nature biotechnology

ARTICLES

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

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).

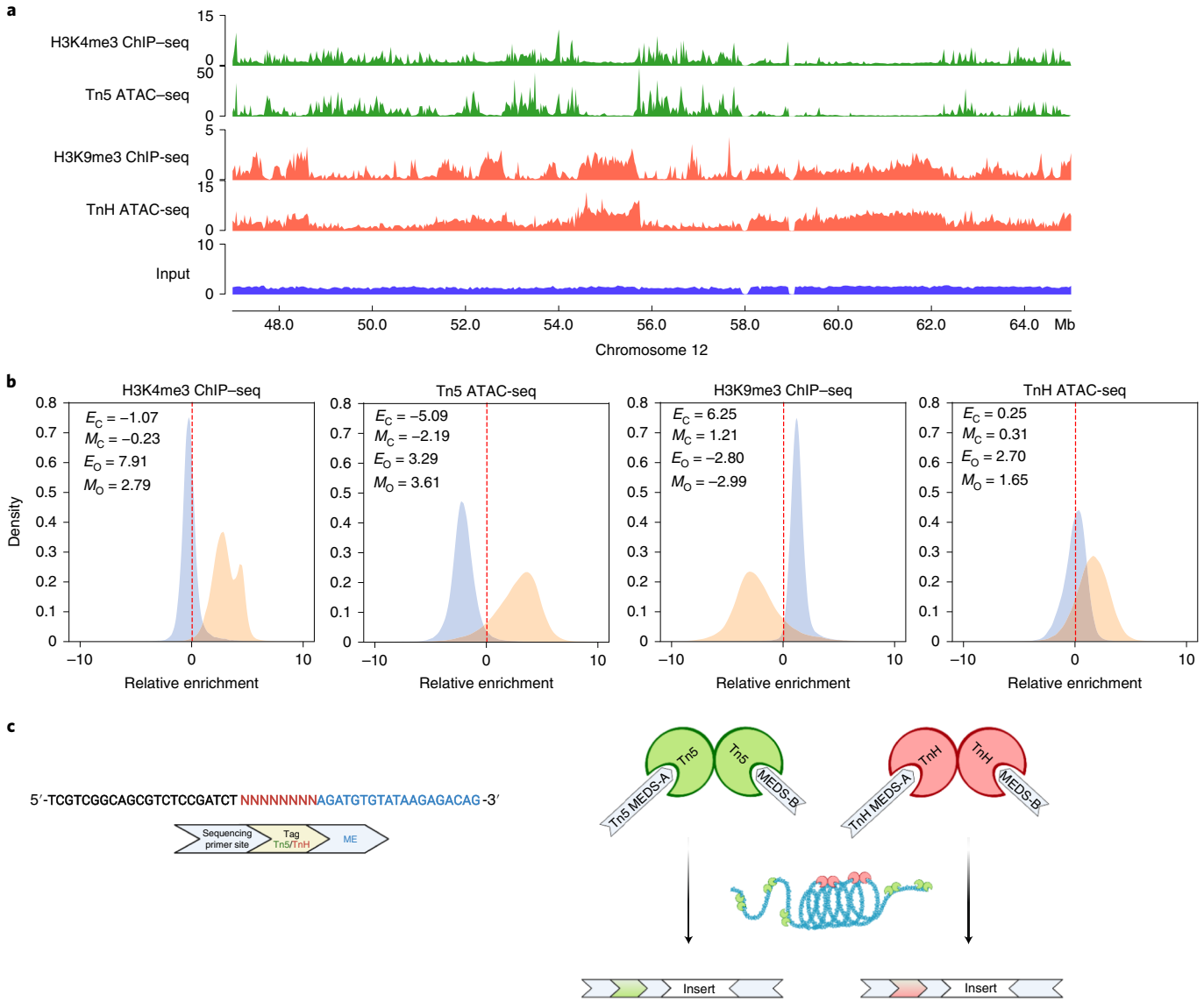


Fig. 1 | The Tn5 transposon is able to target H3K9me3-enriched regions. a, Enrichment profile of H3K4me3-associated (green) and H3K9me3-associated (red) regions obtained by ChIP-seq compared to Tn5 (green) and TnH (red) tagmentation profile obtained by ATAC-seq. The ChIP-seq input track is shown as a 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-

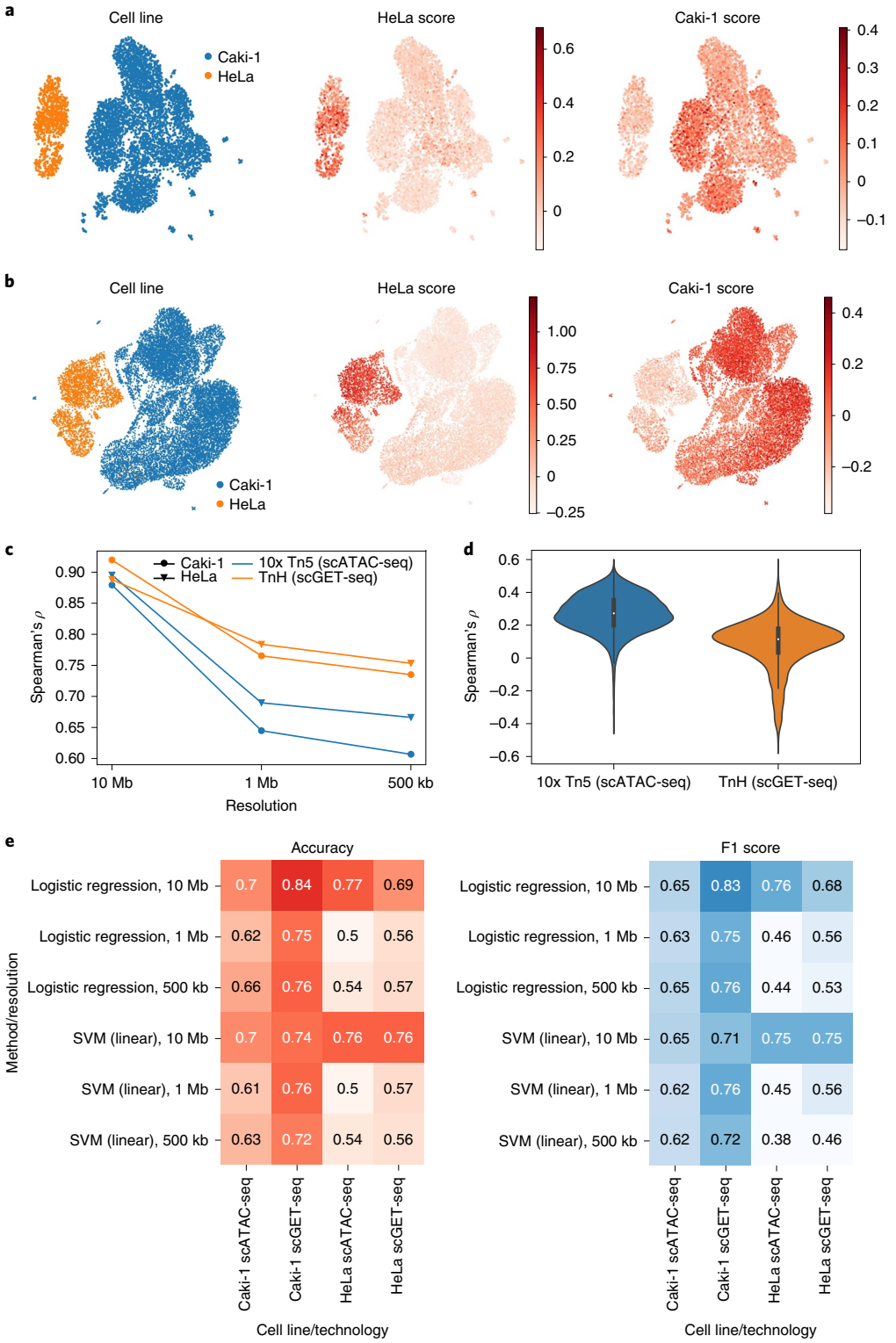


Fig. 2 | Assessment of scGET-seq strategy and genomic copy number at the single-cell level. a, Uniform manifold approximation and projection (UMAP) embedding showing individual cells in a mixture of Caki-1 and HeLa cells at known proportions (80:20) profiled by scGET-seq. Cells are identified according 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.

