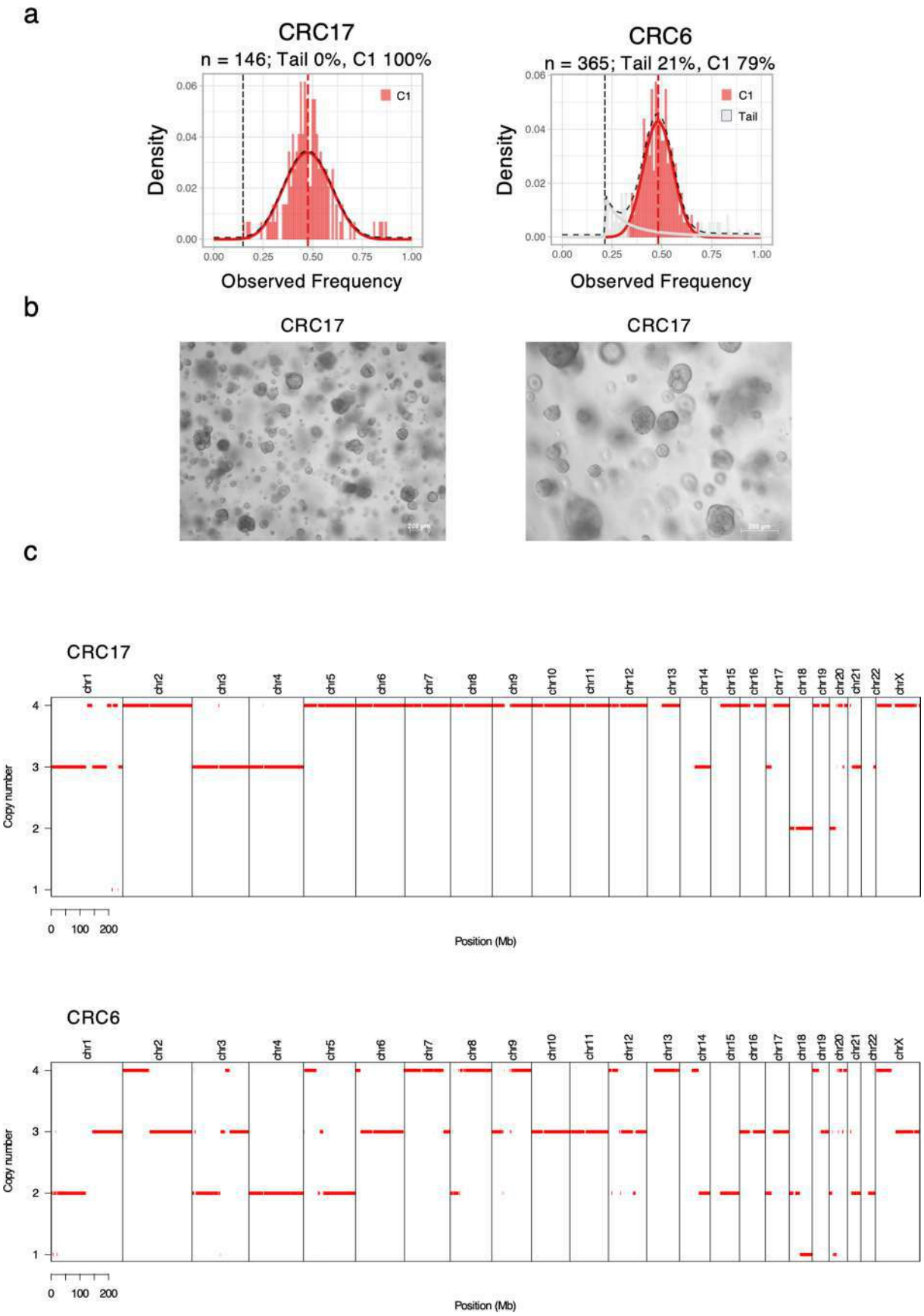


Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Characteristic of scGET-seq data. a Abundance of unique cell barcodes retrieved by scATAC-seq performed on Caki-1 cells using the provided ATAC transposition enzyme (10X Tn5; 10X Genomics) (blue) compared to cell barcodes countable by TnH (orange) or Tn5 (green) alone. scGET-seq performance (Tn5 + TnH) is represented in red. The curves are largely overlapping, indicating no evident bias in single cell identification**; b** Distribution of per-cell normalized coverage over fixed-size genomic bins (5 kb) is reported for 10X Tn5 (blue) and for signal obtained by TnH (orange) and Tn5 (green). While Tn5 is comparable to 10X Tn5, TnH returns higher and less overdispersed per-bin coverages. White dot in boxplots reprents the median, boxes span between the 25th and 75th percentiles, whiskers extend 1.5 times the interquartile range. n = 3363, 1281 and 1537 cells in one experiment; **c** Saturation analysis for selected libraries. Dotted lines show the fitted incomplete Gamma functions on subsampled data; red solid lines show subsampling data from the same libraries**; d** Tn5 (green) and TnH (red) enrichment profiles obtained from scGET-seq (pseudo-bulk) or from ATAC-seq performed by using the two enzymes separately, compared with H3K4me3 (green) and H3K9me3 (red) ChIP-seq data. Data shown refer to experiments performed on Caki-1 cells.

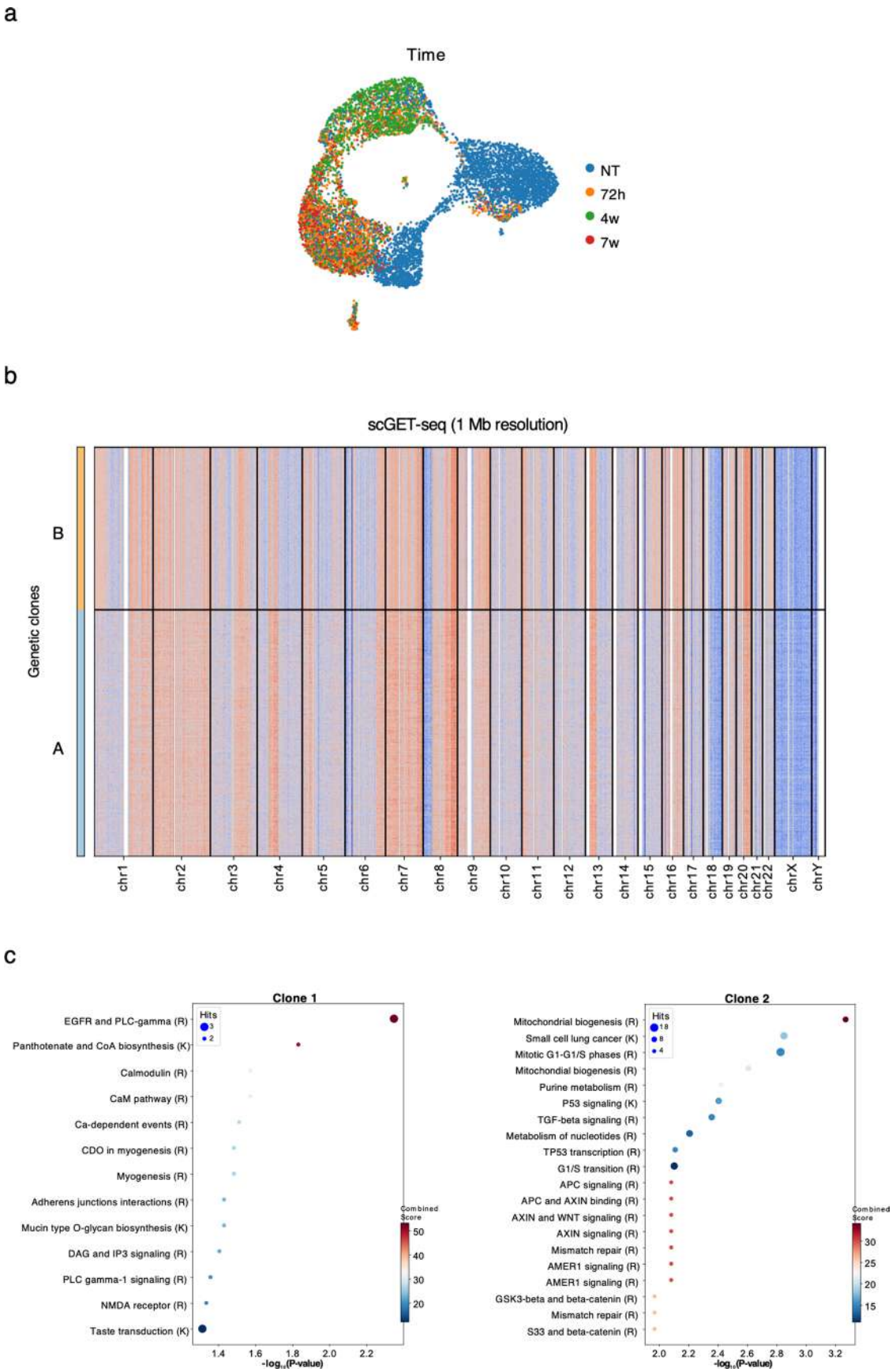


Extended Data Fig. 5 | Copy Number analysis at multiple resolutions. a, Segmentation profiles in individual cells profiled by 10X Tn5 (scATAC-seq) (left panel) or TnH scGET-seq (right panel) at 500 kb. **b**, Segmentation profiles in individual cells profiled by 10X Tn5 (scATAC-seq) (left panel) or TnH scGET-seq (right panel) at 1 Mb. **c**, Segmentation profiles in individual cells profiled by 10X Tn5 (scATAC-seq) (left panel) or TnH scGET-seq (right panel) at 10 Mb. On top of each heatmap the genome-wide coverage of bulk sequencing of corresponding cell lines is represented. Centromeric regions and gaps (in white) have been excluded from the analysis.



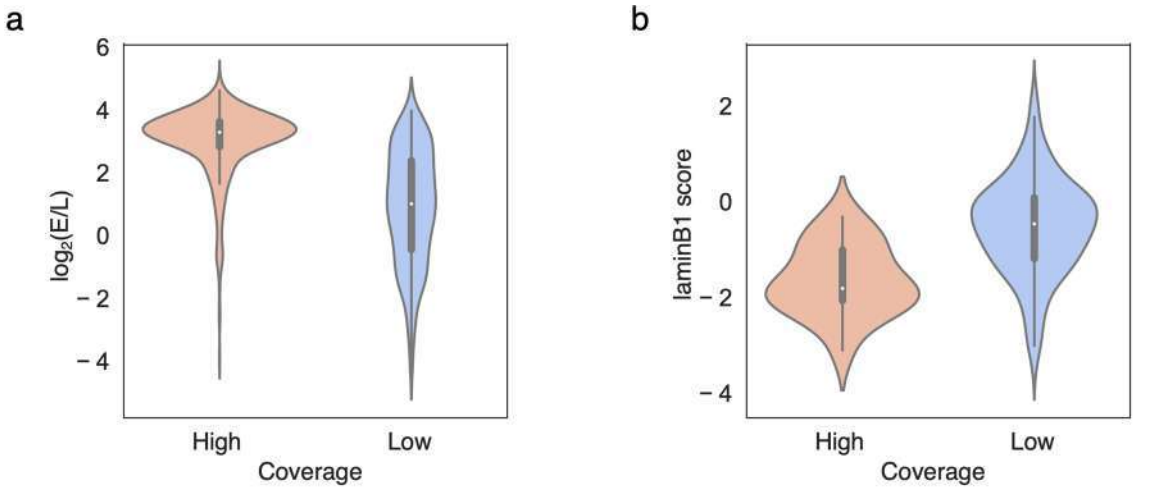
Extended Data Fig. 6 | See next page for caption.

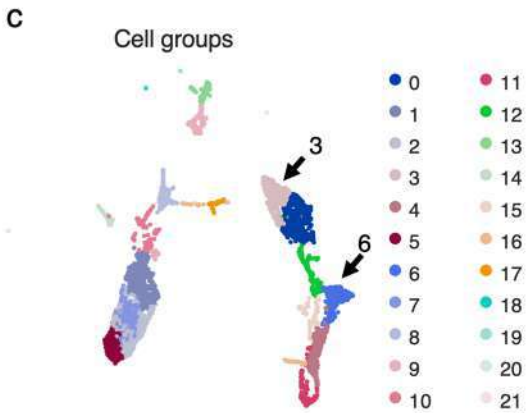
Extended Data Fig. 6 | Characterization of Patient Derived Organoids. a, evaluation of clonal structure of two PDO (CRC6 and CRC17) by exome sequencing; the histogram show the distribution of the cancer cell fraction estimated from the analysis of somatic mutations; in both organoids we observe a monoclonal structure **b**, 5X (left panel) and 10X (right panel) magnification contrast phase images of PDO #CRC17 obtained from a liver metastasis of a CRC patient (n>5); c absolute copy number of CRC17 and CRC6 as revealed by whole exome sequencing; data in panel c are equivalent to barplots over heatmaps in Fig. 3a.



Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | scGET-seq analysis on PDX samples. a, UMAP embedding of individual cells as in Fig. 3, colored by the time PDX were harvested. **b**, Segmentation profiles in individual cells profiled by scGET-seq at 1 Mb resolution expressed as log2(ratio) over the median signal. Cells are clustered according to genetic clones. Red: positive values; Blue: negative values. Centromeric regions (white) have been excluded from the analysis because they correspond to low mapping and not fully characterized regions.



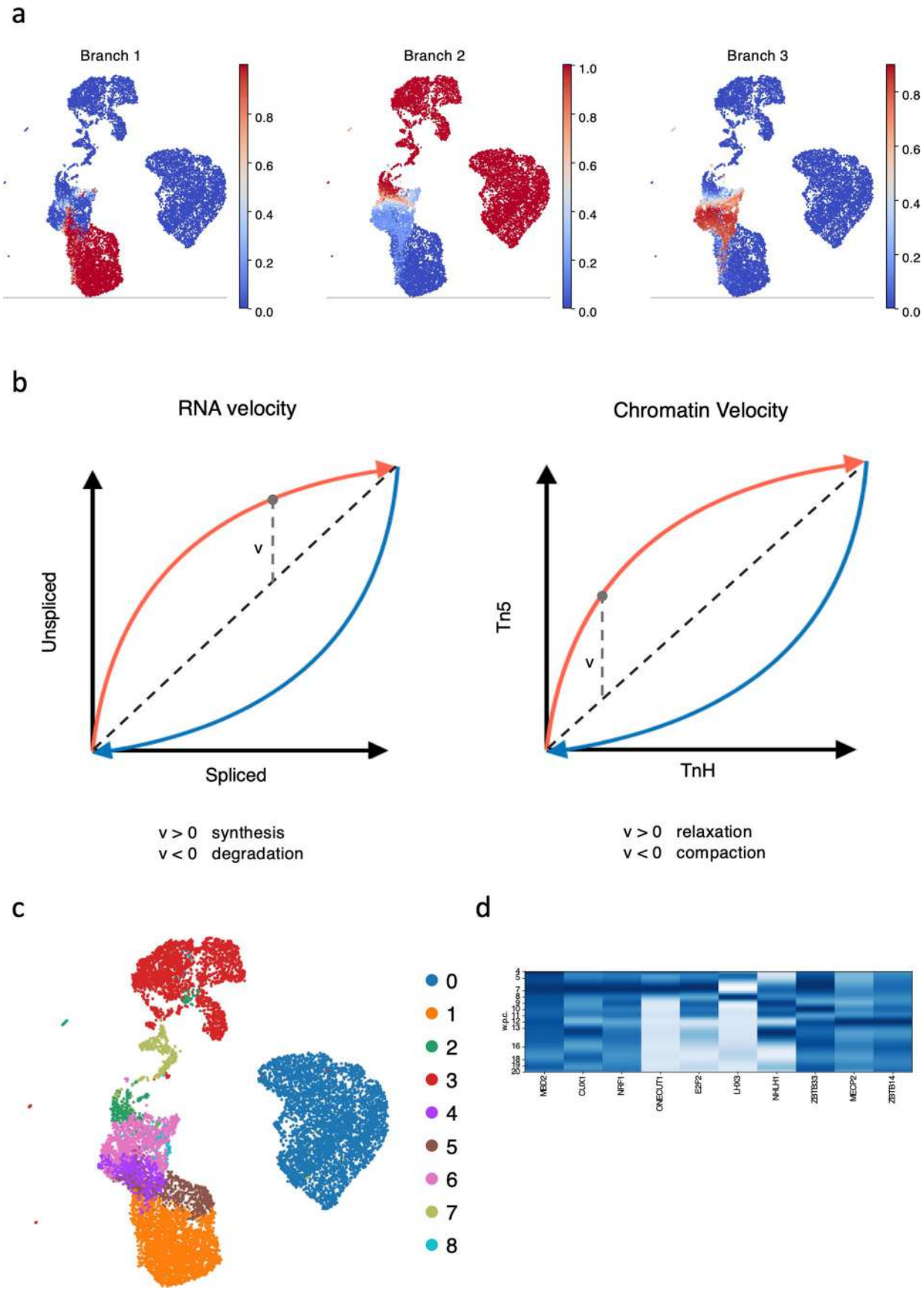


d

			chr2:98658425-98684400			chrUn_JH584304:0-114452		
group	phase	scramble	score	p-value	fdr	score	p-value	fdr
0	G1/S	45.00%	0.198	1.59E-01	1.59E-01	0.198	1.99E-01	1.99E-01
1	G2/M	53.47%	-0.172	1.00E+00	1.00E+00	-0.133	1.00E+00	1.00E+00
2	G2/M	76.09%	-0.178	1.00E+00	1.00E+00	-0.105	1.00E+00	1.00E+00
3	G1/S	71.20%	***0.323***	***2.99E-05***	***2.99E-05***	***0.329***	***3.50E-05***	***3.50E-05***
4	G1/S	31.98%	0.126	1.00E+00	1.00E+00	0.067	1.00E+00	1.00E+00
5	G2/M	47.72%	-0.204	1.00E+00	1.00E+00	-0.146	1.00E+00	1.00E+00
6	G1/S	44.09%	***0.387***	***9.19E-07***	***9.19E-07***	0.138	1.00E+00	1.00E+00
7	G2/M	40.45%	-0.319	2.03E-03	2.03E-03	-0.189	1.00E+00	1.00E+00
8	G2/M	51.27%	-0.361	2.81E-02	2.81E-02	-0.287	1.00E+00	1.00E+00
9	G1/S	48.68%	-0.698	1.71E-15	1.71E-15	-0.523	3.50E-07	3.50E-07
10	G2/M	42.54%	-0.17	1.00E+00	1.00E+00	-0.073	1.00E+00	1.00E+00
11	G1/S	55.87%	-0.589	6.80E-10	6.80E-10	-0.457	1.02E-04	1.02E-04
12	G1/S	29.24%	0.121	1.00E+00	1.00E+00	0.163	1.00E+00	1.00E+00
13	G2/M	39.46%	-1.339	4.69E-48	4.69E-48	-0.528	2.86E-05	2.86E-05
14	G2/M	59.84%	-0.24	1.00E+00	1.00E+00	-0.28	1.00E+00	1.00E+00
15	G1/S	30.63%	0.07	1.00E+00	1.00E+00	-0.47	2.96E-02	2.96E-02
16	G1/S	56.04%	0.595	1.41E-03	1.41E-03	0.576	3.58E-03	3.58E-03
17	G2/M	28.92%	0.511	8.87E-02	8.87E-02	0.468	5.97E-01	5.97E-01
18	G1/S	0.00%	-1.168	4.28E-03	4.28E-03	-0.089	1.00E+00	1.00E+00
19	G1/S	0.00%	-0.084	1.00E+00	1.00E+00	-0.129	1.00E+00	1.00E+00
20	G2/M	0.00%	0.298	1.00E+00	1.00E+00	-0.011	1.00E+00	1.00E+00
21	G2/M	100.00%	-0.839	1.00E+00	1.00E+00	0.138	1.00E+00	1.00E+00

Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | scGET-seq profiling of NIH-3T3 cells knocked-down for Kdm5c. a, Distribution of early-to-late ratio of 2-stage Repli-seq data for NIH-3T3 cells. Violin plots represent the value of log2(E/L) values over DHS regions which are differential in the high-vs-low coverage cells in Fig. 4a (Mann-Whitney U = 36169.5, p = 1.403e-84). White dot in boxplots represents the median, boxes span between the 25th and 75th percentiles, whiskers extend 1.5 times the interquartile range. n = 35438 regions. **b**, Distribution of lamin-B1 DamID scores for NIH-3T3 cells. Violin plots represent the value of DamID scores over DHS regions which are differential in the high-vs-low coverage cells in Fig. 4a (Mann-Whitney U = 723.0, p = 4.621e-6). White dot in boxplots represents the median, boxes span between the 25th and 75th percentiles, whiskers extend 1.5 times the interquartile range. n = 35438 regions. **c**, UMAP embedding of individual cells coloured by cell groups, identified by Leiden algorithm with resolution parameter set to 0.2. **d**, Results of the linear model calculating the group-wise differences between TnH and Tn5 enrichment. For each group we reported the coefficient of the model, the p-value and the Benjamini-Hochberg corrected p-value. Values are reported for the two genomic regions including the Major primers (see text). Barplot indicates the proportion of shScr-treated for each cell group.



