Published in final edited form as:
Nat Genet. 2020 July 01; 52(7): 709–718. doi:10.1038/s41588-020-0645-y.

Single-cell analysis of clonal maintenance of transcriptional and epigenetic states in cancer cells

Zohar Meir[1], **Zohar Mukamel**[1], **Elad Chomsky**[1], **Aviezer Lifshitz**[1], **Amos Tanay**[1]

[1]Faculty of Mathematics and Computer Science and Department of Biological Regulation, Weizmann Institute of Science, Rehovot Israel

Abstract

Propagation of clonal regulatory programs contributes to cancer development. It is poorly understood how epigenetic mechanisms interact with genetic drivers to shape this process. Here we combine single-cell analysis of transcription and DNA methylation with a Luria-Delbrück experimental design to demonstrate the existence of clonally stable epigenetic memory in multiple types of cancer cells. Longitudinal transcriptional and genetic analysis of clonal colon cancer cell populations reveals a slowly drifting spectrum of epithelial-to-mesenchymal transcriptional identities that is seemingly independent of genetic variation. DNA methylation landscapes correlate with these identities but also reflect an independent clock-like methylation loss process. Methylation variation can be explained as an effect of global *trans*-acting factors in most cases. However, for a specific class of promoters, in particular cancer testis antigens (CTA), de-repression is correlated with and likely driven by loss of methylation in *cis*. This study indicates how genetic sub-clonal structure in cancer cells can be diversified by epigenetic memory.

Introduction

The ability of cells to maintain their molecular identity through mitotic cell divisions is essential for the establishment of functionally coherent and stable clonal cell populations. During carcinogenic transformation, selection for beneficial driver mutations contributes to the basic clonal and sub-clonal structure of the evolving tumor[1]. However, this selective process provides limited flexibility for rapid adaptation and diversification in the context of a dynamic stromal environment, immune interactions or following treatment. Non-genetic mechanisms can modulate cellular states and enhance such flexibility if their diversification can persist and form an epigenetic memory. Pioneer systematic screenings for transcriptional memory[2–4] demonstrated the potential for clonally stable transcriptional phenotypes in

Correspondence to: Amos Tanay.

Correspondence: Amos Tanay amos.tanay@weizmann.ac.il.

Authors Contribution

ZMe and AT designed the study with help from ZMu. ZMe performed experiments with help from ZMu. EC helped with MARS-seq and automation. ZMe and AT analyzed data with help from AL. ZMe and AT wrote the manuscript with input from all authors.

Conflict of Interest

The authors declare no conflict of interest.

Europe PMC Funders Author Manuscripts

mammalian systems. Nonetheless, it remains difficult to measure the extent of such transcriptional memory and to distinguish transient transcriptional heterogeneity within cell populations from stable and clonally transmitted states.

Theoretical and experimental models of commitment and memory through specific gene network architecture in unicellular organisms[5], and systems for analyzing stable mitotic transmission of epigenetic states in mammals[6–8] provide a basis for understanding the mechanisms underlying the formation and maintenance of epigenetic memory. DNA methylation is the best studied epigenetic mechanism for stable memory formation, and the ability of cells to copy their methylation makeup to daughter cells is well established[9–10]. The correlation between specific DNA methylation patterns and cell-type-specific transcriptional programs has been also demonstrated[11]. But the role of DNA methylation in regulating clonal heterogeneity within diversifying cell populations is still unclear. Methylation changes with ageing[12], cellular senescence[13], and transformation[14–15]. Errors in replicating methylation marks, which cause epimutations, can accumulate to create global, replication-dependent reduction in methylation levels that is associated with clock-like dynamics[6]. Such changes are unlikely to have a direct functional impact, and are observed as a background process in both normal and cancer tissues. Other cancer-linked methylation changes may be driven by modulation in the activity of *trans*-acting factors, including recurrent genetic mutations in the methylation machinery itself[16]. Here again, the role of DNA methylation as a carrier of molecular memory in *cis* is limited since it is only indicating the activity of an aberrant epigenetic modulator in *trans*. Given these broad background dynamics and *trans*-acting effects, it remains difficult to identify cases in which methylation is effectively maintaining repressed/de-repressed switches in *cis* at gene promoters or distal regulatory elements.

To address some of these difficulties, we implemented a Luria-Delbrück experimental design and compared single-cell transcriptional and epigenetic distributions in cancer cells to the distributions of mean gene expression and methylation across clones originating from the same cell populations. Using high-throughput and precise single-cell genomic technology, this resulted in thousands of single-cell profiles and hundreds of matching clonal profiles, revealing broad clonally stable transcriptional diversity in immortalized fibroblasts and in lung and colon cancer cells. Genetic profiling and longitudinal analysis indicate that the observed clonal diversity represents epigenetic memory. Analysis of DNA methylation characterizes the epigenetic makeup underlying repression and de-repression of key clonal gene modules, in particular along a spectrum of epithelial-to-mesenchymal transition (EMT) identities in colon cancer cells. DNA methylation in *cis* cannot be implicated unambiguously with a causal role in maintaining (and drifting) gene expression state for most clonal genes. Nevertheless, we observe a class of promoters, including several cancer testis antigens (CTA), for which correlation indicates a causal *cis*-regulatory role for DNA methylation. In conclusion, our in vitro Luria-Delbrück assays suggest that epigenetic memory in cancer cell populations operates pervasively and in parallel to genetic drivers, to diversify transcriptional programs and channel cells toward EMT and other tumorigenic transcriptional dynamics.

Results

Luria-Delbrück assays identify clonally stable transcriptional memory in cancer cells

To facilitate detailed exploration of clonal or transient transcriptional and methylation states in cancer cells we followed the classic Luria-Delbrück scheme (Fig. 1a). Clonally stable transcriptional programs or epigenetic states are expected to propagate from founder cells to give rise to homogeneous clones. In this case, inter-clonal variance should recapitulate precisely single-cell heterogeneity in the founding population. In contrast, cell-cycle signatures or other transient fluctuations are averaged out in clones so as to eliminate inter-clonal variance (Extended Data Fig. 1a). We FACS-sorted and cultured single cells of two lung cancer cell lines (NCI-H1299 and A549), one colorectal cancer (HCT116) and the non-cancerous WI38 fibroblast cell line. We analyzed 841 clones (median coverage of 103,000 UMIs), that were expanded for 9-10 doublings (Extended Data Fig. 1b) and compared those to single-cell RNA-seq profiles obtained from the matching founding populations (Extended Data Fig. 1c-f). Clonal populations (500 to 1,000 cells) allowed sampling of 50 cells in replicates, ensuring quantitative estimation of clonal RNA concentrations (Extended Data Fig. 1g-n). Comparison of pooled RNA from cells and clones showed high degree of concordance and suggested the bias or selective constraint associated with the single-cell cloning procedure itself was limited (Extended Data Fig. 2a-c).

As expected, cell-cycle-linked single-cell variation was observed pervasively in all studied cell lines (Fig. 1b-d, Extended Data Fig. 2d-g and Supplementary Table 1). This prompted us to implement a strategy for discovery of cell-cycle independent gene-gene correlation modules, by comparing correlations in the raw count matrix and a randomized matrix obtained by shuffling RNA counts over cells with similar cell-cycle characteristics (see Methods and Extended Data Fig. 2h,i). Interestingly, we found strong cell-cycle independent transcriptional variation in all cell lines, suggesting the existence of either transient transcriptional dynamics, genetic sub-clonal population structure, and/or non-genetic clonal population structure. Our Luria-Delbrück scheme allowed us to distinguish transient from clonal population structure (Fig. 1e-g), confirming that the S-phase and M-phase gene modules observed in single-cell analysis were indeed transient, but suggesting other gene modules to be clonally stable (Extended Data Fig. 2j-o and Supplementary Table 2). For example, in HCT116 we discovered a clonally stable epithelial gene module, marked by *EpCAM* expression, and an anti-correlated EMT-like gene module, marked by *ZEB1* and *VIM* expression (Extended Data Fig. 3 and Supplementary Fig. 1). H1299 cells showed variable expression of a module including *ID1*, *ID2* and *ID3*, WI38 fibroblasts variably expressed a Collagen/fibronectin gene module, and A549 cells showed continuous variation of several additional gene modules (Extended Data Fig. 4 and Supplementary Table 3). Taken together, these findings showed that at least within 9-10 cell divisions, all examined cell populations featured a strong component of clonal and cell-cycle independent transcriptional stability.

Characterizing a continuum of clonal epithelial identities in HCT116 cells

Both single cells and clones of HCT116 cells showed a consistent distribution of epithelial gene expression (Fig. 1e) and expression of genes linked with EMT (Fig. 1h), adhering to

the Luria-Delbrück principle of clonal memory. This suggested this system can be an effective model to test the possible genetic or epigenetic basis for clonal transcriptional memory in our cancer cells. Genes linked with high or low *EpCAM* expression (Fig. 1i, Supplementary Fig. 1d and Supplementary Table 4) defined two continuous and anti-correlated spectra of epithelial and EMT-like identities across all clones (Fig. 1j and Supplementary Fig. 1e). At the extreme low-end of this spectrum we identified 4% of the clones (and consistently 4% of the single cells) with particularly high expression of *VIM* (defining a population we denote below as VIM-high clones). The continuum of epithelial and EMT-like transcriptional states were both correlated with specific transcription factors (TFs), suggestive of possible regulatory networks diversifying the memorized clone states (Fig. 1k).

To elicit specific molecular mechanism that support such memory, we performed genetic, transcriptional and epigenetic analysis on single-cell clones propagated for up to 168 days. We first generated hundreds of clones that were profiled after 10 and 18 days at low resolution (Fig. 2a,b). We then selected six clones representing high and low *EpCAM*-expressing states, as well as a VIM-high state, and followed them for additional 150 days. For these clones we performed exome sequencing in two time points (d = 78, 168) and discovered that the high mutational load in the mutagenic HCT116 system gave rise to polymorphisms that are not shared between clones, even if their epithelial signatures are similar (Fig. 2c). Exome analysis of clones selected using a similar strategy from two additional cancer cell lines (Extended Data Fig. 5) revealed highly robust genetic subclones underlying one transcriptional module in H1299 cells, but no evidence for genetic basis for variation in all other transcriptional modules detected.

To further characterize clonal transcriptional dynamics, we next sampled 7,590 single cells from the six selected HCT116 clones after 33, 62, 98 and 148 days. We modeled the transcriptional space of single cells in these clones (Fig. 2d) and tracked evolving variation over this space within the clonal populations. This first reconfirmed the clonality of epithelial identity in the HCT116 system, showing some clones (1d12, 4e1, 7b11) maintain a distribution of high and low *EpCAM*-expressing states across many dozens of cell divisions (Fig. 2e,f). Clonal stability was however imperfect, as expected from a non-genetic mechanism, and we observed decrease in epithelial gene expression progressively in clone 3b3 and a reciprocal effect in clone 7a2 (Fig. 2g). Together, these experiments support the idea that epithelial transcriptional heterogeneity in the system has a non-genetic basis, permitting (slowly) drifting identities, and convergence of multiple clones toward similar transcriptional fates.

Replication-linked loss of methylation underlies genome-wide clonal epigenetic diversity

We next performed single-cell analysis of global DNA methylation profiles (1,022 cells at low depth) and targeted methylation profiling (enriching over 70-fold the coverage for 318,783 select sites) of 251 9-days old HCT116 clones (Extended Data Fig. 6a-d) for which transcriptional data were also available (Supplementary Table 5). We focused initially on global genome methylation dynamics at low- and high-CpG content loci (LCG and HCG). Interestingly, LCG sites, that represent the majority of genomic territories and are highly

methylated in normal cells, showed surprising quantitative variation in average methylation for both clones and single cells (Fig. 3a and Extended Data Fig. 6e-k, 72-83.5% for single cells, 68.3-78.9% for clones). A specific small subpopulation of cells and clones (3.7% and 4.7%, respectively) showed major LCG hypomethylation (46.5-62%), and was shown to coincide with the VIM-high clone population we defined based on gene expression (Fig. 3b). But beyond this small subpopulation, we identified only weak negative correlations between gene expression and LCG methylation (Fig. 3c and Supplementary Table 6). In contrast to the lack of expression correlation, we observed lower clonal methylation (Fig. 3d) and higher inter-clonal variance (Fig. 3e) for LCG sites associated with late time of replication, compared to early replicating LCG sites. Overall, these data are consistent with several recent reports[17], suggesting variation in LCG methylation in cancer may originate from the accumulation of replication-linked epimutations. Such epimutations are more frequent in late-replicating domains and may cause progressive loss of methylation in loci that are originally highly methylated. The process is pervasive, and is initially unlinked to transcriptional perturbation or transcriptional memory. It can however predispose specific genes to stochastic de-repression as we demonstrate below.

High CpG content sites show clonal instability and turnover linked with cell proliferation

High CpG-content sites (HCG) are normally protected from methylation and are observed at CpG islands within gene promoters and distal regulatory elements. Similarly to LCG sites, we observed significantly high inter-clonal variation in HCG methylation levels (Fig. 3f). But unexpectedly the two global methylation trends were uncorrelated (Fig. 3g) and HCG methylation correlated extensively to gene expression (Fig. 3h). HCG methylation lacked association with genomic time of replication at regions with overall low methylation (Fig. 3i,j). Many of the top HCG correlated genes, both negative and positive, were associated with cell proliferation (Supplementary Table 6), suggesting possible linkage between clone proliferation rate and poor maintenance of HCG methylation.

We next used our target set of enhancers and promoters (Extended Data Fig. 6l-n) to study the intra- and inter-clonal distribution of methylation at higher depth. We contrasted two models of methylation transmission per locus (Fig. 3k): one assuming continuous methylation turnover within a growing clone, leading to convergence to a clonal distribution that is independent from the founding cell's epigenome, and the other assuming single cells are sampling two epi-alleles and transmitting them to all offsprings. Using data on all clones, we assayed the degree of clonal coherence of each locus in our target set (Fig. 3l) defining "hot" (marked red) and "cold" loci (marked blue) as those showing high turnover or high degree of clonal persistence, respectively. Remarkably, we observe average methylation in "cold" sites to be correlated with LCG methylation, and average methylation in "hot" sites to be correlated with HCG methylation (Fig. 3m).

Importantly, the unexpected variation and independence of the HCG and LCG methylation signatures we characterize here are observed pervasively in TCGA tumor methylation data (Extended Data Fig. 6o-q). In summary, we demonstrated two types of methylation regimes that globally affect most CpG loci in the genome. The first involves replication-linked progressive hypomethylation in LCG sites that are normally methylated. The second

involves epigenetic instability at high CpG content sites that are normally unmethylated, in correlation with perturbed expression of many cell-cycle genes.

Clonal DNA methylation signatures are correlated with epithelial transcriptional identity

To search further for an epigenetic basis for the broad transcriptional epithelial identity spectrum we observed above, we normalized methylation of enriched enhancers and promoters given clones' global HCG and LCG methylation levels (Extended Data Fig. 6a-c) and clustered the normalized profiles to reveal additional inter-clonal epigenomic structure (Extended Data Fig. 7d). Two of the observed gene clusters were associated with the epithelial identity spectrum (Extended Data Fig. 7e,f). We therefore screened directly for promoters with differential methylation between high and low *EpCAM*-expressing clones (Fig. 3n). We detected strong anti-correlation between expression and methylation in EpCAM-high and EpCAM-low promoters (red and blue points, Fig. 3n, $P = 8 \times 10^{-5}$, D = 0.3, Kolmogorov-Smirnov (KS) two-tailed test). Weak hypomethylation was observed in the *EpCAM* locus itself ($P = 0.01$, $X^2 = 6.5$, chi-squared test), and stronger reduction in methylation was shown for additional promoters of induced genes (Fig. 3o). We expanded this analysis to putative enhancer loci (Extended Data Fig. 7g), identifying 53 and 30 hypomethylated enhancers in EpCAM-high and EpCAM-low clones, respectively (Supplementary Table 7). For example, a hypomethylation hotspot observed for putative enhancer in chromosome 20 is correlated with EpCAM-high expression for at least 5 genes within 400 kb around it (Extended Data Fig. 8a,b). Interestingly, the dynamics of CpGs in this enhancer are classified as "cold" by the model described above, supporting their possible role as epigenetic memory carriers (Extended Data Fig. 8c,d, see Supplementary Table 8 for hot/cold classification of all EpCAM-linked enhancer CpGs).

Genes that Escaped Mitotically INherited Inhibition (GEMINIs)

We next developed a screen for genes with high clonal expression in one or more clones, but very low expression in at least 90% of the other clones. These represented loci that escaped repression rarely, in a clonally stable fashion, and not within the context of a larger co-regulated gene module (see Methods and Extended Data Fig. 8e). We reasoned such genes may be strong candidates for *cis*-acting epigenetic control. We detected 206 rare clonal de-repression events of 98 different genes in 97 distinct clones and define these as Genes that Escaped Mitotically INherited Inhibition (GEMINIs) (Fig. 4a). We found GEMINIs are encoding for different functions, including transcription factors (RUNX2, GRHL2, ELF3, TCF4 and ATOH8), long non-coding RNAs (LOC100287225, BC037861), transmembrane proteins (TM4SF18, CHODL), phosphatases (CCDC8) and more. Interestingly, five of the GEMINIs (PAGE1, PAGE4, MAGEA1, MAGEA11 and MAGEB1) belong to a subset of Cancer-Testis Antigens (CTAs), previously annotated as germline restricted genes that are de-repressed in different cancers[18]. Average expression of GEMINIs in clones and cells is generally consistent (Fig. 4b), but GEMINIs are noisier in single cells compared to control genes with similar expression levels ($P = 0.01$, KS test, Fig. 4c).