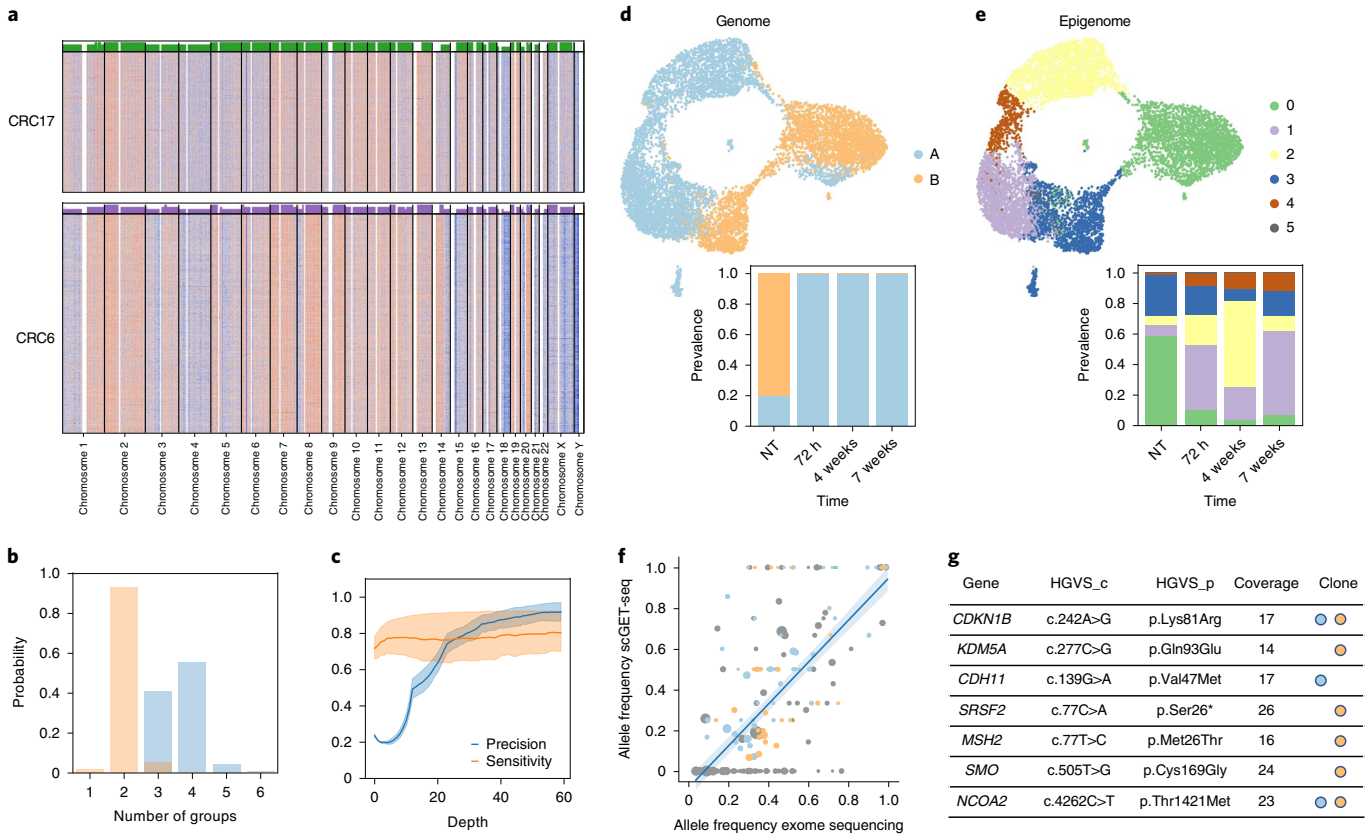


Fig. 3 | Analysis of human-derived samples by scGET-seq. a, Segmentation profile in individual cells profiled by scGET-seq of two PDOs (CRC6 and CRC17). The heat maps show the genomic landscape of two discovered clones assigned to each organoid. scGET-seq data are expressed as normalized $\log_2$ (ratio) 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

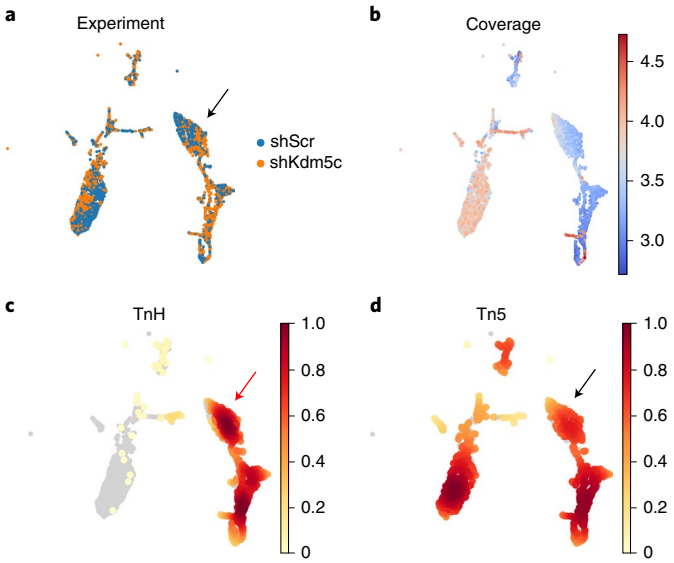


Fig. 4 | scGET-seq profiling of NIH-3T3 cells after *Kdm5c* knockdown. **a**, UMAP embedding showing the location of cells transfected with shKdm5c or shScr. **b**, UMAP embedding of individual cells colored by read coverage. Two main clusters appear depending on the coverage. **c**,**d**, UMAP embedding 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

