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



DOI:10.1038/s41587-021-01031-1

['[AU: Please shorten the main text to our maximum allowed length of 4500 words, excluding abstract, figure legends, methods and references.]

[AU: Subheadings are limited to 59 characters (incl. spaces). Please check throughout]

**[AU: Please note that we only allow 6 main display items. I would suggest combining either figure 1 and 2 or 3 and 4 as these seem to fit best thematically. It might be necessary to move a panel or two to the extended data to keep the figures to a reasonable size]**

AU: Please add code availability statement (see below, page 42)

**Editorial summary**

Single-cell mapping of heterochromatin and euchromatin defines chromatin velocity.

# Chromatin Velocity reveals epigenetic dynamics by single-cell profiling of heterochromatin and euchromatin [AU:OK? ok]

Martina Tedesco [§ 1,2], Francesca Giannese [§ 3], Dejan Lazarević [3], Valentina Giansanti [3,4], Dalia Rosano [2,5], Silvia Monzani [6], Irene Catalano [7,8], Elena Grassi [7,8], Eugenia R. Zanella [8], Oronza A. Botrugno[2], Leonardo Morelli [3], Paola Panina Bordignon [1,9], Giulio Caravagna[10], Andrea Bertotti

[7,8], Gianvito Martino [1,9], Luca Aldrighetti [11], Sebastiano Pasqualato [6], Livio Trusolino [7,8], Davide Cittaro* [3], Giovanni Tonon* [2, 3]

1. Università Vita-Salute San Raffaele, Milano, Italy.
2. Functional Genomic 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. (Present Address) Department of Surgery and Cancer, Imperial College, London, UK.
6. Biochemistry and Structural Biology Unit, Department of Experimental Oncology, IEO, IRCCS European Institute of Oncology, Milano, Italy.
7. Department of Oncology, University of Torino School of Medicine, Candiolo, Torino, Italy
8. Candiolo Cancer Institute FPO- IRCCS, Candiolo, Torino, Italy.
9. Neuroimmunology Unit, Institute of Experimental Neurology, Division of Neuroscience, IRCCS San Raffaele Hospital, Milano, Italy.
10. Department of Mathematics and Geosciences, University of Trieste, Italy.
11. Hepatobiliary Surgery Division, IRCCS San Raffaele Hospital, Milano, Italy.

§: These Authors contributed equally.

***Corresponding authors: Davide Cittaro, cittaro.davide@hsr.it, Giovanni Tonon, tonon.giovanni@hsr.it**

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 an hybrid transposase including the chromodomain of the heterochromatin protein 1-$\alpha$ (HP1$\alpha$) , involved in heterochromatin assembly and maintenance through its binding to H3K9me3 and developed a single-cell method, scGET-seq (genome and epigenome by transposases sequencing), that unlike scATAC-seq comprehensively probes both open and closed chromatin, concomitantly recording the underlying genomic sequences [AU: Please briefly describe in a bit more detail how the method works and how it differs from previous methods. Abstract word count limit is 160 words]. We tested scGET-seq in cancer-derived organoids and 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, which 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. [AU: OK? ok]

# Introduction

Cancers are characterized by extensive inter-patient and intra-tumour heterogeneity, down to the single cell level[1]. This fuels clonal evolution, leading to treatment resistance[2], the leading cause of death for cancer patients. The mechanisms underlying such resistance are still largely unknown, especially for standard chemotherapeutic and immunotherapeutic regimens. Increasingly detailed analysis of cancer genomes, before and after treatment, have so far failed to identify genetic causes which 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 ensuing after treatment, could finally 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 then 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 on the cellular status of open chromatin. However, the epigenetic modifications of large portions of the genome which exert 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 regulate a large array of transposable elements and ncRNAs[11–13]. Heterochromatin is assembled and maintained through the tri-methylation of the lysine 9 on histone 3 (H3K9me3) [12,14] and its accurate regulation is essential for the cells, for example towards the definition of cell identity[12,13] and the maintenance of genomic integrity[15].

While single-cell transcriptomic analysis has fostered ground-breaking insights on the biology of healthy and diseased tissues, including cancer[16,17], a tool which comprehensively audits, at the single cell level, both the genomic and the epigenetic landscape to our knowledge 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 Immuno-Precipitation (TAM-ChIP) approach, which combines the antibody-mediated targeting of chromatin immune-precipitation 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 recognizing the histone mark H3K9me3 (or H3K4me3, as control), in line with a recent report[18], which was then bound by a secondary antibody conjugated to Tn5. H3K4me3 TAM-ChIP-seq profiles mirrored the corresponding H3K4me3 ChIP-seq profiles. Instead, when a Tn5-secondary antibody complex recognizing H3K9me3-specific primary antibody was used, Tn5 tagmented H3K9me3-enriched compacted chromatin regions (Extended Data Fig. 1b), results confirmed by Real Time-qPCR (Extended Data Fig. 1c).

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

## Hybrid CD (HP1α)-Tn5 targets H3K9me3 chromatin regions

TAM-ChIP towards H3K9me3 was only partially effective in guiding Tn5 transposase towards 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 Tn5 natural tropism. To this end, we extensively reviewed proteins and domains targeting H3K9me3. We finally selected heterochromatin protein 1-α (HP1α), one of the hallmark proteins involved in heterochromatin assembly and maintenance, which specifically binds H3K9me3, through its chromodomain (CD)[19–21].

We generated a hybrid protein, whereby the HP1α CD was cloned alongside Tn5 (Extended Data. Fig. 2a). In order to link the chromodomain with Tn5 transposase, we took advantage of the natural linker that connects the chromodomain and the chromoshadow domain of HP1α, which we extended with two artificial linkers of different length (TnH#1-4, Extended Data Fig. 2a). All four hybrid constructs were as efficient as the native Tn5 (either the commercial Nextera enzyme or in-house produced, from now on, Tn5) to fragment and insert oligos on genomic DNA (Extended Data Fig. 2b).