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | scGET-seq profiling of a developmental model of iPSC. a, UMAP embedding of individual cells colored by the probability of being included in a trajectory branch estimated by Palantir. Three major branches have been identified, roughly corresponding to the three cell types profiled in this study. **b**, Schematic representation of the phase portraits underlying Chromatin Velocity. In RNA-velocity, the time derivative of the unspliced/spliced RNA is used to estimate synthesis or degradation of RNA; in Chromatin Velocity, the same procedure is applied on Tn5/TnH data to estimate chromatin relaxation or compaction. **d**, UMAP embedding of individual cells colored by cell clusters. **e**, Heatmap shows average expression profiles of TF with the top 10 most negative on PLS2 during the early brain development. Darker color indicates higher expression. w.p.c.: weeks post conception.

nature research

Corresponding author(s): Giovanni Tonon, Davide Cittaro

Last updated by author(s): Apr 7, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Real time qPCR analysis was performed using Viia 7 Real Time PCR System (Applied Biosystems). Cells were counted using TC20 automated cell counter (BioRad). Tagmentation products, cDNA traces and final libraries were evaluated using 4200 Tapestation System (Agilent). Cells were encapsulated using Chromium Platform (10X Genomics). High-throughput sequencing was performed on Miseq or Novaseq6000 platforms (Illumina).
Data analysis	Real-time qPCR data were analyzed using GraphPad Prism version 7 for Mac, GraphPad Software, San Diego, California USA. Demultiplex was performed using bcl2fastq (Illumina), cellranger-atac (10X Genomics) or cellranger (10X Genomics). Tn5 and TnH read tags were separated using tagdust (v2.33). Read tags were aligned to reference genome using bwa mem (v0.7.12). Reads were deduplicated using samblaster. Genome tracks were created using bamCoverage from the deepTools suite. Peaks for H3K4me3 ChIP were called using MACS (v.2.2.7). Peaks for H3K9me3 ChIP were called using SICER (v2). Single cell alignments were processed with custom software available at https://github.com/dawe/scatACC. Single cell data were processed using scanpy, schist and scvelo. Mutation analysis was performed using freebayes. Clonal analysis was performed using MOBSTER

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All relevant data are included in the manuscript. Sequencing data are deposited on the ArrayExpress database with the following ID: E-MTAB-9648 (ChIP-seq, bulk ATAC-seq and GET-seq), E-MTAB-10218 (fibroblast, iPSC, NPC scGET-seq), , E-MTAB-10220 (fibroblast, iPSC, NPC scRNA-seq), E-MTAB-9650 (Caki-1-HeLa scGET-seq), E-MTAB-9651 (shKdm5c and shScr NIH-3T3 scGET-seq), E-MTAB-9659 (PDX scGET-seq), E-MTAB-10219 (patient derived organoids scGET-seq)

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No power analysis have been performed before the experiment. In scGET-seq the nuclei suspensions were prepared in order to get the following number of nuclei for each experimental condition, as target nuclei recovery: 5,000 (fibroblast, iPSC, NPC), 20,000 (Caki-1-HeLa), 10,000 (shKdm5c and shScr NIH-3T3), 20,000 (PDX), 5,000 (PDO). In scRNA-seq the cell suspensions were prepared in order to retrieve 5,000 cells.
Data exclusions	In TAM-ChIP-qPCR one of the two H3K4me3 biological replicate was excluded because no significant signal was detected for any condition. Raw data for TAM-ChIP-seq are not available because of a storage failure and subsequent data loss.
Replication	TAM-ChIP was performed on two biological replicates for each condition (H3K4me3, H3K9me3 and NoAb). For each biological replicate three technical replicates were analyzed in Real-Time qPCR. Representative sequencing tracks are shown for TAM-ChIP, ATAC-seq and GET-seq.
Randomization	For in vivo drug treatment of PDX models, mice were randomized into treatment arms that received either placebo or cetuximab.
Blinding	For in vivo drug treatment of PDX models, tumor growth was monitored and tumor volumes were calculated; operators were blinded during measurements.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☐	☒ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☐	☒ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	Ab anti-H3K9me3 (ab8898 Abcam), Ab anti-H3K4me3 (07-473 Millipore) were used for TAM-ChIP and for ChIP experiments.
Validation	All antibodies were validated as described in Rondinelli et al. JCI 2015.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	All established cell lines were purchased from American Type Culture Collection (ATCC), except for HEK293T cell line that was used only for lentiviral production and was a kind gift from Prof. Luigi Naldini (San Raffaele Telethon Institute for Gene Therapy, Milan).
Authentication	NIH-3T3 and HeLa cell lines were genotyped using Cell ID™ System (Promega) for STR validation. Caki-1 cell line was genotyped by using established methods described in Keats et al. Blood (2007) 110 (11): 2485, based on Multiplex PCR Kit (206143, Qiagen) for testing of copy number variation.
Mycoplasma contamination	All cell lines were regularly tested for mycoplasma contamination and resulted negative.
Commonly misidentified lines (See ICLAC register)	We realise that both HeLa and Caki-1 cells are included in the ICLAC. However as reported above for both cell lines we conducted extensive genotyping assays, confirming their identities.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	For in vivo drug treatment of PDX models, 6-week-old male and female NOD (nonobese diabetic)/SCID (severe combined immunodeficient) mice were used.
Wild animals	not applicable
Field-collected samples	not applicable
Ethics oversight	Animal procedures were approved by the Italian Ministry of Health (authorization 806/2016-PR).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Human research participants

Policy information about studies involving human research participants

Population characteristics	Fibroblasts were isolated from two healthy donors. Donor A is a female, year of birth 1979; donor B is a female, year of birth 1981. No genotypic information is available. Organoids were derived from liver metastatic colorectal cancer from two patients. CRC17 is mutated in KRAS.
Recruitment	As for the differentiation experiment, donor A and donor B were recruited as part of a project on multiple sclerosis. Donor A is the dizygotic twin of a patient with relapsing-remitting multiple sclerosis; donor B is the monozygotic twin of a patient with relapsing-remitting multiple sclerosis. As for the PDO experiment, patients were recruited from a study on liver metastatic colorectal cancer patients (ACC_ORG).
Ethics oversight	Studies approved by Comitato Etico Ospedale San Raffaele (BANCA-INSPE 09/03/2017, ACC_ORG 19/06/2019)

Note that full information on the approval of the study protocol must also be provided in the manuscript.

ChIP-seq

Data deposition

☒ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☒ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links *May remain private before publication.*	https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-9648
Files in database submission	H3K4me3_kapa_1.bigwig H3K4me3_kapa_2.bigwig H3K9me3_kapa_1.bigwig H3K9me3_kapa_2.bigwig Input_kapa_1.bigwig Input_kapa_2.bigwig
Genome browser session (e.g. UCSC)	Not available

Methodology

Replicates	Two biological replicates were used for ChIP experiments.
Sequencing depth	All ChIP-seq data were sequenced with 150bp paired end strategy. Input were sequenced in single read at 150bp. H3K4me3_kapa_1 total: 19961410, uniquely mapped: 18382805 H3K4me3_kapa_2 total: 33548470, uniquely mapped: 30773239 H3K9me3_kapa_1 total: 43563908, uniquely mapped: 37961313 H3K9me3_kapa_2 total: 35281046, uniquely mapped: 31667718 Input_kapa_1 total: 31171148, uniquely mapped: 27898938 Input_kapa_2 total: 29526698, uniquely mapped: 26648442
Antibodies	Ab anti-H3K9me3 (ab8898 Abcam), Ab anti-H3K4me3 (07-473 Millipore)
Peak calling parameters	H3K4me3 peaks were called with the following parameters: callpeak -f BAM -g hs --keep-dup 1 --llocal 1000000 --slocal 50000 --nomodel --extsize 150. H3K9me3 peaks were called with default parameters.
Data quality	Not applicable
Software	H3K4me3 data were analyzed with MACS (v2.2.7). H3K9me3 data were analyzed with SICER (v2).

ANNEX II: FAST ANALYSIS OF SCATAC-SEQ DATA USING A PREDEFINED SET OF GENOMIC REGIONS.

Check for updates

METHOD ARTICLE

REVISED Fast analysis of scATAC-seq data using a predefined set of genomic regions [version 2; peer review: 2 approved]

Valentina Giansanti [1,2], Ming Tang[3], Davide Cittaro [2]

[1]Department of Informatics, Systems and Communication, University of Milano-Bicocca, Milan, Italy
[2]Center for Omics Sciences, IRCCS San Raffaele Institute, Milan, Italy
[3]FAS informatics, Harvard University, Cambridge, MA, USA

v2 **First published:** 20 Mar 2020, **9**:199
https://doi.org/10.12688/f1000research.22731.1
Latest published: 28 May 2020, **9**:199
https://doi.org/10.12688/f1000research.22731.2

Abstract

Background: Analysis of scATAC-seq data has been recently scaled to thousands of cells. While processing of other types of single cell data was boosted by the implementation of alignment-free techniques, pipelines available to process scATAC-seq data still require large computational resources. We propose here an approach based on pseudoalignment, which reduces the execution times and hardware needs at little cost for precision.
Methods: Public data for 10k PBMC were downloaded from 10x Genomics web site. Reads were aligned to various references derived from DNase I Hypersensitive Sites (DHS) using *kallisto* and quantified with *bustools*. We compared our results with the ones publicly available derived by *cellranger-atac*. We subsequently tested our approach on scATAC-seq data for K562 cell line.
Results: We found that *kallisto* does not introduce biases in quantification of known peaks; cells groups identified are consistent with the ones identified from standard method. We also found that cell identification is robust when analysis is performed using DHS-derived reference in place of *de novo* identification of ATAC peaks. Lastly, we found that our approach is suitable for reliable quantification of gene activity based on scATAC-seq signal, thus allows for efficient labelling of cell groups based on marker genes.
Conclusions: Analysis of scATAC-seq data by means of *kallisto* produces results in line with standard pipelines while being considerably faster; using a set of known DHS sites as reference does not affect the ability to characterize the cell populations.

Keywords

single cell, scATAC-seq, pseudoalignment

Open Peer Review

Reviewer Status ✓ ✓

	Invited Reviewers	
	1	**2**
version 2 (revision) 28 May 2020		✓ report
version 1 20 Mar 2020	✓ report	? report

1. **Iros Barozzi**, Imperial College London, London, UK

2. **Qiangfeng Cliff Zhang**, Tsinghua University, Beijing, China

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Davide Cittaro (cittaro.davide@hsr.it)

Author roles: Giansanti V: Data Curation, Formal Analysis, Resources, Software, Writing – Original Draft Preparation; **Tang M**: Data Curation, Formal Analysis, Resources, Software, Writing – Original Draft Preparation; **Cittaro D**: Conceptualization, Investigation, Methodology, Project Administration, Supervision, Writing – Original Draft Preparation

Competing interests: No competing interests were disclosed.

Grant information: DC and VG are supported by the Accelerator Award: A26815 entitled: "Single-cell cancer evolution in the clinic" funded through a partnership between Cancer Research UK and Fondazione AIRC. MT is supported by NIH grants 1U19MH114830 and 1U19MH114821.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2020 Giansanti V *et al*. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Giansanti V, Tang M and Cittaro D. **Fast analysis of scATAC-seq data using a predefined set of genomic regions [version 2; peer review: 2 approved]** F1000Research 2020, **9**:199 https://doi.org/10.12688/f1000research.22731.2

First published: 20 Mar 2020, **9**:199 https://doi.org/10.12688/f1000research.22731.1

REVISED **Amendments from Version 1**

In order to show that reference-based analysis of scATAC-seq data is not suitable only for well defined cell groups, as in PBMC, this version of the manuscript extends the analysis to a sparser and more homogeneous dataset (K562 cells). In addition, we analyzed computational resources needed to run the exemplified processes.

Any further responses from the reviewers can be found at the end of the article

Introduction

Recent technological advances in single-cell technologies resulted in a tremendous increase in the throughput in a relatively short span of time[1]. The increasing number of cells that could be analyzed prompted a better usage of computational resources. This has been especially true for the post-alignment and quantification phases. As a consequence, it is today feasible to run the analysis of single cell data on commodity hardware with limited resources[2], even when the number of observables is in the order of hundreds of thousands. Conversely, the analysis steps from raw sequences to count matrices lagged for some time. Alignment to the reference genome or transcriptome is largely dependent on classic aligners, without any specific option to handle single-cell data, with the notable exception of the latest implementation of STARsolo in the STAR aligner[3].

More recently, analysis of Next generation sequencing (NGS) data benefits from technologies based on *k*-mer processing, allowing alignment-free sequence comparison[4]. Most of these technologies require a catalog of *k*-mers expected to be in the dataset and, hence, subject of quantification. RNA-seq analysis relies on the quantification of gene/transcript abundances and, while it is possible to perform *de novo* characterization of unknown species in every experiment, it is common practice[5,6] to rely on a well-defined gene model such as GENCODE[7] to quantify expressed species. It is then possible to efficiently perform alignment-free analysis on transcripts to quantify gene abundances. Tools implementing this approach such as *kallisto*[8] or *salmon*[9] have been quickly adopted on a wide scale. Moreover, a recent implementation of *kallisto* extended its capabilities to the analysis of single cell RNA-seq data[10] by direct handling of cell barcodes and UMIs, allowing the analysis of such data in a streamlined way. Of notice, a scRNA-seq oriented implementation of *salmon* has been recently developed[11].

Analysis of epigenetic features by ATAC-seq requires the identification of enriched peaks along the genome sequence. This is typically achieved using peak callers such as MACS[12], with tuned parameters. Since ATAC-seq signal mirrors DNA accessibility as mapped by DNase-seq assays[13] and catalogs of DNase I Hypersensitive Sites (DHS) are available[14,15], it should be possible to perform reference-based ATAC-seq analysis in a way much similar to what is performed for RNA-seq analysis. In this paper we show it is indeed possible to perform single-cell ATAC-seq analysis using *kallisto* and *bustools*, with minor tweaks, using an indexed reference of ~1 million known DHS sites on the human genome.

Methods

Single cell ATAC-seq data

Single cell ATAC-seq data for PBMC were downloaded from the 10x Genomics public datasets (https://support.10xgenomics.com/single-cell-atac/datasets/1.1.0/atac_v1_pbmc_10k) and include sequences for 10k PBMC from a healthy donor. We used the Peak by cell matrix HDF5 (filtered) object as our ground truth.

Raw sequences for single cell ATAC-seq data for K562 cell line were downloaded from Short Read Archive (GEO ID GSE112200).

Generation of *kallisto* index

We downloaded the DNase I Hypersensitive Sites (DHS) interval list for hg19 genome from the Regulatory Elements DB[16]. Intervals closer than 500 *bp* were merged using bedtools[17].

We extracted DNA sequences for DHS intervals and indexed corresponding fasta files using *kallisto index* (v0.46.0) with default parameters, resulting in an index for the full DHS set (iDHSfull) and an index for the merged set (iDHS500). The same procedure was performed for the peak set identified by *cellranger-atac* and distributed along with the data (iMACS).

Processing of Chromium 10x data

kallisto requires the definition of the unique molecular identifiers (UMI) and cellular barcodes (CB) in a specific fastq file. For standard Chromium scRNA-seq data, these are substrings of R1 and RNA is sequenced in R2. Chromium scATAC-seq reads are not structured in the same way: paired end genomic reads are in R1 and R3, R2 includes only the 16 *bp* cellular barcode. In addition, `kallisto bus` expects only a single read with genomic information. Therefore we simulated appropriate structures in three different ways:

1. by adding 12 random nucleotides and mapping the R1 file (forward read):
```
kallisto bus -x 10xV2 modified_R1.fastq.gz
pbmc_10k_R1.fastq.gz
```
2. by extracting sequences of different length *n* (5, 10, 15, 20) from the 5' of R3 (reverse read) and mapping the R1 file:
```
kallisto bus -x 1,0,16:2,0,n:0,0,0
pbmc_10k_R1.fastq.gz
pbmc_10k_R2.fastq.gz
pbmc_10k_R3.fastq.gz
```
3. by extracting sequences of different length *n* (5, 10, 15, 20) from the 5' of R1 and then mapping the R3 file:
```
kallisto bus -x 1,0,16:2,0,n:0,0,0
pbmc_10k_R3.fastq.gz
pbmc_10k_R2.fastq.gz
pbmc_10k_R1.fastq.gz
```

We will refer to the second set of simulation as *n-fwd* and to the third set as *n-rev*, where *n* is the number of nucleotides considered as UMI. We also applied two different summarization strategies for *bustools count* step. In the first approach, pseudocounts are not summarized, the number of features matches the size of the index; in the second approach, summarized, we

let *bustools map* counts on iDHSfull to the merged intervals (Figure 1A).

Processing of Fluidigm C1 data

Reads were aligned to reference genome (hg19) using bwa mem (v0.7.12)[18], deduplication was performed using samblaster (v0.1.21)[19]. Only R2 were aligned in *bwa SE* configuration. Individual BAM files were merged using samtools and peaks were called from the pseudo-bulk using MACS2 (v2.2.7.1)[12] (paired end options: -q 0.1 --nomodel --shift 0, single read options: -q 0.1 --nomodel --shift -100 --extsize 200). Quantification was performed using bedtools multicov (-q 15).

kallisto quant was run with default parameters for paired end data. Only R2 were processed in *kallisto quant SE* with specific options (--single -l 300 -s 20). Individual counts from abundance files were merged using tximport[20].

In order to perform *kallisto bus* analysis we generated a set of 288 random CB which were used to create 288 matched fastq files. Once all read pairs and cellular barcodes have been concatenated into R1, R2 and CB fastq files, we ran *kallisto bus* with the same strategy used for PBMC data (-x 1,0,16:2,0,10:0,0,0).

Analsyis of single-cell data

Counts matrices were analysed using Scanpy (v1.4.2)[2] with standard parameters. In PBMC data, we filtered out cells that had less than 200 regions and regions that were not at least in 10 cells. In K562 data we only excluded regions that were not shared by at least 20 cells. The count matrices were normalized and log transformed. The highly variable regions were selected and the subsetted matrices processed to finally clusterized the data with the Leiden algorithm[21], setting resolution parameter to 0.2. Marker peaks were selected using Wilcoxon rank-sum test. Adjusted Mutual Information (AMI) was used to evaluate the concordance between the 10x and matrices derived from kallisto.

Cellular barcodes were extracted using UMITools[22], setting the expected number of cells to 10,000.

