Single cell analysis reveals dynamics of transposable element transcription following epigenetic de-repression

David Brocks[1], Elad Chomsky[1], Zohar Mukamel[1], Aviezer Lifshitz[1], and Amos Tanay[1]

[1] Weizmann Institute of Science, Department of Applied Mathematics and Computer Science and Department of Biological Regulation

ABSTRACT

Spontaneous or pharmacological loss of epigenetic repression exposes thousands of promoters encoded in transposable elements (TEs) for pervasive transcriptional activation. Despite the potential impact of TE de-repression on genome regulation, it remains unknown how TE responses vary between epigenetically targeted cancer cells and whether their de-repression is stochastic or co-regulated with genic transcriptional programs. Here, we develop an approach to quantify TE transcription initiation at single cell and locus resolution to link it with genic programs in epigenetically relaxed cancer cell populations. We find combinatorial TE expression patterns are associated with cell cycle stages, stress response signatures and immune response pathways. The data suggest single cell heterogeneity in TE expression is also driven by the diversity of epigenomic contexts and sequence polymorphisms within large TE families. Our approach thereby facilitates the identification of a previously underestimated source of regulatory heterogeneity in cells with relaxed epigenetic repression.

INTRODUCTION

Transposable elements (TEs) are DNA sequences that multiply in the germline via vertical transmission. They have been evolutionary successful in colonizing many eukaryotic genomes and their massive expansion lead to partial neo-functionalization, in particular as regulatory elements[1,2]. Inactivating mutations have neutralized the threat of insertional mutagenesis for most human TEs. Nevertheless, retrotransposons (LINEs, SINEs, and LTRs) that require an RNA intermediate for their life-cycle encode strong promoter activity that poses the risk of spurious transcriptional activation and subsequent epigenetic perturbation across numerous genomic loci[3,4]. Host cells therefore utilize multiple epigenetic surveillance mechanisms, such as DNA methylation and histone modifications, to prevent the pervasive mis-expression of TEs[5,6].

While TEs are targeted by common mechanisms for epigenetic repression, their ancient promoters are diverse at multiple levels. For example, TE sequences are originating from many classes and families, which continue to diverge even after their immobilization, creating diverse repertoires of *cis*-elements that can promote binding by different *trans*-factors[7,8]. The genomic and epigenomic contexts of TEs are also diverse, with specific families enriched in repressed or active genomic domains, within proximity to other regulatory elements, or away from them[9]. It remains however largely unknown how these factors contribute to the potentially complex and locus-specific regulation of TEs, especially once epigenetic repression is eroded or targeted therapeutically[10–12].

TE de-repression is inherently difficult to characterize since it couples together a large number of distinct genomic loci subjected to stochastic epigenetic erosion. Moreover, analysis of TE's average transcriptional response over millions of cells and thousands of loci made it so far impossible to define heterogeneous and locus specific models for the cis- and trans- regulation of TEs within subpopulations of de-repressed cells. Single cell RNA-sequencing (scRNA-seq) is rapidly revolutionizing our understanding of transcriptional heterogeneity and dynamics in multiple contexts, including cancer[13–16]. The application of single cell technologies to study TE regulation is therefore promising to facilitate a much better understanding of the phenomenology of their transcriptional heterogeneity and the mechanisms regulating it. However, popular scRNA-seq protocols fail to provide information about the origin of transcription when non-canonical reference models, such as the one underlying poorly characterized TEs, are to be considered.

Combining classical gene expression signatures with the noisy and stochastic output from de-repressed TE loci also calls for the development of dedicated computational strategies.

Here, we exploit the benefit of 5' scRNA sequencing and an integrated TE and genic analysis pipeline, to map active transcription start sites (TSSs) within TEs *de novo* and model their co-regulation within sub-populations of single cells. We applied our new approach to study two cancer cell lines following treatment with a clinically established DNA hypomethylating agent and HDAC inhibitor, a drug combination known to induce massive expression of specific TE families[12]. Unexpectedly, we found highly heterogeneous overall TE activity levels in different populations of single cells. This heterogeneity is driven by specific groups of co-regulated TEs, which in turn were strongly linked with fundamental biological processes, including the cell cycle, stress responses, and type-I interferon signaling. Expression patterns of loci from genetically similar families were correlated with specific sequence polymorphisms in their putative promoters, as well as with the transcriptional permissiveness of their genomic and epigenomic context. 5' scRNA-seq highlights the necessity to deconvolute TEs' transcriptional output to the single cell and locus level. Using this new technology it is now possible to investigate poorly appreciated aspects of genome regulation in epigenetically de-repressed cells, not only following epigenetic therapy, but during specific embryonic stages[17], aging[18], and carcinogenesis[19].

RESULTS

Epigenetic therapy induces heterogeneous TE de-repression in single cancer cells. We cultured lung and colon cancer cell lines (H1299 and HCT116, respectively) and performed 5' single cell sequencing to map the transcriptional response to an established epigenetic drug regimen combining a low-dose DNA hypomethylating agent and an HDAC inhibitor (DACSB). After exclusion of low-quality cell barcodes (**Fig S1A**), we retained over 15,000 single cell profiles from treated and untreated cells, quantifying between 7972 - 17328 unique molecular identifiers (UMIs) mapping to known genes per condition (**Fig 1A**). We also identified molecules mapping unambiguously to individual TE loci, and confirmed that the treatment triggered on average a 1.56 – 1.72 (H1299 and HCT116) fold transcriptional induction of TEs in both cell lines (**Fig 1B**). This allowed us to quantify TE transcription at single cell resolution with high accuracy based on 19 – 7037 (12 – 6444 for HCT116) TE UMIs per cell post treatment. Interestingly, both cell lines showed profound intercellular heterogeneity in their TE transcriptional load (7.53 and 7.87-fold

change between 5 and 95 percentile in H1299 and HCT116, respectively**),** and in particular in LTR activity (10.5 and 9.35 fold variation**, Fig 1C**). The surprising extent of global TE expression heterogeneity motivated further in-depth analysis of locus- and family- specific variability in TE de-repression, and the factors contributing to it.

Validating 5' single cell RNA-seq for precise TSS identification.

To validate the precision of our new 5' scRNA-seq pipeline for mapping TSSs, we studied its performance against reference gene TSSs. We found that the maximally covered genomic bins within well-characterized genes were highly consistent with the existing TSS annotation (data on untreated H1299 cells, **Fig 1D**). To systematically identify active genic TSSs, we screened and clustered genomic bins with high relative coverage (**Fig S1B**) defining a set of 11,448 active TSSs in 10,999 genes (with 431 genes linked with more than one active TSS, **Fig 1E-G,** See HCT116 data in **Fig S1 C-E**). Our predicted TSSs accounted for ~62% of the scRNA-seq UMIs overall and the remaining UMIs could be attributed to transcriptional noise or technical effects (e.g. incomplete processing of transcripts using library preparation). We found that the fraction of non-TSS molecules was correlated with the overall size of the spliced transcript (**Fig 1H and Fig S1F**). Moreover, we observed significant cell-to-cell variation in the rate of non-TSS specific UMIs (**Fig 1I**), suggesting the possibility of variable transcriptional noise at single cell resolution. Indeed, we observed mild positive correlation between the non-TSS expression and the overall TE burden for untreated single cells ($r=0.128$ for H1299 and $r=0.127$ for HCT116, both $p<<0.01$). We further confirmed the robustness of TSS de-novo discovery from 5' scRNA-seq data between cell lines and treatment conditions (**Fig S1G**), thereby allowing the application of a similar strategy to characterize transcription initiation outside genic models.