We then determined whether TnH#1-4 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 1b and Extended Data Fig. 2d and 2e). We identified the construct TnH#3, from now on TnH, as the most efficient (Fig. 1b and Extended Data Fig. 2d and 2e).

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 sequentially the two enzymes (Extended Data Fig. 3b) in the transposition reaction. The sequential use of native Tn5, followed by TnH, provided the most comprehensive mapping of the two chromatin profiles.

All 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 eu- and heterochromatin. We call this method GET-seq (**g**enome and **e**pigenome by **t**ransposases **seq**uencing).

## 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), replacing 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, using 10X Tn5, TnH, Tn5, or a combination of TnH and Tn5 (scGET-seq) in Caki-1 cells, and found that the 4 profiles were overlapping (Extended Data Fig. 4a). We next explored the portion of the genome which was captured by each transposase. TnH had the higher mean distribution of coverage per cell, with a smaller standard deviation, when compared with 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 pseudo-bulks and compared with 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 the cancer cell lines HeLa (20%) and Caki-1 (80%), 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 clusters (Fig. 2a), in accordance with standard scATAC-seq results (Fig. 2b).

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

## Genomic copy number variants at single cell level

The definition of genomic copy number variants (CNVs) using scATAC-seq remains imprecise since only accessible chromatin regions are surveyed by this approach and the remaining genomic sequences could only be imputed from adjacent regions[22].

As TnH targets also H3K9me3-enriched chromatin regions, we tested whether it could be harnessed also to define CNVs. Whole genome sequencing (WGS) revealed several CNVs in both cell lines (Fraction of Genome Altered, FGA: Caki-1 = 0.475, HeLa = 0.508). The correlation between the genomic profiles obtained with WGS and the average pseudo-bulk profile obtained from single-cell data was much higher for the TnH signal, when compared with 10X Tn5, 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). Increasing the resolution, up to 500 kb, scGET-seq remained reliable while the ability of scATAC-seq to identify CNVs degraded, as large swaths of the genome were excluded from the analysis (Extended Fig. 5a and 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 sequencing 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 (LR) and a Support Vector Machine with linear kernel (SVM). We calculated their accuracy and F1-score 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).

In all, these data show the feasibility of single cell profiling by GET-seq, which allows for a more precise description of genomic features with respect to scATAC-seq.

## scGET-seq identifies clonality in patient-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, patient derived organoids (PDOs). We thus used a tumour matched-normal design to generate whole-exome data derived from two hepatic metastases of primary colorectal tumours. The analysis of somatic single nucleotide variants and allele-specific copy numbers showed high-level 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 17 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 2 clusters, corresponding to CRC6 and CRC17. Notably, only a minority of the cells assessed were misclassified (Extended data Table S1). A similar analysis on Tn5-derived reads showed a tendency for overclustering and of cell misclassification (Fig. 3b and Extended data Table S1). We finally explored the accuracy of variant calling (*i.e.* 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).

All 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 patient derived xenograft (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, alongside with features of plasticity that are not driven by genomic modifications but most likely by chromatin reshaping[26,27]. We hence followed cancer evolution in one PDX model throughout several weeks of treatment with the clinically approved EGFR antibody cetuximab (Extended Data Fig. 7a). Analysis of genomic segmentation by scGET-seq revealed 2 major clones in the absence of treatment (Fig. 3d and Extended Data Fig. 7b). Conversely, cells were separated into 6 different clones when assessing the pre-treatment 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). In 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 blockade. To this end, we performed functional enrichment analysis using the genes associated to the regions differentially affected in the various clones (Extended Data Table S2). In the epigenetic clones most associated with resistance, there was a significant enrichment on pathways linked to with refractoriness to EGFR inhibitors, including the phospholipase C pathway[28], TGFβ 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 remodelling phenomena[26,27].

As scGET-seq includes sequences for portion of the genome that are eluded by conventional ATAC-seq, we next sought to determine whether we could also define single nucleotide variations (SNV) within single cells. While 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 tumour, 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 which affected established cancer genes (tier 1, COSMIC Cancer Gene Census, version 92[31], Extended Data Table S3), 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,

$p$=3.57·10$^{-3}$, 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.

In all, these results suggest that scGET-seq could be used to comprehensively assess the tumour 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 dynamic between accessible and compacted chromatin at the single-cell level. We have recently demonstrated that the 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 knock-down. We identified two neatly distinguished cell groups, including 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 cell, during or after DNA replication, while 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[10], where we analysed the distribution of Repli-seq[33–35] signal over differentially enriched DNase I hypersensitive sites (DHS) regions between high- and low-coverage cells. We found that high coverage cells are characterized by 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 laminB1 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 which undergo compaction during the cell cycle, through the acquisition of H3K9me3 residues. As Kdm5c acts, and heterochromatin assembly occurs, during the middle/late S phase we focused on the G1/S cell cycle phase[32,36]. The signal emerging from Tn5 was weaker on G1/S cells where Kdm5c was not knocked down (Fig. 4a and d, black arrow, compared with TnH, 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 on G1/S cells, irrespectively 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 interaction term when looking for groupwise specific markers. We found that the TnH enrichment was significantly higher than Tn5 in groups 3 and 6 (Extended Data Fig. 8c and d), where indeed shScr cells are present in 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.

All 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 (FIB) obtained from two unrelated healthy subjects, undergoing reprogramming into induced pluripotent stem cells (iPSC), and of iPSC undergoing differentiation into neural progenitor cells (NPC). In parallel, we performed scRNA-seq analysis on cells from the same samples.

Low dimensional representation of single cell data from scGET-seq and scRNA-seq separated FIB, iPSC and NPC into three distinct populations (Fig. 5a and b). Notably, UMAP representations of both scGET-seq and scRNA-seq data showed that iPSC and NPC were in close proximity, while FIB were isolated from the other two populations, with the exception of a small subset of FIB and to a lesser extent NPCs clustering alongside iPSC exclusively in the scGET-seq data (black arrow in Fig. 5a).