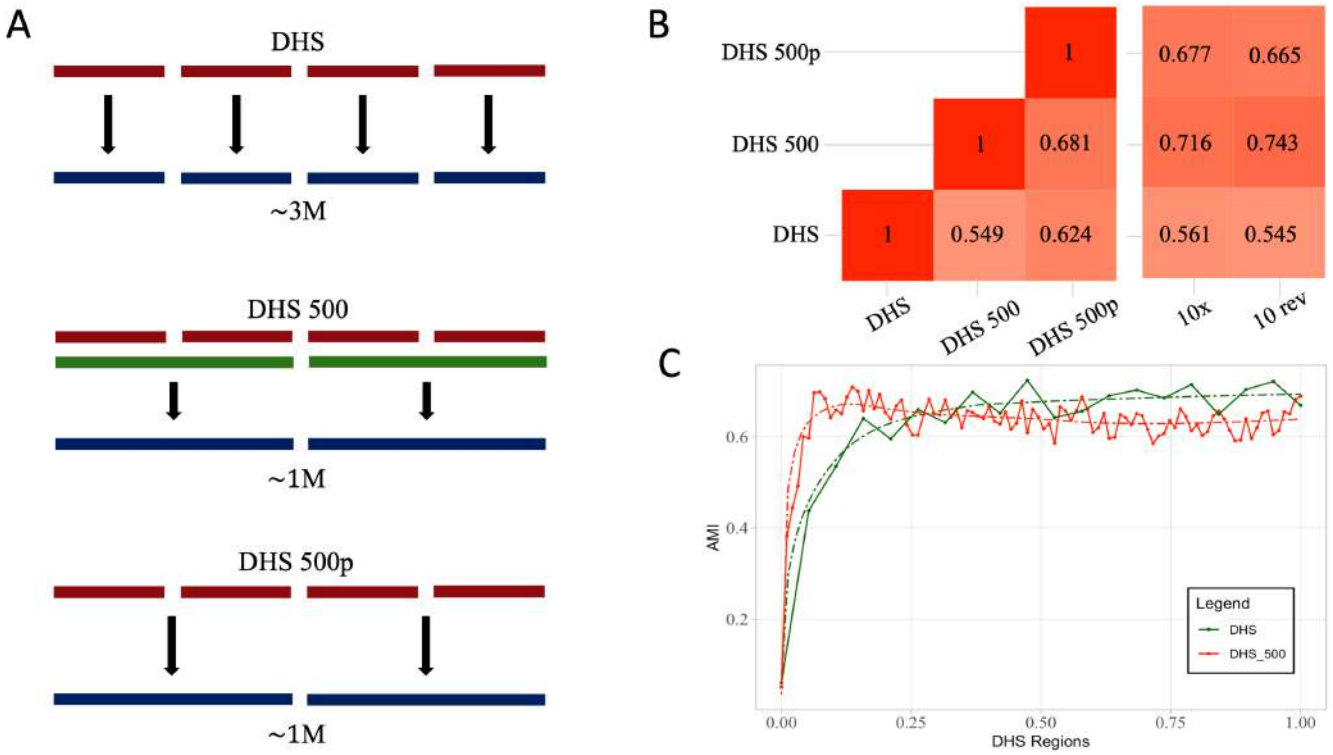


Figure 1. (**A**) Graphical depiction of processing of pseudoalignment over DHS, based on three DHS derived indices. The first (DHS) generated by *kallisto* on ~3 *M* DNase I sites, the second (DHS500) by merging regions closer than 500 *bp* and the last (DHS500p) by projecting the result of DHS index to DHS500 using bustools capabilities. (**B**) Heatmaps representing MI scores for the DHS derived matrices. The heatmap on the left reports the pairwise MI values between DHS, DHS500 and DHS500p strategies. The heatmap on the right represents MI values comparing the DHS derived strategies to the *cellranger-atac* (10x) results or 10- *rev* strategy. DHS500 strategy achieves the highest scores. (**C**) AMI values comparing DHS (green line) and DHS500 (red line) strategies to *cellranger-atac* at different thresholds on the number of regions considered in the analysis. When approximately 50,000 regions are included, the AMI stabilizes at its maximum. Dashed lines represent the fit curves.

The PBMC matrices derived from *kallisto* and *cellranger-atac* were also imported into Seurat V3[23]. Gene activity score was calculated using the CreateGeneActivityMatrix function or directly summarized by *kallisto*. The annotated 10*k* PMBC scRNA-seq Seurat object was downloaded from the link available in their v3.1 ATAC-seq Integration Vignette (https://satijalab.org/seurat/v3.1/atacseq_integration_vignette.html).

Cell labels from the scRNA-seq data were transferred to scATACseq data using TransferData function based on the gene activity score. All the analyses were carried out using standard parameters. Jaccard similarities were evaluated using the scclusteval (v0.1.1) package[24].

Results

Limitations of *kallisto*-based analysis

At time of writing, *kallisto* does not natively support scATAC-seq analysis, though it can be applied to any scRNA-seq technology which supports CB and UMI. According to the *kallisto* manual, the technology needs to be specified with a tuple of indices indicating the read number, the start position and the end position of the CB, the UMI and the sequence respectively. In this sense, the technology specifier for standard 10x scRNA-seq with v2 chemistry is 0,0,16:0,16,26:1,0,0 (see *kallisto* manual for details). Using this logic, a single fastq file contains sequence information and UMI is always required. scATAC-seq from 10x genomics is sequenced in paired-end mode and there is no definition of UMI, reads are deduplicated after genome alignment.

kallisto requires an index of predefined sequences to perform pseudoalignment, typically transcript. When applied to scATAC-seq analysis, it does not allow for any epigenomic analysis, including the identification and quantification of enriched regions. Therefore, we computed an index on the genomic sequences for the 80,234 peaks identified by *cellranger-atac* and distributed along with fastq files. This ensures that the subsequent analysis were performed on the same regions and allowed us to quantify the bias, if any, introduced by *kallisto*.

kallisto primary analysis on PBMC data

We tested different strategies to overcome the technical limits and the absence of UMI. We evaluated concordance of different approaches using AMI between cell groups identified with equal processing parameters. Analysis based on *cellranger-atac* results is considered as ground truth. Results are reported in Table 1.

We tested two main strategies: first, the R1 is pseudoaligned and the initial nucleotides of R2, cut at different thresholds, are used as UMI (pseudoUMI hereafter). As UMI is needed for deduplication, we reasoned that a duplicate in scATAC-seq should be identified by the same nucleotides in the first portion of the read, where quality is higher. We observed generally high values of AMI, with the notable exception of pseudoUMI 5 *nt* long. Since basecall qualities are generally higher for R1 and *kallisto* does not use qualities in pseudoalignment, we tested the strategy where R2 is mapped and R1 is used to derive pseudoUMI. Again, 5 *nt* pseudoUMI raised the worst results, while AMI values were slightly higher than the forward configuration. In particular, we noticed the highest AMI values when R2 is used and pseudoUMI is 10 *nt* long (*AMI* = 0.7625). Second, we tested a configuration using R1 as sequence and 10 *nt* UMI randomly generated. Interestingly, concordance remains in line with previous experiments (*AMI* = 0.7272).

Table 1. Comparision of *cellranger-atac* and *kallisto* analysis. The table reports Adjusted Mutual Information between single cell cluster assignments on *cellranger-atac* data and *kallisto* analysis. Different strategies to evaluate pseudoUMI are reported. All simulations raised high AMI values, both in the forward and reverse approach, except for the pseudoUMI of length 5. The 10-Reverse configuration reached the highest score.

Comparison	Forward	Reverse
10x vs 5nt	0.1854	0.1733
10x vs 10nt	0.7434	0.7625
10x vs 15nt	0.7571	0.7398
10x vs 20nt	0.7356	0.7520
10x vs Random	0.7272	*None*

These data indicate that *kallisto* is able to properly quantify enrichments in scATAC-seq and does not introduce a considerable bias.

Analysis of DHS as reference

One major limitation of a *kallisto*-based approach to scATAC-seq is the lack of peak calling routines and the need of an index of sequences for pseudoalignments. Hence, we reasoned that we could use any collection of regions that putatively would be target of ATAC-seq experiments. Since ATAC-seq is largely overlapping DHS, we exploited regions defined in the ENCODE project[25]. The DHS data provided by ENCODE includes 2,888,417 sites. We generated an additional dataset by merging regions closer than 500 *bp* into 1,040,226 sites. We performed pseudoalignment on the full dataset, on the merged dataset and on the full dataset summarized by *bustools* (Figure 1A, see *Methods*). Pairwise comparison between performances of the three methods reveals lower values of AMI (Figure 1B). Comparison with 10x data and the configuration 10-*rev* previously performed shows high values of AMI when considering merged DHS intervals (*AMI* = 0.7164 and 0.743 respectively). When pseudoalignmets are performed on the full DHS set, performance degrades to lower AMI values. Since the number of DHS intervals is considerably higher than the typical number of regions identifiable by ATAC-seq, we tested the trend of AMI at different cutoffs on the number of DHS included in the analysis (Figure 1C). AMI reaches a plateau when approximately 50,000 regions are included into the analysis. This defines a reasonable target for filtering during preprocessing

stages of scATAC-seq data. Taken together, these findings support the suitability of using *kallisto* for identification of cell identities in scATAC-seq without any prior knowledge of the epigenetic status of single cells.

Identification of marker regions

A crucial step in the analysis of scATAC-seq data is the identification of marker peaks which can be used to functionally characterize different clusters. We tested the ability of our reference-based approach to identify differential DNase I hypersensitive sites that are overlapping or close to peaks identified with standard analysis. To this end, we first matched cell groups from DHS500 to groups identified after *cellranger-atac*. We selected the top 1,000 peaks marking each DHS500 group and evaluated the concordance by mutual distance to the top 1,000 significant markers in the matched groups ($p < 0.05$), we could identify significant markers only in five matched clusters. We found that the large majority of peaks (>= 80%) were overlapping between the two strategies or closer than 20 *kb* (Figure 2). These results confirm the substantial equivalence between the standard strategy and the reference-based one.

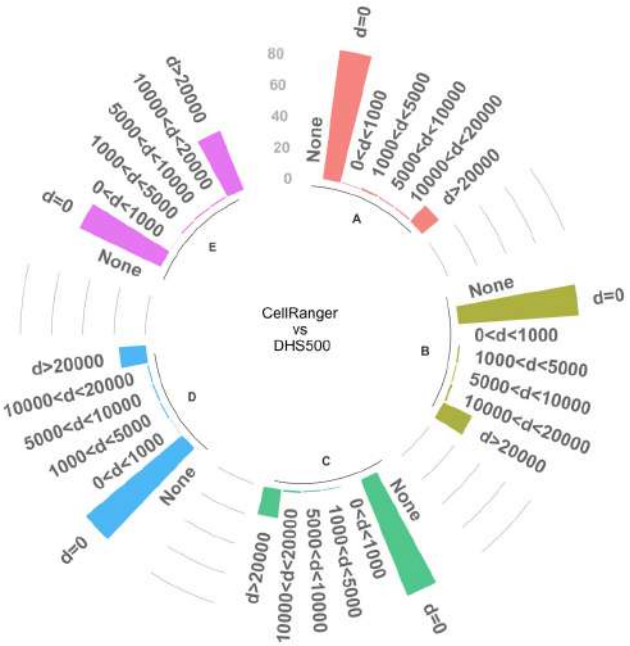


Figure 2. Analysis of peak concordance. The bars represent the proportion of marker peaks that are in common between DHS500 and *cellranger-atac*-based strategies at different distance thresholds. Only the top 1,000 significant peaks ($p < 0.05$) were included in the analysis; the graph reports results for the 5 cell clusters (**A**–**E**) that contain the required amount of significant markers. The chart also reports the proportion of peaks without any match (*None*).

Integration with scRNA-Seq data and cluster labeling

In addition to the analysis of technical suitability of *kallisto* for the analysis of scATAC-seq data, we investigated its validity in extracting biological insight. To this end, we performed a more detailed analysis of PBMC data by label transferring using Seurat V3[23], with the hypothesis that different approaches could lead to mislabeling of cells clusters. Matching is performed with the help of Gene Activity Scores calculated as sum of scATAC-seq counts over gene bodies extended 2 *kb* upstream the TSS, Seurat's default approach. We applied the same transferring protocol on data derived from *cellranger-atac* counts and from the DHS500 approach (Figure 3), finding no relevant differences in the UMAP embeddings. A detailed quantification of cluster matches reveals a slight deviance in the characterization of NK subpopulations (Figure 4A). In addition to scores calculated by Seurat, we tested the ability of *bustools* summarization step to project and sum scATAC-seq values into Gene Activity using the identical mapping to extended gene bodies. We found that gene activity score obtained by kallisto is similar to Seurat's CreateGeneActivityMatrix (Figure 4B) in terms of cell labeling, with the additional advantage of reduced run time.

Analysis of K562 cell lines

PBMC mixtures among the *de facto* standards in single cell benchmarks; it may be argued that the heterogeneity of the mixture can be easily revealed, implying that the differences between cell populations are large enough to be spotted also with suboptimal approaches. We analyzed scATAC-seq data for 288 K562 cells[26], profiled on a Fluidigm C1 apparatus, to test the consistency of our approach on a supposedly homogeneous population. Since sequences are available for each cell separately, we could extend our tests to the standard *kallisto* quantification procedure (*kallisto quant*), performing separated cell-based pseudoalignments. We explored seven different strategies, either based on paired-end reads (*bwa PE + MACS*, *bwa PE + DHS, kallisto quant PE*) or single reads (*bwa SE + MACS, bwa SE + DHS, kallisto bus* and *kallisto quant SE*). We tested single read modality to accomplish a fair comparisions with *kallisto bus.* In our tests, *bwa PE + MACS* resembles a typical approach for the analysis of such data (as in 26). Strategies based on *kallisto* and strategies named with *DHS* make use of the DHS500 set of regions.

Overall, we found a high concordance among all strategies. Two major cell groups could be identified using the equal processing parameters (Figure 5A) and cells were found generally classified into consistent groups (Figure 5B), with the notable exception of *bwa SE + MACS*. Excluding the latter, AMI ranges between 0.69 and 0.97. Interestingly, the comparison between *bwa PE + MACS* and *bwa PE + DHS* (*AMI*=0.74) suggests that the major source of differences is the set of regions, not the alignment and quantification method. The concordance between marker regions, measured by Jaccard's coefficient, reveals

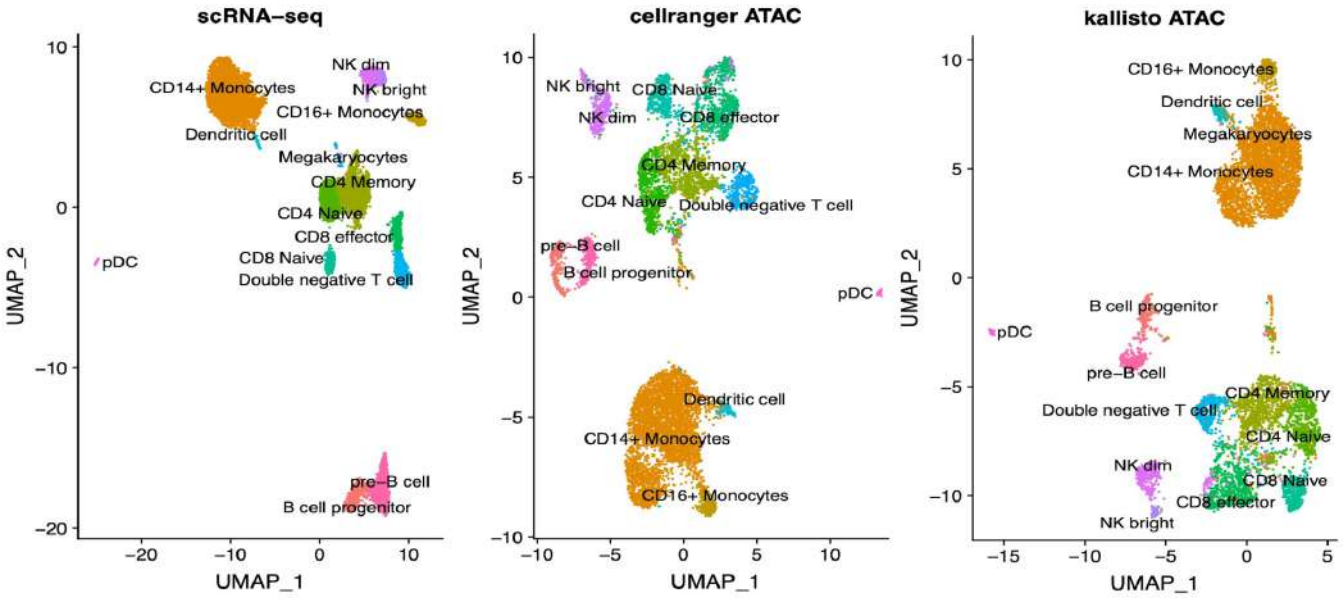


Figure 3. Results of label transfer from reference populations. The UMAP plot on the left represents scRNA-seq data of 10k PBMC as returned by Seurat vignette. The UMAP plots in the middle and on the right represent scATAC-seq analysis on *cellranger-atac* or *kallisto* analysis respectively. Cell clusters are consistent in their topology in the three plots, indicating the validity of *kallisto* for this kind of analysis.

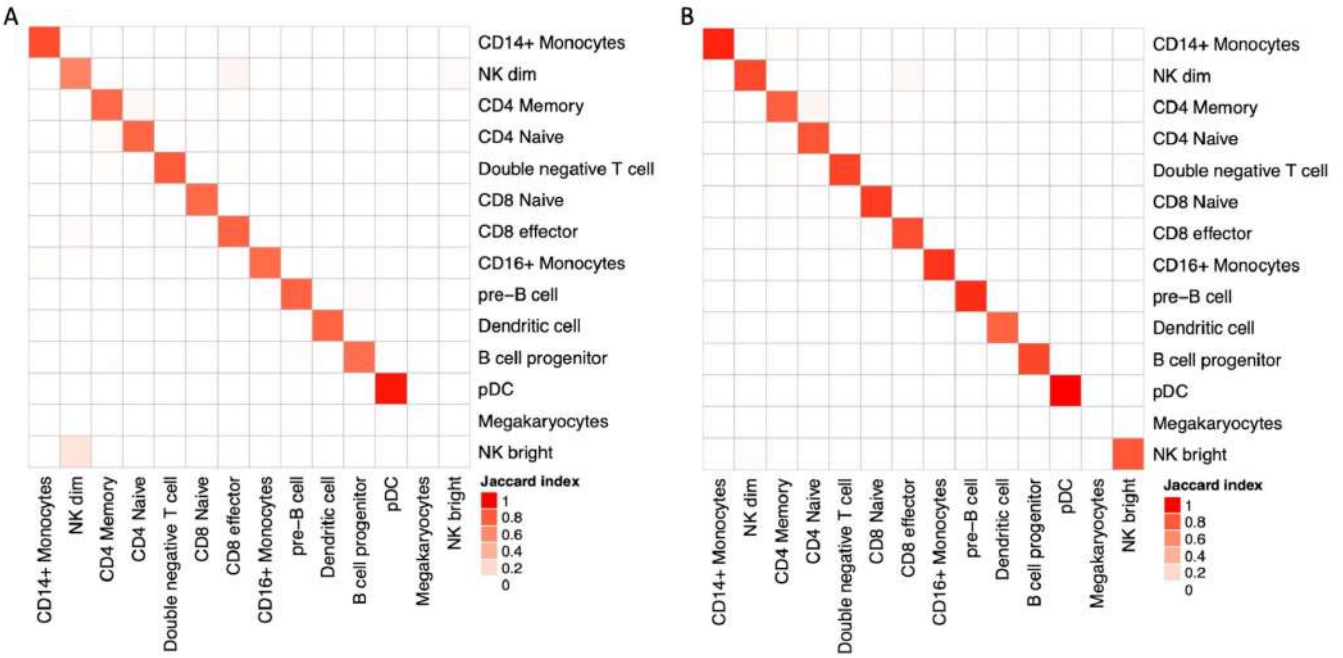


Figure 4. Analysis of Gene Activity Scores. (**A**) Pairwise Jaccard similarity between cell annotations as a result of label transfer from RNA-seq data using Gene Activity Score evaluated by Seurat. Concordance between results after *cellranger-atac* (rows) and DHS500 (columns) are largely comparable, with the notable exception of NK subpopulations. (**B**) Pairwise Jaccard similarity between cell annotations on DHS500 when Gene Activity Score is computed by Seurat (rows) or by *bustools* summarization step (columns).