Single cell RNA-seq defines putative TSSs in consensus models of TE families

Encouraged by the robust identification of genic TSSs based on 5' scRNA-seq data, we adapted our approach to define transcriptional activity originating in TEs. We focused solely on UMIs that uniquely mapped to individual genomic TE loci (between 86.17 and 88.96% of all TE UMIs, **Fig S2A**), which allowed the unbiased quantification of transcriptional activity for most TE families (**Fig S2B**). To improve power in detecting bona-fide TSS activity within TEs, we initially aggregated UMIs over consensus sequence models for 1076 TE families (between 3 to 228,527 loci per model) representing the four

major TE classes (LTR, LINE, SINE, and DNA) (**Fig 2A)**. We binned the consensus sequence model of each family and analyzed relative enrichment of bins as a proxy for specific TSS activity (**Fig 2B**). This led to the identification of 722 bins in 442 families with putative TSS activity in H1299 cells (862/506 in HCT116). As shown in **Fig 2C**, mapping UMIs on the TE consensus model for families with at least one specific bin reflected a highly specific pattern of localized consensus transcription initiation (see **Fig S2C** for case examples). We validated that this effect is not correlated with variable mapping efficiency in degenerate TE sequences (**Fig S2D**). The degree of TSS specificity was markedly different among broad classes of TEs. Within LTRs, 72.7% of the total molecules mapped to bona-fide TSS bins, compared to 5.62, 24.2, and 8.25 percent in LINEs, SINEs and DNA TEs, respectively (78.9 / 6.7 / 47.4 / 13.1 in HCT116) **(Fig 2D and Fig S2E**). The overall transcriptional density (molecules per kb), in non-TSS TE bins was comparable between the classes (**Fig 2E and Fig S2F)**, with the exception of few LTR families showing higher density. This showed elevation in transcriptional output was indeed specific for TSSs in certain families, and could not be explained by broad patterns of genomic de-repression without promoter specificity. In conclusion, we observed specific transcriptional initiation in consensus models of de-repressed TE families (and LTRs in particular), linking transcripts from multiple TE loci into consistent and coordinated TE transcriptional patterns.

TE families are transcribed from thousands of loci

To better understand potential intra-TE family expression dynamics, we next investigated the contribution of individual TE copies to the TE consensus signal. We found that most of the active families aggregated transcriptional contribution from a large number of expressed loci (329 out of 442 expressed families with over 10 contributing loci in H1299 405 out of 507 in HCT116) (**Fig 2F and Fig S2G**). For highly expressed families, we also confirmed that the specificity of the TSS bin was completely consistent between copies (**Fig 2G and Fig S2H**). On the other hand, for all TE families, the transcriptional response was incomplete and a large number of TE loci remained repressed (**Fig 2H and Fig S2I**). For the most consistently active family (LTR12C), we observed some transcriptional activity in 2208/2310 (in H1299/HCT116) out of 2735 loci, but also noted that half of all UMIs were contributed by only 3.17/3.38% (n=70/78) of the loci. To further validate the inferred TE expression patterns, we confirmed a high degree of consistency in the activity of TE loci between the two treated cancer cell lines (**Fig S2J**), in particular for loci

associated with high expression. In conclusion, we defined 136,366 genomic loci from 442 TE families (170,746 from 506 in HCT1160) as sites of potential transcriptional initiation following epigenetic therapy. The activity in these sites generated 0 – 23.9% (0 – 38% HCT116) of the total RNA molecule count profiled per cell, suggesting a major influence on the genomic landscape of transcriptional initiation which goes beyond canonical assembly of transcriptional machineries at genic TSSs. This highlights the importance of understanding stochastic and regulated mechanisms driving the specification and activation of TE TSSs at single cell and single locus resolution.

TE metacells uncover co-regulatory TE modules

Having identified transcriptional output from thousands of TE loci at single-cell resolution, we next searched for cell subpopulations that share TE expression patterns, as a way to define TE co-regulatory modules. Analysis of normalized single cell transcriptional variance identified high variance TE TSS loci in treated cells (**Fig 3A**). High variance TE TSSs also showed a detailed and non-homogeneous correlation structure (**Fig S2K**) within the single cell cohort, confirming TE expression is not a mere reflection of variable global TE load following treatment. Furthermore, TE clusters that were correlated in their expression patterns also showed a high degree of pairwise sequence similarity and often mapped to phylogenetically related families (**Figure 3B,** lower right)**.** We organized correlated TEs into groups and annotated the derived TE modules (10 for each cell line) based on their dominant TE class/superfamily associations. This resulted in the identification of multiple ERV modules (comprising mostly solitary LTRs) with a complex correlation structure, a HCT116-specific LINE/DNA module, as well as 2 modules (3 in HCT116) defined by weaker association of multiple Alu elements.

We adapted the MetaCell framework[20] to identify TE metacells as groups of single cells showing highly similar TE co-expression patterns. Analysis of TE metacells and the distribution of TE expression over them showed that the rich correlation structure among TE TSSs is reflecting an organization of the single cell cohort into groups representing distinctive combinatorial TE expression signatures (**Figure 3C**). Importantly, although the observed metacells differed considerably in their overall TE burden (**Figure 3C**, top panel), some of the co-expressed TE TSS modules behave antagonistically between metacells. This was the case even for modules grouping together the same superfamily of ERV elements (e.g. modules ERV9-IV and ERV9-VII in H1299) (**Fig S3A**), confirming metacells were identifying combinatorial TE regulation rather than approximating different degrees

of overall TE activity. These findings demonstrate regulated TE expression dynamics in treated cells and suggest the presence of inter- and intra-family specific transcriptional regulation.

Gene expression correlates with specific TE activation signatures

The observed TE subpopulation dynamics hint towards coordinated control of TE modules by factors acting in *trans*. In that case, TE metacells, which are statistically derived based solely on TE expression, should be distinguishable also by their gene expression signatures. Consistent with this idea, we observed remarkably rich gene expression signatures in TE metacells (**Fig 4A-C**). Globally, we detected 6518 (3950 in HCT116) (chi-squared, FDR < 0.01) genes with significant differential expression between TE metacells. More specifically, genes enriched in specific TE metacells were functionally and spatially related. For example, in H1299 we observed KRT8, KRT18 and additional genes overexpressed in TE metacells #22 and #8, expressing TE module ERV9-II (**Fig S3B**). In HCT116, expression of early-immediate genes (JUN, ARC) were enriched in TE metacells #1 expressing TE module ERV1/ERVK (**Fig S3C**). The LINE/DNA module was highly specific for HCT116 metacell #13, comprising presumably apoptotic cells based on their high mitochondrial and autophagy-related gene expression signature (ATG10, DNASE1) (**Fig S3D**). Importantly, we identified a group of TE metacells in H1299 (#12-14, #21) that were enriched for expression of loci within chromosomal locus 19q3 (**Fig S3E and F**), suggesting potential copy number heterogeneity as the underlying mechanism for up-regulation of TEs and genes in these metacells. However, most other TEs were not linked with specific chromosomal domains, and included a balanced mixture of chromosomal loci that are unlikely to be explained by sub-clonal structure bearing specific chromosomal aberrations (**Fig S3G**).

To complement the detection of these combinatorial gene expression patterns in TE metacells, we searched for gene expression signatures that correlate with overall TE load in single cell resolution (**Fig S3H**). We found various immune-related genes (e.g. IL16) to be positively associated with overall TE load, in line with recent reports on de-repressed ERVs triggering a viral-like interferon (IFN)-response[10,11,21]. Indeed, overall IFN-I intensity was correlated with total TE and especially LTR load at the single cell level (**Fig S3I and J**). However, the effect size was modest with at most 1.61 and 1.5 fold change in IFN-I expression between TE metacells in H1299 and HCT116, respectively. In summary, our analysis defines a rich subpopulation structure in treated cells, involving combinatorial

activity of different TE TSSs, and correlated patterns of regulated gene expression. These data suggest a cancer cell's state may regulate classical gene and TE promoters simultaneously.

TE activation patterns are cell-cycle dependent

Some of the most notable gene expression signatures observed for TE metacells involved co-regulation of multiple cell cycle genes (e.g. MKI67, UBE2C, BIRC5, PTTG1). To follow up on these observations and characterize systematically potential cell cycle regulation of TE expression, we generated cell cycle metacell models using a selected set of genes in each cell lines (**Fig 4D and S4A**). We then projected the total expression of TE modules on the cell cycle metacells (**Fig 4E-F**). This resulted in observation of unexpectedly pervasive cell cycle correlation of TE expression patterns. In H1299, some TE modules (ERVL-MaLR/ERVL, ERV9-III, and ERV9-VII) peaked in expression at G1/G0, while other TEs modules (ERV9-IV, and ERV9-VI) showed positive correlation of their expression with specific stages in S phase. Analogous dynamics (albeit less specific) were inferred for HCT116 cells. Interestingly, analysis of the distributions of genomic time of replication for loci in different TE modules showed remarkable diversity as well. We observed an overall enrichment of active Alu elements at early replicating domains consistent with the general preference of SINEs to active genomic loci. More surprisingly, we detected distinct time of replication distributions for different ERV9 TE modules, for example showing in H1299 the ERV9-III module as early replicating and the ERV-IV module as late replicating. The correlation between TE cell cycle dynamics and time of replication was imperfect – although the two modules with the latest time of replication in H1299 (ERV-III, ERV-VII) showed decrease in expression during replication. Interestingly, we observed mild cell cycle correlations for some of the identified TE modules even in untreated cells, suggesting this effect is generally applicable, even when TEs are only weakly and sporadically de-repressed (**Fig S4B**). Together, these data indicate that chromosomal dynamics during replication can act together with different cellular programs to regulate specific patterns of TE activity.