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 which are crucial for the biology of each population, such as collagen for FIB, L1TD1 for iPSC[37] and PRTG for NPC[38] (Fig. 5c and Extended Data table S4), 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 bifurcation alongside developmental paths[39]. To this end, a tool has been recently devised, Palantir[39], which is able to capture these dynamics from scRNA-seq data. When we applied Palantir to the GET-seq data set, 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 iPSC and the subset of FIB and NPC clustering alongside iPSC (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 (Extended data Table S5), 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 to these regions to generate a differentiation signature, which we then applied to scRNA-seq data. This signature highlighted in the scRNA-seq data a subset of NPC as well as FIB sharing similar features (red arrows in Fig.5f).

In all, these results suggest that GET-seq is able to capture the epigenetic diversity arising during developmental processes and to 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 which 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 a more relaxed chromatin, while the opposite is indicative of chromatin compaction (Extended Data Fig. 9b). We found that this approach, which we named Chromatin 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 iPSC, NPC and FIB from which most differentiation processes appeared to arise (Fig. 6a and 5d). Also, RNA-velocity revealed that the subset of FIB 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. The 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 Extended Data Table S6).

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 TF in specific cell trajectories. We applied a Projection to Latent Structures regression analysis (PLS)[45] fitting the cell TF scores to cell clusters (Extended Fig. 89c and Extended Data Table S7) that clearly separated FIB on one site, and NPC and iPSC on the other. Several TFs already implicated in FIB development and maintenance were included, such as FOSL2[46], TP63[47], and NFE2L2[48]. Conversely, NPCs and iPSC were strongly enriched for TFs which are key for neural differentiation, namely NHLH1[49] and MECP2, whose mutations lead to mental retardation[50]. MECP2, MBD2 e 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 towards maturing ESC-derived neurons through the engagement of stage-specific enhancers[54].

As PLS1 seems to be associated to 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[55], focusing on the early time points (4-20 weeks post conception). With the exception of DUX4, which was not profiled in that dataset, we found that TF with the most negative loading on PLS1 have a single peak of expression in the early stages of brain development (Fig. 6g) and are abruptly downregulated afterwards. Similarly, TF 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.

All together, we posit that Chromatin Velocity captures epigenetic transitions underlying crucial biological processes and illuminates the hidden transcription factor 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 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, which defines vectors of cellular fate and predict 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 number 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 expedite 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 a multidimensional information at the epigenetic level. Bulk analysis has revealed that in development cells undergo epigenetic changes, such as modulation in the opening of open and closed chromatin, which precedes and prepares gene expression modifications[58–63]. Therefore, it stands to reason to anticipate that RNA- and chromatin velocity are going to capture non-superimposable biological processes.

Retracing the specific engagement of TF from scRNA-seq experiments is challenging[64]. Leveraging on a detailed description of the epigenome analysis provides more robust data and reduces variability, allowing the genome-wide identification of TFs, thus 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.

# References

1. McGranahan, N. & Swanton, C. Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. *Cell* vol. 168 613–628 (2017).
2. Greaves, M. Evolutionary determinants of cancer. *Cancer Discovery* **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.e7 (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. *Nature reviews. Genetics* **10**, 336–42 (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* vol. 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. *Nature methods* **10**, 1213–8 (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. *Molecular Cell* **67**, 725–727 (2017).

12. Ninova, M., Fejes Tóth, K. & Aravin, A. A. The control of gene expression and cell identity by H3K9 trimethylation. *Development (Cambridge, England)* **146**, dev181180 (2019).

13. Nicetto, D. *et al.* H3K9me3-heterochromatin loss at protein-coding genes enables developmental lineage specification. *Science (New York, N.Y.)* **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 (New York, N.Y.)* **292**, 110–3 (2001).

15. Peters, A., O'Carroll, D. & Scherthan, H. Loss of the Suv39h Histone Methyltransferases Impairs Mammalian Heterochromatin and Genome Stability. *Cell* **107**, 323–37 (2001).

16. Patel, A. P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science (New York, N.Y.)* **344**, 1396–401 (2014).

17. Aldridge, S. & Teichmann, S. A. Single cell transcriptomics comes of age. *Nature Communications* **11**, 9–12 (2020).

18. Henikoff, S., Henikoff, J., Kaya-Okur, H. & Ahmad, K. Efficient chromatin accessibility mapping in situ by nucleosome-tethered tagmentation. *eLife* **9**, (2020).

19. Jacobs, S. A. & Khorasanizadeh, S. Structure of HP1 chromodomain bound to a lysine 9-methylated histone H3 tail. *Science (New York, N.Y.)* **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**, 116–20 (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. *Nature Biotechnology* **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. *bioRxiv* (2020) doi:10.1101/2020.03.26.007138.

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. *Nature Communications* **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. *Science Translational Medicine* **12**, eaax8313 (2020).

28. Laurent-Puig, P., Lievre, A. & Blons, H. Mutations and response to epidermal growth factor receptor Inhibitors. *Clinical Cancer Research* **15**, 1133–1139 (2009).

29. Wang, C. *et al.* Acquired resistance to EGFR TKIs mediated by TGFb1/integrin B3 signaling in EGFR-mutant lung cancer. *Molecular Cancer Therapeutics* **18**, 2357–2367 (2019).

30. Hu, T. & Li, C. Convergence between Wnt-β-catenin and EGFR signaling in cancer. *Molecular Cancer* **9**, 1–7 (2010).

31. Sondka, Z. *et al.* The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nature Reviews Cancer* **18**, 696–705 (2018).

32. Rondinelli, B. *et al.* Histone demethylase JARID1C inactivation triggers genomic instability in sporadic renal cancer. *Journal of Clinical Investigation* **125**, 4625–4637 (2015).

33. Peric-Hupkes, D. *et al.* Molecular Maps of the Reorganization of Genome-Nuclear Lamina Interactions during Differentiation. *Molecular Cell* **38**, 603–613 (2010).