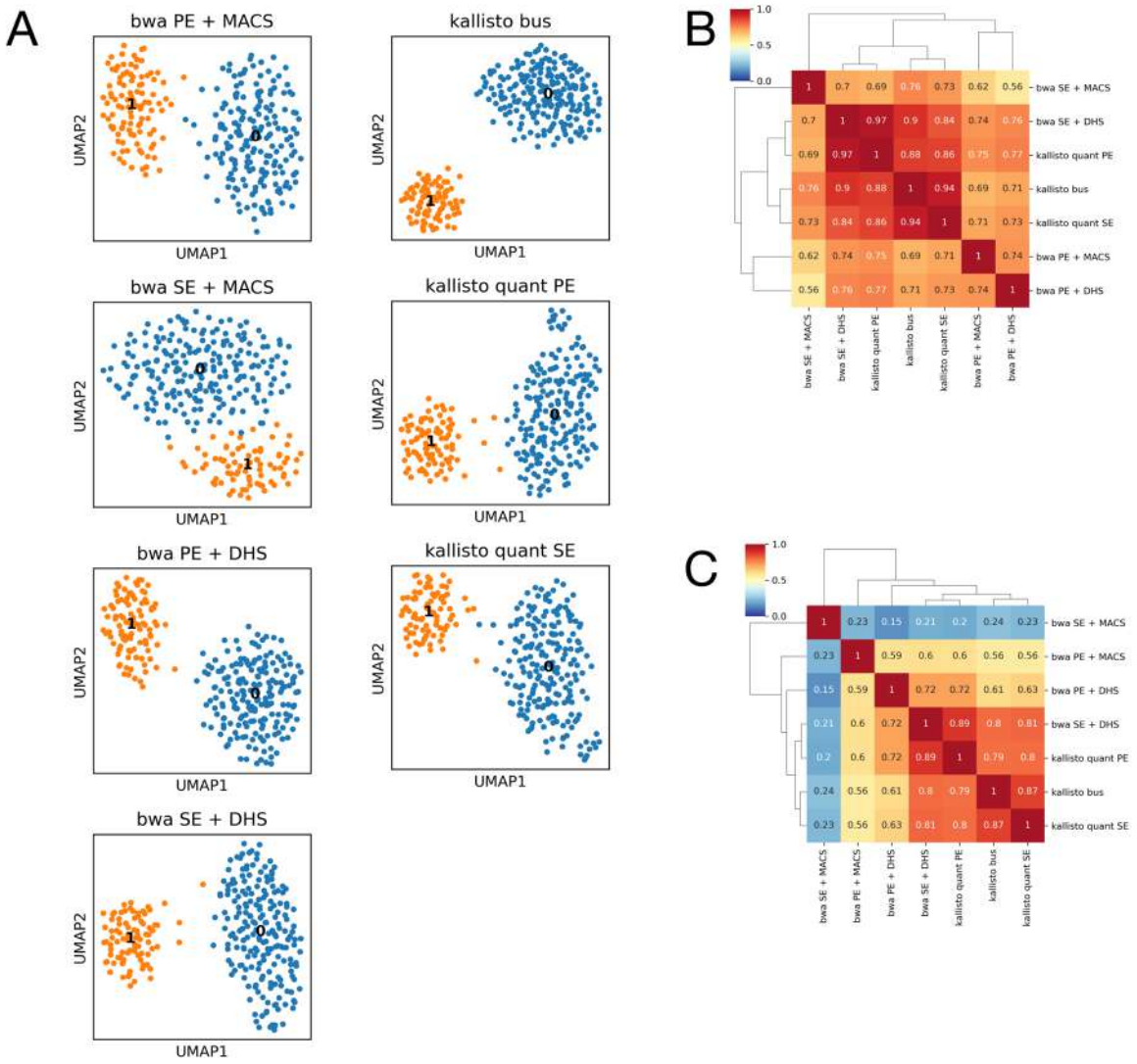


Figure 5. Analysis of K562 cell lines. Comparison of multiple standard- and reference-based approaches on scATAC-seq of K562 cell line. (**A**) UMAP embeddings for the multiple approaches described in the text. All cases identify two major subpopulations. (**B**) Pairwise Adjusted Mutual Information between all the approaches described in the main text. High AMI values indicate that all the approaches are consistent in identifying cell propertis. (**C**) Pairwise Jaccard's coefficients between marker peaks identified in each analysis. All approaches are able to identify a similar set of regions marking cell groups, with the exception of *bwa SE + MACS* which relies on a larger set of spurious regions.

a similar configuration, again with the notable exception of *bwa SE + MACS* (Figure 1C). This last approach is possibly biased by spurious ATAC peaks identified when only single reads are used: in this case MACS2 identified 17,125 peaks (average score 46.079), while in paired end configuration it identified 5,120 peaks (average score 65.919). Peaks shared by both the analyses have high quality (average score 86.104) while peaks specific of peaks identified after *bwa SE + MACS* are indeed low quality (average score 31.039). These findings indicate that single read mode is not suitable for *de novo* identification of ATAC peaks.

In all, analysis on less heterogeneous data confirm the suitability of *kallisto*-based and, more in general, reference-based approaches for the analysis of scATAC-seq experiments.

Computational resources

One of the most obvious advantages in using *kallisto* in place of alignment-based methods is the reduction of resources required to process raw sequences into a count matrix. We compared runtimes of the various approaches used through this work. First, we compared *cellranger-atac* pipeline and *kallisto* on a machine equipped with 12 CPU (Intel X5650@2.67GHz) and 72 Gb RAM using the PBMC dataset. While it took 46:49:57 hours for *cellranger-atac* to complete the analysis, its total runtime includes several post-processing and analysis steps. To make a fair comparison, we focused on pipeline steps that are mirrored in both the approaches (alignment, barcode assignment and counting) and the steps that are prerequisites to them (Figure 6A). To this end, we also considered in the *kallisto*

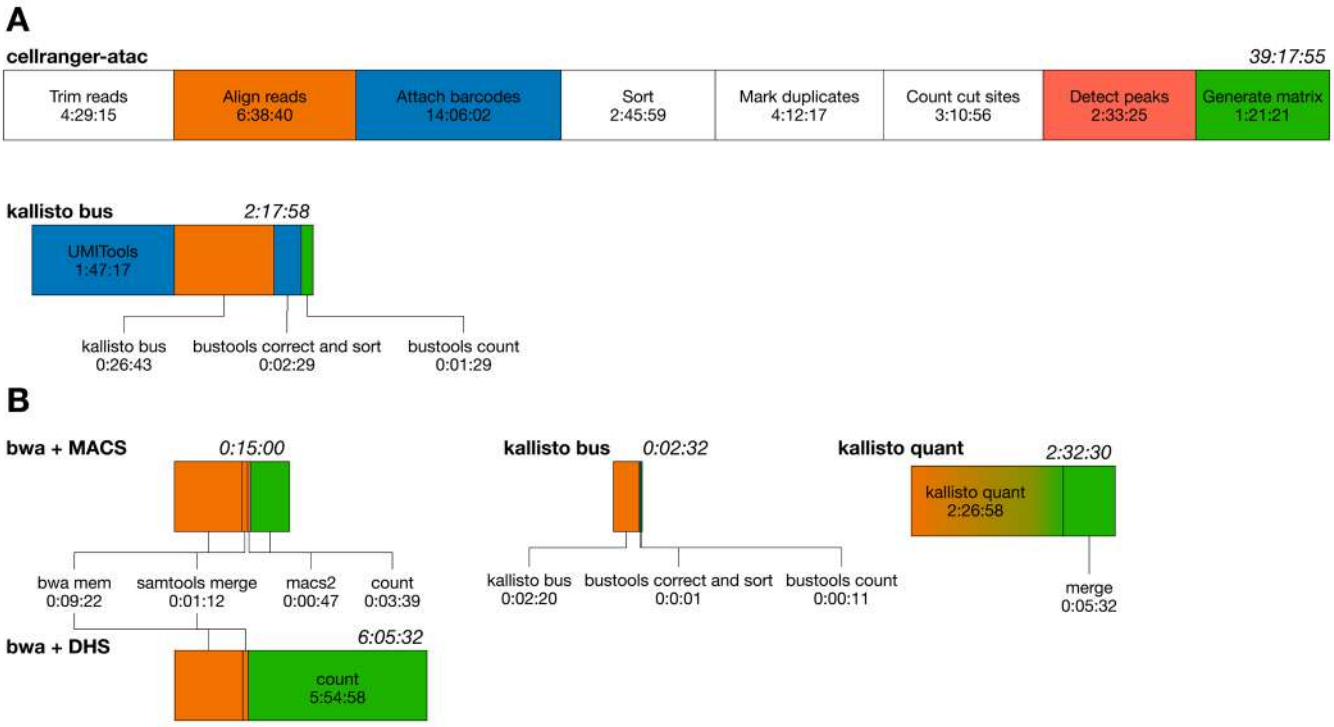


Figure 6. Runtime analysis. Graphical representation of runtimes for the datasets processed in this paper. Each box represents a separate step in a pipeline, box size is proportional to runtime in logarithmic scale. Colors in each box maps logically equivalent steps mirrored in different pipelines. (**A**) Runtimes of *cellranger-atac* and *kallisto bus* on the PBMC 10k dataset. White boxes indicate steps that are not mirrored in both the analysis. (**B**) Runtimes of all the approaches used in the analysis of K562 data. The gradient in *kallisto quant* indicates a hybrid step, which performs mapping and quantification. *bwa SE* pipelines have been excluded from the chart.

runtime an external application to identify valid cellular barcodes (*UMITools*). This step can be replaced by any tool capable to return a list of valid cellular barcodes. The total effective time of *kallisto* is approximately 17x shorter, also because many processing steps are not required (initial trimming and BAM processing) or missing by design (peak calling). Our results are consistent with previous estimates on scRNA-seq data[27]. In addition to reduced runtimes and pipeline simplicity, usage of *kallisto* implies reduced disk usage (12 Gb vs 40 Gb).

Analysis of the K562 datasets show reduced runtimes due to the smaller number of cells and sequences. Comparisons have been performed on the same 12 CPU platform, running 3 cells in parallel, 4 threads each, for *kallisto quant* and *bwa*-based pipelines. Coherently with the PBMC dataset, *kallisto bus* analysis is approximately 7x shorter than the default approach (Figure 6B). Note, however, that raw sequences are generated for separate cells: alignment could be performed on as many computing units as the number of cells themselves. As an example, one could run 288 parallel alignments, reducing the total alignment step by a factor 96x (5.8s), assuming no impact on the I/O subsystem. The quantification step of *bwa*-based approach is impacted by the size of the peak list, which was three orders of magnitude smaller for *bwa PE + MACS* (5,120). A special case is the *kallisto quant* approach: we found the pseudoalignment step being much slower than the *bwa* counterpart. By looking at execution logs, we noticed that *kallisto* spends a large time in loading the reference in memory, this is repeated for each cell separately. *kallisto bus* loads the reference one time only, with beneficial impact on speed. As for disk usage, *kallisto bus* requires less space than *bwa PE + MACS* (393 Mb vs 1.2 Gb), while *kallisto quant* needs considerably more space (14 Gb), due to the 'abundance.tsv' text files produced by default during processing.

Lastly, it should be noticed that *kallisto* memory requirements in building the index are proportional to the number of *k*-mers found. The DHS500 database is composed by 682,100,489 *k*-mers and RAM allocation peaks at 37 Gb during indexing. The process itself takes 37.5 hours to complete.

Discussion/conclusions

Analysis of differential chromatin properties, through ATAC-seq and other quantitative approaches, relies on the identification of peaks or enriched regions, It is often achieved with the same statistical framework used in analysis of differential gene expression[28,29]. Identification of peaks is a key difference between the two approaches. *De novo* discovery of unannotated transcripts has been shown to be possible in early times of NGS[30], but the large majority of analysis is performed on gene models. Conversely, analysis of epigenomes involves identification of regions of interest, although a large catalogues

of such regions have been provided by several projects, such as the ENCODE project[31], the BluePrint project[32] or the GeneHancer database[33]. In single cell analysis, for both scRNA-seq and scATAC-seq, identification of novel features may be an issue, especially because of the low coverage at which single cells are profiled. To our knowledge, this work is the first to test the feasibility of a reference-based approach to ATAC-seq analysis, with a special focus on single cell ATAC-seq. In combination, we tested the suitability of *kallisto* to quantify scATAC-seq, which maximizes the performances of the whole process. Our results suggest that identification of cell groups using a reference-based approach is not different from a standard pipeline. Not only cells could be classified in a nearly identical way, but also differential features are largely matched between the analysis. The most obvious advantage is the gain in speed and efficiency: once reads have been demultiplexed, *kallisto* analysis requires short execution times, in the order of minutes, with limited hardware resources. This advantage has been known for a while and, in fact, it has been demonstrated that it can be used on Rock64 hardware[34]. We also anticipate that adoption of a reference-based strategy comes with additional advantages: in particular, functional annotations and gene associations are available for known regulatory regions[25] and, more recently, for DNase I Hypersensitive Sites[15]. In the analysis of K562 cells, we highlighted a degradation of performances when a spurious region list is used, in our case peaks identified by MACS using single reads only. While best practices for ATAC-seq analysis are available[35], adoption of a reference-based approach could improve stability of results and their reproducibility.

Of course, our strategy has limitations that come from the unavailability of read positioning on the genome: in addition to the impossibility of identifying novel peaks, it is not possible to perform some ATAC-specific analysis, such as nucleosome positioning or footprinting of transcription factors in accessible regions. Indeed, these two can be overcome if standard alignment is used in place of pseudoalignment. Another limitation is the large amount of memory needed to index the DHS reference. Although indexing cannot be performed on less performing hardware, prebuilt indexes can be distributed as it is currently done for many aligners. As concluding remark we would like to underline that, although we showed that *kallisto* can be effectively used for analysis of scATAC-seq data, we are aware that it has not been conceived for that purposes; its interface needs some tweaks to work. For this reason, we advocate the development of tools which support scATAC-seq natively and other tools for postprocessing and data visualization.

Data availability

Source data

Single cell ATAC-seq data for 10k PBMCs dataset were downloaded from the 10x Genomics public datasets (https://support.10xgenomics.com/single-cell-atac/datasets/1.1.0/atac_v1_pbmc_10k). Access to the data is free but requires registration. Raw sequences for K562 cells were downloaded from the Gene Expression Omnibus under the accession ID GSE112200 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE112200).

Extended data

Zenodo: vgiansanti/Kallisto-scATAC v1.1. https://doi.org/10.5281/zenodo.3834767[36].

This project contains a detailed explanation of the procedures described in this work and the list of DHS sites; this is also available at https://github.com/vgiansanti/Kallisto-scATAC.

Extended data are available under the terms of the Creative Commons Attribution 4.0 International license (CC- BY 4.0).

Acknowledgements

The authors want to thank the people and supervisors who supported their work, in particular Giovanni Tonon, Catherine Dulac and Tim Sackton.

References

1. Svensson V, Vento-Tormo R, Teichmann SA: **Exponential scaling of single-cell RNA-seq in the past decade.** *Nat Protoc.* 2018; **13**(4): 599–604.
PubMed Abstract | Publisher Full Text

2. Wolf FA, Angerer P, Theis FJ: **SCANPY: large-scale single-cell gene expression data analysis.** *Genome Biol.* 2018; **19**(1): 15.
PubMed Abstract | Publisher Full Text | Free Full Text

3. Dobin A, Davis CA, Schlesinger F, *et al.*: **STAR: ultrafast universal RNA-seq aligner.** *Bioinformatics.* 2013; **29**(1): 15–21.
PubMed Abstract | Publisher Full Text | Free Full Text

4. Zielezinski A, Vinga S, Almeida J, *et al.*: **Alignment-free sequence comparison: benefits, applications, and tools.** *Genome Biol.* 2017; **18**(1): 186.
PubMed Abstract | Publisher Full Text | Free Full Text

5. Van den Berge K, Hembach KM, Soneson C, *et al.*: **RNA sequencing data: hitchhiker's guide to expression analysis.** *Annu Rev Biomed Data Sci.* 2019; **2**(1): 139–173.
Publisher Full Text

6. Conesa A, Madrigal P, Tarazona S, *et al.*: **A survey of best practices for RNA-seq data analysis.** *Genome Biol.* 2016; **17**(1): 13.
PubMed Abstract | Publisher Full Text | Free Full Text

7. Harrow J, Frankish A, Gonzalez JM, *et al.*: **GENCODE: the reference human genome annotation for The ENCODE Project.** *Genome Res.* 2012; **22**(9): 1760–1774.
PubMed Abstract | Publisher Full Text | Free Full Text

8. Bray NL, Pimentel H, Melsted P, *et al.*: **Near-optimal probabilistic RNA-seq quantification.** *Nat Biotechnol.* 2016; **34**(5): 525–527.
PubMed Abstract | Publisher Full Text

9. Patro R, Duggal G, Love MI, *et al.*: **Salmon provides fast and bias-aware quantification of transcript expression.** *Nat Methods.* 2017; **14**(4): 417–419.
PubMed Abstract | Publisher Full Text | Free Full Text

10. Melsted P, Ntranos V, Pachter L: **The barcode, UMI, set format and BUStools.** *Bioinformatics.* 2019; **35**(21): 4472–4473.
PubMed Abstract | Publisher Full Text

11. Srivastava A, Malik L, Smith T, *et al.*: **Alevin efficiently estimates accurate gene abundances from dscRNA-seq data.** *Genome Biol.* 2019; **20**(1): 65.
PubMed Abstract | Publisher Full Text | Free Full Text

12. Zhang Y, Liu T, Meyer CA, *et al.*: **Model-based analysis of ChIP-Seq (MACS).** *Genome Biol.* 2008; **9**(9): R137.
PubMed Abstract | Publisher Full Text | Free Full Text

13. Buenrostro JD, Giresi PG, Zaba LC, *et al.*: **Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position.** *Nat Methods.* 2013; **10**(12): 1213–1218.
PubMed Abstract | Publisher Full Text | Free Full Text

14. Thurman RE, Rynes E, Humbert R, *et al.*: **The accessible chromatin landscape of the human genome.** *Nature.* 2012; **489**(7414): 75–82.
PubMed Abstract | Publisher Full Text | Free Full Text

15. Meuleman W, Muratov A, Rynes E, *et al.*: **Index and biological spectrum of accessible dna elements in the human genome.** *bioRxiv.* 2019.
Publisher Full Text

16. Sheffield NC, Thurman RE, Song L, *et al.*: **Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions.** *Genome Res.* 2013; **23**(5): 777–788.
PubMed Abstract | Publisher Full Text | Free Full Text

17. Quinlan AR: **BEDTools: The Swiss-Army Tool for Genome Feature Analysis.** *Curr Protoc Bioinformatics.* 2014; **47**: 11.12.1–34.
PubMed Abstract | Publisher Full Text | Free Full Text

18. Li H: **Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.** arXiv.org. 2013.
Reference Source

19. Faust GG, Hall IM: ***SAMBLASTER*: fast duplicate marking and structural variant read extraction.** *Bioinformatics.* 2014; **30**(17): 2503-2505.
PubMed Abstract | Publisher Full Text | Free Full Text

20. Soneson C, Love MI, Robinson MD: **Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. [version 2; peer review: 2 approved].** *F1000Res.* 2015; **4**: 1521.
PubMed Abstract | Publisher Full Text | Free Full Text

21. Traag VA, Waltman L, van Eck NJ: **From Louvain to Leiden: guaranteeing well-connected communities.** *Sci Rep.* 2019; **9**(1): 5233.
PubMed Abstract | Publisher Full Text | Free Full Text

22. Smith T, Heger A, Sudbery I: **UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy.** *Genome Res.* 2017; **27**(3): 491-499.
PubMed Abstract | Publisher Full Text | Free Full Text