GEMINIs constitute a set of de-repression events that are uncorrelated to each other and uncorrelated to observed *trans*-factors regulatory changes. We hypothesized that the regulation of their clonal expression signature may therefore be mediated by *cis*- epigenetic

effects. Promoters of GEMINIs were found to be natively more methylated in comparison to matched controls with similar CpG content and overall expression levels ($P = 2 \times 10^{-6}$, D = 0.35, KS test). GEMINIs tended to be positioned away from constitutively expressed genes ($P = 8 \times 10^{-8}$, D = 0.22, KS test) (Fig. 4d). Most importantly, data from clones showed nearly 50% reduction in promoter methylation in *cis* for de-repressed loci ($P = 9 \times 10^{-13}$, $X^2 = 51$, chi-squared test, Fig. 4e,f), compatible with a mono-allelic loss of methylation and suggesting allele-specific de-repression. In addition, we observed GEMINIs expression is more associated with the "cold" LCG clonal methylation trend (Fig. 4g, left) and show lower correlation with the "hot" HCG trend (Fig. 4g, right, see also Extended Data Fig. 8f). These data identify a class of loci in which DNA methylation is not only tightly correlated to de-repression in *cis*, but is likely driving it. Stochastic loss of methylation at these sites (and subsequent clonal maintenance of this hypomethylation) stabilize a de-repressed state in specific clones.

GEMINIs are induced in methylation-impaired cells

We next profiled 4,523 single cells and 319 short-term clonal populations derived from methylation-impaired HCT116 cells (double knockout of DNMT1 and DNMT3B DNA-methyltransferases or DKO). As expected, DKO cells displayed severe whole-genome reduction in DNA methylation when compared to parental HCT116 single cells (Extended Data Fig. 8g-m). We observed a high degree of concordance of cell-cycle dependent transcriptional states between the DKO system and the wild-type (Extended Data Fig. 8n-p), and also identified cell-cycle independent co-variating gene modules in DKO cells and clones, including the same epithelial gene module observed in wild-type cells (Extended Data Fig. 9a-c). DKO cells were biased toward lower EpCAM module expression compared to wild-type cells, but the distribution in single cells and clones was conserved (Extended Data Fig. 9d-f and Supplementary Fig. 2). This suggested the EpCAM module gene expression remains largely coordinated even when methylation is impaired. We also find that DKO cells maintain a gene module that includes *ZEB1* and several HOXB genes, which was anti-correlated to expression of the epithelial module, as well as instability of Interferon type-I (IFN-I) and DNA damage response (DDR) gene modules (Supplementary Fig. 3). Together the data show transcriptional memory for co-regulated gene modules in HCT116 can be independent from a fully functional DNA methylation machinery.

Using loci identified as GEMINI in the wild-type, we next tested if repression of GEMINI is maintained in the DKO system. Strikingly, 40 out of the 98 GEMINIs found in WT were pervasively expressed in DKO clonal populations (FDR corrected q-value < 0.001, KS test, Fig. 4h, right panel). De-repression was significant compared to control genes that were repressed in the wild type ($P = 2 \times 10^{-3}$, D = 0.2, two-tailed KS test, Fig. 4i). Notably, de-repression was in general observed more often for low CpG content repressed promoters that were methylated in the wild-type, compared to controls that were not methylated ($P = 4 \times 10^{-3}$, $X^2 = 8.1$, chi-squared test, Fig. 4j). We profiled 251 additional clones on the background of a single DNMT3b knockout (Supplementary Fig. 4a), showing milder, but still noticeable de-repression in some of the GEMINIs (Supplementary Fig. 4b). We hypothesize that the in *cis* de-repression we have observed in our clonal populations applies for other systems as well, and specifically may be occurring in colon tumors. We thereby

analyzed bulk RNA-seq experiments of nearly 400 colorectal adenocarcinoma tumors from the TCGA database, focusing on repressed genes and computing the ratio between maximally observed de-repression of each gene to its expression in the top five percentile of the cohort. Interestingly, we found that rare de-repression is indeed enriched for genes identified in our screen as GEMINIs ($P = 5 \times 10^{-5}$, D = 0.28, KS test, Fig. 4k).

Correlated clonal de-repression within topological associating domains

Some of the clonally stable gene modules we detected above were defined through co-expression of spatially linked genes within a single topologically associating domain (TAD). We therefore screened systematically for clonal co-expression of genes within TADs in the HCT116 system (Methods and Extended Data Fig. 10a-c), using shuffled controls and accounting for possible chromosomal dosage effects using pBAT coverage statistics (Methods and Supplementary Fig. 5). This approach identified 149 genes in 89 TADs with statistical support (FDR < 0.25) for intra-TAD co-expression, including the embryonic globin genes *HBE1* and *HBG* (Extended Data Fig. 10d-f and Supplementary Fig. 6a,b). Similar analysis in WI38 and A549 cells identified additional putative cases for TAD-linked clonal co-expression (Extended Data Fig. 10g-l, Supplementary Fig. 6c-f and Supplementary Table 9). In-depth analysis of the co-expression around the beta-globin genes in HCT116 cells (Fig. 5a), suggested a bimodal clonal distribution of transcription in these loci (denoted as HB-high and HB-low). We observed weak de-repression of several OR genes in HB-high clones (Fig. 5b,c). De-repression in the HB TAD was uncorrelated with the EpCAM expression signature (Fig. 5d), but showed remarkable long-term stability (Fig. 5e). Another notable spatial clonal expression pattern involved 11 keratin associated proteins (KRTAP) organized in a 130-kb cluster on chromosome 17 (Fig. 5f-j). HCT116 DKO clones showed conserved bimodal HB and KRTAP TAD expression (Supplementary Fig. 6g,h). Importantly, de-repression of the globin and KRTAP clusters was observed in a large fraction of the clones (47% and 23%, respectively). Nevertheless the transcriptional output per single cell was low for most cells even in de-repressed clones (Fig. 5k,l). Moreover, in some cases TAD clonal output was also correlated with the expression of an additional factor in *trans* (as for *TCF25* and the HB expression, Extended Data Fig. 10m,n), suggesting de-repression requires a permissive TAD state but also additional driving factors. In summary, spatially correlated clonal expression patterns raise the hypothesis that TADs can toggle between clonally inactive and active states, where in the active state multiple genes within a TAD are predisposed for de-repression, and in the inactive state complete repression is secured. This mechanism can thereby diversify clonal expression patterns of spatially (and functionally) linked genes.

Discussion

We studied the clonal propagation of transcriptional and epigenetic identity in cancer cells using a Luria-Delbrück design in which single-cell distributions of RNA expression and DNA methylation are compared to the analogous distributions in hundreds of short-term clones. We characterized stable transcriptional heterogeneity in all studied cell types, which in all except for one case appeared to be unlinked with a sub-clonal genetic structure. In colon cancer cells, we identified clonally stable variation in epithelial gene expression and a

reciprocal variation in expression of EMT genes. Longitudinal analysis of select clones using additional single-cell analysis in four time points showed that slow drift in the epithelial/EMT identity is observed for up to 150 days. These data pave the way to analysis of the dynamics underlying non-genetic clonal memory, and the role of epigenetic mechanisms implementing it.

Non-genetic clonal memory is first (and perhaps foremost) implemented by gene regulatory circuits using feedback to enable semi-stable transcriptional states, but our studies here show that in addition to these mechanisms, stable epigenetic variation between cells contribute to clonal memory. DNA methylation is the best understood clonally persistent epigenetic mark. It is generally assumed methylation propagates from mother to daughter cells through the housekeeping methylation machinery. Our data on single cells and clonal methylation distributions suggest more complex modes of methylation dynamics. First, most of the genome, being rarely targeted by epigenetic modulators or other *trans*-factors, indeed shows high degree of epigenetic persistence that is driven by the housekeeping machinery. Despite this persistence, we show here variability in average methylation between single cells and clones at such loci. We hypothesize this is the result of the accumulation of epimutations with replicative age[17]. Second, and unexpectedly, we show that methylation of CpG islands (or high CpG content loci in general) is governed by a different and much more dynamic process. Clones and single cells are shown to be variable in their ability to protect CpG islands from methylation accumulation. We define this molecular phenotype as “epigenetic instability” and show it to be uncorrelated with the replication-age effects that are observed in the rest of the genome.

We previously suggested that methylation dynamics can be governed by clonal persistence in somatic cells and turn-over in embryonic stem cells[6]. This model can now be generalized, predicting that in cancer cells protection from DNA methylation at promoters and regulatory elements primarily involves turnover. CpG island hypermethylation can then result from change in the efficiency of the protective turnover mechanism, rather than from progressive accumulation of epimutations with time.

The regulatory consequences of the two types of proposed methylation dynamics are global – affecting thousands of loci together rather than tinkering gene expression of individual genes or programs. But the combination of stochastic and regulated changes in methylation at loci with low turnover rate gives rise to an epigenetic landscape that is specific to each clone, resulting in many opportunities to diversify transcription in a clonally stable fashion. One example demonstrated here is the methylation change at “cold” promoters and enhancers of epithelial/EMT genes, that may participate in regulating the spectrum of epithelial identities in colon cancer cells. A second example involve GEMINIs - cases of rare, but clonally stable promoter hypomethylation and gene de-repression events that are enriched for CTA loci. A similar mechanism of clonal epigenetic change predisposing transcriptional dynamics is suggested by analysis of correlated de-repression of genes residing in the same TAD. In this case, TADs can repress effectively their associated genes when present in their “epigenetically closed” state, but are allowing their occasional de-repression in clones that switch to the “epigenetically open” state.

Together, the data here highlight the ability of non-genetic mechanisms to stably diversify transcription in cancer cells. This in turn can facilitate opportunities for adaptation that run in parallel (and in coordination) to the genetic evolutionary process driving carcinogenic transformation. An immediate challenge is to evaluate the impact of such non-genetic mechanisms in vivo. Initial analysis of TCGA data[19] suggests DNA methylation dynamics can be generalized from the ex vivo systems we analyzed here to tumors in vivo. Additional profiling of tumors using new single-cell epigenomic and multi-omic strategies will be essential for fully appreciating how clonally stable epigenetic changes drive long-term regulatory programs and how such changes may affect tumor function, response to therapy, or metastasis.

Methods (including online methods)

Cell lines

HCT116 parental colorectal cell line (HD PAR-033), as well as KO (DNMT3B$^{-/-}$, HD R02-023) and DKO (DNMT1$^{exons3\text{-}5/exons3\text{-}5}$; DNMT3B$^{-/-}$, HD R02-022) were obtained from Horizon Discovery Ltd., Cambridge, UK. A549 cells were obtained from the NCI-60 cell panel, NCI-H1299 cells were purchased from ATCC (ATCC® CRL-5803™) and WI38-hTERT embryonic lung fibroblasts (WI38) were generated as described in Milyavsky et al. [21]. All Cells were cultured on 100 × 20 mm culture dishes (Corning, 353063) in heat-inactivated medium and split at a ratio of 1:10 every 2–3 days using 0.05% trypsin-EDTA solution C (Biological Industries, Israel; 03-053-1B). McCoy's 5A medium (Biological Industries, Israel; 01-075-1A) was used for HCT116 WT, KO and DKO, DMEM medium (Dulbecco's Modified Eagle Medium, Biological Industries, Israel; 01-050-1A) was used for WI38, DMEM/F-12 (HAM) medium (Biological Industries, Israel; 01-170-1A)) was used for A549 cells and RPMI medium (Biological Industries, Israel; 01-100-1A) was used for NCI-H1299 cells. All media were supplemented with 10% Fetal Bovine Serum (GibcoTm FBS, 10270-106), 0.4% Penicillin-Streptomycin (Biological Industries, Israel; 01-031-1C) and 1% L-glutamine (Biological Industries, Israel; 01-020-1A). Modified medium was filtered through a 0.22-μm filter (Corning, 430769) prior to culture.

Single-cell isolation and clonal expansion

In parallel to sorting of single cells for MARS-seq analysis, individual cells were selected by FACS (SORP FACSAriaII cell sorter) using a 70-μm diameter nozzle into 96-well culture plates (Corning-Costar, 3596), which already contained 100 μl of conditioned media in each well. Conditioned media were taken from cycling populations of the respective cell line, centrifuged and filtered through a 0.22-μm filter (Corning, 430769). Subsequent to sorting, plates were transferred immediately to a 37 °C incubator for culture. 48 hours post sorting, 100 μl of conditioned media were added to each well. Clonal expansion was terminated when populations reached approximately 500 cells (8-10 cell divisions), in average after 9-10 days for HCT116 parental, NCI-H1299 and A549 cells, 12 days for parental and DKO cells and 21 days for WI38 cells. All clonal populations were examined by microscope prior to their harvesting. Small clones (~200 cells or less) were discarded, as well as wells that were suspected of having more than one founder cell in them.

Harvesting clonal populations

In order to process each clonal population for MARS-seq[22] (transcriptome analysis) and PBAT-capture (methylome analysis), clones were detached by 30 µl trypsin-incubation for 2 min. (HCT116 parental, A549, H1299 and WI38 cells) or 4 min. (HCT116 DKO cells) in 37 °C, washed by 100 µl PBS (Biological Industries, 02-023-1A) and suspended in 10 µl ultra-pure water (Biological Industries, 01-866-1A) and 0.005% RNase inhibitor (RNasin plus, Promega, N2611). The 10 µl suspension of each clone was then transferred to a skirted twin.tec 96-well PCR plate, (Eppendorf, 0030 128.648) on dry ice. When a 96-well plate was filled, replicates of 1 µl from each clone (two for HCT116 parental and HCT116 DKO, and four replicates for H1299 and A549 cells) were transferred into a barcoded (8 nM oligo-dT barcodes concentration) twin.tec 384-well (Eppendorf, 0030 128.508) MARS-seq plate that already contained 2 µl of lysis-buffer (0.1% Triton, 0.005% RNasin and ultra-pure water). The 384-well plates were then transferred into -80°C until its cDNA library preparation by MARS-seq. The remaining 8 µl from each HCT116 clone were kept in -20 °C for further methylome profiling by PBAT-capture. For WI38 clones, the 10-µl suspension of each clone was used for 6 technical replicates in 384-wells barcoded MARS-seq plates. For A549 and H1299 clones, out of the 10 µl of each clone – 5 µl were immediately transferred into 96-wells culturing plates (Corning-Costar, 3596) that contained 200 µl of the respective growth media and the rest were used for MARS-seq. A549 and H1299 clones that were further growing in 96-well culture plates were then routinely split until selection (according to their transcriptome profiles obtained by MARS-seq), resulting in six A549 and seven NCI-H1299 clonal populations assayed by whole-exome sequencing.

Single-cell transcriptome profiling

For MARS-seq scRNA protocol, single cells were sorted into 384-well plates containing 2 µl of barcoded RT primers in concentration of 8 nM in each well. Downstream library preparation was done according to Jaitin et al. 2014, and by using randomized UMI sequence of 8 base-pairs (allowing maximal count of ~65k UMIs per gene per well). For NCI-H1299 scRNA, we used data of ‘5 10x libraries described in Brocks et al. 2019[23]. HCT116 WT and HCT116 DKO scRNA libraries based on 10x Genomic platform, were generated using Chromium™ Single Cell 3’ Library & Gel Bead Kit v2 (PN-120237), Chromium Single cell 3’ Chip Kit V2 (PN-120236) and Chromium i7 Multiplex Kit (PM-120262), following the manufacturer’s instructions (10x Genomics®, Inc.). Two samples of 5,000 cells were loaded per each, HCT116 PAR and HCT116 DKO-033 libraries were sequenced paired-end 150 bp on Nextseq 500 to mean depth of 68,257 and 45,259 reads per cell for PAR and DKO cells, respectively.

Multiplexed transcriptional analysis of clonal populations

Lysed clonal populations were manually transferred in two replicates of 1 µl into MARS-seq barcoded plates. Library preparation was identical to that of single cells, with the single exception of extending the initial evaporation time on 95°C before RT1 from 3 min to 4 min, in order to compensate for the additional volume in each well.

Single-cell methylation data

We used data on HCT116 single-cell DNA methylation that was derived as controls for analyzing tumor samples (Mukamel et al., in preparation), using a variant of a previously described approach[24]. To quantify the distribution of average methylation in the founding single-cell population, we generated an ultra-low-depth library of 1,022 cells (1,045 – 52,740 uniquely mapped molecules per cell). For QC, we filtered verified rate of incomplete Cytosine conversion (or alternatively CHH methylation) is lower than 2%. We then binned CpG loci into groups according to the CpG content in the 500 bp around the locus, and estimated average methylation for independent bins (as shown in Extended Data Fig. 9e). The correlation between such bins allow some validation on the estimation noise of the average, which is verified to be robust as expected from the sampling depth of each cell. We note that for quantitative methylation analysis we rely on targeted analysis of clonal populations as described below.

Targeted methylation analysis

We used the PBAT-capture protocol combining Post-Bisulfite Adaptor Tagging (PBAT) and hybridization to an RNA probe library (capture) as described in Mukamel et al. (in preparation), which we briefly summarize here. The remaining 8 μl of each clonal population was treated with 4 μl RNase-A (20 mg/ml, ThermoFisher Scientific 1910121, diluted 1:3 in water, 30 min. in 37 °C), 3 μl Proteinase-K (20 mg/ml, ThermoFIsher Scientific 25530-49, diluted 1:1 in water, 30 min. in 65 °C), then stored in -20 °C in 96-well plate.

65 μl of Lightning Conversion Reagent (ZYMO REASEARCH, D5032-1) was added to 15 μl of each clone in each well, and bisulfite conversion was performed according to the EZ-96 DNA Methylation-Lightning MagPrep kit (D5046, Zymo) manufacturer's protocol, where 100 μl of M-Binding buffer and MagnaBeads mix in a ratio of 9:1 was added to each sample. Converted DNA samples were eluted in 38 μl elution buffer and were subjected to the first tagging step (PBAT1).