34. Hiratani, I. *et al.* Global reorganization of replication domains during embryonic stem cell differentiation. *PLoS Biology* **6**, 2220–2236 (2008).

35. Marchal, C. *et al.* Genome-wide analysis of replication timing by next-generation sequencing with E/L Repli-seq. *Nature Protocols* **13**, 819–839 (2018).

36. Rondinelli, B. *et al.* H3K4me3 demethylation by the histone demethylase KDM5C/JARID1C promotes DNA replication origin firing. *Nucleic Acids Research* **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. *Journal of Neuroscience* **30**, 4428–4439 (2010).

39. Setty, M. *et al.* Characterization of cell fate probabilities in single-cell data with Palantir. *Nature Biotechnology* **37**, 451–460 (2019).

40. Wang, C. *et al.* Reprogramming of H3K9me3-dependent heterochromatin during mammalian embryo development. *Nature Cell Biology* **20**, 620–631 (2018).

41. Nicetto, D. & Zaret, K. S. Role of H3K9me3 heterochromatin in cell identity establishment and maintenance. *Current Opinion in Genetics and Development* **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. *Nature Cell Biology* **22**, 767–778 (2020).

43. Novo, C. L. *et al.* The pluripotency factor Nanog regulates pericentromeric heterochromatin organization in mouse embryonic stem cells. *Genes and Development* **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. *Chemometrics and Intelligent Laboratory Systems* **58**, 109–130 (2001).

46. Eferl, R. *et al.* Development of pulmonary fibrosis through a pathway involving the transcription factor Fra-2/AP-1. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 10525–10530 (2008).

47. Soares, E. & Zhou, H. Master regulatory role of p63 in epidermal development and disease. *Cellular and Molecular Life Sciences* **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. *Proceedings of the National Academy of Sciences of the United States of America* **89**, 38–42 (1992).

50. Lombardi, L. M. *et al.* MECP2 disorders : from the clinic to mice and back Find the latest version : MECP2 disorders : from the clinic to mice and back. *Journal of Clinical Investigation* **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 playa redundant function in neural differentiation. *PLoS ONE* **4**, (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 Research* **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. *Nature Communications* **10**, 1–10 (2019).
57. Wu, S. J. *et al.* Single-cell analysis of chromatin silencing programs in development and tumor progression. *bioRxiv* 2020.09.04.282418 (2020) doi:10.1101/2020.09.04.282418.
58. Stadhouders, R. *et al.* Transcription factors orchestrate dynamic interplay between genome topology and gene regulation during cell reprogramming. *Nature Genetics* **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. *Current Opinion in Genetics and Development* **64**, 21–25 (2020).

61. Li, D. *et al.* Chromatin Accessibility Dynamics during iPSC Reprogramming. *Cell Stem Cell* **21**, 819-833.e6 (2017).

62. Schwarz, B. A. *et al.* Prospective Isolation of Poised iPSC Intermediates Reveals Principles of Cellular Reprogramming. *Cell Stem Cell* **23**, 289-305.e5 (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.e9 (2019).

64. Lin, C., Ding, J. & Bar-Joseph, Z. Inferring TF activation order in time series scRNA-Seq studies. *PLoS Computational Biology* **16**, 1–19 (2020).

## Acknowledgements

We thank all the members of the COSR and Tonon laboratory for discussions, support and for critical reading of the manuscript. We are grateful to Elena Brambilla and Francesca Ruffini for the preparation of the iPSC and NPC cells, and Dr. Alessia Mira for assistance in the preparation of the organoids. We would like to thank Stefano de Pretis for the thoughtful discussions about chromatin velocity. We are grateful to Gabriele Bucci for providing raw exome sequencing data and Paolo Dellabona for the coordination of the metastatic colon cancer sample collection and analysis. We also thank Drs. Gabellini, Bianchi, Agresti and Biffo for helpful discussions and for reviewing the manuscript. AB and LT are members of the EurOPDX Consortium. This work was partially supported by the Italian Ministry of Health with Ricerca Corrente and 5x1000 funds (SM and SP), by AIRC, Associazione Italiana per la Ricerca sul Cancro, Investigator Grants 20697 (to AB) and 22802 (to LT), AIRC 5x1000 grant 21091 (to AB and LT), AIRC/CRUK/FC AECC Accelerator Award 22795 (to LT), European Research Council Consolidator Grant

724748 – BEAT (to AB), H2020 grant agreement #754923 COLOSSUS (to LT), H2020 INFRAIA grant agreement #731105 EDIReX (to AB), Fondazione Piemontese per la Ricerca sul Cancro-ONLUS, 5x1000 Ministero della Salute 2014, 2015 and 2016 (to LT), AIRC investigator grant (to GT) and by the Italian Ministry of Health with 5x1000 funds, Fiscal Year 2014 (to GT), AIRC5x1000 ID. 22737 (to GT) and the AIRC/CRUK/FC AECC Accelerator Award "Single Cell Cancer Evolution in the Clinic", A26815 (AIRC number programme 2279)(to GT).

## Author contributions

M.T. performed experiments and analyzed the data. F.G. devised the methodology and experimental design, performed experiment and analyzed data. D.L. devised the methodology. V.G. performed bioinformatic analysis. D.R. performed experiments and provided experimental assistance and expertise. L.R. performed bioinformatic analysis. S.M. performed cloning and transposases production. I.C. and E. Z. performed in vivo experiments. O.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 fibroblast reprogramming and iPSC differentiation experiments. A.B. designed and supervised *in vivo* experiments and reviewed the data. G.V.M. designed and supervised the fibroblast 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 transposases 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.

