scientific reports

Check for updates

OPEN

scTEM-seq: Single-cell analysis of transposable element methylation to link global epigenetic heterogeneity with transcriptional programs

Kooper V. Hunt[1,2], Sean M. Burnard[1,2], Ellise A. Roper[1,2], Danielle R. Bond[1,2], Matthew D. Dun[1,2], Nicole M. Verrills[1,2], Anoop K. Enjeti[1,2,3,4] & Heather J. Lee[1,2]✉

Global changes in DNA methylation are observed in development and disease, and single-cell analyses are highlighting the heterogeneous regulation of these processes. However, technical challenges associated with single-cell analysis of DNA methylation limit these studies. We present single-cell transposable element methylation sequencing (scTEM-seq) for cost-effective estimation of average DNA methylation levels. By targeting high-copy SINE Alu elements, we achieve amplicon bisulphite sequencing with thousands of loci covered in each scTEM-seq library. Parallel transcriptome analysis is also performed to link global DNA methylation estimates with gene expression. We apply scTEM-seq to KG1a acute myeloid leukaemia (AML) cells, and primary AML cells. Our method reveals global DNA methylation heterogeneity induced by decitabine treatment of KG1a cells associated with altered expression of immune process genes. We also compare global DNA methylation estimates to expression of transposable elements and find a predominance of negative correlations. Finally, we observe co-ordinated upregulation of many transposable elements in a sub-set of decitabine treated cells. By linking global DNA methylation heterogeneity with transcription, scTEM-seq will refine our understanding of epigenetic regulation in cancer and beyond.

Single-cell analysis of DNA methylation has revealed epigenetic heterogeneity in development and disease, and parallel transcriptomic analyses are allowing this heterogeneity to be understood in the context of genomic regulation[1,2]. For example, single-cell analysis of DNA methylation, chromatin accessibility and gene expression has demonstrated that active epigenetic remodelling is required for endoderm and mesoderm specification during gastrulation[3]. In contrast, the ectoderm lineage is epigenetically primed in the epiblast and serves as a default differentiation pathway. Similar analyses have been applied to colorectal cancer revealing relationships between somatic copy number alterations, DNA methylation and gene expression[4]. While genetic sub-lineages were found to have distinct epigenetic profiles, comparison between primary and metastatic sites suggested that epigenetic reprogramming was not essential for tumour dissemination.

The studies described above demonstrate the power of linking DNA methylation heterogeneity with genetic and transcriptional heterogeneity. However, technical challenges continue to limit the implementation of single-cell DNA methylation analyses. Most methods rely on bisulphite conversion to distinguish methylated from unmodified cytosines. This chemistry provides single-nucleotide resolution but is incompatible with available high-throughput droplet barcoding technologies. Thus, single-cell analysis of DNA methylation is currently limited to low-throughput multi-well plate assays that are relatively high cost. Furthermore, genome-wide bisulphite sequencing (BS-seq) requires ten times as many reads as RNA sequencing (RNA-seq), meaning that studies on thousands of cells are usually cost-prohibitive. Finally, the sparse data obtained from single-cell BS-seq (scBS-seq) and single-cell RNA-seq (scRNA-seq) libraries poses a major challenge to multi-omic studies hoping to identify individual loci where DNA methylation correlates with gene expression. In the study of colorectal cancer

[1]School of Biomedical Sciences and Pharmacy, College of Health Medicine and Wellbeing, University of Newcastle, Callaghan, NSW, Australia. [2]Medical Genetics, Level 3 West, Hunter Medical Research Institute, Lot 1 Kookaburra Circuit, New Lambton Heights, NSW 2305, Australia. [3]NSW Health Pathology North, New Lambton Heights, NSW, Australia. [4]Calvary Mater Newcastle, Waratah, NSW, Australia. ✉email: Heather.Lee@newcastle.edu.au

mentioned above[4], promoters with differential DNA methylation between primary tumour and metastatic sites were identified, but no correlations to expression of the associated genes were reported. Indeed, the most exciting findings from this study were related to global changes in DNA methylation, as opposed to locus-specific effects.

We reasoned that assessment of global DNA methylation in single cells would be a useful alternative to genome-wide analyses in contexts such as embryonic development and cancer, and reckoned that transposable element (TE) methylation could be exploited for this purpose. TEs are conserved DNA sequences capable of replicating and inserting into new positions in the genome. Discovered by Barbara McClintock in 1950[5], TEs are estimated to make up around half of the human genome[6]. Poly-A retrotransposons Long Interspersed Element 1 (LINE-1) and Short Interspersed Element Alu (SINE Alu) account for almost 25% of the genome and are some of the only active or 'hot' TEs still capable of transposing in our genome[7,8]. Active retrotransposition causes genome instability, and because of this mutagenic potential, TEs are epigenetically silenced by high DNA methylation levels in internal promoters. Since TEs are so abundant in mammalian genomes, global changes in DNA methylation are correlated to changes in TE methylation in early embryonic development[9], primordial germ cell development[10], induced pluripotent stem cell (iPSC) reprogramming[11] and cancer[12]. Indeed, even in the single-cell analysis of colorectal cancer discussed above, lineage-specific global DNA hypomethylation was associated with an over-representation of TE sequences (LTRs, LINEs)[4].

These observations justify the use of TEs as surrogate measures for global DNA methylation levels, and LINE-1 and SINE Alu elements are common targets for bisulphite conversion-based analysis[13,14]. Here we adapt this approach for cost-effective analysis of global DNA methylation levels in a method called single-cell transposable element methylation sequencing (scTEM-seq). To achieve this, we perform targeted amplification of bisulphite converted SINE Alu and LINE-1 sequences. We apply scTEM-seq in acute myeloid leukaemia (AML) cells and detect global DNA methylation heterogeneity following treatment with the hypomethylating agent (HMA), decitabine (DAC). Parallel analysis of gene expression in the same single cells identifies links to immune processes, translation and induction of TE expression.

Results

To investigate whether TEs might serve as surrogate measures for global DNA methylation levels in single-cell data, we first interrogated genome-wide scBS-seq data from a colorectal cancer patient (CRC01)[4]. We observed very strong correlations between DNA methylation within TE annotations and global methylation averages for both LINE-1 ($R^2 = 0.88$, $p < 2.2^{-16}$) and SINE Alu ($R^2 = 0.91$, $p < 2.2^{-16}$) families (Supplementary Fig. S1A). Furthermore, TE methylation was sufficient to identify sub-clonal differences in global DNA methylation (Supplementary Fig. S1B)[4]. This demonstrates that TE methylation in single-cell data can highlight biologically interesting heterogeneity in cancer cells.