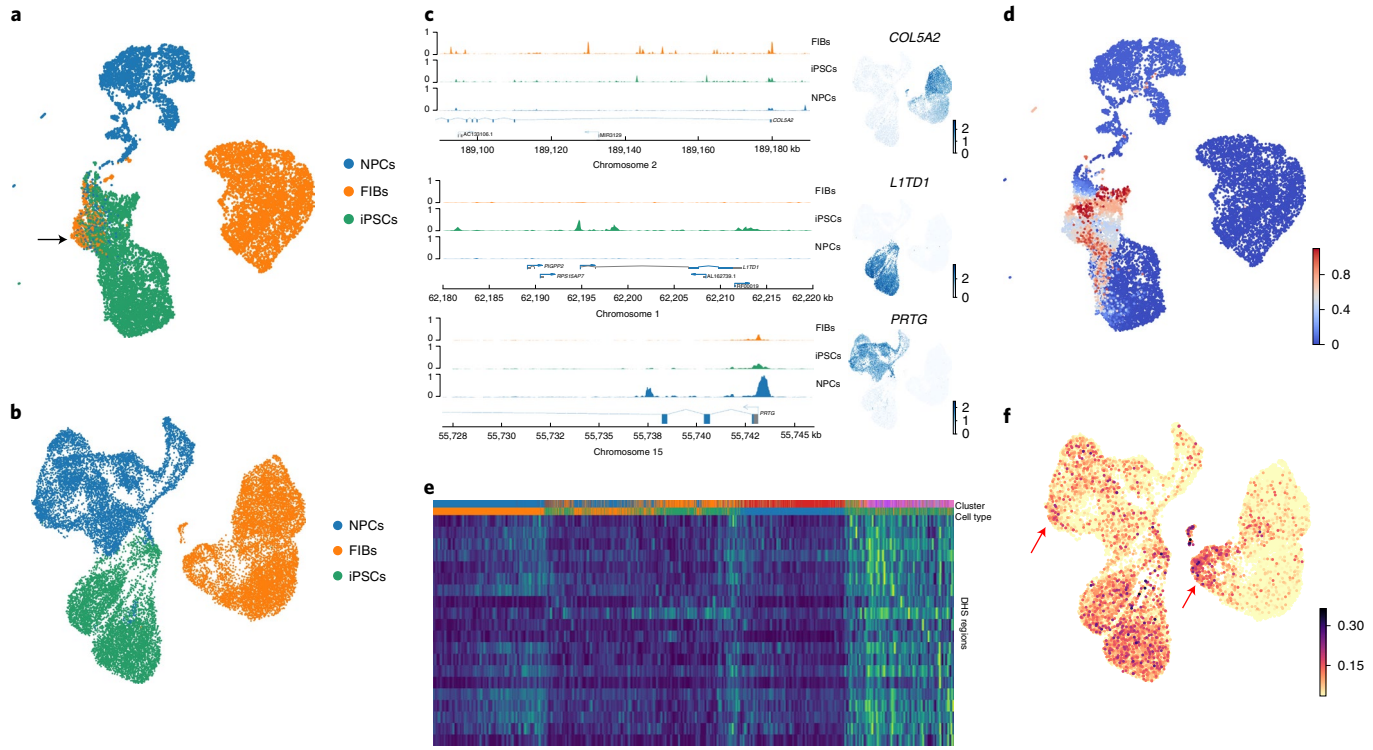


Fig. 5 | scGET-seq defines cell identity and developmental trajectories of FIBs, iPSCs and NPCs. a, UMAP embedding showing scGET-seq profiling of human FIBs, iPSCs and NPCs. The black arrow shows a small subset of FIBs and NPCs clustering alongside iPSCs. **b**, UMAP embedding showing scRNA-seq profiling 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

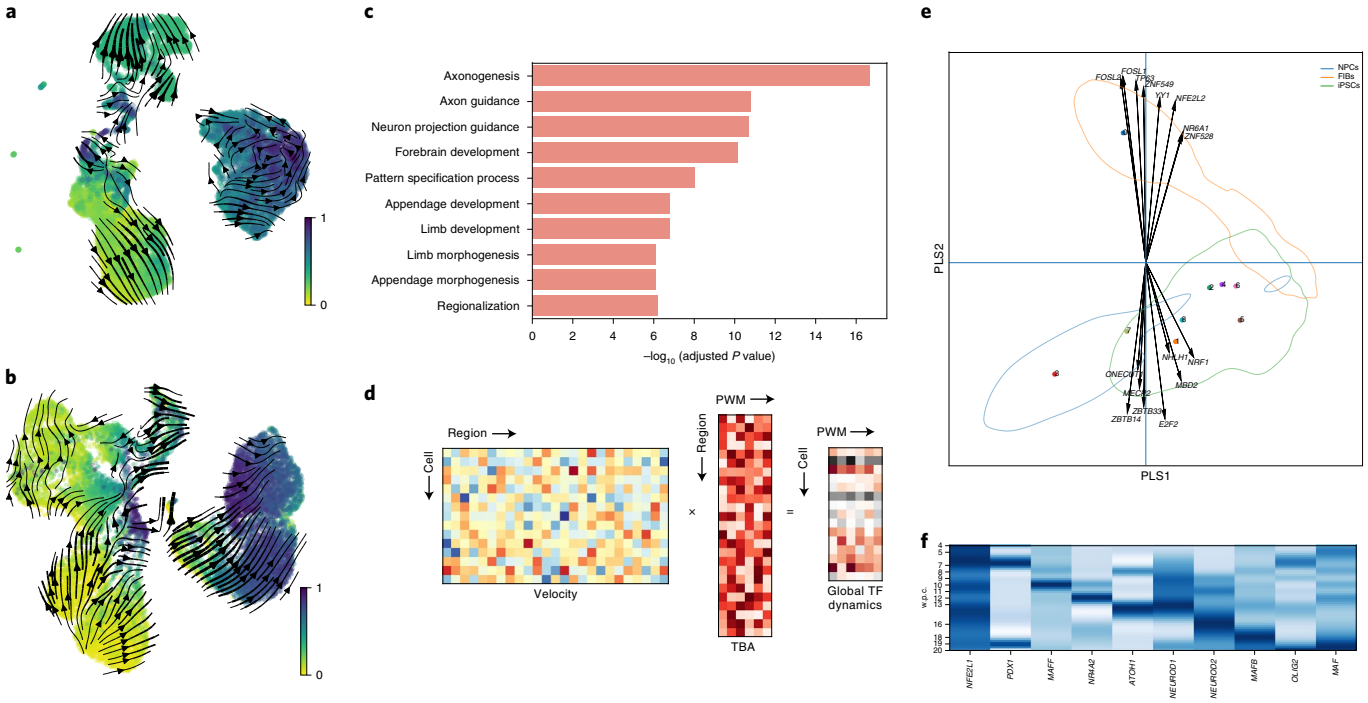


Fig. 6 | Chromatin Velocity. a, UMAP embedding of differentiating single cells profiled by scGET-seq. Cells are colored by velocity pseudotime, and arrow streams indicate the Chromatin Velocity extracted using scvelo. **b**, UMAP embedding of differentiating single cells profiled by scRNA-seq. Cells 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).