# Figure Legends

**Figure 1 - Tn5 transposon is able to target H3K9me3-enriched regions. a,** Enrichment profile of H3K4me3 (green) and H3K9me3 (red) -associated regions obtained by ChIP-seq compared to Tn5 (green) and TnH (red) tagmentation profile obtained by ATAC-seq. ChIP-seq input track is shown as control (violet). **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 log2(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 scheme of the GET-seq transposon structure. Standard Tn5ME-A oligo was replaced by 49 nt oligos composed by 22 nt for Read 1 sequencing primer binding, 8 nt tags to discriminate Tn5 from TnH tagmentation products, and standard 19-bp ME sequence for transposase binding (created with BioRender.com). Data shown refer to experiments performed on Caki-1 cells.

**Figure 2 - Assessment of scGET-seq strategy and genomic copy number at the single-cell level. a,** UMAP embedding showing individual cells in a mixture of Caki-1/HeLa at known proportions (80:20) profiled by scGET-seq. Cells are identified according to a signature calculated on specific DHS identified from bulk studies. **b,** UMAP embedding showing individual cells in a mixture of Caki-1/HeLa 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 between the segmentation profile of Caki-1 and HeLa cells at increasing resolution. Signal from bulk sequencing is compared to average cell signal obtained in single cell profiling. scGET-seq (orange) shows consistently higher correlation compared to standard scATAC-seq (blue). **d,** Spearman's correlation between the segmentation profiles and the density of regulatory elements in the GeneHancer catalog. White dot in boxplots reprents the median, boxes span between the $25^{th}$ and $75^{th}$ percentiles, whiskers extend 1.5 times the interquartile range. n=323 regions. **e,** Heatmap 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.

**Figure 3 – Analysis of Patient Derived Samples by scGET-seq a**, segmentation profile in individual cells profiled by scGET-seqof two PDO (CRC6 and CRC17). The heatmaps 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 1Mb windows with respect to the average per-cell coverage. Centromeric regions and genome gaps were excluded from the analysis and colored in white. Barplots on top of each heatmap 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, the solid line represents the geometric mean. Sensitivity is calculated as TP/(TP + FN), Precision is calculated as TP/(TP + FP), where TP = alleles correctly identified; 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 (panel a) or epigenome analysis (panel b). Below each UMAP embedding, a barplot represents the abundance of subpopulations over time.; f Scatterplot 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 panel d; grey dots are mutations shared by two clones (Pearson $r$=0.712, p=7.93e-38, $n$=389); g Representative mutations of COSMIC Hallmark genes found in scGET-seq data which were not present in the primary tumor. Each mutation is associated to the corresponding genetic clone using the appropriate color code.

**Figure 4 - - scGET-seq profiling of NIH-3T3 cells knocked-down for Kdm5c. a** UMAP embedding showing the location of cells transfected with shKdm5c or shScr. **b,** UMAP embedding of individual cells colored by the 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, panel c, or Tn5, panel d. The two signals are unevenly distributed and tend to localize where higher amounts of shScr cells are. All these data refer to experiments performed on NIH-3T3 cell line.

**Figure 5 – scGETseq defines cell identity and developmental trajectories of FIB, iPSC and NPC. a,** UMAP embedding showing scGET-seq profiling of human fibroblasts (FIB), induced Pluripotent Stem Cells (iPSC) and Neural Precursor Cells (NPC). Black arrow shows a small subset of FIB and NPCs clustering alongside iPSC. **b,** UMAP embedding showing scRNA-seq profiling of the same cell populations derived from the same samples as in panel a. **c,** the 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 types as in panels 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,** heatmap 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, 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 to regions depicted in panel. The red arrows show the subsets of NPC and FIB that share similar features with iPSC.

**Figure 6 - Chromatin velocity**. **a,** UMAP embedding of differentiating single cells profiled by scGET-seq. Cells are colored by velocity pseudotime, 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, arrow streams indicate the RNA velocity extracted using scvelo. **c,** Selected terms enriched for genes associated to 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 calculated for all PWM in 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 Fig. S8c. Arrows indicate the loading of the top 4 PWM in each quadrant. The colored contours indicate the density estimates of the three cell types. **g,** Heatmap shows average expression profiles of TF with the top 10 most negative on PLS1 during the early brain development. Darker color indicates higher expression. w.p.c.: weeks post conception.

# Online Methods

### CELL CULTURE

All established cell lines were purchased from American Type Culture Collection (ATCC), except for HEK293T cell line that was a kind gift from Prof. Luigi Naldini (San Raffaele Telethon Institute for Gene Therapy, Milan). 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 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 from 10,000,000 of Caki-1 cells crosslinked with 38% formadheide; fixation was stopped with 0.125 M glycine. Sonication was then performed on Covaris E220 with the following parameters: total time 6 min, 175 Peak Incident Power, 200 cycles per burst. 8 μg of sonicated chromatin was used as input for each experimental condition. No Antibody (No Ab), Ab anti-H3K9me3 (ab8898 Abcam), Ab anti-H3K4me3 (07-473 Millipore). ChIP-seq, performed as already described in [32], were used as reference for TAM-ChIP-seq (Ab anti-H3K9me3 (ab8898 Abcam) and Ab anti-H3K4me3 (07-473 Millipore) have been used).

### TAM-ChIP – qPCR

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. In TAM-ChIP-qPCR one of the two H3K4me3 biological replicate was excluded because no significant signal was detected for any condition. For each TAM-ChIP condition, 10 ng of final libraries were used as input. Water was used as negative control. Real time PCR 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[32]) and used at a final concentration of 400 nM. To determine the enrichment obtained, we normalized TAM-ChIP-qPCR data for No Ab sample. Primers are listed below.

| **Primer** | **Forward sequence** | **Reverse sequence** |
|---|---|---|
| FAM5B | GCGCCTTCCTTACTTCCATG | AGTGGCCATCTCATTTCCCA |
| NTF3 | AAAGGCCTTGGTCCCAGA | ATTGAAGGAACGCAGCCC |
| CACNA1E | GAGGGAGGAGAAAGCCGA | TTGTCCAGACCAGCCCTT |