First-strand synthesis was performed in a 50 μl reaction that contained 5 μl NEB buffer 2, 2 μl dNTPs (10 mM) and 4 μl PBAT1 oligo (4 μM, Supplementary Table 10). Prior to addition of enzyme, the reaction mix was incubated at 65 °C for 3 min. followed by 4 min. at 4 °C and pause break to add 2 μl Klenow exo– (M0212L, NEB). Graduate increase of temperature +1 °C/15 sec. to 37 °C for 90 min, followed by heat inactivation of the enzyme at 70 °C for 10 min. Removing excess of oligonucleotides was done by adding 1.5 μl exonuclease I (M0293L, NEB) for 45 min. at 37 °C, followed by DNA purification by 0.8× Agencourt Ampure XP beads (A63881, Beckman Coulter) and elution in 38 μl.

Second strand synthesis was performed similarly to the first strand synthesis, using 4 μl PBAT2 oligo (4 μM, Supplementary Table 10), incubating at 95 °C for 45 sec, followed by 4 min at 4 °C and pause at 4 °C for adding 2 μl Klenow exo- during incubation at 4 °C for 5 min., +1 °C/15 sec. to 37 °C, 37 °C for 90 min., 70°C for 10 min. Excess of oligonucleotides removed by adding 1.5 μl exonuclease I, cleaning with 0.8× beads and elution in 22 μl elution buffer. Tagged products were then amplified for library preparation in 14 PCR cycles

in 25 μl Kapa HiFi Hot start ready mix kit (KK2601, KAPABIOSYSTEMS) following the manufacturer's protocol, and by using 3 μl (from a 10 μM stock) of 1:1 PBAT_PCR_For and PBAT_PCR_Rev primers mix (Supplementary Table 10). The reaction mix was then cleaned with 0.7x beads and eluted in 25 μl 10 mM Tris pH 8.

Pools of indexed libraries in same molarity (30-40 clones in each pool) were concentrated by 1x Agencourt Ampure XP beads cleanup to a volume of 10 μl (~75 ng from each clone, to 2.5 μg in pool). Each 10 μl pool was subjected to Mybait capture reaction (MYcroarray) according to the manufacturer's instructions and by adding the following modifications: The amount of baits in the hybridization mix was replaced by 2 μl baits and therefore the hybridization mix was in aliquots of 16 μl. Captured products were washed 4 times for 10 min. with washing buffer 2.2. Following amplification, the product was purified by 0.7x beads, eluted in 10 μl elution buffer and was subjected to an additional round of capture, 4 washes and 14 cycles of amplification. Final libraries were pooled and sequenced on an Illumina Nextseq system using the 150 bp high output sequencing kit.

For PBAT-capture we used a probe library designed for colon cancer analysis (Mukamel et al., in preparation), targeting 47,871 loci, out of which 9,275 were designed to enrich for promoters, 17,120 for enhancers and 21,476 sampled different non-functional epigenomic context for control.

Mapping and low-level analysis of Capture-PBAT reads

Sequenced reads were aligned to the GRCh37 (hg19) reference genome (UCSC, February 2009) using an in-house script based on Bowtie2. In cases where the two reads ends were not aligned in a concordant manner, the reads were discarded. Reads that were mapped to the same genome coordinates where marked as duplicates and were used only once. The level of incomplete conversion was estimated from loci in CHH contexts, and we have validated all libraries show lower than 2%.

Identifying copy number aberrations using PBAT data

To assess copy number trends in the data we used the total number of PBAT reads that were not mapped to on-target loci. For global comparison of HCT116 single cells to normal tissues, or of subset of the clones as shown in Supplementary Figure 5, we collected coverage statistics on bins of 1 Mb bins that showed less than 100 mapped reads in both pools were discarded. The ratio between normalized coverage was then computed and visualized over the chromosome coordinates. For analysis of individual clones, we used binning to entire chromosomal arms and compared the $\log_2$ ratio between the fractions of reads mapped to each arm in each clone, to the median of the coverage of all clones. This analysis was done separately for the two batches of clones we processed.

Mapping and low-level analysis of single cells and clonal RNA-seq data

Mapping MARS-seq reads was performed using the standard MARS-seq pipeline as described[22]. We filtered wells with less than 1,000 UMIs from further processing and verified the estimated level of empty well contamination was less than 4% in all cases. For MARS-seq libraries of clonal populations, we pooled UMIs from replicate experiments

(wells) for 300 HCT116 WT, 319 DKO, 251 KO, 266 NCI-H1299, 208 A549 and 67 WI38 clones, according to the 384-well plate design of each. For 10X single-cell analysis we used the CellRanger software for de-multiplexing, barcode processing and alignment of 10X reads. The total UMI threshold was determine by CellRanger at 4,053 for HCT116-WT cells and 2,995 for DKO cells. Further analysis of the 10X data was performed using Metacell as described below.

scRNA cell-cycle modelling

We used the MetaCell package (Baran et al.[20]), to generate cell-cycle models for the HCT116 WT, H1299, WI38 and HCT116 DKO data. For HCT116 as an example, we selected 722 candidate genes with high correlation to either MKI67, PTTG1, GINS2 or ACTB and clustered them using analysis of the correlation of UMI for single cells following downsampling to 6,422 UMIs. Supervised analysis of the derived metacells identified 320 genes correlated with either S or M-phase genes (See summary of S and M core genes of all cell lines in Supplementary Table 1). We generated a restricted count matrix including only these 320 cell-cycle genes and used the Metacell pipeline with parameters K=100 and minimal meta-cell size=50, without outlier filtering, to generate cell-cycle metacells. These were annotated using analysis of the total expression of the S-phase and M-phase gene modules. A similar protocol was used for all other cell lines. Full assignment of cells to Metacells throughout this manuscript can be found in the source code companion to this paper.

Identification of cell-cycle independent gene correlation

In order to identify cell-cycle independent transcriptional dynamics in single cells, while avoiding various biases emerging when normalizing gene counts, we developed a randomization approach aiming at estimation of the degree of correlation between any gene pair, given the cell-cycle trend alone. This was implemented by first constructing the Metacell balanced KNN similarity graph based on expression of cell-cycle genes only. This was followed by randomization of a downsampled UMI matrix by drawing the molecule count for each cell and gene from one of its 20 most similar neighbors (which were defined based on cell-cycle genes alone). We identified only genes that had at least one high correlation prior to randomization (>0.1 in HCT116 WT, >0.12 in WI38, >0.16 in H1299 and >0.15 HCT116 DKO), but had lower maximal correlation following cell-cycle randomization (at least 0.02 reduction in HCT116 WT and WI38, 0.03 in HCT116 DKO and 0.1 in H1299 cells). See Extended Data Figures 2k,m,o, and 9a, and a full summary of all cell-cycle independent genes in Supplementary Table 2.

We then performed further normalization to consider sampling depth. Given the sparse nature of the scRNA data, the probability of any pair of genes to display high correlation score inversely depends on their sampling variance, which can in turn be predicted by the total number of UMIs captured for the gene. To take this into account, we analyzed for each gene ***g*** the correlations of its original (un-normalized) UMI vector with all other genes' UMI vectors. We sorted all genes based on their total number of UMIs, and computed the empirical trend predicting correlation to gene ***g*** from total number of UMIs in any other gene. This was done using a simple rolling mean analysis with window size of 101 genes.

We subtracted this empirical trend to report for each gene pair (***g***, ***g'***) two version of a normalized correlation value (based on ***g*** and ***g'*** empirical trends). This analysis is exemplified for the *EpCAM* gene in HCT116 WT cells in Extended Data Figure 2i. Depth-adjusted correlation of cell-cycle independent genes in cells and clones are summarized in Supplementary Table 3.

Metacell analysis of long-term clones scRNA-seq

7,590 HCT116 single-cell profiles were obtained by MARS-seq for six selected clones in four different time points. We generated a Metacell model for these cells using 68 gene features that were selected to have normalized variance higher than 0.3 and at least 50 UMIs in a downsampled matrix. We filtered from the feature list any of the 320 cell-cycle genes described above. Metacell was then applied using K=100 to create cell graph and minimal metacell size = 30 cells, using 600 bootstrap iterations and generating 55 metacells ranging in size between 85 to 281 cells. These were visualized using the Metacell 2D projection as shown in the text.

Whole-exome sequencing (WES) of single-cell derived clonal populations

DNA from approximately 5-10 × 10^6 cells of the seven H1299, six A549, five HCT116 (at two different time-points) single-cell derived clonal populations, as well as of polyclonal cell population that was grown in parallel to clones, was extracted by Quick-DNA™ Universal Kit (Zymo, D4069). Exome capture sequencing of extracted DNA was done by IDTxGen lockdown human panel (Admera Health, LLC South Plainfield, NJ). Initial processing of the PE 150-bp reads was done by GATK v3.5 and mapped to hg19 reference genome, following GATK best practices (https://software.broadinstitute.org/gatk/best-practices/), and using Mutect2 module for SNP detection in each one of the clones in comparison to matching polyclonal population, when screening only homozygous sites in polyclonal samples (allowing default parameter of up to 3% variant allele frequency in each polyclonal reference population).

Defining genomic and epigenomic features

For all analyses, we used the fraction of CG in the closest 500 bp as a measure for CG content. Promoters were defined as regions spanning up to 500 bp of a gene's transcription start site (TSS). For definition of enhancers, we used ChIP-seq data of H3K4me1 in HCT116 cells (GSM945858), and classified them as regions that reside up to 500 bp from a peak of H3K4me1 (95^{th} percentile) and located at least 2 kb away from TSS to classify genomic time-of-replication we used ENCODE data of Repli-seq experiment performed on MCF7 cells (GSM923442), relying on the considerable conservation of this effect between cell lines. For Figure 3e,j, we considered CpGs with value lower than 60 (28^{th} quantile of CpGs) as CpGs of late-replicating regions. Genomic regions that reside in a distance of 100 bp or greater from our PBAT probe-set intervals were classified as off-target regions, for example in Extended Data Figure 6b (see the source code companion for detailed information on probe-set intervals).

Comparing models for persistent methylation patterns of individual loci in clonal populations (Fig. 3k,l)

We analyzed 12,536 CpGs covered at least 6 molecules in at least 60 clones. These were downsampled to retain exactly 6 methylation calls for 60 clones, such that the inter-clonal methylation variance could be computed robustly without coverage bias. We modeled the expected variance in methylation for loci govern by clonal dynamics involving no memory by assuming independent sampling from a variable that is methylated with probability p (i.e. $V^{\text{mix}}(\hat{p}) = 6\hat{p} * (1 - \hat{p})$, where $\hat{p}$ is the empirical average methylation). A model assuming perfect memory for a locus with empirical methylation average $\hat{p}$ was estimated by sampling methylation of two founding epi-alleles independently with probability $\hat{p}$ (i.e. assuming Hardy-Weinberg equilibrium), and then sample six methylation calls for the clone randomly from the selected two epi-alleles. The variance of this two-step process (computed by exhaustive summation on the number of reads sampled from the first epi-allele) was defined as $V^{\text{pers}}(\hat{p})$.

Given the two models, we computed for each CpG a *deviation from persistency* value by subtracting its empirical inter-clonal variance from its expected $V^{\text{pers}}(\hat{p})$. As shown in Figure 3l, we defined “cold” CpGs as those that show minor deviation from the expected inter-clonal variance according to the perfect memory model (absolute deviation is lower than 0.3), and “hot” CpGs as those showing lower variance than expected variance according to this model (difference higher than 3). We limited this screen to partially methylated CpGs, where we have the most power to distinct between the methylation dynamics regimes (average methylation in the population p between 0.3 and 0.7).

Testing for differential expression and identifying GEMINIs

GEMINIs were defined as genes that are almost always repressed but show one or few clones with significantly high level of expression. To screen for GEMINIs we used clone RNA profiles that were downsampled to 10k UMIs for WT, DKO and KO clones. We identified for each gene the maximal expression level over clones and computed its fold-change to the 90% expression percentile, assigning a “GEMINI score” to every gene (as shown in Fig. 4a). GEMINIs were restricted to basal expression of up to 1 UMI / 10k UMIs, where basal expression of each gene computed as the average expression across all clones, except the 10% of clones with highest expression of it. GEMINIs also required to have at least 4 UMIs in their maximal clone, and at least 5 UMIs overall. In practice, genes with the highest ratio showed total expression of 5-590 UMIs (median of 46), and maximal expression of 4-106 (median 7). We then assigned a “repressed” or “de-repressed” states for each clone regarding each GEMINI, requiring at least 50% of maximal expression level for a particular GEMINI to define a “de-repressed” state in clone. The number of clones showing de-repression was 1 for 69 genes, 2 for 11 genes and over 2 for 18 genes.

Hi-C analysis and Shaman normalization

We re-analyzed HCT116 data from Rao et al. 2017, using the Shaman package as described (Olivares-Chauvet et al. 2016, Bonev et al. 2017). TADs were called using computation of insulation profiles as described before (Bonev et al. 2017, see example in Extended Data

Fig. 10c). Overall 3,690 TADs were defined ranging in size between 51 kb - 38 Mb (median = 397 kb), and we focused our analysis to TADs with at least four annotated TSSs. Each TAD's total expression was defined as the total UMIs per clones from genes with their TSSs within it. We computed correlation between each gene's expression to all TADs total expression across all clones. To compute the correlation between a gene to the TAD it resides in ("TAD Auto-correlation") we first subtracted the gene's expression from its total TAD's output. To screen for surprising gene-TAD expression correlations, we subtracted from the TAD Auto-correlation score the 20th highest correlation of the gene with any TAD:

$$TAD_{Auto_{correlation}score_{gene}} = TAD\ Auto\ correlation - TAD\ top\ correlation\_20_{gene}$$

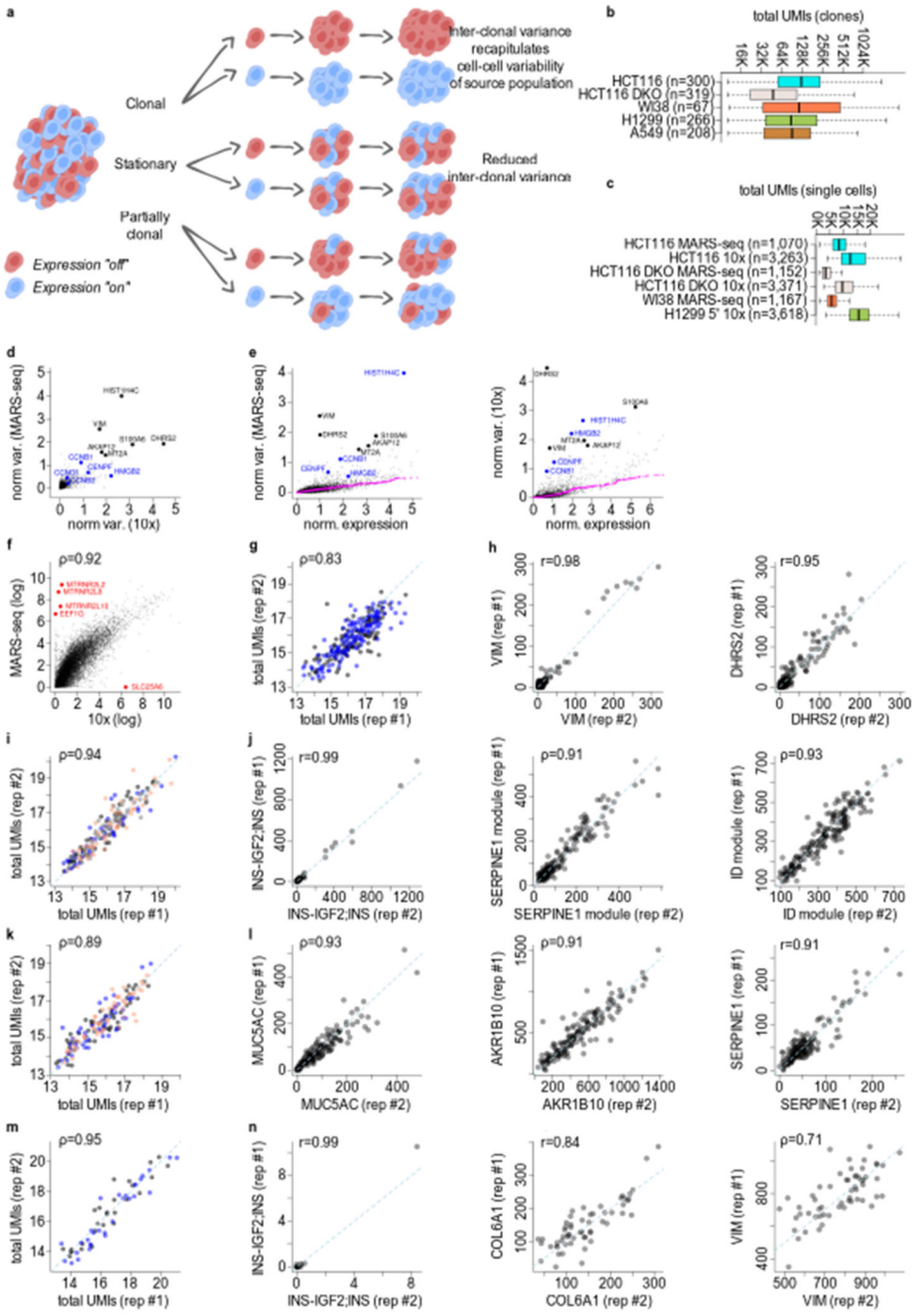
To visualize normalized contact maps of specific genomic regions (as in Fig. 5f), we used SHAMAN package to first normalize local cis-decay over observed contacts by using the *shaman_shuffle_and_score_hic_mat()* function, and then to plot the resulted map with function *shaman_plot_map_score_with_annotations ()*. SHAMAN source code is available at https://github.com/tanaylab/shaman.