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

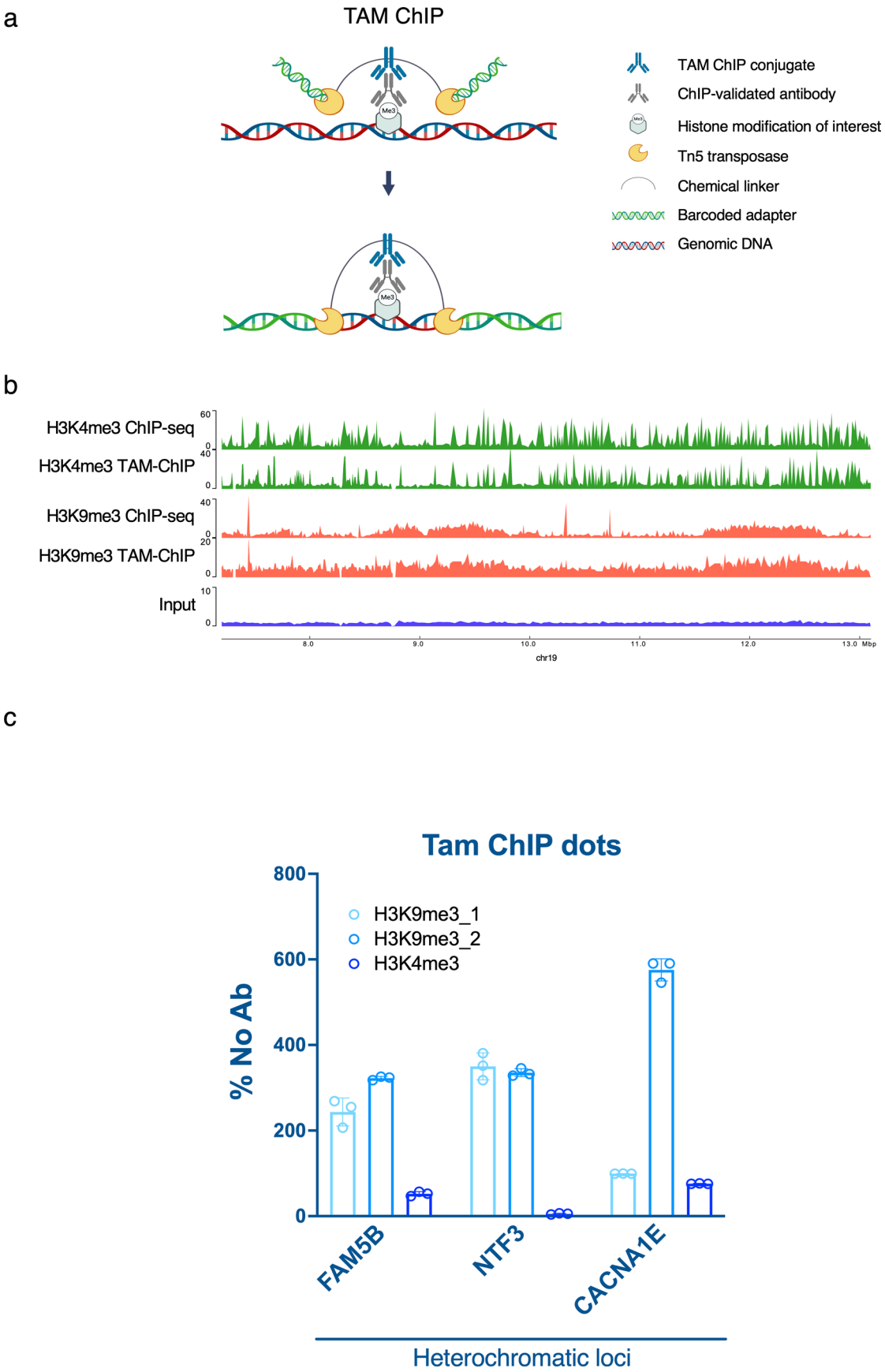
Extended data is available for this paper at https://doi.org/10.1038/s41587-021-01031-1.

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.

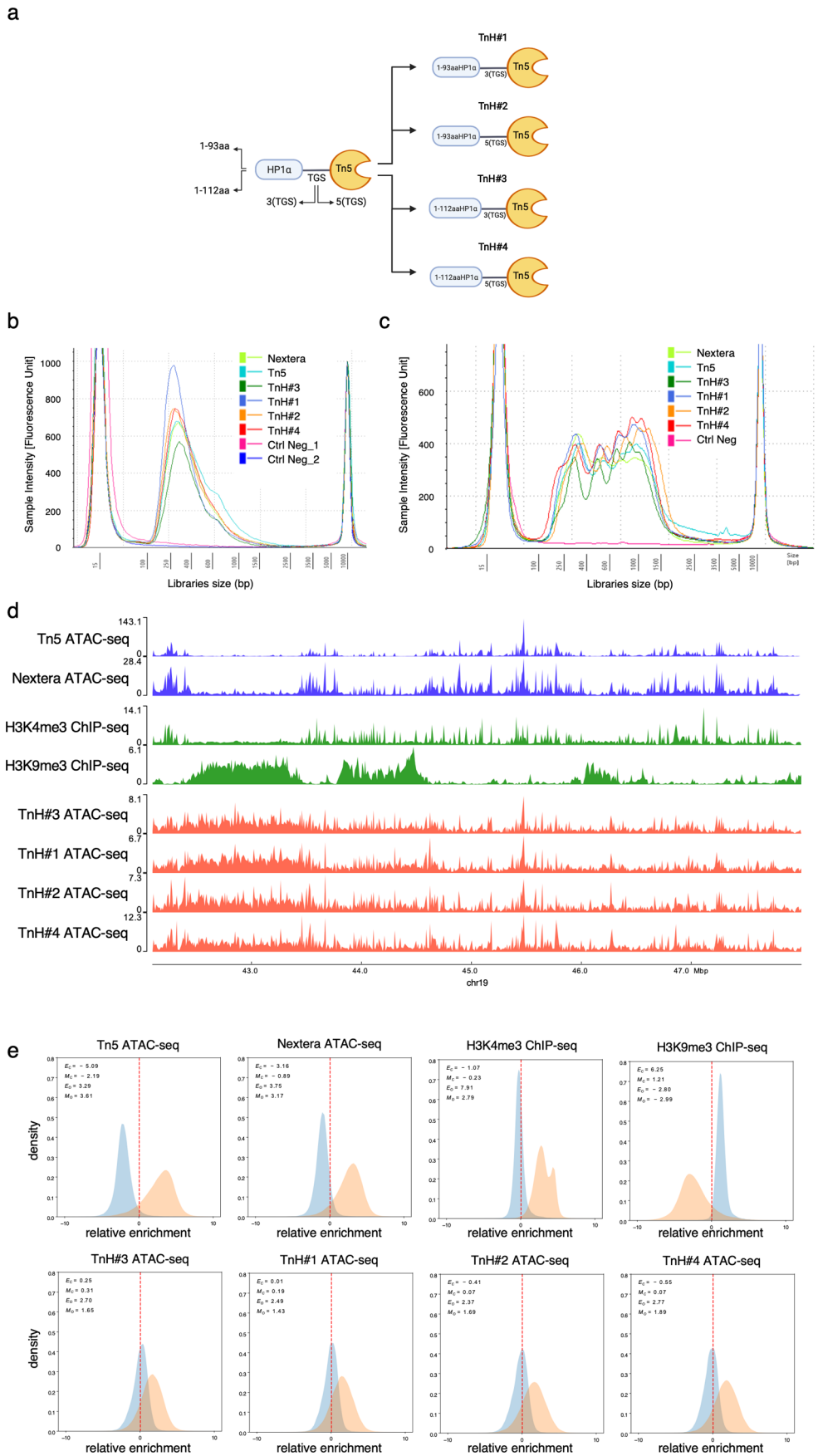
Peer review information *Nature Biotechnology* thanks Kun Zhang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work