Epigenomic context and sequence composition are predictive of TE de-repression

The analysis of TE metacells and TE TSS modules suggested that some of the TE de-repression dynamics are correlated with the activity of specific trans-factors. However, since even highly responsive TE families were characterized by only partial de-repression

of the family's loci, we next aimed to define sequence features and genomics contexts that predispose loci to become de-repressed. Focusing on families contributing at least 5 loci to TE modules, we first found strong correlation between the distance of a TE locus to the nearest expressed genic TSS, and its de-repression predisposition (**Fig 5A**, **Fig S5**). This effect was linked with enriched de-repression probability of loci in the active chromosome compartment, and lower de-repression in loci within lamina associated domains (**Fig 5B**). Analysis of DNA methylation data before and after treatment[12] showed that de-repressed TE loci are significantly more methylated before treatment than loci that are not de-repressed (**Fig 5C**). After treatment, TE methylation levels are decreasing on the responsive TEs, converging to levels similar to those initially observed in the non-responsive loci prior to treatment. This result is suggesting the non-responsive TE loci are relying on mechanisms other than DNA methylation for repression, while the de-repressed loci show tight correlation, and likely a causal link between methylation and repression.

In addition to genomic context and epigenomic markup, a previous report suggests sequence divergence between TEs within the same family can be predictive of de-repression trends[12]. To test this, we clustered 2666 (excluding 66 loci on chromosome Y) sequences from the LTR12C family according to their sequence, deriving two distinct clusters varying by hyper-variable regions at multiple positions and in particular close to the identified TSS (**Fig 5D**). We then hierarchically classified elements in each of the two clusters according to their time of replication and analyzed the overall expression in each group, as well as the distribution of expression over TE metacells (**Fig 5E**). The data showed substantially strong expression in genetic cluster #2 compared to #1 ($p << 0.01$ in both cell lines). Within the genetic clusters, we still observed a strong correlation between time of replication and activity patterns (all $p << 0.01$), highlighting how broad genomic context and sequence composition additively contribute to the de-repression propensity within a TE family.

DISCUSSION

We performed 5' single-cell transcriptional profiling to *de novo* identify TSS activity in epigenetically de-repressed cancer cells. Using a computational approach that leverages the high genomic copy-number and sequence conservation of TE families, we robustly pinpointed the exact sites and degree of transcription initiation in thousands of loci representing all major transposon classes. This uncovered a remarkably rich and dynamic landscape of TE cell-to-cell variation in supposedly homogenous cancer populations. We

observed different overall transcriptional dynamics for the four major TE classes. SINE retrotransposons largely showed a pattern of expression consistent with a more sporadic and spurious de-repression. Specific families of DNA transposons and LINEs were linked to a presumed pre-apoptotic phenotype, suggesting broad-scale loss of genomic control. On the other hand, we observed structured activity of LTR retrotransposons, which was linked to specific stages of the cell cycle and additional genic pathways. Coordinated activity of multiple loci from the same TE family was observed systematically, but in some cases, and contrary to our expectations, loci of the same LTR family could be clustered into modules showing distinct combinatorial expression patterns. This data shows that even in cells with largely perturbed epigenetic integrity, TE de-repression is highly regulated, suggesting dynamic availability of trans-factors (through cell cycle fluctuation or more stable transcriptional states) can be read by TE promoters and give rise to complex patterns of combinatorial response.

As exemplified for members of the massively de-repressed LTR12C TE family, intra-family de-repression variability is associated with at least three layers of regulation. First, activated loci rely on a favorable broad genomic context and proximity to activator machinery at constitutively active TSSs. Second, de-repression involve change in local epigenetic composition (in particular DNA methylation) of the TE TSSs. Finally, specific sequence characteristics within the TE family consensus module are linked with locus-specific regulation, possibly ensuring or preventing effective recruitment of the activatory trans-machinery. Single cell RNA-seq can uncover the combinatorics of these layers and how they affect TE regulation in general, improving our understanding of the previously described cell-type and disease-specific TE expression landscapes[22–24].

Understanding the genomic rules of TE de-repression is particularly important in the context of epigenetic therapy, where a subset of ERVs is reported to mimic a viral-like interferon response and to serve as a rich source of potentially immunogenic neoantigens[10–12]. We found the expression of several LTR families to be correlated with overall interferon type-I gene expression at single cell resolution, motivating further studies on the interplay between TE activity, intra-cellular regulation of immune response, and the efficacy of epigenetic and immunotherapy. Epigenetic relaxation is also observed during massive genome-wide DNA demethylation during embryogenesis, germline development, or when cells progressively lose their canonical DNA methylation in ageing[25–27]. The pervasive regulation of TE dynamics we observed in this study suggest the possibility that

a large number of un-appreciated hotspots of genomic activity can become activated in such processes. Incorporating single cell TE activity profiles into models of epigenetic control and gene regulation will provide a powerful tool for understanding TE sub-functionalization toward developmental enhancers and alternative promoters, as well as in the context of aberrant de-repression leading to disease.

DATA ACCESS

Raw and processed data were deposited in the gene expression omnibus repository under accession GSE121309.

MATERIAL AND METHODS

Cell culture and treatment

HCT116 parental colorectal cell lines (HD PAR-033) were obtained from Horizon Discovery Ltd, Cambridge, UK and NCI-H1299 non-small cell lung cancer cells were provided by the courtesy of Prof. Moshe Oren. H1299 cells were cultured in RPMI 1640 (Gibco 21875034I) and HCT116 cells were grown in heat-inactivated McCoy's 5A medium (Biological Industries, Israel; 01-075-1A), both supplemented with 10% Fetal Bovine Serum (Gibco FBS 10270-106), 0.4% Penicillin-Streptomycin (Biological Industries, Israel; 01-031-1C) and 1% L-glutamine (Biological Industries, Israel; 01-020-1). Conditioned media were filtered through a 0.22 microns filter (Corning, 430769) prior to culture. Cells were split at a ratio of 1:10 every 2–3 days using 0.05% trypsin-EDTA solution C (Biological Industries, Israel; 03-053-1B). Treatment with the DNA hypomethylating agent 5-Aza-2′-deoxycytidine (Sigma-Aldrich, A3656; Lot #SLBS4457) and HDAC inhibitor SB939 (Cayman Chemical, 10443; Lot# 0435336-34) was performed as previously described[12]. To prepare single-cell suspensions, cells were processed according to the instructions of the 10x Genomics Single Cells Protocols Cell Preparation Guide (10x Genomics, CG00053 Rev C). About 4000 cells per cell line and treatment condition were used as input for further processing.

5' single-cell RNA sequencing

Libraries were generated using the Chromium Single Cell 5' Library & Gel Bead (10x Genomics, PN-1000014) according to the manufacturer's instructions and sequenced on the Illumina NextSeq 550 using the High Output v2 kit (Illumina, FC-404-2004).

Curation and binning of gene and TE models

Gene models were based on Gencode release 28 (GRCh38.p12) obtained from (https://www.gencodegenes.org/releases/). Exon coordinates of all transcript isoforms were reduced to a single set per gene symbol using the reduce function of the R package GenomicRanges. To minimize mapping ambiguity for exons shared between different genes, we then created a set of non-overlapping exon coordinates with the disjoin function of the same package and retained all gene symbols for shared intervals. After elongating the 5' end of exons containing annotated TSSs to 100bps or to the next exon, we binned the resulting coordinates into 20bp intervals.

We downloaded genomic TE alignments generated by RepeatMasker open-4.0.5[28] with the repeat library 20140131 from (http://www.repeatmasker.org/species/hg.html). After binning all family consensus sequences into 20bp intervals, we used the obtained alignments to map genomic TE positions onto their relative bin ID on the corresponding consensus. Rare sequence polymorphisms not represented in the consensus model were assigned to the preceding bin ID. We then grouped positions with identical family and bin ID into genomic intervals. To minimize misassignment of genic UMIs onto TE loci, we finally excluded TE bins located within genic exons (except those overlapping any genic reference TSS) to generate the final TE coordinate set.