TCGA analysis

All RNA-seq and methylome datasets that were downloaded from the TCGA repository (http://cancergenome.nih.gov/) and re-analyzed in this manuscript are listed in Supplementary Table 11.

Europe PMC Funders Author Manuscripts

Extended Data

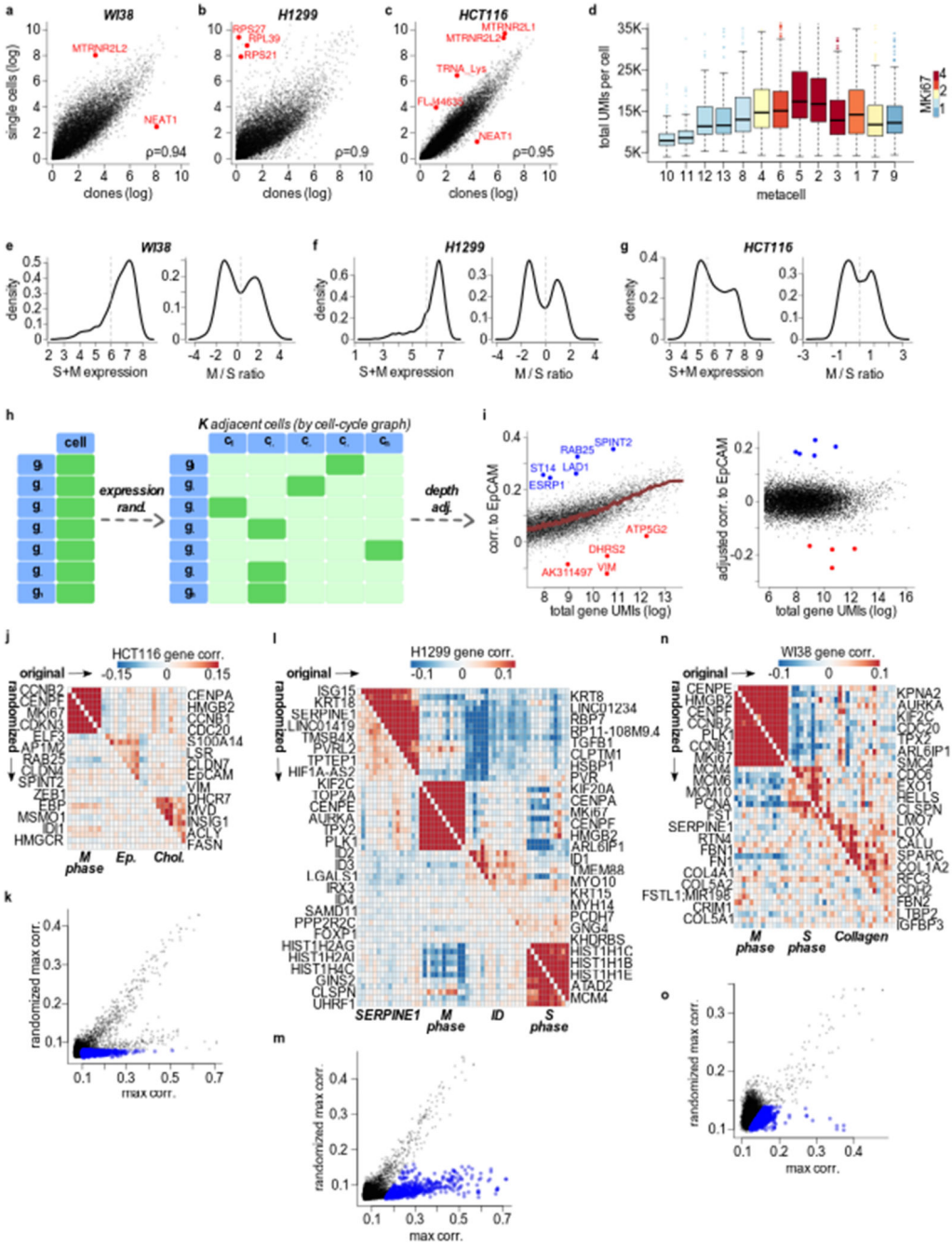


Extended Data Fig. 1. MARS-seq in short-term clonal populations.

a, Schematic readout of transcriptional memory test using a Luria-Delbrück design. **b,** Distributions of the total number of UMIs obtained per clone in different cell-lines. n = number of clones profiled. **c,** Distributions of total UMIs obtained per cell in different cell-lines. n = number of cells profiled. **d,** Normalized expression variability ($\log_2$(variance/mean)) per gene in singles cells obtained by 10x (x-axis) and MARS-seq (y-axis). Genes

with high normalized variance are annotated. Blue - cell-cycle markers. **e,** Normalized gene expression variance in HCT116 cells. Selected variable genes (black) and cell-cycle markers (blue) are annotated. Purple line is showing a roll-median trend. For both plots, cells are down-sampled to 6K UMIs. **f,** Normalized pooled expression of common 17,949 genes in single cells obtained by 10x (x-axis) and MARS-seq (y-axis). Expression values were computed as $\log_2$ of UMIs / 10K UMIs. Five top differential genes are annotated in red. **g,** Total $\log_2$ UMI counts in two MARS-seq technical replicates of 260 HCT116 well covered clonal populations (>10K UMIs in both replicates). Color of dots indicates first (black) or second (blue) culturing batches. **h,** Normalized expression of selected variable genes and gene-modules in technical replicates of HCT116 clones. **i-j,** as in **g-h**, for 199 H1299 clonal populations. All replicates were covered by at least 5K UMIs (random pairs of quadruple experiments are shown) **k-l,** As in **g-h**, for 157 A549 clonal populations where all replicates covered by at least 5K UMIs. **m-n,** As in **g-h**, for 57 WI38 clonal populations where all replicates are covered by at least 5K UMIs. Three randomly selected replicates were selected and summed to represent a single technical. ρ values represent Spearman's correlations and r values represent Pearsons's.

Europe PMC Funders Author Manuscripts



Extended Data Fig. 2. Identification of cell-cycle independent transcription variation of HCT116, H1299 and WI38 single cells.

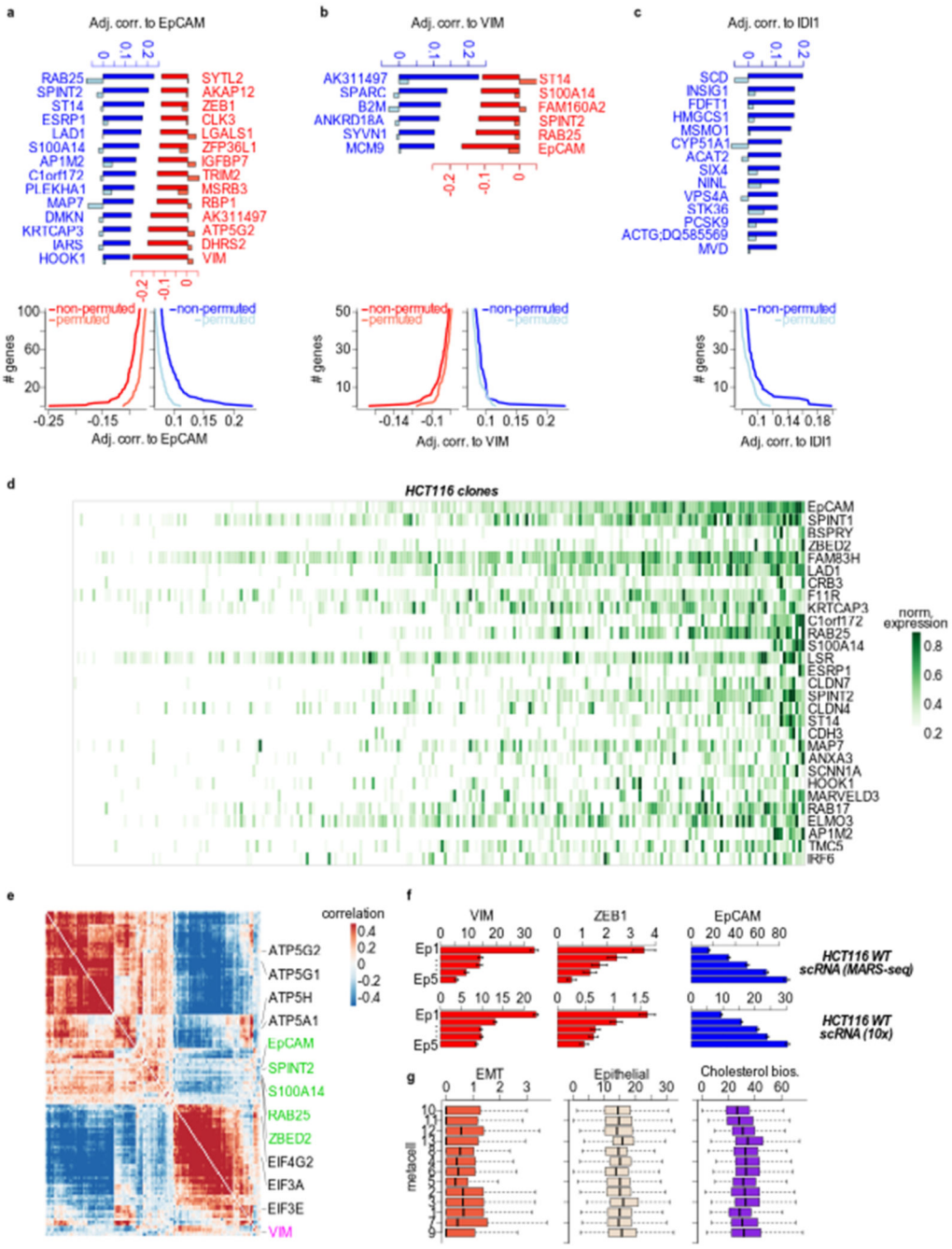
a-c, Normalized pooled expression in clonal populations (x-axis) and single-cells (y-axis) in WI38 (left), H1299 (center) and HCT116 (right) cells. Expression values computed as $\log_2$ of UMIs / 10K UMIs. Genes with high differential expression in each system are displayed by red dots and annotated. n (WI38, HCT116) = 27,052, n (H1299) = 17,698. **d,** Distributions of total number of molecules per cell in inferred cell-cycle metacells of HCT116. Colors are as in Fig. 1b. **e,** $\log_2$ total expression signatures (left) and ratio of cell-

Europe PMC Funders Author Manuscripts

cycle phases (right) in WI38 cells. Right panel shows only cells that were annotated as replicating in left panel. **f,** as in **e** for H1299 cells. **g,** as in **e** for HCT116 cells. A full list of genes used in this assay for all cell lines is in Supplementary Table 1. **h,** Illustration of expression randomization in each cell according to cell-cycle based cell-cell similarity graph. **i,** Showing scRNA gene profiles correlation with EpCAM expression, controlled by each gene total expression (left), with a running median shown in red. Following subtraction of the trendline, correlations are generally independent of gene sampling depth (right). **j,** Matrix of gene-gene correlations in HCT116 cells before (upper triangle) and after (lower triangle) cell-cycle based randomization. Showing selected cell-cycle related (Supplementary Table 1) and unrelated (Supplementary Table 3) gene modules. Number of analyzed cells defined in Extended Data Fig. 1c. **k,** Maximal correlation value of each gene with another gene before (x-axis) and after (y-axis) cell-cycle based randomization. Loss of correlation (blue dots) indicates that the co-expression patterns of this gene were independent of the cell-cycle, thus eliminated by the randomization. **l-m,** as in **j-k**, for NCI-H1299 cells. **n-o,** as in **j-k**, for WI38 cells.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

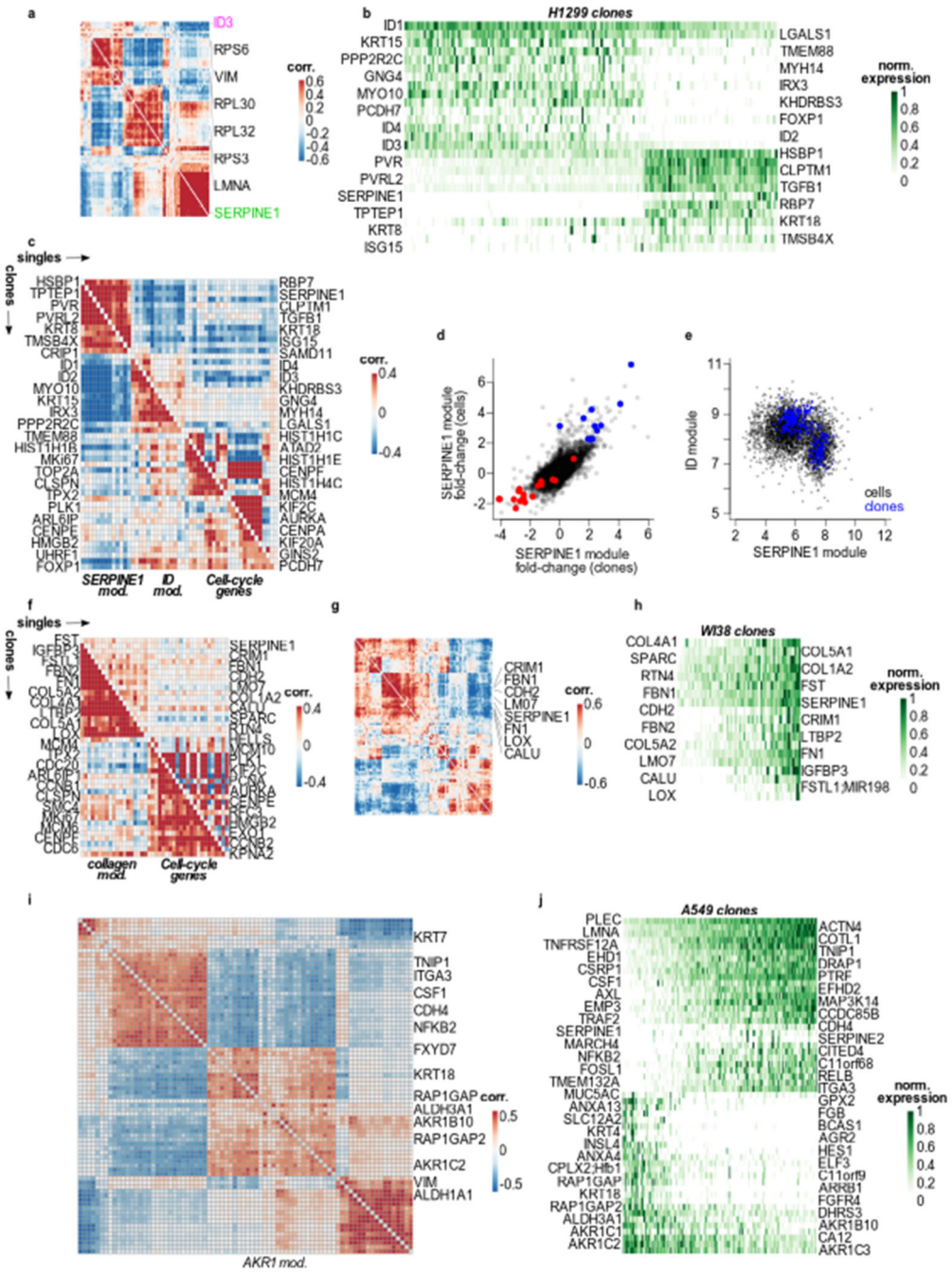


Extended Data Fig. 3. Cell-cycle independent gene modules in HCT116 cells.

a-c, Spearman's correlations (depth-adjusted) of HCT116 scRNA-seq gene profiles without (blue, red) and following (light blue, tomato) permutation. Bar graphs show top positively (left) and negatively (right) correlated genes with EpCAM, VIM and IDI1, with the respective distributions of original and permutated correlations shown at the bottom. **d,** Normalized expression of epithelial module genes in HCT116 clones. For each gene (row), expression is divided by maximal value observed in clones. Displaying 233 clones (columns) covered by at least 50K UMIs. **e,** Matrix showing clustered gene-gene

correlations of all genes defined to maintain strong cell-cycle independent co-variation in Extended Data Fig. 2K (and summarized in Supplementary Table 2). Labels of genes related to epithelial module shown in **d** are colored in green, and its anti-correlated gene Vimentin (VIM) is colored in magenta. **f,** As in Fig. 1j for clones, we grouped cells obtained by MARS-seq (top) and 10x (bottom) into five bins based on expression of the EpCAM gene module (Ep5 consisting of cells with highest module expression). Bars are showing mean expression of each bin for EpCAM gene (blue) and for genes negatively enriched in EpCAM high cells (red). Error-bars represents standard error of binomial distribution. **g,** Distribution of normalized expression of Cholesterol (purple), Epithelial (antique-white) and EMT genes (red), binned and ordered according to the cell-cycle associated HCT116 metacells shown in Fig. 1b.