We adapted the scBS-seq protocol and achieved amplification of SINE Alu and LINE-1 sequences following bisulphite conversion of single-cell DNA samples (Fig. 1A, Supplementary Fig. S2A). LINE-1 primers used in previous studies[15], and SINE Alu primers designed against an AluYa5 consensus sequence, were modified to be compatible with amplicon sequencing (see "Methods" and Supplementary Table S1). In initial experiments, SINE Alu primers (Sine.Alu_F, Sine.Alu_R) delivered greater amplicon yield and library complexity than LINE-1 primers (LINE.L1_F, LINE.L1_R), consistent with the higher copy-number of SINE Alu elements (Supplementary Fig. S2A). A second generation of 28 SINE Alu primer sequences were then designed (Supplementary Fig. S2, Supplementary Tables S1), and unique primer pairs were arranged in a 96-well plate (Supplementary Table S2). An 8 bp index was included in each primer, such that every library carried a dual index in the adaptor sequence, and a second internal dual index at the start of sequencing reads. This means that up to 18,432 scTEM-seq libraries can be pooled for sequencing. In addition, a 0–5 N spacer was included in these primers, and the direction of amplicon sequencing was reversed in 50% of primer pairs, to ameliorate the technical challenges of sequencing low-diversity amplicon libraries. For all further experiments, SINE Alu elements were targeted in scTEM-seq analysis.

We applied our optimised scTEM-seq protocol to Acute Myeloid Leukaemia (AML) cells treated with Decitabine (DAC); a hypomethylating agent (HMA) used to treat elderly AML patients[16]. DAC is a cytidine analogue that is incorporated into DNA and causes genome-wide loss of DNA methylation by blocking its maintenance during DNA replication. While some studies have demonstrated durable responses in some patients, clinical use of this epigenetic therapy is limited by variability in patient response[16].

KG1a AML cells were treated with and without 100 nM DAC for 72 h, and single cells were collected by FACS for scTEM-seq. Negative controls (no cell) were also included to monitor DNA contamination in reagents. Average amplicon yield from single cell samples was 16.10 ng/µl (4.08 SD (standard deviation)), compared to 1.12 ng/µl (1.03 SD) in negative controls (Supplementary Fig. S3A). scTEM-seq libraries achieved unique alignment rates of 67.23% (5.11 SD) (Supplementary Fig. S3C), and efficient bisulphite conversion was confirmed by very low non-CpG methylation rates (DNA methylation in CHG trinucleotide contexts was 0.67%, 0.2 SD) (Supplementary Table S3). Information was recovered from between 1000 and 6000 unique SINE Alu annotations for each cell, despite low sequencing depth (14,000–37,000 raw reads per cell) (Fig. 1B). Further analysis confirmed that scTEM-seq reads were predominantly focused on AluY elements, though other SINE Alu families were also represented in the data (Supplementary Fig. S4).

In untreated KG1a cells, scTEM-seq showed uniformly high levels of DNA methylation, with an average of 85.4% (1.65 SD). In DAC treated cells, a heterogeneous loss of DNA methylation was observed, with levels ranging from 29 to 69% (average 41.86%, 10.46 SD) (Fig. 1C). To assess the accuracy of these DNA methylation measurements, we first compared our scTEM-seq results to genome-wide methylation levels in bulk sequencing libraries prepared from matched populations of cells. The average methylation rate for all CpGs covered in bulk libraries was 78.58% for untreated cells and 43.87% for DAC treated cells. As expected, CpGs within

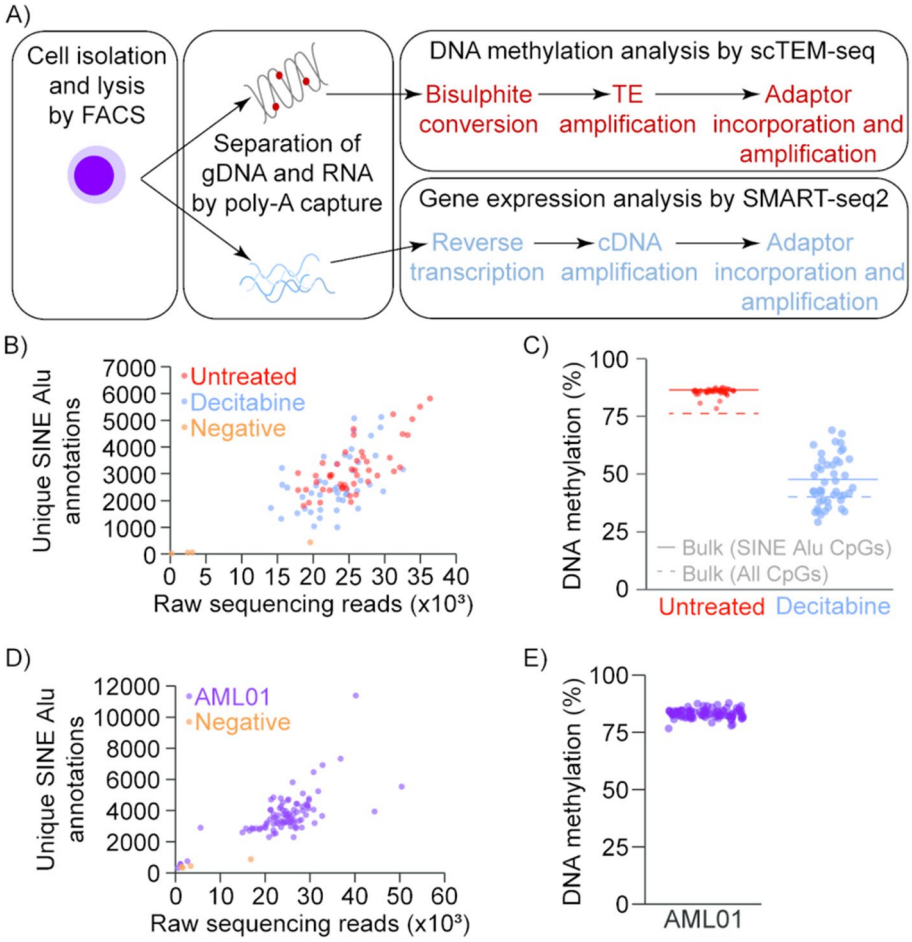

Figure 1. scTEM-seq accurately measures DNA methylation at TE sites. **(A)** Schematic representation of combined scTEM-seq and scRNA-seq workflow. **(B)** Unique SINE Alu sites measured in KG1a cells compared to raw sequencing reads. **(C)** DNA methylation levels as measured by scTEM-seq in KG1a cells with and without DAC treatment. Coloured lines show average DNA methylation levels at SINE Alu sites for each treatment group measured in bulk samples. DAC treated KG1a cells show a heterogeneous loss of DNA methylation. **(D)** Unique SINE Alu sites measured in AML01 patient blasts compared to raw sequencing reads. **(E)** DNA methylation levels in untreated AML01 patient blasts measured by scTEM-seq.

SINE Alu sites had higher average methylation levels at 86.48% and 47.63% in untreated and DAC treated cells, respectively (Fig. 1C). For untreated cells, 42 of 46 scTEM-seq libraries had methylation estimates within ± 2% of the expected value based on bulk libraries (86.48%). Thus, SINE Alu analysis by scTEM-seq provides accurate DNA methylation estimates.

To validate our observation of DAC induced DNA methylation heterogeneity, we compared scTEM-seq to an established method. The range and variance of DNA methylation values were similar for genome-wide single-cell bisulphite sequencing (scBS-seq)[17] and scTEM-seq libraries (Supplementary Fig. S5A). Furthermore, scTEM-seq analysis of HL60 cells treated with and without DAC showed similar patterns of DNA methylation to KG1a cells (Supplementary Fig. S5B).