**Tn5 transposase production**

Tn5 transposase was produced as previously described[65] using pTXB1-Tn5 vector (Addgene, Plasmid #60240). For hybrid transposases, the DNA fragment encoding human HP1α was derived from the pET15b-HP1α (pHP1α-pre) vector[66], kindly provided by Dr. Hitoshi Kurumizaka. According to the cloning strategy, 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. Construct amino acid sequences are detailed in Supplementary Data 1

**Transposon assembly**

Assembly of standard and modified pre-annealed Mosaic End Double-Stranded (MEDS) oligonucleotides, Tn5MEDS-A, Tn5MEDS-B and TnHMEDS-A was performed in solution following published protocol[67]. For single cell GET-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, Tn5ME-A.8) and other four sequences for TnHME-A (TnHME-A.4, TnHME-A.5, TnHME-A.9, TnHME-A.10). A Read 1 primer binding site was reconstituted adding 8 nt (TCCGATCT) upstream the Tn5/TnH tag.  Modified Tn5ME-A sequences are reported in Supplementary Data 1
Creation of functional transposon was performed following previously published protocol[65].

**Bulk tagmentation reaction and ATAC-seq**

Bulk tagmentation was performed on Caki-1 genomic DNA (gDNA) following published protocol[65]. Specifically, 500 ng of gDNA was incubated for 7 min at 55 °C with 1 µL of functional transposon in 1X TAPS-PEG8000 buffer in a final 20 µL volume. As control, a parallel reaction was carried out on Caki-1 gDNA but using the Nextera DNA Library Prep Kit according to the manufacturer's protocol. Reactions were stopped adding SDS at a final concentration of 0.05% and incubated for 5 min at room temperature (RT). 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 sec at 98 °C followed by 13 cycles of 45 sec at 98 °C, 30 sec at 55 °C, 30 sec at 72 °C . Libraries were then purified using 1X volume of Ampure XP beads (Beckman-Coulter) and checked for fragment distribution on TapeStation (Agilent).
ATAC-seq was performed following published protocols[9] with minor modifications.
Briefly, 100,000 Caki-1 cells pellets were washed in 100 µL cold 1X PBS, centrifuged for 10 min at 500 *g at 4 °C, and permeabilized in 100 µL of cold lysis buffer (10 mM Tris·Cl, pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% (v/v) 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 (5x TAPS-PEG8000 buffer mixed with 10 µL of 1.39 µM of functional transposon in a final volume of 100 µL). As control, a parallel reaction was carried out on 100,000 Caki-1 cells pellets 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 adding SDS at a final concentration of 0.05%. After 5 min of incubation at RT, reactions were purified using QIAquick Gel Extraction Kit (Qiagen) and eluted in 15 µL of EB buffer. 5 µL of this reaction was used as input for indexing PCR as described before.

Libraries were sequenced on Illumina platforms with 2x50 bp sequencing protocol.

**Single cell ATAC-seq and GET-seq**

Single-cell ATAC-seq was performed on 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 suspension was prepared in order to get 10,000 nuclei as target nuclei recovery.

Single cell GET-seq was performed as previously described but replacing the provided ATAC transposition enzyme (10X Tn5; 10X Genomics) with a sequential combination of Tn5 and TnH functional transposons, in the transposition mix assembly step. Specifically, a transposition mix contained 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 total of 1 h incubation.

When scGET-seq was performed on 20:80 proportion of HeLa:Caki-1 cells, nuclei suspension was prepared in duplicate in order to get 10,000 nuclei as target nuclei recovery for each replicate.

Final libraries were loaded on Novaseq6000 platform (Illumina) to obtain 50,000 reads/nucleus with 2x50 bp read length. For GET-seq, the sequencing target was 100,000 reads/nucleus; and a custom Read 1 primer was added to the standard Illumina mixture (5'-TCGTCGGCAGCGTCTCCGATCT-3').

**Single cell RNA-seq**

Single-cell RNA-seq was performed on 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 Novaseq6000 platform (Illumina) to obtain 50,000 reads/cells.

**Kdm5c Knock-Down experiment**

Lentiviral vectors were produced by transfecting HEK293T cells (a kind gift from Prof. Luigi Naldini, San Raffaele Telethon Institute for Gene Therapy, Milan) with pLK0.1 plasmid containing shRNAs targeting Kdm5c (shKdm5c, CCGGGCAGTGTAACACACGTCCATTCTCGAGAATGGACGTGTGTTACACTGCTTTT ) or scramble (shScr)[32].

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 VSV-G, 6.25 µg of REV, 15 ug of ADV plasmid, was prepared and filled up to 1125 µl with 0.1X TE/dH2O (2:1); after 30 min of incubation on rotation, 125 µl of 2.5 M $CaCl_2$ were added to the mix and, after 15 min of incubation, the precipitate was formed by dropwise addition of 1,250 µl of 2X 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 fresh medium/dish supplemented with 16 µl of NAB/dish. After 30 h the medium containing viral particles was collected, filtered with 0.22 µm filter and and stored at –80 °C in small aliquots to avoid freeze-thaw cycles.

NIH-3T3 cells were transduced in 6 well-plate format. To this end, 2 ml of shKdm5c/shScr lentiviral vector supplemented with Polibrene (final concentration 8 µg/ml) were added to actively cycling (50% confluency) NIH-3T3; one well of untransduced cells was used as negative control. After 24 h transduced cells were splitted in a 10 cm dish and Puromycin selection (final concentration 4 µg/ml) was performed. 48 h post selection half of transduced cells were detached, washed twice with cold 1X PBS and tested for gene knock-down by Real Time (RT)-PCR as described below. Upon validation of knock-down, 72 h post selection, all the remaining cells were collected and subjected to scGET-seq as already described. Nuclei suspension was prepared in order to get 10,000 nuclei as target nuclei recovery.