Europe PMC Funders Author Manuscripts

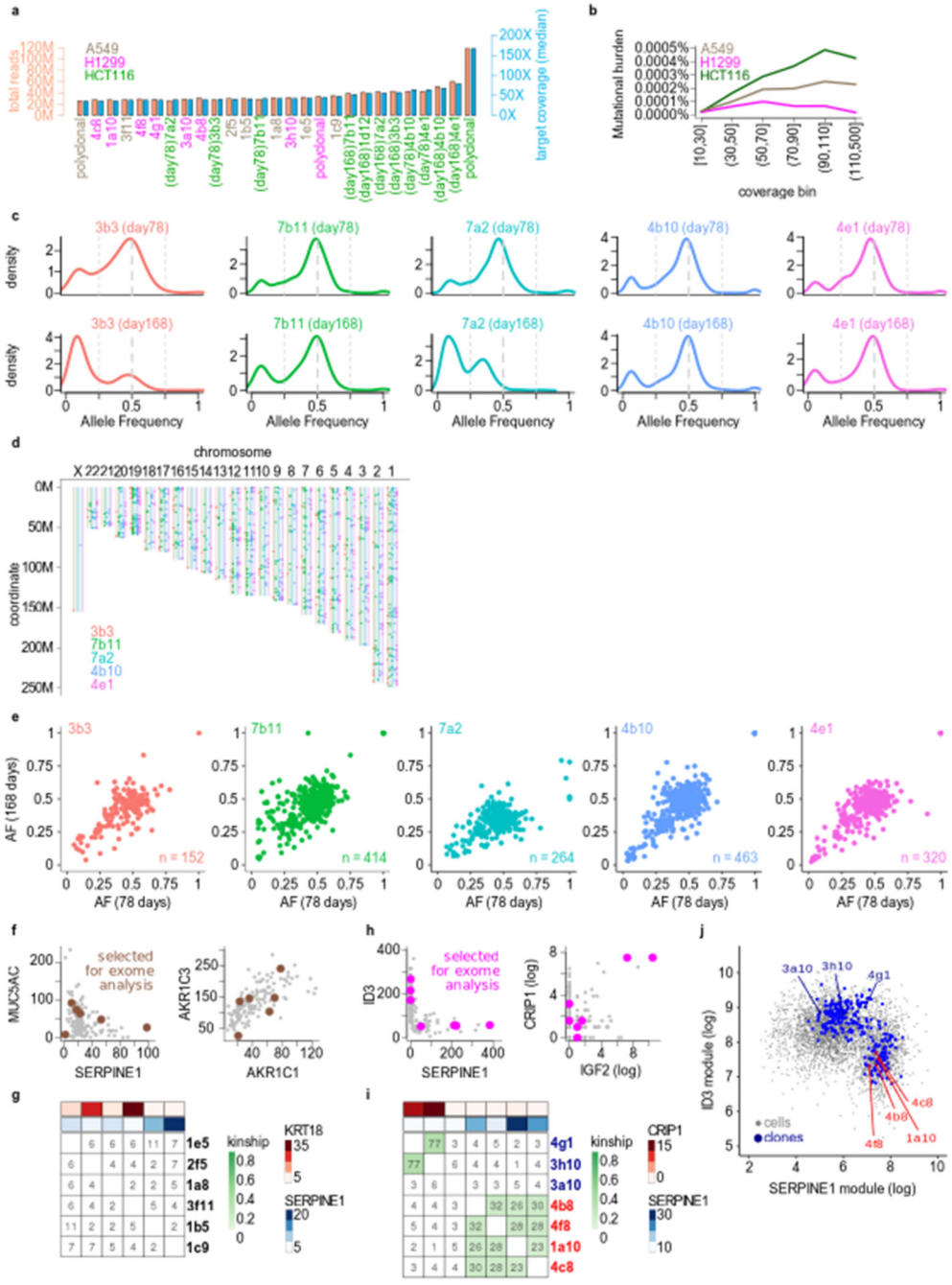


Extended Data Fig. 4. Cell-cycle independent gene modules in H1299, WI38 and A549 cells.

a, As shown in Extended Data Fig. 3e for HCT116, clustering gene-gene correlations of all H1299 cell-cycle independent genes labeled in Supplementary Fig. 3d (and summarized in Supplementary Table 2). Number of cells and clones analyzed from each cell-line are defined in Extended Data Fig. 1b,c. **b,** As in Extended Data Fig. 3d, showing normalized expression of ID and SERPINE1 gene modules in all H1299 single-cell derived clones**. c,** Comparison of NCI-H1299 gene-gene correlation over single cells (upper triangle) and clones (lower triangle). **d,** As in Supplementary Fig. 1d, showing for each gene the $\log_2$-ratio

of relative expression in high SERPINE1 (top 25% H1299 cells and 20% H1299 clones) and low SERPINE1 (lower 30% cells and 40% clones) cells (x-axis) and clones (y-axis). Labeling genes of the ID module (red dots) and of the SERPINE1 gene module (blue dots). **e,** Total output (log-normalized expression) of SERPINE1 gene module (x-axis) and ID gene module (y-axis) in cells and clones shows a bi-modal, clonally conserved population structure in the NCI-H1299 system. **f,** As in **c,** Comparison of WI38 gene-gene correlation over single cells (upper triangle) and clones (lower triangle). **g,** As in **a**, clustering gene-gene correlations of all WI38 cell-cycle independent genes labeled in Extended Data Fig. 2o. Showing black labels for collagen module genes. **h,** As in **b**, showing normalized expression of Collagen module genes in WI38 clones. **i,** Gene-gene correlation of most variable genes in A549 clones. Labels of selected gene are shown on right. **j,** As in Extended Data Fig. 3d, showing normalized expression of variable genes in A549 clones.

Europe PMC Funders Author Manuscripts



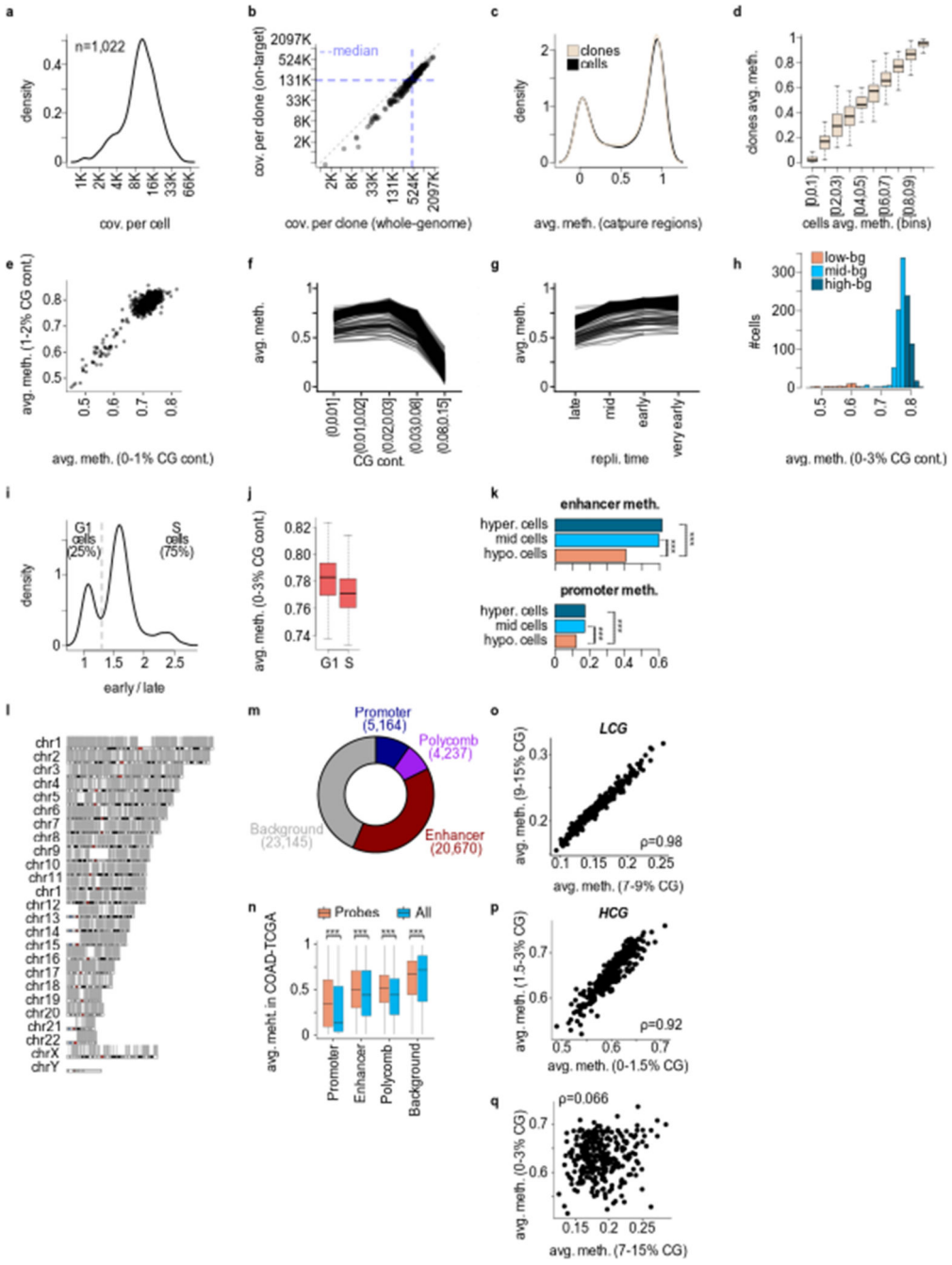
Extended Data Fig. 5. Longitudinal whole exome sequencing (WES) analysis of selected clonal populations.

a, Coverage Summary of 27 Whole Exome Sequencing (WES, see Methods) experiments. Total reads obtained per sample (orange) and median on-target coverage per base (blue) are shown. Other stats and WES quality control are available in Supplementary Tables 12,13. **b,** Fraction of SNPs detected per coverage bin in different cell lines (mutational burden). Calls from all clones were aggregated per cell line. Coverage per base was obtained by DepthOfCoverage module of GATK v3.5. **c,** Allele frequency (AF) distributions of detected

Europe PMC Funders Author Manuscripts

variants in HCT116 clones sampled after 78 days (top) and 168 days (bottom). **d,** Spatial distribution of SNPs detected in HCT116 clones. **e,** Comparison of allele frequencies in five HCT116 single-cell derived clones after 78 days (x axis) and 168 days (y axis). If selection was greatly affecting the process, allele frequencies were not expected to remain stable as observed in practice. **f,** Expression of marker genes in six A549 clones that were selected for exome analysis (colored in brown). **g,** Similar to Fig. 2c, kinship analysis of A549 clones. Rows above column show normalized expression of KRT18 (red) and SERPINE1 (blue) genes in each clone. **h,** Selection of seven NCI-H1299 clones (colored in magenta). **i,** Kinship analysis as in Fig. 2c for NCI-H1299 clones. **j,** Normalized expression of the SERPINE1 and ID modules in H1299. Single cells represented by small grey dots. Clones profiled by WES are labeled in black and red (as in panel **i**). Note the concordance between the genetic and transcriptional sub-clonal structure for these cells.

Europe PMC Funders Author Manuscripts

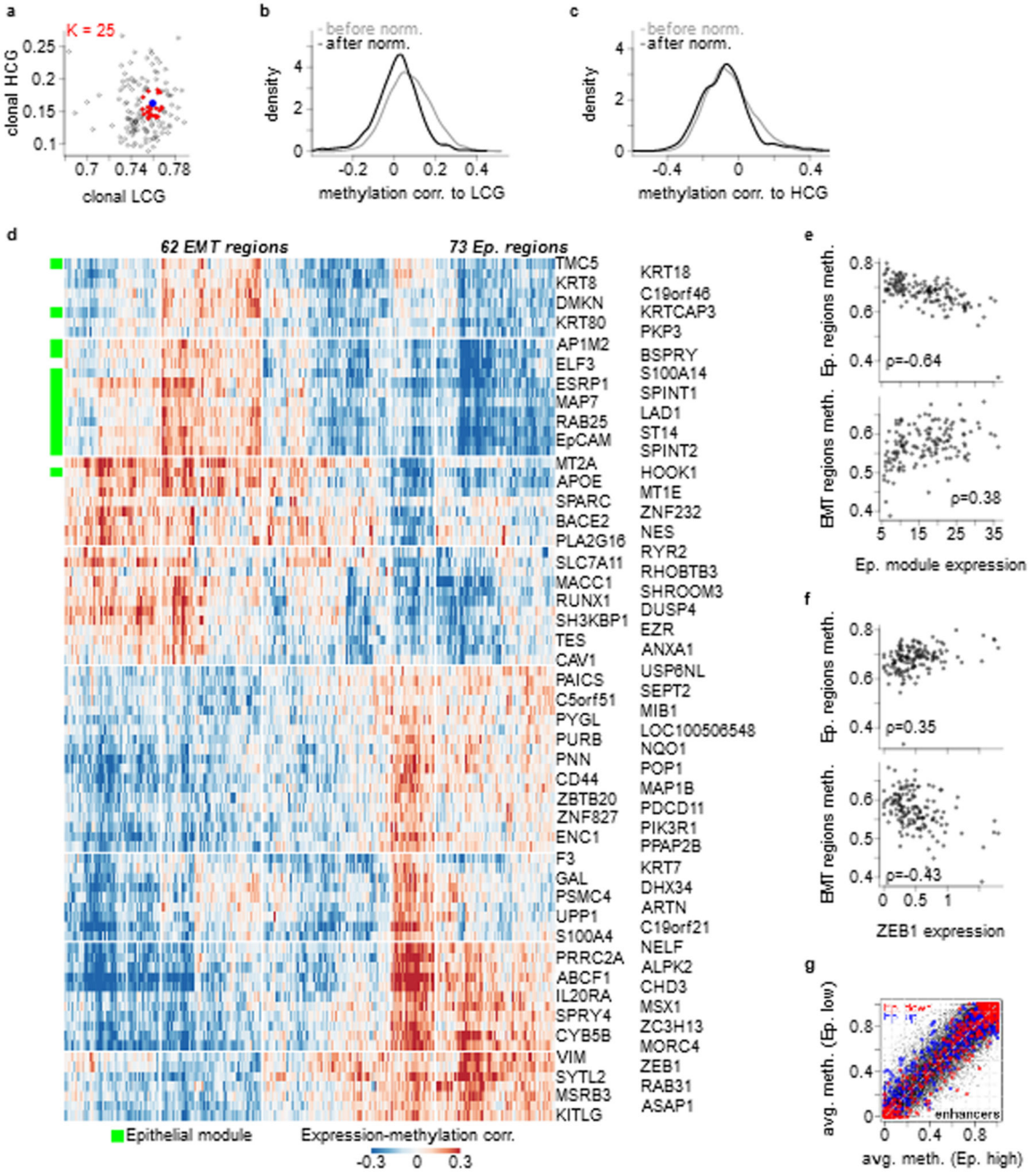


Extended Data Fig. 6. scPBAT and PBAT-capture of HCT116 clones using a colon cancer oriented probe-set.

a, Distribution of methylation calls in low-depth HCT116 single-cell PBAT analysis. **b,** Whole genome coverage and on-target coverage for HCT116 clonal populations. Coverage = total number of methylation calls. **c,** Density plot of pooled average methylation of on-target regions in single cells (black line) and clonal populations (antique-white line). **d,** Distributions of pooled averaged methylation of on-target regions in clones, grouped by their respective pooled average methylation in single cells (regions with > 50 calls in single cells

Europe PMC Funders Author Manuscripts

and > 500 calls in clones, n = 341, 69, 49, 49, 39, 32, 67, 77, 186, 595). **e,** Pooled average methylation of individual cells in very low (0-1%) and low (1-2%) CG-content regions. **f,** Trends showing the correspondence between DNA methylation and CpG content for 1,022 single cells. **g,** Similar to **f**, showing correlation with genomic replication time. **h,** Single cell methylation average in regions with low CG-content (0-3%), defining classification into low, mid and high-bg cells **i,** Distribution of the $\log_2$ ratio of coverage of genomic sequences in early- and late-replicating regions. Vertical dashed grey line is defining the threshold for classifying single cells into G1 and S phase. **j,** Distribution of average methylation per cell in genome-wide low CpG regions (0-3%) for cells inferred to be in G1 and S phases in panel **i**. $n_{G1} = 254$, $n_S = 767$. **k,** Average promoter and enhancer (Methods) methylation in groups of single cells For all groups, n > 6000, chi-squared test, in all cases $P < 2*10^{-16}$. **l,** Genomic spatial distribution of colon PBAT-capture probe set. **m,** Number of regions covered by the probe set, stratified by genomic context. **n,** Distribution of methylation of covered regions in TCGA colon cancer (COAD). Shown is average methylation of CpGs that reside within (blue) and outside (orange) the PBAT-capture probe set, grouped by genomic context (for all comparisons n > 6000, two-tailed KS test: D > 0.11, $P < 2*10^{-16}$). **o-q,** Average methylation of 293 TCGA colon cancer tumors (COAD), in different ranges of CpG content.