We also performed a bootstrapping analysis of scTEM-seq data to test the stability of DNA methylation estimates at low read counts. A slight bias toward increased methylation estimates at low read count was observed, especially in cells with high methylation rates (Supplementary Fig. S5). This is likely to result from more efficient amplification of methylated sequences, which is known to influence bisulphite sequencing libraries[18]. Nonetheless, methylation estimates were stable within a 3% range for all sub-samples of > 5000 aligned reads, demonstrating that scTEM-seq is a reliable measure of DNA methylation.

To test scTEM-seq in primary human cells, we applied our analysis on sorted blasts from an AML patient. Amplicon yield, alignment rates and bisulphite conversion were comparable to KG1a samples, and 88 of 92 libraries passed quality control with representation of > 1000 unique SINE Alu elements (Supplementary Table S4). This patient had not received hypomethylating agent therapy, and DNA methylation at SINE Alu elements was consistently high in these cells (84.74%, 2.15 SD) (Fig. 1E).

Prior separation of gDNA and RNA allowed us to prepare matched scRNA-seq libraries from each cell using the SMART-seq2 protocol[17,19] (Fig. 1A, Supplementary Tables S5 and S6). We then correlated DNA methylation

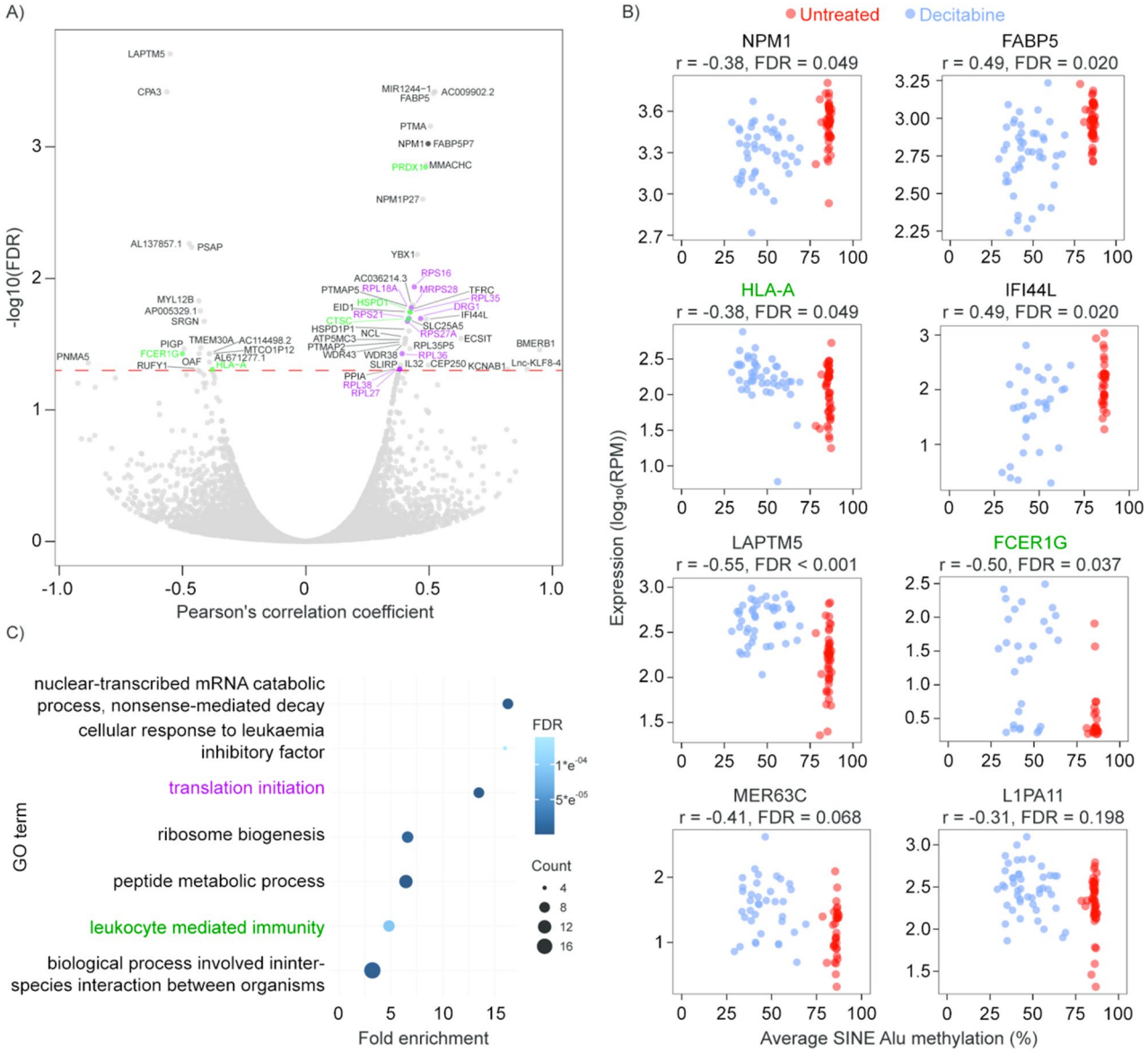

Figure 2. Correlations between average DNA methylation levels and gene expression. **(A)** Volcano plot showing Pearson's correlation between average DNA methylation in SINE Alu elements and gene expression in the KG1a dataset. Genes involved in 'translational initiation' and 'leukocyte mediated immunity' are highlighted in purple and green, respectively. **(B)** Select examples showing expression levels of an individual gene and average DNA methylation levels in our treated and untreated KG1a cells. Examples include 6 genes (*NPM1*, *FABP5*, *HLA-A*, *IFI44L*, *LAPTM5* and *FCER1G*) and 2 TEs (MER63C and L1PA11). The Pearson's correlation coefficient (r) and false discovery rate (FDR) for each correlation are shown. RPM = reads per million. **(C)** Gene ontology (Panther) results for statistically overrepresented biological pathways in all genes with expression correlated to DNA methylation (FDR < 0.05). For related terms, only the pathway with the highest number of correlated genes is displayed for simplicity.

levels to changes in gene expression. In our KG1a data, expression of 60 genes was correlated to the average DNA methylation at SINE Alu sites (FDR < 0.05 after multiple testing correction), with the majority (43) showing positive associations (Fig. 2A). For example, interferon induced protein *IFI44L* was down-regulated in cells with lower SINE Alu methylation, whereas major histocompatibility complex (MHC) I component *HLA-A* was up-regulated in cells with lower SINE Alu methylation (Fig. 2B). Gene ontology analysis on all genes with a significant correlation to DNA methylation (FDR < 0.05) revealed over-represented of pathways including: translational initiation, leukocyte mediated immunity, and biological process involved in interspecies interaction between organisms (Fig. 2C). These results are consistent with the ability of HMAs to induce differentiation[20], and inhibit translation[21], in AML cells.