**Gene Knock-down validation by Real Time (RT)-qPCR**

Total RNA was isolated using Trizol (Invitrogen, Carlsbad, CA, USA) and purified using RNeasy mini kit (Qiagen); cDNA was generated using 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). 10 ng of cDNA were used as input, water was used as negative control. Amplification was performed using previously validated primers[32] and used at a final concentration of 400 nM except for major that were used 200 nM. Primers for minor ncRNA were taken from [68] and were used at a final concentration of 400 nM.

**Patient-derived colorectal cancer organoids (PDOs)**

Samples from 2 patients with liver metastatic gastrointestinal cancers were obtained upon written informed consent, in line with protocols approved by the San Raffaele Hospital Istitutional Review Board, and following procedures in accordance with the Declaration of Helsinki of 1975, as revised in 2000. PDOs cultures were established as previously reported[69]. Briefly, fresh tissues were minced immediately after surgery, conditioned in PBS/5mM EDTA and digested for 1h at 37°C in a solution composed of 2X TrypLE™ Select Enzyme (Thermofisher) in PBS/1mM EDTA with DNAse I (Merck) addition.. Release of the cells was facilitated by pipetting. Dissociated cells were collected, suspended in 120µl growth factor reduced (GFR) Matrigel™ (Corning™ 356231, FisherScientific), seeded in single domes in 24-well flat bottom cell culture plate (Corning) and, after dome solidification, covered with 1ml of complete human organoid medium[69] and medium replaced every two/three days. For scGET-seq analysis PDOs were dissociated to single cells by combining mechanical (pipetting) and enzymatic digestion after 20 min incubation at 37 °C in a solution of 1X TrypLE™ Select Enzyme in PBS/1mM EDTA, washed in 1X PBS and processed as previously described.

**Patient-derived colorectal cancer xenografts (PDXs)**

*Specimen collection and annotation* - EGFR blockade responsive colorectal cancer and matched normal samples were obtained from one patient that underwent liver metastasectomy at the Azienda Ospedaliera Mauriziano Umberto I (Torino). The patient 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 NOD (nonobese diabetic)/SCID (severe combined immunodeficient) mice as previously described[69]. Once tumors reached an average volume of ~400 mm3, mice were randomized into 4 treatment arms that received either placebo or cetuximab (Merck, 20 mg/kg twice weekly, intraperitoneally) as follows: i) untreated; ii) cetuximab 72 hours; iii) cetuximab 4 weeks; iv) cetuximab 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 case of untreated tumors; n = 2 for 72 hours; 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. In order 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 $4/3\pi \cdot (d/2)2 \cdot D/2$, 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 (DOI 10.1007/s10916-012-9891-6). Animal procedures were approved by the Italian Ministry of Health (authorization 806/2016-PR).

*Single cell GET-seq on PDX*A - At the end of treatments, mice were sacrificed and tumors 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 single-cell GET-seq as already described. Nuclei suspension was prepared in order to get 10,000 nuclei as target nuclei recovery for each replicate.

### Fibroblast reprogramming towards iPSC and iPSC differentiation towards NPC

Dermal fibroblasts (FIB) obtained from skin biopsies of two different healthy subjects (A and B) were cultured in fibroblast medium and reprogrammed with the Sendai virus technology (CytoTune-iPS Sendai Reprogramming Kit, ThermoFisher, Waltham, MA, USA) to generate Human induced pluripotent Stem Cells (iPSC) clones. iPSC clones were individually picked, expanded and maintained in mTeSR1 on hESC-qualified Matrigel. Human iPSC-derived neural progenitor cells (NPC) were generated following the standard protocol based on a dual-smad inhibition[70]. Briefly, iPSCs were differentiated in NPC 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 GSK3b 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 FIB, iPSC and NPC derived from patient 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 HMM model. Data preprocessing pipeline is available at https://github.com/leomorelli/scGET
Reads for ChIP-seq, GET-seq, scGET-seq experiments were aligned to reference genome (hg38 or mm10) using bwa mem v0.7.12[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 BPM normalization. H3K4me3 enriched regions were identified using `MACS` v2.2.7[76], H3K9me3 enriched regions were identified using `SICER` v2[77], using default parameters.

**Definition of epigenome reference sets**

We segmented the genome according to DNAseI Hypersensitive Sites (DHS), as previously described[78]. Briefly, we downloaded the index of DHS for human[79] and mouse genome[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 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 `scatACC` repository. Each count matrix was processed using `scanpy` v1.4.6 or v1.6.0[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 covariate for linear regression and data were then scaled with a maximum value set to 10. Neighborhood was evaluated using Batch balanced KNN[82], cell groups were identified with Leiden algorithm[83] for cell lines or `schist`[84] choosing the hierarchy level that maximizes modularity. In order to extract a unique representation of four datasets, we applied graph fusion using `scikit-fusion`[85]: we first extracted a 20-components UMAP reduction of each view, then we built a relation graph where all views are connected to a 20-components Latent Space (LS). Matrix factorization was run with 1,000 iterations 5 times. The resulting LS 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 10 datasets (4 depicted in Figure 2a and 6 randomly chosen) at different proportions (0.1x, 0.2x, 0.5x) 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 $s = 0.815 \cdot N_{cells} + 0.406 \cdot (1-d) + 0.2316$, where $N_{cells}$ is the number of cells divided by 1000 and $d$ is the fraction of duplicated reads.

**Analysis of HeLa/Caki-1 cell identity**

To identify cell identity in Caki-1/HeLa mixture, we downloaded publicly available bulk ATAC-seq 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 using `edgeR`[87] using RLE normalization and contrasting HeLa vs Caki-1 by exact test. We selected HeLa specific regions by filtering for FDR < 1e-3, logCPM > 3 and logFC > 0 (*i.e.* 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.
2-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 log2-ration between early stage (E) and late stage (L) was calculated. Entries in Tn5-dhs list were assigned the average log2(E/L) value over its interval.
LaminB1 DamID data for NIH-3T3 cells were also downloaded from UCSC genome browser tables, converted to bigwig format and lifted over mm10 assembly coordinates using `Crossmap`[88]. Average value of LaminB1 data over Tn5-dhs regions was assigned as described above.
Differences in distribution of log2(E/L) and LaminB1 values were evaluated by Mann-Whitney U-test.