Data processing

Read pairs missing the constant part of the template switch oligo (defined by a hamming distance to TTTCTTATATGGG greater than 4) were excluded before subsequent processing with the cell-ranger analysis software version 2.1.0. Alignment was performed against the human GRCh38-1.2.0 reference using non-default Cell Ranger parameters (chemistry = SC5P-PE; r1_length = 100; r2_length = 50) and alignment-specific parameter (--outFilterMultimapScoreRange: 2). UMI tools[29] (--per-cell; --extract-umi-method = tag; --buffer-whole-contig; --cell-tag = CR; --umi-tag = UR; --paired) was used to remove duplicates. Cell barcodes identified by Cell Ranger with low mitochondrial or ribosomal counts were excluded (**Fig S1A**). Coordinates of the first position of read1 in properly and uniquely aligned read pairs were tested for overlaps with binned exon or TE intervals, summarized for all cells, and used for downstream analysis.

Alignment benchmark

Ambiguous reads (up to 10 distinct genomic alignments) were tested for overlaps with the curated TE bin set. Reads were assigned to the TE family consensus bin with maximal

count (ties were resolved randomly). Mappability score was defined as the percentage of uniquely mapped from all UMIs (including multi-mappers).

TSS mapping

UMI counts were aggregated for all cells per TE family or gene bin. To account for inter- and intrafamily bin copy-number and size variations, TE aggregate bin UMI counts were normalized by a bin's genomic copy-number and median width, and regularized by the families' median bin copy-number. Genic bins with a minimal UMI to total gene expression ratio of 0.1 and at least 10 raw UMIs were marked as putative TSSs and combined if less than 100bp apart. For TE bins, we calculated the UMI enrichment over the neighboring 8 bins (with an offset of 1 in both directions) and assessed statistical significance based on the Poisson distribution. TE bins with at least 20 raw UMIs, fold-change > 3, and FDR corrected p-value < 0.01 were selected. While both genic TSS and non-TSS UMIs were retained (distinguished by an |# TSS suffix), only TE TSS bins were kept as features before summarizing molecule counts in a UMI matrix $U = [u_{fi}]$ on features f and cells i. For gene-centric analyses, non-TSS UMI counts for genes that were associated with a significantly enriched TSS were ignored.

Metacell analysis

Transposable elements

We used the MetaCell package[30] for TE feature selection, grouping of cells, and to calculate gene enrichment along those groups, with the following adaptations. No preliminary cell filtering was applied to TE UMI counts. Markers to model cell-to-cell similarity were selected based on a normalized variance threshold above 0.1, a total UMI count of at least 10 molecules, and second highest UMI count of 2 after downsampling. After constructing an initial similarity graph with a K = 100 parameter, we sampled 500 metacell covers with a minimal cell size of 50, each organizing 75% of cells into coherent groups in the graph. The resulting pairwise co-clustering of all single cells was used as input to construct the final similarity graph and robust metacell covers were inferred with a K = 30, alpha = 2, and minimal cell size of 50.

Cell cycle

Cell cycle-related genes were defined based on their correlation to one of the marker genes PCNA, E2F1, TOP2A, MKI67, UBE2S, ATAD2, ARL6IP1, AURKA, or RRM2 (r >=

0.14 to 0.25). Clustering correlated genes based on downsampled single-cell UMI count matrices, was followed by manual selection of gene clusters that consisted of cohesive cell cycle gene modules. This procedure was repeated for four different down-sampling depths (n=10000 to 5000) to finally select 63 - 119 cell cycle regulated genes for further analysis. Metacell analysis of a UMI matrix consisting of these genes alone (aiming at larger metacell sizes by setting min_mc_size = 100) was then performed, organizing all cells along a putative cell cycle trajectory. This approach provided adequate power for testing subsequent correlation of TEs with different cell cycle phases, and we therefore did not aim at more continuous modelling of the transcriptional dynamics as suggested by more direct modelling strategies[31,32].

Time of replication and lamin association

Repli-seq and DAM-ID normalized coverage files were obtained from the 4D nucleus[33] data portal (https://data.4dnucleome.org/) for HCT116 and H1-hESC cells, respectively. Replicates were averaged and resulting coverage values were percent rank transformed.

Multiple sequence alignment and genetic divergence

MAFFT[34] version 7.394 was used to align sequences with a gap opening penalty and offset of 1.1 and 0, respectively. Positions with a gap frequency greater than 95% on the alignment were removed. Distances between TE loci were computed using the dist.dna function of the R package ape using the indelblock substitution model. Ward's linkage was used to group loci. A locus' position-wise divergence from the consensus was defined as 1 – that bases' consensus frequency (including gaps as a fifth base).

Selection of IFN-I genes

All annotated IFN-I genes from the interferome[35] database were downloaded and further filtered for genes that were consistently upregulated with a minimal average fold-change of 2 across at least 2 independent studies. We compared expression between untreated and treated cells for the remaining 1220 putative IFN-I regulated genes and defined a final set of 164 (H1299) and 127 (HCT116) genes with increased (log2 fold-change larger than 2) expression following treatment.

DNA methylation analysis

DNA methylation values for untreated and treated H1299 cells were downloaded from the gene expression omnibus repository under accession GSE81322. Hg19 CpG coordinates were lifted to hg38 using the liftOver function from the Genomic Ranges R package.

Statistical analysis

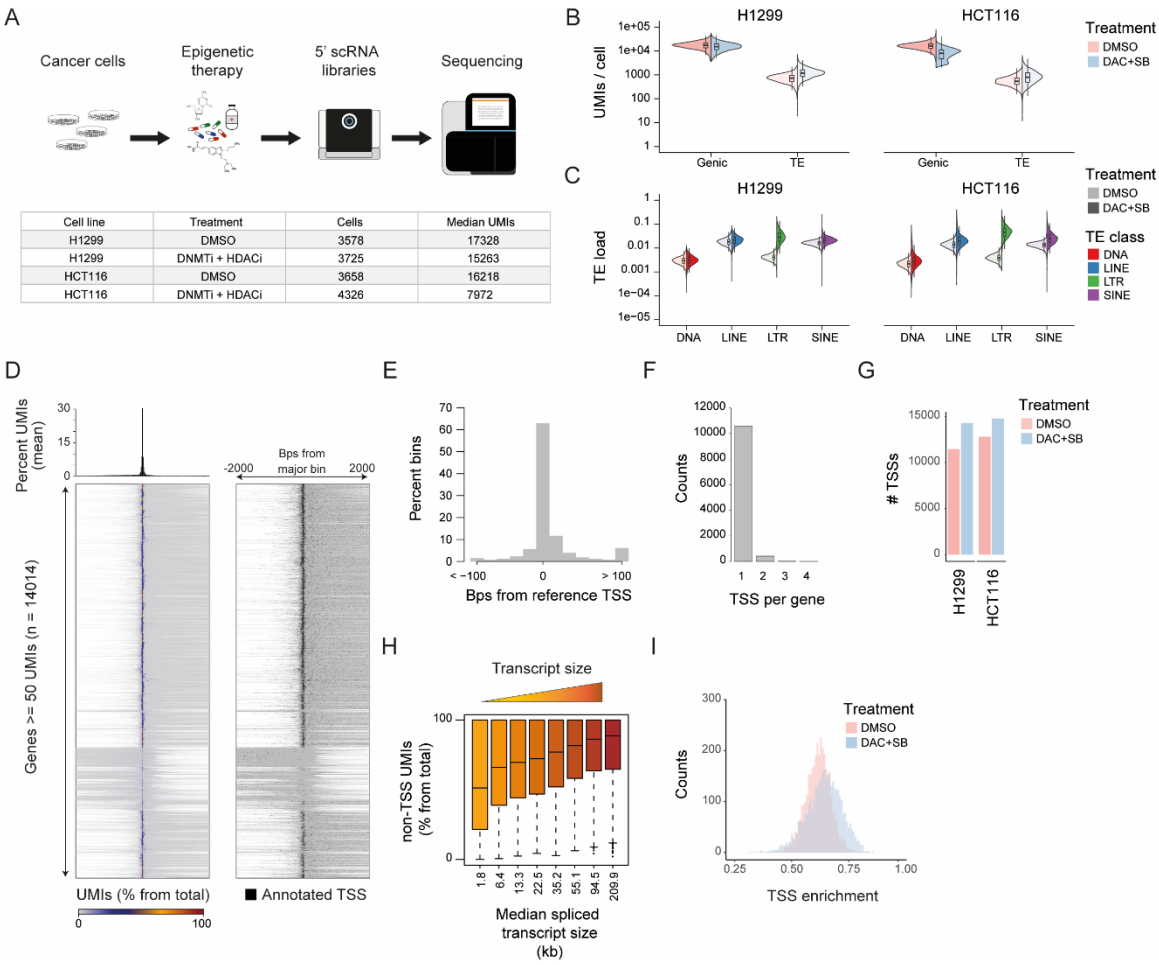
All statistical analyses were performed using the R statistical environment, version 3.4.2. Unless stated otherwise, box plot center lines indicate data medians, box limits indicate the 25th and 75th percentiles, whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles, and outliers are not shown. For group-wise comparison of two distributions from different samples or treatments, the two-tailed nonparametric Wilcoxon and Mann-Whitney test was used. Unless stated otherwise, p-values < 0.05 were considered statistically significant and significance levels are depicted as follows: * $P < 0.05$, ** $P < 0.05$, *** $P < 0.005$. The Benjamini and Hochberg method was used to correct for multiple hypothesis testing.