HMAs have also been shown to act through a 'viral mimicry' process whereby loss of DNA methylation induces transcription of TEs (e.g. endogenous retroviruses, LINEs and SINEs) and a subsequent type 1 interferon response in effected cells[22–24]. To test whether scTEM-seq could link epigenetic heterogeneity induced by DAC

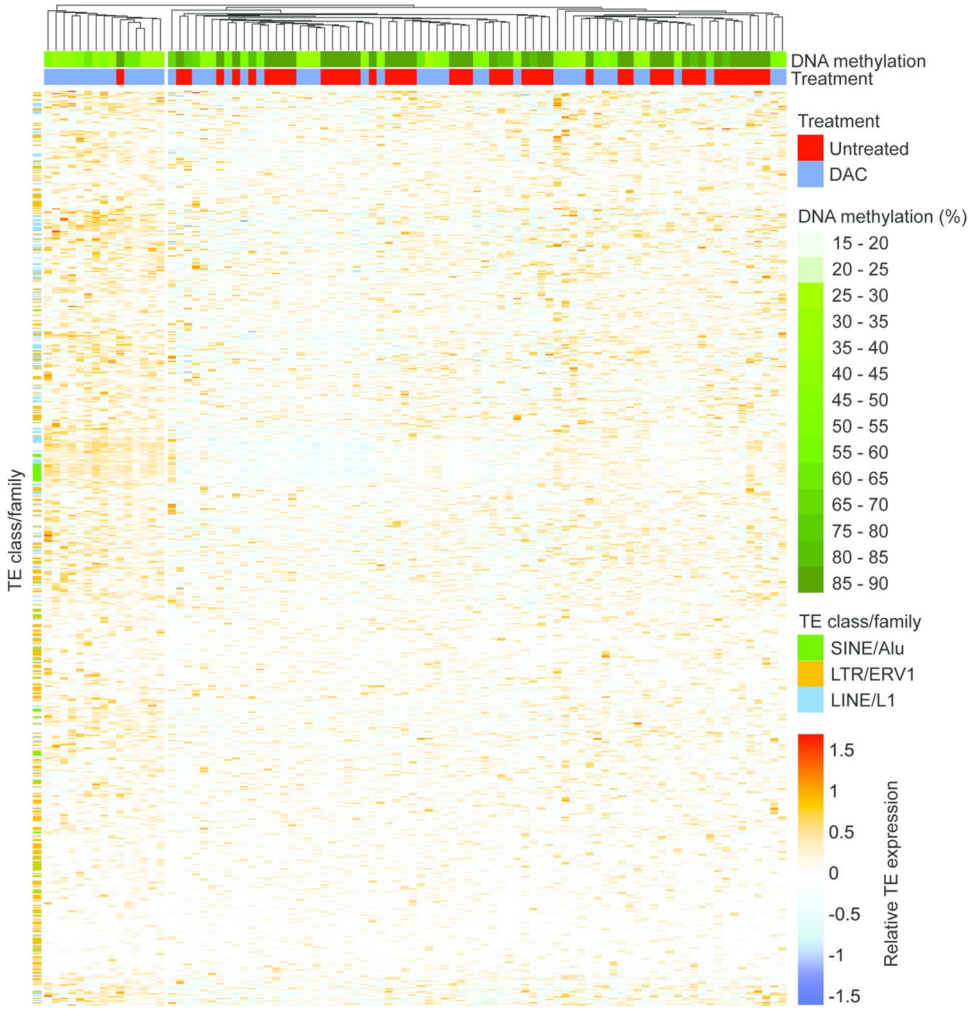

Figure 3. Coordinated up-regulation of TE transcription is observed in a subset of DAC treated KG1a cells. TE families with altered expression between untreated (red) and DAC treated (blue) KG1a cells were identified by differential expression analysis using DEseq2. The heatmap shows the relative expression of all TEs from significant families (adjusted $p < 0.05$) following normalisation by variance stabilisation transformation (vst) (DESeq2) and mean centering. Both rows (TEs) and columns (cells) are clustered by Euclidean distance. Global DNA methylation percentages for each cell are indicated (green scale at top) and selected TE families are highlighted (left). In total, 11 TE families reached the significance threshold (Family:Class; acro:Satellite, ERV1:LTR, ERVK:LTR, L1:LINE, Alu:SINE, ERVL:LTR, ERVL-MaLR:LTR, TcMar-Tigger:DNA, hAT-Charlie:DNA, MIR:SINE, L2:LINE), corresponding to 834 TE elements. A sub-cluster of mostly DAC treated cells (left) have high expression of TEs.

to expression of TEs, we assessed the abundance of TE sequences in RNA-seq data. KG1a cells showed a clear increase in TE expression levels after DAC treatment (Supplementary Figure S6A), and a bias toward negative correlations between TE expression and DNA methylation (although no TEs had significant correlations after multiple testing correction, see examples in Fig. 2B). A trend toward negative correlations was also observed in AML01 and HL60 datasets, with 4 TE families showing significant (FDR < 0.05) correlations in HL60 cells (Supplementary Fig. S6B). To further investigate TE expression patterns, we performed clustering analysis of TE families that were differentially expressed after DAC treatment (Fig. 3, Supplementary Figs. S7 and S8). In KG1a cells, we observed a subgroup of mostly DAC treated cells with co-ordinated up-regulation of many TEs, especially LINE-1 and SINE Alu families (Fig. 3). A similar pattern was observed in HL60 cells (Supplementary Fig. S8). Interestingly, cells with high TE expression could not be distinguished from other DAC treated cells based on global DNA methylation alone (KG1a: 46.4%, 10.9 SD vs 46.8%, 10.4 SD , respectively) (Supplementary Fig. S9), suggesting that other factors must regulate TE expression in the absence of DNA methylation.

Discussion

TEs have been widely targeted for surrogate measures of global DNA methylation. We have adapted this approach to single cells, developing a cost-effective alternative to genome-wide techniques[17,25–29]. While other studies have amplified loci of interest in bisulphite converted DNA from single cells[30–34], ours is the first to target TEs.

We demonstrate that methylation of SINE Alu elements in single cells compares well to global DNA methylation levels using in silico analysis of published data (Supplementary Fig. S1A) and by comparing scTEM-seq to matched bulk sequencing libraries (Fig. 2C), and established single-cell techniques (Supplementary Fig. S5A). SINE Alu methylation over-estimates global DNA methylation levels (e.g. by 6.8% for untreated KG1a cells) (Fig. 1C), which can be explained by the well-characterised enrichment of TEs in hypermethylated regions[35]. However, scTEM-seq accurately estimates SINE Alu methylation in untreated KG1a cells and detects changes in DNA methylation after DAC treatment (Fig. 1C). In untreated KG1a cells, scTEM-seq libraries had homogeneous SINE Alu methylation rates that were typically within ± 2% of the expected value from bulk measurements in a matched population of cells. Furthermore, down-sampling scTEM-seq libraries revealed that methylation estimates are stable, even at low read depth, for both treated and untreated cells (Supplementary Fig. S5C). scTEM-seq has several advantages over comparable genome-wide methods such as single-cell bisulphite sequencing (scBS-seq; Supplementary Table S6)[25]. scTEM-seq libraries are prepared using sequence-specific primers, rather than random-priming oligos, leading to reduced oligo contamination and improved alignment rates. Indeed, the unique alignment rates for scTEM-seq libraries are surprisingly high considering that repetitive loci are inherently difficult to map in the reference genome. Improved alignment rates confer a cost saving by reducing wastage from sequencing runs; however, an even greater advantage is obtained by reducing the sequencing demand. Whereas scBS-seq libraries require ~ 20 million raw reads per cell to obtain genome-wide information, scTEM-seq libraries can provide a global estimate of DNA methylation from ~ 20 thousand raw reads. Thus, the sequencing cost is 3 orders of magnitude lower for scTEM-seq libraries. Obviously, this reduced cost comes with a considerable loss of information. However, locus-specific analysis of DNA methylation is also difficult in genome-wide libraries, due to the low coverage obtained in each cell (e.g. 10–40% of the genome). Like scBS-seq and other plate-based methods, scTEM-seq is compatible with parallel analysis of gene expression in the same single cell. This allows epigenetic heterogeneity to be linked to transcriptional output. Thus, scTEM-seq will increase the scale of single-cell studies in biological contexts where global changes in DNA methylation are of interest.