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



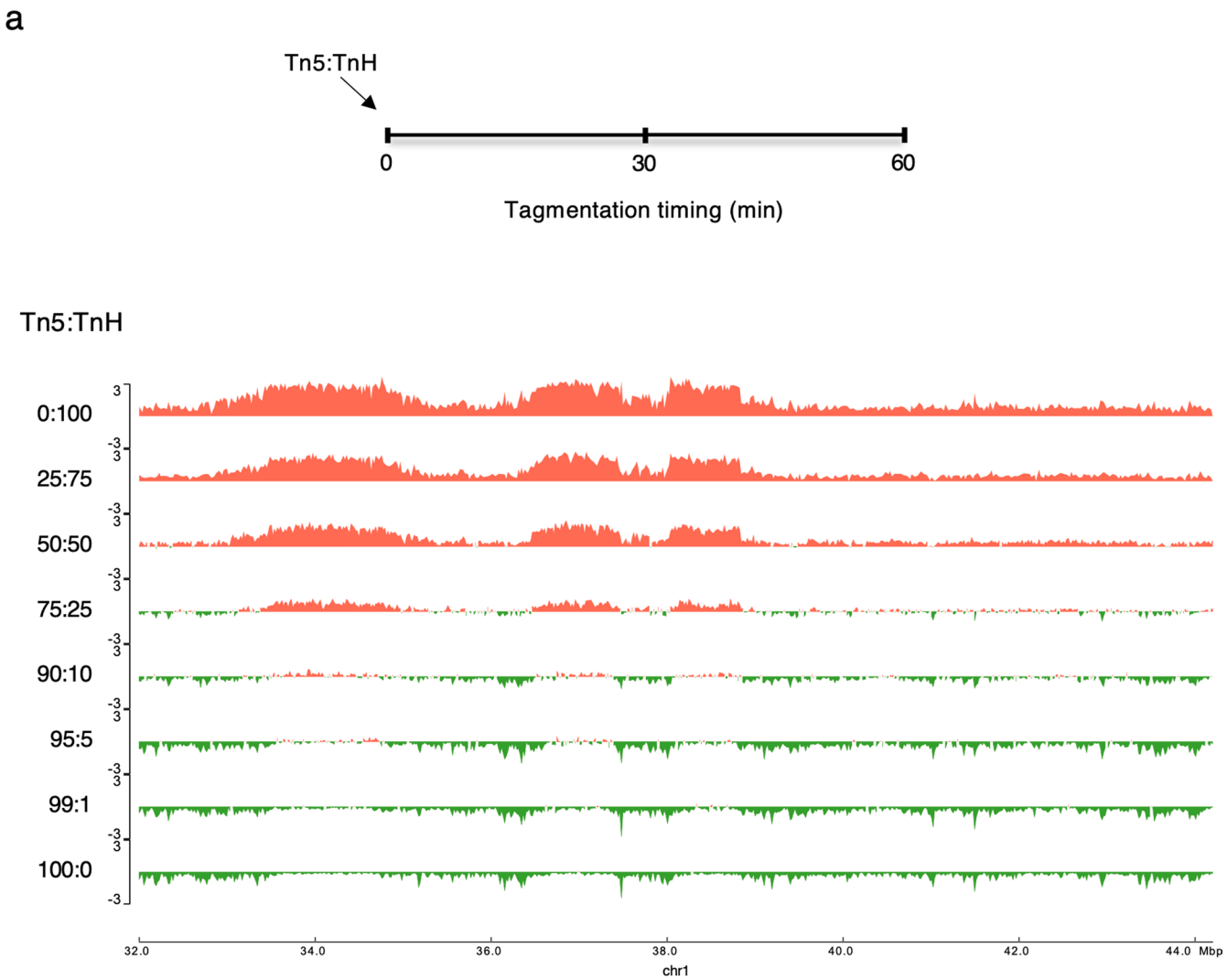
Extended Data Fig. 1 | See next page for caption.

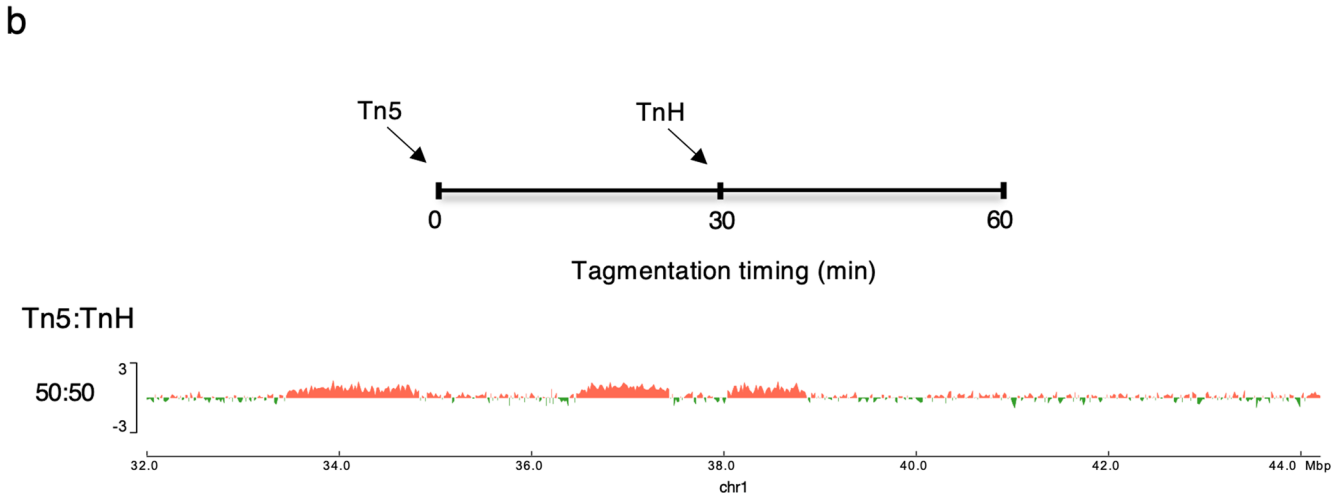
Extended Data Fig. 1 | Tn5 transposase is able to tagment compacted chromatin featuring H3K9me3. a, General scheme of TAM-ChIP technique (created with BioRender.com). A primary antibody (ChIP-validated antibody, dark grey) binds to a specific histone modification (light grey) over the genome (blue-red). 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.