**Analysis of Copy Number Alterations**

Copy Number Alteration were derived from TnH data quantified over the entire genome, binned at 5 kbp resolution. Counts were extracted using `peak_count.py` script from the `scatACC` repository. After that, data were processed by collapsing values into larger bins at different resolutions (10 Mb, 1Mb, 500 kb). The value of each bin is divided by the average per-cell read count; we apply linear regression of per-bin GC content and mappability [89], and finally express values as log2 of the scaled residuals. Cell clustering was performed using `schist` applied on the kNN graph built with `bbknn` and using correlation as 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 CNV in Caki-1:HeLa cells**

We created a ground truth dataset by calling copy number alterations in Caki-1 and HeLa cells with Control-FREEC [89] on Whole Genome Sequencing data. We binned the resulting segments according to the desired resolution in single cell experiments (10Mb, 1Mb and 500kb), 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 log2ratio matrices into a training and a test set in 70:30 proportion. We trained a Logistic Regression classifier and a Support Vector Machine with the one-vs-rest strategy and increasing 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 organoids Whole Exome Sequencing data**

Reads were aligned to hg38 reference genome using bwa, reads were then processed using bwa. Alignment 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 CNV was performed using CNAqc[91], clonal deconvolution was performed using MOBSTER and BMix [92] with default parameters.

**Analysis of mutations**

Reads for Tn5 and TnH data were separated to individual BAM files using `separate_bam.py` script from the `scatACC` repository. Known somatic mutations were genotyped using `freebayes` v.1.3.2 [93] (parameters: -@ exome_somatic.vcf.gz -C 2 -F 0.01). Only variants with depth > 1 were then considered for the analysis.
Variant calling without priors was performed using `freebayes` using the same thresholds. VCF files were annotated using `snpEff` v4.3p[94] using GRCh38.86 annotation model. Known cancer variants were annotated using COSMIC catalog[95]. Variants were then filtered for depth > 10, 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 95-th percentile were considered as marker regions.

**Analysis of scRNA-seq data**

Reads were demultiplexed using cellranger (v4.0.0). Identification of valid cellular barcodes and UMIs was performed using `umitools` with default parameters for 10x v3 chemistry. Reads were aligned to 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[98]. Count matrices were imported into `scanpy`, doublet rate was estimated using scrublet[99]. Count matrix was filtered (min_genes = 200, min_cells=5, pct_mito<20) before normalization and log-transformation. kNN graph was built using `bbknn`. RNA velocity was estimated using `scvelo` dynamical modeling with latent time regularization.

**Total Binding Affinity analysis**

For each DHS region selected for likelihood, we extracted the 500bp sequence flanking summits there included, as annotated in the DHS index. We downloaded the HOCOMOCO v11 list of PWM was downloaded[100] and calculated the Total Binding Affinity as defined in[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 `PLSCanonical` function from the python `sklearn.cross_decomposition` library, using cell groups as targets for the matrix transformation.

**References for the Methods section**

65. Picelli, S. *et al.* Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Research* **24**, 2033–2040 (2014).
66. Machida, S. *et al.* Structural Basis of Heterochromatin Formation by Human HP1. *Molecular Cell* **69**, 385-397.e8 (2018).
67. Reznikoff, W. S. Transposon Tn5. *Annual Review of Genetics* **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 Discovery* **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 Research* **27**, 491–499 (2017).
72. Lassmann, T. TagDust2: A generic method to extract reads from sequencing data. *BMC Bioinformatics* **16**, 1–8 (2015).
73. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* **1303.3997v2**, (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 Research* **42**, 187–191 (2014).

76. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biology* **9**, R137 (2008).

77. Breeze, C. E. *et al.* Atlas and developmental dynamics of mouse DNase I hypersensitive sites. *bioRxiv* (2020) doi:10.1101/2020.06.26.172718.

78. Giansanti, V., Tang, M. & Cittaro, D. Fast analysis of scATAC-seq data using a predefined set of genomic regions. *F1000Research* **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*. *Current Protocols in Bioinformatics* (2014). doi:10.1002/0471250953.bi1112s47.

81. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome biology* **19**, 1–5 (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. *Scientific Reports* **9**, 1–12 (2019).

84. Morelli, L., Giansanti, V. & Cittaro, D. Nested Stochastic Block Models Applied to the Analysis of Single Cell Data. *bioRxiv* (2020) doi:10.1101/2020.06.28.176180.

85. Žitnik, M. & Zupan, B. Data fusion by matrix factorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **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.e22 (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 research* **46**, e120 (2018).

90. Favero, F. *et al.* Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Annals of oncology : official journal of the European Society for Medical Oncology* **26**, 64–70 (2015).

91. Househam, 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. *bioRxiv* 2021.02.13.429885 (2021) doi:10.1101/2021.02.13.429885.

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. *arXiv:1207.3907 [q-bio]* (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 Research* **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. *BioRxiv* (2021) doi:10.1101/2021.05.05.442755.
98. Harrow, J. *et al.* GENCODE: The reference human genome annotation for the ENCODE project. *Genome Research* **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 systems* **8**, 281-291.e9 (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 Research* **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. *Molecular Biology and Evolution* **28**, 2173–2183 (2011).
102. Morelli, L. & Cittaro, D. *scGET: pre-release of scGET repository*. (Zenodo, 2021). doi:10.5281/zenodo.5095040.
103. Cittaro, D. *scatACC: Version 0.1*. (Zenodo, 2021). doi:10.5281/zenodo.5095157.

# 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.

## Code availability

Code necessary to preprocess scGET-seq data is available at https://github.com/leomorelli/scGET[102] and https://github.com/dawe/scatACC [103]. Illustrative code snippets for post processing are reported in Supplementary Data S2.