In this study, DAC treatment of KG1a AML cells led to heterogeneous loss of DNA methylation and altered expression of many genes (Fig. 2). For example, *HLA-A* and *FCER1G* were negatively correlated to DNA methylation only 72 h after initial treatment, possibly signifying increased monocytic differentiation in cells that have lost DNA methylation[20]. Furthermore, we were able to link epigenetic heterogeneity to expression of TEs, suggesting that variable activation of viral mimicry pathways could influence treatment response. We identified a subgroup of DAC treated KG1a cells with co-ordinated up-regulation of many TE families. This group of cells could not be distinguished based on DNA methylation levels alone, suggesting that loss of methylation is insufficient for activation of viral mimicry. In cells that do not up-regulate TEs, other epigenetic processes may substitute for the suppressive effects of DNA methylation, or transcriptional activators required for TE expression may be absent. Interestingly, recent reports have implicated the histone methyltransferase SETDB1 in suppression of TEs and tumour immunogenicity, including effects in hypomethylated cell line models[36,37].

A major limitation for the clinical use of hypomethylating agents is the variability in patient response. Although azacitidine has been shown to improve survival compared to conventional care, a large proportion of patients receive little or no benefit[38]. Changes in global DNA methylation levels during treatment, measured in bulk tumour samples, have not been able to predict patient response to hypomethylating agents[39,40]. Expression of subsets of evolutionarily young TEs, however, correlates with improved prognosis[23]. Using scTEM-seq, we can take these studies a step further and explore how heterogeneity of DNA methylation and expression of TE subtypes within a tumour contribute to patient prognosis.

We also applied scTEM-seq to primary patient blasts, revealing homogeneous levels of DNA methylation. We did not identify correlations between DNA methylation levels and gene expression in this small set of cells (data not shown). However, we did note a bias toward increased TE expression in cells with lower DNA methylation levels (Supplementary Fig. S6B). This is consistent with previous observations that DNA methylation proximal to TE sites correlates with their expression across different cancer types[41]. Future studies will apply scTEM-seq to many cells from numerous patients to test whether variation in TE methylation may lead to intra-tumoural heterogeneity in TE expression.

scTEM-seq is also relevant to several contexts in stem cell and developmental biology. iPSC reprogramming is a heterogeneous process in which global epigenetic remodelling accompanies reactivation of pluripotency networks[42,43]. Variable DNA methylation in iPSCs raises concerns regarding their safety in clinical regenerative medicine since incorrect reprogramming could lead to cancerous growth[44]. Thus, scTEM-seq may be a useful tool to understand the heterogeneity and assess the quality of iPSCs. Ultimately, scTEM-seq will find applications in many aspects of medicine and biology. The reduced complexity and cost of this approach will also allow multi-dimensional single-cell analysis to be used more often and at scale.

Methods

Cell lines and patient samples. KG1a cells (ATCC, catalog #CCL-246.1) were cultured in Iscove's Modified Dulbecco's Medium (IMDM) (Sigma-Aldrich, catalog # I3390) with 10% fetal bovine serum (FBS). HL60 cells (ATCC, catalog #CCL-240) were cultured in Iscove's Modified Dulbecco's Medium (IMDM) (Sigma-Aldrich, catalog # I3390) with 10% fetal bovine serum (FBS) and 4 mM glutamax (Life Technologies, catalog # 35050061). Routine mycoplasma testing was performed using the MycoAlert Mycoplasma Detection Kit (Lonza, catalog #LT07-318), and cell line validation was performed by the Australian Genome Research Facility using custom microsatellite analysis. Cell lines were treated with 100 nM 5-aza-2'-deoxycitidine (decitabine, DAC) every 24 h (0, 24 and 48 h) and harvested at 72 h.

Experiments involving human samples were approved by the human ethics committees of the Hunter New England Area Health service, and the University of Newcastle, and all methods were performed in accordance

with the relevant guidelines and regulations. The AML patient included in this study (AML01) was recruited at diagnosis through the Calvary Mater Newcastle Hospital, with written informed consent. The patient was a 60-year-old male, diagnosed with secondary AML following chronic myelomonocytic leukaemia. Clinical assessment revealed a complex karyotype including an isochromosome 17q, and mutations in the *ASXL1*, *SETBP1* and *SRSF2* genes. Enriched mononuclear cells were purified from peripheral blood using Lymphoprep density gradient medium (StemCell, catalog # 7851) and SepMate tubes (StemCell, catalog # 85450), and cryopreserved.

Cell sorting. KG1a cells were stained using the PE Annexin V Apoptosis Detection Kit (BD Life Science, catalog # 559763). Live cells (Annexin V^-/7-AAD^-) were sorted into individual wells of a 96 well plate containing lysis buffer 2.5µL RLT Plus Lysis Buffer (QIAGEN, catalog # 1053393) with 1U/µL SUPERase-In (ThermoFisher, catalog # AM2696). Before sorting, bulk KG1a samples of 1,000,000 cells were collected from both the untreated and treated populations for comparison with single cells. HL60 cells were stained with Propinium Iodide (PI) (ThermoFisher, catalog # P1304MP) and live cells (PI^-) were sorted into 96 well plate containing lysis buffer 2.5µL RLT Plus Lysis Buffer with 1U/µL SUPERase-In.

Cryopreserved primary human cells were resuspended in thawing media (IMDM, 20% FBS), washed twice and resuspended. The cells were then rested for 1 h at 37 °C before preparation for flow cytometry. Cells (1×10^6/100 µl) were stained with 1.5 µg/mL propidium iodide (PI, Sigma-Aldrich, P1304MP), 1:20 CD45-PECy7 (2D1, Life Technologies, catalog # 25-9459-42), 1:20 CD33-FITC (WM-53, Life Technologies, catalog # 11-0338-42) and 1:20 CD19-BV711 (SJ25C1, BD Biosciences, catalog # 563036). Single blasts (PI^-/$CD45^{dim}$) were collected in 2.5µL RLT Plus Lysis Buffer containing 1U/µL SUPERase-In in 96 well plates.