23. Stuart T, Butler A, Hoffman P, *et al.*: **Comprehensive Integration of Single-Cell Data.** *Cell.* 2019; **177**(7): 1888–1902.e21.
PubMed Abstract | Publisher Full Text | Free Full Text

24. Tang M: **crazyhottommy/scclusteval: second release for citing.** *Zenodo.* 2020.
Publisher Full Text

25. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, *et al.*: **Integrative analysis of 111 reference human epigenomes.** *Nature.* 2015; **518**(7539): 317–330.
PubMed Abstract | Publisher Full Text | Free Full Text

26. Chen X, Litzenburger UM, Wei Y, *et al.*: **Joint single-cell DNA accessibility and protein epitope profiling reveals environmental regulation of epigenomic heterogeneity.** *Nat Commun.* 2018; **9**(1): 4590.
PubMed Abstract | Publisher Full Text | Free Full Text

27. Melsted P, Booeshaghi AS, Gao F, *et al.*: **Modular and efficient pre-processing of single-cell RNA-seq.** *BioRxiv.* 2019.
Publisher Full Text

28. Anders S, Huber W: **Differential expression analysis for sequence count data.** *Genome Biol.* 2010; **11**(10): R106.
PubMed Abstract | Publisher Full Text | Free Full Text

29. Yan F, Powell DR, Curtis DJ, *et al.*: **From reads to insight: a hitchhiker's guide to ATAC-seq data analysis.** *Genome Biol.* 2020; **21**(1): 22.
PubMed Abstract | Publisher Full Text | Free Full Text

30. Robertson G, Schein J, Chiu R, *et al.*: ***De novo* assembly and analysis of RNA-seq data.** *Nat Methods.* 2010; **7**(11): 909–912.
PubMed Abstract | Publisher Full Text

31. ENCODE Project Consortium: **An integrated encyclopedia of DNA elements in the human genome.** *Nature.* 2012; **489**(7414): 57–74.
PubMed Abstract | Publisher Full Text | Free Full Text

32. Adams D, Altucci L, Antonarakis SE, *et al.*: **BLUEPRINT to decode the epigenetic signature written in blood.** *Nat Biotechnol.* 2012; **30**(3): 224–226.
PubMed Abstract | Publisher Full Text

33. Fishilevich S, Nudel R, Rappaport N, *et al.*: **GeneHancer: genome-wide integration of enhancers and target genes in GeneCards.** *Database (Oxford).* 2017; **2017**: bax028.
PubMed Abstract | Publisher Full Text | Free Full Text

34. Tan QW, Mutwil M: **Inferring biosynthetic and gene regulatory networks from *Artemisia annua* RNA sequencing data on a credit card-sized ARM computer.** *Biochim Biophys Acta Gene Regul Mech.* 2019; **1863**(6): **194429.**
PubMed Abstract | Publisher Full Text

35. Yan F, Powell DR, Curtis DJ, *et al.*: **From reads to insight: a hitchhiker's guide to ATAC-seq data analysis.** *Genome Biol.* 2020; **21**(1): 22.
PubMed Abstract | Publisher Full Text | Free Full Text

36. Giansanti V, Cittaro D: **vgiansanti/kallisto-scatac v1.1.** *Zenodo.* 2020.
http://www.doi.org/10.5281/zenodo.3834767

Open Peer Review

Current Peer Review Status: ✔ ✔

Version 2

Reviewer Report 15 June 2020

https://doi.org/10.5256/f1000research.26547.r64049

© **2020 Zhang Q.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

✔ **Qiangfeng Cliff Zhang**
MOE Key Laboratory of Bioinformatics, Beijing Advanced Innovation Center for Structural Biology, Center for Synthetic and Systems Biology, School of Life Sciences, Tsinghua University, Beijing, China

I am happy that all my concerns are addressed and I like the method.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bioinformatics, genomics, RNA structure, Genome structure, AI algorithms in biomedicine.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 06 May 2020

https://doi.org/10.5256/f1000research.25099.r62150

© **2020 Zhang Q.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

? **Qiangfeng Cliff Zhang**
MOE Key Laboratory of Bioinformatics, Beijing Advanced Innovation Center for Structural Biology, Center for Synthetic and Systems Biology, School of Life Sciences, Tsinghua University, Beijing,

China

Comments:
The work by Giansanti *et al*. presents a novel and smart idea for scATAC-seq data analysis. It demonstrates the possibility of using a reference-based, pseudo-alignment method to reduce the computational requirement for scATAC-seq data analysis, with only a little sacrifice on precision. The idea is inspired by the using of pesudoalignment for bulk and single-cell RNA-seq quantification. Here they showed that with some tweaking of the input sequencing reads, they could use kallisto to analyze scATAC-seq data on a pre-defined set of DNase hypersensitive sites. They compared their results with the standard protocol (e.g. *cellranger-atac*) for peak quantification, single cell clustering, marker peaks identification, and gene activity score calculation.

The results very nicely revealed the consistency on peak quantification between *kallisto*-based method and *cellranger-atac*. The cell clusterings were almost identical between the new reference-based method and canonical mapping strategy. And the gene activity scores by two different methods also agreed well with each other. The approach presented in this study thus could be a very efficient way for scATAC-seq data analysis.

The following are a few comments/questions:

1. The method was only tested with one dataset - PMBC. In fact, single cell ATAC-seq data is usually very sparse. The PMBC dataset used in this study is of relatively high quality. The method remains to be tested on more datasets, especially on those of more sparse, lower-quality.

2. The key advantage of the method is presumably the much improved computational efficiency – there may be other advantages brought by the reference-based method. However, there is no results/statistics on the running time and memory usage in the manuscript. From the description, the improvement should be dramatic. I think it would be very nice to include a table or a figure to demonstrate the increase of computational efficiency. This could be a very helpful way to convince potential users.

3. As in the above, this whole strategy is so different. It is thus possible for the method to be used for some other scATAC-seq data analysis with advantages not only in computational efficiency. It would be good for the authors to explore.

4. The manuscript is well organized with the core ideas clearly described. But the presentation could be improved - there are a lot of very long sentences unnecessarily connected by “and”, “while”, etc.

5. The legend for Fig 1A says “The first (DHS) generated by kallisto on ~2M DNase I sites … ”, but according to the figure and the main text, it should be“~3M”?

Is the rationale for developing the new method (or application) clearly explained?
Yes

Is the description of the method technically sound?

Yes

Are sufficient details provided to allow replication of the method development and its use by others?
Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bioinformatics, genomics, RNA structure, Genome structure, AI algorithms in biomedicine.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 20 May 2020

Davide Cittaro, IRCCS San Raffaele Institute, Milan, Italy

Thank you for reviewing our manuscript and for the helpful comments. We have addressed major and minor points as detailed in the point-by-point response:

The method was only tested with one dataset - PMBC. In fact, single cell ATAC-seq data is usually very sparse. The PMBC dataset used in this study is of relatively high quality. The method remains to be tested on more datasets, especially on those of more sparse, lower-quality.

We agree that the PBMC dataset is of high quality. It was used as it could be considered a *de facto* standard in single cell analysis as it includes several populations at different degrees of separations (*i.e.* B-cells and T-Cells are well separated, while NK and CD8 are less clearly distinguished). We also would like to point out that it is difficult to identify low quality scATAC-seq datasets for two reasons: one is the relative novelty of this technique and the other is the positive bias in publications, which generally lack of negative or low-quality results. Nevertheless, we tried to address this question analyzing data for K562 cell line. Cell lines are supposedly more homogeneous, data were obtained on a low-throughput platform (Fluidigm C1). We believe that it could be considered a good example of “lower quality” dataset, compared to the PBMC, at least considering the information content. We show that our strategy is consistent with standard approaches based on alignment and peak identification, we can identify the same level of residual heterogeneity.

The key advantage of the method is presumably the much improved computational efficiency – there may be other advantages brought by the reference-based method. However, there is no results/statistics on the running time and memory usage in the manuscript. From the description, the improvement should be dramatic. I think it would be very nice to include a table or a figure to demonstrate the increase of computational efficiency. This could be a very helpful way to convince potential users.

Thank you for this comment. We benchmarked *kallisto+bustools* and compared it to *cellranger-atac*, the default application for 10x data. We added a dedicated section in the main text, which shows the large reduction in required resources. Note that the *cellranger-atac* pipeline includes several steps that are common in downstream analysis (such as Seurat or Scanpy). In order to make it fair, as explained in the text, we did not consider these steps in the comparison. In addition, we added runtime analysis for the approaches used in the analysis of K562 data.

As in the above, this whole strategy is so different. It is thus possible for the method to be used for some other scATAC-seq data analysis with advantages not only in computational efficiency. It would be good for the authors to explore.

Our work has been mainly motivated by the reduced resources that are needed by a kallisto-based approach, as we predict the number of scATAC-seq experiments will increase as well as the number of cells profiled. We anticipated additional advantages of a reference-based strategy in the first version of our manuscript, *e.g.* the availability of promoter-enhancer/gene interactions which could be readily applied to scATAC-seq data. During the revision process we had the opportunity to perform the analysis with non-optimal conditions (*i.e.* peak identification from single end reads instead of paired end), which led to slightly different results. This serendipitous finding suggests that our strategy, not relying on *de novo* identification, improves the stability of cell characterization and, therefore, the reproducibility of results. We added these observations in the discussion.
Of course, the usage of standardized reference could pave the way to a new class of processing steps not currently performed. As an example, one could identify a set of regions known to be generally accessible (or not) to perform standardized QC. Another example could be the identification of regions that could be used to score the cell cycle phases in scATAC-seq data, much like what is normally done with scRNA-seq data. We feel that all these examples require a deeper analysis, which is beyond the scope of this work, and any undemonstrated procedure would be, at best, greatly speculative. Our aim was to show general consistence between diverse approaches, which we believe has been demonstrated.

The manuscript is well organized with the core ideas clearly described. But the presentation could be improved - there are a lot of very long sentences unnecessarily connected by "and", "while", etc.

Thank you for this comment, we modified the text to increase readability.

The legend for Fig 1A says "The first (DHS) generated by kallisto on ~2M DNase I sites … ", but according to the figure and the main text, it should be"~3M"?

Thank you for spotting the typo in the figure legend. We corrected accordingly.

Competing Interests: Nothing to disclose

Reviewer Report 30 March 2020

https://doi.org/10.5256/f1000research.25099.r61566

© 2020 Barozzi I. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Iros Barozzi
Department of Surgery and Cancer, Imperial College London, London, UK

In their paper "Fast analysis of scATAC-seq data using a predefined set of genomic regions" Giansanti *et al.* suggest an efficient strategy to analyse scATAC-seq data using *kallisto* and *bustools*.

The paper is clearly written, the proposed strategy is well conceived and tested, and it will be useful for many researchers in the field of regulatory genomics. Clear advantages of this strategy are the reduced requirements in terms of computational resources and shorter execution times, when compared to other pipelines such as *cellranger-atac*. This comes at a cost, most notably the chance of missing signals at regions that are not present in the reference set. Nevertheless, in my opinion evaluations about this being a limitation has to be made on a case-by-case basis, and the authors clearly pointed this out (among other limitations) in the discussion. The authors also provide access to the full code, datasets and documentation to reproduce the analyses.

A wide range of parameters was tested, both in terms of handling and modifying the input sequences to make them suitable for *kallisto*, and in terms of pre- vs post- processing the genomic partition considered for indexing. Combinations that return results that are highly concordant with those obtained with *cellranger-atac* were highlighted. The authors then demonstrated the robustness of the biological inferences made using their strategy by showing a very large overlap with the results achieved by *cellranger-atac* (in terms of different groups of regions marking distinct clusters and clusters annotation based on label transferring from scRNA-seq data).

I am wondering if a natural application of this strategy would simplify the characterization of chromatin state at highly repetitive regions of mammalian genomes (e.g. indexing a database of transposable elements). This task would otherwise be quite difficult to handle explicitly with pipelines such as *cellranger-atac*.

I only have two minor comments:

- Can the authors provide more details about the analysis described in the paragraph "Identification of marker regions"? How were the cell groups defined? How were the top 1,000 peaks for each group selected/identified?

- Fig. 1C: description of the blue curve seems to be missing.

Is the rationale for developing the new method (or application) clearly explained?
Yes

Is the description of the method technically sound?
Yes

Are sufficient details provided to allow replication of the method development and its use by others?
Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Genomics; Transcriptional Regulation; Single-cell Transcriptomics.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 20 May 2020

Davide Cittaro, IRCCS San Raffaele Institute, Milan, Italy

Thank you for reviewing our work and for the comments. We have addressed your minor concerns as follows.

Can the authors provide more details about the analysis described in the paragraph "Identification of marker regions"? How were the cell groups defined? How were the top 1,000 peaks for each group selected/identified?

We apologize for lack of clarity in the manuscript. Cell groups were identified with the Leiden method, while markers were identified with Wilcoxon rank-sum test. The complete list of instructions used in the analysis is part of the repository linked in the main text, nevertheless we modified the text adding these specific details.

Fig. 1C: description of the blue curve seems to be missing.

Thank you for pointing this out. The blue line represented the fit of the DHS data. We

acknowledge colouring scheme was not appropriate and, moreover, the fit DHS500 data was missing. In the revised manuscript we modified the figure accordingly.

Competing Interests: Nothing to disclose

Comments on this article

Version 2

Reader Comment 15 Jun 2020

Charles Warden, City of Hope National Medical Center, Duarte, CA, USA

Hi,

Thank you for posting this article.

However, I think there is still at least 1 typo:

"Anal**sy**is of single-cell data" --> "Anal*y*sis of single-cell data"

Hopefully, I believe that the F1000 system provides a good way to revise the paper without a formal correction.

Best Wishes,
Charles

Competing Interests: I do not disclose any conflicts of interest.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research

ANNEX III: NESTED STOCHASTIC BLOCK MODELS APPLIED TO THE ANALYSIS OF SINGLE CELL DATA

Morelli *et al. BMC Bioinformatics* (2021) 22:576
https://doi.org/10.1186/s12859-021-04489-7

BMC Bioinformatics

RESEARCH Open Access

Nested Stochastic Block Models applied to the analysis of single cell data

Leonardo Morelli[1,2], Valentina Giansanti[1,3] and Davide Cittaro[1*]

*Correspondence:
cittaro.davide@hsr.it
[1] Center for Omics Sciences, IRCCS San Raffaele Institute, Milan, Italy
Full list of author information is available at the end of the article

Abstract

Single cell profiling has been proven to be a powerful tool in molecular biology to understand the complex behaviours of heterogeneous system. The definition of the properties of single cells is the primary endpoint of such analysis, cells are typically clustered to underpin the common determinants that can be used to describe functional properties of the cell mixture under investigation. Several approaches have been proposed to identify cell clusters; while this is matter of active research, one popular approach is based on community detection in neighbourhood graphs by optimisation of modularity. In this paper we propose an alternative and principled solution to this problem, based on Stochastic Block Models. We show that such approach not only is suitable for identification of cell groups, it also provides a solid framework to perform other relevant tasks in single cell analysis, such as label transfer. To encourage the use of Stochastic Block Models, we developed a python library, `schist`, that is compatible with the popular `scanpy` framework.

Background

Transcriptome analysis at single cell level by RNA sequencing (scRNA-seq) is a technology growing in popularity and applications [1]. It has been applied to study the biology of complex tissues [2, 3], tumor dynamics [4–7], development [8, 9] and to describe whole organisms [10, 11].

A key step in the analysis of scRNA-seq data and, more in general, of single cell data, is the identification of cell populations, that is groups of cells sharing similar properties. Several approaches have been proposed to achieve this task, based on well established clustering techniques [12, 13], consensus clustering [14–16] and deep learning [17]; many more have been recently reviewed [18, 19] and benchmarked [20]. As the popularity of single cell analysis frameworks `Seurat` [21] and `scanpy` [22] raised, methods based instead on graph partitioning became the *de facto* standards. Such methods require the construction of a cell neighbourhood graph (e.g. by k Nearest Neighbours, kNN, or shared Nearest Neighbours, sNN). Encoding cell-to-cell similarities into graphs has practical advantages beyond clustering, as many algorithms for graph analysis can be applied and interpreted in a biological way. A notable example is the analysis of cell trajectories which can be derived from the analysis of Markov processes traversing the

© The Author(s), 2021. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

NN graph [23, 24]. In another context, computation of RNA moments in scRNA velocity is also based on the NN graph structure [25]. Arguably, the biggest utility of NN structure is the possibility to identify cell groups by partitioning the graph into communities; this is typically achieved using the Louvain method [26], a fast algorithm for optimisation of graph modularity. While fast, this method does not guarantee the identification of internally connected communities. To overcome its limits, a more recent approach, the Leiden algorithm [27], has been implemented and it has been quickly adopted in the analysis of single cell data, for example by `scanpy` [22] and `PhenoGraph` [28]. In addition to Newman's modularity [29], other definitions currently used in single cell analysis make use of a resolution parameter [30, 31]. In lay terms, resolution works as a threshold on the density within communities: lowering the resolution results in less and sparser communities and *vice versa*. Identification of an appropriate resolution has been recognised as a major issue [32], also because it requires the definition of a mathematical property (clusters) over biological entities (the cell groups), with little formal description of the latter. In addition, the larger the dataset, the harder is to identify small cell groups, as a consequence of the well-known resolution limit [33]. Moreover, it has been demonstrated that random networks can have modularity [34] and its optimisation is incapable of separating actual structure from those arising simply of statistical fluctuations of the null model. Lastly, it is a common error to assume that the resolution parameter reflects a hierarchical structure of the communities in the graph when, in general, this is not rigorously true. Additional solutions to cell group identification from NN graphs have been proposed, introducing resampling techniques [35, 36] or clique analysis [37]. It has been proposed that high resolution clustering, e.g. obtained with Leiden or Louvain methods, can be refined in agglomerative way using machine learning techniques [38].

An alternative solution to community detection is the Stochastic Block Model, a generative model for graphs organised into communities [39]. In this scenario, identification of cell groups requires the estimation of the proper parameters underlying the observed NN graph. According to the microcanonical formulation [40], the parameters are partitions and the matrix of edge counts between them. Under this model, nodes belonging to the same group have the same probability to be connected together. It is possible to include node degree among the model parameters [41], to account for heterogeneity of degree distribution of real-world graphs. A Bayesian approach to infer parameters has been developed [42] and implemented in the `graph-tool` python library (https://graph-tool.skewed.de). There, a generative model of network $\boldsymbol{A}$ has a probability $P(\boldsymbol{A}|\boldsymbol{\theta}, \boldsymbol{b})$ where $\boldsymbol{\theta}$ is the set of parameters and $\boldsymbol{b}$ is the set of partitions. The likelihood of the network being generated by a given partition can be measured by the posterior probability

$$P(\boldsymbol{b}|\boldsymbol{A}) = \frac{P(\boldsymbol{A}|\boldsymbol{\theta}, \boldsymbol{b})P(\boldsymbol{\theta}, \boldsymbol{b})}{P(\boldsymbol{A})} \tag{1}$$

and inference is performed by maximising the posterior probability. The numerator in Eq. 1 can be rewritten exponentiating the description length

$$\Sigma = -\ln P(\boldsymbol{A}|\boldsymbol{\theta}, \boldsymbol{b}) - \ln P(\boldsymbol{\theta}, \boldsymbol{b}) \tag{2}$$

so that inference is performed by minimising the information required to describe the data (Occam's razor); `graph-tool` is able to efficiently do this by a Markov Chain Monte Carlo approach [43]. SBM itself may fail to identify small groups in large graphs, hence hierarchical formulation has been proposed [44]. Under this model, communities are agglomerated at a higher level in a block multigraph, also modelled using SBM. This process is repeated recursively until a graph with a single block is reached, creating a nested Stochastic Block Model (nSBM).

In this work we propose nSBM for the analysis of single cell data, in particular scRNA-seq data. This approach identifies cell groups in a statistical robust way and, moreover, it is able to determine the likelihood of the grouping, thus allowing model selection. In addition, it is possible to measure the confidence of assignment to groups, a measure that can be exploited in various analysis tasks.

We developed `schist` (https://github.com/dawe/schist), a python library compatible with `scanpy`, to facilitate the adoption of Stochastic Block Models in single-cell analysis.