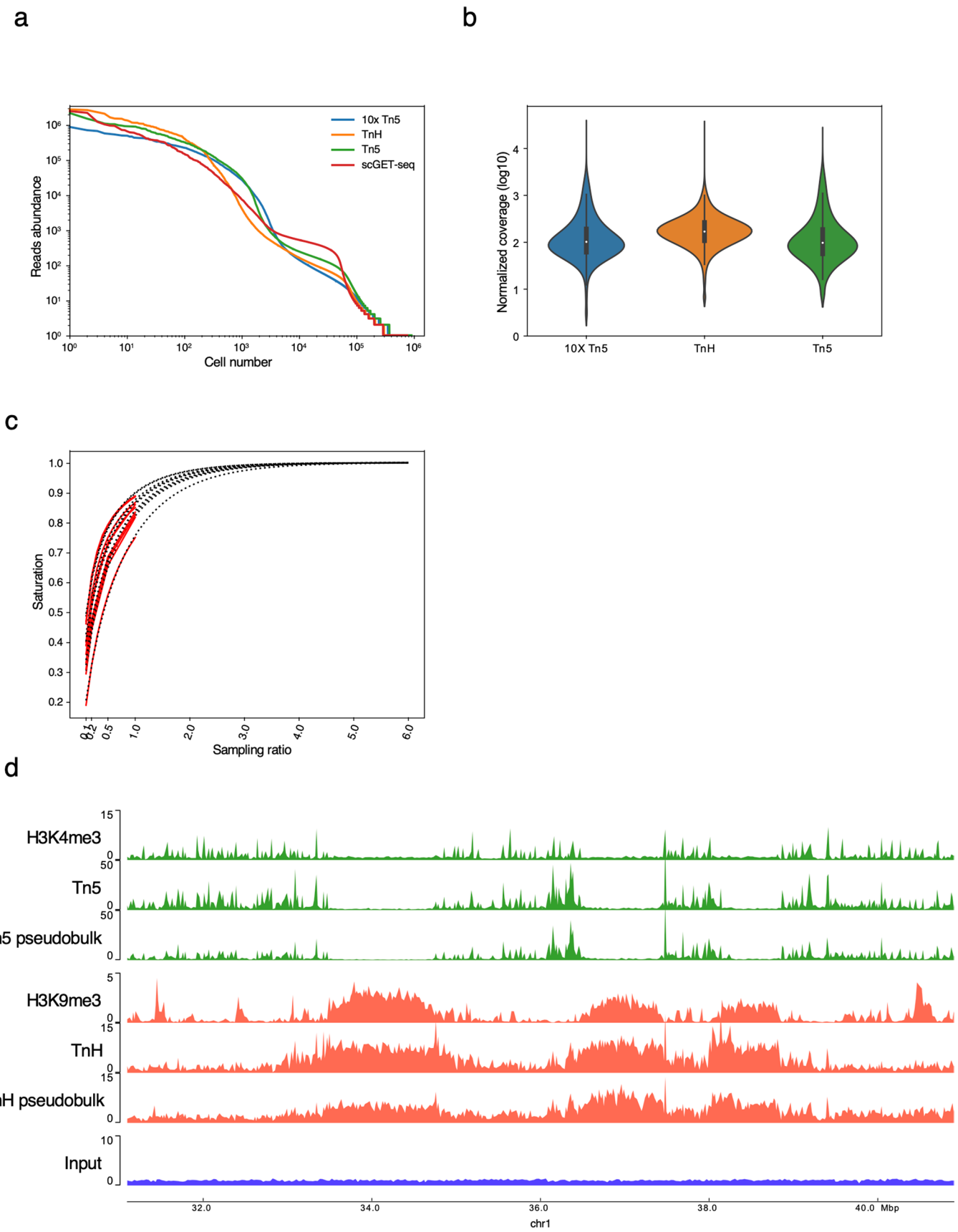
Extended Data Fig. 2 | Hybrid CD (HP1α)-Tn5 targets H3K9me3 chromatin regions. a, Two different lengths of HP1α polypeptide (spanning amino acids 1-93 and 1-112) were linked to Tn5, using either a 3 or 5 poly-tyrosine–glycine–serine (TGS) linker, resulting in four hybrid construct, TnH#1-4. TnH#1 made 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.





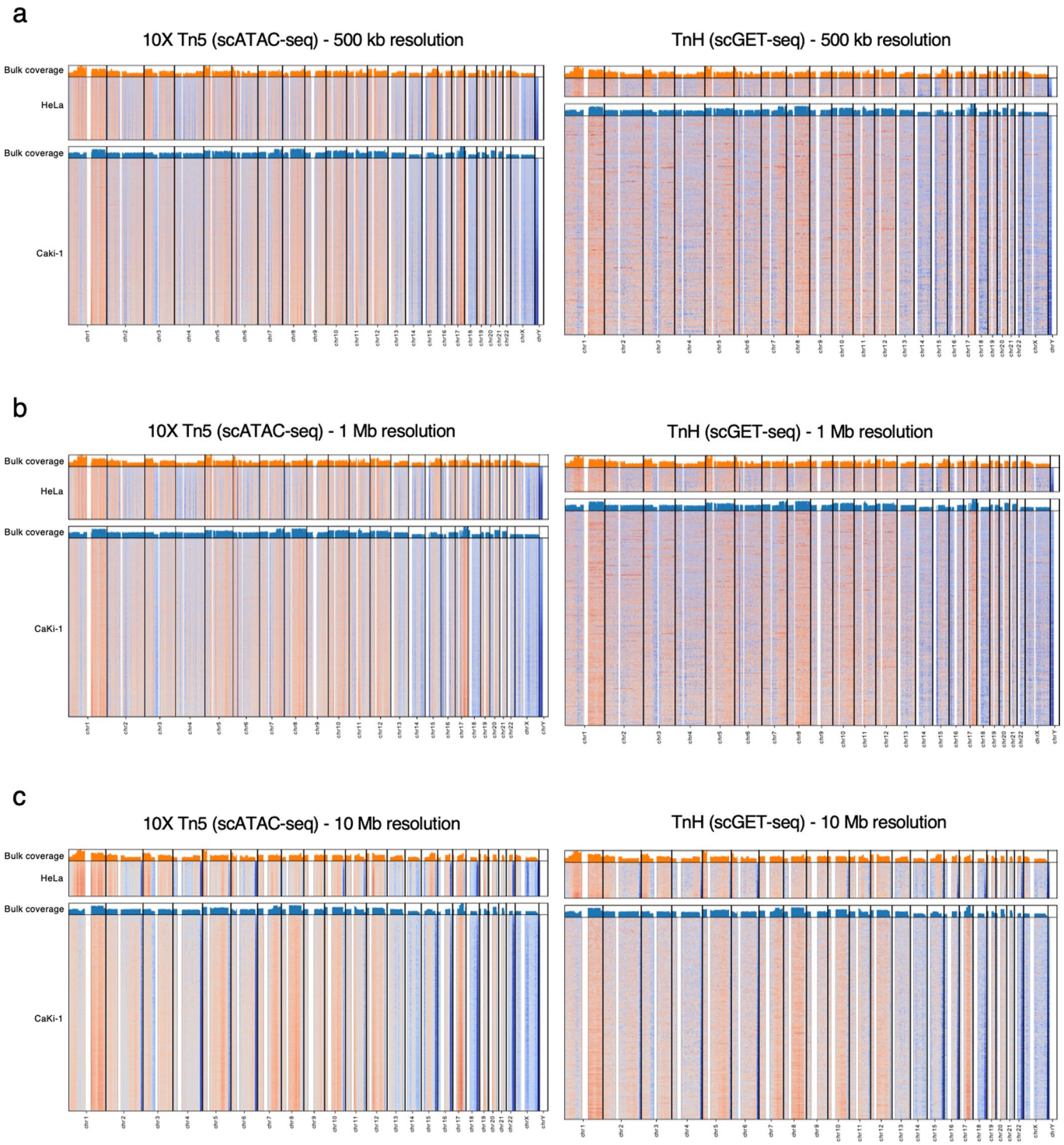
Extended Data Fig. 3 | See next page for caption.

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.