Library preparation. We utilised the G&T-seq protocol to separate genomic DNA and RNA from the single-cell samples[45]. Genomic DNA from each cell was purified and bisulphite conversion was performed as described[17], with minor modifications. Bisulphite conversion was carried out using the EZ-96 DNA Methylation-Direct MagPrep Kit (Integrated Sciences, catalog # D5054) with half volumes of the manufacturer's instructions. Bisulphite converted DNA was eluted directly from MagBeads into PCR-mix, and amplification of TEs was performed with MagBeads still in the well. PCR cycling conditions used were 95 °C for 5 min (1 cycle), 98 °C for 20 s, 53 °C for 15 s, 72 °C for 1 min (35 cycles), and 72 °C for 10 min (1 cycle). PCR mix used 7.5 µl 1 × KAPA HiFi hotStart Uracil + ReadyMix (Millennium, catalog # ROC-07959079001) and 0.3 µM primer mix. Primers were targeted to SINE Alu and LINE-1 consensus sequences and included a partial adaptor sequence at the 5' end to enable later indexing with NEBNext dual index oligos (Supplementary Fig. S2A, Supplementary Tables S1 and S2). Second generation primers also included a spacer of 0–5 N, and an 8 bp index sequence between the adaptor and SINE Alu priming sequence. After amplification libraries were purified using a 1.2 × volume of AMPure XP beads (Beckman Coulter, catalog # A63881). All libraries were then quantified using the Qubit dsDNA HS kit (Life Technologies), normalised and pooled to a single tube. Pools were then added to 0.8 µM NEBNext dual index oligos (Genesearch, catalog # E7780S) and 14.5 µl 1 × KAPA HiFi HotStart ReadyMix (Millennium, catalog # ROC-07958935001) for indexing and adaptor addition. PCR cycling conditions used were 98 °C for 45 s (1 cycle), 98 °C for 15 s, 65 °C for 30 s, 72 °C for 30 s (5 cycles), and 72 °C for 5 min (1 cycle). Pools were then purified using 0.9 × volume of Ampure XP beads, normalised and combined for sequencing. Matched scRNA-seq libraries were prepared as described[9,17]. For AML01, 4 columns (30 samples and 2 negative controls) were excluded prior to sequencing due to low library quality after an error in library preparation.

A post-bisulphite adaptor tagging (PBAT) approach[46] was used to prepare bulk genome-wide sequencing libraries from matched populations of cells. Libraries were prepared as described[47], with minor modifications. The 6NR adaptor 2 oligo used during second strand synthesis was modified (5'-CAGACGTGTGCTCTTCCG ATCTNNNNNN-3') to be compatible with NEBNext dual index oligos that were used for library amplification.

Sequencing. Sequencing of bisulphite reads was performed using the Illumina MiSeq platform. Low read depth is required, so for data in this paper sequencing kits with only 4 million reads were used for 192 cells. Library loading concentrations of 8-10 pM were used with a 1% PhiX spike-in. We achieved on average 23,000 read pairs per sample.

scRNA-seq Libraries were sequenced using the NextSeq platform with a loading concentration of 1.5 pM and a 1% PhiX spike-in. We excluded all cells with alignment rates under 80%. With approximately 1,000,000 reads per cell, we measured between 6300 and 15,000 genes in all of our single cell KG1a scRNA-seq libraries (Supplementary Table S5). Gene numbers measured in AML01 cells were more modest, with between 2800 and 5200 genes in cells passing quality control (Supplementary Table S6).

PBAT libraries were sequenced using the MiSeq platform. These libraries were prepared with the intention of measuring global DNA methylation levels and as such were also sequenced with low read depth (~ 100,000 reads per bulk sample).

Data processing and analysis (scTEM-seq). After initial demultiplexing of primary Illumina indexes, Cutadapt (v2.10)[48] was used to demultiplex pools based on custom secondary indexes (Supplementary Table S1). Commands –g and -G were used to pass named forward and reverse index lists as a .fasta file to Cutadapt. Bisulphite reads were trimmed using Trim Galore (v0.6.5)[49]. 10 bp was trimmed from both the 5' and 3' ends to remove remaining adapter sequences from reads. Reads were mapped to Bowtie2 (v 2.4.1)[50] indexed human genome (GRCh38) using Bismark (v0.22.3) in non-directional and paired-end mode[51]. The methylation extraction module from Bismark was then used to produce coverage files for methylation analysis.

Coverage of annotated transposable elements was measured in scTEM-seq data using SeqMonk (v1.46.0)[52]. We excluded cells with coverage of less than 1000 annotated TE sites (or 500 for HL60 cells) using Repbase

annotations. Methylation levels were calculated from .cov files using the mean of all CpG sites covered (Figs. 1C,D, 2B, 3 and Supplementary Figs. S7 and S8).

Data processing and analysis (PBAT). PBAT libraries were trimmed using Trim Galore to remove 9 bp from the 5' end of all reads. Reads were mapped using Bismark in non-directional and paired-end mode. Unmapped reads were re-aligned in single-end mode to account for chimeric reads seen in PBAT libraries[53]. After producing coverage files with the Bismark methylation extraction module, paired and single end alignments for each sample were merged into a single file using the cat (concatenate) command. Downstream analysis was performed using SeqMonk. Genome wide cytosine methylation levels was averaged over 3000 bp tiles. SINE Alu methylation levels were measured over annotated Alu sites using Repbase annotations.

Data processing and analysis (scRNA-seq). scRNA-seq data was trimmed using Trim Galore, with default setting in paired-end mode. Hisat2[54] (v2.1.0) and Samtools[55] (v1.10) were used to convert, map and align unique and ambiguous reads to the human reference genome build GRCh38 from raw fastq reads into bam format. TEtranscripts[56] was used to obtain raw gene and transposable element counts from the unique and ambiguously aligned reads using the GTF files for 1) TEs (http://labshare.cshl.edu/shares/mhammelllab/www-data/TEtranscripts/TE_GTF/) and 2) genes (https://asia.ensembl.org/info/data/index.html; release 101 from the FTP server) in GRCh38 ensembl format. TEtranscripts was run in a Conda[57] environment setup with Python (v3.7.7)[58], Pysam (v0.16.0.1)[59], R-base (v4.0.3) and Bioconductor-Deseq2 (v1.28.0)[60].

Correlation of gene and TE expression to DNA methylation (Fig. 2, Supplementary Fig. S6) was performed using R[61]. Transcripts with at least 2 reads in 10 cells were included in analysis. Read counts for scRNA-seq data were normalised per million reads for each sample and log transformed. Cor.test function using Pearson's method was used to correlate gene and TE transcript counts with DNA methylation levels. P-values for significance of correlation were adjusted for false discovery rates using the p.adjust function and fdr method. Gene ontology was performed on genes of interest from correlation analysis using Panther[62] statistical overrepresentation analysis. Panther's GO biological process complete dataset was used for gene annotation, and expressed genes (at least 10 reads in 2 cells) were used as a reference list for the statistical overrepresentation analysis. Correlation, boxplots, and gene ontology results were plotted using ggplot2 (v3.3.5)[63].

Differential expression analysis was performed in R using DESeq2 (v1.32.0)[60] on genes and TEs at the family level (sum of TE element counts) on cells passing initial library QC and excluding features (genes and TEs) with less than 5 reads in at least 3 cells. Default parameters were used in DESeq2 with the significance threshold set at p adjusted < 0.05. Heatmapping was performed on all TE elements belonging to the 'significantly differentially expressed' TE families. Genes and TE counts (at the element level) were normalised by variance stability transformation (vst) (DESeq2), and the subset of TE elements were extracted, mean centred, and pheatmap (v1.0.12)[64] was used to produce the heatmaps with clustering by Euclidean distance on both rows (TEs) and columns (cells), with additional labels for treatment, corresponding global methylation levels and the TE 'family' each 'element' belongs.

Data availability

Sequencing data has been deposited in GEO database under accession number GSE171029.