Results

Overview of `schist`

`schist` is a convenient wrapper to the `graph-tool` python library, written in python and designed to be used with `scanpy`. The most prominent function is `schist.inference.nested_model()` which takes a `AnnData` object as input and fits a nested Stochastic Block Model on the *k*NN graph built with `scanpy` functions (e.g. `scanpy.tools.neighbors()`). When launched with default parameters, `schist` fits a model which maximises the posterior probability of having a set of cell groups (or blocks) given a graph. `schist` then annotates cells in the data object with all the groups found at each level of a hierarchy. Given the large size of the NN graph in real-world experiments, it is possible that a single solution represents local minima of the fitting process. In addition, it is possible that multiple solutions are equally acceptable to represent the graph partitioning and a better description is given by the consensus over such solutions [45]. To overcome these issues, `schist` fits multiple instances in parallel and returns the inferred consensus model, alongside the marginal probabilities for each cell to belong to a specific group (*cell marginals*). Moreover, the Stochastic Block Model has no constraints on what type of modular structure is fitted, meaning that groups are not necessarily identified only by assortativity (i.e. cells are mostly connected within the same group). When assortativity is thought to be the dominant pattern another model (the Planted Partition Block Model, PPBM [46]), also implemented in `schist`, is better suited to find statistically significant assortative communities, also eliminating the need to set a resolution parameter as required in standard community detection by maximisation of modularity.

Analysis of the impact of noise

One of the most relevant difference between the SBM and other methods to cluster single cells is that it relies on robust statistical modelling. In this sense, the number of groups identified strictly mirrors the amount of information contained in the data. An important consequence is that absence of information (i.e. maximal entropy) can be

properly handled. To show this property we performed a simple experiment on a randomised *k*NN graph. We collected data for 3k PBMC (available as preprocessed data in `scanpy`, Fig. 1A) and shuffled the edges of the prebuilt *k*NN graph, this to keep the general graph properties unchanged. We tested that the degree distribution does not change after randomisation (Kolmogorov-Smirnov $D = 0.0733$, $p = 0.703$). We found that the default strategy, based on maximisation of modularity, identifies 24 cell groups at default resolution, whereas `schist` does not identify any cell group, at level 0 (Fig. 1B).

Only by reducing resolution to $\gamma < 0.6$ we were able to obtain a single partition by modularity (Additional file 1: Fig. S1). Of course, this experiment is a deliberate extreme case. The quality of grouping proposed by a standard approach can be disputed in many ways, and the UMAP embedding indeed reflects the absence of any information. Nevertheless, real-world data may include an unknown amount of random noise. Hence, it is important to identify cell groups that are not artefacts arising from processing and that do reflect the information contained in the dataset. To understand the impact of

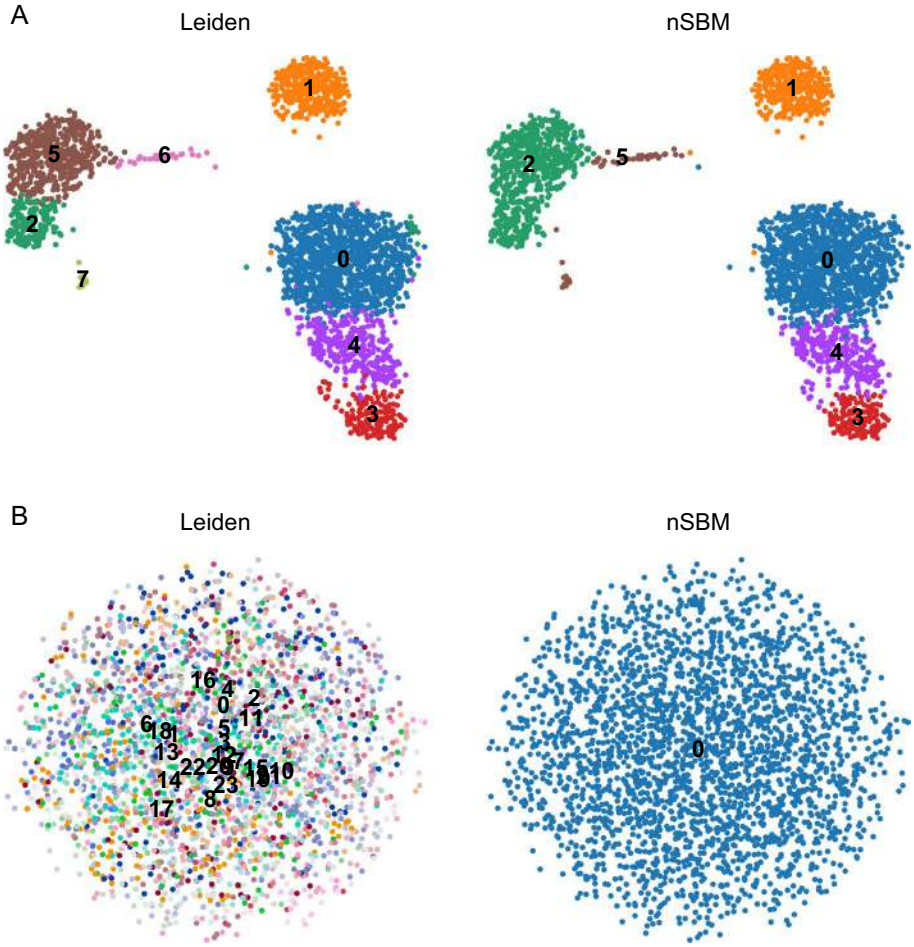


Fig. 1 Modeling randomised data. **A** UMAP embedding of 3k PBMC after standard processing with Leiden approach (left) or nSBM (right). Cell grouping is consistent for both the approaches (Adjusted Rand Index *ARI* = 0.869). **B** UMAP embedding of the same data after randomisation of *k*NN graph edges. In this case `schist` does not return any cell grouping, while optimisation of modularity finds up to 24 different cell groups

random noise on structured data, we considered the same PBMC dataset and added white noise to the normalised counts at increasing levels of σ, ensuring that the noise level is modelled after the feature-wise distribution of detected genes. We then compared partitions to the original annotation by Adjusted Rand Index (*ARI*), we found that `schist` is more robust to perturbations and that, again, optimising the modularity results in overestimation of the number of communities at high noise levels (Additional file 1: Table S1). Of note, the concordance with original annotations drops at $\sigma \geq 1.5$ for both the approaches.

`schist` correctly identifies cell populations

To benchmark `schist`, we tested it on scRNA-seq mixology data [47], a dataset explicitly developed to benchmark single cell analysis tools without the need to simulate data. In particular, we used the mixture of 5 cell lines profiled with Chromium 10x platform. At a first evaluation of the UMAP embedding, all lines appear well separated. Only the lung cancer line H1975 shows a considerable degree of heterogeneity and appears to be split into two cell groups (Fig. 2A). Using default parameters, `schist` is able to identify correct cell groups ($ARI = 0.829$), with a further split in H2228 cell line (Fig. 2B), whereas Leiden method clusters the dataset into 10 groups ($ARI = 0.549$, Fig. 2C). `schist` correctly identifies H1975 groups as a single entity at level 1 of the nSBM hierarchy. We then sought to check if an independent agglomerative method, SCCAF [38], was able to recover cell line groupings starting from both partition schemes. Given the ground truth, the cell lines, SCCAF is able to assess the maximal accuracy that can be achieved in the dataset (0.992). When trained with this target accuracy, SCCAF precisely reconstructs the original cell line annotations starting from `schist` partitions with high accuracy (Fig. 2D). When Leiden partitions are set as input, SCCAF merges H2228 and

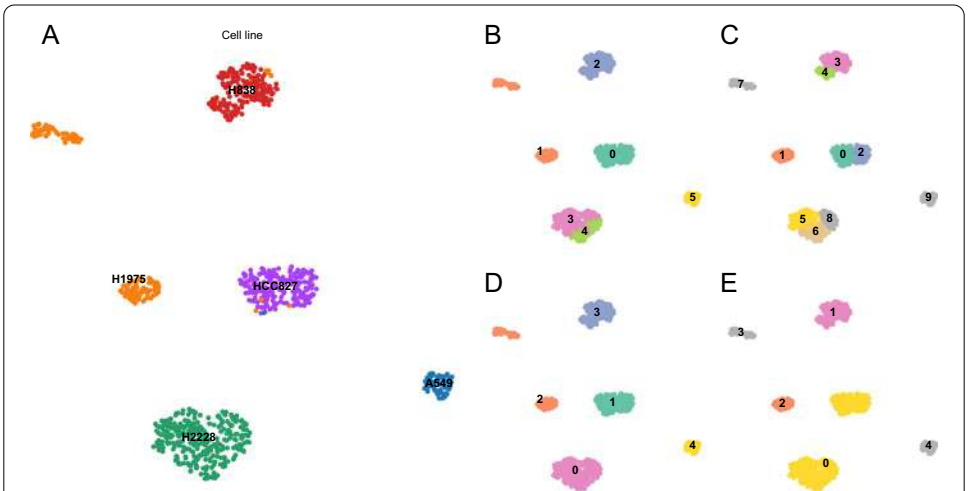


Fig. 2 Analysis of a mixture of 5 known cell lines. **A** UMAP embedding coloured and labeled by cell line identity. H1975 cells (orange) show two distinct groups reflecting an internal heterogeneity. **B** UMAP embedding showing cells coloured by level 1 of the hierarchy proposed by the nested Stochastic Block Model. **C** UMAP embedding showing cells coloured according to the Leiden method at resolution $\gamma = 1$. **D** UMAP embedding coloured by the classification made by SCCAF when partitions in (**B**) are used. **E** UMAP embedding coloured by the classification made by SCCAF when partitions in (**C**) are used

HCC827 cells into a single cluster and keeps H1975 cells split into two groups (Fig. 2E), highlighting potential limitations of this approach.

In another experiment, we analysed data from the Tabula Muris project [48] mixing four different tissues as previously performed [49] (i.e. skin, spleen, large intestine and brain, Additional file 1: Fig. S2A). In this experiment we expect higher heterogeneity than controlled cell lines, however `schist` is able to correctly identify the original tissues (Additional file 1: Fig. S2C), which are again almost perfectly classified after SCCAF is applied (Additional file 1: Fig. S2D). Similarly to the cell line experiment, optimisation of modularity isolates cell clumps evident in UMAP embedding (Additional file 1: Fig. S2E) which could not be correctly merged after SCCAF iteration (Additional file 1: Fig.S2F). In all, these data support the suitability of `schist`, hence of nested Stochastic Block Models, for cell group identification in single cell studies.

Hierarchy modelling complies with biological properties

When grouping is performed by optimisation of modularity, there is often the implicit assumption that the resolution parameter reflects a hierarchical structure of the graph, i.e. communities are consistently grouped at lower resolutions. Not only this assumption is wrong, but it may also lead to spurious groupings in real experiments, whereas a nSBM inherently encodes hierarchies by merging communities in a tree. The improper use of resolution parameter may lead to two types of errors: grouping of cells that are in fact distinct and creating an inconsistent hierarchy.

To show this we took advantage of public spatial RNA dataset of a coronal section of murine brain tissue profiled with 10X Visium H&E technology [50], as provided by the recently introduced package `SquidPy` [51]. We chose to stick to the given tissue annotation by the package authors. At default resolution, Leiden clustering resolves the tissue structure, as does the first level of the nSBM hierarchy (Fig. 3). When resolution is decreased (e.g. $\gamma = 0.5$), the dentate gyrus is incorrectly merged to the hippocampus, whereas `schist` correctly identifies the pyramidal layer.

In another context, we tested the effect on the interpretation of the hierarchy varying the resolution parameter. We analysed data for hematopoietic differentiation [52], previously used to benchmark the consistency of cell grouping with differentiation trajectories by graph abstraction [53] (Additional file 1: Fig. S3A). Data show three major branchings (Erythroids, Neutrophils and Monocytes) stemming from the progenitor cells, mostly recapitulated by level 2 of the hierarchy computed by `schist` (Fig. 4). Not only the hierarchic model recapitulates the branching trajectories, also the cell groups appear to be consistent with the estimated pseudotime (Additional file 1: Fig. S3B). Conversely, the Leiden method at default resolution identified 24 groups. By lowering the γ parameter we observed cell groups that merge and split at different resolutions disrupting the hierarchy (Additional file 1: Fig. S4).

In all, these data show that the common intuition that γ parameter acts as a thresholding factor over a hierarchy is wrong. Not only the hierarchy is not conserved, but also very different cell types may be mixed in spurious clusters. By using nSBM, `schist` is able to represent hierarchical relations in appropriate way. Moreover, the hierarchy appears to be more robust in aggregating different cell types at coarser scales.

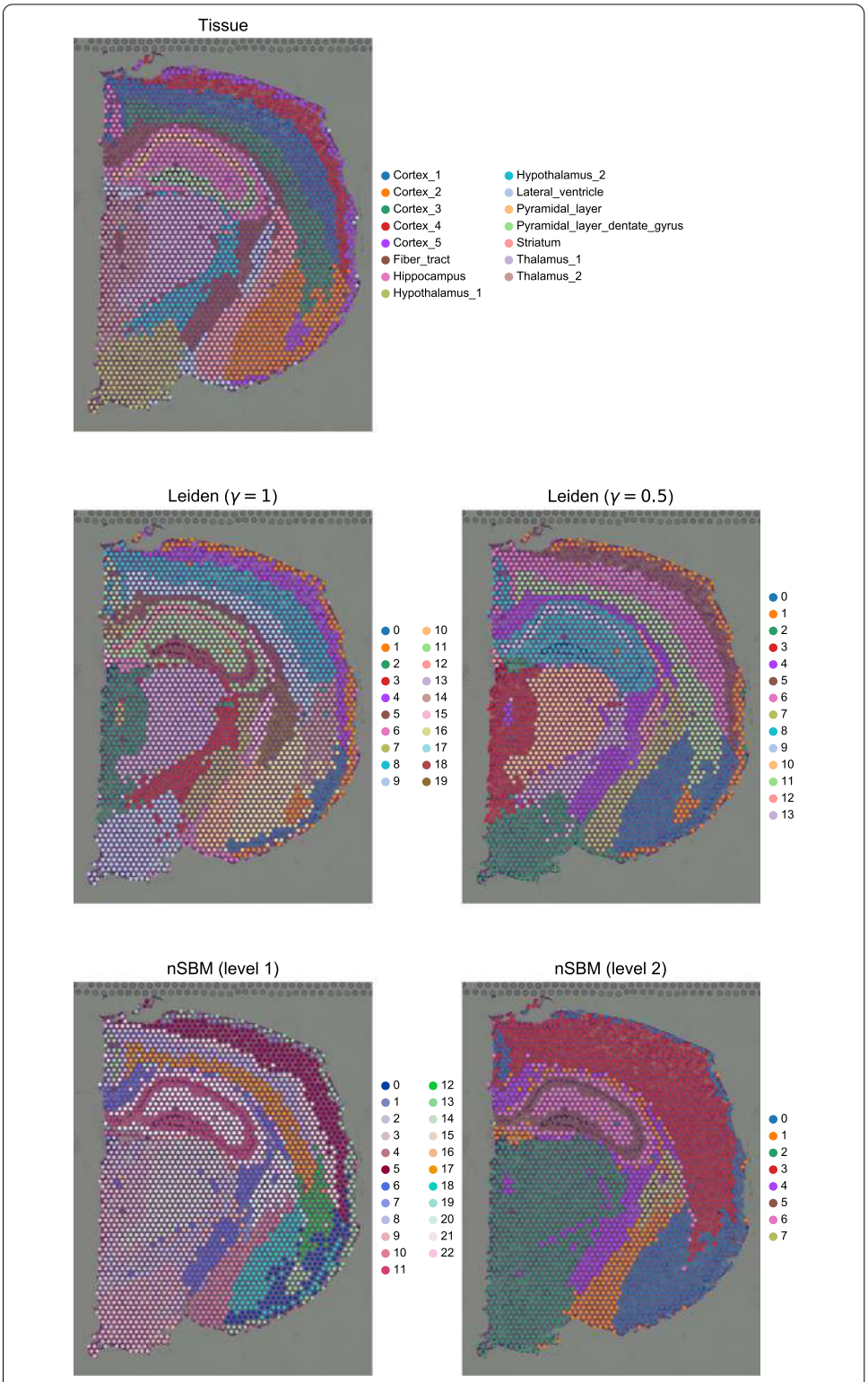


Fig. 3 Analysis of spatial transcriptomics of a coronal section of mouse brain by Visium H&E. In the first panel, original tissue annotation is given. Tissues are well defined at default resolution for the standard approach. When resolution is decreased to $\gamma = 0.5$, cells are aggregated breaking the histological types, e.g. cells from the dentate gyrus are merged to the hippocampus. When a nSBM is applied, the structure of the pyramidal layer is maintained at different levels of the hierarchy

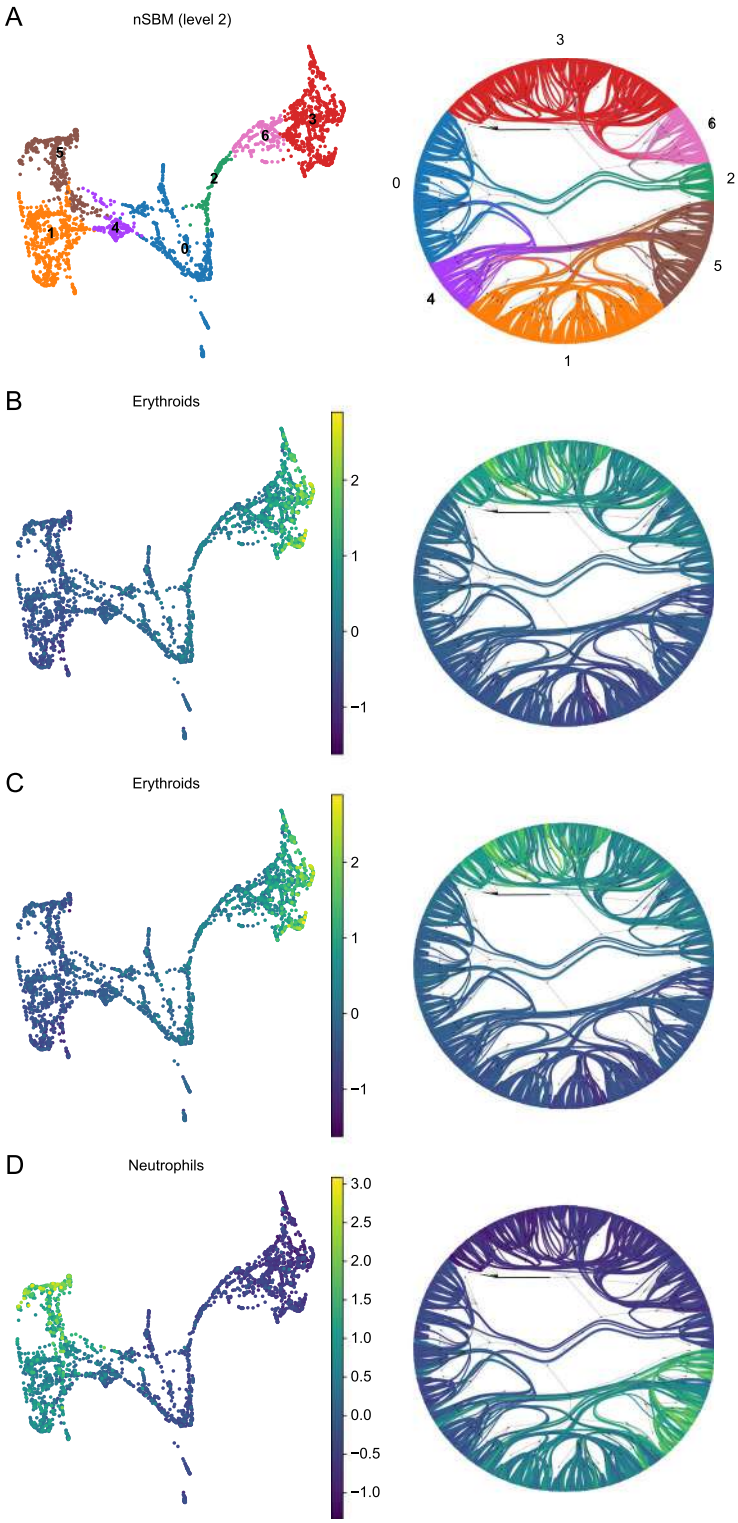


Fig. 4 Analysis of hematopoietic differentiation. Each panel presents a low dimensional embedding of single cells next to a radial tree representation of the nSBM hierarchy. Cells are coloured according to groupings at level 2 of the hierarchy, group 0 marks the most primitive population (**A**). In subsequent panels, cells are coloured using a signature of Erythroid lineage (**B**), Monocytes (**C**) or Neutrophils (**D**)

Cell marginals can be used to assess the data quality

By computing the consensus among multiple models, `schist` returns the marginal probability for each cell to belong to a specific cluster at each level of the hierarchy. Ideally, all cells should always be assigned with $p = 1$ to a cluster. When the uncertainty is maximal, cells are assigned to clusters randomly with $p = 1/B_i$, where B_i is the number of groups for the i-th level in the hierarchy. We sought to check if these probabilities could be interpreted in terms of data quality.