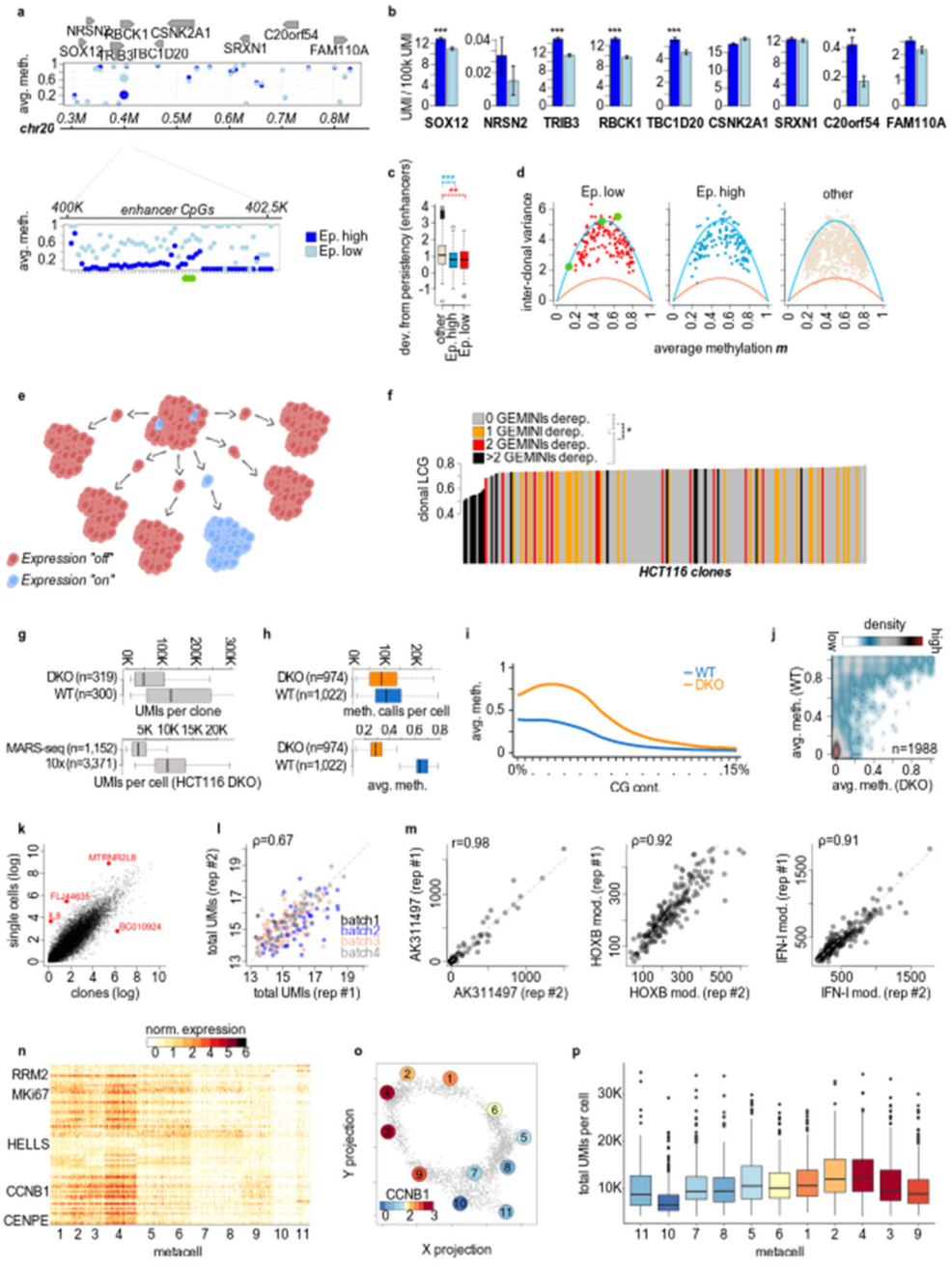
Extended Data Fig. 7. Clonal methylation at functional regions is association with epithelial transcriptional output in HCT116.

a, Example for selection of clones for KNN-based normalization of DNA methylation over the clonal HCG (y axis) and LCG (x axis) space. Red dots indicate the K = 25 nearest neighboring clones used to normalize methylation of the selected clone (shown as blue dot). **b,** Distribution of correlations between average methylation of capture regions in clones to average methylation of clones in Low-CG (LCG) loci before (grey) and after (black) KNN normalization. **c,** same as **b** for High-CG (HCG) loci. **d,** Clustering of Spearman's cross-

Europe PMC Funders Author Manuscripts

correlation between gene expression and normalized average methylation of capture regions over 251 HCT116 clones covered simultaneously by RNA-seq and PBAT-capture. Green annotation of genes indicates epithelial genes. **e,** Epithelial transcriptional output per clone (x-axis) and clonal average methylation (y-axis) in 73 capture regions defined in **d** as Epithelial regions (Ep. regions, top) and 62 capture regions defined in **d** as EMT related regions (bottom) in 155 HCT116 clones, covered by at least 50K UMIs and 50K on-target methylation calls. **f,** As in **e**, showing expression of EMT related gene Zinc finger E-box-binding homeobox 1 (ZEB1) and methylation in Epithelial (top) and EMT (bottom) associated capture regions defined in **d**. **g,** Pooled average methylation of enhancer CpGs in EpCAM-high and -low clones, highlighting enhancers of epithelial up- (blue) or down-regulated (red) genes.

Europe PMC Funders Author Manuscripts

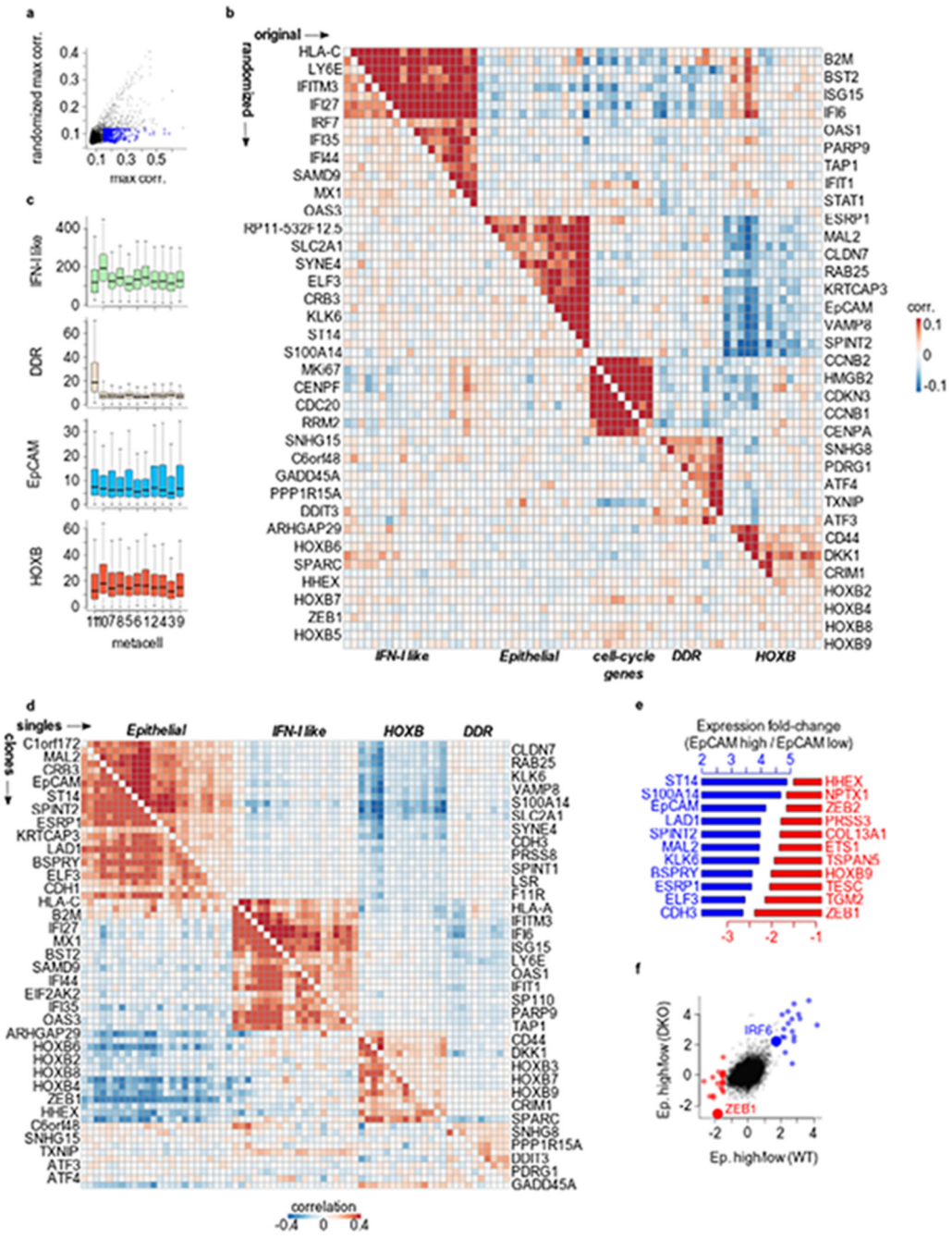


Extended Data Fig. 8. GEMINIs rationale and cell-cycle modelling of DKO HCT116 cells.

a, Average methylation of EpCAM-high (blue, n = 51) and EpCAM-low (light-blue, n = 51) clones over a region of chromosome 20 (Top panel - 5kb bins, lower magnification: single CpGs). Green dots mark “cold” CpGs as defined in Fig. 3l. **b,** Bars indicate pooled expression levels in EpCAM-high and -low clones for genes within the genomic region shown in **a**. Chi-squared *P* values: TRIB3, RBCK1 < $2*10^{-16}$, SOX12 = $7*10^{-6}$, TBC1D20 = $2*10^{-5}$, C20orf54 = $1.3*10^{-3}$, CSNK2A1 = $3*10^{-3}$. **c,** Distribution of deviation from persistency (blue trend in Fig. 3l, see Methods) of enhancer CpGs. Ep-high and Ep-low

Europe PMC Funders Author Manuscripts

represent CpGs with differential methylation of 0.1 or higher in EpCAM high vs. low clones. n_{other} = 767, $n_{Ep.high}$ = 122, $n_{Ep.low}$ = 152. Two-tailed KS test, Ep-high: D = 0.2, P = $4*10^{-4}$. Ep-low: D = 0.16, $P = 3*10^{-3}$. **d,** Showing inter-clonal variance (Fig 3l) for enhancer CpGs colored as in **c**. Green - epithelial-related CpGs in chr20 as in panel a. **e,** Schematics of the screen for GEMINIs. **f,** Bars indicate clones' LCG average methylation, color-coded by the number of GEMINIs de-repressed in it. Two-tailed KS test (D = 0.22, $P = 0.039$), comparing LCG average methylation for clones with and without GEMINIs (excluding VIM-high clones). **g,** Coverage depth of DKO transcriptome. **h,** Statistics of single-cell-PBAT methylation profiles of 974 DKO cells (orange boxes) and 1,022 WT cells (blue boxes). **i,** Pooled average methylation in WT (blue line) and DKO (orange line) cells, as a function of genomic CpG content. **j,** Distribution of pooled methylation of DKO and WT HCT116 cells (x-axis), showing 1,988 CpGs with n > 8 calls in both pools. **k,** Normalized pooled expression (log_2 UMI / 10K Umis) in DKO clonal populations (x-axis) and DKO single-cells (y-axis). Genes with highest differential expression are highlighted. **l-m,** Reproducibility of technical replicates in MARS-seq for 203 DKO clones, showing total UMI counts (log_2 transformed) in two MARS-seq technical replicates and normalized expression of selected variable genes and gene-modules. Rho (ρ) represents Spearman's correlation coefficient and r represents Pearson's. **n-p,** Cell-cycle analysis of 3,371 single DKO cells, as shown in Fig. 1b and Extended Data Fig. 2d for wild-type HCT116.



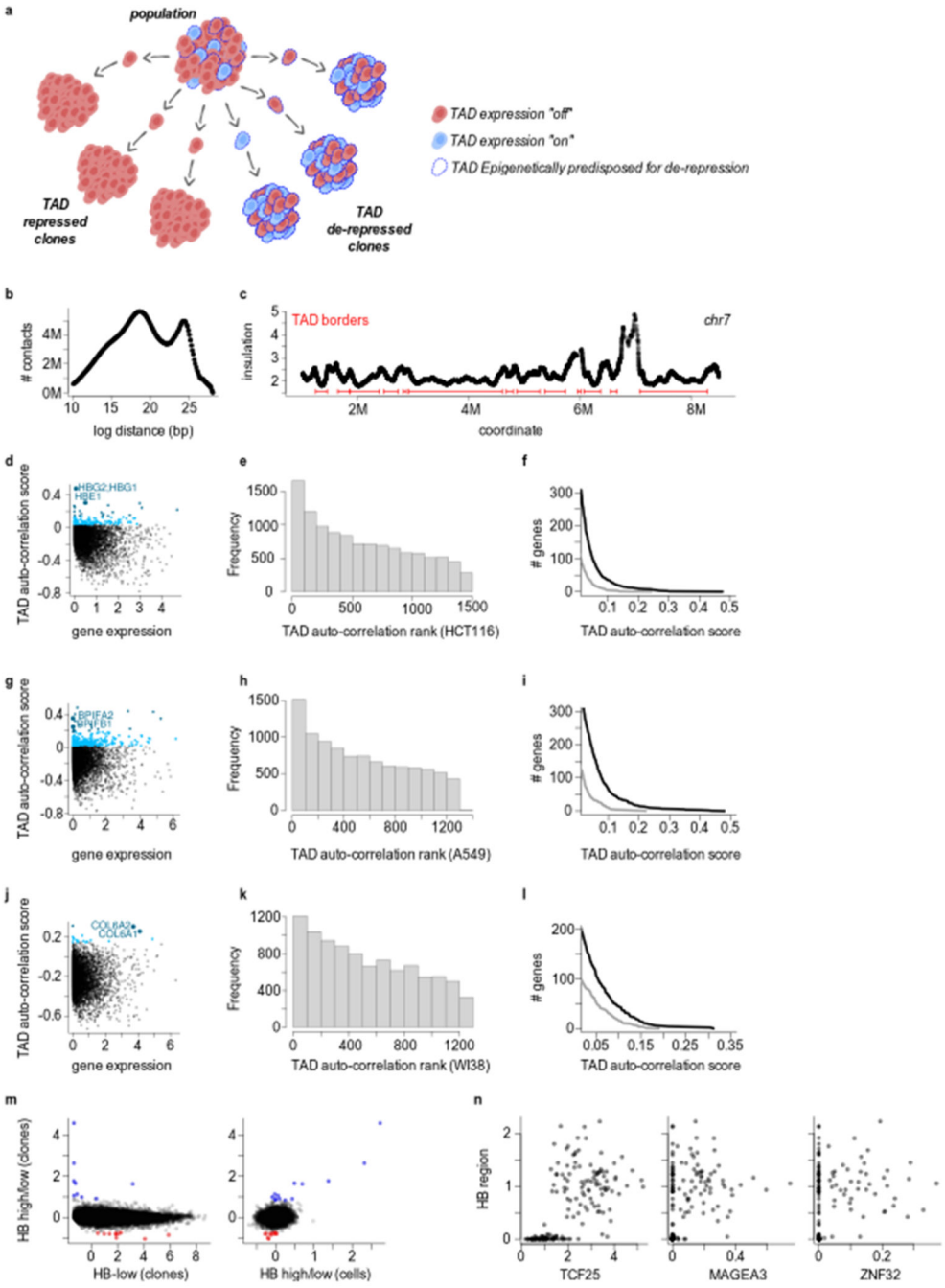
Extended Data Fig. 9. Identification of cell-cycle independent transcriptional variance in DKO HCT116 single cells.

a, Maximal correlation values of each gene with another gene before (x-axis) and after (y-axis) cell-cycle based randomization of DKO cells (blue dots indicate genes maintaining cell-cycle variance, for full list see Supplementary Table 2). **b,** Matrix of gene-gene Spearman's correlations in DKO cells before (upper triangle) and after (lower triangle) cell-cycle based randomization. **c,** Distribution of gene module expression per cell, classified by cell-cycle associated metacells in DKO (as defined in Extended Data Fig. 8n-p). **d,** Matrix of

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

gene-gene Spearman's correlations in DKO single cells (upper triangle) and DKO clones (lower triangle), indicating cell-cycle independent gene modules summarized in Supplementary Table 3. **e,** Genes with highest (blue bars) and lowest (red bars) expression change between EpCAM high and low DKO clones. **f,** Comparing gene expression fold-change of EpCAM high and low clones in HCT116 WT (x-axis) and DKO (y-axis).



Extended Data Fig. 10. Screening for in-TAD transcriptional memory in HCT116, A549 and WI38 cells.

a, Schematics of the screen for TAD de-repression. Clones can maintain deterministic repression of transcription in TADs that are “closed”. De-repression of a TAD in a clone can result in stochastic (possibly uncorrelated) de-repression of genes within it. **b,** Distribution of contact distances for 488M HCT116 Hi-C contacts. **C**. TADs are defined between two picks of insulation (y-axis), as exemplified here for a segment of chromosome 7. **d,** Showing log mean expression in HCT116 clones (x-axis) and TAD auto-correlation scores (y-axis, see Methods). Genes showing statistically significant (positive) auto-correlation are labeled (light-blue for FDR < 0.25 and dark-blue for FDR < 0.05), for full list see Supplementary Table 9). **e,** We computed the correlation of expression between all genes to all TADs, and for each gene we measured the rank of its TAD auto-correlation. Shown is the distribution of these TAD auto-correlation ranks (value of 1 means the gene’s own TAD was the most correlated to it). **f,** Cumulative distribution of TAD auto-correlations in HCT16 clones, for observed data (black line) and for shuffled data (randomly assigning genes to TADs, grey line). **g-i,** Same screen as in **d-f** for A549 clones. **j-l,** Same screen as in **d-f** for WI38 clones. **m,** Showing fold-change expression of genes in HB-high vs. HB-low HCT116 clones (y-axis), over expression in HB-low clones (x-axis, left) fold-change in HB-high single cells vs. HB-low cells (x-axis, right). **n,** Expression across HCT116 clones of selected genes that correlate with expression of the HB gene module (x-axis), compared to expression from genes in the beta-globin chromosomal domains (y-axis).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Omer Schwartzman for assistance with WES analysis, Pedro Olivares-Chauvet for help with analysis of Hi-C data and all members of the Tanay group for fruitful discussions. We also wish to thank Efrat Hagai, Gitit Levi-Cohen, Ayala Sharp and Ziv Porat from WIS flow cytometry unit. Research was supported by the European Research Council (scAssembly), the Israel Science Foundation, and the Chan Zuckerberg Initiative. A.T. is a Kimmel investigator.

Data Availability

Raw and transformed data are available at GEO and SRA (Accession GSE144357).

Code Availability

Metacell[20] code is available in https://tanaylab.github.io/metacell/. Source code, additional metadata and a vignette exemplifying cell-cycle normalization with MetaCell are available at https://github.com/tanaylab/Meir_et_al_nat_gen_2020_clonemem.