ACKNOWLEDGEMENTS

We thank L. Velten and C. Plass for helpful feedback and discussions and Z. Meir and A. Sebe-Pedros for critical proofreading of the manuscript. We are also grateful for the support of H. Keren-Shaul and other members of the Advanced Sequencing Technologies facility at the Weizmann Institute of Science. D.B. is supported by an EMBO long-term fellowship (ALTF 900-2017). Research was funded by the European Research Council (grant scAssembly), by the flight attendant medical research institute and the ISF I-core program. We thank B. van Steensel and D. Gilbert for providing access to lamin association and time of replication data generated as part of the 4DN NIH consortium.

FIGURES

Figure 1



Cell line	Treatment	Cells	Median UMIs
H1299	DMSO	3578	17328
H1299	DNMTi + HDACi	3725	15263
HCT116	DMSO	3658	16218
HCT116	DNMTi + HDACi	4326	7972

Figure 1 | Single-cell RNA-seq identifies the transcription start site of genes

A) Summary of the applied experimental workflow.

B) Violin plot showing the single-cell density distribution of total genic and TE UMI counts for untreated (red) and treated (blue) H1299 (left) and HCT116 (right) cells.

C) Single-cell density distribution of TE load (TE UMIs / total UMIs) in four classes, before (light colors) and after (dark colors) treatment.

D) Left: Color coded spatial UMI density matrix for a pool of untreated H1299 cells, showing number of mapped UMIs for 20bp bins (columns) within 14,014 genes (rows), aligned around the maximally covered bin. Right: showing the same genomic bins and using the same alignment, but depicting annotated TSSs. Gene models with low total coverage (<50 UMIs) or those comprising ambiguous exons shared by multiple gene symbols were excluded.

E) Histogram of the genomic distance of selected bins to the nearest reference TSSs for untreated H1299 cells.

F) Counts of distinct putative TSS bins per gene model in untreated H1299 cells.

G) Final number of identified distinct TSS bins in untreated (red) and treated (blue) cells.

H) Gene-wise TSS enrichment stratified by groups of increasing spliced transcript size (from yellow to red) based on untreated H1299 aggregate counts.

I) Single-cell distribution of TSS enrichment (proportion of UMIs mapping to identified TSSs) before (red) and after (blue) treatment shown for untreated H1299 cells.

Figure 2

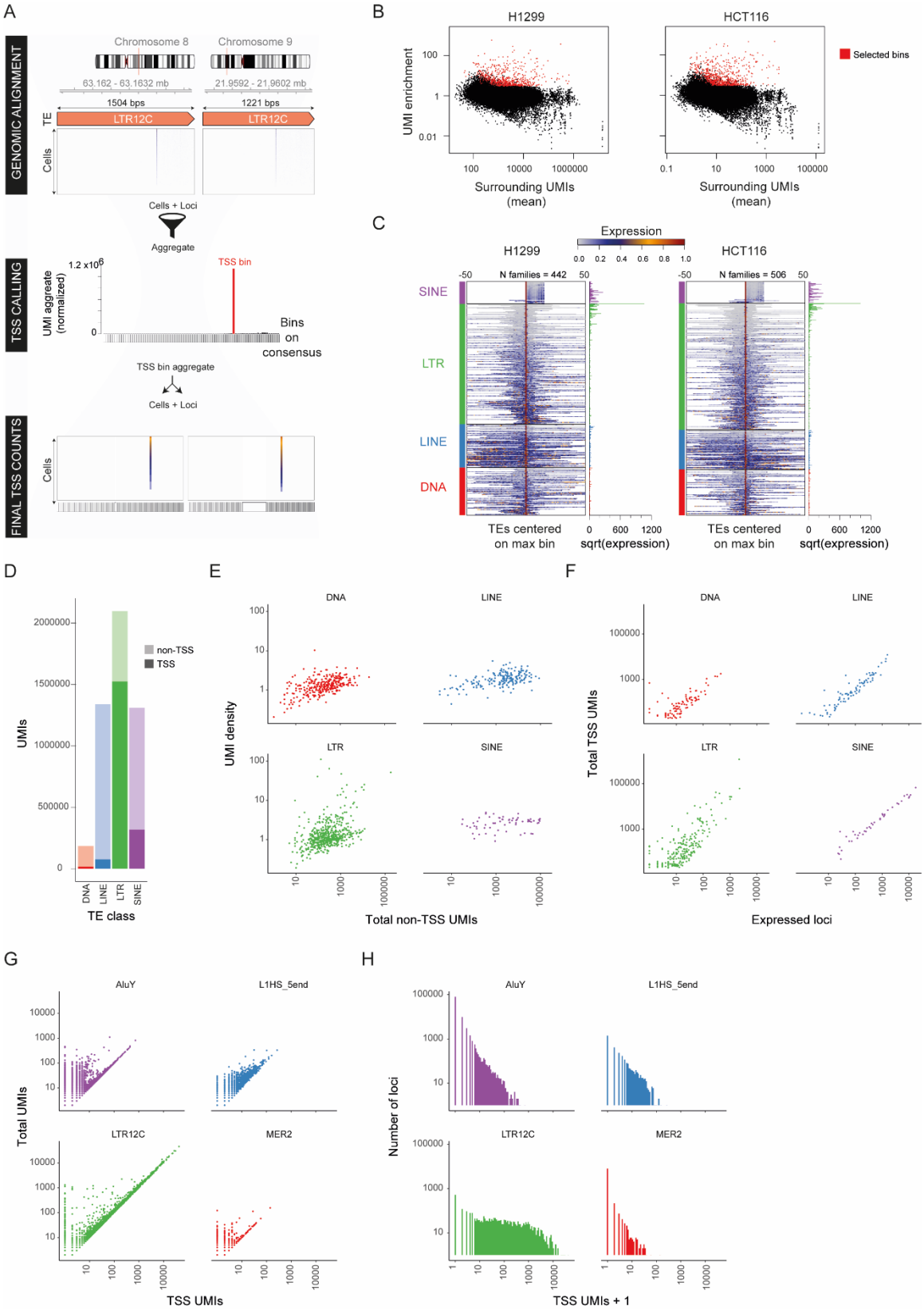

A
GENOMIC ALIGNMENT
Chromosome 8
Chromosome 9
63,162 - 63,1632 mb
21,9592 - 21,9602 mb
1504 bps
1221 bps
TE
LTR12C
LTR12C
Cells
Cells + Loci
Aggregate
TSS CALLING
UMI aggreate (normalized)
1.2 x10^6
0
TSS bin
Bins on consensus
TSS bin aggregate
Cells + Loci
FINAL TSS COUNTS
Cells
B
H1299
HCT116
UMI enrichment
Selected bins
Surrounding UMIs (mean)
Surrounding UMIs (mean)
C
Expression
H1299
HCT116
N families = 442
N families = 506
SINE
LTR
LINE
DNA
TEs centered on max bin
sqrt(expression)
TEs centered on max bin
sqrt(expression)
D
UMIs
non-TSS
TSS
DNA
LINE
LTR
SINE
TE class
E
DNA
LINE
LTR
SINE
UMI density
Total non-TSS UMIs
F
DNA
LINE
LTR
SINE
Total TSS UMIs
Expressed loci
G
AluY
L1HS_5end
LTR12C
MER2
Total UMIs
TSS UMIs
H
AluY
L1HS_5end
LTR12C
MER2
Number of loci
TSS UMIs + 1

Figure 2 | Identification of transcription start sites in TEs

A) Strategy to identify putative TE TSSs. UMIs aligning unambiguously to single TE copies are aggregated from all cells and all loci of a given family. Binning UMIs over the consensus model is performed in 20bp resolution, as exemplified for two LTR12C copies (top). Consensus TSS bins are selected by their UMI enrichment compared to neighboring consensus bins (middle). UMIs from Identified TSS bins are then projected back to specific loci and single cells. Downstream analyses are then performed on signatures of transcription for individual TE loci and single cells (bottom).