Received: 13 September 2021; Accepted: 28 March 2022
Published online: 06 April 2022

References

1. Clark, S. J., Lee, H. J., Smallwood, S. A., Kelsey, G. & Reik, W. Single-cell epigenomics: Powerful new methods for understanding gene regulation and cell identity. *Genome Biol.* **17**, 72. https://doi.org/10.1186/s13059-016-0944-x (2016).
2. Shema, E., Bernstein, B. E. & Buenrostro, J. D. Single-cell and single-molecule epigenomics to uncover genome regulation at unprecedented resolution. *Nat. Genet.* **51**, 19–25. https://doi.org/10.1038/s41588-018-0290-x (2019).
3. Argelaguet, R. *et al.* Multi-omics profiling of mouse gastrulation at single-cell resolution. *Nature* **576**, 487–491. https://doi.org/10.1038/s41586-019-1825-8 (2019).
4. Bian, S. *et al.* Single-cell multiomics sequencing and analyses of human colorectal cancer. *Science* **362**, 1060–1063. https://doi.org/10.1126/science.aao3791 (2018).
5. McClintock, B. The origin and behavior of mutable loci in maize. *Proc. Natl. Acad. Sci.* **36**, 344 (1950).
6. de Koning, A. P. J., Gu, W., Castoe, T. A., Batzer, M. A. & Pollock, D. D. Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet.* **7**, e1002384. https://doi.org/10.1371/journal.pgen.1002384 (2011).
7. Brouha, B. *et al.* Hot L1s account for the bulk of retrotransposition in the human population. *Proc. Natl. Acad. Sci.* **100**, 5280. https://doi.org/10.1073/pnas.0831042100 (2003).
8. Sassaman, D. M. *et al.* Many human L1 elements are capable of retrotransposition. *Nat. Genet.* **16**, 37–43. https://doi.org/10.1038/ng0597-37 (1997).
9. Guo, H. *et al.* The DNA methylation landscape of human early embryos. *Nature* **511**, 606–610. https://doi.org/10.1038/nature13544 (2014).
10. Seisenberger, S. *et al.* The dynamics of genome-wide DNA methylation reprogramming in mouse primordial germ cells. *Mol. Cell* **48**, 849–862. https://doi.org/10.1016/j.molcel.2012.11.001 (2012).
11. Milagre, I. *et al.* Gender differences in global but not targeted demethylation in iPSC reprogramming. *Cell Rep.* **18**, 1079–1089. https://doi.org/10.1016/j.celrep.2017.01.008 (2017).
12. Bock, C. *et al.* Quantitative comparison of DNA methylation assays for biomarker development and clinical applications. *Nat. Biotechnol.* **34**, 726–737. https://doi.org/10.1038/nbt.3605 (2016).
13. Yang, A. S. *et al.* A simple method for estimating global DNA methylation using bisulfite PCR of repetitive DNA elements. *Nucleic Acids Res.* **32**, e38–e38. https://doi.org/10.1093/nar/gnh032 (2004).
14. Lisanti, S. *et al.* Comparison of methods for quantification of global DNA methylation in human cells and tissues. *PLoS ONE* **8**, e79044. https://doi.org/10.1371/journal.pone.0079044 (2013).