References

1. Hanahan D, Weinberg RA. Hallmarks of Cancer: The Next Generation. Cell. 2011; 144:646–674. [PubMed: 21376230]
2. Zwemer LM, et al. Autosomal monoallelic expression in the mouse. Genome Biology. 2012; 13:R10. [PubMed: 22348269]

3. Iberg-Badeaux A, et al. A Transcription Factor Pulse Can Prime Chromatin for Heritable Transcriptional Memory. Mol Cell Biol. 2017; 37
4. Shaffer SM, et al. Memory sequencing reveals heritable single cell gene expression programs associated with distinct cellular behaviors. 2018; doi: 10.1101/379016
5. Vardi N, Levy S, Assaf M, Carmi M, Barkai N. Budding yeast escape commitment to the phosphate starvation program using gene expression noise. Curr Biol. 2013; 23:2051–2057. [PubMed: 24094854]
6. Shipony Z, et al. Dynamic and static maintenance of epigenetic memory in pluripotent and somatic cells. Nature. 2014; 513:115–119. [PubMed: 25043040]
7. Choi M, et al. Epigenetic memory via concordant DNA methylation is inversely correlated to developmental potential of mammalian cells. PLoS Genet. 2017; 13
8. Arand J, et al. Selective impairment of methylation maintenance is the major cause of DNA methylation reprogramming in the early embryo. Epigenetics Chromatin. 2015; 8
9. Li E, Bestor TH, Jaenisch R. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. Cell. 1992; 69:915–926. [PubMed: 1606615]
10. Liu X, et al. UHRF1 targets DNMT1 for DNA methylation through cooperative binding of hemi-methylated DNA and methylated H3K9. Nature Communications. 2013; 4:1563.
11. Roadmap Epigenomics Consortium. et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015; 518:317–330. [PubMed: 25693563]
12. Horvath S, Raj K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. Nature Reviews Genetics. 2018; 19:371–384.
13. Cruickshanks HA, et al. Senescent cells harbour features of the cancer epigenome. Nat Cell Biol. 2013; 15:1495–1506. [PubMed: 24270890]
14. Irizarry RA, et al. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nat Genet. 2009; 41:178–186. [PubMed: 19151715]
15. Klutstein M, Nejman D, Greenfield R, Cedar H. DNA Methylation in Cancer and Aging. Cancer Research. 2016; 76:3446–3450. [PubMed: 27256564]
16. Shih AH, Abdel-Wahab O, Patel JP, Levine RL. The role of mutations in epigenetic regulators in myeloid malignancies. Nature Reviews Cancer. 2012; 12:599–612. [PubMed: 22898539]
17. Zhou W, et al. DNA methylation loss in late-replicating domains is linked to mitotic cell division. Nature Genetics. 2018; 50:591–602. [PubMed: 29610480]
18. Maxfield KE, et al. Comprehensive functional characterization of cancer–testis antigens defines obligate participation in multiple hallmarks of cancer. Nat Commun. 2015; 6
19. The Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012; 487:330–337. [PubMed: 22810696]
20. Baran Y, et al. MetaCell: analysis of single-cell RNA-seq data using K-nn graph partitions. Genome Biol. 2019; 20:206. [PubMed: 31604482]
21. Milyavsky M, et al. Prolonged Culture of Telomerase-Immortalized Human Fibroblasts Leads to a Premalignant Phenotype. Cancer Research. 2003; 63:7147–7157. [PubMed: 14612508]
22. Jaitin DA, et al. Massively parallel single cell RNA-Seq for marker-free decomposition of tissues into cell types. Science. 2014; 343:776–779. [PubMed: 24531970]
23. Brocks D, Chomsky E, Mukamel Z, Lifshitz A, Tanay A. Single cell analysis reveals dynamics of transposable element transcription following epigenetic de-repression. bioRxiv. 2019; doi: 10.1101/462853
24. Luo C, et al. Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. Science. 2017; 357:600–604. [PubMed: 28798132]

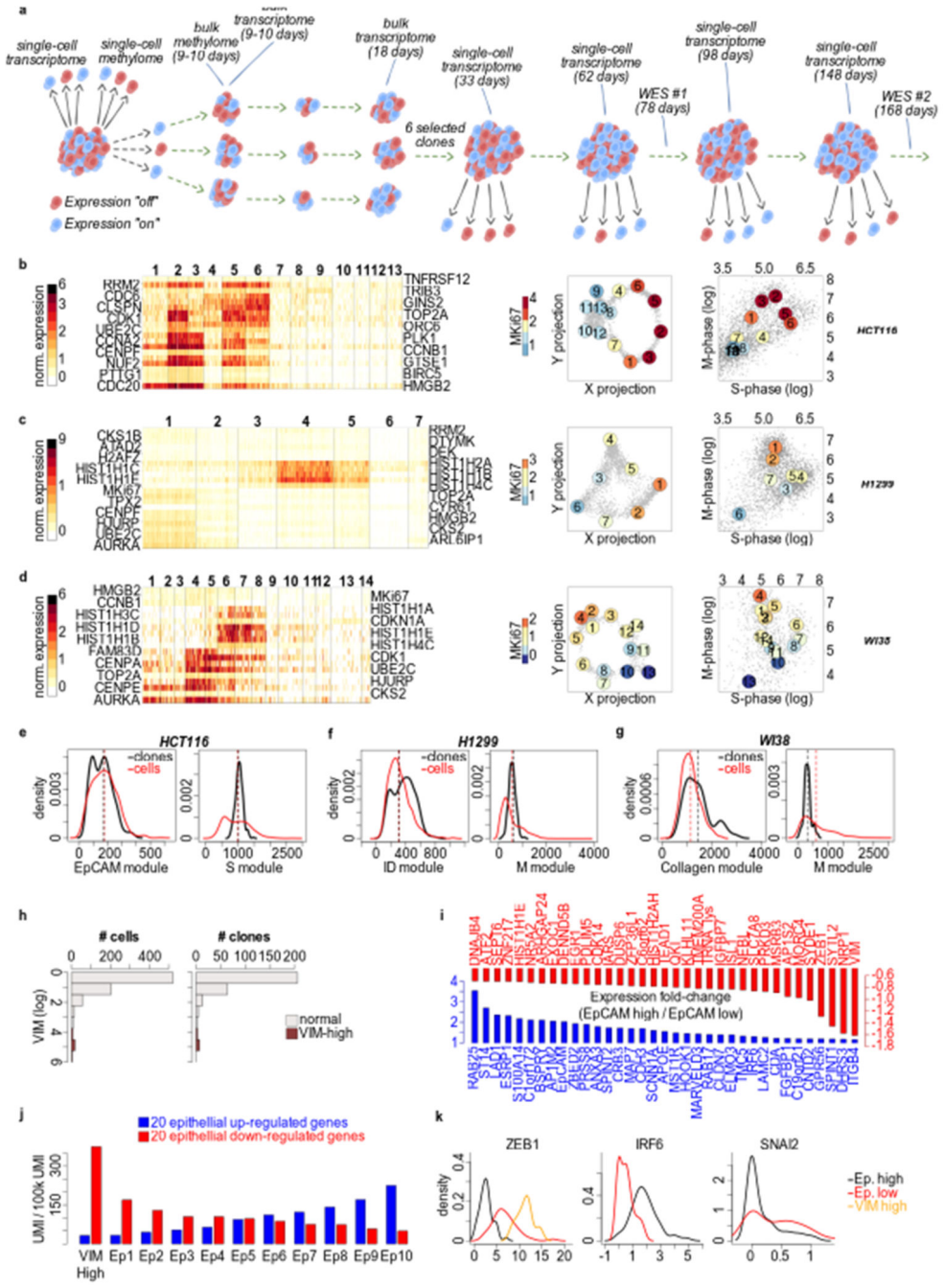


Figure 1. A Luria-Delbrück design for testing transcriptional and epigenetic memory.

a, Schematic overview of our experimental design. Green dashed arrows: culturing steps. Black dashed arrows: sorting of single cells into conditioned media. Non-dashed arrows/ lines: processing steps. WES: whole-exome sequencing. **b,** Left: Expression of selected genes over 3,255 HCT116 single cells (columns) grouped into metacells (top labels) according to similarity in cell-cycle gene expression. Center: 2D Projection of the cell-cycle metacell model. Metacells)large ovals) are color coded according to the expression intensity of the cell-cycle marker MKi67, cells are shown as small gray dots. Right: Comparing

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

expression of M-phase and S-phase genes (Supplementary Table 1) for cells and metacells. **c,** As **b**, for 3,584 NCI-H1299 single cells. **d,** As in **b**, for 1,172 WI38 single cells. **e,** Normalized expression (UMI per 100k UMIs) distribution in HCT116 cells and clones of epithelial (EpCAM) and S-phase gene modules (as detailed in Supplementary Table 1 and Supplementary Table 3). **f,** As **e**, for ID module and M-phase in NCI-H1299 cells. **g,** As **e**, for Collagen module and M-phase in WI38 cells. **h,** Distribution of VIM expression (log_2 of UMI per 10k UMIs) in HCT116 single cells (left) and clones (right). **i,** log_2 expression fold changes for genes enriched in EpCAM high clones (blue, top 30% of clones, n = 51) and EpCAM low clones (red, lowest 30% of clones, n = 51), after exclusion of 11 VIM-high clones. For all genes shown here, FDR corrected q-value « 0.001, chi-squared test. **j,** Shown is the total UMIs for genes positively (blue) and negatively (red) enriched in EpCAM-high clones, in clones grouped based on expression of the *EpCAM* gene module. **k,** Density plots of expression in EpCAM-high (black line), EpCAM-low (red line) and 11 VIM-high clones (orange line) for selected genes.

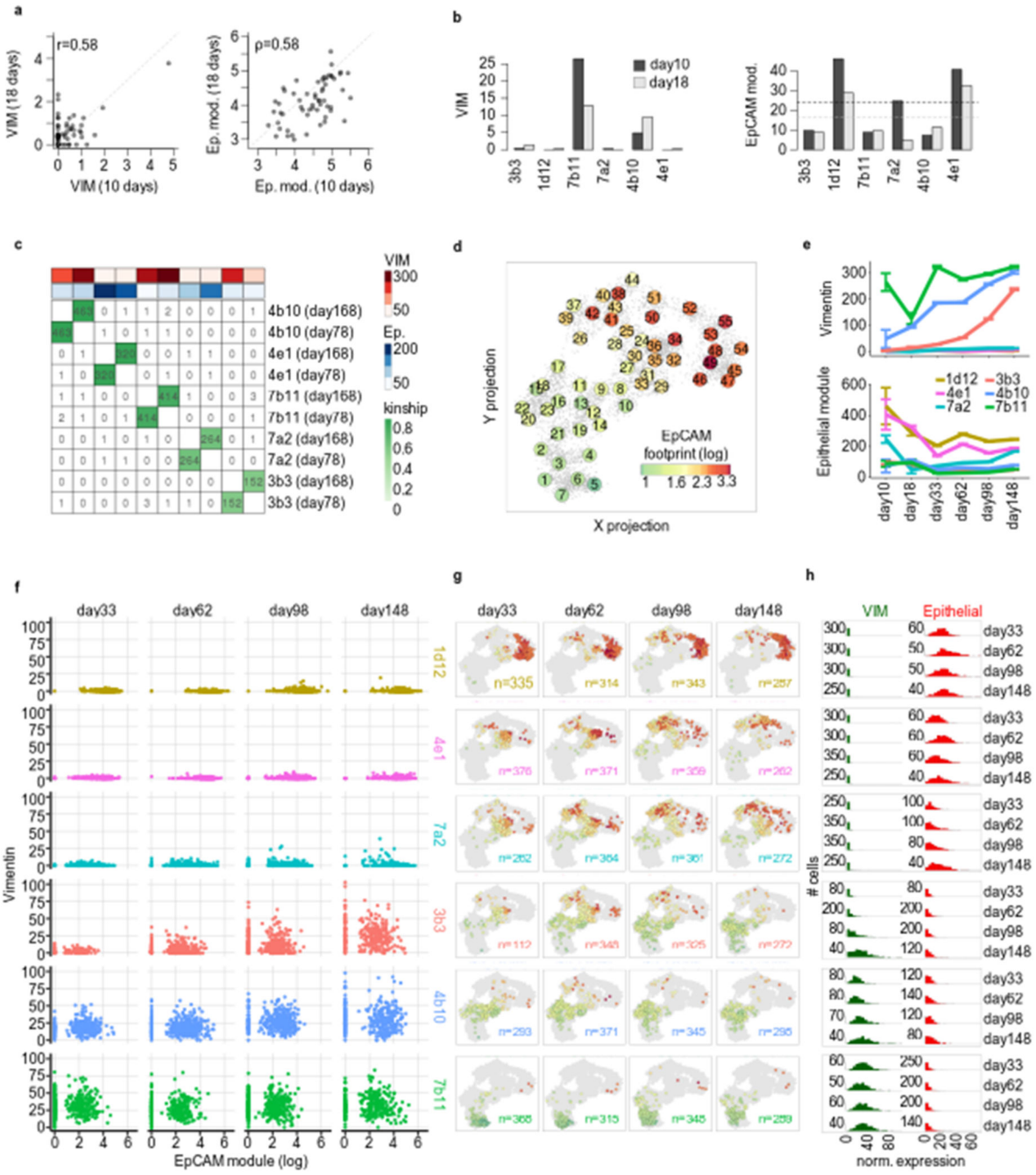


Figure 2. Long-term clonal maintenance of Epithelial and VIM-high transcriptional states.

a, Expression (log_2 count per 10k UMI), of VIM (left panel) and the EpCAM module (right panel) for clonal populations that were sampled twice, after 10 and 18 days (> 15 k UMI per sample in both). **b,** Expression (note linear scale) of VIM and the EpCAM module in six clones selected for further longitudinal analysis. Dashed horizontal line represents median expression over all clones sampled after 10 (grey, n = 59) and 18 (dark grey, n = 59) days. **c,** Analysis of genetic kinship between HCT116 clones. Text in each cell shows the absolute count of shared SNPs between two clones. Upper bars show VIM (red) and EpCAM gene

module (blue) expression (pooled single cells RNA in the closest time point)**. d,** Metacell 2D projection for single-cell RNA-seq data from longitudinal analysis of six clones. Colors represent the level of *EpCAM* expression. **e,** Average expression of *VIM* (upper panel) and epithelial genes (lower panel) in tracked clones over six time-points (clonal RNA-seq at day = 10 and 18; pool of single-cell RNA-seq at day = 33, 62, 98 and 148). See panel **g** below for the number of cells at each time-point. Error-bars represent SE of binomial sampling, based on total sampled UMIs per clone per time-point. **f-h,** For each clone (row), showing total epithelial and VIM transcriptional output per cell by time-point (**f**); 2D projection of clones' single cells, coloring according to the EpCAM module expression intensity (**g**) and the changes in VIM and the EpCAM gene module expression distributions over time (**h**).

Europe PMC Funders Author Manuscripts

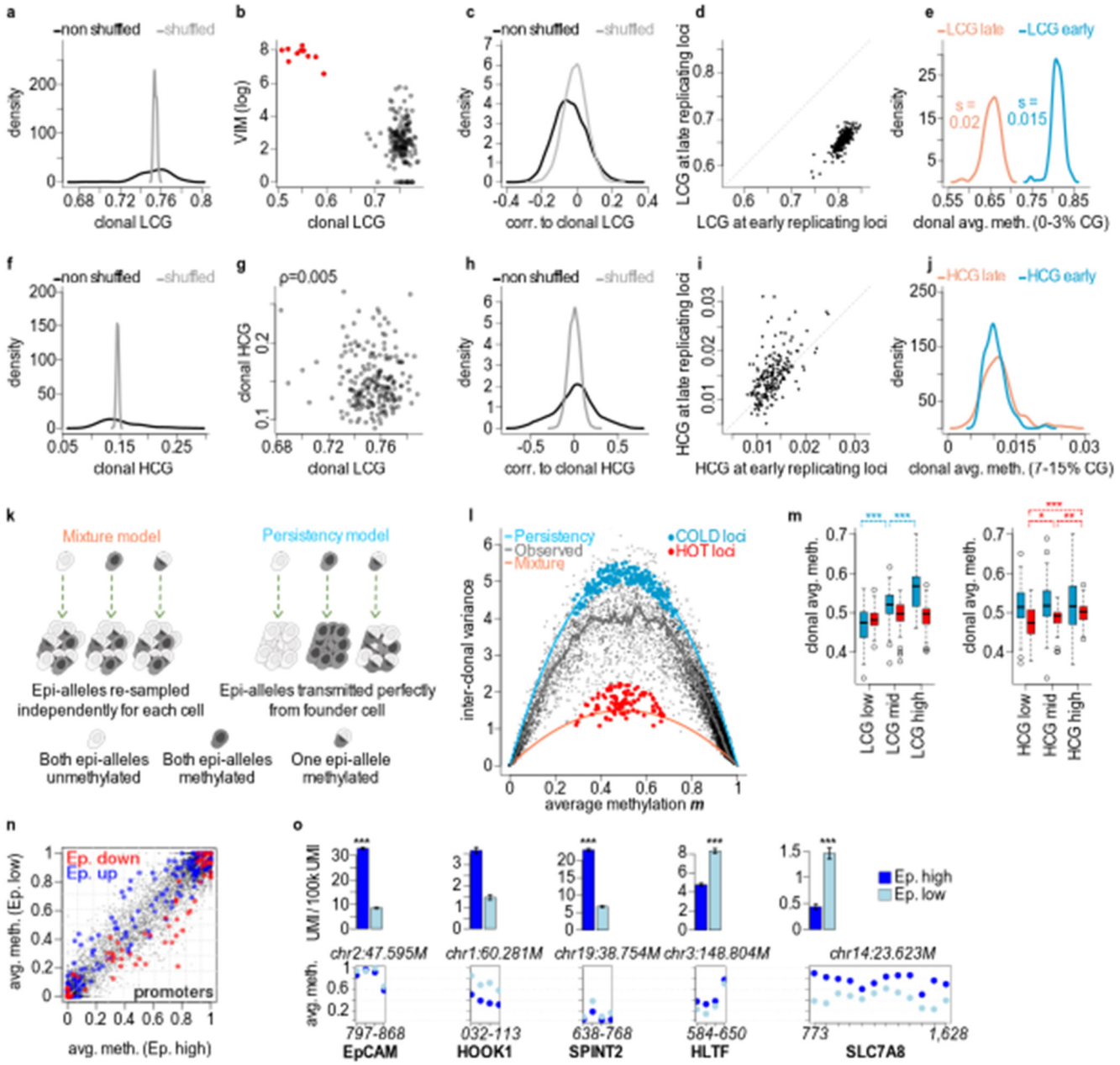


Figure 3. Hot and cold dynamics of clonal methylation.