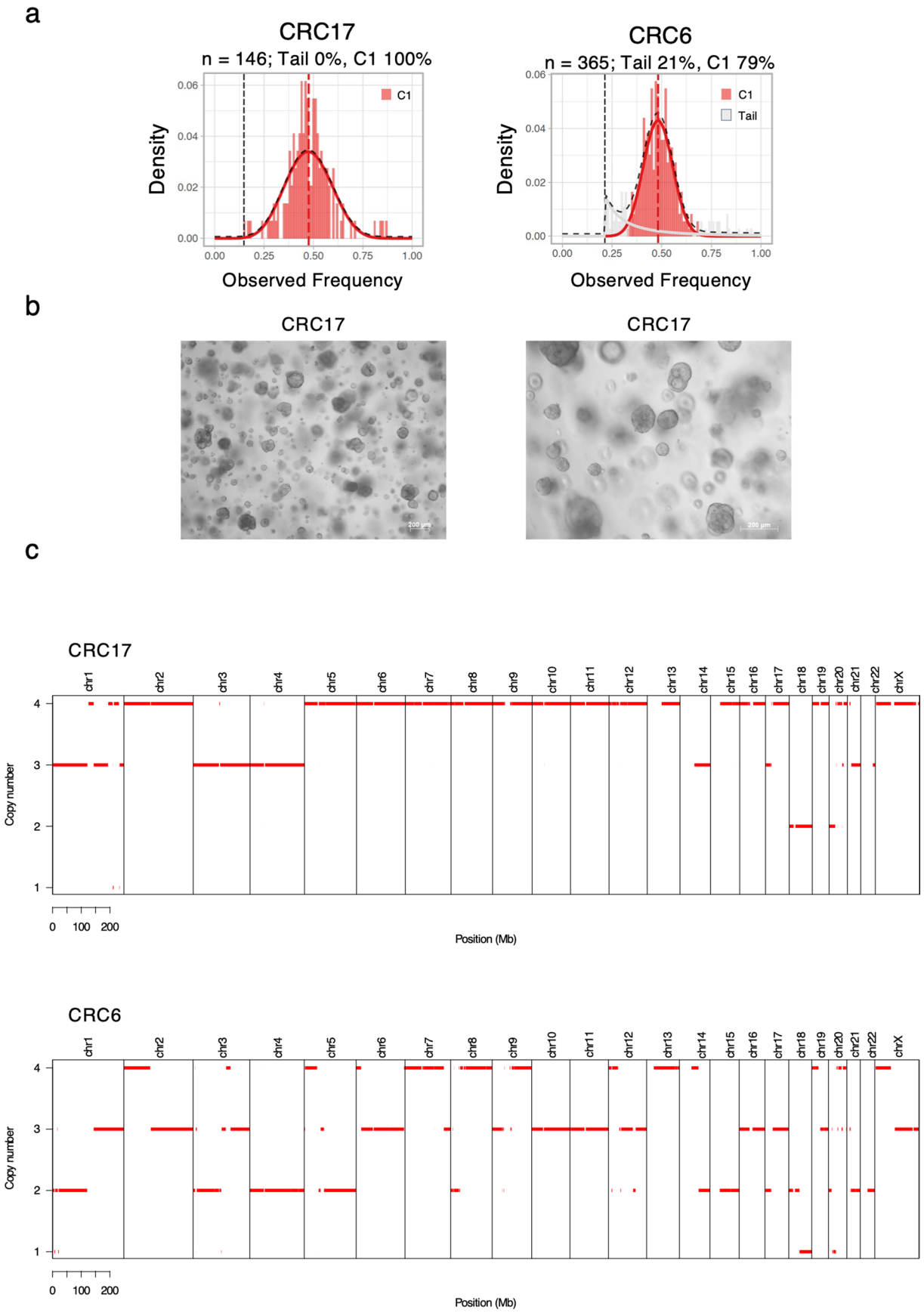


Extended Data Fig. 4 | See next page for caption.

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.

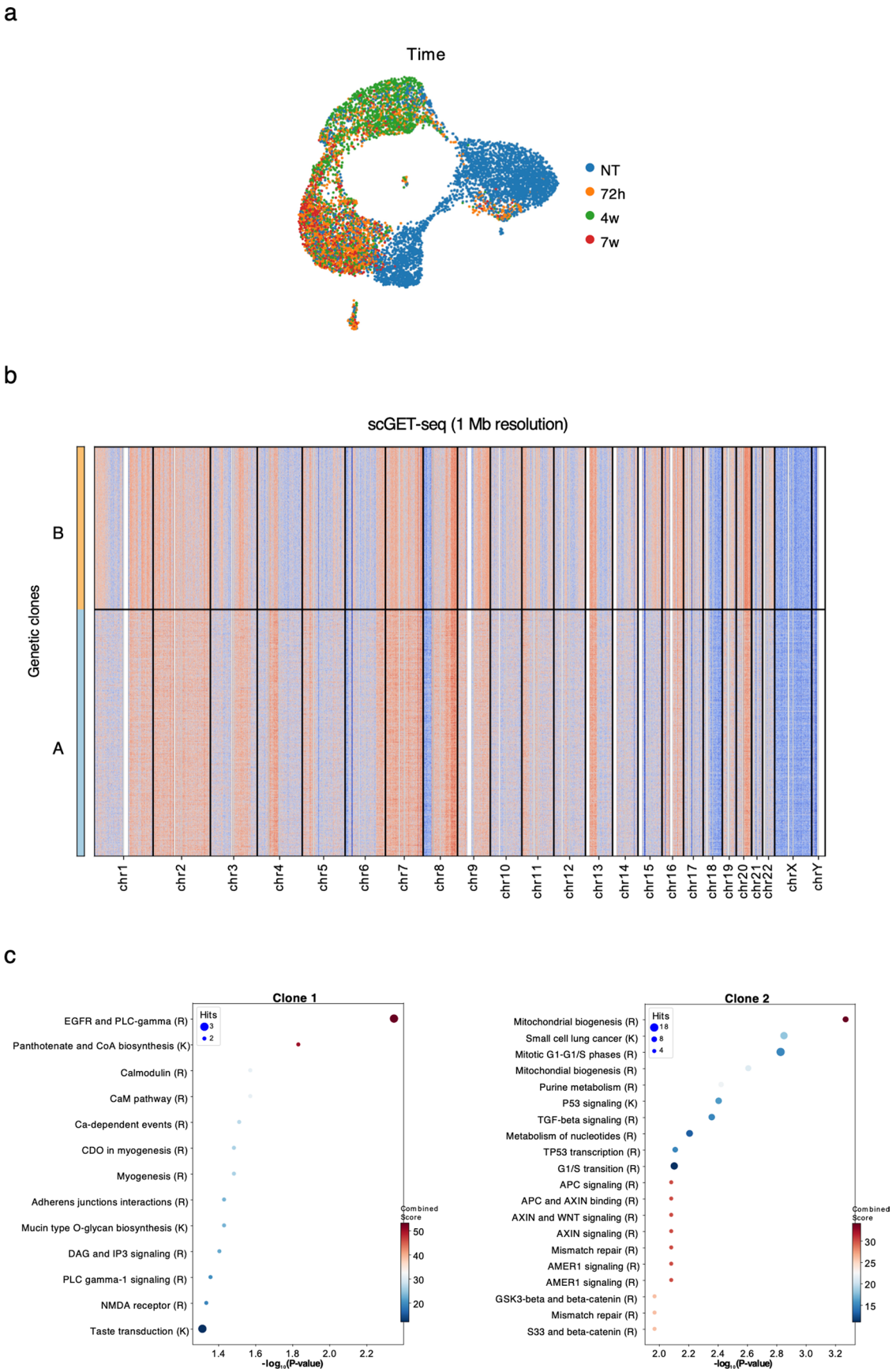


Extended Data Fig. 5 | Copy Number analysis at multiple resolutions. a, Segmentation profiles in individual cells profiled by 10X Tn5 (scATAC-seq) (left panel) or TnH scGET-seq (right panel) at 500 kb. **b**, Segmentation profiles in individual cells profiled by 10X Tn5 (scATAC-seq) (left panel) or TnH 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.