15. Issa, J.-P.J. *et al.* Safety and tolerability of guadecitabine (SGI-110) in patients with myelodysplastic syndrome and acute myeloid leukaemia: A multicentre, randomised, dose-escalation phase 1 study. *Lancet Oncol.* **16**, 1099–1110. https://doi.org/10.1016/S1470-2045(15)00038-8 (2015).
16. Fennell, K. A., Bell, C. C. & Dawson, M. A. Epigenetic therapies in acute myeloid leukemia: where to from here?. *Blood* **134**, 1891–1901. https://doi.org/10.1182/blood.2019003262 (2019).
17. Angermueller, C. *et al.* Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat. Methods* **13**, 229–232. https://doi.org/10.1038/nmeth.3728 (2016).
18. Ji, L. *et al.* Methylated DNA is over-represented in whole-genome bisulfite sequencing data. *Front. Genet.* https://doi.org/10.3389/fgene.2014.00341 (2014).
19. Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10**, 1096–1098. https://doi.org/10.1038/nmeth.2639 (2013).
20. Koschmieder, S. *et al.* Decitabine and Vitamin D3 differentially affect hematopoietic transcription factors to induce monocytic differentiation. *Int. J. Oncol.* **30**, 349–355. https://doi.org/10.3892/ijo.30.2.349 (2007).
21. Scheller, M. *et al.* Hotspot DNMT3A mutations in clonal hematopoiesis and acute myeloid leukemia sensitize cells to azacytidine via viral mimicry response. *Nat. Cancer* **2**, 527–544. https://doi.org/10.1038/s43018-021-00213-9 (2021).
22. Roulois, D. *et al.* DNA-demethylating agents target colorectal cancer cells by inducing viral mimicry by endogenous transcripts. *Cell* **162**, 961–973. https://doi.org/10.1016/j.cell.2015.07.056 (2015).
23. Ohtani, H. *et al.* Activation of a subset of evolutionarily young transposable elements and innate immunity are linked to clinical responses to 5-azacytidine. *Can. Res.* **80**, 2441–2450. https://doi.org/10.1158/0008-5472.Can-19-1696 (2020).
24. Liu, M. *et al.* Vitamin C increases viral mimicry induced by 5-aza-2'-deoxycytidine. *Proc. Natl. Acad. Sci. U S A* **113**, 10238–10244. https://doi.org/10.1073/pnas.1612262113 (2016).
25. Smallwood, S. A. *et al.* Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat. Methods* **11**, 817–820. https://doi.org/10.1038/nmeth.3035 (2014).
26. Farlik, M. *et al.* Single-cell DNA methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics. *Cell Rep.* **10**, 1386–1397. https://doi.org/10.1016/j.celrep.2015.02.001 (2015).
27. Gravina, S., Dong, X., Yu, B. & Vijg, J. Single-cell genome-wide bisulfite sequencing uncovers extensive heterogeneity in the mouse liver methylome. *Genome Biol.* **17**, 150. https://doi.org/10.1186/s13059-016-1011-3 (2016).
28. Mulqueen, R. M. *et al.* Highly scalable generation of DNA methylation profiles in single cells. *Nat. Biotechnol.* **36**, 428–431. https://doi.org/10.1038/nbt.4112 (2018).
29. Luo, C. *et al.* Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. *Science* **357**, 600–604. https://doi.org/10.1126/science.aan3351 (2017).
30. El Hajj, N. *et al.* Limiting dilution bisulfite (pyro)sequencing reveals parent-specific methylation patterns in single early mouse embryos and bovine oocytes. *Epigenetics* **6**, 1176–1188. https://doi.org/10.4161/epi.6.10.17202 (2011).
31. Lorthongpanich, C. *et al.* Single-cell DNA-methylation analysis reveals epigenetic chimerism in preimplantation embryos. *Science* **341**, 1110 (2013).
32. Kantlehner, M. *et al.* A high-throughput DNA methylation analysis of a single cell. *Nucleic Acids Res.* **39**, e44–e44. https://doi.org/10.1093/nar/gkq1357 (2011).
33. Pixberg, C. F. *et al.* Analysis of DNA methylation in single circulating tumor cells. *Oncogene* https://doi.org/10.1038/onc.2016.480 (2017).
34. Gravina, S., Ganapathi, S. & Vijg, J. Single-cell, locus-specific bisulfite sequencing (SLBS) for direct detection of epimutations in DNA methylation patterns. *Nucleic Acids Res.* **43**, e93–e93. https://doi.org/10.1093/nar/gkv366 (2015).
35. Pehrsson, E. C., Choudhary, M. N. K., Sundaram, V. & Wang, T. The epigenomic landscape of transposable elements across normal human development and anatomy. *Nat. Commun.* **10**, 5640. https://doi.org/10.1038/s41467-019-13555-x (2019).
36. Griffin, G. K. *et al.* Epigenetic silencing by SETDB1 suppresses tumour intrinsic immunogenicity. *Nature* https://doi.org/10.1038/s41586-021-03520-4 (2021).
37. Irwin, R. *et al.* UHRF1 suppresses viral mimicry through both DNA methylation-dependent and -independent mechanisms. *bioRxiv*: 2020.2008.2031.274894. https://doi.org/10.1101/2020.08.31.274894 (2020).
38. Fenaux, P. *et al.* Azacitidine prolongs overall survival compared with conventional care regimens in elderly patients with low bone marrow blast count acute myeloid leukemia. *J. Clin. Oncol.* **28**, 562–569. https://doi.org/10.1200/jco.2009.23.8329 (2010).
39. Unnikrishnan, A. *et al.* Integrative genomics identifies the molecular basis of resistance to azacitidine therapy in myelodysplastic syndromes. *Cell Rep.* **20**, 572–585. https://doi.org/10.1016/j.celrep.2017.06.067 (2017).
40. Yang, A. S. *et al.* DNA methylation changes after 5-aza-2'-deoxycytidine therapy in patients with leukemia. *Cancer Res.* **66**, 5495–5503. https://doi.org/10.1158/0008-5472.CAN-05-2385 (2006).
41. Kong, Y. *et al.* Transposable element expression in tumors is associated with immune infiltration and increased antigenicity. *Nat. Commun.* **10**, 5228. https://doi.org/10.1038/s41467-019-13035-2 (2019).
42. Nguyen, Q. H. *et al.* Single-cell RNA-seq of human induced pluripotent stem cells reveals cellular heterogeneity and cell state transitions between subpopulations. *Genome Res.* **28**, 1053–1066. https://doi.org/10.1101/gr.223925.117 (2018).
43. Lee, D.-S. *et al.* An epigenomic roadmap to induced pluripotency reveals DNA methylation as a reprogramming modulator. *Nat. Commun.* **5**, 5619. https://doi.org/10.1038/ncomms6619 (2014).
44. Shao, X., Zhang, C., Sun, M.-A., Lu, X. & Xie, H. Deciphering the heterogeneity in DNA methylation patterns during stem cell differentiation and reprogramming. *BMC Genomics* **15**, 978. https://doi.org/10.1186/1471-2164-15-978 (2014).
45. Macaulay, I. C. *et al.* G&T-seq: Parallel sequencing of single-cell genomes and transcriptomes. *Nat. Methods* **12**, 519–522. https://doi.org/10.1038/nmeth.3370 (2015).
46. Miura, F., Enomoto, Y., Dairiki, R. & Ito, T. Amplification-free whole-genome bisulfite sequencing by post-bisulfite adaptor tagging. *Nucleic Acids Res.* **40**, e136–e136. https://doi.org/10.1093/nar/gks454 (2012).
47. Rulands, S. *et al.* Genome-scale oscillations in DNA methylation during exit from pluripotency. *Cell Syst.* **7**, 63-76.e12. https://doi.org/10.1016/j.cels.2018.06.012 (2018).
48. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**(3), 2011. https://doi.org/10.14806/ej.17.1.200 (2011).
49. Grosselin, K. *et al.* High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states in breast cancer. *Nat. Genet.* **51**, 1060–1066. https://doi.org/10.1038/s41588-019-0424-9 (2019).
50. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359. https://doi.org/10.1038/nmeth.1923 (2012).
51. Krueger, F. & Andrews, S. R. Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572. https://doi.org/10.1093/bioinformatics/btr167 (2011).
52. Viswanathan, R., Cheruba, E. & Cheow, L. F. DNA analysis by restriction enzyme (DARE) enables concurrent genomic and epigenomic characterization of single cells. *Nucleic Acids Res.* **47**, e122. https://doi.org/10.1093/nar/gkz717 (2019).
53. Krueger, F. PBAT and single-cell (scBS-Seq) libraries may generate chimeric read pairs. https://sequencing.qcfail.com/articles/pbat-libraries-may-generate-chimaeric-read-pairs/ (2016).
54. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915. https://doi.org/10.1038/s41587-019-0201-4 (2019).

55. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352 (2009).
56. Jin, Y., Tam, O. H., Paniagua, E. & Hammell, M. TEtranscripts: A package for including transposable elements in differential expression analysis of RNA-seq datasets. *Bioinformatics* **31**, 3593–3599. https://doi.org/10.1093/bioinformatics/btv422 (2015).
57. Inc., A. *Anaconda Software Distribution* https://docs.anaconda.com/ (2020).
58. Van Rossum, G. A. D. J., Fred L. *Python Reference Manual*. https://www.python.org/ (1995).
59. Andreas Heger, K. J. A. C. *pysam: Htslib Interface for Python*. https://github.com/pysam-developers/pysam (2009).
60. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550. https://doi.org/10.1186/s13059-014-0550-8 (2014).
61. 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. https://doi.org/10.1038/s41587-019-0206-z (2019).
62. Mi, H. *et al.* PANTHER version 16: A revised family classification, tree-based classification tool, enhancer regions and extensive API. *Nucleic Acids Res.* **49**, D394–D403. https://doi.org/10.1093/nar/gkaa1106 (2021).
63. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. https://ggplot2.tidyverse.org (2016).
64. Kolde, R. *pheatmap: Pretty Heatmaps*. https://CRAN.R-project.org/package=pheatmap (2019).

Acknowledgements

We thank Dr Shuhui Bian (Peking University) for providing the genetic sub-lineages of CRC01 patient cells analysed in Supplementary Fig. S1. We thank Dr Carlos Riveros (Hunter Medical Research Institute) for consultation regarding data processing pipelines related to this work, and Dr Shalin Naik (Walter and Eliza Hall Institute) for suggestions regarding primer design. We also thank Professor Geoffrey Faulkner and Dr Sandra Richardson (University of Queensland) for helpful discussions regarding transposable element biology, and Ms Nicole Cole (Hunter Medical Research Institute) for assistance with fluorescence activated cell sorting.

Author contributions

Conception and design of the study: H.J.L. and K.V.H.; Experiments and data collection: K.V.H. E.A.R., D.R.B.; Provision of biological material and clinical data from patients: M.D.D., N.M.V. and A.K.E.; Data analysis and figure preparation: K.V.H., S.M.B. and H.J.L.; Interpretation of results: K.V.H. and H.J.L.; Drafting, revising and approval of the manuscript: all authors.

Funding

Funding was provided by National Health and Medical Research Council (Grant No. GNT1143614),Cancer Institute NSW (Grant No. ECF171145),Ian Potter Foundation (Grant No. 20180029),Australian Research Council (Grant No. DP200102903).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-09765-x.

Correspondence and requests for materials should be addressed to H.J.L.

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

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