a, Distribution of clonal average methylation in low CpG content (LCG, 0-3% CG content) loci (observed vs. shuffled control, excluding VIM-high clones as defined in Fig. 1h). **b,** *VIM* expression by clonal LCG methylation. VIM-high clones are colored in red. **c,** Distribution of Spearman correlations between LCG methylation and gene expression over all genes. Controls are based on shuffling clonal LCG values. **d,** Clonal LCG methylation in early- and late-replicating genomic domains. **e,** Distributions of early and late replicating loci methylation over clones, indicating by s the standard deviations. **f,** As in **a**, for high CpG content (HCG, 7-15% CG content) loci. **g,** clonal LCG vs HCG average methylation, indicating lack of correlation. **h,** As in **c**, for HCG methylation. **i,** As in **d**, for HCG

methylation (but excluding loci with overall average methylation higher than 0.3). **j,** As in **e**, for HCG methylation. **k,** We simulated two alternative methylation dynamics in clonal population assuming zero memory (left, mixture model) and perfect memory (right, persistency model). **l,** Shown are inter-clonal methylation variance vs. average methylation across well covered loci (running median is defined by a gray curve). Blue and orange lines depict the variance predicted by the persistency and mixture models (see Methods). Red and Blue dots mark partially methylated loci showing empirically “hot” (high turnover, mixture model) and “cold” (low turnover, persistency model) dynamics, respectively. **m,** We grouped 192 clones with sufficient coverage by their LCG (left) or HCG (right) methylation (minimal group size = 54, excluding VIM-high clones). Boxplots depict distribution of average methylation in hot and cold loci across the groups. In all boxplots throughout this manuscript we used R version 3.5.3 defaults for boxplot() function – where middle line indicates median, box limits are quantiles, and whiskers are 1.5 × IQR. Kolmogorov-Smirnov two-tailed test, LCG-high to LCG-mid: D = 0.39, $P = 3 \times 10^{-5}$. LCG-low to LCG-mid: D = 0.49, $P = 9 \times 10^{-7}$. HCG-high to HCG-mid, D = 0.31, $P = 3 \times 10^{-3}$. HCG-low to HCG-mid: D = 0.29, $P = 0.01$. HCG-high to HCG-low: D = 0.39, $P = 1 \times 10^{-4}$. **n,** Pooled average methylation of promoter CpGs in EpCAM-high (n = 51) and –low (n = 51) clones, highlighting promoters with up-regulated (blue) or down-regulated (red) expression in EpCAM-high clones (D = 0.3, $P = 8 \times 10^{-5}$, KS two-tailed, $n_{blue} = 257$, $n_{red} = 88$). **o,** Promoter methylation in clones showing high (n = 51) and low (n = 51) EpCAM expression. Bars showing average expression and error-bars represent SE of binomial sampling. Chi-squared test, all P values $< 2 \times 10^{-15}$. Panels below bars indicate chromosomal coordinates and show average methylation of covered CpGs in EpCAM-high (blue dots) and -low (light blue dots).

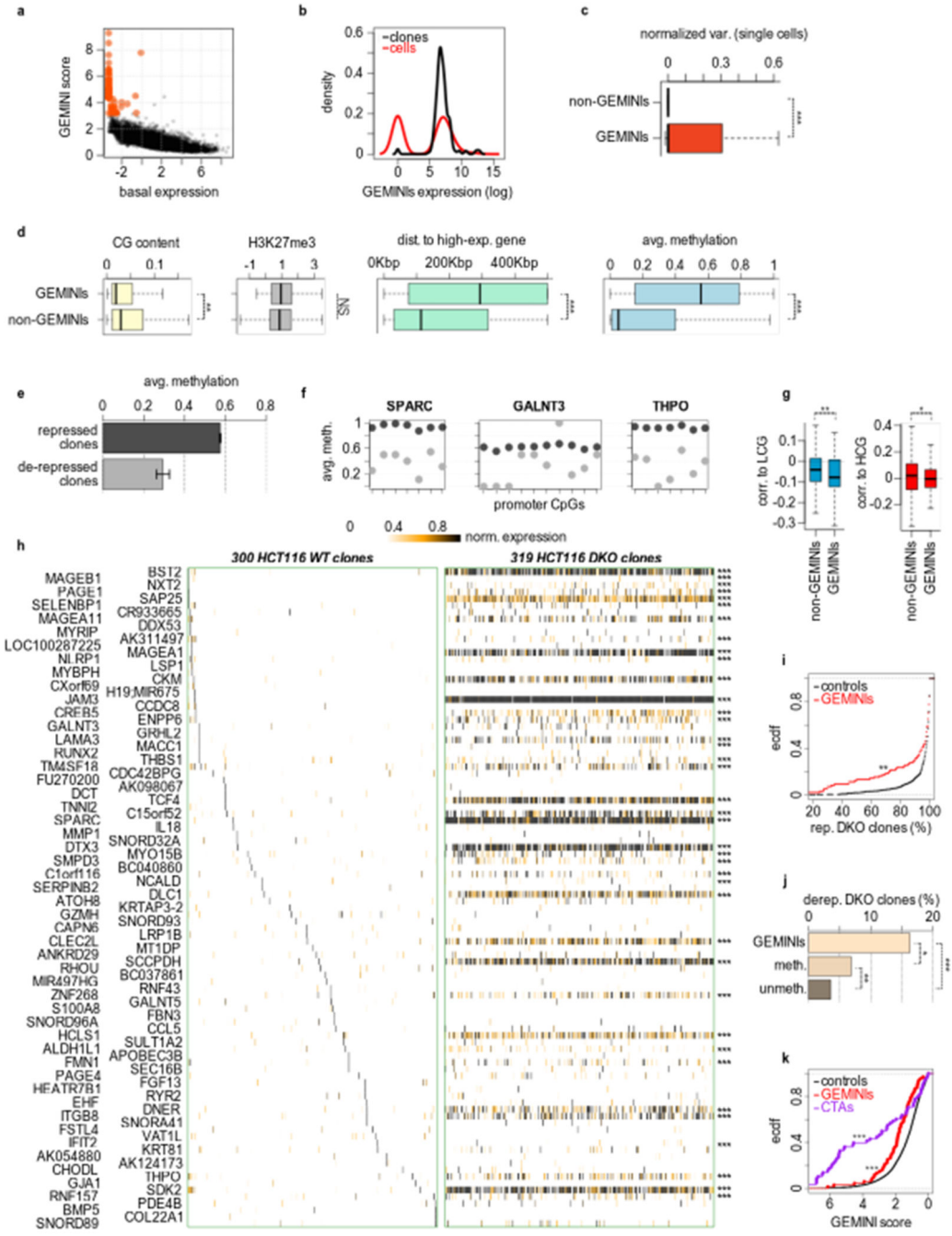


Figure 4. Screening for Genes that Escaped Mitotically Inherited Inhibition (GEMINIs).

a, 98 GEMINIs were selected based on genes with low basal expression (less than 1 UMI / 10k UMIs) and high maximum de-repression (red dots, see Methods). For each gene, showing averaged basal expression across HCT116 clones (x-axis, see Methods), and GEMINI score indicating rare de-repression in few clones (y-axis, see Methods). **b,** Density plot of overall GEMINI pooled expression per cell (red line) and clone (black line). **c,** Distributions of normalized variances in single cells ($\log_2$ of variance-to-mean ratio) for 98 GEMINIs and 969 randomized matched controls with similar expression levels and

Europe PMC Funders Author Manuscripts

promoter CG content. Two-tailed KS test, D = 0.21, $P = 9 \times 10^{-4}$. **d,** Distributions of genomic features of GEMINIs and randomized controls matched for expression levels. High-exp gene: TSS within top 20 expression percentiles. Two-tailed KS test, from left to right: CG (D = 0.13; $P = 8 \times 10^{-3}$, compared to matched-controls of expression only), Distance (0.22; 8×10^{-8}), Methylation (0.35; 2×10^{-6}). **e,** We annotated each clone with a "repressed", or "de-repressed" state regarding each one of the GEMINIS (see Methods). Bars showing average methylation of pooled GEMINI promoters in their repressed (black bar, n = 20,797) and de-repressed clones (grey bar, n = 201). Error-bars represent sampling SE, based on total methylation calls in de-repressed or repressed clones. Chi-squared test, $\chi^2 = 51$, $P = 9 \times 10^{-13}$. **f,** Average methylation of selected GEMINIs in their de-repressed clone. **g,** Distribution of gene expression correlation to clonal LCG (blue boxplots, as in Fig. 3c), and HCG (red boxplot, as in Fig. 3h) for GEMINIs and for matched controls. Two-sided KS test, LCG (D = 0.19; $P = 5 \times 10^{-3}$), HCG (D = 0.16; $P = 0.035$). **h,** Left panel: GEMINIs (rows) expression in 300 HCT116 WT clones (columns). Expression levels are normalized by maximal expression value of each of the 98 GEMINIs. Order of columns is determined by clonal LCG average methylation. Right panel: Expression of GEMINIs in DKO clones. Expression is normalized by the maximal expression of each GEMINI in its de-repressed WT clone (*** FDR corrected q-value < 0.001, KS test for comparison of GEMINIs expression in HCT116 WT and HCT116 DKO clones). **i,** Cumulative distribution for the fraction of repressed DKO clones for GEMINIs (red line) and for the matched randomized subset of control genes with similar CG-content and expression levels in WT (black line). De-repression threshold was set to half of the maximal normalized expression in WT clone. Two-sided KS test, D = 0.2, $P = 2 \times 10^{-3}$. **j,** Comparison of de-repression in DKO for methylated and unmethylated genes. Showing data only on genes that were considered repressed in HCT116 WT (< 0.25 UMIs / 10k UMIs in at least 80% of the clones). Pooled methylation data on clones used to define methylated (> 0.9, n = 365) and unmethylated (< 0.9, n = 4,160) promoters. Chi-squared test: GEMINI/meth: $\chi^2_{.} = 5.9$; $P = 0.015$; GEMINIS/unmeth (32; 2×10^{-8}); meth/unmeth (8.1; 4×10^{-3}). **k,** Cumulative distribution of genes showing rare de-repression (see Methods for definition of GEMINI score) in 399 colon adenocarcinoma RNA-seq samples (obtained from TCGA). GEMINI score compared expression in maximal tumor to 95'th expression quantile. Showing only genes that were generally repressed in both HCT116 clones and TCGA samples. CTAs represent 77 annotated Cancer-Testis Antigens. Two-tailed KS test, Geminis (D = 0.28, $P = 5 \times 10^{-5}$), CTA (0.42; 2×10^{-9}).

Europe PMC Funders Author Manuscripts

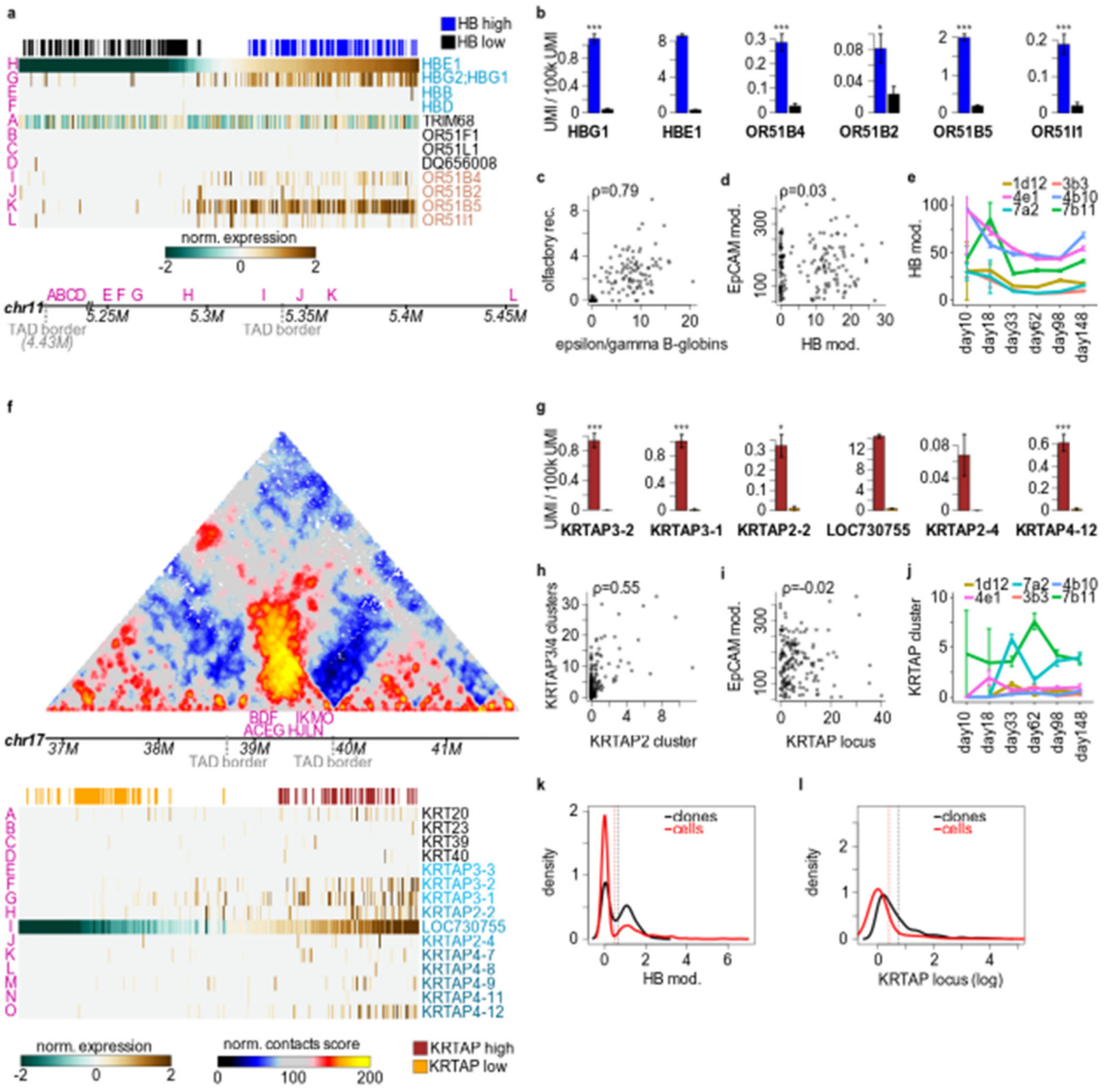


Figure 5. Evidence for in-TAD memory.

a, Normalized expression of genes spanning the beta-globin TAD on chromosome 11. Heatmap is showing gene (rows) expression over clones (columns), normalized to the gene's median expression across clones. Top bars annotate clones as HB-high and -low, (VIM-high clones are excluded). Spatial map using one letter encoding (left of heatmap) is shown at the bottom, also indicating TAD borders as grey dashed vertical lines. **b,** Pooled average expression in HBE1-high (n = 76, blue) and HBE1-low (n = 77, black) clones defined in **a**. Error bars represent SE of sampling a binomial distribution. Chi-squared *P* values: HBG1,OR51B5 < 2×10^{-16}, OR51B4 = 5×10^{-12}, OR51B2 = 0.018, OR51I1 = 9×10^{-8}. **c,**

Europe PMC Funders Author Manuscripts

Expression of embryonic (HBE1) and fetal (HBG2, HBG1) beta-globin genes (x-axis) vs. expression of adjacent cluster of olfactory receptors (y-axis). Dots represent clones and Spearman correlation is indicated here and in other figure panels. n = 168 clones covered by >100k UMIs. **d,** Expression of the HB (x-axis) and EpCAM (y-axis) modules, indicating lack of correlation. **e,** Temporal change in HB module expression for six clones, similar to Figure 2e. Error bars represent sampling SE, based on total sampled UMIs per clone per time-point. **f,** As in **a**, for KRTAP region on chromosome 17. Spatial map of genes is drawn on top, and the SHAMAN-normalized contact frequency map is depicted as a triangle respective to the region's linear coordinates (HCT116 Hi-C data obtained from Rao et al. 2017, see Methods). **g,** As in **b,** Bars indicate average pooled expression in LOC730755-high (brown, n = 50) and LOC730755-low (orange n = 55) clones defined in **f**. Error bars represent SE of binomial sampling. Chi-squared *P* values: KRTAP3-1, KRTAP3-2 < 2×10^{-16}, KRTAP2-2 = 0.03, KRTAP4-12 = 3×10^{-16} **. h,** Comparing total clonal expression of four genes in the KRTAP2 sub-cluster (x-axis) and 8 genes in the KRTAP3, 4 sub-cluster (y-axis). Sample size as in **c**. **i,** Comparing clonal expression of KRTAP-associated genes (x-axis) and the EpCAM module (y-axis). **j,** Similar to **e**, for KRTAP-associated genes. **k,** Distribution of HB region expression (genes E-L in **a**) in clones and single cells. **l,** Distribution of KRTAP region expression (genes E-O in **f**) in clones and single cells. Vertical dashed lines in **k-l** indicate the overall normalized expression in clones (black) and cells (red).