We devised a simple metric, *cell stability*, that is defined by the fraction of levels for which the marginal probability is higher than $1 - 1/B_i$. To do so, we only consider levels with at least two groups, hence excluding the root of the tree. We tested this metric on four datasets from [54] with different quality levels (iCELL8, MARS-seq, 10XV3 and Quartsz-seq2) (Additional file 1: Fig. S5). By taking a summary metric, e.g. the mean $\overline{S}$ or the fraction of cells with $S > 0.5$, we observed that it correlates with the data quality (Table 1).

These data suggest that measures of uncertainty of cell clustering can be useful for general quality control assessment. In addition to this, we foresee they could be used to isolate cells with specific patterns.

Cell affinities can be used for label transfer

The modelling approach we adopted allows the estimation of the information required to describe a graph given any partitioning scheme, not limited to the solution given by the model itself. Differences in entropy can be used to perform model selection, hence we can choose which model better describes the data. We sought to exploit this property to address the task of annotating cells according to a reference sample. To this end we analysed datasets from [54], which includes mixtures of human PBMC and HEK293T cells profiled with various technologies. We chose cells profiled with 10X V3 platform as reference dataset and performed annotation on cells profiled with Quartz-seq2 or MARS-seq. These are at the extremes of the capability to distinguish cell types, so they provide good benchmark configurations for this task.

After preprocessing raw data according to the parameters given in [54], we integrated each dataset with 10XV3 into a unified representation using Harmony [55], and computed the *k*NN graph. In each merged dataset, we retained cell type annotations for 10X cells, while we assigned a "Unknown" label to all cells derived from the other technology (i.e. MARS-seq or Quartz-seq2). We then calculated the *cell affinity* matrix, that is we computed the difference in entropy that can be observed by assigning each cell to each annotation cluster, this being either one of the original cell types

Table 1 Cell stability as indicator of data quality

Dataset	$\overline{S}$	$S > .5$
iCELL8 [54]	0.368	0.312
MARS-seq [54]	0.579	0.536
Chromium 10x [54]	0.716	0.728
Quartz-seq2 [54]	0.705	0.739

Table shows summary metrics derived from the Cell Stability calculated for various datasets. $\overline{S}$ is the average Cell Stability over all cells, $S > .5$ indicates the fraction of cells with Cell Stability higher than 0.5

or "Unknown". Once the matrix has been computed, each cell from the query data is assigned to the group with the highest likelihood. The rationale behind this approach is that if cells belong to the same annotation group, then more information is required to describe the graph if they were annotated as different cell types; hence, cells from the query datasets should retain their "Unknown" label if and only if there is not enough evidence to associate them to another group. We compared the accuracy of the outcome to *k*NN classification, given by the closest entry in the *k*NN graph, and to `ingest`, a tool included in `scanpy` based on *k*NN classification of UMAP embeddings. Analysis of a well defined dataset, such as Quartz-seq2, reveals that the three approaches are equally good in classifying unknown cells (Fig. 5, central column), with accuracies ranging from .870 to .927. When data are noisy, instead, *k*NN-based methods show low accuracy and a tendency to assign the most represented cell group (HEK293T) to the unlabelled cells. This misannotation is particular evident for `ingest`, in which only CD4 T cells and HEK cells are transferred, resulting in the lowest accuracy (0.243). Conversely, `schist` is able to assign correct labels with higher accuracy (0.641). Moreover, *k*NN methods assign a label to each cell, whereas `schist` does not relabel cells if there are no sufficient evidence (e.g. the "Unknown" state is the most likely). Interestingly, we found that for the largest part of cells without assigned label, the second choice by affinity ranking was indeed the appropriate one (Additional file 1: Fig. S6).

Choice of an optimal hierarchy level

`schist` fits a hierarchical model of communities into a graph. When it comes to analysis of single cell data, it means that the cells are best described by the hierarchy itself and that cells *can* be grouped consistently at each level of the tree. In addition, the size of groups at the deepest level scales as $O(N/\log N)$ [44], where N is the number of cells. Given the current throughput in single cell experiments ($\sim$10k cells), the number of groups is difficult to handle. For this reason, in most of single cell experiments, it is preferable to identify an optimal level of the hierarchy that best resembles the cell properties at the scale they can be validated.

A possible strategy is based on Random Matrix Theory, as suggested by the authors of the SC3 package [15], for which a suitable number of clusters, $\hat{k}$, is determined by the number of eigenvalues of the $\boldsymbol{Z}^\top \boldsymbol{Z}$ matrix (where $\boldsymbol{Z}$ is the normalised count matrix) significantly different at $p < .001$ from the appropriate Tracy-Widom distribution. According to this strategy, the optimal level i^k is the one that minimises the number of partitions and $\hat{k}$:

$$i^k = \underset{x}{\operatorname{argmin}} |B_x - \hat{k}| \tag{3}$$

where B_x is the number of non empty partitions at level x.

An alternative strategy is to evaluate the behaviour of modularity at different hierarchy levels. While `schist` does not optimise the graph modularity Q, we observed that this tends to be maximal for the level better describing known cell populations, so the optimal level i^Q is

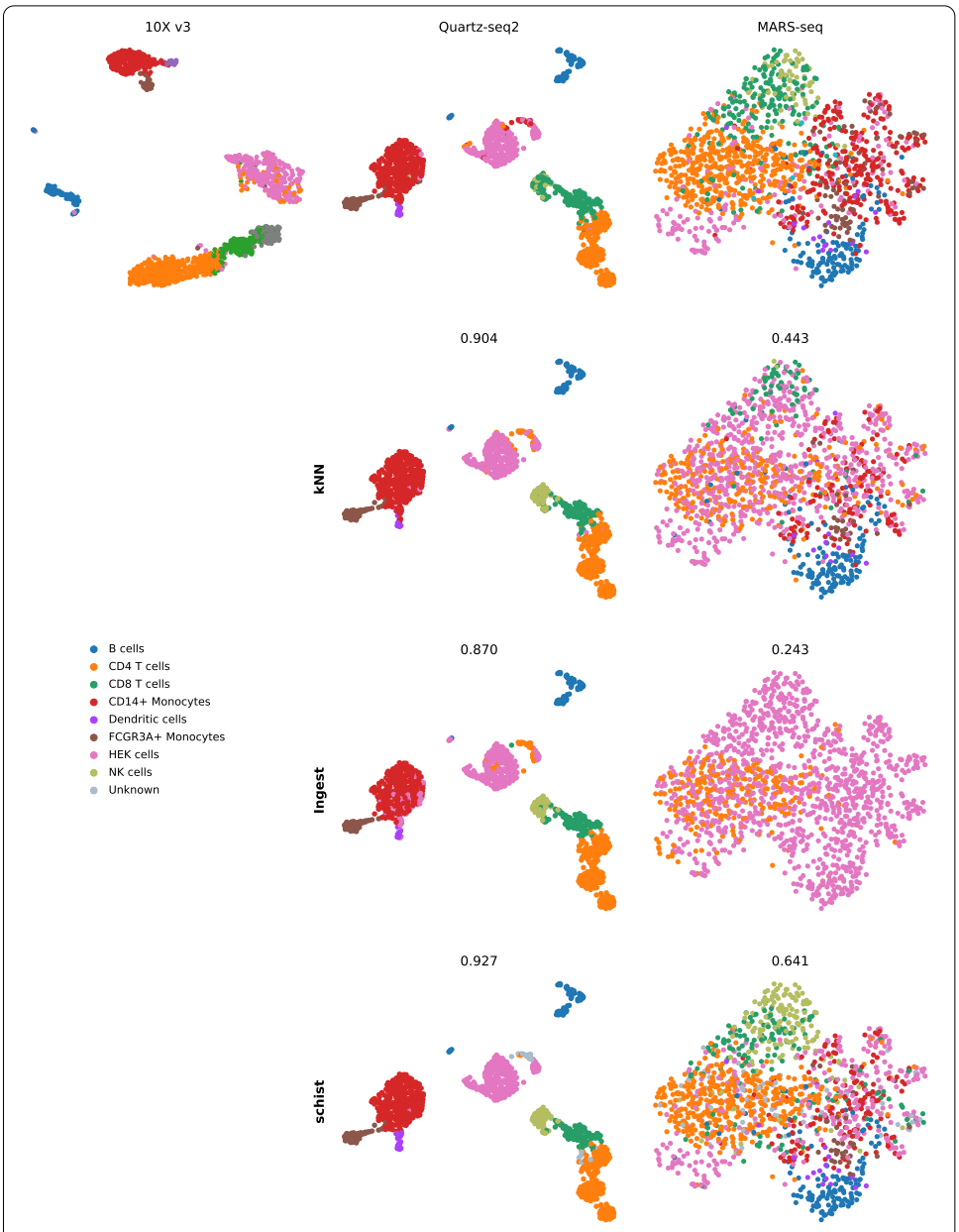


Fig. 5 Label transfer using SBM. The first line reports UMAP embeddings for datasets profiled with Chromium 10X V3, Quartz-seq2 and MARS-seq, each annotated by known cell types. Quartz-seq2 and MARS-seq were reannotated using *k*NN method, `scanpy.tools.ingest()` or `schist`. The accuracy of each label transfer task is reported above the corresponding UMAP

$$i^{Q} = \underset{x}{\operatorname{argmax}} |Q_x| \tag{4}$$

Where Q_x is modularity at level x. We collected values arising from both the approaches for some datasets used in this work (Table 2 and Additional file 1: Fig. S7)

As expected, the larger the network, the higher the optimal level. For relatively small datasets (i.e. less than 10k cells), the first level of the hierarchy contains a number of groups in line with how many observable populations are. Notwithstanding, cell groups

Table 2 Selection of the optimal level in the nSBM hierarchy

Dataset	Cells	D	$\hat{k}$	i^k	B_k	i^Q	B_Q
sc-mixology [47]	860	5	21	1	6	1	6
Chromium 10x [54]	1523	8	43	0	58	1	13
Quartz-seq2 [54]	1266	8	37	0	62	1	12
MARS-seq [54]	1401	9	9	1	16	1	16
iCELL8 [54]	1830	9	20	1	21	2	6
Mouse brain [50]	2688	15	8	2	8	1	23
Planaria [10]	21,612	51*	34	2	22	3	10

For each dataset we report the number of groups D that were given by the authors. The optimal level selection should recover a number of groups in the order of magnitude of D. Value of D in Planaria dataset is derived from manual curation of Louvain clustering. $\hat{k}$: number of groups according to RMT, i^k: level selected according to RMT criterion, B_k: number of partitions at level i^k, i^Q: level at which modularity is maximal, B_Q: number of groups at level i^Q

identified at leach level may have a biological interpretation. In particular, groups at deepest levels (0 or 1) may be relevant when studying rare populations. For example, in the hematopoiesis dataset shown in Additional file 1: Fig. S3A, groups 11 (DC) and 19 (Lymph) cannot be distinguished from nSBM level 1 and up; a closer investigation to level 0, however, revealed that these cells are clearly separated (Additional file 1: Fig. S3D). To better understand the role of deepest levels, we performed an additional analysis of a single cell dataset of mouse crypt cells [56], which was also covered in a recent paper proposing GapClust as an optimal approach to identify rare cell populations [57]. We sought to identify the four rare populations identified by GapClust. We could distinguish all but the erythrocyte group (R3) at level 1 of the hierarchy (Fig. 6 and S8), suggesting that exploring nSBM levels with appropriate community size is a valid method to spot rare populations. Of note, modularity optimisation could not pinpoint Tuft cells in appropriate way (Additional file 1: Fig. S8C), not even at high resolution, hence prompting the development of specific approaches such as GapClust.

As the size and number of communities is strictly dependent on the *k*NN graph generation, we investigated how different parameters (i.e. number of principal components and number of neighbors) affect the partition structure (Additional file 1: Fig. S9). We found, as a general pattern, that increasing the number of neighbors results in more granular structure at level 0, with different solutions being consistent (Additional file 1: Fig. S10), suggesting that higher number of neighbors provides richer description of the dataset. The number of PCs used to evaluate cell-to-cell distance influences the variability of community sizes; the consistency among different solutions is high when a sufficient number of PCs is chosen, data suggest that for large datasets more PCs should be included to include adequate fraction of overall variability.

Analysis of runtimes

Minimisation of a nSBM is a process that requires a large amount of computational resources. While the underlying `graph-tool` library is efficient in exploring the solution space using a multiflip MCMC sampling strategy, the number of required iterations before convergence is considerable and the running time scales linearly with the number of edges. Moreover, to collect a consensus partition, we minimise multiple models (default: 100) that need to be averaged. To give a reference, we report runtimes for some

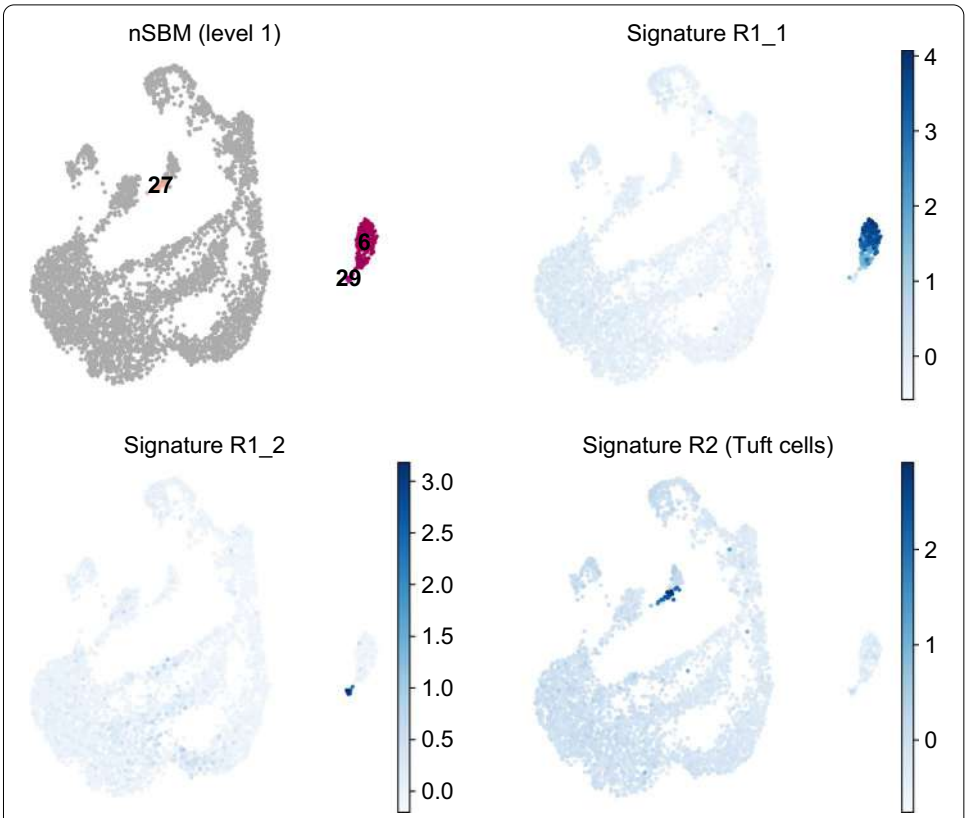


Fig. 6 Identification of rare populations. UMAP embeddings of mouse crypt cells highlighting the groups found at level 1 of nSBM hierarchy corresponding to rare populations identified by GapClust [57]. Each rare population is colored by a signature calculated from published gene lists (see methods)

Table 3 Time required to run different partitioning strategies implemented in `schist` on various datasets

Dataset	Cells	Edges	Leiden	PPBM	nSBM
sc-mixology [47]	860	9186	00:06	00:13	00:36
Quartz-seq2 [54]	1266	14,603	00:10	00:19	00:45
MARS-seq [54]	1401	21,756	00:20	00:34	02:14
iCELL8 [54]	1830	30,636	00:23	00:40	03:02
Chromium 10x [54]	1523	21,447	00:14	00:26	01:07
Hematopoiesis [52]	2730	15,444	00:37	01:27	05:52
Mouse Cortex [58]	3005	54,460	00:59	00:53	07:32
Endocrinogenesis [59]	3696	74,670	01:15	01:29	10:56
Baron Pancreas [60]	8569	294,480	03:51	07:35	1:33:40
Airzani Liver [61]	10,368	354,440	04:13	09:47	1:42:23
Tabula Muris [48]	12,434	265,610	03:07	10:00	1:23:35
Planaria [10]	21,612	173,667	05:41	13:52	1:20:40

All approaches fit 100 models. Number of nodes and edges refer to the structure of the *k*NN graph as built by `scanpy`. Times are expressed in MM:SS

example datasets in Table 3 on a commodity hardware (Intel i7@2.8 GHz, 32 GB RAM). Compared to Leiden approach, nSBM requires at best $\sim 6\times$ times more, and $\sim 30\times$ at worst. A reasonably fast alternative to the nSBM is the Planted Partition Block Model

(PPBM), for which we also report runtimes. The PPBM [46] is able to find statistically significant assortative modules and eliminates the resolution parameter; differently from nSBM, PPBM is not hierarchic.

Conclusions

Identification of cells sharing similar properties in single cell experiments is of paramount importance. A large number of approaches have been described, although the standardisation of analysis pipelines converged to methods that are based on modularity optimisation. We tackled the biological problem using a different approach, nSBM, which has several advantages over existing techniques. As random data may have modular structure [34], an important property of our approach is that it does not overfit data by finding partitions when, in fact, there are not. Another important advantage is that the hierarchical definition of cell groups eliminates the choice of an arbitrary threshold on clustering resolution. In addition, we showed that the hierarchy itself could have a biological interpretation, implying that the hierarchical model is a valid representation of the cell ensemble. We performed experiments to evaluated the impact of parameters to build the *k*NN graph on the final partitions. We found that our solutions were consistent across parameters; we also found that the more information is included during graph generation, the more granular the final description. Our results suggest that the number of principal components used to evaluate the cell-to-cell distance may have an impact on the final results and that the number of components to include depends on the data size and heterogeneity; while intuitive, this finding is in contrast with what has been observed for other PCA-based methods [18], whereas has an impact on probabilistic methods [49].

The Bayesian formulation of Stochastic Block Models provides the possibility to perform inference on a graph for any partition configuration, thus allowing reliable model selection using an interpretable measure, entropy. We exploited this property to perform label transfer with high accuracy and with the possibility to discard cells with unreliable assignments. In all, `schist` facilitates the adoption of nSBM by the bioinformatics community and exposes a robust framework to perform tasks that go beyond the principled identification of cell clusters.

The major drawback of adopting this strategy is the substantial increase of runtimes. As observed, model minimisation is many times slower than the extremely fast Leiden approach. It should be noted that `schist` initialises multiple models that are treated by multiple concurrent processes. `graph-tool` itself supports CPU-level parallelisation for some of its tasks. These optimisations are well suited for clustered computing infrastructure. Further development, possibly including GPU-level parallelisation, is surely required to accomodate the large size of datasets that are being produced.