B) Enrichment scores for TE consensus bins (Y axis), by total UMI counts on neighboring bins for H1299 (left) and HCT116 (right) cells. Bins identified as putative TSSs are highlighted in red.

C) Aggregate normalized UMI count distribution along the consensus models for families with at least 1 detected TSS bin in H1299 (n=442, left) and HCT116 (n=506, right) cells. Family profiles (rows) are centered around the TSS bin with highest raw expression and arranged by TE class (colored boxes) and TSS bin enrichment. UMI aggregate counts of the maximum TSS bin are shown in the right.

D) Number of TSS (dark colors) and non-TSS (light colors) UMIs mapping to different TE classes in treated H1299 cells.

E) Family-wise H1299 UMI density (defined as non-TSS UMIs per genomic kilobase) over total number of non-TSS UMIs.

F) Family-wise total number of TSS UMIs by the number of active loci (at least 1 UMI) in H1299 cells grouped by TE class.

G) Per locus total number of UMIs as a function of TSS-only UMIs for representative TE families. For each TE class, the maximally expressed family in H1299 cells was selected.

H) Distribution of total UMIs per locus for the highest expression multi-copy TE family of each class.

Figure 3

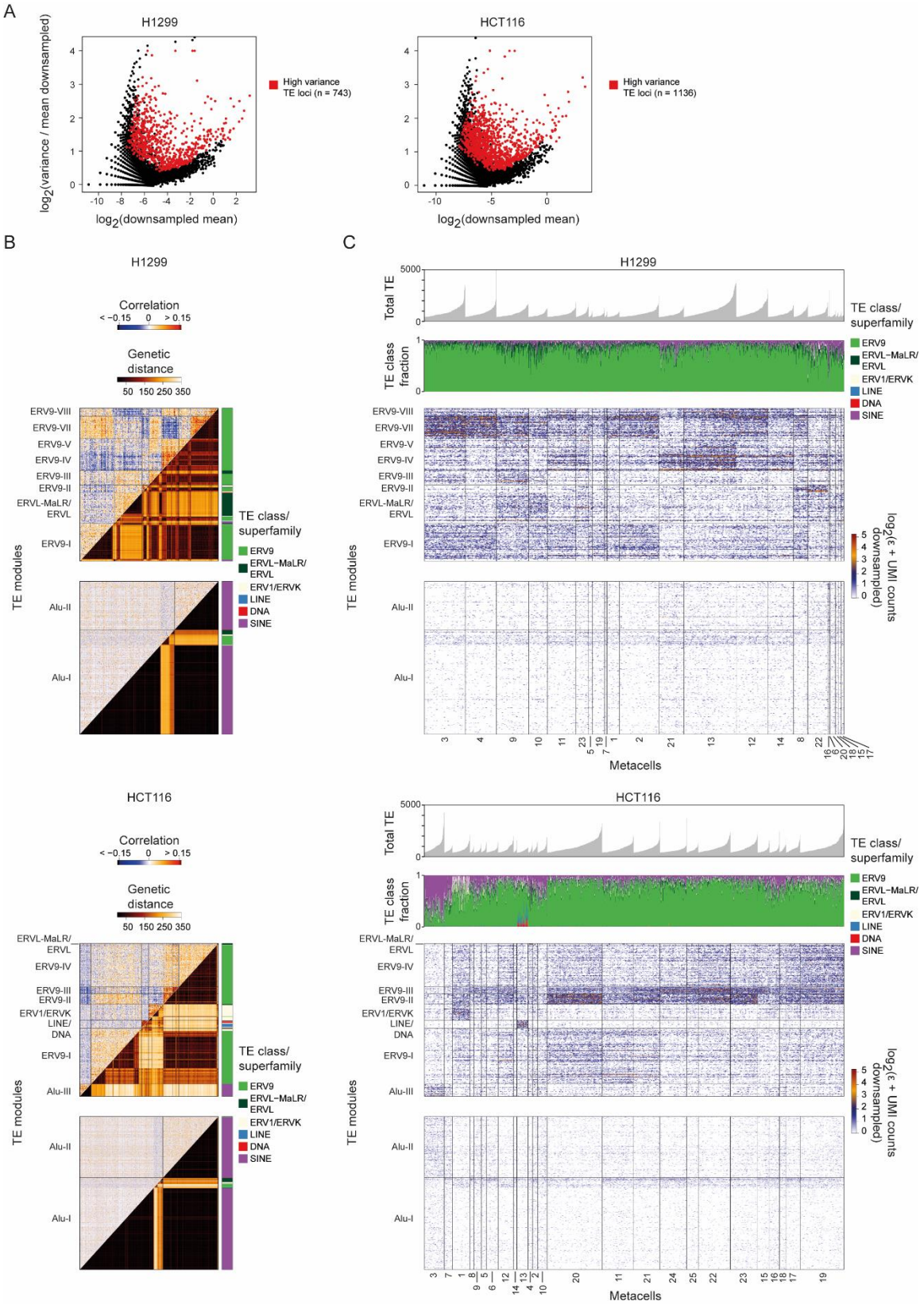

A
H1299
HCT116
High variance TE loci (n = 743)
High variance TE loci (n = 1136)
log2(variance / mean downsampled)
log2(downsampled mean)
B
C
Correlation
< −0.15 0 > 0.15
Genetic distance
50 150 250 350
TE modules
TE class/ superfamily
ERV9
ERVL−MaLR/ ERVL
ERV1/ERVK
LINE
DNA
SINE
Total TE
TE class fraction
log2(ε + UMI counts downsampled)
Metacells

Figure 3 | Meta-grouping of cells with consistent TE expression identifies structured variation in TE transcription

A) Normalized variance vs. mean UMI counts in down-sampled TE TSSs UMI matrices.
B) Single-cell expression correlation (upper triangle) and pairwise genetic divergence (lower triangle) for 10 TE TSSs modules in H1299 (top) and HCT116 (bottom) cells. SINE and non-SINE modules are plotted separately for clarity. Color annotation represent a loci's TE class or superfamily membership.
C) H1299 (upper) and HCT116 (lower) single-cell expression matrix showing the rolling mean (five cells) of downsampled and log2 transformed UMI counts. Rows represent individual loci arranged by TE modules and columns correspond to individual cells grouped by metacells and arranged by total TE load. The two upper bar plots indicate a cell's absolute TE load and the contribution of each TE class to it.