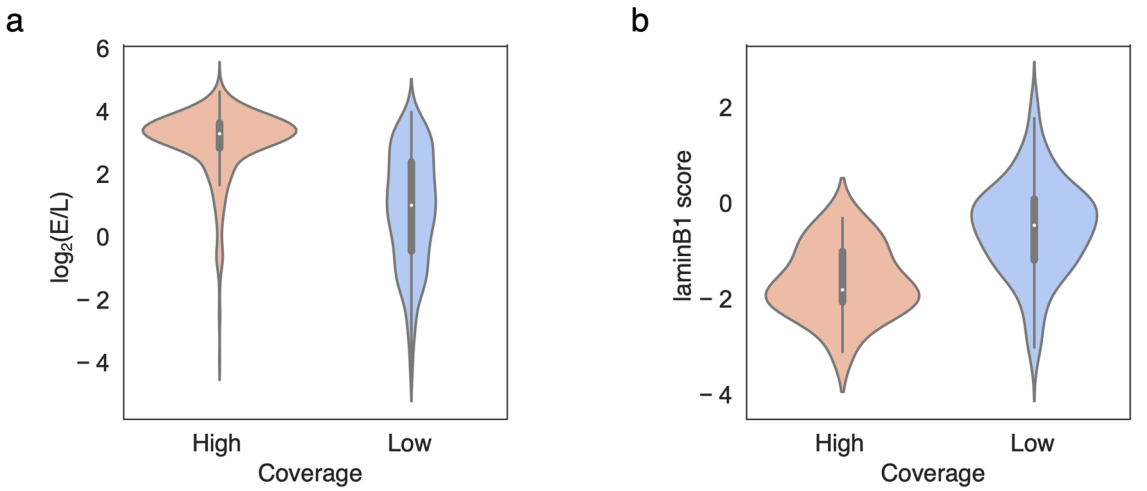
Extended Data Fig. 6 | See next page for caption.

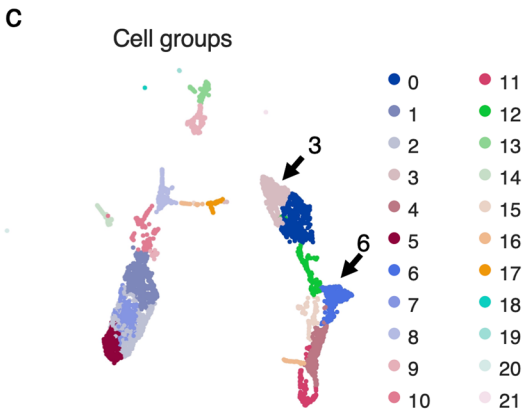
Extended Data Fig. 6 | Characterization of Patient Derived Organoids. a, evaluation of clonal structure of two PDO (CRC6 and CRC17) by exome sequencing; the histogram show the distribution of the cancer cell fraction estimated from the analysis of somatic mutations; in both organoids we observe a monoclonal 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.



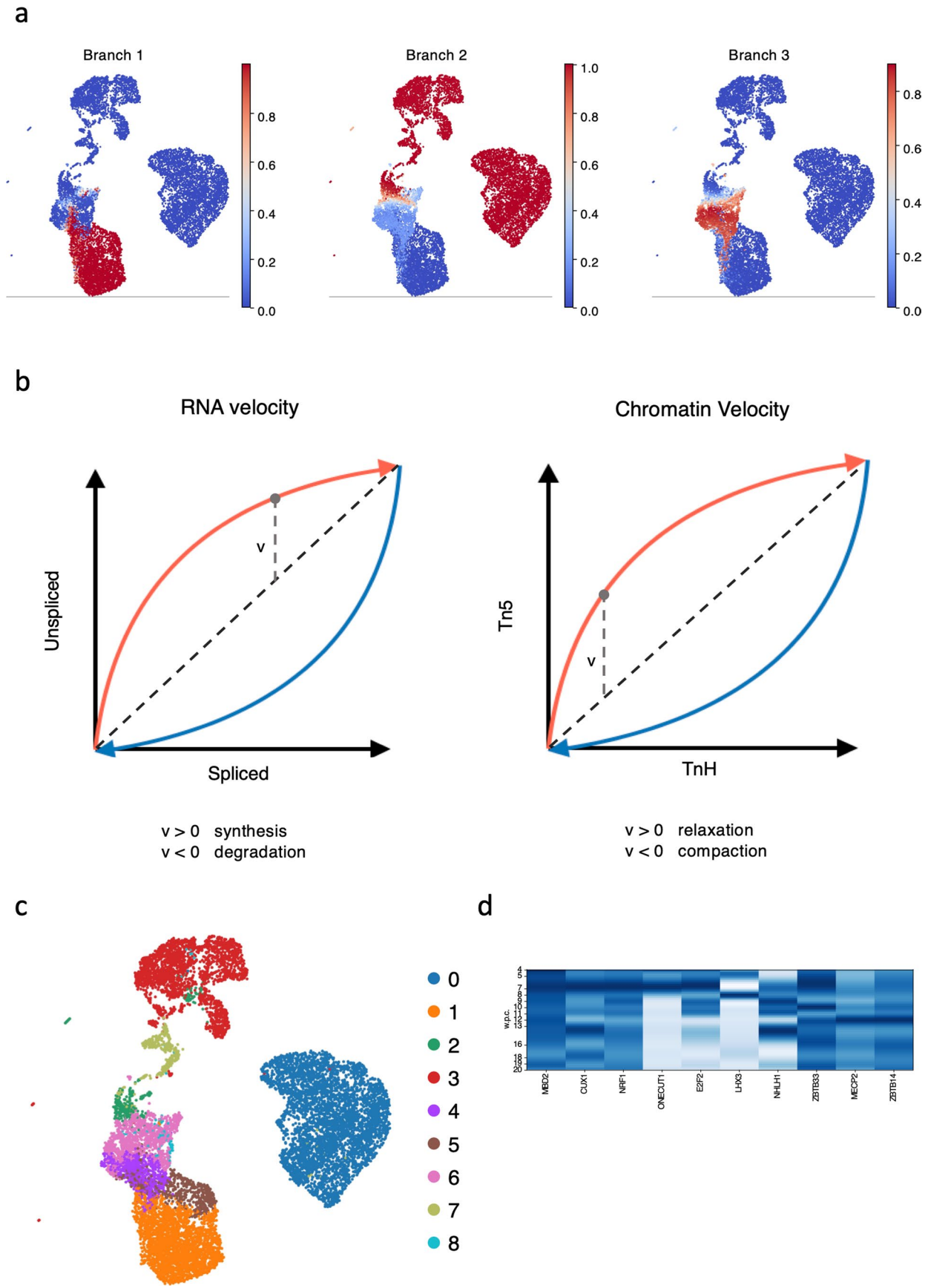


d

			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).