Materials and methods

Unless differently stated, all the analysis were produced using `scanpy` v1.7.1 [22] and `schist` v0.7.6 and the corresponding dependencies. All models were initialised 100 times, herein including Leiden partitioning for which we also calculated the consensus partition.

Analysis of randomized data

Data were retrieved in `scanpy` environment using `scanpy.datasets.pbmc3k_processed()` function. The random *k*NN graph was obtained shuffling the node labels of each edge. UMAP embedding was recomputed after randomisation using the shuffled graph. To generate data with white noise we computed the genewise means (μ_g) and standard deviations (σ_g) of log-normalized counts excluding 0 values. We generated random values using μ_g and $k\sigma_g$, $k \in \{0.5, 1, 1.5, 2\}$, and added to original expression values.

Analysis of cell mixtures

Data and metadata for five cell mixture profiled by Chromium 10x were downloaded from the sc-mixology repository (https://github.com/LuyiTian/sc_mixology). Cells with less than 200 genes were excluded, as genes detected in less than 3 cells. Cells with less than 5% of mitochondrial genes were retained for subsequent analysis. Data were normalised and log-transformed; number of genes and percentage of mitochondrial genes were regressed out. *k*NN graph was built with default parameters (50 components and 15 nearest neighbours). Data were assessed by SCCAF using cell line annotation. Mean cross-validated accuracy was set as target for all the models.

Analysis of Tabula Muris data

Data for FACS isolated cells sequenced with Smart-seq2 were downloaded from the Tabula Muris consortium [49] (https://doi.org/10.6084/m9.figshare.5975392), analysis was restricted to Skin, Spleen, Large Intestine and Brain-Myeloid count matrices. Cells with less than 200 genes were excluded, as genes detected in less than 3 cells. Data were normalised and log-transformed. Merged data were then processed using Harmony [55] by the `scanpy.external.pp.harmony_integrate()` function with default parameters. *k*NN graph was built on integrated data using 50 components and 30 nearest neighbours. Data were assessed by SCCAF using tissue annotation. Mean cross-validated accuracy was set as target for all the models.

Analysis of visium H&E data

Data were retrieved using `squidpy.datasets.visium_hne_adata()` built-in function, without further processing. Leiden clustering was performed using `schist.inference.leiden()` function, allowing for 100 initialisations, with resolutions $\gamma = 1$ and $\gamma = 0.5$.

Analysis of hematopoietic differentiation

Data were retrieved using `scanpy`'s built-in functions and were processed as in [53], except for *k*NN graph built using 30 principal components, 30 neighbours and diffmap as embedding. Gene signatures were calculated with `scanpy.tools.score_genes()` using the following gene lists

- Erythroids: Gata1, Klf1, Epor, Gypa, Hba-a2, Hba-a1, Spi1
- Neutrophils, Elane, Cebpe, Ctsg, Mpo, Gfi1
- Monocytes, Irf8, Csf1r, Ctsg, Mpo

Processing of PBMC data from various platforms

Count matrices were downloaded from GEO using the following accession numbers: GSE133535 (Chromium 10Xv3), GSE133543 (Quartz-seq2), GSE133542 (MARS-seq) and GSE133541 (iCELL8). Data were processed according to the methods in the original paper [54]. Briefly, cells with less than 10,000 total number of reads as well as the cells having less than 65% of the reads mapped to their reference genome were discarded. Cells in the 95th percentile of the number of genes/cell and those having less than 25% mitochondrial gene content were included in the downstream analyses. Genes that were expressed in less than five cells were removed. Data were normalised and log-transformed, highly variable genes were detected at minimal dispersion equal to 0.5. Neighbourhood graph was built using 30 principal components and 20 neighbours.

Analysis of crypt cells data

Count matrix for untreated crypt cells (GSM3308718) was downloaded from GEO. Cells with less than 200 genes and genes detected in less than 2 cells were excluded from the analysis. After normalization and log-transformation, highly variable genes were selected with a cutoff on the mean expression equal to 0.05. Rare subpopulations were first highlighted with `scanpy.tools.score_genes()` using signatures published in [57]:

- R1_1: Cd8a, Cd3g, Ccl5, Gzma, Gzmb, RGs1, Nkg7, Cd7, Fcer1g
- R1_2: H2-Aa, H2-Ab1, H2-Eb1, Cd74, Ly6d, Ebf1, Cd79a, Mef2c
- R2: Krt18, Cd24a, Adh1, Cystm1, Aldh2, Dclk1, Sh2d6, Rgs13, Hck, Trpm5
- R3: Alas2, Hbb-bs, Hba-a1, Hbb-bt

Label transfer

Processed data for MARS-seq or Quart-seq2 platforms were merged to data for 10X V3. Merged data were then processed using Harmony [55] by the `scanpy.external.pp.harmony_integrate()` function with default parameters. Cells not belonging to the 10X data were assigned an "Unknown" label. We calculated cell affinity to each annotation label using `schist.tl.calculate_affinity()` function. We assigned the most affine annotation only to "Unknown" cells. For *k*NN-based procedure, we built a *k*NN graph on the merged data using `pynndescent` library on the 10XV3 subset of cells in the merged data, then we assigned "Unknown" cells to the closest entry in the graph. Assignment by `scanpy.tools.ingest()` was performed using default parameters.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-021-04489-7.

Additional file 1. Supplementary Table S1; Supplementary Figures S1–S10.

Acknowledgements
We would like to thank Tiago de Paula Peixoto (Central European University, ISI Foundation) and Giovanni Petri (ISI Foundation) for the discussions and the precious hints. We also would like to thank all people at COSR, in particular Giovanni Tonon and Paolo Provero.

Authors' contributions
LM, VG and DC performed the analysis and wrote the manuscript. LM and DC wrote the software. DC conceived the study. All authors read and approved the final manuscript.

Funding
This work has been supported by Accelerator Award: A26815 entitled: "Single-cell cancer evolution in the clinic" funded through a partnership between Cancer Research UK and Fondazione AIRC.

Availability of data and materials
No datasets were generated in the current study. The original third-party datasets that were analysed are included in the corresponding publications [10, 47, 52, 54, 58–61]

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Center for Omics Sciences, IRCCS San Raffaele Institute, Milan, Italy. [2]Università Vita-Salute San Raffaele, Milan, Italy. [3]Department of Informatics, Systems and Communication, University of Milano-Bicocca, Milan, Italy.

Received: 12 May 2021 Accepted: 19 November 2021
Published online: 30 November 2021

References

1. Svensson V, Vento-Tormo R, Teichmann SA. Exponential scaling of single-cell RNA-seq in the past decade. Nat Protoc. 2018;13(4):599–604. https://doi.org/10.1038/nprot.2017.149.
2. Guo J, Grow EJ, Mlcochova H, Maher GJ, Lindskog C, Nie X, et al. The adult human testis transcriptional cell atlas. Cell Res. 2018;28(12):1141–57. https://doi.org/10.1038/s41422-018-0099-2.
3. Vento-Tormo R, Efremova M, Botting RA, Turco MY, Vento-Tormo M, Meyer KB, et al. Single-cell reconstruction of the early maternal-fetal interface in humans. Nature. 2018;563(7731):347–53. https://doi.org/10.1038/s41586-018-0698-6.
4. Rozenblatt-Rosen O, Regev A, Oberdoerffer P, Nawy T, Hupalowska A, Rood JE, et al. The Human tumor atlas network: charting tumor transitions across space and time at single-cell resolution. Cell. 2020;181(2):236–49. https://doi.org/10.1016/j.cell.2020.03.053.
5. Tirosh I, Izar B, Prakadan SM, Wadsworth MH, Treacy D, Trombetta JJ, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science. 2016;352(6282):189–96. https://doi.org/10.1126/science.aad0501.
6. Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science. 2014;344(6190):1396–401. https://doi.org/10.1126/science.1254257.
7. Neftel C, Laffy J, Filbin MG, Hara T, Shore ME, Rahme GJ, et al. An integrative model of cellular states, plasticity, and genetics for glioblastoma. Cell. 2019;178(4):835-849.e21. https://doi.org/10.1016/j.cell.2019.06.024.
8. 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;360(6385):176–82. https://doi.org/10.1126/science.aam8999.
9. Wagner DE, Weinreb C, Collins ZM, Briggs JA, Megason SG, Klein AM. Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. Science. 2018;360(6392):981–7. https://doi.org/10.1126/science.aar4362.
10. Plass M, Solana J, Wolf FA, Ayoub S, Misios A, Glažar P, et al. Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. Science. 2018. https://doi.org/10.1126/science.aaq1723.

11. Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, et al. The human cell atlas. eLife. 2017. https://doi.org/10.7554/eLife.27041.
12. Wang B, Zhu J, Pierson E, Ramazzotti D, Batzoglou S. Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. Nat Methods. 2017;14(4):414–6. https://doi.org/10.1038/nmeth.4207.
13. Lin P, Troup M, Ho JWK. CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data. Genome Biol. 2017;18(1):59. https://doi.org/10.1186/s13059-017-1188-0.
14. Huh R, Yang Y, Jiang Y, Shen Y, Li Y. SAME-clustering: single-cell aggregated clustering via mixture model ensemble. Nucleic Acids Res. 2020;48(1):86–95. https://doi.org/10.1093/nar/gkz959.
15. Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T, et al. SC3: consensus clustering of single-cell RNA-seq data. Nat Methods. 2017;14(5):483–6. https://doi.org/10.1038/nmeth.4236.
16. Ranjan B, Schmidt F, Sun W, Park J, Honardoost MA, Tan J, et al. scConsensus: combining supervised and unsupervised clustering for cell type identification in single-cell RNA sequencing data. BMC Bioinform. 2021;22(1):186. https://doi.org/10.1186/s12859-021-04028-4.
17. Li X, Wang K, Lyu Y, Pan H, Zhang J, Stambolian D, et al. Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis. Nat Commun. 2020;11(1):2338. https://doi.org/10.1038/s41467-020-15851-3.
18. Krzak M, Raykov Y, Boukouvalas A, Cutillo L, Angelini C. Benchmark and parameter sensitivity analysis of single-cell RNA sequencing clustering methods. Front Genet. 2019;10:1253. https://doi.org/10.3389/fgene.2019.01253.
19. Kiselev VY, Andrews TS, Hemberg M. Challenges in unsupervised clustering of single-cell RNA-seq data. Nat Rev Genet. 2019;20(5):273–82. https://doi.org/10.1038/s41576-018-0088-9.
20. Duò A, Robinson MD, Soneson C. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. F1000Research. 2018;7:1141. https://doi.org/10.12688/f1000research.15666.2.
21. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol. 2018;36(5):411–20. https://doi.org/10.1038/nbt.4096.
22. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 2018;19(1):15. https://doi.org/10.1186/s13059-017-1382-0.
23. Setty M, Kiseliovas V, Levine J, Gayoso A, Mazutis L, Pe'er D. Characterization of cell fate probabilities in single-cell data with Palantir. Nat Biotechnol. 2019;37(4):451–60. https://doi.org/10.1038/s41587-019-0068-4.
24. Lange M, Bergen V, Klein M, Setty M, Reuter B, Bakhti M, et al. Cell rank for directed single-cell fate mapping. BioRxiv. 2020. https://doi.org/10.1101/2020.10.19.345983.
25. Bergen V, Lange M, Peidli S, Wolf FA, Theis FJ. Generalizing RNA velocity to transient cell states through dynamical modeling. Nat Biotechnol. 2020;38(12):1408–14. https://doi.org/10.1038/s41587-020-0591-3.
26. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech Theory Exp. 2008;2008(10):P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008.
27. Traag VA, Waltman L, van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep. 2019;9(1):5233. https://doi.org/10.1038/s41598-019-41695-z.
28. Levine JH, Simonds EF, Bendall SC, Davis KL, Amir EAD, Tadmor MD, et al. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. Cell. 2015;162(1):184–97. https://doi.org/10.1016/j.cell.2015.05.047.
29. Newman MEJ, Girvan M. Finding and evaluating community structure in networks. Phys Rev E Stat Nonlinear Soft Matter Phys. 2004;69(2 Pt 2):026113. https://doi.org/10.1103/PhysRevE.69.026113.
30. Traag VA, Van Dooren P, Nesterov Y. Narrow scope for resolution-limit-free community detection. Phys Rev E. 2011. https://doi.org/10.1103/PhysRevE.84.016114.
31. Reichardt J, Bornholdt S. Statistical mechanics of community detection. Phys Rev E. 2006. https://doi.org/10.1103/PhysRevE.74.016110.
32. 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;21(1):31. https://doi.org/10.1186/s13059-020-1926-6.
33. Fortunato S, Barthélemy M. Resolution limit in community detection. Proc Natl Acad Sci USA. 2007;104(1):36–41. https://doi.org/10.1073/pnas.0605965104.
34. Guimerà R, Sales-Pardo M, Amaral LAN. Modularity from fluctuations in random graphs and complex networks. Phys Rev E. 2004. https://doi.org/10.1103/PhysRevE.70.025101.
35. Baran Y, Bercovich A, Sebe-Pedros A, Lubling Y, Giladi A, Chomsky E, et al. MetaCell: analysis of single-cell RNA-seq data using K-nn graph partitions. Genome Biol. 2019;20(1):206. https://doi.org/10.1186/s13059-019-1812-2.
36. Tang M, Kaymaz Y, Logeman BL, Eichhorn S, Liang ZS, Dulac C, et al. Evaluating single-cell cluster stability using the Jaccard Similarity Index. Bioinformatics. 2020. https://doi.org/10.1093/bioinformatics/btaa956.
37. Xu C, Su Z. Identification of cell types from single-cell transcriptomes using a novel clustering method. Bioinformatics. 2015;31(12):1974–80. https://doi.org/10.1093/bioinformatics/btv088.
38. Miao Z, Moreno P, Huang N, Papatheodorou I, Brazma A, Teichmann SA. Putative cell type discovery from single-cell gene expression data. Nat Methods. 2020;17(6):621–8. https://doi.org/10.1038/s41592-020-0825-9.
39. Holland PW, Laskey KB, Leinhardt S. Stochastic blockmodels: first steps. Soc Netw. 1983;5(2):109–37. https://doi.org/10.1016/0378-8733(83)90021-7.
40. Peixoto TP. Nonparametric Bayesian inference of the microcanonical stochastic block model. Phys Rev E. 2017;95(1–1):012317. https://doi.org/10.1103/PhysRevE.95.012317.
41. Karrer B, Newman MEJ. Stochastic blockmodels and community structure in networks. Phys Rev E Stat Nonlinear Soft Matter Phys. 2011;83(1 Pt 2):016107. https://doi.org/10.1103/PhysRevE.83.016107.
42. Peixoto TP. Parsimonious module inference in large networks. Phys Rev Lett. 2013;110(14):148701. https://doi.org/10.1103/PhysRevLett.110.148701.
43. Peixoto TP. Efficient Monte Carlo and greedy heuristic for the inference of stochastic block models. Phys Rev E Stat Nonlinear Soft Matter Phys. 2014a;89(1):012804. https://doi.org/10.1103/PhysRevE.89.012804.
44. Peixoto TP. Hierarchical block structures and high-resolution model selection in large networks. Phys Rev X. 2014b;4(1):011047. https://doi.org/10.1103/PhysRevX.4.011047.

45. Peixoto TP. Revealing consensus and dissensus between network partitions. Phys Rev X. 2021;11(2):021003. https://doi.org/10.1103/PhysRevX.11.021003.
46. Zhang L, Peixoto TP. Statistical inference of assortative community structures. Phys Rev Res. 2020;2(4):043271. https://doi.org/10.1103/PhysRevResearch.2.043271.
47. Tian L, Dong X, Freytag S, Lê Cao KA, Su S, JalalAbadi A, et al. Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. Nat Methods. 2019;16(6):479–87. https://doi.org/10.1038/s41592-019-0425-8.
48. Consortium TM, et al. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. Nature. 2018;562(7727):367–72. https://doi.org/10.1038/s41586-018-0590-4.
49. Raimundo F, Vallot C, Vert JP. Tuning parameters of dimensionality reduction methods for single-cell RNA-seq analysis. Genome Biol. 2020;21(1):212. https://doi.org/10.1186/s13059-020-02128-7.
50. Gracia Villacampa E, Larsson L, Kvastad L, Andersson A, Carlson J, Lundeberg J. Genome-wide spatial expression profiling in FFPE tissues. BioRxiv. 2020. https://doi.org/10.1101/2020.07.24.219758.
51. Palla G, Spitzer H, Klein M, Fischer DS, Schaar AC, Kuemmerle LB, et al. Squidpy: a scalable framework for spatial single cell analysis. BioRxiv. 2021. https://doi.org/10.1101/2021.02.19.431994.
52. Paul F, Arkin Y, Giladi A, Jaitin DA, Kenigsberg E, Keren-Shaul H, et al. Transcriptional heterogeneity and lineage commitment in myeloid progenitors. Cell. 2015;163(7):1663–777. https://doi.org/10.1016/j.cell.2015.11.013.
53. 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;20(1):59. https://doi.org/10.1186/s13059-019-1663-x.
54. 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;38(6):747–55. https://doi.org/10.1038/s41587-020-0469-4.
55. 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;16(12):1289–96. https://doi.org/10.1038/s41592-019-0619-0.
56. Ayyaz A, Kumar S, Sangiorgi B, Ghoshal B, Gosio J, Ouladan S, et al. Single-cell transcriptomes of the regenerating intestine reveal a revival stem cell. Nature. 2019;569(7754):121–5. https://doi.org/10.1038/s41586-019-1154-y.
57. Fa B, Wei T, Zhou Y, Johnston L, Yuan X, Ma Y, et al. GapClust is a light-weight approach distinguishing rare cells from voluminous single cell expression profiles. Nat Commun. 2021;12(1):4197. https://doi.org/10.1038/s41467-021-24489-8.
58. Zeisel A, Muñoz-Manchado AB, Codeluppi S, Lönnerberg P, La Manno G, Juréus A, et al. Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science. 2015;347(6226):1138–42. https://doi.org/10.1126/science.aaa1934.
59. Bastidas-Ponce A, Tritschler S, Dony L, Scheibner K, Tarquis-Medina M, Salinno C, et al. Comprehensive single cell mRNA profiling reveals a detailed roadmap for pancreatic endocrinogenesis. Development. 2019. https://doi.org/10.1242/dev.173849.
60. Baron M, Veres A, Wolock SL, Faust AL, Gaujoux R, Vetere A, et al. A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. Cell Syst. 2016;3(4):346-360.e4. https://doi.org/10.1016/j.cels.2016.08.011.
61. Aizarani N, Saviano A, Mailly L, Durand S, Herman JS, et al. A human liver cell atlas reveals heterogeneity and epithelial progenitors. Nature. 2019;572(7768):199–204. https://doi.org/10.1038/s41586-019-1373-2.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

BMC

CODE AVAILABILITY

MOWGAN is available at https://github.com/vgiansanti/MOWGAN. Tutorials for the different application can be found in the same repository.

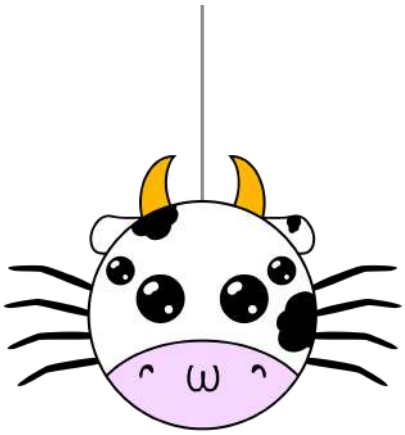


Figure 25 MOWGAN logo.

ACKNOWLEDGMENTS

I am grateful to AIRC/CRUK/FC AECC Accelerator Award “Single Cell Cancer Evolution in the Clinic” A26815 (AIRC number program 2279) for supporting me during the PhD. I would also like to thank all the members at Center for Omics Sciences at IRCCS Ospedale San Raffaele (COSR) who have hosted me in these three years.