Figure 4

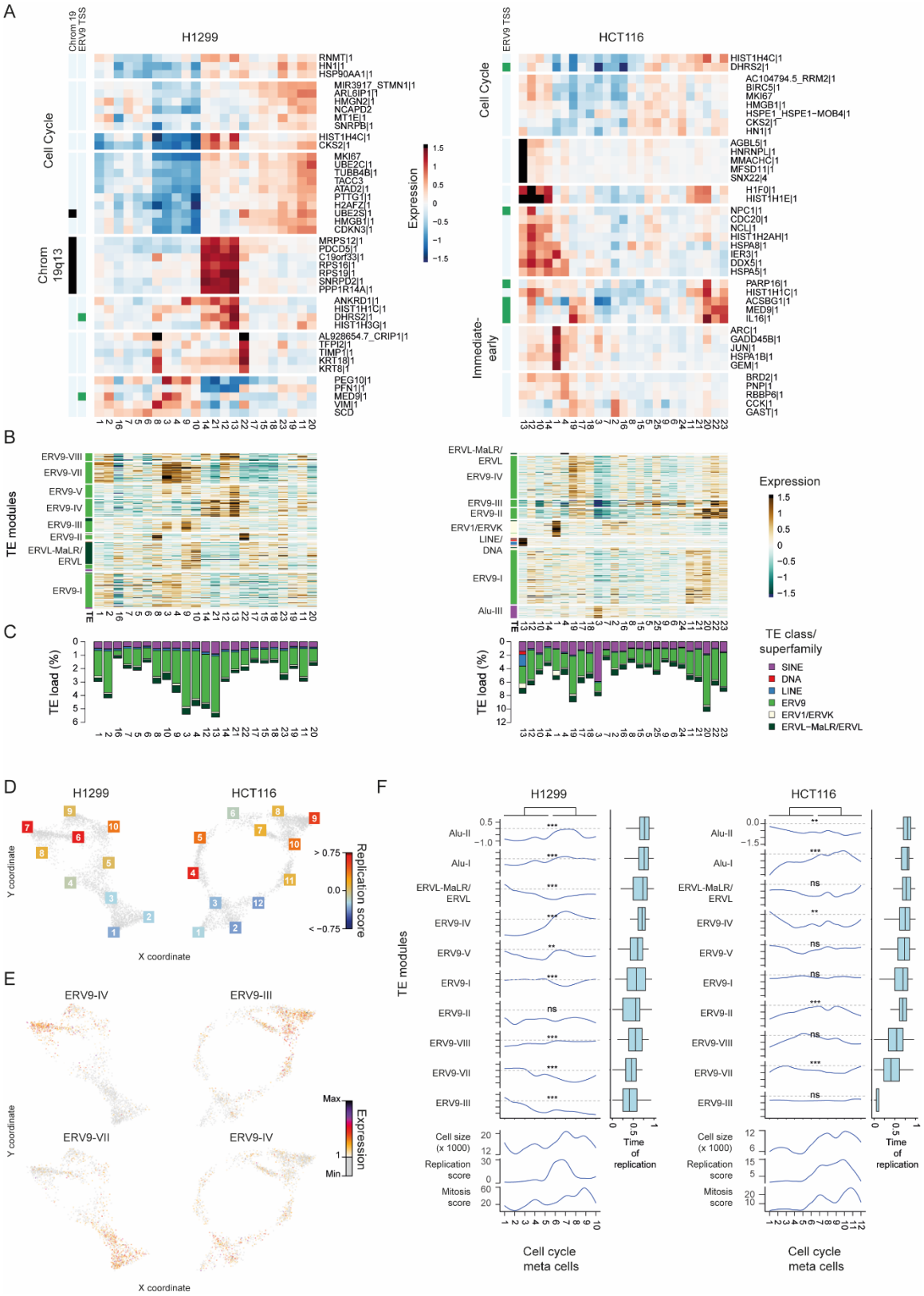

A
H1299
HCT116
Chrom 19 ERV9 TSS
ERV9 TSS
Cell Cycle
Chrom 19q13
Immediate-early
Expression
B
TE modules
ERV9-VIII
ERV9-VII
ERV9-V
ERV9-IV
ERV9-III
ERV9-II
ERVL-MaLR/ ERVL
ERV9-I
ERV1/ERVK
LINE/ DNA
Alu-III
TE
C
TE load (%)
TE class/ superfamily
SINE
DNA
LINE
ERV9
ERV1/ERVK
ERVL-MaLR/ERVL
D
Y coordinate
X coordinate
Replication score
E
Max
Min
F
Alu-II
Alu-I
Cell size (x 1000)
Replication score
Mitosis score
Time of replication
Cell cycle meta cells

Figure 4 | TE meta-cells are linked to the cell cycle and other gene-modules

A) Shown are enrichment values for genes (rows) over TE metacells (columns). Left color bars indicate genic TSS overlap with ERV9 family members (green) and mapping to chromosome 19 in H1299. Metacell numbers (bottom) are consistent with Figure 3.

B) Enrichment values for TE loci (rows) across metacells (columns). Rows are grouped by module (as in Fig 3). Colored bars represent the superfamily/class membership of a locus. Alu modules are not shown.

A) Average TE load per metacell. Colors indicate the contribution of individual TE classes to the total.

B) Single-cell 2D projection based on cell cycle transcriptome signatures. Grey dots represent individual cells. The color of the corresponding metacell ID (squares positioned at the centroid) is based on the average expression footprint of presumed mitosis-specific genes (top 10 genes with highest footprint in mitotic metacell).

C) Single-cell 2D projection as before but colors indicate summarized expression intensity of indicated TE modules.

D) Cell cycle metacells (columns) were arranged according to their estimated cell cycle phase from left (G0/G1) to right (mitotic exit) in a supervised fashion. Upper rows show the smoothed enrichment of aggregated TE modules across metacells. TE modules were arranged by their median time of replication (right boxplots). Mitosis and replication expression scores (lower rows) summarize downsampled UMI counts of the top 10 genes in the respective cell cycle gene module. Significance is based on Wilcoxon rank-sum test comparing cells from metacells #1-5 (#1-6 in HCT116) against #6-10 (#7-12 in HCT116).

Figure 5

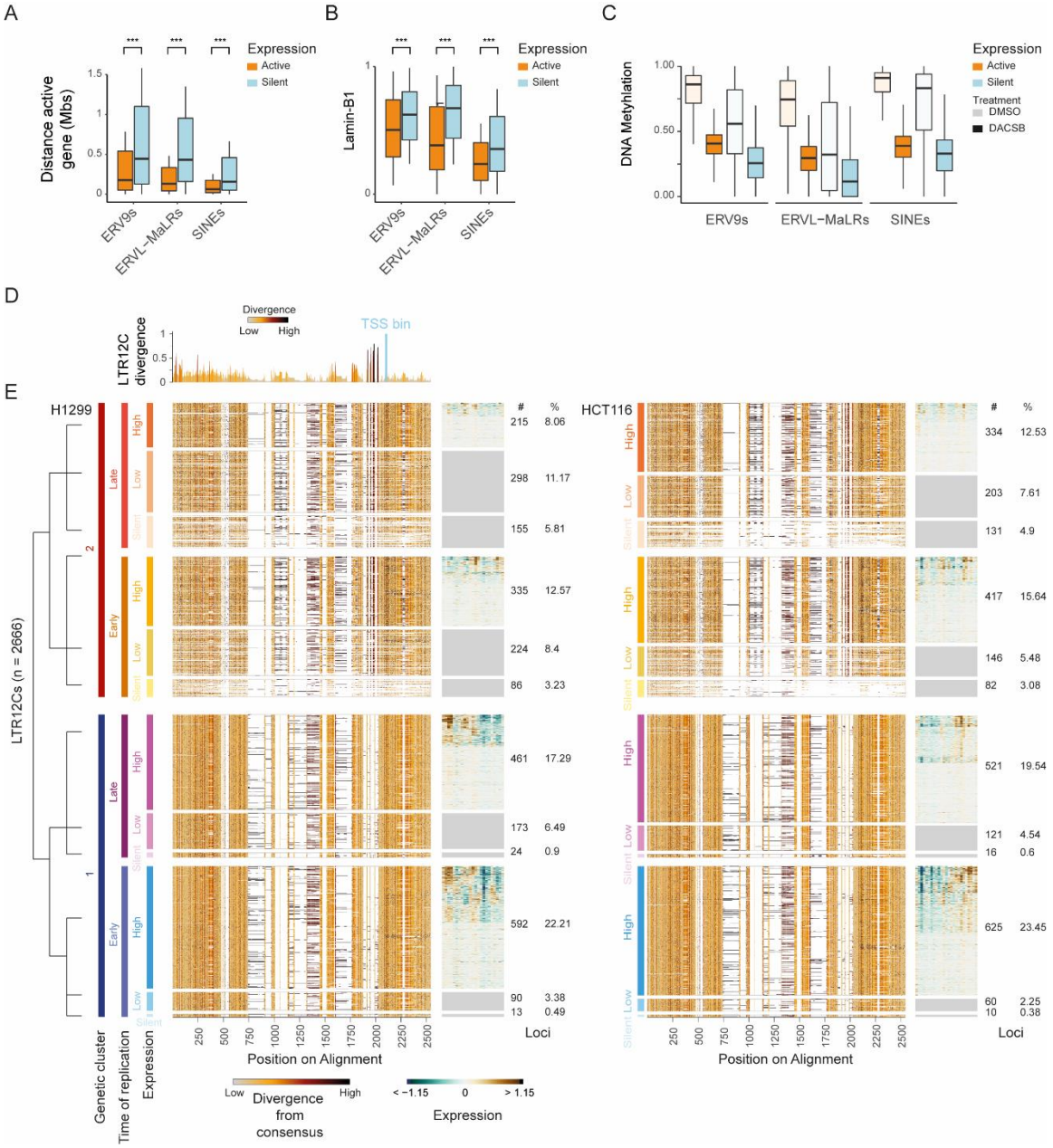


Figure 5 | Sequence composition and genomic context explain TE transcriptional activation and variability

A) Boxplot showing the distance to the nearest identified genic TSS for active (>10 TSS UMIs, orange) and silent (0 TSS and non-TSS UMIs, blue) TE loci grouped by TE class/superfamily membership in H1299 cells. Boxplot whiskers extend 0.5 times the interquartile range from the 25th and 75th percentiles and loci more than 2 mega bases away are not shown.

B) Lamina-association of active (orange) and silent (blue) TE loci grouped by TE class/superfamily in H1299 cells.

C) Distribution of average methylation for active (orange) and silent (blue) TE loci before (light colors) and after treatment (dark colors) grouped TE class/superfamily membership in H1299 cells.

D) Sequence divergence between two genetic subgroups of LTR12C elements defined as the maximal delta in base frequencies per position. Colors indicate the degree of divergence.

E) Multiple sequence alignment (left) and TE metacell expression (right) for 2666 LTR12C loci hierarchically grouped by genetic sequence (clusters #1 and #2), time of replication (early and late), and transcriptional activity (silent (0 UMIs), low (less than 10 TSS UMIs) and high (at least 11 TSS UMIs)). Loci (rows) within a group are arranged according to K-means clustering of TE metacell footprints. Numbers to the right indicate the number of loci and their percentage from total for each group. Colors either represent the position-wise genetic divergence from the consensus or the TE metacell footprint. Alignment is identical for H1299 (left panel) and HCT116 (right panel) cells, but the order may differ according to cell-line specific expression dynamics.

REFERENCES

1. Rebollo, R., Romanish, M. T. & Mager, D. L. Transposable Elements: An Abundant and Natural Source of Regulatory Sequences for Host Genes. *Annu. Rev. Genet.* (2012). doi:10.1146/annurev-genet-110711-155621

2. Feschotte, C. Transposable elements and the evolution of regulatory networks. *Nature Reviews Genetics* (2008). doi:10.1038/nrg2337

3. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory activities of transposable elements: From conflicts to benefits. *Nature Reviews Genetics* (2017). doi:10.1038/nrg.2016.139

4. Fuentes, D. R., Swigut, T. & Wysocka, J. Systematic perturbation of retroviral LTRs reveals widespread long-range effects on human gene regulation. *Elife* (2018). doi:10.7554/eLife.35989

5. Karimi, M. M. *et al.* DNA methylation and SETDB1/H3K9me3 regulate predominantly distinct sets of genes, retroelements, and chimeric transcripts in mescs. *Cell Stem Cell* (2011). doi:10.1016/j.stem.2011.04.004

6. Slotkin, R. K. & Martienssen, R. Transposable elements and the epigenetic regulation of the genome. *Nature Reviews Genetics* (2007). doi:10.1038/nrg2072

7. Sundaram, V. *et al.* Widespread contribution of transposable elements to the innovation of gene regulatory networks. *Genome Res.* (2014). doi:10.1101/gr.168872.113

8. Jordan, I. K., Rogozin, I. B., Glazko, G. V. & Koonin, E. V. Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends in Genetics* (2003). doi:10.1016/S0168-9525(02)00006-9

9. Hua-Van, A., Le Rouzic, A., Maisonhaute, C. & Capy, P. Abundance, distribution and dynamics of retrotransposable elements and transposons: Similarities and differences. *Cytogenetic and Genome Research* (2005). doi:10.1159/000084975

10. Chiappinelli, K. B. *et al.* Inhibiting DNA Methylation Causes an Interferon Response in Cancer via dsRNA Including Endogenous Retroviruses. *Cell* (2015). doi:10.1016/j.cell.2015.07.011

11. Roulois, D. *et al.* DNA-Demethylating Agents Target Colorectal Cancer Cells by Inducing Viral Mimicry by Endogenous Transcripts. *Cell* (2015). doi:10.1016/j.cell.2015.07.056

12. Brocks, D. *et al.* DNMT and HDAC inhibitors induce cryptic transcription start sites encoded in long terminal repeats. *Nat. Genet.* **49,** 1052–1060 (2017).

13. Tanay, A. & Regev, A. Scaling single-cell genomics from phenomenology to mechanism. *Nature* **541,** 331–338 (2017).

14. Shapiro, E., Biezuner, T. & Linnarsson, S. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Reviews Genetics* (2013). doi:10.1038/nrg3542

15. Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C. & Teichmann, S. A. The Technology and Biology of Single-Cell RNA Sequencing. *Molecular Cell* (2015). doi:10.1016/j.molcel.2015.04.005

16. Bock, C., Farlik, M. & Sheffield, N. C. Multi-Omics of Single Cells: Strategies and Applications. *Trends in Biotechnology* (2016). doi:10.1016/j.tibtech.2016.04.004

17. Rodriguez-Terrones, D. & Torres-Padilla, M. E. Nimble and Ready to Mingle: Transposon Outbursts of Early Development. *Trends in Genetics* (2018). doi:10.1016/j.tig.2018.06.006

18. Issa, J. P. Aging and epigenetic drift: A vicious cycle. *Journal of Clinical Investigation* (2014). doi:10.1172/JCI69735

19. Jones, P. A., Issa, J. P. J. & Baylin, S. Targeting the cancer epigenome for therapy. *Nature Reviews Genetics* (2016). doi:10.1038/nrg.2016.93

20. Sebé-Pedrós, A. *et al.* Cnidarian Cell Type Diversity and Regulation Revealed by Whole-Organism Single-Cell RNA-Seq. *Cell* (2018). doi:10.1016/j.cell.2018.05.019

21. Strick, R., Strissel, P. L., Baylin, S. B. & Chiappinelli, K. B. Unraveling the molecular pathways of DNA-methylation inhibitors: Human endogenous retroviruses induce the innate immune response in tumors. *Oncoimmunology* (2016). doi:10.1080/2162402X.2015.1122160

22. Faulkner, G. J. *et al.* The regulated retrotransposon transcriptome of mammalian cells. *Nat. Genet.* (2009). doi:10.1038/ng.368

23. Göke, J. *et al.* Dynamic transcription of distinct classes of endogenous retroviral elements marks specific populations of early human embryonic cells. *Cell Stem Cell* (2015). doi:10.1016/j.stem.2015.01.005

24. Hashimoto, K. *et al.* CAGE profiling of ncRNAs in hepatocellular carcinoma reveals widespread activation of retroviral LTR promoters in virus-induced tumors. *Genome Res.* (2015). doi:10.1101/gr.191031.115

25. Shipony, Z. *et al.* Dynamic and static maintenance of epigenetic memory in pluripotent and somatic cells. *Nature* (2014). doi:10.1038/nature13458

26. Santos, F., Hendrich, B., Reik, W. & Dean, W. Dynamic reprogramming of DNA methylation in the early mouse embryo. *Dev. Biol.* (2002). doi:10.1006/dbio.2001.0501

27. Messerschmidt, D. M., Knowles, B. B. & Solter, D. DNA methylation dynamics during epigenetic reprogramming in the germline and preimplantation embryos. *Genes and Development* (2014). doi:10.1101/gad.234294.113

28. Smit, A., Hubley, R. & Green, P. RepeatMasker Open-4.0. 2013-2015 . *http://www.repeatmasker.org* (2013).

29. Smith, T., Heger, A. & Sudbery, I. UMI-tools: Modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res.* (2017). doi:10.1101/gr.209601.116

30. Baran, Y. *et al.* MetaCell: analysis of single cell RNA-seq data using k-NN graph partitions. *bioRxiv* (2018). doi:10.1101/437665

31. Liu, Z. *et al.* Reconstructing cell cycle pseudo time-series via single-cell transcriptome data. *Nat. Commun.* (2017). doi:10.1038/s41467-017-00039-z

32. Scialdone, A. *et al.* Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods* (2015). doi:10.1016/j.ymeth.2015.06.021

33. Dekker, J. *et al.* The 4D nucleome project. *Nature* (2017). doi:10.1038/nature23884

34. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* (2013). doi:10.1093/molbev/mst010

35. Rusinova, I. *et al.* INTERFEROME v2.0: An updated database of annotated interferon-regulated genes. *Nucleic Acids Res.* (2013). doi:10.1093/nar/gks1215